Editorial de la Universidad Tecnológica Nacional

Maestría en Transporte y Logística

Metodología para la estimación del TMDA
(Tránsito Medio Diario Anual)
mediante conteos de tránsito esporádicos en la zona central de la República Argentina.

Autor Ing. José Julián Rivera

Director MBA Ing. Edgardo Alberto Masciarelli

Editorial de la Universidad Tecnológica Nacional – edUTecNe
http://www.edutecne.utn.edu.ar
edutecne@utn.edu.ar

Universidad Tecnológica Nacional
Facultad Regional Santa Fe

Maestría en Transporte y Logística

Metodología para la estimación del TMDA (Tránsito Medio Diario Anual) mediante conteos de tránsito esporádicos en la zona central de la República Argentina.

por Ing. José Julián Rivera

Director MBA Ing. Edgardo Alberto Masciarelli

Jurado de Tesis
Ing. Roberto Cruz
Dr. Omar Chiottti
Ms. Ing. Graciela Berardo

Febrero 2007

DEDICATORIA DEL AUTOR

Esta tesis va dedicada a mis familiares, amigos y a todas aquellas personas que me han dado continuamente fuerzas para su concreción. Especialmente a mi esposa María Eugenia, por haber sabido disimular tantas ausencias durante el cursado de la maestría, a mis padres por inculcarme constantemente la cultura del estudio y a Gerardo Botasso, y demás compañeros de trabajo, por su apoyo y hacer posibles los tiempos necesarios durante la cursada de la maestría y desarrollo de la tesis.

INDICE

Resumen

El *TMDA* (Tránsito Medio Diario Anual) es una forma de valoración del volumen de tránsito empleada en un sinnúmero de aplicaciones viales y de estudios relacionados. Por definición su obtención implica que deben medirse los volúmenes pasantes por la vía en análisis durante todo el año calendario, lo cual no es factible en muchos de los estudios que requieren su cuantificación.

Para subsanar esta problemática, se suele adoptar lo que puede denominarse como la "metodología clásica", que contempla la obtención del *TMDA* mediante el uso complementado de conteos esporádicos sobre la vía en análisis con series históricas de vías cercanas de similares características. De esta forma se incluye como requisito principal que su aplicación sea efectuada por un profesional capacitado en la materia, como único medio para reducir la subjetividad que implica el decidir sobre la validez o no del empleo de una serie, el cual generalmente no se encuentra disponible (o incluso no resulta justificable) en muchas de las aplicaciones del *TMDA*.

El presente estudio atiende a esta problemática mediante el desarrollo de una metodología objetiva, que permite, mediante la valoración de parámetros medibles de las condiciones de borde de la vía, la obtención de curvas de corrección para los conteos esporádicos para su extrapolación al *TMDA*, con aplicabilidad en la región conformada por las provincias argentinas de Buenos Aires, Córdoba, Santa Fe, Entre Ríos y La Pampa.

Para esto la metodología emplea modelos obtenidos por regresión de los datos históricos recolectados en el área en estudio. Razón por la cual se genera un fuerte análisis de manejo estadístico y de la modelización por regresión, que sirve de base a la aplicación de los datos relevados hasta la obtención de los modelos finales.

Como último paso se realiza el análisis de validación de la metodología mediante su aplicación en diversas tipologías de vías y comparación de resultados con los valores reales y los obtenidos mediante la metodología clásica, y se analiza el empleo de técnicas alternativas para el desarrollo de los modelos, generándose también en este sentido el análisis comparativo. Los resultados de ambos análisis permiten concluir que mediante la metodología desarrollada pueden obtenerse en su área de aplicación y en forma objetiva valores de *TMDA* confiables.

Reconocimientos del autor

Quiero expresar mi reconocimiento a las personas e instituciones que aportaron desinteresadamente la bibliografía de consulta, los datos de tránsito y las opiniones técnicas para la elaboración del presente estudio, sin los cuales su concreción seguramente no hubiera sido posible:

- Al Director de la tesis, Ing. Edgardo Masciarelli, del ISIT de la Universidad Nacional de Córdoba.
- A la Dra. Ana Rosa Timoschiuk, de la UTN Facultad Regional Santa Fe.
- Al Ing. Pablo Arranz, del ISIT de la Universidad Nacional de Córdoba.
- A los profesionales de la Dirección de Señalización Luminosa del Gobierno de la Ciudad de Buenos Aires.
- Al Ing. Ricardo Montes de Oca, de la concesionaria AUFE.
- A la Sra. Magali Fernández, de la concesionaria Autopistas del Oeste.
- A la Lic. Victoria Fasano, especialista en técnicas de regresión matemática.
- Al Ing. Daniel Bortolin, consultor particular especialista en tránsito.
- A la Arq. Alejandra Barczuk, de la concesionaria Autopistas del Sol.
- Al Ing. Sergio Peirone, de la UTN Facultad Regional Rafaela.
- Al Ing. Marcelo David, de la Dirección de Vialidad de la Provincia de Santa Fe.
- A los profesionales de la Auditoria General de la Nación.

Lista de tablas

Listado de figuras

4.34. Gráfico de caja y bigotes para los intervalos de confianza de los coeficientes diarios por regresión

4.35. Gráfico de caja y bigotes para los intervalos de confianza de los coeficientes mensuales por valores medios

4.36. Gráfico de caja y bigotes para los intervalos de confianza de los coeficientes mensuales por regresión

a.1. Nube de puntos que ajusta bien a la recta

a.2. Nube de puntos para la cual el ajuste lineal no resulta adecuado

a.3. Nube de puntos sin relación lineal entre variables

a.4. Nube de puntos con claros indicios de heterocedasticidad

a.5. Nube de puntos con datos atípicos

a.6. Nube de puntos con posibilidad de inclusión de variable binaria

a.7. Modelo $Y = \exp(\alpha_0 + \alpha_1 x)$

a.8. Modelo $Y = 1/(\alpha_0 + \alpha_1 X)$

a.9. Modelo $Y = \alpha_0 + \alpha_1 \lg X$

a.10. Modelo $Y = \alpha_0 X^{\alpha}{}_1$

a.11. Modelo $Y = \alpha_0 X^{-\alpha}{}_1$

a.12. Modelo $Y = \exp X$

a.13. Gráfico de dispersión matricial

a.14. Gráfico de residuos frente a una variable explicativa

a.15. Modelo heterocedástico

a.16. Gráfico de residuos frente a una variable omitida

a.17. Gráfico de residuos frente a las predicciones

a.18. Gráfico de residuos frente a una variable de clasificación omitida

a.19. Gráfico entre las variables X_1 y X_2

a.20. Gráfico de dos variables regresoras

a.21. Función de Huber

b.1. Existencia de dependencia funcional lineal

b.2. Relación lineal entre variables pequeña

b.3. Dependencia entre variables no lineal

b.4. Ajuste razonable a una recta

b.5. Fuerte dependencia lineal negativa

b.6. Nube con tres observaciones extremas (outliers).

b.7. Influencia del punto A.

b.8. Influencia del punto B.

b.9. Influencia del punto C.

b.10. Efecto de omitir un atributo

b.11. Efecto al omitir un atributo

b.12. Efecto al omitir un atributo

Capítulo 1 – Introducción

1.1. Enfoque del estudio

Los análisis que involucran al tránsito automotor nos plantean generalmente el requisito básico de conocer de manera ajustada su magnitud, o lo que en su forma técnica conocemos como *TMDA* (Tránsito Medio Diario Anual), es decir el volumen promedio diario de tránsito registrado a lo largo de un año calendario sobre una sección de un camino o arteria, concepto sobre el que volvemos más adelante. La siguiente es una muestra de su amplia variedad de aplicaciones.

"...Planeamiento

- Clasificación sistemática de redes de caminos
- Estimación de los cambios anuales en los volúmenes de tránsito
- Modelos de asignación y distribución de tránsito
- Desarrollo de programas de mantenimiento, mejoras y prioridades
- Análisis económicos
- Estimaciones de la calidad del aire
- Estimaciones del consumo de combustibles

Proyecto

- Aplicación a normas de proyecto geométrico
- Requerimientos de nuevos caminos
- Análisis estructural de superficies de rodamiento

Ingeniería de tránsito

- Análisis de capacidad y niveles de servicio en todo tipo de vialidades
- Caracterización de flujos vehiculares
- Necesidad de dispositivos para el control del tránsito
- Estudio de estacionamientos

Logística

- Análisis de recorridos óptimos

Estudio de mercado de combustibles, lubricantes, etc.

Seguridad

Cálculo de índices de accidentes y mortalidad

Evaluación de mejoras por seguridad

Investigación

Nuevas metodologías sobre capacidad

Análisis e investigación de los accidentes y la seguridad

Estudio sobre ayudas, programas o dispositivos para el cumplimiento de las normas de tránsito

Estudios de antes y después

Estudios sobre medio ambiente y la energía

Usos comerciales

Hoteles y restaurantes

Urbanismo

Autoservicios

Actividades recreacionales y deportivas..." [1]

No sólo son numerosos los campos de aplicación del parámetro *TMDA*, sino que en cada uno de ellos puede resultar de una gran importancia en la toma de decisiones, junto con otras características del tránsito. Como ejemplo podemos considerar que "...el diseño de un camino, se encontrará preponderantemente influenciado por dos factores; la *configuración del terreno* que debe atravesar y las *modalidades y exigencias del tránsito* que debe soportar... Será un buen diseño el que, con un costo anual mínimo, tenga en cuenta simultáneamente ambos factores, en la medida de su importancia... Cuando el tránsito es reducido, el diseño del camino deberá estar influenciado por la configuración del terreno, en cambio cuando el tránsito es intenso, las necesidades de los usuarios y las características del tránsito deberán ser los factores preponderantes... El volumen, composición, distribución, velocidad del tránsito... determinan diversas magnitudes del diseño geométrico de un camino, tales como radios y peraltes de curvas horizontales, parámetros de curvas verticales, pendientes, anchos de calzada, etc..."[2]

[1] "Ingeniería de tránsito, fundamentos y aplicaciones", R. Cal y Mayor, J. Cárdenas, Alfaomega 7°ed., México 1995.

[2] "Tránsito medio diario anual 98/99", División Tránsito de la Dirección Nacional de Vialidad, Argentina 2000.

No obstante las amplias posibilidades de aplicación, la determinación y empleo del *TMDA*, y demás parámetros asociados, en Argentina y Latinoamérica no están aun generalizados, tal cual lo advierte el Banco Mundial cuando asegura que "… Aunque el rápido desarrollo de la tecnología ha reducido el costo de las modernas técnicas de gestión de tránsito, muchas ciudades están todavía pobremente organizadas y tienen personal inadecuado para hacer uso efectivo de ellas. Tanto la asistencia técnica como las inversiones son capaces de generar elevados retornos en este campo, siempre y cuando se traten los problemas fundamentales de recursos humanos e institucionales…"[3].

Incluso a nivel nacional, el CIMOP (Consejo Interprovincial de Ministros de Obras Públicas) afirma "...La red vial troncal debe desarrollarse con las redes provinciales y locales (terciarias) de modo tal que tengamos un sistema vial jerarquizado que cubra el territorio y potencie la accesibilidad a las diferentes regiones y jerarquías del sistema de asentamientos humanos. Para su diseño se deben tener en cuenta tres criterios:

• *TMDA*. Flujos actuales o potenciales en la red.

• Necesidad de potenciar la accesibilidad y conectividad entre los asentamientos humanos, privilegiando la conectividad entre las metrópolis regionales y la accesibilidad a las ciudades intermedias.

• Promoción de la integración y la ordenación territorial..."[4]

Dándose a entender que una de las razones de la carencia de ese sistema vial adecuado ha sido justamente el no contar con el conocimiento y empleo acabado de los *TMDA* involucrados.

Tal vez la causa de esta falta de conocimiento, esta traba en la divulgación de su correcto empleo, podamos deducirla de cierta característica fundamental del *TMDA*, la cual es que en este parámetro se promedian volúmenes que son generados en gran parte por actividades no constantes, o que incluso se realizan intermitentemente. Pudiéndose citar entre éstas el estudio, trabajo, vacaciones, esparcimiento, etc. Por esto, "...el tránsito debe ser considerado como un factor dinámico, siendo solamente

[3] "Ciudades en movimiento", Banco Mundial, TWU-44, 2002.
[4] "Una visión estratégica del Transporte en la Argentina", CIMOP, Argentina 2003.

su valor preciso para el período de duración de sus mediciones. Sin embargo, debido a que sus variaciones son generalmente rítmicas y repetitivas, es importante tener un conocimiento de sus características…"[5].

Ya que existe variabilidad en las necesidades que originan el movimiento de las personas (tránsito), existe la necesidad de realizar conteos continuos a lo largo de todo el año calendario, para así arribar al *TMDA* buscado. Siendo justamente esta la razón a nuestro entender que diferencia este parámetro de otros de obtención más inmediata (ancho de calzada, pendientes, velocidades de circulación, etc.).

El conteo continuo a lo largo de un ciclo podemos efectuarlo en innumerables análisis relacionados con estudios de tránsito, transporte o logística de gran envergadura. Pero en tareas de tipo tácticas y operativas (de mediano y corto plazo), en aquéllas en que debemos generar soluciones inmediatas con implicancias en el largo plazo o para las cuales no contamos con los suficientes recursos (equipamiento, personal, tiempo y dinero) se torna imposible. Por tal razón, los profesionales relacionados con la temática, suelen recurrir en estos casos a conteos esporádicos de tránsito para su posterior expansión por medio de registros históricos. Así, "...a nivel de planificación Argentina dispone de información sistemática de la red de contadores permanentes de la DNV (Dirección Nacional de Vialidad), pero para estudios específicos deben programarse relevamientos de tránsito que generalmente en una semana o menos puedan dar una aceptable estimación..."[6]

De esta manera, la ingeniería de tránsito ha tendido a la implementación de los denominados "censos de cobertura", que permiten la extrapolación de las mediciones esporádicas efectuadas en una sección por medio de las curvas establecidas por censos continuos en puntos cercanos al lugar en estudio. La aplicación se efectúa de la siguiente manera.

"...Los censos de tránsito caminero consisten en el relevamiento del volumen de tránsito en los tramos de la red vial en ciertos y determinados puntos de la misma... El objeto de estos conteos es el estimar el Tránsito Medio Diario Anual (*TMDA*) en cada uno de los puntos en que se realicen. Los conteos deberán ser efectuados con

[5] "Ingeniería de transporte", W. Hay, Limusa, México 1998.

[6] "Caracterización de errores de muestreo en censos de volumen y composición", M. Herz, J. Galárraga, M. Maldonado, XIV Congreso Argentino de Vialidad y Tránsito, Argentina 2005.

clasificación según los tipos representativos de vehículos. En general los tipos de vehículos con que se clasifica son automóviles, utilitarios de cuatro ruedas, ómnibus, camiones simples, camiones con semiacoplado o semiremolque. Esta clasificación puede variar según las necesidades, aumentando el número de clases o disminuyéndola. La duración de los conteos estará entre 1 y 7 días, durante las 24 horas. Cuando el conteo es por día el *TMDA* se calcula de la siguiente manera.

$$TMDA = TC \cdot f_d \cdot fe_m \quad (1.1)$$

Siendo:

TC = Tránsito contado a lo largo del día.

f_d = Factor de corrección por el día de realización del conteo.

fe_m = Factor de corrección estacional correspondiente al mes *m* en que se realizó el conteo.

Los factores de corrección diarios se determinan a partir de información obtenida de los contadores permanentes más próximos al sitio y tienen por objeto estimar el promedio diario semanal a partir de conteos de menor duración. Si el conteo es de 7 días no es necesario determinar este factor. En ese caso se estima el Tránsito Medio Diario Semanal (*TMDS*) y el *TMDA* de la siguiente forma, partiendo de los Tránsitos Contados (TC) en cada uno de los días de una semana:

$$TMDS = \frac{1}{7}\sum TC \quad (1.2)$$

y luego:

$$TMDA = TMDS \cdot fe_m \quad (1.3)$$

Cuando el conteo abarque menos de 7 días el *TMDS* se calculará ponderando los promedios de día hábil y fin de semana.

En ciertas ocasiones se suelen realizar conteos de tres días, incluyendo un día hábil (viernes o lunes), un sábado y un domingo, estimándose el *TMDS* de la siguiente manera:

$$TMDS = 1/7 \cdot (5 \cdot TDH + TDS + TDD) \quad (1.4)$$

siendo:

TDH = Tránsito medido en el día hábil (viernes o lunes)

TDS = Tránsito medido durante el día sábado

TDD = Tránsito medido durante el día domingo

Los factores de corrección estacional se obtienen del organismo vial con jurisdicción en el tramo, o bien se calculan a partir de información de contadores permanentes próximos al lugar. Se deberá tener en cuenta que la DNV (Dirección Nacional de Vialidad) determina los factores de corrección estacional solamente para días hábiles, de manera que el TMC_j deberá ser determinado solamente con días hábiles..." [7].

Esta técnica puede ser bien empleada cuando el análisis es dirigido por un especialista de tránsito, que puede interpretar la validez de relacionar un punto con el otro (en función de la similitud en las necesidades cubiertas por el tramo de vía), profesional generalmente no disponible en estudios que requieren la valoración del *TMDA* para implementaciones que poco tienen que ver con la especialidad (estudios de mercado, logística, accidentología, etc.), sumándose a esta complicación el hecho no menor de que en la práctica sólo se cuenta con este tipo de conteos continuos en zonas urbanas muy desarrolladas o vías rurales de importancia, quedando sin cobertura la inmensa mayoría de las ciudades y rutas secundarias y terciarias que constituyen la red vial de la región.

Cuando la expansión de la muestra es realizada por profesionales que no guardan relación con la ingeniería de tránsito o en función de series poco adecuadas a las circunstancias puntuales del lugar en estudio, se agrega un término de incertidumbre, llegándose a desvirtuar por completo la aplicación posterior de cálculos que sí están sostenidos en datos certeros, obteniéndose en conjunto valores de confiabilidad bajos.

El problema de la confiabilidad en los resultados no se observa solamente en la Argentina. Por ejemplo la AASHTO (American Association of State Highway and Transportation Officials), de reconocido prestigio en el ambiente vial, ha previsto para salvar este problema de la falta de datos en su metodología del año 2002 lo siguiente.

"…El procedimiento de diseño de pavimentos requiere de datos tales como volúmenes de tránsito y espectros de carga por cada tipo de eje... Sin embargo, es necesario recordar que muchas veces las agencias no cuentan con los recursos

[7] "Planeamiento del transporte", L. Girardotti, Fac. de Ing. UBA, Argentina 2003.

suficientes para recolectar datos de tránsito. Por esto, el método define tres niveles claramente determinados de entrada de datos, basados en la cantidad de información disponible. Estos niveles representan la calidad de la estimación que el diseñador puede efectuar de las características futuras del tránsito en la ruta a diseñar…
El alto nivel de exactitud en los datos y en las proyecciones de las cargas de tránsito aplicadas trae como consecuencia pavimentos mucho más confiables, a diferencia de aquellas rutas diseñadas con información de cargas y volúmenes sin un alto nivel de exactitud…" [8]

A partir de lo aquí volcado, hemos buscado poner en relieve ciertas dificultades que presenta la aplicación de los censos de cobertura y la posibilidad de inclusión de mejores metodologías de estimación, pues en la mayoría de las técnicas de aplicación del *TMDA* no se cuenta con refinamientos como el expuesto, de generar diversos niveles de análisis en función de la precisión con que el tránsito ha sido analizado, llevándose indefectiblemente, como ya se mencionara, a bajas confiabilidades.

Toda esta situación ha sido detectada con anterioridad, por eso a nivel mundial existen estudios tendientes a establecer los parámetros de comportamiento del tránsito en busca de calcular el *TMDA* mediante la utilización de conteos esporádicos. Como ejemplo podemos mencionar las curvas de Petroff y Blensly, destacando su particular antigüedad y restricción geográfica.
Es justamente la restricción geográfica lo que hace que no exista un modelo de aplicación generalizada y mucho menos para la región central de la Argentina, lugar propuesto para la realización del estudio. Por esto vale recordar lo enunciado en una de las publicaciones más consultadas a nivel mundial por los especialistas en tránsito, el Manual de Capacidad 2000 de la TRB (Transportation Research Board), que en su capítulo de "Características del tránsito vehicular y factores humanos" sostiene que "… las variables estacionales en la demanda de tránsito reflejan la actividad social y económica del área servida por un camino. Los datos volcados en esta publicación son típicos de la zona estudiada. Sin embargo, estos parámetros varían en función de los hábitos de viaje locales y el medioambiente, los ejemplos no pueden ser usados como un sustituto para la obtención de datos locales…"[9].

[8] "Vialidad II", C. Wahr, Universidad Técnica Federico Santa María, Chile 2003.
[9] "Highway Capacity Manual 2000", Transportation Research Board, National Research Council, EEUU 2000.

1.2. Objetivos, etapas y alcances del trabajo

Por lo que expusimos en el punto anterior, planteamos el presente trabajo, que busca facilitar el empleo de extrapolaciones de los conteos esporádicos de tránsito al *TMDA*, fundadas en:

- parámetros medibles,
- comportamientos conocidos de forma estadística,
- y la posibilidad de aplicación en una amplia región relativamente homogénea, como lo es la zona central de la República Argentina, conformada por las provincias de Buenos Aires, Santa Fe, Córdoba, Entre Ríos y La Pampa.

Intentamos con el estudio generar una herramienta simplificada, constituida por una metodología de relevamiento y algoritmos de aplicación, sostenida en un análisis estadístico de regresión, que pueda ser utilizada como alternativa o reemplazo de los actuales métodos existentes, sin necesidad de extrapolaciones subjetivas generadas por la falta de datos o por no poseer el conocimiento acabado del lugar en estudio. Este planteo guarda concordancia con la línea actual de pensamiento para la región, ya que "...la velocidad de cambio y la inestabilidad económica son a menudo más altas en países en desarrollo como los nuestros que en Europa o EEUU, así, no solo el futuro es más difícil de predecir, sino que se ha pensado que el estilo de países en desarrollo debe cambiarse radicalmente, y para esto se necesitan modelos que debieran:

- Ser fácil de utilizar y requerir pocos recursos escasos.
- Usar información de bajo costo (que sea fácil de recolectar o que esté disponible de otras fuentes).
- Permitir el uso de información histórica, de modo que ésta no sea desechada..."[10]

Para poder llegar a este objetivo hemos planteado los siguientes lineamientos generales, que delimitan las etapas del estudio:

[10] "Modelos de demanda de transporte", Juan de Dios Ortúzar, Universidad Católica de Chile, Alfaomega, Chile 2000.

- El desarrollo se basa en el análisis de datos de tránsito y sus características recabados en diversas vías de la zona en estudio, combinados con datos adicionales del entorno, referidos a lo geográfico, económico y social.
- Para esto recolectamos los datos provenientes de fuentes del más amplio espectro, fijando para ello un horizonte entre el año 1993 y 2003.
- Los datos pasan a conformar una base de datos general homogénea, sobre la que realizamos los análisis estadísticos necesarios para la conformación de bases de datos reducidas, conteniendo las variables explicativas de significancia. Aquí es donde se filtran y adaptan los datos recabados en función de los requisitos particulares del estudio.
- En función de las bases de datos reducidas, se determinan los algoritmos que conforman el modelo por medio de regresión matemática, detectando su ajuste.
- Finalizamos el estudio comparando la aplicación de la metodología desarrollada con otras alternativas, detectando las potenciales ventajas y desventajas comparativas.

Los alcances pueden deducirse entonces de lo ya expresado, resultando:

- Alcance temporal de los datos analizados; los coeficientes son determinados por las series de datos recolectados en los diez ciclos que van desde 1993 a 2003.
- Alcance geográfico; las vías que conforman las redes viales de las provincias de Buenos Aires, Córdoba, La Pampa, Santa Fe y Entre Ríos.
- Alcance metodológico; la obtención de los modelos se efectúa por medio de la aplicación de técnicas de estadística y regresión.

Con estos lineamientos desarrollamos la metodología en sus diversas partes y analizamos la validación y discusión de la misma, sobre todo en lo que hace a su análisis comparativo con la metodología clásica y la aproximación de los valores obtenidos a la realidad.

La conclusión final de todo el análisis generado es que la metodología resulta una herramienta de cálculo del *TMDA* aplicable en el área en estudio, con la que se obtiene sin necesidad de subjetividades resultados confiables.

Capítulo 2 - Marco teórico y descripción metodológica

2.1. Marco teórico del estudio

2.1.1. Otros conceptos del tránsito y su medición

Como inicio del análisis veamos más detalladamente algunas de las características de lo que denominamos tránsito, que consideramos de interés para su desarrollo.
Primeramente nos parece interesante diferenciar entre algunos conceptos que pueden resultar similares, ya que "...el volumen y el flujo son dos medidas que cuantifican la cantidad de tránsito pasante por un punto de un camino durante un intervalo dado de tiempo. Estos términos se definen como:

- Volumen: el número total de vehículos que pasan por una sección dada de un camino durante un intervalo de tiempo dado; los volúmenes pueden estar expresados en año, día, hora o periodos menores.
- Flujo: es el equivalente horario de los vehículos que pasan por una sección de camino dada durante un intervalo dado menor de una hora, usualmente 15 minutos..."[11]

Es claro que el análisis del trabajo se centraliza en el volumen de tránsito, pero como vemos éste puede expresarse en diversas unidades de tiempo en función de los requisitos de la metodología de aplicación del parámetro.
Entre estas formas de expresión surge el concepto de *TMDA*, pues "...el *tránsito medio diario anual* es una medida fundamental del tránsito y en el sentido estricto se define como el volumen de tránsito total anual dividido por el número de días del año..." [12]

[11] "Highway Capacity Manual 2000", Transportation Research Board, National Research Council, EEUU 2000.
[12] "Tránsito medio diario anual 98/99", División Tránsito de la Dirección Nacional de Vialidad, Argentina 2000.

Por otro lado, para analizar las características puntuales del tránsito, debemos comprender que éste es una expresión del transporte automotor carretero, y que por lo tanto arrastra características del concepto general de transporte, algunas de las cuales nos resultan de interés. La que sigue es una forma de enumerar a estas "...características generales del transporte:

a) El transporte es un bien altamente cualitativo y diferenciado: existen viajes con distintos propósitos, a diferentes horas del día, por diversos medios, para variados tipos de carga. Esto implica una enorme cantidad de factores difíciles de analizar y cuantificar (por problemas de seguridad o comodidad, por ejemplo).
b) La demanda de transporte es derivada: los viajes se producen por la necesidad de llevar a cabo ciertas actividades (ej: trabajo, compras, recreación) en el destino.
c) La demanda de transporte está localizada en el espacio.
d) La demanda de transporte es eminentemente dinámica..."[13]

El tránsito debe entonces su variabilidad en el tiempo a su condición de dependiente de una demanda derivada particularizada en propósitos y espacios. Por esto a nivel transporte automotor tenemos que "...las variaciones estacionales en la demanda del tráfico, reflejan la actividad social y económica del área periférica servida por una vía, en donde generalmente se observa que:

- Las variaciones mensuales son mayores en rutas rurales que en rutas urbanas.
- Las variaciones son mayores en rutas rurales que sirven principalmente a tránsito recreacional que en aquellas que sirven principalmente a tránsito comercial.
- Los parámetros de tránsito diario varían por mes del año más severamente en rutas recreacionales.

Estas observaciones llevan a la conclusión que los viajes cotidianos y relacionados con negocios ocurren en forma más uniforme que el tránsito recreacional, que genera grandes variaciones en los volúmenes...

Las variaciones de volumen por día de la semana también se relacionan con el tipo de vía en la que las observaciones son realizadas... los volúmenes de fin de semana

[13] "Modelos de demanda de transporte", Juan de Dios Ortúzar, Universidad Católica de Chile, Alfaomega, Chile 2000.

son menores que en los días laborales para caminos que sirven predominantemente a viajes de negocios, como en vías urbanas... en comparación, los picos de tránsito ocurren en los fines de semana en la mayoría de las rutas rurales y recreacionales... de todos modos, la magnitud de la variación diaria es mayor para rutas recreacionales y menor en rutas urbanas con viajes cotidianos..."[14]

Vemos así una gama de condicionantes que inciden en el tránsito, lo que se complementa con el hecho de que "...la variación del tránsito a través del día, los días de la semana y los meses del año no sigue leyes físicas sino comportamientos humanos, pero con técnicas estadísticas se puede intentar caracterizar los patrones de variación y mejorar nuestro conocimiento para realizar estimaciones..."[15]

Pero, ¿Cómo alcanzamos este conocimiento? La respuesta reside en la oportunidad de medir, ya que "...los aforos continuos proporcionan información muy importante con respecto a los patrones de variación horaria, diaria, periódica o anual del volumen de tránsito. El tránsito tiende a tener variaciones cíclicas predecibles, por lo que a través de una clasificación adecuada de las vialidades y los aforos, es posible establecer el patrón básico de variación del volumen de tránsito para cada tipo de carretera o calle. Más aun, si bien los valores de los volúmenes específicos para determinados periodos (minutos, horas, días) pueden llegar a ser bastante diferentes de un lugar a otro, su proporción en el tiempo con respecto a los totales o promedios, es en muchos casos constante o consistente. Estas propiedades, son las que sustentan el uso de factores de expansión y ajuste en la estimación de volúmenes para otros lugares y otros periodos..."[16]

Complementariamente, "...los censos en estaciones permanentes se realizan con contadores automáticos que operan los 365 días del año en forma continua, y registran en forma horaria la cantidad de vehículos que circulan por el lugar donde se hallan emplazados... Estas estaciones, además de determinar el *valor verdadero* del *TMDA* en el lugar de emplazamiento, tienen dos objetivos:

[14] "Highway Capacity Manual 2000", Transportation Research Board, National Research Council, EEUU 2000.
[15] "Caracterización de errores de muestreo en censos de volumen y composición", M. Herz, J. Galárraga, M. Maldonado, XIV Congreso Argentino de Vialidad y Tránsito, Argentina 2005.
[16] "Ingeniería de Tránsito", R. Cal y Mayor, J. Cárdenas, Alfaomega, Méjico 1995.

- Determinar los patrones de flujo de tránsito (variaciones estacionales, diarias, horarias, etc.)
- Permitir la elaboración de la serie histórica para así determinar la *tendencia en el uso del camino en el largo plazo*

...Los censos de cobertura se realizan en tramos en los que no se efectúan censos permanentes, instalando durante 48 horas, en días hábiles, contadores automáticos de tránsito con registro horario...

...Los censos de clasificación se realizan en estaciones predeterminadas en días hábiles durante 24 horas consecutivas. En estos censos se clasifican manualmente los vehículos según las siguientes siluetas: automóviles, pick-up, ómnibus, camiones simples, camiones con acoplado y semiremolques..."[17]

Como vemos, estas formas expuestas de censos y de clasificación son propias de la *DNV*, ya que otras formas de organización pueden ser empleadas atendiendo a la "...distribución del tránsito por tipo de viaje:

- Tránsito metropolitano comercial; relacionado con viajes de corta distancia en días hábiles, por motivos de trabajo, estudio, comercio zonal, etc.
- Tránsito metropolitano turístico; relacionado con viajes de corta distancia en fines de semana y feriados.
- Tránsito interurbano comercial; relacionado con viajes de media y larga distancia, por todo motivo, durante todo el año, excepto turismo de verano.
- Tránsito interurbano turístico; relacionado con los picos de enero y febrero en rutas de zonas no turísticas..."[18]

2.1.2. El análisis estadístico del tránsito

Dijimos que los censos permiten establecer los patrones y analizar la serie histórica de los datos, entramos de este modo al análisis estadístico del problema y comenzamos a considerar lo que se conoce como serie de tiempo.

"... Se tiene una serie de tiempo cuando se recopila información sobre ciertas variables agregadas (población, ingreso, flujos vehiculares) en distintos instantes de

[17] "Tránsito medio diario anual 98/99", División Tránsito de la Dirección Nacional de Vialidad, Argentina 2000.

[18] "Red de Acceso a Córdoba; Capacidad y Nivel de Servicio para el tránsito actual y su predicción", Instituto Superior de Ingeniería de Transporte, Universidad Nacional de Córdoba, Argentina 1996.

tiempo. Esta información tiene la ventaja de que suele estar institucionalizada, por lo que los datos ampliamente disponibles y las series históricas tienen una longitud interesante. Un requisito importante es que las series sean lo más completas posible, por lo que, previo a su utilización, deben ser "llenadas" con métodos adecuados..."[19]

Estas series pueden ser empleadas en métodos en busca de conclusiones como a las que intentamos llegar con este estudio. Estos "...métodos de series de tiempo son técnicas estadísticas que hacen uso de datos históricos acumulados en un periodo de tiempo. Asumen que lo ocurrido en el pasado continuará ocurriendo en el futuro. Como su nombre sugiere, estos métodos relacionan el pronóstico a un solo momento..."[20]

Con el empleo de las series buscamos la identificación de estos parámetros repetitivos de comportamiento del tránsito, que es una forma de conocer la realidad sobre la vía. Realidad que una vez conocida debe de algún modo poder ser modelada matemáticamente. Hallar ese modelo es la finalidad de la parte central de este estudio.

Pero no buscamos un modelo cualquiera, sino, claro está, uno al cual ingresando con ciertos datos nos permita la obtención del *TMDA*, y que de estos datos el principal sea el volumen de tránsito contado. Para establecer límites a esta búsqueda fijamos a este volumen contado en el nivel diario. Es decir que de ahora en más nuestro dato de tránsito viene expresado en vehículos por día, lo cual es relativamente fácil de obtener hasta incluso con conteos manuales, eliminando de esta forma del estudio el análisis de los volúmenes horarios.

Por ser el tránsito el dato principal de entrada, hablamos de un modelo basado en él.

"...Los modelos basados en conteos de tránsito parecen una idea particularmente interesante, ya que:

- Los conteos son relativamente baratos de obtener (se recolectan con varios usos posibles en mente: diseño de intersecciones, manutención de caminos, etc.).

[19] "Modelos de demanda de transporte", Juan de Dios Ortúzar, Universidad Católica de Chile, Alfaomega, Chile 2000.

[20] "Operations management. Focusing on quality and competitiveness", R. Russel, B. Taylor, Prentice Hall, EEUU 2003.

- Hoy existen técnicas y equipos modernos muy eficientes para contar en forma automática y luego procesar, en forma también automática, la información.
- Contar vehículos, es más sencillo que realizar encuestas (donde hay que realizar entrevistas, completar cuestionarios y codificar respuestas).
- Algunas operaciones de conteo se realizan como parte de la operación normal de organismos de planificación y operación (ej. plazas de peaje).
- La gran mayoría de las actividades de conteo no requiere demorar el tráfico..."[21]

2.1.3. La modelización del tránsito elegida

Existen matemáticamente diversas formas de llegar al modelo buscado. Para este estudio hemos decidido encarar el análisis mediante las técnicas de regresión matemática, sin pretender con esto asegurar que sea la forma óptima de hacerlo, sino una más, tan valedera como cualquiera de las demás opciones existentes. Más adelante, en el capítulo de validación y discusión, vemos la aplicación de otras técnicas y realizamos el análisis comparativo en busca de elementos que nos permitan ratificar esta afirmación.

¿En que se basan los análisis de regresión?

"...Cuando se desea relacionar un conjunto de observaciones acerca del resultado de un experimento (*Y*), con la cantidad que se agregue de un cierto ingrediente (*X*), es natural utilizar técnicas de ajuste –como mínimos cuadrados ordinarios- que entreguen una función que permita interpolar resultados dentro del rango de los datos con el menor error posible. Si no sólo interesa un ajuste mecánico de una curva, sino que la capacidad de realizar inferencias acerca de la población de la cual proviene una muestra, se entra al área de la modelación matemática e interesan conceptos como intervalos de confianza y prueba de hipótesis. El modelo de regresión lineal, sobre la base de una serie de hipótesis provee este tipo de herramienta y es

[21] "Modelos de demanda de transporte", Juan de Dios Ortúzar, Universidad Católica de Chile, Alfaomega, Chile 2000.

consistente con la solución de mínimos cuadrados ordinarios, por lo que posee enorme popularidad tanto en ciencias exactas como en ciencias sociales..." [22]
Como vemos se mantiene un paralelismo con lo que veníamos diciendo, ya que planteamos un modelo al que ingresemos en un principio con datos de tránsito y de variables de entorno de la vía (ingredientes *X*) para llegar a un resultado de *TMDA* (*Y*).

"...Cuando deseamos estimar, basados en datos de una muestra, el valor de una variable *Y* correspondiente a un valor dado de la variable *X*, podemos hacerlo mediante una curva de mínimos cuadrados que ajuste los datos. La curva resultante se llama una *curva de regresión de Y sobre X*...
Si la variable independiente *X* esta relacionada con el tiempo, los datos muestran los valores de *Y* en varios instantes que ordenados en el tiempo se llaman *series de tiempo*. La recta o curva de regresión de *Y* sobre *X* en este caso se suele llamar *curva de tendencia*..."[23]

Llegamos de esta forma a delinear cual es el marco teórico para nuestro estudio, ya que nos hemos detenido en los conceptos que hacen a la noción del tránsito (algunos de los cuales se profundizan más adelante) y hemos recorrido el camino que nos lleva teóricamente a convalidar la idea de modelar la situación mediante regresión matemática.
Son justamente las técnicas de regresión, en su descripción como metodología, la temática de la segunda parte de este capítulo, previo a su empleo en el análisis de los datos.

2.2. Descripción metodológica

Las técnicas de regresión matemática pertenecen al área disciplinar de la estadística. "...La estadística es la única herramienta que permite dar luz y obtener resultados en

[22] "Modelos de demanda de transporte", Juan de Dios Ortúzar, Universidad Católica de Chile, Alfaomega, Chile 2000.
[23] "Estadística", M. Spiegel, Mc Graw Hill, EEUU 1988.

cualquier tipo de estudio, cuyos movimientos y relaciones, por su variabilidad intrínseca, no puedan ser abordadas desde la perspectiva de las leyes deterministas. Esta se ocupa de recoger, clasificar, resumir, hallar regularidades y analizar los *datos*, siempre y cuando la variabilidad e *incertidumbre* sea una causa intrínseca de los mismos; así como de realizar *inferencias* a partir de ellos, con la finalidad de ayudar a la toma de *decisiones* y en su caso formular *predicciones*...
La estadística es inferencial cuando el objetivo del estudio es derivar las conclusiones obtenidas a un conjunto de datos más amplio..."[24]
Comencemos entonces con su análisis.

Cuando se estudia la relación entre una variable de interés, variable respuesta o variable dependiente (Y), y un conjunto de variables regresoras, variables explicativas o independientes (X_1, X_2, ..., X_k), pueden darse las siguientes situaciones:

- Existe una relación funcional entre ellas, en el sentido de que el conocimiento de las variables regresoras determina completamente el valor que toma la variable respuesta,

$$Y = m\,(X_1, X_2, ..., X_k) \quad (2.1)$$

- No existe ninguna relación entre la variable respuesta y las variables regresoras, en el sentido de que el conocimiento de éstas no proporciona ninguna información sobre el comportamiento de la otra.
- El caso intermedio, existe una relación estocástica entre la variable respuesta y las variables regresoras, en el sentido de que el conocimiento de éstas permite predecir con mayor o menor exactitud el valor de la variable respuesta. Por tanto siguen un modelo de la forma,

$$Y = m\,(X_1, X_2, ..., X_k) + \varepsilon_t \quad (2.2)$$

  siendo m la función de regresión desconocida y ε una variable aleatoria de media cero (el error de observación).

Las relaciones estocásticas son las que ocurren en la mayoría de las situaciones y su estudio se corresponde con los modelos de regresión.

El objetivo básico en el estudio de un modelo de regresión es el de estimar la función de regresión, m, y el modelo probabilístico que sigue el error aleatorio ε, o sea

[24] "Bioestadística: Métodos y Aplicaciones", J. Barón López, Universidad de Málaga, España 1998.

estimar la función de distribución F_ε de la variable de error. La estimación de ambas funciones se hace a partir del conocimiento de una muestra de las variables en estudio, $\{((X_{1,i},X_{2,i},\ldots,X_{k,i}),Y_i): i=1,2,\ldots,n\}$.

Una vez estimadas estas funciones se tiene conocimiento de la relación funcional de la variable respuesta con las variables regresoras, dada por la función de regresión que se define como

$$m(x_1,\ldots,x_k)=E(Y/X_1=x_1,\ldots,X_k=x_k). \quad (2.3)$$

pudiéndose estimar y predecir con ésta el valor de la variable respuesta de un individuo del que se conocen los valores de las variables regresoras. Esto es, de un individuo *t* se sabe que $X_1 = x_{1,t},\ldots,X_k = x_{k,t}$, entonces se puede predecir el valor de Y_t y calcular un intervalo de predicción del mismo.

"...Los modelos de regresión se pueden clasificar de dos formas:

- Según la metodología utilizada para su estudio:

  - Modelos de regresión paramétricos, se supone que la función de regresión, *m*, que relaciona a la variable respuesta con las variables regresoras pertenece a una determinada familia paramétrica:

  $$m(\vec{\mathbf{x}})=m(\vec{\alpha},\vec{\mathbf{x}}), \quad (2.4)$$

  donde $\vec{\mathbf{x}}=(x_1,\ldots,x_k)$ y $\vec{\alpha}=(\alpha_1,\ldots,\alpha_p)\in\Theta^p\subset R^p$.

  Por ejemplo, se supone que la familia paramétrica es lineal,

  $$m(\vec{\mathbf{x}})=m(\vec{\alpha},\vec{\mathbf{x}})=\alpha_0+\alpha_1x_1+\ldots+\alpha_kx_k. \quad (2.5)$$

  En este caso, el problema básico es estimar los parámetros $\vec{\alpha}$ de la familia supuesta a partir de las observaciones muestrales. En el ejemplo anterior hay que estimar los parámetros α_0, α_1, α_2,..., α_k. También se debe contrastar la hipótesis de que la función de regresión pertenece a la familia paramétrica supuesta... este enfoque es el que con mayor frecuencia se utiliza en la práctica.

  - Modelos de regresión no paramétricos, es un enfoque alternativo... con este método no se hace ninguna suposición acerca de la forma funcional de la regresión y se estima la función de regresión punto a punto. Esto es, se estima el valor de $m(x_{1,i},x_{2,i},\ldots,x_{i,k})$ en un enrejado (grid) de valores $\{(x_{1,i},x_{2,i},\ldots,x_{k,i})\}_{i=1}^{N}$ de las variables regresoras.

No deben considerarse los métodos de regresión paramétricos y los no paramétricos como competidores sino como métodos complementarios... pues los dos métodos proporcionan información complementaria acerca del problema en estudio...

- Según la forma de recogida muestral

  - Modelos de regresión de diseño fijo, en estos modelos las variables regresoras son valores predeterminados. Este modelo se utiliza en el estudio del comportamiento de una variable respuesta cuando las variables regresoras varían en una determinada dirección. En este caso se debe diseñar y realizar un experimento en el que las variables regresoras se muevan en dicha dirección. Por tanto, en este diseño se controla en todo momento el valor de las variables regresoras.

  - Modelos de regresión con diseño aleatorio, en estos modelos las variables regresoras son variables aleatorias. Se utiliza este modelo cuando se estudia la relación entre la variable respuesta y las variables regresoras a partir de una muestra obtenida de la observación de las variables en unidades de experimentación elegidas al azar. Esto es, el experimentador es un observador pasivo en la recogida muestral y los resultados sólo serán válidos para el rango de variación conjunta de las variables implicadas en el estudio.

El tratamiento matemático en ambos modelos, de diseño fijo y de diseño aleatorio, es similar aunque las conclusiones e interpretación de los resultados varían según sea el caso..."[25]

Para este estudio empleamos el modelo de regresión con diseño fijo debido a las características de la obtención de los datos y a su menor complejidad. Veamos como estudia la estadística estos modelos.

2.2.1. El modelo de regresión lineal simple

El modelo estudia la relación lineal entre la variable respuesta *Y* y la variable regresora *X*, a partir de una muestra $\{(x_i, Y_i)\}_{i=1}^{n}$, que sigue el siguiente modelo:

$$Y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i \qquad i = 1, 2, \ldots, n. \quad (2.6)$$

Por tanto, es un modelo de regresión paramétrico de diseño fijo. En forma matricial

[25] "Modelos Estadísticos aplicados", J. Vilar Fernández, Universidade da Coruña, España 2003.

$$\vec{\mathbf{Y}} = \alpha_0 \vec{\mathbf{1}} + \alpha_1 \vec{\mathbf{X}} + \vec{\varepsilon}, \quad (2.7)$$

donde $\vec{\mathbf{Y}}^t = (y_1, \ldots, y_n)$, $\vec{\mathbf{1}}^t = (1, \ldots, 1)$, $\vec{\mathbf{X}}^t = (x_1, \ldots, x_n)$, $\vec{\varepsilon}^t = (\varepsilon_1, \ldots, \varepsilon_n)$.

y se supone que se verifican las siguientes hipótesis:

- La función de regresión es lineal,

  $$m(x_i) = E(Y/x_i) = \alpha_0 + \alpha_1 x_i, \quad i = 1, \ldots, n, \quad (2.8)$$

  o, equivalentemente, $E(\varepsilon_i) = 0$, $i = 1,...,n$, aunque puede ser que no haya linealidad y $E(\varepsilon_i)$ sea 0.

- La varianza es constante (homocedasticidad),

  $$Var(Y/x_i) = \sigma^2, \quad i = 1, \ldots, n, \quad (2.9)$$

  o, equivalentemente, $Var(\varepsilon_i) = \sigma^2$, $i = 1,...,n$.

- La distribución es normal,

  $$Y/x_i \sim N(\alpha_0 + \alpha_1 x_i, \sigma^2), \quad i = 1, \ldots, n, \quad (2.10)$$

  o, equivalentemente, $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1,...,n$.

- Las observaciones Y_i son independientes. Bajo las hipótesis de normalidad, esto equivale a que la $Cov(Y_i,Y_j) = 0$, si $i \neq j$.

  Esta hipótesis en función de los errores sería "los ε_i son independientes", que bajo normalidad, equivale a que $Cov(\varepsilon_i; \varepsilon_j) = 0$, si $i \neq j$.

En este modelo hay tres parámetros que se deben estimar: los coeficientes de la recta de regresión, α_0 y α_1; y la varianza de la distribución normal, σ^2.

El cálculo de estimadores para estos parámetros puede hacerse por diferentes métodos, siendo los más utilizados el método de máxima verosimilitud y el método de mínimos cuadrados (Reseña teórica 1, Anexo A).

2.2.1.1. Propiedades de los estimadores

Los estimadores del modelo de regresión simple tienen las siguientes propiedades:

- De su primera ecuación canónica se deduce que la recta de regresión pasa por el punto $(\bar{x}, \bar{y})$ que es el centro geométrico de la nube de datos.
- El estimador $\hat{\alpha}_1$ es la pendiente de la recta regresión, se denomina *coeficiente de regresión* y tiene una sencilla interpretación, indica el crecimiento (o decrecimiento) de la variable respuesta *Y* asociado a un incremento unitario en la variable regresora *X*.

- Utilizando las hipótesis de normalidad e independencia la distribución del estimado $\hat{\alpha}_1$ es una normal de media α_1 y varianza σ^2/ns_x^2. Esto es,

$$\hat{\alpha}_1 \sim N\left(\alpha_1, \frac{\sigma^2}{ns_x^2}\right). \quad (2.11)$$

Por tanto la $Var(\hat{\alpha}_1)$

- disminuye al aumentar *n*,
- disminuye al aumentar s_x^2 (varianza marginal)
- disminuye al disminuir σ^2.

- El estimador $\hat{\alpha}_0$ indica el valor de la ordenada en la recta de regresión estimada para $x = 0$ tiene menor importancia y, en muchos casos, no tiene una interpretación práctica. La distribución de $\hat{\alpha}_0$ es una normal de media α_0 y varianza

$$\frac{\sigma^2}{n} + \frac{\sigma^2\bar{x}^2}{ns_x^2} = \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{s_x^2}\right). \quad (2.12)$$

Esto es,

$$\hat{\alpha}_0 \sim N\left(\alpha_0, \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{s_x^2}\right)\right). \quad (2.13)$$

Por tanto la $Var(\hat{\alpha}_0)$ disminuye al disminuir $Var(\hat{\alpha}_1)$ (disminuye al aumentar *n* o al aumentar s_x^2 o al disminuir σ^2). - disminuye al disminuir $\bar{x}^2$.

- Nuevamente, utilizando las hipótesis de normalidad e independencia se obtiene que la distribución del estimador máximo-verosímil de σ^2, viene dada por

$$\frac{n\hat{\sigma}_{MV}^2}{\sigma^2} \sim \chi_{n-2}^2 \quad (2.14)$$

De las ecuaciones canónicas se deduce que los residuos verifican que $\sum_{i=1}^{n} e_i = 0$ y $\sum_{i=1}^{n} e_i x_i = 0$. Por tanto, el número de grados de libertad de los residuos es *n*-2 porque hay *n* residuos relacionados por dos ecuaciones. De donde

$$E\left(\hat{\sigma}_{MV}^2\right) = \frac{n-2}{n}\sigma^2 \Longrightarrow Sesgo\left(\hat{\sigma}_{MV}^2\right) = \frac{2}{n}\sigma^2 \to 0 \quad \text{cuando } n \to \infty. \quad (2.15)$$

y $\hat{\sigma}_{MV}^2$ es un estimador consistente pero sesgado. Por este motivo, como estimador de σ^2 se utiliza la varianza residual**,** $\hat{s}_R^2$ definida como la suma de residuos al cuadrado dividida por el número de grados de libertad

$$\hat{s}_R^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 \quad (2.16)$$

$\hat{s}_R^2$ es un estimador consistente e insesgado.

La relación entre los dos estimadores de la varianza es

$$\hat{\sigma}^2_{MV} = \frac{n-2}{n}\hat{s}^2_R. \quad (2.17)$$

Para tamaños muestrales grandes, ambos estimadores, $\hat{\sigma}_{MV}^2$ y $\hat{s}_R^2$ toman valores muy próximos.

- La distribución de la varianza residual viene dada por

$$\frac{(n-2)\,\hat{s}^2_R}{\sigma^2} \sim \chi^2_{n-2} \quad (2.18)$$

A partir de este estadístico se pueden obtener intervalos de confianza de la varianza poblacional, σ^2. Con nivel de confianza 1-α el intervalo de confianza es

$$\frac{(n-2)\,\hat{s}^2_R}{\chi^2_{n-2}\left(1-\frac{\alpha}{2}\right)} \leq \sigma^2 \leq \frac{(n-2)\,\hat{s}^2_R}{\chi^2_{n-2}\left(\frac{\alpha}{2}\right)} \quad (2.19)$$

- En la práctica, de la distribución de $\hat{\alpha}_1$ aparece σ, que es desconocido, para calcular un intervalo de confianza para este parámetro debemos estimar σ mediante un estimador, $\hat{s}_R^2$. De la distribución de éste se obtiene que la distribución del estadístico pivote ω_1 que sigue la distribución t_{n-2},

$$\omega_1 = \frac{\hat{\alpha}_1 - \alpha_1}{\hat{s}_R} s_x\sqrt{n} \sim t_{n-2}. \quad (2.20)$$

Un intervalo de confianza para α_1 a un nivel de confianza 1-α es

$$\alpha_1 \in \hat{\alpha}_1 \mp \frac{\hat{s}_R}{s_x\sqrt{n}}\, t_{n-2}\left(1-\frac{\alpha}{2}\right) \quad (2.21)$$

donde $t_{n-2}(\theta)$ es un número que verifica que $P(\zeta \leq t_{n-2}(\theta)) = \theta$, siendo ζ una variable aleatoria con distribución t con $n-2$ grados de libertad.

- De forma análoga se puede obtener un intervalo de confianza del parámetro α_0. De las funciones de distribución de $\hat{\alpha}_0$ y $\hat{s}_R^2$ se deduce que la distribución del estadístico ω_0 verifica que

$$\omega_0 = \frac{\hat{\alpha}_0 - \alpha_0}{\hat{s}_R\sqrt{\frac{1}{n}\left(1+\frac{\bar{x}^2}{s_x^2}\right)}} \sim t_{n-2}. \quad (2.22)$$

- Los estimadores $\hat{\alpha}_0$ y $\hat{\alpha}_1$ no son variables aleatorias independientes ya que su covarianza viene dada por

$$Cov\,(\hat{\alpha}_0, \hat{\alpha}_1) = \frac{-\bar{x}\sigma^2}{ns_x^2} \quad (2.23)$$

por tanto, si $\bar{x}$ es positiva, la *Cov* $(\hat{\alpha}_0, \hat{\alpha}_1)$ es negativa, esto es, al crecer $\hat{\alpha}_1$ disminuye $\hat{\alpha}_0$.

- Como ya se ha indicado el parámetro α_0 tiene menor importancia que α_1 y, en algunas situaciones, no tiene una interpretación realista si el cero no es un punto del rango de la *X*. Por ello tiene interés la ecuación de la recta de regresión que utiliza sólo el parámetro α_1. Esta ecuación es la siguiente

$$y_i - \bar{y} = \alpha_1 (x_i - \bar{x}) + \varepsilon_i, \quad (2.24)$$

o bien,

$$\hat{y}_i - \bar{y} = \hat{\alpha}_1 (x_i - \bar{x}) . \quad (2.25)$$

Para ello basta con centrar las dos variables en estudio y calcular la recta de regresión que pasa por el origen de coordenadas.

- La recta de regresión de *X* sobre *Y* es distinta de la recta de regresión de *Y* sobre *X*. En el primer caso se obtiene que

$$\hat{x}_i = \hat{\gamma}_0 + \hat{\gamma}_1 y_i \quad (2.26)$$

donde $\hat{\gamma}_1 = \frac{s_{XY}}{s_Y^2}$ y $\hat{\gamma}_0 = \bar{x} - \hat{\gamma}_1 \bar{y}$.

2.2.1.2. Análisis de contrastes

En los modelos de regresión es de gran interés el análisis de contrastes, ya que "...pueden presentarse en la práctica, situaciones en las que exista una teoría preconcebida relativa a la característica de la población sometida a estudio. Tal sería el caso, por ejemplo si pensamos que un tratamiento nuevo puede tener un porcentaje de mejoría mayor que otro estándar, o cuando nos planteamos si los niños de las distintas comunidades tienen la misma altura. Este tipo de circunstancias son las que nos llevan al estudio de la parcela de la estadística inferencial que se recoge bajo el título genérico de contraste de hipótesis. Implica, en cualquier investigación, la existencia de dos teorías o hipótesis implícitas, que denominaremos hipótesis nula e hipótesis alternativa, que de alguna manera reflejarán esa idea a priori que tenemos y que pretendemos contrastar con la *realidad*..."[26]

[26] "Bioestadística: Métodos y Aplicaciones", J. Barón López, Universidad de Málaga, España 1998.

Para el modelo de regresión lineal simple "...es importante analizar el siguiente contraste

$$C_1 : \left\{ \begin{array}{l} H_0 : \alpha_1 = 0 \\ H_1 : \alpha_1 \neq 0 \end{array} \right\}$$

ya que aceptar H_0 implica que la recta de regresión es $Y_i = \alpha_0 + \varepsilon_i$, por tanto, no existe relación lineal entre las variables X e Y .

Utilizando la distribución, si H_0 es cierto, se sigue que

$$\omega_1|_{H_0} = \hat{t}_1 = \frac{\hat{\alpha}_1}{\sigma\left(\hat{\alpha}_1\right)} = \frac{\hat{\alpha}_1}{\hat{s}_R} s_x \sqrt{n} \sim t_{n-2}. \quad (2.27)$$

Utilizando $\hat{t}_1$ como estadístico del contraste C_1 que es bilateral, se obtiene la siguiente región de aceptación a un nivel de significación α,

$$-\frac{\hat{s}_R}{s_x\sqrt{n}}\, t_{n-2}\left(1-\frac{\alpha}{2}\right) \leq \hat{\alpha}_1 \leq \frac{\hat{s}_R}{s_x\sqrt{n}}\, t_{n-2}\left(1-\frac{\alpha}{2}\right) \quad (2.28)$$

El *p-valor* del contraste C_1 es

$$p - valor = 2\,P\left(\zeta > \frac{\hat{\alpha}_1}{\hat{s}_R} s_x \sqrt{n}\right) \quad (2.29)$$

siendo ζ una variable aleatoria con distribución t_{n-2}. Este contraste se denomina contraste (individual) de la *t*..."[27]

Analicemos ahora el contraste de regresión del modelo, para ello descomponemos la variabilidad de la variable respuesta en variabilidad explicada por el modelo más variabilidad no explicada o residual, esto permite contrastar si el modelo es significativo o no. Bajo la hipótesis de que existe una relación lineal entre la variable respuesta y la regresora, se quiere realizar el siguiente contraste de hipótesis,

$H_0 : \quad E(Y/X = x) = \alpha_0$ (es constante, no depende de x)

frente a la alternativa

$H_1 : \quad E(Y/X = x) = \alpha_0 + \alpha_1 x$ (el modelo lineal es significativo)

por tanto, si se acepta H_0, la variable regresora no influye y no hay relación lineal entre ambas variables. En caso contrario, si existe una dependencia lineal de la variable respuesta respecto a la regresora.

Para todos los datos muestrales se hace la siguiente descomposición

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}), \quad (2.30)$$

[27] "Modelos Estadísticos aplicados", J. Vilar Fernández, Universidade da Coruña, España 2003.

elevando al cuadrado y sumando se obtiene,

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}), \quad (2.31)$$

en base a la ortogonalidad de los vectores se obtiene que los productos cruzados son cero, de donde se sigue la siguiente igualdad (Teorema de Pitágoras) que permite descomponer la variabilidad de la variable respuesta $\sum_{i=1}^{n}(y_i-\bar{y})^2$ en la variabilidad explicada por la recta de regresión $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ más la variabilidad residual o no explicada por el modelo ajustado $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$,

$$\underbrace{\overbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}^{\substack{Suma\ de\ Cuadrados\\ Global\ (scG)}}}_{g.l.=n-1}=\underbrace{\overbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}^{\substack{Suma\ de\ Cuadrados\\ Explicada\ (scE)}}}_{g.l.=1}+\underbrace{\overbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}^{\substack{Suma\ de\ Cuadrados\\ \mathrm{Re}sidual\ (scR)}}}_{g.l.=n-2}$$

En función de esto, estamos en condiciones de escribir la tabla ANOVA (Tabla 2.1).

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Varianzas
Por la recta	$scE=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	1	$\hat{s}_e^2=\frac{scE}{1}$
Residual	$scR=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	n-2	$\hat{s}_R^2=\frac{scR}{n-2}$
Global	$scG=\sum_{i=1}^{n}(y_i-\bar{y})^2$	n-1	$\hat{s}_y^2=\frac{scG}{n-1}$

Tabla 2.1. Tabla ANOVA del modelo de regresión simple

Si H_0 es cierta (la variable X no influye), la recta de regresión es aproximadamente horizontal y se verifica que aproximadamente $\hat{y}_i \approx \bar{y}$, y por tanto $scE \approx 0$. Pero scE es una medida con dimensiones y no puede utilizarse como medida de discrepancia, para resolver este inconveniente se divide por la varianza residual y como estadístico del contraste de regresión se utiliza el siguiente

$$\hat{F}_R=\frac{\hat{s}_e^2}{\hat{s}_R^2}. \quad (2.32)$$

Por la hipótesis de normalidad y bajo H_0 se deduce que el estadístico F_R sigue una distribución F (Contraste de la F) con 1 y n-2 grados de libertad.

$$\hat{F}_R = \frac{\hat{s}_e^2}{\hat{s}_R^2} \sim F_{1,n-2} \quad \text{bajo } H_0. \tag{2.33}$$

Sí el p - $valor = P(F_{1,n-2} \geq \hat{F}_R)$ es grande (mayor que α) se acepta H_0.

El Contraste de la F es un contraste unilateral (de una cola), pero en este modelo proporciona exactamente el mismo resultado que se obtiene por el contraste individual de la t relativo al coeficiente de regresión α_1 (Contraste de la t).

Si para cada valor de la variable explicativa $(X = x_i)$ se dispone de varios valores de la variable respuesta (algo normal en los modelos de regresión de diseño fijo) la muestra es de la siguiente forma $\{(x_i, y_{i,j}) : i = 1, \ldots, k;\ j = 1, \ldots, n_i\}$, que se puede ordenar como en la Tabla 2.2.

X_1	X_2	…	X_k
Y_{11}	Y_{21}	…	Y_{k1}
Y_{12}	Y_{22}	…	Y_{k2}
⋮	⋮	⋮	⋮
Y_{1n1}	Y_{2n2}	…	Y_{knk}

Tabla 2.2. Datos ordenados de la variable respuesta

El tamaño muestral es $n_1 + n_2 + ... + n_k = n$, y para cada valor de $X = x_i$, $i = 1,2,...,k$ se puede calcular la media condicionada muestral de la variable respuesta:

$$\bar{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}, \quad i = 1, 2, \ldots k, \tag{2.34}$$

lo que permite descomponer los residuos de la siguiente forma

$$e_{ij} = (y_{ij} - \hat{y}_i) = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i), \quad i = 1, 2, \ldots k, \quad j = 1, \ldots, n_i. \tag{2.35}$$

Un razonamiento análogo al realizado anteriormente permite descomponer la variabilidad no explicada como sigue,

$$
\begin{aligned}
scR &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i} (\bar{y}_{i.} - \hat{y}_i)^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^{k} n_i (\bar{y}_{i.} - \hat{y}_i)^2 .
\end{aligned}
\tag{2.36}
$$

Ahora la descomposición de la variabilidad total es la siguiente,

$$
\begin{aligned}
scG &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \left[\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2\right] + \left[\sum_{i=1}^{k} n_i (\hat{y}_{i.} - \bar{y}_{..})^2\right] \\
&= \left[\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^{k} n_i (\bar{y}_{i.} - \hat{y}_i)^2\right] + \left[\sum_{i=1}^{k} n_i (\hat{y}_{i.} - \bar{y}_{..})^2\right] = \\
&= scR + scE = scR(1) + scR(2) + scE.
\end{aligned}
\tag{2.37}
$$

En base a esta igualdad se puede construir la Tabla 2.3, más completa que la anterior.

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Varianzas
Recta	$scE = \sum_{i=1}^{k} n_i (\hat{y}_i - \bar{y}_{..})^2$	1	$\hat{s}_e^2 = \frac{VE}{1}$
$ScR(1)$	$scR(1) = \sum_{i=1}^{k} n_i (\bar{y}_{i.} - \hat{y}_i)^2$	k-2	$\hat{s}_{R,1}^2 = \frac{scR(1)}{k-2}$
$scR(2)$	$scR(2) = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$	n-k	$\hat{s}_{R,2}^2 = \frac{scR(2)}{n-k}$
scR	$scR = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2$	n-2	$\hat{s}_R^2 = \frac{scR}{n-2}$
Global	Global $\sum_{i=1}^{n} (y_i - \bar{y})^2$	n–1	$\hat{s}_y^2 = \frac{scG}{n-1}$

Tabla 2.3. Tabla ANOVA del modelo de regresión

A partir de esta tabla ANOVA se puede contrastar la hipótesis de que la función de regresión es lineal frente a la alternativa de que no es lineal, esto es,

$$H_0: \quad E(Y/X = x_i) = \alpha_0 + \alpha_1 x_i \ \text{(la función es lineal)}$$

frente a la alternativa

$$H_1: \quad E(Y/X = x) = m(x) \ \text{(no es una función lineal)}$$

Si H_0 es cierto, las medias condicionadas estarán próximas a la recta de regresión: $\bar{y}_i \approx \hat{y}_i$, y la $scR(1) = \sum_{i=1}^{k} n_i (\bar{y}_i - \hat{y}_i)^2 \approx 0$. De nuevo esta medida tiene dimensiones y

no es válida para utilizar como medida de discrepancia, para resolver el problema se compara con $\hat{s}_{R,2}{}^2$ y el cociente de ambas cantidades se utiliza como estadístico del contraste en estudio.

$$\hat{F}_L = \frac{\hat{s}^2_{R,1}}{\hat{s}^2_{R,2}} \quad (2.38)$$

Bajo la hipótesis de normalidad y H_0 (hipótesis de linealidad) se deduce que $\hat{F}_L$ sigue una distribución $F_{k-2,\,n-k}$ (Contraste de la *F*).

$$\hat{F}_L = \frac{\hat{s}^2_{R,1}}{\hat{s}^2_{R,2}} \sim F_{k-2,n-k} \quad \text{bajo } H_0 \quad (2.39)$$

Este *contraste de linealidad de la F* es unilateral. Si el *p-valor* = $P\left(F_{k-2,n-k} \geq \hat{F}_L\right)$ es grande (mayor que α) se acepta que la curva de regresión es lineal.

2.2.1.3. El coeficiente de determinación

Una vez ajustada la recta de regresión a la nube de observaciones es importante disponer de una medida que mida la bondad del ajuste realizado y que permita decidir si el ajuste lineal es suficiente o se deben buscar modelos alternativos. Como medida de bondad del ajuste se utiliza el coeficiente de determinación, definido como

$$R^2 = \frac{scE}{scG} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (2.40)$$

o bien

$$R^2 = 1 - \frac{scR}{scG} = 1 - \frac{n-2}{n-1} \frac{\hat{s}^2_R}{\hat{s}^2_Y} \quad (2.41)$$

Como $scE \leq scG$, se verifica que $0 \leq R^2 \leq 1$.

El coeficiente de determinación mide la proporción de variabilidad total de la variable dependiente (*Y*) respecto a su media que es explicada por el modelo de regresión. Es usual expresar esta medida en tanto por ciento, multiplicándola por cien.

Por otra parte, teniendo en cuenta que $\hat{y}_i - \bar{y} = \hat{\alpha}_1(x_i - \bar{x})$, se obtiene

$$R^2 = \frac{s^2_{XY}}{s^2_X \, s^2_Y} \quad (2.42)$$

2.2.1.4. El coeficiente de correlación

Dadas dos variables aleatorias cualesquiera *X* e *Y* , una medida de la relación lineal que hay entre ambas variables es el coeficiente de correlación definido por

$$\rho = \frac{Cov\,(X,Y)}{\sigma\,(X)\;\sigma\,(Y)} \quad (2.43)$$

donde $\sigma(X)$ representa la desviación típica de la variable *X* (análogamente para $\sigma(Y)$). Un buen estimador de este parámetro es el coeficiente de correlación lineal muestral (o coeficiente de correlación de Pearson), definido por

$$r = \frac{s_{XY}}{s_X\, s_Y} = signo\,(\hat{\alpha}_1)\,\sqrt{R^2}. \quad (2.44)$$

Por tanto, $r \in [-1, 1]$. Este coeficiente es una buena medida de la bondad del ajuste de la recta de regresión. Evidentemente, existe una estrecha relación entre *r* y $\hat{\alpha}_1$ aunque estos estimadores proporcionan diferentes interpretaciones del modelo:

* *r* es una medida de la relación lineal entre las variables *X* e *Y*.

* $\hat{\alpha}_1$ mide el cambio producido en la variable *Y* al realizarse un cambio de una unidad en la variable *X*.

De las definiciones anteriores se deduce que:

$$s_{XY} = 0 \Leftrightarrow \hat{\alpha}_1 = 0 \Leftrightarrow r = 0. \quad (2.45)$$

Es importante estudiar si *r* es significativo (distinto de cero) ya que ello implica que el modelo de regresión lineal es significativo. Desafortunadamente la distribución de *r* es complicada pero para tamaños muestrales mayores que 30 su desviación típica es $\sigma(r) \simeq 1/\sqrt{n}$, y puede utilizarse la siguiente regla

$$|r| > \frac{2}{\sqrt{n}} \quad \Rightarrow \quad r \text{ es significativo (con } \alpha = 0{,}05)$$

En la interpretación del coeficiente de correlación se debe tener en cuenta que:

- *r*=±1 indica una relación lineal exacta positiva (creciente) o negativa (decreciente),
- *r*=0 indica la no existencia de relación lineal estocástica, pero no indica independencia de las variables ya que puede existir una relación no lineal incluso exacta,
- Valores intermedios de *r* (0 < *r* < 1 ó -1 < *r* < 0) indican la existencia de una relación lineal estocástica, más fuerte cuanto más próximo a +1 (ó -1) sea el valor de *r*.

En el Ejemplo 1 del Anexo B pueden verse diversos casos de ajustes de curvas a nubes de puntos.

2.2.1.5. Las transformaciones

Como ya dijéramos, la hipótesis básica del modelo de regresión lineal simple es

$$E(Y/X = x) = \alpha_0 + \alpha_1 x, \quad (2.46)$$

pero en muchos casos en el gráfico de la variable respuesta frente a la variable regresora puede verse que la relación no es de este tipo. A pesar de ello, el modelo de regresión lineal continúa siendo válido en muchas situaciones porque la relación puede convertirse en lineal por medio de una transformación simple en la variable respuesta *Y* (trabajando con lg *Y* , 1/*Y*, Y^2...), o en la variable regresora, *X*, o en ambas.

Distintos tipos de transformaciones pueden verse en la Tabla 2.4., algunas de las cuales se emplean más adelante en la determinación de los modelos.

	Modelo	Trans. X	Trans. Y
Simple	$Y = \alpha_0 + \alpha_1 X$	$t(x) = x$	$t(y) = y$
Expon.	$Y = \exp(\alpha_0 + \alpha_1 X)$	$t(x) = x$	$t(y) = \ln(y)$
Recípr. Y	$Y = \frac{1}{\alpha_0 + \alpha_1 X}$	$t(x) = x$	$t^{(y)} = \frac{1}{y}$
Recípr. X	$Y = \alpha_0 + \alpha_1 \frac{1}{X}$	$t^{(x)} = \frac{1}{x}$	$t(y) = y$
Rec Doble	$Y = \frac{1}{\alpha_0 + \alpha_1 / X}$	$t^{(x)} = \frac{1}{x}$	$t^{(y)} = \frac{1}{y}$
Logar. *X*	$Y = \alpha_0 + \alpha_1 \ln(X)$	$t(x) = \ln(x)$	$t(y) = y$
Multipl	$Y = \alpha_0 X_\alpha{}^1$	$t(x) = \ln(x)$	$t(y) = \ln(y)$
Raíz C. *X*	$Y = \alpha_0 + \alpha_1 \sqrt{X}$	$t(x) = \sqrt{x}$	$t(y) = y$
Raíz C. *Y*	$\sqrt{Y} = \alpha_0 + \alpha_1 X$	$t(x) = x$	$t(y) = \sqrt{y}$
Curva *S*	$Y = \exp\left(\alpha_0 + \frac{\alpha_1}{X}\right)$	$t^{(x)} = \frac{1}{x}$	$t(y) = \ln(y)$

Tabla 2.4. Transformaciones para la regresión

También pueden observarse las transformaciones en forma gráfica en la Reseña Teórica 3 del Anexo A.

En algunos casos transformar las variables del modelo permite resolver problemas como falta de normalidad o heterocedasticidad. Por ello, si en el análisis de residuos no se observan estos problemas, se puede intentar conseguir la linealidad del modelo transformando solamente la variable regresora *x*. Pero si, por el contrario, se observan estos problemas puede ser necesario transformar las dos variables.

2.2.1.6. Análisis de residuos

Al obtenerse el modelo de regresión, se genera una diferencia punto a punto entre el valor real de la variable dependiente y el que se obtiene por el modelo, que se denomina residuo. El análisis de los residuos es el paso siguiente a la obtención del modelo. Para realizar este análisis veamos los problemas que pueden aparecer al ajustar el modelo.

"...Al ajustar un modelo de regresión lineal simple se pueden presentar diferentes problemas bien porque no existe una relación lineal entre las variables o porque no se verifican las hipótesis estructurales que se asumen en el ajuste del modelo. Estos problemas son los siguientes:

- Falta de Linealidad, porque la relación entre las dos variables no es lineal o porque variables explicativas relevantes no han sido incluidas en el modelo.
- Existencia de valores atípicos e influyentes, existen datos atípicos que se separan de la nube de datos muestrales e influyen en la estimación del modelo.
- Falta de Normalidad, los residuos del modelo no se ajustan a una distribución normal.
- Heterocedasticidad, la varianza de los residuos no es constante.
- Dependencia (autocorrelación), existe dependencia entre las observaciones.

Un primer paso para el estudio de estos problemas es la realización de un estudio descriptivo, analítico y gráfico, de la muestra. En particular el gráfico de puntos de la muestra bidimensional permite detectar algunos problemas como se ponen de manifiesto..." [28]

En la Reseña Teórica 2 del Anexo A, pueden observarse los gráficos de los casos citados.

Veamos una clasificación de los residuos.

- *Residuos ordinarios:* Se define el residuo ordinario asociado a una observación muestral como la diferencia entre la observación (y_i) y la predicción $(\hat{y}_i)$,

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\alpha}_0 + \hat{\alpha}_1 x_i), \quad i = 1, \ldots, n. \tag{2.47}$$

El *i*-ésimo residuo e_i es una variable aleatoria que tiene las siguientes propiedades

$$E(e_i) = 0, \quad Var(e_i) = \sigma^2(e_i) = \sigma^2(1 - h_{ii}), \quad i = 1, \ldots, n. \tag{2.48}$$

Bajo la hipótesis de normalidad se obtiene

$$e_i \sim N\left(0; \sigma^2(1 - h_{ii})\right), \quad i = 1, \ldots, n, \tag{2.49}$$

- *Residuos estandarizados:* De lo expuesto se deduce que $\sigma^2(e_i)$ no es constante, lo que hace difícil identificar las observaciones con residuos grandes. Por ello es usual tipificarlos y se definen los residuos estandarizados como

$$r_i = \frac{e_i}{s_R\sqrt{1 - h_{ii}}}, \quad i = 1, \ldots, n. \tag{2.50}$$

Los residuos estandarizados tienen media cero y varianza próxima a 1, esto permite distinguir a los residuos grandes.

- *Residuos estudentizados:* De lo expuesto también se deduce que existe una relación de dependencia entre el numerador y el denominador de r_i ya que en el cálculo de s_R se utiliza el residuo e_i. Este problema se elimina si se estima la varianza residual a partir de toda la muestra excepto la observación (x_i, y_i). A la varianza residual así obtenida se le denota por $s_{R,}(i)^2$.

Se definen los *residuos estudentizados* como

$$t_i = \frac{e_i}{s_{R,(i)}\sqrt{1 - h_{ii}}}, \quad i = 1, \ldots, n. \tag{2.51}$$

Si *n* es grande los residuos estandarizados y estudentizados toman valores próximos.

Bajo la hipótesis de normalidad se verifica que t_i sigue una distribución *t* con *n*-3 grados de libertad.

[28] "Modelos Estadísticos aplicados", J. Vilar Fernández, Universidade da Coruña, España 2003

Los residuos estudentizados se pueden calcular de forma más sencilla como sigue

$$t_i = \frac{e_i\sqrt{n-3}}{\left[(n-2)\,s_R^2\,(1-h_{ii}) - e_i^2\right]^{1/2}} \in t_{n-3}, \quad i = 1, \ldots, n. \tag{2.52}$$

- *Residuos eliminados:* Se definen los *residuos eliminados* como la diferencia entre lo observado en la respuesta y_i y la predicción cuando se utiliza toda la muestra excepto la observación en estudio y que se denota por $\hat{y}_{i(i)}$,

$$e_{(i)} = y_i - \hat{y}_{i(i)}, \quad i = 1, \ldots, n. \tag{2.53}$$

Entre los residuos ordinarios y los residuos eliminados existe la siguiente relación

$$e_{(i)} = \frac{e_i}{1-h_{ii}}, \quad i = 1, \ldots, n. \tag{2.54}$$

Si la observación (x_i,y_i) tiene una influencia grande en el cálculo de la recta de regresión, los dos residuos e_i y $e_{(i)}$ son diferentes, en caso contrario, serán muy parecidos.

2.2.1.7. Influencia de las observaciones

En el ajuste de una recta de regresión a una muestra bidimensional $\{(x_i, y_i)\}_{i=1}^{n}$, al observar el gráfico de *y* frente a *x*, en algunas ocasiones, existen observaciones (valores extremos) que se separan claramente del resto de la nube de observaciones. Es importante conocer la *influencia* que estos puntos tienen en el cálculo de la estimación de la recta. Es decir, fijada una observación (x_t,y_t) de la muestra, la variación que se produce en la estimación de la recta de regresión al calcularla con toda la muestra excepto con el dato (x_t,y_t) en lugar de hacerlo con toda la muestra. Esto puede verse claramente en el Ejemplo 2 del Anexo B.

2.2.2. El modelo de regresión lineal múltiple

Hasta ahora hemos analizado un situación en donde con una sola variable se puede dar respuesta a una realidad dada, pero esto no siempre es así.

Regresemos a la definición de modelos de regresión vista, que dice que éstos estudian la relación estocástica cuantitativa entre una variable de interés y un conjunto de variables explicativas.

Sea Y la variable de interés, variable respuesta o dependiente y sean $x_1,x_2,...,x_k$ las variables explicativas o regresoras. La formulación matemática de estos modelos es la siguiente

$$Y = m\left(x_1, x_2, \ldots, x_k\right) + \varepsilon \quad (2.55)$$

donde ε es el error de observación debido a variables no controladas.

Como el modelo de regresión lineal general "*supone"* que la función de regresión $m(x_1, x_2, \ldots, x_k)$ es lineal, podemos decir que su expresión matemática es

$$Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k + \varepsilon \quad (2.56)$$

Un primer objetivo en el estudio de este modelo es el de estimar los parámetros del mismo $\alpha_0,\alpha_1,\alpha_2,...,\alpha_k$, y la función de distribución del error F_ε a partir de una muestra de n observaciones, que tendrá la forma

$$\left\{\left(\vec{\mathbf{x}}_{i\cdot}; y_i\right)\right\}_{i=1}^{n} = \left\{\left(\left(x_{i1}, x_{i2}, \ldots, x_{ik}\right); y_i\right)\right\}_{i=1}^{n}. \quad (2.57)$$

De la expresión matemática del modelo de regresión lineal general se deduce que para i = 1,2,...,n se verifica la siguiente igualdad

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_k x_{ik} + \varepsilon_i \quad i = 1, 2, \ldots, n, \quad (2.58)$$

donde ε_i es el error aleatorio o perturbación de la observación i-ésima.

Es interesante escribir el modelo de regresión lineal general en forma matricial.

$$\left\{\begin{array}{ccccccccccc} y_1 & = & \alpha_0\, 1 & + & \alpha_1 x_{11} & + & \alpha_2 x_{12} & + & \ldots & + & \alpha_k x_{1k} & + & \varepsilon_1 \\ y_2 & = & \alpha_0\, 1 & + & \alpha_1 x_{21} & + & \alpha_2 x_{22} & + & \ldots & + & \alpha_k x_{2k} & + & \varepsilon_2 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ y_n & = & \alpha_0\, 1 & + & \alpha_1 x_{n1} & + & \alpha_2 x_{n2} & + & \ldots & + & \alpha_k x_{nk} & + & \varepsilon_n \end{array}\right\}$$

escrito en forma vectorial

$$\vec{\mathbf{Y}} = \alpha_0\, \vec{\mathbf{1}} + \alpha_1 \vec{\mathbf{x}}_{\cdot 1} + \alpha_1 \vec{\mathbf{x}}_{\cdot 2} + \ldots + \alpha_1 \vec{\mathbf{x}}_{\cdot k} + \vec{\boldsymbol{\varepsilon}}, \quad (2.59)$$

escrito en forma matricial

$$\vec{\mathbf{Y}} = \mathbf{X}\, \vec{\alpha} + \vec{\varepsilon} \quad (2.60)$$

donde $\vec{\mathbf{Y}}$ es un vector n-dimensional (matriz $n \times 1$) de la variable respuesta o dependiente,

X es la matriz del diseño de las variables regresoras (matriz $n \times (k+1)$), la primera columna de esta matriz está formada por unos, es la columna asociada con el

parámetro α_0; la columna j+1 contiene la información relativa a la variable x_j, j = 1,...,k, es la columna asociada al parámetro α_j.

$\vec{\alpha}$ es el vector (k+1)-dimensional (matriz (k+1)× 1) de los parámetros del modelo,

$\vec{\varepsilon}$ es el vector n-dimensional (matriz n×1) de las perturbaciones aleatorias.

Desarrollando la ecuación matricial anterior se tiene,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

La fila i-ésima de la matriz **X**, $\vec{\mathbf{x}}_{i\cdot} = (x_{i1}, x_{21}, \dots, x_{ik})$ se corresponde con los datos de las variables regresoras en el individuo i-ésimo, i=1,2,...,n. Por tanto, la información acerca del individuo *i-ésimo* está contenida en el vector $\vec{\mathbf{x}}_{i\cdot}$

La columna j-ésima de la matriz **X**, $\vec{\mathbf{x}}_{\cdot j} = (x_{1j}, x_{2j}, \dots, x_{nj})^t$ se corresponde con los datos de la variable regresora x_j, j=1,2,...,k. La información acerca de la variable *j-ésima* está contenida en el vector $\vec{\mathbf{x}}_{\cdot j}$.

En resumen, las matrices del modelo de regresión lineal múltiple son:

$$\vec{\mathbf{Y}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}, \quad \vec{\alpha} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix}, \quad \vec{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

"...En el estudio del modelo de regresión lineal general se asume que se verifican las siguientes hipótesis:

- La función de regresión es lineal,

$$m(\vec{\mathbf{x}}_{i\cdot}) = m(x_{i1}, x_{i2}, \dots, x_{ik}) = E(Y/x_{i1}, x_{i2}, \dots, x_{ik}) = E(Y/\vec{\mathbf{x}}_{i\cdot}) \quad (2.61)$$
$$= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \dots + \alpha_k x_{ik},\ i = 1,\dots,n,$$

  o, equivalentemente, $E(\varepsilon_i) = 0$, i = 1,...,n.

- La varianza es constante (homocedasticidad),

$$Var(Y/\vec{\mathbf{x}}_{i\cdot}) = Var(Y/x_{i1}, x_{i2}, \dots, x_{ik}) = \sigma^2, \quad i = 1, \dots, n, \quad (2.62)$$

  o, equivalentemente, $Var(\varepsilon_i) = \sigma^2$, i = 1,...,n.

- La distribución es normal,

$$Y/\vec{\mathbf{x}}_{i\cdot} = Y/x_{i1}, x_{i2}, \dots, x_{ik} \sim N\left(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \dots + \alpha_k x_{ik}, \sigma^2\right), \quad (2.63)$$

  o, equivalentemente, $\varepsilon_i \sim N(0, \sigma^2)$, i = 1,...,n.

- Las observaciones Y_i son independientes (bajo normalidad, esto equivale a que la $Cov(Y_i,Y_j) = 0$, si $i \neq j$).

  Esta hipótesis en función de los errores sería "los ε_i son independientes, que bajo normalidad, equivale a que $Cov(\varepsilon_i;\varepsilon_j) = 0$, si $i \neq j$".
- $n > k+1$. En caso contrario no se dispone de información suficiente para estimar los parámetros del modelo.
- Las variables regresoras $x_1,x_2,...,x_k$ son linealmente independientes..."[29]

2.2.2.1. Los estimadores

De la cita bibliográfica anterior también se obtiene que la estimación de los parámetros del modelo de regresión lineal múltiple se realiza planteando $\hat{\alpha}$ como un estimador del vector de parámetros $\vec{\alpha}$. Luego, se define el vector de predicciones como

$$\hat{\mathbf{Y}} = \mathbf{X}\,\hat{\alpha}. \quad (2.64)$$

El vector de residuos se obtiene como

$$\vec{\mathbf{e}} = \vec{\mathbf{Y}} - \hat{\mathbf{Y}}. \quad (2.65)$$

El estimador por mínimos cuadrados de $\vec{\alpha}$ se obtiene minimizando la suma de los residuos al cuadrado. Esto es, se minimiza la siguiente función de $k+1$ variables:

$$\begin{aligned} \Psi(\hat{\boldsymbol{\alpha}}) &= \sum_{i=1}^{n} e_i^2 = \vec{\mathbf{e}}^t\,\vec{\mathbf{e}} = \left(\vec{\mathbf{Y}} - \hat{\mathbf{Y}}\right)^t \left(\vec{\mathbf{Y}} - \hat{\mathbf{Y}}\right) \\ &= \left(\vec{\mathbf{Y}} - \mathbf{X}\,\hat{\boldsymbol{\alpha}}\right)^t \left(\vec{\mathbf{Y}} - \mathbf{X}\,\hat{\boldsymbol{\alpha}}\right) = \left(\vec{\mathbf{Y}}^t - \hat{\boldsymbol{\alpha}}^t\mathbf{X}^t\right)\left(\vec{\mathbf{Y}} - \mathbf{X}\,\hat{\boldsymbol{\alpha}}\right) \\ &= \vec{\mathbf{Y}}^t\vec{\mathbf{Y}} - \vec{\mathbf{Y}}^t\mathbf{X}\,\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^t\mathbf{X}^t\vec{\mathbf{Y}} + \hat{\boldsymbol{\alpha}}^t\mathbf{X}^t\mathbf{X}\,\hat{\boldsymbol{\alpha}}. \end{aligned} \quad (2.66)$$

Derivando respecto a $\hat{\alpha}$ e igualando a cero, se obtienen las ecuaciones de regresión

$$\mathbf{X}^t\vec{\mathbf{Y}} = \mathbf{X}^t\mathbf{X}\,\hat{\alpha}, \quad (2.67)$$

de donde se deduce el siguiente estimador por mínimos cuadrados

$$\hat{\alpha} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\vec{\mathbf{Y}} \quad (2.68)$$

Debe tenerse en cuenta que para calcular este estimador es necesario que la matriz $\mathbf{X}^t\mathbf{X}$ sea invertible. Esto está garantizado por la sexta hipótesis del modelo.

La matriz $\mathbf{X}^t\mathbf{X}$ es una matriz $(k+1)$x$(k+1)$ cuya expresión es la siguiente

[29] "Modelos Estadísticos aplicados", J. Vilar Fernández, Universidade da Coruña, España 2003

$$\mathbf{X}^t\mathbf{X} = \begin{pmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \dots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \dots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \dots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{ik}x_{i1} & \sum_{i=1}^n x_{ik}x_{i2} & \dots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}.$$

La matriz $\mathbf{X}^t\mathbf{Y}$ es una matriz $(k+1) \times 1$ que viene dada por

$$\mathbf{X}^t\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \sum_{i=1}^n x_{i2}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_i \end{pmatrix}.$$

Si se trabaja con todas las variables centradas se obtiene otra forma interesante de expresar el modelo de regresión lineal.

$$(Y_i - \bar{Y}) = \alpha_1 (x_{i1} - \bar{x}_1) + \alpha_2 (x_{i2} - \bar{x}_2) + \ldots + \alpha_k (x_{ik} - \bar{x}_k) + \varepsilon_i \quad i = 1, 2, \ldots, n \tag{2.69}$$

donde $\bar{Y}, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ son las medias muestrales de las variables $Y, x_1, x_2, \ldots, x_k$.

Razonando como antes, se obtiene el siguiente estimador por mínimos cuadrados del vector $\vec{\mathbf{a}} = (\alpha_1, \alpha_2, \ldots, \alpha_k)^t$

$$\hat{\mathbf{a}} = \left(\tilde{\mathbf{X}}^t\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^t\tilde{\mathbf{Y}} = \mathbf{S}_{XX}^{-1}\mathbf{S}_{XY} \tag{2.70}$$

donde $\tilde{\mathbf{X}}$ es la matriz del diseño de las variables regresoras centradas (matriz $n \times k$)

$$\tilde{\mathbf{X}} = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \dots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \dots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \dots & x_{nk} - \bar{x}_k \end{pmatrix}$$

$\mathbf{S}_{XX}$ es la matriz de covarianzas de $(x_1, x_2, \ldots, x_k)$, esto es,

$$\mathbf{S}_{XX} = \begin{pmatrix} Var(x_1) & Cov(x_1,x_2) & \dots & Cov(x_1,x_k) \\ Cov(x_2,x_1) & Var(x_2) & \dots & Cov(x_2,x_k) \\ \vdots & \vdots & \vdots & \vdots \\ Cov(x_k,x_1) & Cov(x_k,x_2) & \dots & Var(x_k) \end{pmatrix}$$

Y $\mathbf{S}_{XY}$ es el vector de covarianzas de Y con $(x_1, x_2, \ldots, x_k)$,

$$\mathbf{S}_{XY} = \begin{pmatrix} Cov(x_1,Y) \\ Cov(x_2,Y) \\ \vdots \\ Cov(x_k,Y) \end{pmatrix}$$

En el estudio del modelo de regresión lineal múltiple con *k* variables regresoras a partir de una muestra de *n* observaciones se considera el subespacio vectorial π de R^n, de dimensión (*k*+1), generado por los vectores $\{\vec{\mathbf{1}}, \vec{\mathbf{x}}_{\cdot 1}, \vec{\mathbf{x}}_{\cdot 2}, \ldots, \vec{\mathbf{x}}_{\cdot k}\}$ (columnas de la matriz de diseño **X**). El problema de ajustar un modelo de regresión lineal múltiple se puede interpretar geométricamente como el problema de encontrar en este subespacio vectorial π el vector $\hat{\mathbf{Y}}$ (vector de predicciones) lo más próximo al vector de la variable respuesta, $\vec{\mathbf{Y}}$. Esto es, encontrar el vector $\hat{\mathbf{Y}}$ que minimice el módulo del vector de residuos, $\vec{\mathbf{e}} = \vec{\mathbf{Y}} - \hat{\mathbf{Y}}$ (la suma de los residuos al cuadrado). La resolución de este problema viene dada por el vector proyección ortogonal del vector $\vec{\mathbf{Y}}$ en el subespacio π considerado. Por tanto,

$$\hat{\mathbf{Y}} = \mathbf{H}\vec{\mathbf{Y}} \quad (2.71)$$

siendo *H* la matriz de proyección (hat matrix) en el subespacio π.

El estimador por mínimos cuadrados $\hat{\alpha}$ viene dado por las coordenadas del vector de predicciones $\hat{\mathbf{Y}}$ en el subespacio π respecto a la base $\{\vec{\mathbf{1}}, \vec{\mathbf{x}}_{\cdot 1}, \vec{\mathbf{x}}_{\cdot 2}, \ldots, \vec{\mathbf{x}}_{\cdot k}\}$.

De esta interpretación geométrica se deduce que los residuos verifican las siguientes (*k*+1) restricciones

$$\vec{\mathbf{e}} \perp \vec{\mathbf{1}} \Rightarrow \sum_{i=1}^{n} e_i = 0$$

$$\vec{\mathbf{e}} \perp \vec{\mathbf{x}}_{\mathbf{j}} \Rightarrow \sum_{i=1}^{n} e_i x_{ij} = 0 \quad j = 1, 2, \ldots, k \quad (2.72)$$

por tanto, los residuos tienen *n*-(*k*+1) grados de libertad.

Dado que

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\alpha} = \mathbf{X}\ \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\vec{\mathbf{Y}} = \mathbf{H}\vec{\mathbf{Y}}. \quad (2.73)$$

Por tanto la matriz de proyección sobre el subespacio π es

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t \quad (2.74)$$

Por tanto la matriz $\mathbf{H} = (h_{ij})_{i,j=1}^{n}$ se obtiene a partir de la matriz del diseño **X**, es una matriz *n*×*n* y juega un papel muy importante en el modelo de regresión lineal.

En el estudio del modelo de regresión múltiple tiene gran interés la suma de residuos al cuadrado que representa la variabilidad no explicada por el modelo (*scR*). A partir de este valor se obtiene el estimador de la varianza σ^2.

Una forma sencilla de calcular *scR* es la siguiente: el vector de residuos se puede expresar como

$$\vec{\mathbf{e}} = \vec{\mathbf{Y}} - \hat{\mathbf{Y}} = \vec{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\alpha}} = \vec{\mathbf{Y}} - \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\vec{\mathbf{Y}} \Rightarrow \quad (2.75)$$

$$\vec{\mathbf{e}} = \left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\right)\vec{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\,\vec{\mathbf{Y}}$$

Utilizando esto, el módulo de $\vec{\mathbf{e}}$ al cuadrado es

$$\sum_{i=1}^{n} e_i^2 = |\vec{\mathbf{e}}|^2 = \vec{\mathbf{e}}^t\vec{\mathbf{e}} = \left(\vec{\mathbf{Y}} - \hat{\mathbf{Y}}\right)^t\vec{\mathbf{e}} = \left(\vec{\mathbf{Y}} - \hat{\mathbf{Y}}\right)^t(\mathbf{I} - \mathbf{H})\,\vec{\mathbf{Y}} \Rightarrow \quad (2.76)$$

$$\sum_{i=1}^{n} e_i^2 = \vec{\mathbf{Y}}^t(\mathbf{I} - \mathbf{H})\,\mathbf{Y} - \hat{\mathbf{Y}}^t(\mathbf{I} - \mathbf{H})\,\vec{\mathbf{Y}},$$

dado que $\hat{\mathbf{Y}} \perp \vec{\mathbf{e}}$, el segundo término de la expresión es cero, por tanto

$$\hat{\mathbf{Y}}^t(\mathbf{I} - \mathbf{H})\,\mathbf{Y} = \hat{\mathbf{Y}}^t\vec{\mathbf{e}} = 0, \quad (2.77)$$

de donde se sigue que

$$\sum_{i=1}^{n} e_i^2 = \vec{\mathbf{e}}^t\vec{\mathbf{e}} = \vec{\mathbf{Y}}^t(\mathbf{I} - \mathbf{H})\,\mathbf{Y} = \vec{\mathbf{Y}}^t\vec{\mathbf{Y}} - \vec{\mathbf{Y}}^t\mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\vec{\mathbf{Y}} = \vec{\mathbf{Y}}^t\vec{\mathbf{Y}} - \hat{\alpha}^t\mathbf{X}^t\vec{\mathbf{Y}}, \quad (2.78)$$

o equivalentemente

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n}\left(\hat{\alpha}_0\sum_{i=1}^{n} y_i + \hat{\alpha}_1\sum_{i=1}^{n} x_{i1}y_i + \hat{\alpha}_2\sum_{i=1}^{n} x_{i2}y_i + \ldots + \hat{\alpha}_k\sum_{i=1}^{n} x_{ik}y_i\right)$$

Esta expresión es muy útil para el cálculo de *scR*. Debe tenerse en cuenta que el cálculo de la matriz $\mathbf{X}^t\vec{\mathbf{Y}}$ ya se utilizó en el calculo del estimador $\hat{\alpha}$.

Los estimadores en la regresión lineal múltiple tienen las siguientes propiedades:

- *Estimador de los coeficientes del modelo lineal*: El estimador del vector $\vec{\alpha}$ por el método de mínimos cuadrados es

$$\hat{\alpha} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} \quad (2.79)$$

por la hipótesis de normalidad, es igual al estimador que se obtiene utilizando el método de máxima verosimilitud.

Este estimador verifica las siguientes propiedades:

- El estimador $\hat{\alpha}$ es insesgado o centrado: $E(\hat{\alpha}) = \vec{\alpha}$.
- La matriz de varianzas-covarianzas del estimador $\hat{\alpha}$ es

$$Var\,(\hat{\boldsymbol{\alpha}}) = E\left((\hat{\boldsymbol{\alpha}} - E\,(\hat{\boldsymbol{\alpha}}))\,(\hat{\boldsymbol{\alpha}} - E\,(\hat{\boldsymbol{\alpha}}))^t\right) = \sigma^2\left(\mathbf{X}^t\mathbf{X}\right)^{-1} = \left(\sigma_{ij}^2\right)_{i,j=0}^{k}. \quad (2.80)$$

De donde se deduce que los estimadores $\hat{\alpha}_i$ y $\hat{\alpha}_j$ ($i \neq j$) no son incorrelados ya que $\sigma_{ij}^2 = Cov(\hat{\alpha}_i, \hat{\alpha}_j) \neq 0$, con $i,j = 0,1,...,k$ y, por tanto, no son independientes. En particular, la varianza del estimador $\hat{\alpha}_i$ viene dada por

$$\sigma_{ii}^2 = Cov(\hat{\alpha}_i, \hat{\alpha}_i) = Var(\hat{\alpha}_i) = \sigma_i^2 = \sigma^2 q_{ii}, \quad i = 0, 1, \ldots, k, \quad (2.81)$$

siendo q_{ii} el elemento i-ésimo de la matriz $(\mathbf{X}^t\mathbf{X})^{-1}$.

- El estimador $\hat{\alpha}$ tiene distribución normal multivariante de orden $k+1$,

$$\hat{\alpha} \sim N_{(k+1)}\left(\vec{\alpha}, \sigma^2 (\mathbf{X}^t\mathbf{X})^{-1}\right). \quad (2.82)$$

- El estimador $\hat{\alpha}_i$ del parámetro α_i tiene la siguiente distribución normal

$$\hat{\alpha}_i \sim N(\alpha_i, \sigma^2 q_{ii}) \quad i = 0, 1, \ldots, k. \quad (2.83)$$

El parámetro α_i indica la influencia de la variable regresora x_i en la variable respuesta Y, representa el incremento que se produce en la variable respuesta por un crecimiento unitario en la variable regresora x_i.

Debe tenerse en cuenta que el valor de α_i está condicionado al modelo de regresión múltiple con el que se está trabajando y si se cambia el modelo (se eliminan variables regresoras o se introducen nuevas variables) el coeficiente α_i, asociada a la variable regresora x_i, también cambia.

Aceptar que el valor de α_i es cero equivale a aceptar que la variable x_i no está relacionada linealmente con la variable Y.

Si se conoce la varianza del modelo σ^2, utilizando las distribuciones expuestas se pueden calcular intervalos de confianza de los parámetros α_i, individuales o conjuntos (regiones de confianza del vector paramétrico $(\alpha_{j_1}, \alpha_{j_2}, \ldots, \alpha_{j_h})$, con $j_1,j_1,...,j_h \in \{0, 1, 2, \ldots, h\}$) o hacer contrastes de simplificación sobre estos parámetros. En la práctica casi nunca se conoce el parámetro σ^2 y es necesario estimarlo.

- *El estimador de la varianza:* Una hipótesis básica del modelo es que los errores son normales y homocedásticos, por tanto, $Var(\varepsilon_i)=\sigma^2$, $i=1,...,n$, el parámetro σ^2 normalmente es desconocido y es necesario estimarlo. El estimador de este parámetro es la varianza residual, definida como "el coeficiente entre la suma de residuos al cuadrado (scR) y el número de grados de libertad del modelo (gl)",

$$\hat{s}_R^2 = \frac{1}{n-(k+1)} \sum_{i=1}^{n} e_i^2. \quad (2.84)$$

El estimador $\hat{s}_R{}^2$ es distinto del estimador que se obtiene por máxima verosimilitud, $\sigma_{MV}{}^2$, dado por

$$\hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^{n} e_i^2. \quad (2.85)$$

La relación entre ambos estimadores es la siguiente:

$$\hat{s}^2_R = \frac{n}{n-(k+1)}\hat{\sigma}^2_{MV}. \quad (2.86)$$

El estimador $\hat{s}_R{}^2$ tiene la ventaja, respecto a $\hat{\sigma}_{MV}{}^2$, de ser insesgado.

Utilizando la hipótesis de normalidad se obtiene la siguiente relación que permite conocer la distribución de $\hat{s}_R{}^2$,

$$\frac{(n-(k+1))\,\hat{s}^2_R}{\sigma^2} \sim \chi^2_{n-(k+1)} \quad (2.87)$$

De esto se obtiene que un intervalo de confianza de σ^2 con un nivel de confianza 1-α es

$$\frac{(n-(k+1))\,\hat{s}^2_R}{\chi^2_{n-(k+1)}\left(1-\frac{\alpha}{2}\right)} \leq \sigma^2 \leq \frac{(n-(k+1))\,\hat{s}^2_R}{\chi^2_{n-(k+1)}\left(\frac{\alpha}{2}\right)} \quad (2.88)$$

donde $\chi_{n-}(k+1)^2(\theta)$ el número que verifica que $P\left(\xi \leq \chi^2_{n-(k+1)}(\theta)\right) = \theta$, siendo ξ una variable aleatoria con distribución $\chi_{n-}(k+1)^2$.

Sobre los coeficientes del modelo de regresión lineal múltiple se pueden realizar algunas inferencias.

De la distribución de $\hat{\alpha}_i$ dada se deduce

$$\hat{\alpha}_i \sim N\left(\alpha_i, \sigma^2 q_{ii}\right) \Rightarrow \frac{\hat{\alpha}_i - \alpha_i}{\sigma\sqrt{q_{ii}}} \sim N(0,1) \quad i = 0, 1, \ldots, k. \quad (2.89)$$

Como σ^2 no se conoce, se sustituye por su estimador $\hat{s}_R{}^2$, lo que permite obtener el siguiente estadístico

$$\omega_i = \frac{\hat{\alpha}_i - \alpha_i}{\sigma(\alpha_i)} = \frac{\hat{\alpha}_i - \alpha_i}{\hat{s}_R\sqrt{q_{ii}}}. \quad (2.90)$$

Además se deduce que la distribución de ω_i es $t_{n-(k+1)}$

$$\omega_i = \frac{\hat{\alpha}_i - \alpha_i}{\hat{s}_R\sqrt{q_{ii}}} \sim t_{n-(k+1)}, \quad i = 0, 1, \ldots, k. \quad (2.91)$$

Utilizando esto se obtiene que un intervalo de confianza para α_i a un nivel de confianza 1-α es

$$\alpha_i \in \hat{\alpha}_i \mp \hat{s}_R\sqrt{q_{ii}}\; t_{n-(k+1)}\left(1-\frac{\alpha}{2}\right), \quad i=0,1,\ldots,k, \tag{2.92}$$

donde $t_{n-}(k+1)(\theta)$ es el número que verifica que $P\left(\zeta \le t_{n-(k+1)}(\theta)\right) = \theta$, siendo ζ una variable aleatoria con distribución $t_{n-}(k+1)$.

2.2.2.2. Análisis de contrastes

El estadístico ω_i también puede utilizarse para realizar contrastes de hipótesis acerca de si la variable explicativa x_i influye "individualmente" o no en la variable respuesta Y (contrastes de simplificación). Aceptar que $\alpha_i = 0$ equivale a aceptar que la variable x_i no está relacionada linealmente con la variable Y, por tanto no debe estar en el modelo.

Se desea hacer el siguiente contraste (contraste individual de la t)

$$C_i \equiv \left\{ \begin{array}{lcl} H_0 \equiv & \alpha_i & = \quad 0 \\ H_1 \equiv & \alpha_i & \neq \quad 0 \end{array} \right\} \quad i=0,1,\ldots,k$$

Utilizando lo ya expuesto, si H_0 es cierto, se obtiene

$$\omega_i|_{H_0} = \hat{t}_i = \frac{\hat{\alpha}_i}{\hat{s}_R\sqrt{q_{ii}}} \sim t_{n-(k+1)}, \quad i=0,1,\ldots,k, \tag{2.93}$$

$\hat{t}_i$ representa la "discrepancia" entre la información que proporciona la muestra y la información que proporciona la hipótesis nula H_0.

Como el *p-valor* de este contraste bilateral es

$$p-valor = 2\,P\left(t_{n-(k+1)} > \left|\hat{t}_i\right| = \left|\frac{\hat{\alpha}_i}{\hat{s}_R\sqrt{q_{ii}}}\right|\right) \tag{2.94}$$

La región de aceptación del contraste a un nivel de significación α es

$$-\hat{s}_R\sqrt{q_{ii}}\; t_{n-(k+1)}\left(1-\frac{\alpha}{2}\right) \le \hat{\alpha}_i \le \hat{s}_R\sqrt{q_{ii}}\; t_{n-(k+1)}\left(1-\frac{\alpha}{2}\right) \tag{2.95}$$

"...El siguiente teorema de Gauss-Markov justifica la utilización de los estimadores mínimos cuadráticos, ya que, en este contexto, indica que estos estimadores son los "mejores" (los más eficaces) dentro de la clase de los estimadores lineales

insesgados. El teorema afirma que e*n la estimación del modelo de regresión lineal* $\vec{\mathbf{Y}} = \mathbf{X}\,\vec{\alpha} + \vec{\varepsilon}$ *si las perturbaciones* ε_i *son incorreladas, de igual varianza e independientes de las variables explicativas. Entonces los estimadores mínimo-cuadráticos* $(\hat{\alpha})$ *son "óptimos" o de mínima varianza dentro de la clase de los estimadores centrados que son funciones lineales de las observaciones,* y_i.

El Teorema de Gauss-Markov asegura que los estimadores mínimo-cuadráticos son los "mejores" dentro de la clase de estimadores que son insesgados y funciones lineales de las observaciones, pero no garantiza que estos estimadores sean mejores que otros estimadores que no pertenezcan a la clase anterior.

Por otra parte, al comparar estimadores se está utilizando el criterio de Error Cuadrático Medio (*ECM*), siendo

$$ECM\,(\hat{\alpha}_i) = E\,(\hat{\alpha}_i - \alpha_i)^2 = Sesgo^2\,(\hat{\alpha}_i) + Var\,(\hat{\alpha}_i)\,. \quad (2.96)$$

En la clase de los estimadores insesgados, el sesgo es cero. Por tanto

$$ECM\,(\hat{\alpha}_i) = E\,(\hat{\alpha}_i - \alpha_i)^2 = Var\,(\hat{\alpha}_i)\,. \quad (2.97)$$

Si los estimadores mínimo-cuadráticos son los de menor varianza también son los de menor *ECM*. Pero puede ocurrir que existan estimadores sesgados con menor varianza que los estimadores mínimo-cuadráticos de forma que tengan menor *ECM*.

Finalmente debe tenerse en cuenta que en este teorema no se exigen hipótesis sobre la distribución de los ε_i, tan solo que sean independientes y con la misma varianza..." [30]

Veamos ahora como descomponer la variabilidad de la variable de interés *Y* cuando se ajusta un modelo de regresión múltiple:

- *El contraste conjunto de la* ***F:*** Razonando como en el modelo de regresión lineal simple, en cada observación muestral se puede hacer la siguiente descomposición

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \qquad i = 1, 2, \ldots n. \quad (2.98)$$

En base a las propiedades geométricas del modelo y utilizando el Teorema de Pitágoras, se obtiene

[30] "Modelos Estadísticos aplicados", J. Vilar Fernández, Universidade da Coruña, España 2003

$$\overbrace{\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{g.l.=n-1}}^{\substack{Suma\ de\\ Cuadrados\\ Global\ (scG)}} = \overbrace{\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{g.l.=k}}^{\substack{Suma\ de\\ Cuadrados\\ Explicada\ (scE)}} + \overbrace{\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{g.l.=n-(k+1)}}^{\substack{Suma\ de\\ Cuadrados\\ \mathrm{Re}sidual\ (scR)}}$$

De esta igualdad se construye la correspondiente tabla ANOVA, Tabla 5.

Fuente de Variación	*Suma de Cuadrados*	*Grados de Libertad*	Varianzas
Por la recta	$scE = \sum_i (\hat{y}_i - \bar{y})^2$	k	$\hat{s}_e^2 = \frac{scE}{k}$
Residual	$scR = \sum_i (y_i - \hat{y}_i)^2$	$n-(k+1)$	$\hat{s}_R^2 = \frac{scR}{n-(k+1)}$
Global	$scG = \sum_i (y_i - \bar{y})^2$	$n-1$	$\hat{s}_Y^2 = \frac{scG}{n-1}$

Tabla 2.5. Tabla ANOVA del modelo de regresión múltiple

De esta tabla ANOVA se deduce el siguiente contraste acerca de la influencia "conjunta" del modelo de regresión en la variable respuesta.

- *Contraste de regresión múltiple de la F:* El contraste que se desea resolver es el siguiente

$$C_M \equiv \left\{ \begin{array}{ll} H_0 \equiv & \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0 \\ H_1 \equiv & \text{algún } \alpha_i \neq 0 \quad \text{para algún } i \end{array} \right\}$$

Si H_0 es cierto ninguna de las variables regresoras influye en la variable respuesta (el modelo no influye). En este supuesto se verifica que

$$\hat{y}_i \approx \bar{y} \Rightarrow scE \approx 0, \quad (2.99)$$

por ser ésta una medida absoluta se compara con la varianza residual, lo que lleva a utilizar como estadístico del contraste el siguiente

$$\hat{F}_M = \frac{\hat{s}_e^2}{\hat{s}_R^2}. \quad (2.100)$$

Bajo la hipótesis nula y por la hipótesis de independencia se sigue que $\hat{F}_M$ sigue una distribución F (Contraste de la F) con k y $n-(k+1)$ grados de libertad,

$$\hat{F}_M|_{H_0} = \frac{\hat{s}_e^2}{\hat{s}_R^2} \sim F_{k,n-(k+1)}. \quad (2.101)$$

De donde se deduce que *p-valor* del contraste es

$$p - valor = P\left(F_{k,n-(k+1)} \geq \hat{F}_M\right), \quad (2.102)$$

donde $F_{k,n-(k+1)}$ denota una variable aleatoria que sigue una distribución *F* con *k* y *n*-(*k*+1) grados de libertad. El contraste de la *F* es unilateral (de una cola) y generaliza el contraste de regresión expuesto para el modelo de regresión lineal simple.

Sí el valor crítico (*p-valor*) del contraste es grande (mayor que el nivel de significación α) se acepta H_0, que el modelo de regresión no es influyente y debe buscarse un modelo alternativo.

- Contrastes individuales de la F: El contraste individual de la *t* que permite contrastar la influencia individual de la variable x_i se deduce de la distribución del estimador $\hat{\alpha}_i$, pero también puede hacerse por medio de una tabla ANOVA, estudiando el incremento que se produce en la suma de cuadrados explicada por el modelo al introducir la variable regresora en estudio x_i.

Para ello, si se desea contrastar la influencia de la variable x_i, se ajusta el modelo de regresión completo, con las *k* variables regresoras y se calcula la suma de cuadrados explicada por el modelo (*scE*(*k*)). A continuación, se ajusta el modelo de regresión con *k*-1 variables, todas excepto la variable x_i. Se calcula la suma de cuadrados explicada por este modelo (*scE*(*k*-x_i)). Ahora se define la suma de cuadrados incremental debida a x_i como el valor

$$\triangle scE\left(x_i\right) = scE\left(k\right) - scE\left(k - x_i\right) \geq 0 \quad (2.103)$$

Este valor indica el aumento de la variabilidad explicada por el modelo al introducir la variable x_i. Para contrastar la influencia individual o no de x_i, se realiza el siguiente contraste,

$$C_i \equiv \left\{ \begin{array}{rcl} H_0 \equiv & \alpha_i & = & 0 \\ H_1 \equiv & \alpha_i & \neq & 0 \end{array} \right\} \quad i = 0, 1, \ldots, k$$

Como estadístico del contraste se utiliza

$$\hat{F}_i = \frac{\dfrac{\triangle scE\left(x_i\right)}{1}}{\hat{s}_R^2\left(k\right)}, \quad i = 0, 1, \ldots, k. \quad (2.104)$$

Bajo la hipótesis nula se verifica que $\hat{F}_i$ sigue una distribución *F* (Contraste individual de la *F*) con 1 y *n*-(*k*+1) grados de libertad.

$$\hat{F}_i|_{H_0} \sim F_{1,n-(k+1)}, \quad i = 0, 1, \ldots, k \quad (2.105)$$

Evidentemente, si H_0 es cierto, $\triangle scE(x_i) \approx 0$ y $\hat{F}_i$ tomará valores pequeños. Por tanto este contraste es unilateral siendo el *p-valor* del contraste el siguiente

$$p - valor = P\left(F_{1,n-(k+1)} \geq \hat{F}_i\right), \quad i = 0, 1, \ldots, k. \tag{2.106}$$

Este contraste proporciona exactamente el mismo resultado que el contraste individual de la *t*, ambos dan igual *p-valor*. Sin embargo este método presenta la ventaja adicional de poder utilizarse para contrastar la influencia de un subconjunto de *l* variables explicativas, con $l \leq k$, $\{x_{j_1}, x_{j_2}, \ldots, x_{j_l}\}$. En este caso el estadístico del contraste es

$$\hat{F}_I = \frac{\frac{\triangle scE(l)}{l}}{\hat{s}_R^2(k)} \sim F_{l,n-(k+l)} \tag{2.107}$$

"...En un modelo de regresión múltiple al hacer los contrastes sobre la influencia individual de cada una de las variables regresoras y el contraste sobre la influencia conjunta del modelo de regresión ajustado, pueden darse diversas situaciones (Tabla 2.6.).

Caso	Contraste Conjunto (F)	Contraste Individual
1	Significativo	Todos Significativos
2	Significativo	Alguno Significativo
3	Significativo	Ninguno Significativo
4	No Significativo	Todos Significativos
5	No Significativo	Alguno Significativo
6	No Significativo	Ninguno Significativo

Tabla 2.6. Posibles resultados del Contraste de la *F* en la regresión múltiple

Caso 1. Todas las variables explicativas influyen en la variable respuesta.

Caso 2. Influyen algunas variables explicativas, otras no.

Caso 3. Las variables explicativas son muy dependientes entre sí. Entonces, conjuntamente influyen, pero los coeficientes individuales tienen varianzas muy altas en relación con el valor de las estimaciones que son no significativas. Este problema se denomina multicolinealidad y se soluciona eliminando algunas variables regresoras del modelo.

Caso 4. Es otro caso de multicolinealidad, las variables son muy dependientes pero con una fuerte correlación negativa. Es poco frecuente.

Caso 5. Análogo al anterior.

Caso 6. Ninguna de las variables regresoras influye en la variable respuesta o la influencia no la detecta la muestra tomada..."[31]

2.2.2.3. Los coeficientes de correlación

Al ajustar un modelo de regresión múltiple a una nube de observaciones es importante disponer de alguna medida que permita medir la bondad del ajuste. Esto se consigue con los coeficientes de correlación múltiple.

En el estudio de la recta de regresión definimos el coeficiente de correlación lineal simple (o de Pearson) entre dos variables X e Y , como

$$r(X,Y) = \frac{s_{(X,Y)}}{s_X \; s_Y}, \quad (2.108)$$

donde $s_{(X,Y)}$ es la covarianza muestral entre las variables X e Y ; s_X y s_Y son las desviaciones típicas muestrales de X e Y , respectivamente.

En general cuando se ajusta un modelo estadístico a una nube de puntos, una medida de la bondad del ajuste es el coeficiente de determinación**,** definido por

$$R^2 = \frac{scE}{scG} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (2.109)$$

"...El grado de correlación existente entre tres o más variables se llama correlación múltiple...

A menudo es importante medir la correlación entre una variable dependiente y una variable independiente particular, cuando todas las demás variables se suprimen (indicado con frecuencia con la frase *quedando iguales las restantes*). Esto se consigue definiendo un *coeficiente de correlación parcial*..."[32]

Vemos entonces que si el modelo que se ajusta es un modelo de regresión lineal múltiple, a R se le denomina coeficiente de correlación múltiple y representa el porcentaje de variabilidad de la Y que explica el modelo de regresión.

Como $scE \leq scG$, se verifica que $0 \leq R^2 \leq 1$. Si $R^2=1$ la relación lineal es exacta y si $R^2=0$ no existe relación lineal entre la variable respuesta y las variables regresoras.

[31] "Modelos Estadísticos aplicados", J. Vilar Fernández, Universidade da Coruña, España 2003
[32] "Estadística", M. Spiegel, Mc Graw Hill, EEUU 1988.

El coeficiente de correlación múltiple *R* es igual al coeficiente de correlación lineal simple entre el vector variable respuesta $\vec{\mathbf{Y}}$ y el vector de predicciones $\hat{\mathbf{Y}}$,

$$R = r\left(\vec{\mathbf{Y}}, \hat{\mathbf{Y}}\right). \quad (2.110)$$

El coeficiente de correlación múltiple *R* presenta el inconveniente de aumentar siempre que aumenta el número de variables regresoras, ya que al aumentar *k* (número de variables regresoras) disminuye la variabilidad no explicada, algunas veces de forma artificial lo que puede ocasionar problemas de multicolinealidad.

"...no hay límite para la cantidad de variables que pueden aparecer en el modelo, siempre y cuando estas no estén relacionadas linealmente entre sí (multicolinealidad)..."[33]

Si el número de observaciones *n* es pequeño, el coeficiente R^2 es muy sensible a los valores de *n* y *k*. En particular, si $n = k + 1$ el modelo se ajusta exactamente a las observaciones. Por ello y con el fin de penalizar el número de variables regresoras que se incluyen en el modelo de regresión, es conveniente utilizar el coeficiente de determinación corregido por el número de grados de libertad**,** $\bar{R}^2$**.** Este coeficiente es similar al anterior, pero utiliza el cociente de varianzas en lugar del cociente de sumas de cuadrados. Para su definición se tiene en cuenta que

$$R^2 = \frac{scE}{scG} = 1 - \frac{scR}{scG} \quad (2.111)$$

Cambiando las sumas de cuadrados por varianzas se obtiene el coeficiente de determinación corregido por el número de grados de libertad, $\bar{R}^2$, definido como sigue

$$\bar{R}^2 = 1 - \frac{\hat{s}_R^2}{\hat{s}_Y^2} = 1 - \frac{\dfrac{1}{n-(k+1)}\displaystyle\sum_{i=1}^{n} e_i^2}{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad (2.112)$$

Ahora es fácil deducir la siguiente relación entre los dos coeficientes de determinación

$$\bar{R}^2 = 1 - \left(1 - R^2\right)\frac{n-1}{n-(k+1)} \Rightarrow \bar{R}^2 \leq R^2 \quad (2.113)$$

[33] "Modelos de demanda de transporte", Juan de Dios Ortúzar, Universidad Católica de Chile, Alfaomega, Chile 2000.

También es fácil relacionar el estadístico del contraste de regresión múltiple con el coeficiente de determinación, obteniendo

$$\hat{F}_M = \frac{\hat{s}_e^2}{\hat{s}_R^2} = \frac{R^2}{1-R^2}\frac{n-(k+1)}{k} \quad (2.114)$$

Consideremos ahora a $\{X_1, X_2, \ldots, X_k\}$ que es un conjunto de variables aleatorias, el coeficiente de correlación parcial entre X_i y X_j es una medida de la relación lineal entre las variables X_i y X_j una vez que se ha eliminado en ambas variables los efectos debidos al resto de las variables del conjunto $\{X_1, X_2, \ldots, X_k\}$. Al coeficiente de correlación parcial entre X_1 y X_2 se le denotará por $r_{12\cdot 3\ldots k}$.

Para una mejor interpretación de este concepto, considérese el conjunto de cuatro variables $\{X_1, X_2, X_3, X_4\}$, se desea calcular el coeficiente de correlación parcial entre las variables X_1 y X_2. Para ello, se procede de la siguiente forma,

- Se calcula la regresión lineal de X_1 respecto de X_3 y X_4

  $$X_1 = \alpha_0 + \alpha_3 X_3 + \alpha_4 X_4 + e_{1\cdot 34} \quad (2.115)$$

  donde $e_{1\cdot 34}$ son los residuos del ajuste lineal realizado.
- Se calcula la regresión lineal de X_2 respecto de X_3 y X_4

  $$X_2 = \beta_0 + \beta_3 X_3 + \beta_4 X_4 + e_{2\cdot 34} \quad (2.116)$$

  donde $e_{2\cdot 34}$ son los residuos del ajuste lineal realizado.
- El coeficiente de correlación parcial entre X_1 y X_2 es el coeficiente de correlación lineal simple entre las variables $e_{1\cdot 34}$ y $e_{2\cdot 34}$,

  $$r_{12\cdot 34} = r\left(e_{1\cdot 34}, e_{2\cdot 34}\right) \quad (2.117)$$

Por tanto, el coeficiente de correlación lineal se define siempre dentro de un conjunto de variables y no tiene interpretación ni sentido si no se indica este conjunto de variables.

Consideremos el conjunto de variables $\{Y, X_1, X_2\}$, entonces se verifica la siguiente relación entre los coeficientes de correlación lineal simple y el coeficiente de correlación parcial,

$$r_{Y1\cdot 2} = \frac{r_{Y1} - r_{Y2}\, r_{12}}{\sqrt{(1-r_{Y2}^2)(1-r_{12}^2)}} \quad (2.118)$$

En un modelo de regresión múltiple

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_k X_k + \varepsilon \quad (2.119)$$

se puede calcular fácilmente el coeficiente de correlación parcial entre la variable respuesta Y y una variable regresora X_i controlado por el resto de variables

regresoras. Para ello se utiliza el estadístico del contraste individual de la *t* respecto a la variable X_i y que se definió anteriormente como

$$\hat{t}_i = \frac{\hat{\alpha}_i}{\sigma(\hat{\alpha}_i)} = \frac{\hat{\alpha}_i}{\hat{s}_R\sqrt{q_{ii}}}, \quad i = 1, 2, \ldots, k, \tag{2.120}$$

obteniéndose la siguiente relación

$$r^2_{Yi\cdot C} = \frac{\hat{t}_i^2}{\hat{t}_i^2 + n - (k+1)}, \tag{2.121}$$

donde $C = \{1, 2, \ldots, i-1, i+1, \ldots, k\}$ el conjunto de índices de todas las variables regresoras excepto el índice *i*.

2.2.2.4. La multicolinealidad

Analicemos, al igual que como lo hicimos con la regresión lineal simple, los principales problemas que se pueden presentar en la construcción de un modelo de regresión múltiple:

- Multicolinealidad: las variables regresoras son muy dependientes entre sí, y es difícil separar su contribución individual al modelo. Como consecuencia los parámetros del modelo son muy inestables, con varianzas muy grandes.
- Error de especificación: el modelo de regresión no proporciona un buen ajuste a la nube de observaciones. Esto puede ser por diferentes motivos: la relación no es lineal, existen variables explicativas relevantes que no han sido incluidas en el modelo, etc. Por ello, cuando se dispone de un conjunto amplio de posibles variables explicativas, es importante disponer de algoritmos que seleccionen el subconjunto más adecuado de variables explicativas que se deben incorporar al modelo de regresión, así como de medidas que midan la bondad del ajuste.
- Falta de Normalidad: los residuos no son normales.
- Heterocedasticidad: la varianza no es constante.
- Existencia de valores atípicos o heterogéneos: existen datos atípicos que se separan de la nube de datos muestrales que pueden influir en la estimación del modelo de regresión o que no se ajustan al modelo.
- Dependencia (autocorrelación): existe dependencia entre las observaciones.

Veamos el caso de la multicolinealidad, en el modelo de regresión lineal múltiple

$$\vec{\mathbf{Y}} = \mathbf{X}\vec{\alpha} + \vec{\varepsilon}, \quad (2.122)$$

el estimador por mínimos cuadrados $\hat{\alpha}$ se obtiene resolviendo el sistema de ecuaciones

$$\left(\mathbf{X}^t\mathbf{X}\right)\hat{\alpha} = \mathbf{X}^t\mathbf{Y} \Rightarrow \hat{\alpha} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}. \quad (2.123)$$

Por tanto, para calcular $\hat{\alpha}$ es necesario invertir la matriz $(\mathbf{X}^t\mathbf{X})$. Se pueden dar las siguientes situaciones:

- Una (o más) de las columnas de la matriz de diseño **X**, $\vec{\mathbf{x}}_{.j}$, es una combinación lineal exacta de las otras columnas, esto es, una variable explicativa es combinación lineal de las otras. Entonces el *rang*(**X**)<*k*+1, el $|\mathbf{X}^t\mathbf{X}|$=0 y no existe $(\mathbf{X}^t\mathbf{X})^{-1}$. Por tanto el sistema $(\mathbf{X}^t\mathbf{X})\hat{\alpha}$=$\mathbf{X}^t\mathbf{Y}$ no tiene solución única. No se puede estimar unívocamente el vector $\hat{\alpha}$. Este sería el caso extremo de multicolinealidad que en la práctica no se suele dar.
- El caso opuesto al anterior se da cuando las variables regresoras son ortogonales. Esto es,

$$\vec{\mathbf{x}}_{.j}\vec{\mathbf{x}}_{.k} = \sum_{i=1}^{n} x_{ij}x_{ik} = 0 \quad \text{si} \quad i \neq j, \quad i,j = 1,2,\ldots k. \quad (2.124)$$

  En este caso los resultados del modelo de regresión se pueden interpretar sin ambigüedad. La matriz $\mathbf{X}^t\mathbf{X}$ es diagonal y la matriz $Var(\hat{\alpha})$ también es diagonal, lo que implica que los estimadores $\hat{\alpha}_i$, i = 1,2,...k, son incorrelados. El signo de $\hat{\alpha}_i$ es igual al signo del coeficiente de correlación $r(x_i, Y)$, y la contribución de la variable regresora x_i a R^2 es independiente de las otras variables regresoras que están incluidas en el modelo de regresión, esto es, si se elimina alguna variable regresora o se añade una nueva (ortogonal), la contribución de x_i es la misma.
- En la mayoría de las situaciones lo que ocurre es una situación intermedia entre los dos casos extremos anteriores. Esto es, existe una cierta relación entre las variables explicativas lo que hace que los estimadores $\hat{\alpha}_i$ estén correlacionados. Si está relación es muy fuerte porque dos o más variables regresoras "están próximas" a una relación de linealidad del tipo

$$\nu_1\vec{\mathbf{x}}_1 + \nu_2\vec{\mathbf{x}}_2 + \ldots + \nu_k\vec{\mathbf{x}}_k = \vec{\mathbf{0}}, \quad (2.125)$$

  siendo $\nu_1,\nu_2,\ldots,\nu_k$ números no todos iguales a cero. Entonces se tiene un problema de multicolinealidad.

Aunque exista problema de multicolinealidad, se puede ajustar y estimar el modelo de regresión lineal, pero con mucha variabilidad, en el sentido de que las varianzas de los estimadores de los coeficientes del modelo son muy altas, lo que afecta al estudio del modelo.

Desde otro punto de vista, comparando la $Var(\hat{\alpha}_i)$ cuando se utiliza el modelo de regresión lineal múltiple (*RLM*) con dos regresores y cuando se utiliza el modelo de regresión lineal simple (*RLS*) de un solo regresor. Se obtiene que

$$Var\,(\hat{\alpha}_i/RLM) = \frac{Var\,(\hat{\alpha}_i/RLS)}{1-r_{12}^2}, \quad i = 1, 2, \tag{2.126}$$

si existe alta multicolinealidad $1\text{-}r_{12}^2 \gtrsim 0$ y, por tanto, $Var(\hat{\alpha}_i/RLM) >> Var(\hat{\alpha}_i/RLS)$.

La última ecuación se generaliza para un modelo de regresión lineal con *k* variables regresoras, de la siguiente forma

$$Var\,(\hat{\alpha}_i/RLM) = \frac{Var\,(\hat{\alpha}_i/RLS)}{1-r_{i\text{-}resto}^2}, \quad i = 1, 2, \ldots k, \tag{2.127}$$

donde $r_{i\,resto}^2$ es el coeficiente de correlación múltiple entre la variable explicativa x_i y el resto de variables explicativas.

Se denomina factor de incremento de la varianza al número

$$FIV\,(x_i) = \frac{1}{1-r_{i\text{-}resto}^2}, \quad i = 1, 2, \ldots k. \tag{2.128}$$

Por tanto,

$$Var\,(\hat{\alpha}_i/RLM) = FIV\,(x_i)\,Var\,(\hat{\alpha}_i/RLS), \quad i = 1, 2, \ldots k, \tag{2.129}$$

De aquí se deduce que $Var(\hat{\alpha}_i/RLM) < Var(\hat{\alpha}_i/RLS)$, lo que implica que el modelo de regresión lineal simple estima con mayor precisión la influencia de la variable x_i en la variable respuesta que el modelo de regresión múltiple.

Si existe multicolinealidad, el $FIV(x_i)$ es muy grande y $Var(\hat{\alpha}_i/RLM)$ es mucho mayor que $Var(\hat{\alpha}_i/RLS)$.

De todo lo anterior se deduce que en un problema de regresión múltiple con fuerte multicolinealidad se verificará:

- Los estimadores $\hat{\alpha}_i$ tendrán varianzas muy altas y estarán muy correlacionados entre sí.
- Por la alta variabilidad de los estimadores $\hat{\alpha}_i$ puede ocurrir que los contrastes individuales (contrastes de la *t*) sean no significativos mientras que el contraste conjunto (contraste de la *F*) sea altamente significativo.

- La multicolinealidad normalmente afecta a unas variables y a otras no, por tanto, puede afectar a unos parámetros del modelo y a otros no.
- La multicolinealidad no afecta a las predicciones ($\hat{\mathbf{Y}}$), residuos ($\vec{\mathbf{e}}$), y varianza poblacional (σ^2).
- En resumen la multicolinealidad es un problema de la muestra de la que se quiere obtener más información de la que contiene.
- Se resuelve el problema de multicolinealidad eliminando del modelo las variables explicativas dependientes. Esto es, se deben eliminar del modelo aquellas variables que proporcionan una información que se obtiene de otras variables ya incluidas en el modelo.

En la Reseña Teórica 5, Anexo A, se citan las formas de detectar la mulicolinealidad.

2.2.2.5. Los residuos

Como hemos indicado anteriormente, el análisis de los residuos es básico para chequear si se verifican las hipótesis del modelo de regresión. Por ello, a continuación se exponen las propiedades matemáticas de los mismos. Consideremos el modelo de regresión lineal múltiple

$$\vec{\mathbf{Y}} = \mathbf{X}\,\vec{\alpha} + \vec{\varepsilon}. \quad (2.130)$$

Los residuos mínimo-cuadráticos vienen dados por

$$e_i = y_i - \hat{y}_i \quad i = 1, \ldots, n \quad (2.131)$$

o en forma matricial

$$\vec{\mathbf{e}} = \vec{\mathbf{Y}} - \hat{\mathbf{Y}}. \quad (2.132)$$

Como $\hat{\mathbf{Y}}$= $\mathbf{H}\vec{\mathbf{Y}}$, siendo $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ la matriz de proyección ortogonal. Es fácil probar que la matriz **H** es idempotente $(\mathbf{HH} = \mathbf{H})$ y simétrica $(\mathbf{H}^t = \mathbf{H})$. En base a esto

$$\vec{\mathbf{e}} = \vec{\mathbf{Y}} - \hat{\mathbf{Y}} = \vec{\mathbf{Y}} - \mathbf{H}\vec{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\vec{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})(\mathbf{X}\,\vec{\alpha} + \vec{\varepsilon}) \Rightarrow$$

$$\vec{\mathbf{e}} = \mathbf{X}\vec{\alpha} + \vec{\varepsilon} - \mathbf{HX}\vec{\alpha} - \mathbf{H}\vec{\varepsilon} = (\mathbf{I} - \mathbf{H})\vec{\varepsilon}, \quad (2.133)$$

donde se utilizó que $\mathbf{HX} = \mathbf{X}$. Se calcula la matriz de varianzas de los residuos,

$$Var(\vec{\mathbf{e}}) = (\mathbf{I} - \mathbf{H})\, E\left(\vec{\varepsilon}\vec{\varepsilon}^t\right)(\mathbf{I} - \mathbf{H})^t = \sigma^2 (\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})^t = \sigma^2 (\mathbf{I} - \mathbf{H}). \quad (2.134)$$

Por tanto, e_i es una variable aleatoria con distribución

$$e_i \sim N\left(0, \sigma^2 (1 - h_{ii})\right), \quad i = 1, \ldots, n, \quad (2.135)$$

donde h_{ii} es el valor de influencia de $\vec{\mathbf{x}}_i$ que mide la "distancia estadística" de $\vec{\mathbf{x}}_i$ a $\bar{\mathbf{x}}$. Un residuo "grande" indica que la observación está lejos del modelo estimado y, por

tanto, la predicción de esta observación es mala. Las observaciones con residuos grandes se denominan observaciones atípicas o heterogéneas (outliers).

Como los residuos tienen varianza variable y son dimensionados (tienen las unidades de la variable *Y*), normalmente se tipifican

$$\frac{e_i}{\sigma\sqrt{1-h_{ii}}}, \quad i = 1, \ldots, n, \tag{2.136}$$

los residuos tipificados siguen una distribución normal estándar, pero como σ^2 es desconocido, se sustituye por su estimador, la varianza residual $\hat{s}_R^2$ y se obtienen los residuos estandarizados, definidos como

$$r_i = \frac{e_i}{\hat{s}_R\sqrt{1-\nu_{ii}}}, \quad i = 1, \ldots, n, \tag{2.137}$$

Por la hipótesis de normalidad los residuos estandarizados siguen una distribución *t* con *n*-(*k*+1) grados de libertad. Como ya se indicó en el estudio del modelo de regresión lineal simple, en el cálculo de r_i existe el problema de que hay una relación de dependencia entre el numerador y el denominador de r_i. Para evitar esto, con mayor esfuerzo computacional, se calcula para cada *i*, *i* = 1,...,*n*, el estimador $\hat{s}_{R(i)}$, la varianza residual del modelo de regresión obtenido a partir de la muestra en la que se ha eliminado la observación $(\vec{\mathbf{x}}_i, Y_i)$. Ahora se definen los residuos estudentizados como

$$t_i = \frac{e_i}{\hat{s}_{R(i)}\sqrt{1-\nu_{ii}}} \sim t_{(n-1)-(k+1)} \quad i = 1, \ldots, n \tag{2.138}$$

Los residuos estudentizados siguen una distribución *t* con (*n*-1)-(*k*+1) grados de libertad. Si el tamaño muestral (*n*) es grande, los residuos estandarizados y los estudentizados son casi iguales y muy informativos, pudiéndose considerar grandes los residuos estandarizados tales que $|r_i| > 2$.

En la Reseña Teórica 6 del Anexo A, se pueden observar formas gráficas para el análisis de los residuos, las que se emplean en algunos casos más adelante.

2.2.2.6. Análisis de errores

Todo lo comentado acerca de las hipótesis básicas de normalidad, homocedasticidad e independencia del modelo lineal es válido aquí. En la Reseña Teórica 7 del Anexo A, se realiza un repaso de las características más importantes.

En cuanto a la hipótesis básica del modelo de regresión lineal múltiple que es que la variable respuesta *Y* se puede expresar como una combinación lineal de *k* variables

explicativas más un término de error ε. Se supone que el término de error es independiente de las *k* variables explicativas o, equivalentemente, que cualquier otra variable explicativa no incluida en el modelo y que pueda explicar a la variable *Y* es independiente de las variables explicativas del modelo. En la práctica no siempre es posible incluir todas las variables relevantes, bien porque alguna de estas variables no se considera relevante o porque no se puede medir. Otras veces se incluyen erróneamente variables irrelevantes o se especifica una relación lineal que no lo es. Todo ello conduce a especificar incorrectamente el modelo, resultando importante determinar la influencia de tales especificaciones incorrectas y tenerlas en cuenta en los resultados.

"...Los errores pueden clasificarse de la siguiente manera:

a) Errores de medición, codificación y digitación de los datos (que son típicamente mayores en países en desarrollo), que crecen con el refinamiento o sofisticación de las variables que deben medirse, pero que pueden reducirse invirtiendo más dinero en supervisión y entrenamiento, y en verificación de datos.
b) Errores de muestreo, provenientes de la consideración de muestras finitas en lugar de la población completa; estos errores tienden a ser proporcionales a la raíz cuadrada del número de observaciones (esto es, para reducirlos a la mitad seria necesario cuadruplicar la muestra), por lo que su reducción puede ser muy costosa.
c) Errores de especificación (por ejemplo, omisión de una variable relevante, forma funcional errónea, presencia de hábito o inercia de comportamiento), debidos a que ningún modelo puede pretender representar la realidad en forma exacta; un mejor modelo, más sofisticado, puede reducir este tipo de errores mediante una mayor inversión en la etapa de recolección y procesamiento de los datos.
d) Errores de calibración y predicción: los primeros provienen de la utilización de técnicas de calibración parcialmente inexactas (la generalidad de los modelos se resuelve en forma iterativa y no tiene solución matemática exacta) y los segundos, de errores en la predicción a futuro de las variables independientes del modelo.

e) Errores de transferencia, al usar un modelo desarrollado para A (cierta área o época) en B (otra área o época), aun con los ajustes necesarios.
f) Errores de agregación: la agregación no es un problema sencillo; sin embargo, si se estima un modelo agregado, los errores pueden ser mayores. Existen distintos tipos de agregación; por ejemplo, de información básica y de alternativas..."[34]

Debemos considerar los siguientes errores de especificación al ajustar un modelo de regresión múltiple:

- Omitir una variable relevante, alguna variable regresora de gran importancia no se ha incluido en el modelo. Este problema produce:

  - Que los estimadores mínimo cuadráticos $\hat{\alpha}$ sean sesgados y con mayor varianza salvo que la variable excluida sea ortogonal a las variables regresoras del modelo.

  - Que la varianza residual $\hat{s}_R^2$ sea un estimador sesgado por exceso ya que los errores son mayores de lo que serían si se hubiera incluido la variable excluida, sobre todo si esta variable es ortogonal a las variables regresoras, ya que entonces su influencia en la variable *Y* es mayor.

  - Como $\hat{s}_R^2$ es muy grande los intervalos de confianza de los parámetros del modelo son mayores de lo que deberían y los contrastes individuales de la *t* llevan a considerar como no significativas a variables regresoras que si lo son.

- Incluir una variable irrelevante, que no influye en la variable respuesta *Y* o que la información que proporciona sobre esta variable ya está contenida en las otras variables regresoras. Las consecuencias de este problema son las siguientes:

  - Si la variable irrelevante incluida depende de las otras variables regresoras se tiene un problema de multicolinealidad. Aumenta la varianza de los $\hat{\alpha}$, y los contrastes individuales de la *t* tienden a considerar como no significativas a variables regresoras que si lo son.

  - Si la variable irrelevante incluida es ortogonal a las otras variables regresoras, el efecto es menor, se pierde eficacia porque se pierde un grado de

[34] "Modelos de demanda de transporte", Juan de Dios Ortúzar, Universidad Católica de Chile, Alfaomega, Chile 2000.

libertad al aumentar una variable regresora que no aporta variabilidad explicada, pero para tamaños muestrales grandes el efecto es mínimo.

- Especificar una relación lineal que no lo es, proporciona malos resultados, sobre todo fuera del rango de valores observados porque una relación no lineal en un estrecho intervalo de observación se puede aproximar por una lineal. Las graves consecuencias de este error son las siguientes:

  - Los estimadores $\hat{\alpha}$ son sesgados y su varianza se calcula mal.

  - La varianza residual se calcula mal y los contrastes individuales de la *t* no son válidos.

  - Las predicciones del modelo son malas, sobre todo fuera del rango de valores de las observaciones.

Los errores de especificación se detectan utilizando los gráficos de residuos. Se deben tener en cuenta especialmente:

- El gráfico de residuos (e_i) frente a predicciones $(\hat{y}_i)$.
- El gráfico de residuos (e_i) frente a una variable explicativa (x_{ij}).
- El gráfico de residuos (e_i) frente a una variable explicativa omitida $(x_{i,omit})$. En muchas ocasiones se intuye que se debería incluir un término cuadrático o una interacción (producto) de variables explicativas, siendo razonable hacer el gráfico de los residuos frente a variables como x_{ij}^2 o $x_{ij} \cdot x_{ik}$.
- El gráfico de residuos (e_i) frente a la variable índice o tiempo (i) si las observaciones son recogidas secuencialmente y se sospecha que el tiempo puede ser una variable regresora.

2.2.2.7. Selección de las variables explicativas

En muchas situaciones se dispone de un conjunto grande de posibles variables regresoras, una primera pregunta es saber si todas las variables deben entrar en el modelo de regresión y, en caso negativo, se quiere saber qué variables deben entrar y qué variables no deben entrar en el modelo de regresión.

Intuitivamente parece bueno introducir en el modelo todas las variables regresoras significativas (según el contraste individual de la *t*) al ajustar el modelo con todas las variables posibles. Pero este procedimiento no es adecuado porque en la varianza del modelo $scR/(n-k-1)$ influye el número de variables del modelo, así como la $Var(\hat{\alpha}_i)$

crece al aumentar el número de regresores. Además puede haber problemas de multicolinealidad cuando hay muchas variables regresoras.

Para responder a estas preguntas se dispone de diferentes procedimientos estadísticos. Bajo la hipótesis de que la relación entre las variables regresoras y la variable respuesta es lineal existen procedimientos "paso a paso" (o stepwise) que permiten elegir el subconjunto de variables regresoras que deben estar en el modelo. También existen medidas de la bondad de ajuste de un modelo de regresión que permiten elegir entre diferentes subconjuntos de variables regresoras el "mejor" subconjunto para construir el modelo de regresión. Para la utilización de estas medidas de bondad de ajuste no es necesaria la hipótesis de linealidad. La utilización combinada de los algoritmos de selección de las variables regresoras y los criterios de bondad de ajuste permiten seleccionar adecuadamente el modelo de regresión que se debe utilizar. En todo caso, una vez elegido el modelo de regresión, antes de utilizarlo, se debe de contrastar que se verifican las hipótesis estructurales del modelo y si no se verifican, se debe reformular el modelo.

Los procedimientos para seleccionar las variables regresoras que deben entrar en el modelo se pueden observar en la Reseña Teórica 8 del Anexo A.

2.2.3. Conceptos complementarios

Las características de los modelos de regresión analizadas en este capítulo son aquellas que utilizamos más adelante en la aplicación de los datos, pero no son las únicas que han sido estudiadas por la estadística. Para complemento, ver la Reseña Teórica 9 del Anexo A.

Por último, nos parece adecuado resaltar que "...un modelo es más complejo que otro si tiene más operaciones del mismo tipo, o si tiene operaciones más explosivas en cuanto a error, o si tiene mayor número de variables. Se supone que un modelo se hace más complejo para reducir su error de especificación (e_s). Sin embargo, a medida que la especificación mejora, hay más variables que medir y mayores problemas en cuanto a su facilidad de medición; por lo tanto, se puede esperar que el error de medición (e_m) aumente. En la mayoría de los casos, éste crece rápido al principio, pero luego la curva de error se vuelve asintótica (se aplana).

Si definimos el error predictivo total $E=(e_m^2 + e_s^2)^{1/2}$, podemos ver que el mejor punto de predicción (el mínimo E) no corresponde al punto de máxima complejidad dado el e_m asociado... para predecir puede ser mejor un modelo más sencillo y robusto si los datos son de mala calidad. Sin embargo, para aprender y para entender el fenómeno, siempre va a ser más adecuado el modelo con la especificación más correcta..."[35]

[35] "Modelos de demanda de transporte", Juan de Dios Ortúzar, Universidad Católica de Chile, Alfaomega, Chile 2000.

Capítulo 3 - Análisis de datos

3.1.Obtención de los datos

Hemos justificado hasta ahora la necesidad de desarrollar un modelo para la estimación del *TMDA* en base a conteos diarios, hemos citado la forma habitual de realizar este cálculo cuando se cuenta con los coeficientes (o factores) de corrección y descrito los modelos de regresión, todo esto como pasos indispensables previo a la aplicación de los datos.

3.1.1. Análisis de formas

Volvamos ahora al estudio desde el punto de vista de la estadística de las series de tiempo.

"...Una *serie en el tiempo* es un conjunto de observaciones tomadas en instantes específicos, generalmente a intervalos iguales. Matemáticamente se define por los valores Y_1, Y_2,... de una variable Y en tiempos t_1, t_2,... así pues, Y es una función de t. Es interesante pensar una serie de tiempo como un punto moviéndose con el paso del tiempo resultado de una combinación de fuerzas económicas, sociológicas, psicológicas o de otros tipos. La experiencia con muchos ejemplos de series en el tiempo ha revelado ciertos *movimientos* o *variaciones características* que aparecen a menudo, y cuyo análisis es de gran interés por muchas razones, una de ellas el problema de *predicción* de futuros movimientos.

Los movimientos característicos de series en el tiempo se pueden clasificar en cuatro tipos principales, a menudo llamados *componentes* de una serie en el tiempo:

1) Movimientos a largo plazo o seculares; se refieren a la dirección general en la que el gráfico de una serie en el tiempo parece progresar en un largo periodo de tiempo, a veces se indica por una *curva de tendencia*.

2) Movimientos característicos o variaciones cíclicas; estas se refieren a las oscilaciones a largo término en torno a una recta o curva de tendencia. Estos *ciclos* pueden ser *periódicos* o no, es decir pueden seguir o no esquemas repetidos en intervalos iguales de tiempo.
3) Movimientos estacionales o variaciones estacionales; estos se refieren a los esquemas idénticos o casi idénticos que una serie en el tiempo parece seguir durante meses correspondientes en años sucesivos. Tales movimientos se deben a sucesos recurrentes que tienen lugar anualmente, tales como el brusco aumento de precios al consumo antes de la navidad.
4) Movimientos irregulares o aleatorios; estos se refieren a los movimientos esporádicos de las series en el tiempo debidos a sucesos de azar, tales como inundaciones, huelgas o elecciones. Si bien se puede suponer que tales sucesos producen variaciones que pierden su influencia tras poco tiempo, cabe la posibilidad de que sean tan intensos que den lugar a nuevos movimientos cíclicos de otro tipo...”[36]

El análisis de series en el tiempo consiste en describir matemáticamente los movimientos componentes que están presentes en ella. Supongamos que la serie en el tiempo tiene por variable *Y* el producto de varias variables *T, C, S* e *I* que producen los movimientos de tendencia, cíclicos, estacionales o irregulares, respectivamente.

$$Y = T \times C \times S \times I = TCSI \quad (3.1)$$

El investigar los factores *T, C, S* e *I*, se conoce a menudo como una *descomposición* de una serie en el tiempo. De estas cuatro componentes nos interesan en este análisis especialmente los factores *T* y *S.*

El método de los mínimos cuadrados se puede utilizar para hallar la ecuación de la curva de tendencia adecuada (*T*).

Para determinar el factor estacional *S* debemos estimar como varían los datos de la serie en el tiempo de mes a mes en un año típico. Así, definimos al índice estacional como un conjunto de números que muestran los valores relativos de una variable durante los meses del año, pudiéndose calcular por:

1) Método de porcentaje medio; en este método expresamos los datos de cada mes como porcentajes del promedio anual.

[36] “Estadística”, M. Spiegel, Mc Graw Hill, EEUU 1988.

2) Método del porcentaje de tendencia; en este método expresamos los datos para cada mes como porcentajes de valores de tendencia mensuales.
3) Método del promedio móvil en porcentaje; en este método calculamos un promedio móvil de 12 meses.
4) Método de la relación de enlace; en este método expresamos los datos para cada mes como un porcentaje de los datos para los meses previos; estos porcentajes mensuales se llaman *relaciones de enlace* porque relacionan cada mes con el precedente.

Por otro lado, la teoría del tránsito establece que "...los volúmenes de tránsito futuro se derivan a partir del tránsito actual *TA* y del incremento del tránsito *IT*... de acuerdo a esto se puede plantear:

$$TF = TA + IT \quad (3.2)$$

el tránsito actual *TA*, se puede establecer a partir de aforos vehiculares sobre las vialidades... el incremento del tránsito *IT* es el volumen de tránsito que se espera en el año futuro..."[37]

Esta forma de análisis lleva en la práctica a una falencia en cuanto al estudio de series desde la estadística, lo cual es analizado a continuación y es la causa de una nueva manera de encarar la temática, ya que es posible incluir otra conceptualización en los análisis de tránsito, que consideramos resulta una mejor adaptación a la realidad, y que involucra un nuevo paso en el tratamiento de los datos.

Habitualmente, tal vez por una cuestión de cálculos, cuando se aplica el concepto de incremento de tránsito se tiende a la simplificación de considerarlo como un hecho escalonado año a año. Es decir que conceptualmente se considera que durante el ciclo no existe un crecimiento propio del tránsito, lo que simplifica el análisis de series por anularse la componente generada por la tendencia. La interpretación gráfica de lo expuesto la podemos ver en la Figura 3.1, donde la *TCT* es la tasa de crecimiento del tránsito para los respectivos ciclos.

[37] "Ingeniería de tránsito, fundamentos y aplicaciones", R. Cal y Mayor, J. Cárdenas, Alfaomega 7°ed., México 1995.

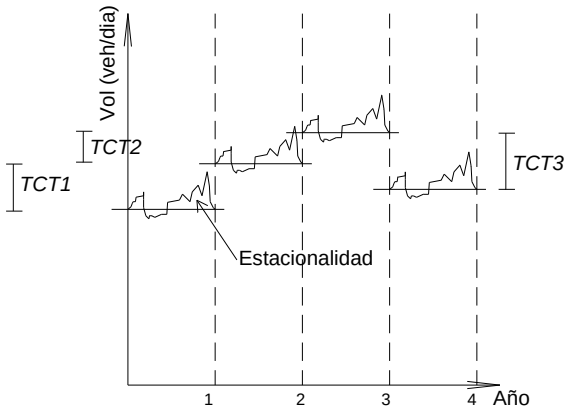

Fig. 3.1. Análisis tradicional del tránsito

En cambio, una visión análoga de lo que dicta la teoría de series de tiempos, seria la que se observa en la Figura 3.2, en donde se da una curva de tendencia (por ejemplo lineal) al largo plazo de la serie, de la cual se desprenden las estacionalidades.

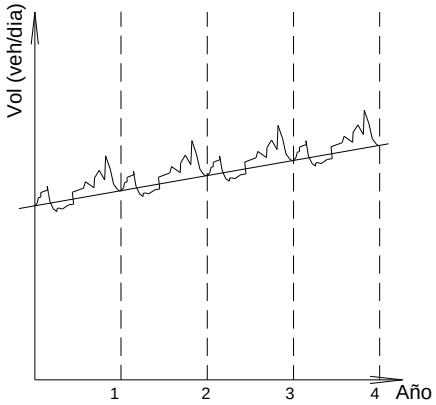

Fig. 3.2. Análisis según la estadística

El considerar los incrementos de tránsito en forma anual como valores de referencia, según se hace tradicionalmente, es una técnica por demás empleada y difundida, que tiene una buena adaptación para su empleo y que el profesional relacionado con la temática acepta intuitivamente. Lo que en cambio resulta difícil de aceptar es la idea de llevar al crecimiento del tránsito en un hecho escalonado en el tiempo. Resulta más realista el aceptarlo como algo gradual dentro del propio ciclo anual, en forma proporcional (crecimiento lineal), en donde la pendiente de lo que podríamos denominar “tendencia anual” lo refleja. Atendiéndose a lo que ilustramos en la Figura 3.3.

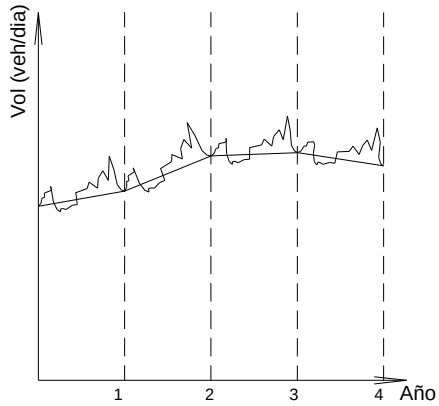

Fig. 3.3. Análisis propuesto

El tomar como base estos conceptos es lo que lleva a la necesidad de retrotraer los datos de tránsito para su comparación al día 1 del año, para que luego de ser realizados los cálculos necesarios puedan ser expandidos en función de la tasa de crecimiento.

Esta técnica nos lleva a situaciones como las que observamos en la Figura 3.4, en donde se ven con líneas punteadas las series correspondiente a los coeficientes de corrección mensuales para una misma vía pero en dos años diferentes. La línea punteada superior corresponde a un año con tasa de crecimiento del tránsito positiva, mientras que la línea inferior corresponde a uno con tasa negativa, ambas situaciones muy comunes en la base de datos recolectada para el estudio. Una vez desafectadas las series por las tasas de crecimiento, es decir como si la tasa fuera 0, obtenemos la serie de línea continua intermedia. Así, vemos como dos series que podrían suponerse en un principio responden a distintas demandas características, son en realidad comparables cuando de ellas se elimina un elemento que puede interpretarse como coyuntural, como lo es el valor en si de la tasa de crecimiento del tránsito.

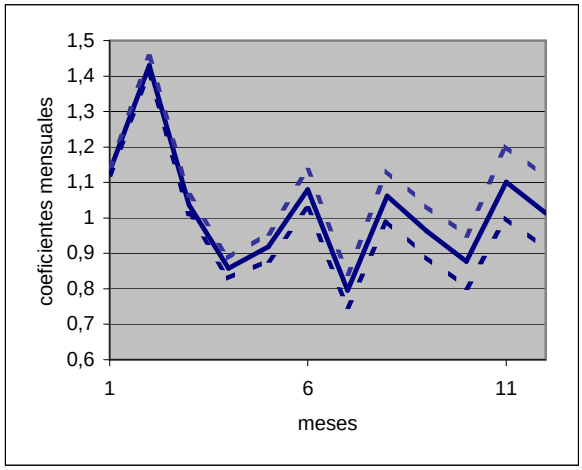

Fig. 3.4. Series de datos con crecimiento descontado

Todas estas consideraciones agregan un término adicional a la forma tradicional cálculo del *TMDA*, ya que al encontrarse los datos librados de su "tendencia anual" es necesario incluirla al final del cálculo. Por ello, en rasgos generales, podemos aventurar que el modelo en desarrollo lleva la siguiente forma:

$$TMDA = TD_0 \; x \; ALG_{COEF.D} \; x \; ALG_{COEF.M} \; x \; ALG_{TCT} \quad (3.3)$$

Donde:

TD_0 = Es el tránsito diario determinado por el censo esporádico, descontado el crecimiento de tránsito registrado en el año hasta el día en que se realiza la medición.

$ALG_{COEF.D}$ = Es el algoritmo que permite obtener el coeficiente de corrección diario que lleva el valor de TD_o a la media mensual, en donde la variable independiente es aquella que toma valor 1 para el domingo, 2 para el lunes, …, y 7 para el sábado.

$ALG_{COEF.M}$ = Es el algoritmo que permite obtener el coeficiente de corrección mensual que lleva el valor de la media mensual a la media anual, en donde la variable independiente es aquella que toma valor 1 para enero, 2 para febrero, …, y 12 para diciembre.

ALG_{TCT} = Es el algoritmo que permite llevar a ese valor medio anual al *TMDA* por afectarlo de la tasa de crecimiento de tránsito supuesta para ese año en estudio.

Cabe aclarar que esta forma de modelo arrastra características del modelo tradicional, pues "...se destaca que en el modelo de estimación del *TMDA* que puede denominarse tradicional, existe cierta inconsistencia formal, ya que se utilizan factores de desestacionalización multiplicativos y los promedios se realizan por medias aritméticas, cuando resultaría más convenientes el uso de medias geométricas..."[38]

Si bien esto es así, decidimos seguir por esta línea de trabajo, pues llevar el análisis a otro tipo de modelaciones existentes daría como resultado consideraciones de difícil aceptación, aun para profesionales con cierta formación matemática, perdiéndose la característica de uso difundido que pretendemos y planteamos para el modelo desde un principio.

3.1.2. Delimitación del área de estudio y antigüedad de los datos

Como ya dijéramos para el presente trabajo hemos establecido como área en estudio la conformada por las provincias de Buenos Aires, Córdoba, Santa Fe, La Pampa y Entre Ríos. Estas provincias de la región central de la Argentina son seleccionadas por conformar una región relativamente homogénea, cuando se la analiza desde el punto de vista socioeconómico.

Además establecemos estos límites, si bien en casos extremos las diferencias pueden resultar importantes, por ciertas similitudes en aspectos geográficos y climáticos de las zonas abarcadas, similitudes que se tornan notorias si se realiza la comparación con otras provincias pertenecientes al cordón montañoso de los Andes, al extremo sur patagónico o al norte subtropical.

"...Mediante el estudio del relieve y del clima es posible determinar las distintas regiones geográficas argentinas. Pero además de la forma y de las características de la superficie terrestre, es necesario tener en cuenta, como factor determinante, el tipo de actividad económica que se desarrolla. La semejanza y homogeneidad del relieve, el clima, la flora, la fauna, el suelo, los recursos naturales y el uso que el hombre da a la tierra contribuyen a definir la extensión y los límites aproximados –y a veces transitorios– de una unidad geográfica..."[39]

---

[38] "Estimación de cambios en el volumen de tránsito a causa del cobro de peaje en rutas de acceso a Córdoba", P. Arranz, F. Marhuenda, E. Masciarelli, XIV Congreso Argentino de Vialidad y Transito, Argentina 2005.

[39] "La Región Pampeana", R. Lima Coimbra, monografía, UNCPBA, Argentina 2003.

Como unidad de análisis fijamos al partido (o departamento) en los que se encuentran divididas las provincias, esta decisión surge como un balance entre la precisión deseada en el estudio y la exactitud alcanzable con los datos disponibles.

"...Los términos exactitud y precisión implican conceptos independientes pero complementarios. La exactitud se relaciona con el alcanzar una respuesta correcta, mientras que la precisión se relaciona con la magnitud del rango de estimación del parámetro en cuestión...

Como un ejemplo de exactitud se puede considerar un método aplicado para estimar una medida de performance. Si la medida de performance es la demora, un método exacto puede proveer una estimación muy aproximada de la demora actual bajo condiciones de campo. La precisión es una estimación del rango aceptable para una perspectiva de análisis proveyendo una estimación exacta..."[40]

Complementario a esto resulta interesante observar que "...la mayoría de los procedimientos estadísticos utilizados en la estimación de modelos asumen, implícita o explícitamente, que tanto los datos como la forma funcional del modelo se conocen en forma exacta. En la práctica, sin embargo, a menudo se violan estas condiciones; por lo demás, aun si se satisficieran, existirían errores en las predicciones de los modelos debido simplemente a la inexactitud de los valores estimados para las variables exógenas en el año de diseño...

Como el objetivo último de la modelación es normalmente la predicción, un problema importante que deben enfrentar los diseñadores de modelos es definir la combinación más adecuada de complejidad computacional y precisión de los datos, para lograr un nivel de exactitud de los resultados en relación al presupuesto del estudio..."[41]

Pasemos al tratamiento de los límites para las series históricas de datos a emplearse. Como comenzamos el análisis de datos a finales del año 2004, y por cuestiones de antigüedad y disponibilidad, decidimos emplear para la obtención del modelo las series comprendidas entre los años 1993 y 2003, destinando los datos

[40] "Highway Capacity Manual 2000", Transportation Research Board, National Research Council, EEUU 2000.

[41] "Modelos de demanda de transporte", Juan de Dios Ortúzar, Universidad Católica de Chile, Alfaomega, Chile 2000.

correspondientes al año 2004, si los hubiera, exclusivamente a la validación de los algoritmos resultantes.

Con estos límites establecidos procedemos a la consulta de diversas fuentes para la obtención de registro, resultando fructífera la consulta con:

- Subsecretaria de Tránsito y Transporte del Gobierno de la Ciudad de Buenos Aires
- Dirección Nacional de Vialidad
- Coviares
- Autopistas del Oeste
- Autopistas del Sol
- UTN Bahía Blanca
- Municipalidad de Concordia
- AUFE
- Puente Subfluvial Hernán Darias
- Dirección de Vialidad de Santa Fe
- Dirección de Vialidad de Buenos Aires
- Municipalidad de Rafaela
- Auditoria General de la Nación
- Instituto Superior del Transporte de la UNC
- LEMaC UTN La Plata
- INDeC
- Particulares

Este relevamiento permite obtener el mapa de cobertura que se observa en la Figura 3.5.

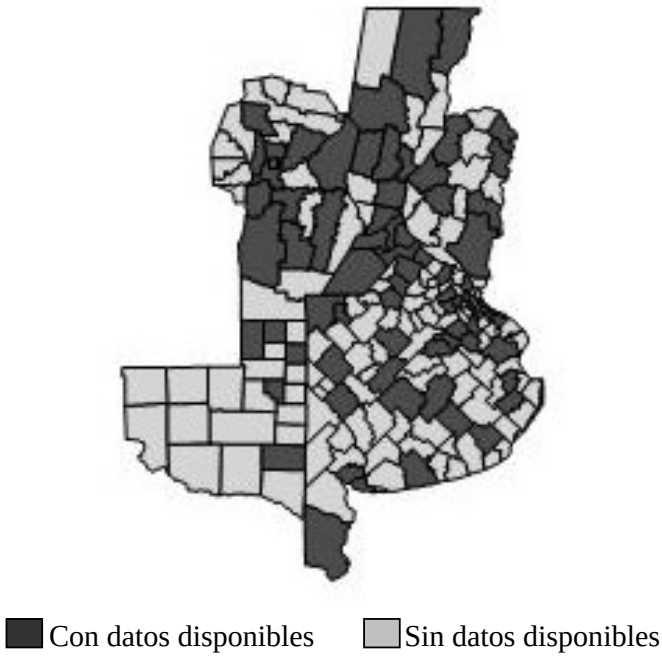

Fig. 3.5. Mapa de cobertura de los datos recabados

3.1.3. Elaboración de la matriz homogénea

Los datos recolectados de las fuentes citadas se encuentran expresados de muy diversas formas, poniendo en evidencia la falta de un procedimiento generalizado de orden nacional al respecto.

Esta heterogeneidad en las características de los datos nos genera la necesidad de una fuerte tarea de reconversión para su inclusión en una matriz general, que por su nueva condición denominamos homogénea.

En realidad esta matriz puede ser también analizada como un grupo de matrices, ya que por la estructura elegida para el trabajo debemos contar como resultado final con matrices que permitan efectuar las regresiones para la obtención de:

- Algoritmos de tasa de incremento de tránsito
- Algoritmos de coeficientes de corrección diaria
- Algoritmos de coeficientes de corrección mensual

Para la elaboración de estas matrices establecemos como campos, aunque algunos de estos luego no sean empleados en el modelo final, los siguientes:

- Denominación de la vía
- Denominación del punto de registro en la vía

- Localidad de ubicación del punto de registro
- Provincia en la que se encuentra la localidad
- Fuente del dato
- Característica de urbano o rural del punto
- Características de turística o comercial de la vía
- Existencia o no de peaje en el tramo
- Coeficientes de corrección diaria (en siete campos, 1 para domingo y 7 para sábado)
- Coeficientes de corrección mensual (en doce campos, 1 para enero y 12 para diciembre)
- Año del registro
- Incremento de tránsito registrado durante ese año
- Clasificación del tránsito en automóviles y camionetas, ómnibus, camión liviano y camión pesado

A estos registros se suman los datos recabados de diversas variables que reflejan la actividad socioeconómica del área en estudio, según lo detallamos más adelante cuando describimos la obtención del algoritmo para la tasa de incremento del tránsito.

Para la conformación de la matrices numéricas sobre las que realizamos las regresiones establecemos:

- Para la característica de urbanidad (urbano o rural) incluimos una variable que toma valor 1 cuando el entorno es urbano y 0 cuando es rural. Consideramos entorno rural cuando no se genera con densidad la actividad residencial y/o comercial, es decir que la accesibilidad se da desde sectores de actividad rural no cotidiana.
- Para el uso, incluimos una variable que toma valor 1 cuando es comercial y valor 0 cuando es turístico. Consideramos uso turístico cuando éste es el preponderante en época de vacaciones, es decir que la vía sirve evidentemente de vinculación a plazas turísticas (por ejemplo la Autovía 2, que sirve al tránsito entre Capital Federal y la Costa Atlántica, o la Ruta Provincial Nº 5 que sirve al tránsito entre Córdoba y Zona Serrana).
- Para la existencia de peaje incluimos una variable binaria que toma valor 1 si existe cobro de peaje y 0 si no existe.

- Para la clasificación del tránsito incluimos una variable dada por el porcentaje de vehículos livianos circulantes (automóviles y camionetas).

Cabe recordar en este punto que para el cálculo de los coeficientes de corrección diarios y mensuales efectuamos con cada grupo de datos, y en función de su forma de expresión, las operaciones que permiten luego asegurar que:

$$TMDA_0 = TD_0 \; x \; COEF_{DIARIO} \; x \; COEF_{MENSUAL} \quad (3.4)$$

Donde tanto $TMDA_0$ y TD_0 son valores con tendencia discriminada.

3.2. Empleo de los datos

3.2.1. Obtención de los algoritmos para el incremento del tránsito

Como hemos dicho, a los datos incluidos en las matrices a ser empleadas en las regresiones se les ha discriminado la tendencia. Esto nos lleva a la necesidad de incorporar posteriormente un término que considere el incremento del tránsito.

"...Para obtener estimativos confiables de los volúmenes vehiculares que circularán en el futuro se utilizarán modelos, los cuales son alimentados utilizando parámetros socioeconómicos (como la población total, la población económicamente activa, la población ocupada y los vehículos registrados)..."[42]

El incremento de tránsito es estimable entonces en función de variables socioeconómicas exógenas, lo cual puede expresarse matemáticamente como cuando se asegura que "...el modelo general de demanda del transporte responde a una estructura multiplicativa del tipo:

$$\Delta T = (\Delta VAR_1)^{\alpha} (\Delta VAR_2)^{\beta} \quad (3.5)$$

donde:

ΔT = es la variación porcentual del tránsito considerado

ΔVAR_1, ΔVAR_2 = representan la variación de las variables explicativas o causales del aumento de tránsito

α y β = son las elasticidades del tránsito respecto a las variables independientes

...incluyendo parámetros como:

ΔPOB = variación porcentual de las poblaciones que sirve cada ruta

ΔPAR = variación porcentual del parque de automóviles

ΔPBI = variación porcentual del Producto Bruto Interno Nacional..."[43]

El problema que se nos plantea ahora es saber qué variables socioeconómicas elegir para explicar el incremento del tránsito, resultando a la vez posibles de ser recabadas con relativa facilidad.

"...Respecto a las variables exógenas, se analizaron diversas posibilidades en base a las siguientes pautas:

a) Que la posible variable a incluir en el estudio tuviera series de duración y desagregación acorde a la serie temporal de tránsitos
b) Que la posible variable a incluir en el estudio fuera más fácil de predecir que el propio tránsito...

Entre las posibles variables identificadas a nivel nacional y regional, se encontraron algunas que no tenían la desagregación adecuada, que resultaban solamente de mediciones en aglomerados urbanos, que no poseían una longitud acorde o que estaban afectadas de falta de datos, etc., en la mayoría de los casos resultaban de más difícil predicción que el mismo tránsito. Entre todas ellas, el *PBI* (Producto Bruto Interno) a precios constantes, en desagregación trimestral resultó la más indicada. Aun cuando el PBI resulta, en si misma, una variable cuya predicción puede resultar azarosa, pueden encontrarse más fácilmente referencias acerca de su probable evolución en publicaciones de organismos e instituciones privadas..."[44]

En resumen, la predicción de las tasas de crecimiento del tránsito en lugares en los que ya se cuenta con datos previos es relativamente sencilla, ya que puede establecerse una tendencia observando los ciclos anteriores. Cuando el análisis se

[42] "Ingeniería de tránsito, fundamentos y aplicaciones", R. Cal y Mayor, J. Cárdenas, Alfaomega 7°ed., México 1995.
[43] "Censos y proyecciones de tránsito de la red de accesos a Córdoba", Instituto Superior de Ingeniería de Transporte, Universidad Nacional de Córdoba, Argentina 1996.
[44] "Estudio econométrico y pronóstico del tránsito que pasa por casillas de peaje en concesiones viales de Argentina", P. Arranz, E. Masciarelli, F. Marhuenda, ISIT, Universidad Nacional de Córdoba, Argentina 2004.

realiza en cambio por profesionales no directamente vinculados a la temática o en lugares en donde no se cuenta con series históricas de tránsito, razones que son justamente las causas del estudio, la predicción se debe realizar en forma subjetiva o empleando otros parámetros relacionados de más sencilla predicción, o que son comúnmente supuestos en estudios socioeconómicos para una amplia gama de estimaciones.

Buscamos entonces correlacionar la tasa de crecimiento del tránsito con diversas variables socioeconómicas explicativas, pero cuando efectuamos el análisis sobre estos parámetros registrados en las localidades del área en estudio y en ciclos seleccionados para el trabajo, nos encontramos con diversos inconvenientes.

En un principio obtenemos del *INDEC* (Instituto Nacional de Estadísticas y Censos) datos discriminados por año de las variables de: *PIB*, Importaciones, Consumo Privado, Consumo Público y Exportaciones. Pero los mismos no se encuentran discriminados por provincias, ni mucho menos por localidades.

También recabamos datos en cuanto a la variación de la población en las diversas localidades, pero sólo contamos con los registros producidos por los censos de 1991 y 2001, razón por la cual la variable no es aplicable, pues carecemos de valores intermedios.

Otro dato que obtenemos es el de la evolución de la canasta familiar, que nuevamente se da discriminado por año, pero no por provincia o localidad.

Finalmente, y luego de descartar otras series de datos por razones similares a las ya enunciadas, obtenemos la serie correspondiente a la tasa de actividad discriminada por año y en las diversas regiones de las provincias. Mediante estos datos confeccionamos una matriz de variaciones de tasa de empleo, para ser utilizada en la regresión junto con los datos de tasa de crecimiento del tránsito.

3.2.1.1. Variación del empleo como variable independiente

Comencemos planteando las variables para la regresiones:

X = variación tasa de empleo (variable independiente)

Y = crecimiento tránsito (variable dependiente)

Con las cuales podemos construir el gráfico de la Figura 3.6, en donde cada par de datos es un punto de la nube de puntos general.

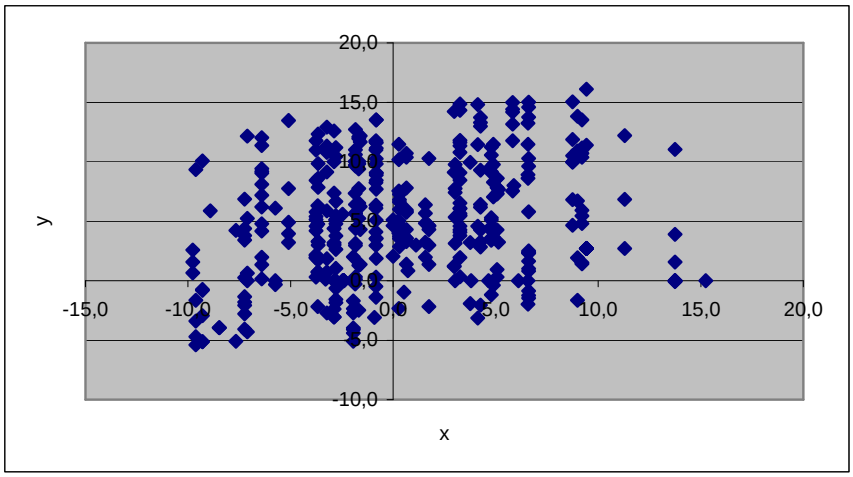

Fig. 3.6. Gráfico de variación tasa de empleo vs crecimiento del tránsito

Debemos ahora analizar la validez de las muestras disponibles, para esto consideramos que "...para analizar la muestra podemos construir una tabla que incluya las medidas de tendencia central, medidas de variabilidad, y medidas de forma de ésta. Son de particular interés entre estas medidas los coeficientes de asimetría y curtosis estandarizados que pueden utilizarse para determinar si la muestra procede de una distribución normal. Los valores de estos estadísticos fuera del rango de -2 a +2 indican alejamiento significante de normalidad que tendería a invalidar cualquier test estadístico con respecto a la desviación normal..."[45]

Siguiendo lo establecido en este párrafo realizamos el análisis estadístico de los datos obtenidos de tasa de crecimiento de tránsito, resultando:

Frecuencia = 518

Media = 5,2417

Varianza = 81,6331

Desviación típica = 9,0351

Asimetría tipificada = 0,296483

Curtosis tipificada = 2,31466

En este caso, el valor del coeficiente de asimetría estandarizado está dentro del rango esperado para los datos de una distribución normal y el valor del coeficiente de

[45] "Estadística Básica Aplicada", A. Fernández Morales y B. Lacomba Arias, Ágora Universidad, España 2004.

curtosis estandarizado se encuentra levemente fuera del rango. Este análisis nos permite tener además gráficas de dispersión y de caja y bigotes, que dan una referencia visual de la distribución de la muestra, los que se observan en la Figura 3.7 y la Figura 3.8.

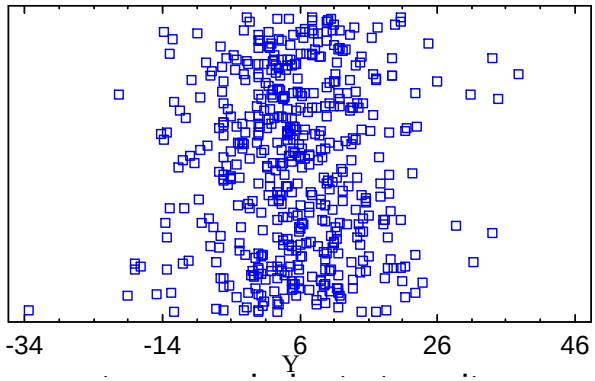

Fig. 3.7. Gráfico de dispersión de la tasa de crecimiento del tránsito

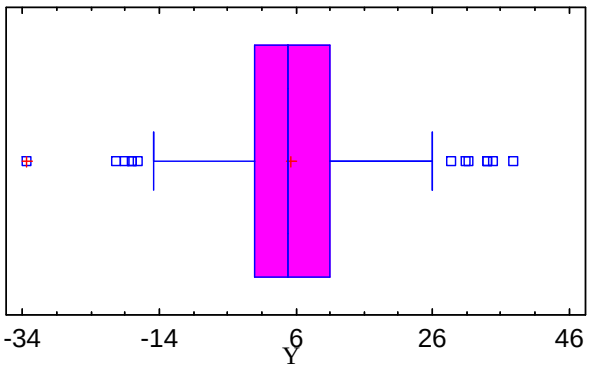

Fig. 3.8. Gráfico de caja y bigotes de la tasa de crecimiento del tránsito

No obstante la distribución de nuestros datos se nos presenta levemente leptocúrtica, decidimos no modificar la muestra hasta no analizar los ajustes obtenidos con las regresiones.

Analizamos estadísticamente ahora los datos de la variación del empleo, resultando:

Frecuencia = 518

Media = 0,17278

Varianza = 30,6671

Desviación típica = 5,53779

Asimetría tipificada = 2,24557

Curtosis tipificada = -2,3667

En este caso se supera levemente el umbral de asimetría y curtosis, pero con idéntico fin que en el caso anterior, decidimos continuar con la muestra para el análisis de regresión.

Las gráficas de dispersión y de caja y bigotes también nos permiten observar la distribución de estos datos, tal cual observamos en la Figura 3.9 y la Figura 3.10.

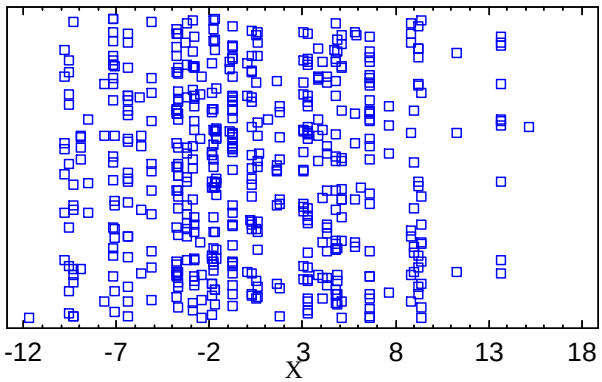

Fig. 3.9. Gráfico de dispersión de la variación del empleo

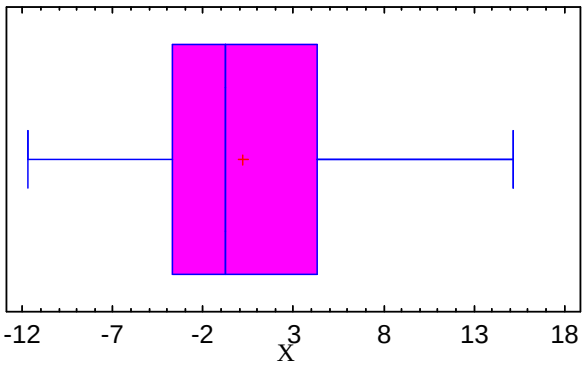

Fig. 3.10. Gráfico de caja y bigotes de la tasa de variación del empleo

Validadas las muestras desde el punto de vista de su normalidad, pasamos al análisis de regresión.

La lectura de la nube de puntos obtenida en el gráfico de tasa de crecimiento de tránsito vs. variación de empleo (Figura 3.6), nos permite establecer que no existe a simple vista una clara forma de relación entre ambas variables. Por esto encaramos el estudio analizando la regresión lineal simple, para lo que utilizamos el "módulo de regresión" del programa Microsoft Excel, obteniendo como resultado:

- Función de regresión $Y = 0{,}21\ X + 4{,}92$
- Coeficiente de correlación múltiple 0,23

- Coeficiente R^2 de determinación 0,05
- Coeficiente R^2 ajustado 0,05
- Error típico 4,78

Como vemos los resultados obtenidos no son buenos, ya que R^2 es muy bajo (debería acercarse a 1) y el *r* se encuentra cercano a 0 (debería acercarse a –1 o 1).

Debemos observar ahora si analizando los residuos obtenidos podemos encontrar las causas de tan bajo ajuste. El gráfico de los residuos frente a las predicciones $(\hat{y}_i, e_i)$ es el que proporciona una mayor información acerca del cumplimiento de las hipótesis del modelo, como lo podemos ver en los siguientes casos:

- No se detecta ningún problema. (Figura 3.11)

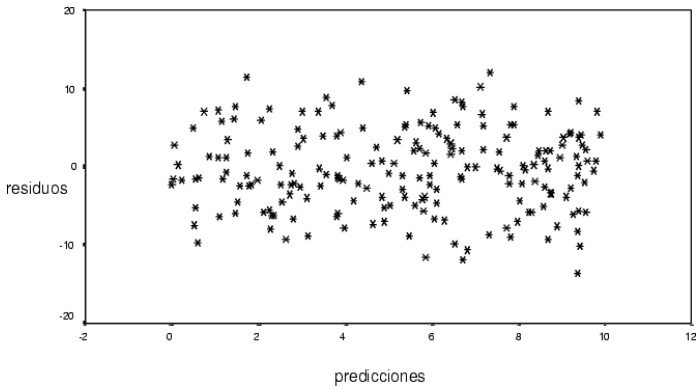

Fig. 3.11. Ejemplo de gráfico de residuos sin indicios de problemas

- El ajuste lineal no es adecuado. (Figura 3.12)

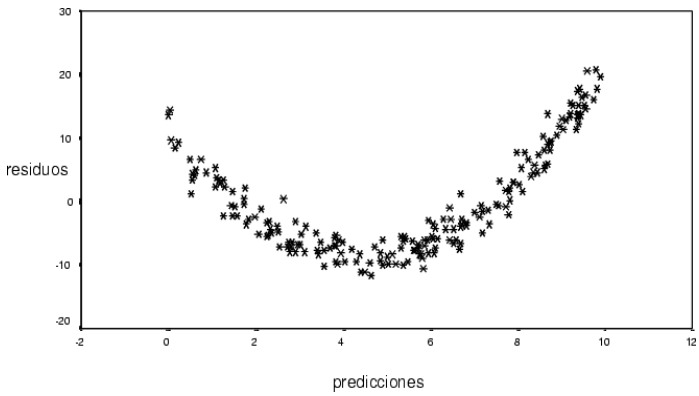

Fig. 3.12. Ejemplo de gráfico de residuos con ajuste lineal no adecuado

- Ajuste lineal mal calculado. (Figura 3.13)

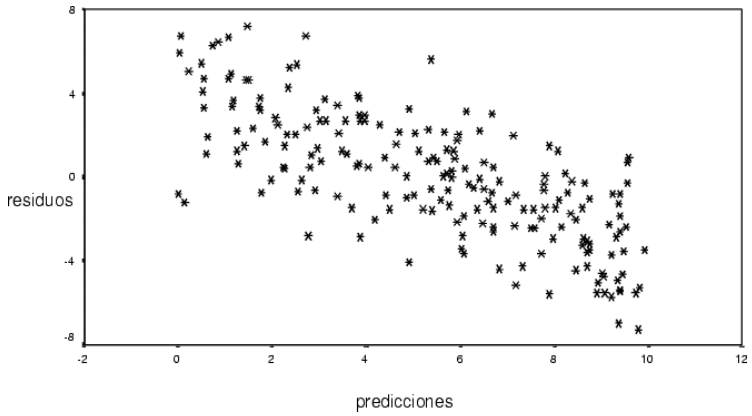

Fig. 3.13. Ejemplo de gráfico de residuos con ajuste mal calculado

- Existe heterocedasticidad. (Figura 3.14)

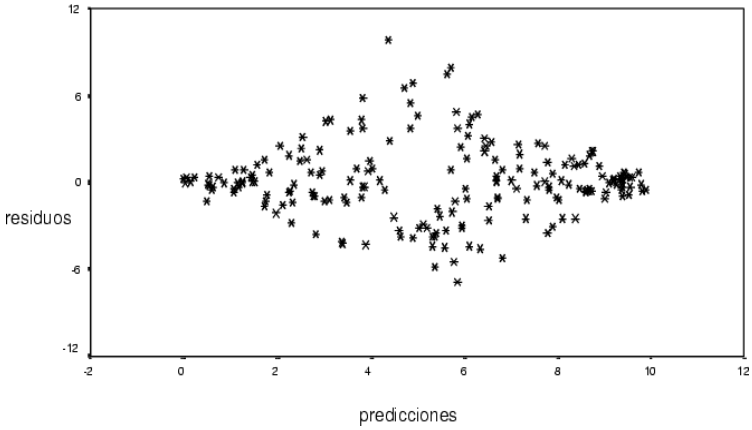

Fig. 3.14. Ejemplo de gráfico de residuos con heterocedasticidad

- Existencia de datos atípicos. (Figura 3.15)

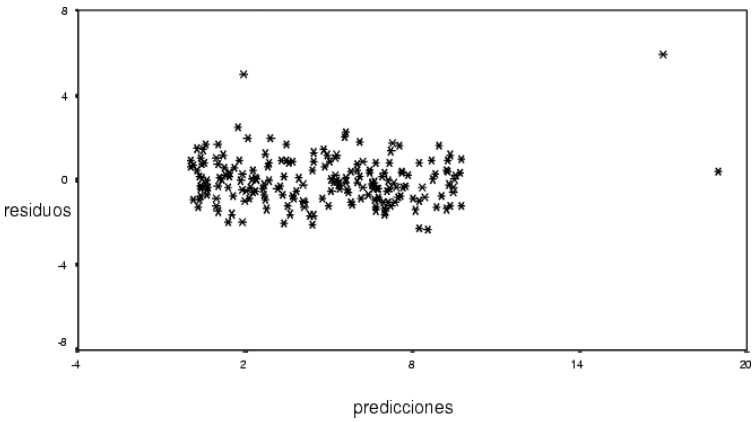

Fig. 3.15. Ejemplo de gráfico de residuos con datos atípicos

Cuando efectuamos el análisis con el gráfico de residuos vs. predicciones de esta regresión, Figura 3.16, no observamos la existencia de ninguno de estos problemas enunciados.

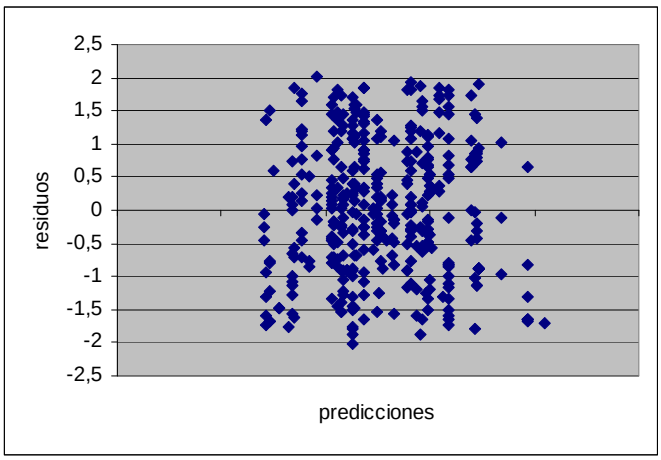

Fig. 3.16. Gráfico de residuos vs predicciones

Pasamos entonces a analizar si la baja correlación se debe a datos atípicos (outliers). Normalmente se considera que una observación es un dato atípico si tiene un residuo estandarizado mayor que 2 ($|r_i| > 2$), otras veces se pide que $|r_i| > 3$. En cualquier caso es una elección subjetiva y cuanto mayor sea $|r_i|$ más atípica es la observación. Los datos atípicos son de gran importancia porque su inclusión o no en la muestra puede hacer que varíe mucho la recta de regresión estimada.

Aquí se observa que ninguno de los datos registra un error estandarizado con valor absoluto mayor de 2, razón por la cual no deberían descartarse datos.

Llegamos así a la conclusión de que con estos datos no podemos establecer una relación lineal entre ambas variables, por tal razón continuamos el análisis transformando las variables para linealizar, en busca de mejores resultados.

Como hemos visto una forma de transformación habitual consiste en tomar logaritmo de la variable independiente. Como esta variable presenta valores negativos, desplazamos el origen de la variable en 10 unidades (valor mayor al máximo registro negativo) y aplicamos la regresión lineal simple. Los resultados obtenidos son:

- Función de regresión $Y = 4{,}15\ log(X+10) + 1{,}17$

- Coeficiente de correlación múltiple 0,29
- Coeficiente R^2 de determinación 0,09
- Coeficiente R^2 ajustado 0,08
- Error típico 4,71

Si bien en este caso obtenemos mejores resultados, estos se encuentran muy lejos de acercarse a valores aceptables.

Al analizar los residuos estandarizados, encontramos algunos casos en los que se supera el umbral establecido de valor absoluto 2, permitiéndonos suponer la existencia de datos atípicos. Razón por la cual decidimos realizar una nueva regresión descartándolos.

En esta nueva regresión obtenemos como resultados:

- Función de regresión $Y = 4{,}79\ log(X+10) + 0{,}49$
- Coeficiente de correlación múltiple 0,33
- Coeficiente R^2 de determinación 0,11
- Coeficiente R^2 ajustado 0,11
- Error típico 4,56

Como podemos ver, los resultados mejoran levemente, pero sin llegar a umbrales de aceptabilidad. Además al analizar los residuos estandarizados no hallamos valores atípicos, razón por la cual decidimos probar una nueva forma de linealización, basada en tomar logaritmos de la variable independiente y de la variable dependiente (regresión doble-log[46]), tal cual lo recomienda la bibliografía de consulta. Para tomar logaritmos de la variable dependiente también es necesario desplazar el origen de la misma para que no se presenten valores negativos. Como resultados obtenemos:

- Función de regresión $log(Y+10) = 0{,}15\ log(X+10) + 1{,}01$
- Coeficiente de correlación múltiple 0,34
- Coeficiente R^2 de determinación 0,11
- Coeficiente R^2 ajustado 0,11
- Error típico 0,15

Los valores son similares a los ya obtenidos, hallándose también en este caso algunos datos atípicos, que luego de descartados dan como resultado la siguiente regresión:

- Función de regresión $log(Y+10) = 0{,}15\ log(X+10) + 1{,}01$
- Coeficiente de correlación múltiple 0,32

[46] "Economía de mercado, virtudes e inconvenientes", EMVI, Universidad de Málaga, España 2005.

- Coeficiente R^2 de determinación 0,10
- Coeficiente R^2 ajustado 0,10
- Error típico 0,13

Como vemos, no se registran mejoras, por lo cual decidimos recurrir a soluciones computacionales de mayor poder.

Empleamos ahora el programa TCWin, el cual al hacer correr los datos disponibles nos indica las ecuaciones de regresión obtenidas, ordenadas por su R^2. Para este caso el mejor valor de R^2 es de 0,13 y para una ecuación de la forma:

$$Y = a + b\,X + (c/\,X) + d\,X^2 + (e/\,X^2) + f\,X^3 + (g/\,X^3) + h\,X^4 + (i/\,X^4) + j\,X^5 + (k/\,X^5) \quad (3.6)$$

Concluimos, por todo lo expuesto, que con los datos disponibles no podemos establecer un buen modelo de correlación entre la tasa de crecimiento del tránsito y la variación del empleo. Debemos trazar entonces otra línea de trabajo.

3.2.1.2. Variación del parque automotor como variable independiente

Uno de los datos socioeconómicos recabados, que dejamos de lado en un principio, por encontrarse sólo disponible entre los ciclos 1997-2003, es el de registro automotor, recabado en la *DNRPA* (Dirección Nacional de Registros de Propiedad Automotor). Si bien este dato no se presenta completo para los años 1993-2003 establecidos para el estudio, es analizado por ser la única alternativa restante, dada nuestra recolección de datos.

Establecemos entonces:

X = variación de parque automotor (variable independiente) = (automotores en ciclo en estudio – automotores ciclo anterior) . 100 / (automotores ciclo anterior)

Y = crecimiento tránsito (variable dependiente)

Nuevamente, previo al análisis de regresión, efectuamos el análisis estadístico de la muestra, el cual para los datos de tasa de crecimiento del tránsito resulta:

Frecuencia = 64

Media = -2,77969

Varianza = 45,1439

Desviación típica = 6,71892

Mínimo = -14,8

Máximo = 18,3

Rango = 33,1

Asimetría tipificada = 2,1114

Curtosis tipificada = 1,46844

Como vemos la asimetría resulta levemente por encima del umbral establecido, lo cual consideramos no llegará a afectar el análisis. La gráfica de caja y bigotes de la Figura 3.17 nos permite observar la distribución de la muestra.

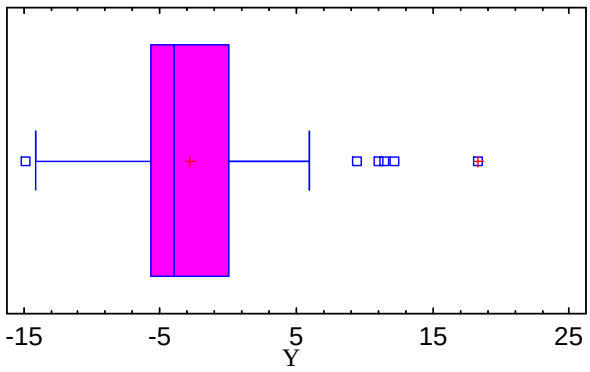

Fig. 3.17. Gráfico de caja y bigotes de la tasa de crecimiento del tránsito

Análogamente efectuamos el análisis estadístico de la variable variación de parque automotor, resultando:

Frecuencia = 64

Media = 2,28571

Varianza = 1,82208

Desviación típica = 1,34985

Mínimo = 0,5

Máximo = 5,5

Rango = 5,0

Asimetría tipificada = 1,91099

Curtosis tipificada = -0,8162

Observamos como tanto la asimetría tipificada y la curtosis se ubican dentro de los valores límites, anexándose al análisis la gráfica de cajas y bigotes de la Figura 3.18.

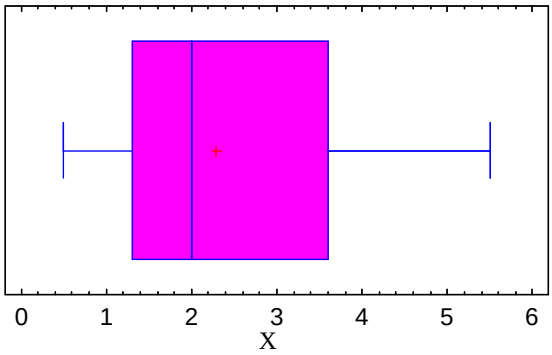

Fig. 3.18. Gráfico de caja y bigotes de la variación del parque automotor

Podemos considerar entonces que la muestra es sensiblemente normal y pasar al análisis de regresión, nuevamente con el "módulo de regresión" del programa Microsoft Excel.

Como la gráfica de tasa de crecimiento de tránsito vs. variación parque automotor, no nos permite establecer a simple vista una relación entre ambas variables, decidimos buscar directamente la linealización de los datos. Tomamos entonces logaritmos de la variable independiente y logaritmos de la variable dependiente (trasladando el origen 20 unidades para evitar valores negativos). Visualmente parece que ahora una regresión lineal puede ser viable, según se observa en la Figura 3.19.

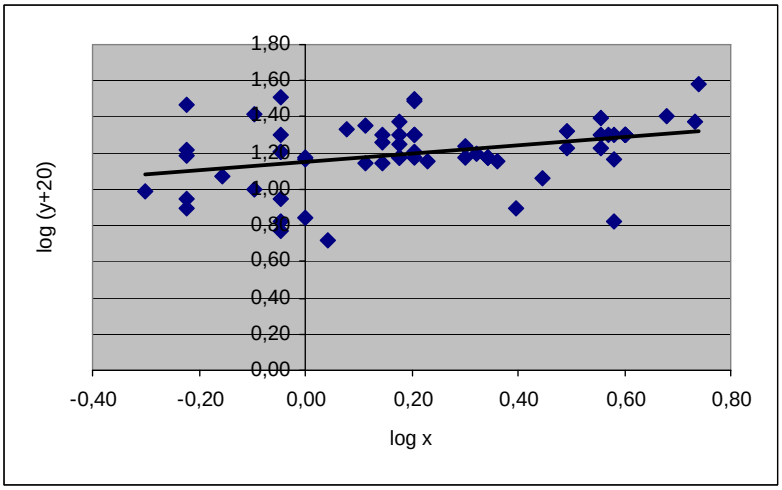

Fig. 3.19. Gráfico variación tránsito vs variación parque automotor, afectados por log.

Pero al realizar las regresiones y tras descartar algunos datos atípicos, los valores que obtenemos son:

- Función de regresión $log(Y+20) = 0{,}19\ log\ X + 1{,}18$

- Coeficiente de correlación múltiple 0,62
- Coeficiente R^2 de determinación 0,39
- Coeficiente R^2 ajustado 0,38
- Error típico 0,07

Estos valores, si bien resultan mucho mejores a los obtenidos con la tasa de empleo, no llegan a los umbrales de aceptabilidad que nos hemos fijado.

Decidimos entonces volver atrás el análisis y aplicar la regresión directamente sobre los valores sin linealizar. Esto nos permite obtener los siguientes valores:

- Función de regresión $Y = 2{,}23\ X - 7{,}99$
- Coeficiente de correlación múltiple 0,75
- Coeficiente R^2 de determinación 0,56
- Coeficiente R^2 ajustado 0,55
- Error típico 2,53

El contraste de la *F* también demuestra la influencia de la linealidad, dándonos un *p-valor* muy por debajo de 0,05.

Aunque ya hemos explicado el contraste de la *F*, tal vez sea conveniente tratar de simplificar un poco más el concepto de su empleo. Con este estadístico lo que se pretende es comparar el modelo propuesto con aquel en donde no aparece la *X* planteada, esto es, ver si realmente hay una dependencia entre *Y* y *X* según lo hemos planteado. Cuanto mayor es el valor del estadístico, mayor la evidencia que juntamos para probar que hay dependencia, y menor *p-valor*. Para un 95 % de confianza si el *p-valor* es menor que 0,05 entonces demostramos esta dependencia.

Con los residuos estandarizados elaboramos la gráfica de la Figura 3.20 y la de la Figura 3.21, que muestran una sensible normalidad de los mismos, con media 0 y desvío estándar 1. La estadística completa de los residuos estandarizados es:

Media = -0,00222222

Varianza = 1,01613

Desviación típica = 1,00803

Mínimo = -2,1

Máximo = 1,9

Rango = 4,0

Asimetría tipificada = 0,817851

Curtosis tipificada = -0,532223

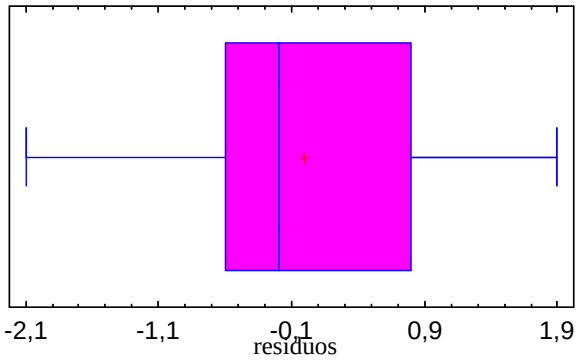

Fig. 3.20. Gráfico de caja y bigotes para los residuos, empleando variación de parque automotor

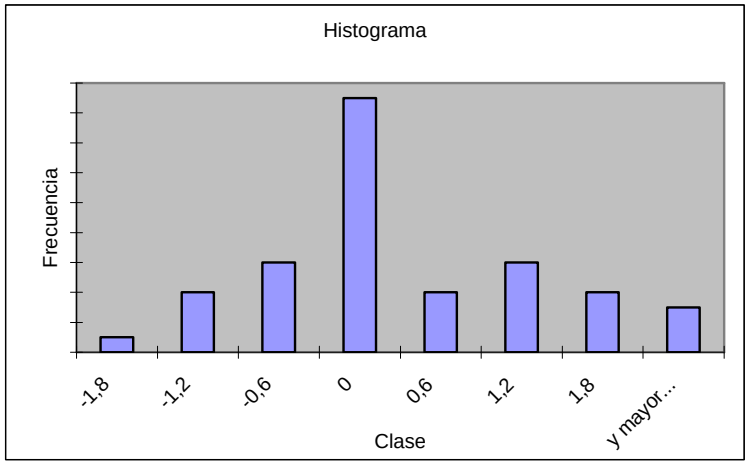

Fig. 3.21. Histograma de los residuos empleando variación del parque automotor

Como dijéramos, el análisis realizado hasta ahora con la variable independiente del parque automotor se lleva adelante con el módulo de regresión del programa Microsoft Excel, veamos los resultados que obtenemos en esta última regresión cuando empleamos el programa Statgraphics Plus, más potente que el anterior.

El gráfico del modelo ajustado obtenido con este programa, Figura 3.22, nos permite observar las bandas para los errores, valores que empleamos más adelante en la discusión del modelo final obtenido.

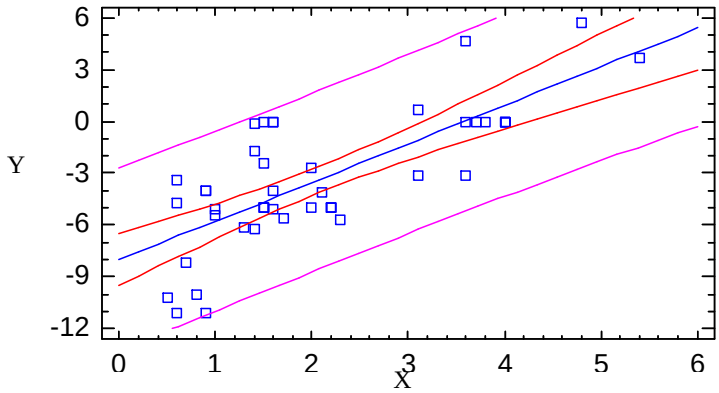

Fig. 3.22. Gráfico del modelo ajustado con bandas para los errores

Los resultados no son malos, pero como el modelo final se compone del producto de los submodelos, debemos hallar en cada uno de estos el menor error posible, en busca de un error general aceptable, tal como ya lo explicáramos.

Volvemos por esto al empleo del programa TCWin, que permite establecer modelos más complejos y ajustados que, al fin y al cabo, podrían ser establecidos por regresión simple mediante las adecuadas linealizaciones, como hemos hecho hasta ahora. Como describimos anteriormente, este programa, al ser cargado con la matriz de datos, nos da como resultado una lista de regresiones en orden decreciente de coeficiente de correlación.

Al analizar nuestra muestra obtenemos dicho listado de ecuaciones, pero descubrimos que éstas poseen similar valor de R^2. Nos surge entonces la pregunta, ¿En caso de distintos modelos de similar ajuste a nuestra nube de puntos, cuál elegimos?

"...Los mecanismos para la selección de un modelo no son fáciles de especificar, ya que dependen en gran medida del tipo de modelo, del contexto de utilización y de las propias características del proceso analizado. Quizás la única norma clara es que ante dos posibles modelos, similares en otros aspectos, preferiremos el que sea más sencillo y que menos suposiciones necesite para su construcción (es lo que se denomina "principio de parsimonia")..."[47]

Basados en este "principio de parsimonia", de la lista obtenida tomamos la ecuación:

[47] "Construcción de modelos de regresión multivariantes", L. Molinero, Alce Ingeniería, España 2002.

- Función de regresión

$$Y = 35{,}596896 - (243{,}628504 / X) + (555{,}412790 / X^2) - (585{,}523100 / X^3) + (283{,}681553 / X^4) - (51{,}088958 / X^5)$$

- Coeficiente R^2 de determinación 0,66
- Coeficiente R^2 ajustado 0,59
- Error típico 2,40
- Estadístico F 11,37 (p-*valor* 0,0000) significativo al 99%

En la Figura 3.23, podemos observar como la función se ajusta a nuestra nube de puntos y en la Figura 3.24 se observa el gráfico de dispersión de los residuos.

Llegamos de este modo a una ecuación que nos arroja valores aceptables de ajuste.

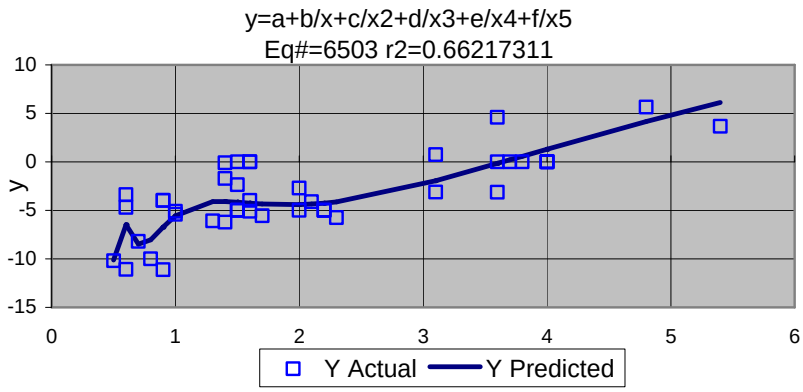

Fig. 3.23. Ajuste de la ecuación a la nube de puntos, empleando variación parque automotor

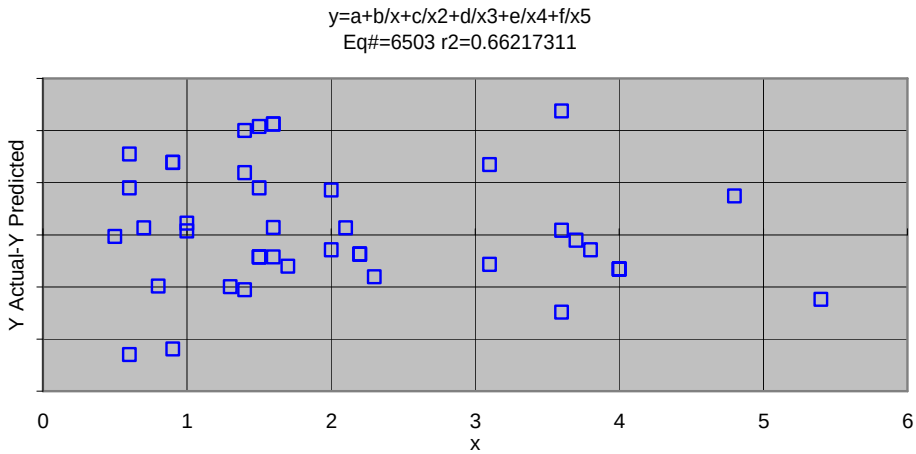

Fig. 3.24. Gráfico de dispersión de los residuos, empleando variación parque automotor

3.2.2. Obtención de los algoritmos para los coeficientes diarios

Para el análisis en busca del algoritmo que nos permita calcular los coeficientes diarios de corrección, fijamos como punto de partida a las siguientes variables:

- X = variable independiente, representa los días de la semana

  X = 1 (día domingo)

  X = 2 (día lunes)

  X = 3 (día martes)

  X = 4 (día miércoles)

  X = 5 (día jueves)

  X = 6 (día viernes)

  X = 7 (día sábado)

- Y = variable dependiente, es el coeficiente diario

Con estas variables en juego, buscamos ahora establecer la validez de la muestra, efectuamos entonces el análisis estadístico de los coeficientes diarios, dándonos como resultado:

Frecuencia = 161

Media = 1,01018

Varianza = 0,0172913

Desviación típica = 0,131496

Mínimo = 0,687

Máximo = 1,469

Rango = 0,782

Asimetría tipificada = 1,88183

Curtosis tipificada = 2,0702

Podemos ver que la curtosis se ubica levemente por encima de los límites establecidos, por lo cual consideramos que no se desvirtúa el análisis por regresión.
En la Figura 3.25, observamos el gráfico de caja y bigotes de la muestra.

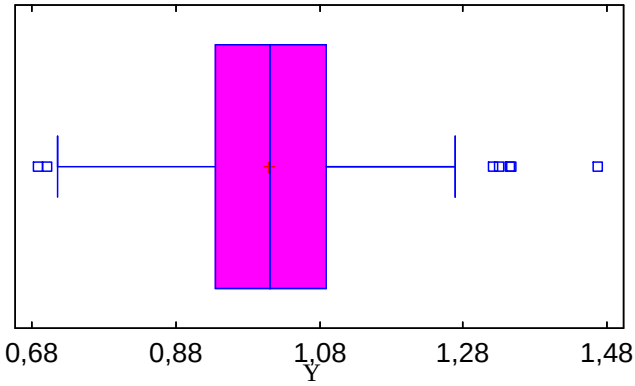

Fig. 3.25. Gráfico de caja y bigotes de los coeficientes diarios

Una vez confirmada la normalidad, analizamos la gráfica de coeficientes diarios vs. días de semana de la Figura 3.26.

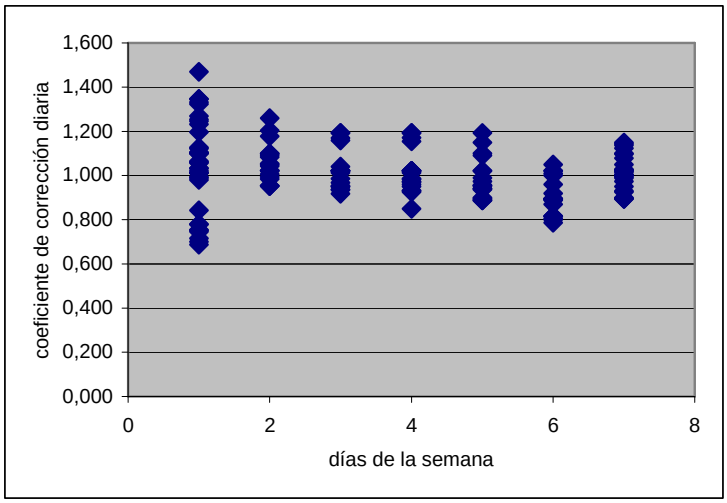

Fig. 3.26. Gráfico de coeficientes diarios vs día de la semana

Aquí no podemos observar claramente grupos de nubes de puntos aislados y podríamos pensar en la realización de una regresión única. Pero en función de los conceptos ya apuntados en este documento y de los que volcamos en el párrafo siguiente, es de esperarse un comportamiento diferente entre dos grandes grupos de vías, las que sirven eminentemente a fines turísticos y las que lo hacen a fines comerciales. En realidad puede también pensarse que existen distintos coeficientes para estas vías en distintas épocas del año (por ejemplo verano, otoño, invierno y primavera), pero el analizar estas alternativas requiere la obtención de series de datos

discriminados en tal sentido, lo cual en nuestro caso no ha sido posible, por lo que decidimos dejar de lado este tipo de planteos.

"...Se han estudiado cuales son los días de la semana que llevan los volúmenes normales de tránsito...

En ciertas carreteras los volúmenes de lunes a viernes son muy estables; y se registran máximos volúmenes durante el fin de semana, ya sea el sábado o domingo, debido a que durante estos días por estas carreteras circula una alta demanda de usuarios de tipo turístico y recreacional...

En otras carreteras los volúmenes máximos se presentan entre semana, al igual que en las calles de la ciudad, donde la variación de los volúmenes de tránsito diario no es muy pronunciada en los días laborales. Ambos casos reflejan el uso comercial de estas vías..."[48]

Decidimos entonces incluir la variable clasificatoria por uso de la vía, previo a los análisis de regresión.

Como ya estableciéramos esta variable toma valor 1 cuando la vía es comercial y 0 cuando es turística. Los gráficos de la Figura 3.27 y la Figura 3.38, reflejan la inclusión de esta clasificación.

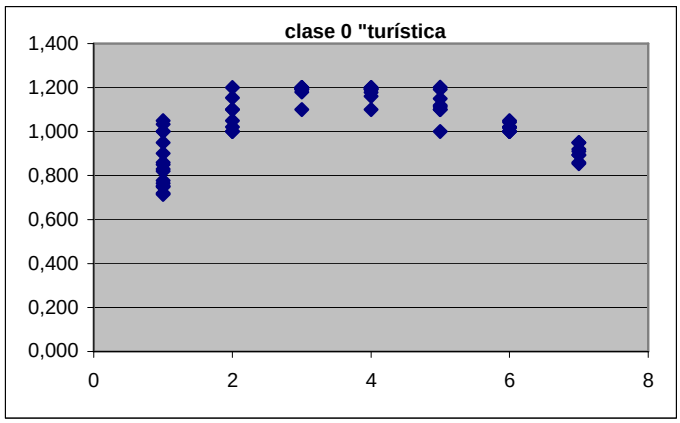

Fig. 3.27. Gráfico de coeficientes diarios para vías turísticas

[48] "Ingeniería de tránsito, fundamentos y aplicaciones", R. Cal y Mayor, J. Cárdenas, Alfaomega 7°ed., México 1995.

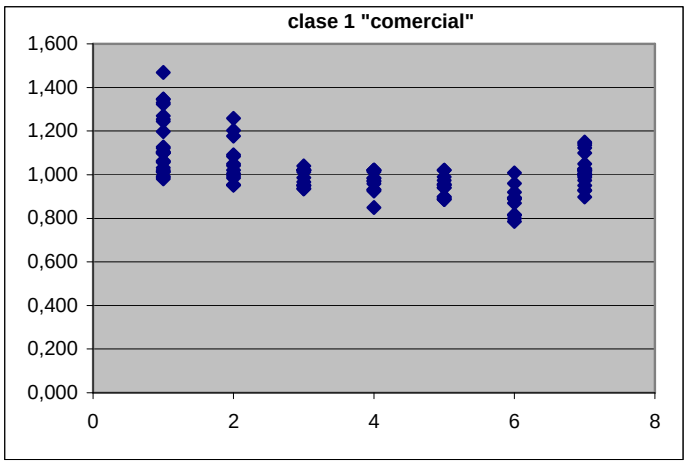

Fig. 3.28. Gráfico de coeficientes diarios para vías comerciales

Podemos ver como la inclusión de la clase realmente genera dos nubes de puntos diferenciables y como éstas convalidan lo asegurado en la consulta bibliográfica. Así, la primera nube presenta una tendencia hacia una parábola cóncava hacia abajo y la segunda aparenta ser una parábola cóncava hacia arriba. Estamos ahora en condiciones de realizar las regresiones por separado que analizamos a continuación.

3.2.2.1. Análisis para vías de uso turístico

Cuando analizamos la Figura 3.27 observamos que la nube de puntos se asemeja a una parábola hacia abajo, mostrando una alta concentración en todos los días, salvo en los domingos que presentan una cierta dispersión de valores.

En función de esta evidente forma de parábola, decidimos comenzar el análisis de regresión aplicando directamente una linealización de la variable independiente que nos permita la obtención de su función. Esto se logra tomando cuadrados de la misma y generando la regresión como si fueran dos las variables independientes (los valores de X y los valores de X^2), es decir como una regresión múltiple. Mediante esta técnica obtenemos los siguientes resultados:

- Función de regresión $Y = -0{,}04\ X^2 + 0{,}32\ X + 0{,}55$
- Coeficiente de correlación múltiple 0,90
- Coeficiente R^2 de determinación 0,82
- Coeficiente R^2 ajustado 0,80
- Error típico 0,07

Estos valores nos permiten observar un muy buen ajuste de esta regresión, pero al analizar los residuos estandarizados se registran algunos casos que superan el umbral establecido del valor absoluto 2, por tal razón decidimos eliminar estos datos y realizar nuevamente la regresión, obteniéndose en este caso como resultados:

- Función de regresión $Y = -0{,}043715\ X^2 + 0{,}363511\ X + 0{,}452025$
- Coeficiente de correlación múltiple 0,97
- Coeficiente R^2 de determinación 0,94
- Coeficiente R^2 ajustado 0,94
- Error típico 0,04

Como puede observarse, el quitar estos pocos datos atípicos nos permite obtener una sensible mejora en los resultados, sin una modificación fuerte de la ecuación. El coeficiente de correlación múltiple, que como viéramos debe acercarse en valor absoluto a 1, es de 0,97. El coeficiente de determinación que debe ser cercano a 1, es de 0,94 y muy pocos de los residuos estandarizados poseen valor absoluto superior a 2, presentando su distribución en la Figura 3.29 una sensible normalidad. El contraste de la *F* también demuestra la tendencia de la linealización con un *p-valor* muy por debajo de 0,05. La estadística completa de los residuos es:

Media = 0,000689655

Varianza = 1,00137

Desviación típica = 1,00069

Mínimo = -1,84

Máximo = 2,3

Rango = 4,14

Asimetría tipificada = 0,35049

Curtosis tipificada = -0,553592

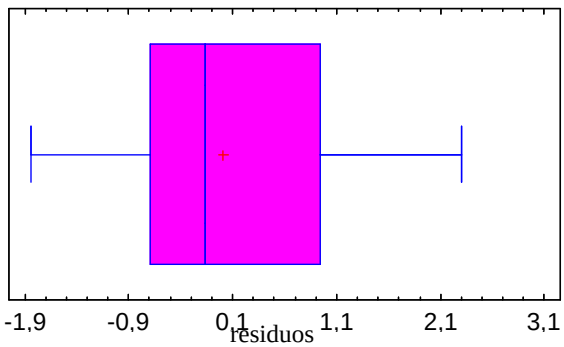

Fig. 3.29. Gráfico de caja y bigotes de los residuos para vías turísticas

En función de estos resultados llegamos a la conclusión de que el algoritmo obtenido para las vías turísticas es válido.

3.2.2.2. Análisis para vías de uso comercial

Cuando analizamos la Figura 3.28, correspondiente a la gráfica de los coeficientes de corrección diarios vs. día de la semana para vías de uso comercial, vimos que se podía intuir una relación asimilable a una parábola cóncava hacia arriba. Por esto, al igual que en el caso de las vías turísticas, comenzamos el análisis de regresión linealizando con la variable independiente para obtener una ecuación cuadrática. Los resultados obtenidos en este experimento son:

- Función de regresión $Y = 0{,}01\ X^2 - 0{,}07\ X + 1{,}12$
- Coeficiente de correlación múltiple 0,64
- Coeficiente R^2 de determinación 0,40
- Coeficiente R^2 ajustado 0,39
- Error típico 0,05

Vemos que los coeficientes de correlación múltiple y de determinación son bajos. Por su parte, el análisis de los residuos estandarizados de la Figura 3.30 no nos permite establecer la existencia de datos atípicos, pero tampoco nos permite observar su normalidad. Esto se ve confirmado con los datos estadísticos completos en donde vemos como la curtosis supera el límite fijado de –2:

Media = 0,000416667
Varianza = 1,00045
Desviación típica = 1,00023
Mínimo = -1,91
Máximo = 1,65
Rango = 3,56
Asimetría tipificada = -0,708985
Curtosis tipificada = -3,51774

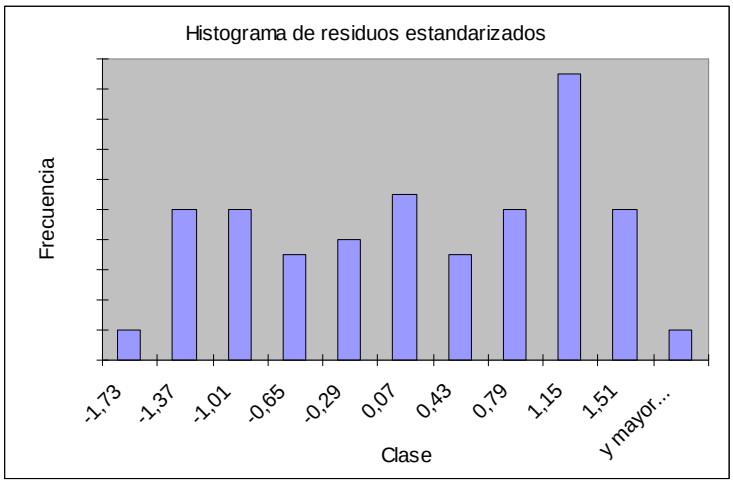

Fig. 3.30. Histograma de los residuos para vías comerciales

Los ajustes alcanzados con esta clase nos permiten deducir que la línea de trabajo seguida no nos conduce a un buen resultado, posiblemente por falta de inclusión de una nueva variable de clasificación.

Para esto volvemos análisis de la Figura 3.28, donde además de la concavidad hacia arriba de la clase para la nube de puntos, observamos cierta dispersión en los valores extremos (es decir 1 y 7).

Tres son las variables clasificatorias que podemos incluir, en función de los datos disponibles. Estas son la urbanidad de la vía, la existencia de peaje o la clasificación del tránsito. El análisis detallado de las series en función de estas tres variables nos lleva a pensar que la variable clasificatoria faltante es la de existencia o no de peaje sobre la vía, con la que se obtienen las nubes de puntos de la Figura 3.31 y de la Figura 3.32.

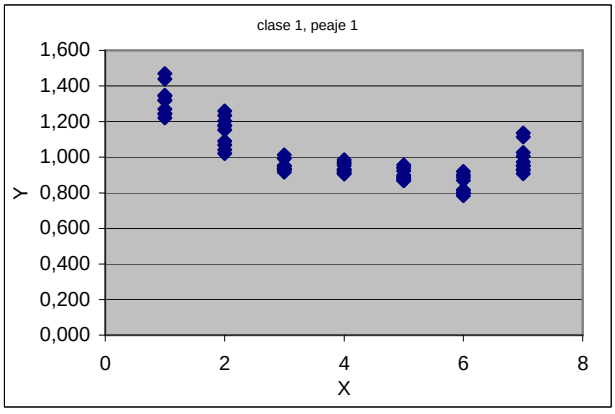

Fig. 3.31. Nube de puntos para los coeficientes diarios en vías comerciales con peaje

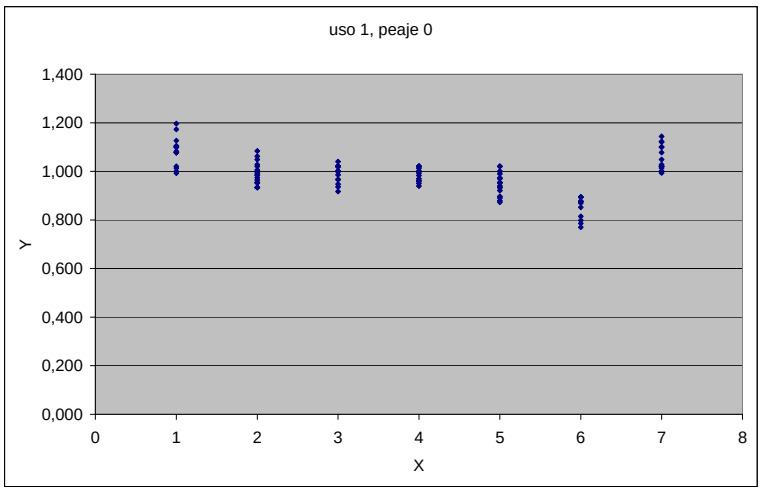

Fig. 3.32. Nube de puntos para los coeficientes diarios en vías comerciales sin peaje

Realicemos el análisis para esta nueva clasificación.

3.2.2.2.1. Análisis en vías comerciales con peaje

La nube de puntos para este caso presenta una clara concavidad hacia arriba, aunque en un sector medio muestra un salto en la función. Esto nos lleva a pensar que un polinomio de grado superior puede ser la mejor forma de regresión, no obstante lo cual experimentamos inicialmente buscando una función polinómica de grado dos, mediante el "módulo de regresión" del programa Microsoft Excel, para analizar su ajuste, obteniendo como resultado:

- Función de regresión $Y = 0{,}03\ X^2 - 0{,}28\ X + 1{,}58$
- Coeficiente de correlación múltiple 0,92
- Coeficiente R^2 de determinación 0,84
- Coeficiente R^2 ajustado 0,83
- Error típico 0,07

Como podemos observar el ajuste nos da valores muy buenos. Complementariamente el análisis de los residuos estandarizados no nos permite observar datos atípicos, no obstante lo cual es de esperarse la existencia de regresiones que ajusten mejor a esta nube.

Buscando estos mejores resultados, volvemos al empleo del programa TCWin. Con los datos ingresados, obtenemos un listado de ecuaciones de regresión ordenadas por su coeficiente de determinación, que para este caso resulta de 0,88 en alrededor de 20 ecuaciones. Guiados nuevamente por el "principio de parsimonia", de entre éstas tomamos la de más sencilla expresión y la analizamos en detalle, obteniendo:

- Función de regresión Y = 0,002781 X^5 - 0,053475 X^4 + 0,378762 X^3 - 1,184775 X^2 + 1,434157 X + 0,758143
- Coeficiente R^2 de determinación 0,88
- Coeficiente R^2 ajustado 0,85
- Error típico 0,06

Estos resultados verifican el buen ajuste obtenido con la regresión de segundo grado, ya que con una mayor sencillez de cálculo se obtienen resultados similares.

El análisis gráfico de la curva ajustada de grado 5, Figura 3.33, y de sus residuos, Figura 3.34, también nos permite establecer a ésta como la función de regresión buscada.

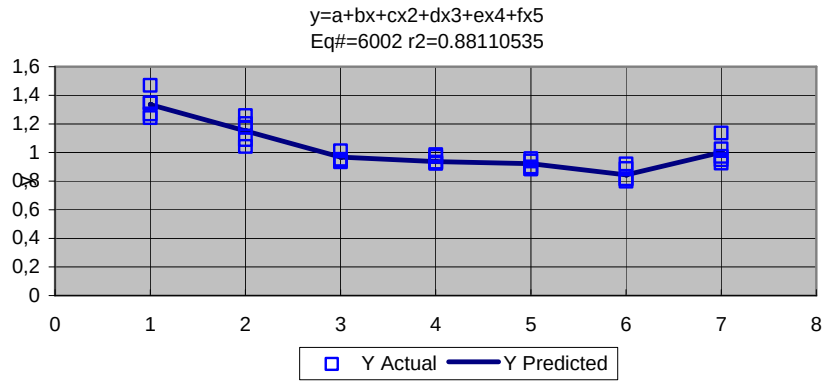

Fig. 3.33. Ajuste de la función polinómica de grado cinco, en vías comerciales con peaje

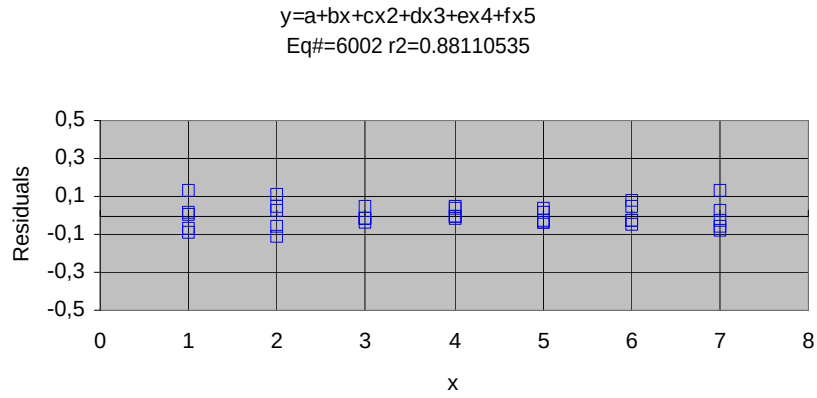

Fig. 3.34. Gráfica de residuos de la función polinómica de grado cinco, en vías comerciales con peaje

3.2.2.2.2. Análisis en vías comerciales sin peaje

La Figura 3.32, correspondiente a la nube de puntos para esta clasificación, nos muestra nuevamente a grandes rasgos una parábola con concavidad hacia arriba, pero con ciertas ondulaciones que nos llevan a pensar que un ajuste polinómico de grado superior tendría que ser apropiado. Por tal razón, decidimos hacer correr directamente los datos con el programa TCWin. Así, obtenemos el listado de ecuaciones con su correspondientes coeficientes de determinación, que para el máximo valor de R^2 obtenido es:

- $R^2 = 0{,}70$
  $Y = a + b\,X + (c/\,X) + d\,X^2 + (e/\,X^2) + f\,X^3 + (g/\,X^3) + h\,X^4$
- $R^2 = 0{,}70$
  $Y = a + b\,\ln X + c\,(\ln X)^2 + d\,(\ln X)^3 + e\,(\ln X)^4 + f\,(\ln X)^5 + g\,(\ln X)^6$
- $R^2 = 0{,}70$
  $Y = a + b\,\ln X + c\,(\ln X)^2 + d\,(\ln X)^3 + e\,(\ln X)^4 + f\,(\ln X)^5 + g\,(\ln X)^6 + h\,(\ln X)^7$
- $R^2 = 0{,}70$
  $Y = a + b\,\ln X + c\,(\ln X)^2 + d\,(\ln X)^3 + e\,(\ln X)^4 + f\,(\ln X)^5 + g\,(\ln X)^6 + h\,(\ln X)^7 + i\,(\ln X)^8 + j\,(\ln X)^9 + k\,(\ln X)^{10}$
- $R^2 = 0{,}70$
  $Y = a + b\,\ln X + c\,(\ln X)^2 + d\,(\ln X)^3 + e\,(\ln X)^4 + f\,(\ln X)^5 + g\,(\ln X)^6 + h\,(\ln X)^7 + i\,(\ln X)^8 + j\,(\ln X)^9$
- $R^2 = 0{,}70$
  $Y = a + b\,\ln X + c\,(\ln X)^2 + d\,(\ln X)^3 + e\,(\ln X)^4 + f\,(\ln X)^5 + g\,(\ln X)^6 + h\,(\ln X)^7 + i\,(\ln X)^8$
- $R^2 = 0{,}70$
  $Y = a + b\,X + (c/\,X) + d\,X^2 + (e/\,X^2) + f\,X^3 + (g/\,X^3) + h\,X^4 + (i/\,X^4) + j\,X^5 + (k/\,X^5)$
- $R^2 = 0{,}70$
  $Y = a + b\,X + (c/\,X) + d\,X^2 + (e/\,X^2) + f\,X^3 + (g/\,X^3) + h\,X^4 + (i/\,X^4) + j\,X^5$
- $R^2 = 0{,}70$
  $Y = a + b\,X + (c/\,X) + d\,X^2 + (e/\,X^2) + f\,X^3 + (g/\,X^3) + h\,X^4 + (i/\,X^4)$
- $R^2 = 0{,}70$

$$Y = a + b\,X + c\,X^2\,\ln X + d\,X^3 + e\,e^x \quad (3.7)$$

La última de todas estas ecuaciones presenta una sencillez notoriamente superior, para un mismo ajuste. Guiados nuevamente por el "principio de parsimonia" ya enunciado decidimos analizarla:

- Función de regresión $Y = 1{,}299385 - 0{,}175416\,X + 0{,}110582\,X^2\,\ln X - 0{,}033388\,X^3 + 0{,}001731\,e^X$
- Coeficiente R^2 de determinación 0,70
- Coeficiente R^2 ajustado 0,67
- Error típico 0,04

La regresión presenta valores admisibles, mostrando su curva de ajuste y de residuos también buen comportamiento, como se ve en la Figura 3.35 y la Figura 3.36.

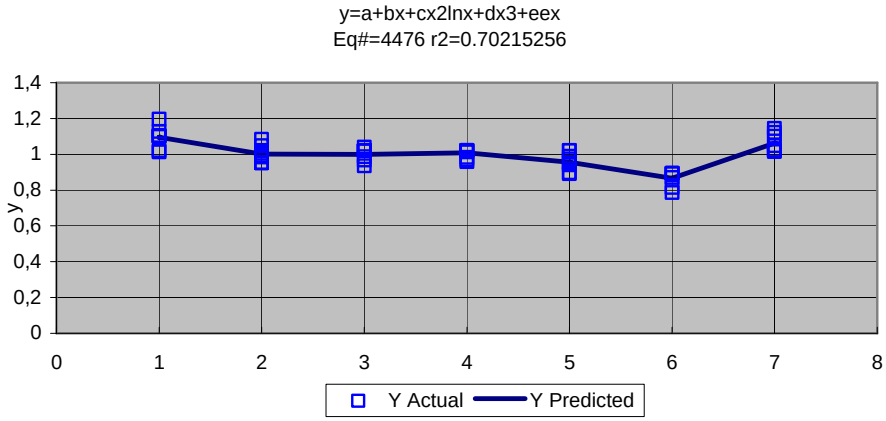

Fig. 3.35. Ajuste de la función obtenida, en vías comerciales sin peaje

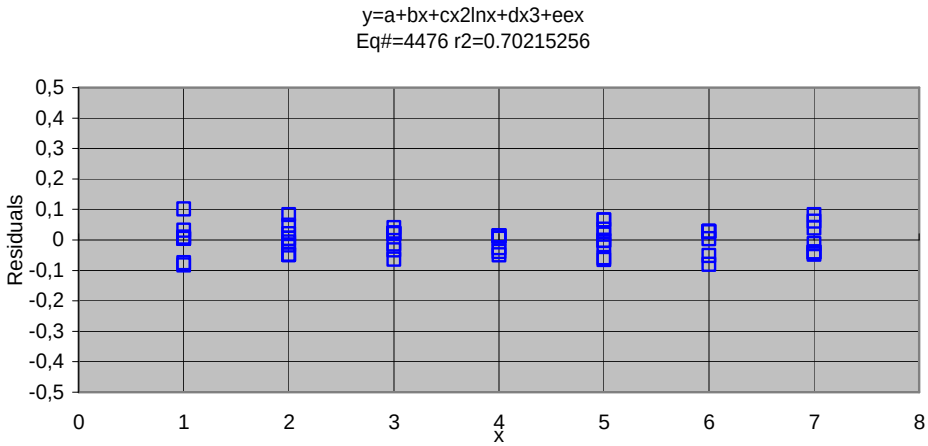

Fig. 3.36. Gráfico de residuos de la función obtenida, en vías comerciales sin peaje

3.2.3. Obtención de los algoritmos para los coeficientes mensuales

El salto conceptual desde el coeficiente diario al mensual podría ser criticado, pero la carencia generalizada de datos hace imposible la obtención de coeficientes más detallados, cuando no lo es incluso la obtención del propio coeficiente mensual.

"...El coeficiente de estacionalidad del mes sólo puede estimarse en el corto plazo por analogía con tramos con coeficientes conocidos, sea por contadores permanentes, sea por censos de cobertura efectuados en distintas épocas del año. Por esta razón, con censos de corta duración la expansión se realiza al *TMDM*, y la corrección al *TMDA* exige disponer del coeficiente del mes de fuente exógena.

Si se releva el volumen de tránsito de todo un mes, el Tránsito Medio Diario del Mes se calcula con la expresión siguiente

$$TMDM = (5 \text{ media día hábil} + \text{media sábado} + \text{media domingo})/7 \quad (3.8)$$

El coeficiente de estacionalidad de la semana dentro del mes es usualmente asumido igual a la unidad, pues el patrón del tránsito es repetitivo en módulos de 7 días, y las diferencias atribuibles a censar en la primera semana del mes o en otra semana suelen ser despreciables..."[49]

Aunque la forma de calcular el *TMDM* citada en este párrafo pueda ser mejorada empleando la ecuación:

$$TMDM = (\text{nº días hábiles} \cdot \text{media día hábiles} + \text{nº días no hábiles} \cdot \text{media de día no hábiles}) / (\text{nº de días hábiles} + \text{nº de días no hábiles}) \quad (3.9)$$

todo parece indicar que lo aquí expresado en cuanto a los patrones de tránsito es valedero. Por esto consideramos que no es necesario contar con un coeficiente intermedio que distinga los volúmenes entre las distintas semanas de un mes. De todos modos sería adecuado que este parámetro, en caso de que las condiciones particulares de un estudio hicieran imperiosa su determinación, tome valor en forma independiente del mes que se trate. Razón por la cual podría incluirse sencillamente como un factor más en la ecuación final para el cálculo del *TMDA*, mediante un algoritmo que posea como variable independiente la ubicación que la semana en la que se realiza el conteo posee dentro del mes.

[49] "Caracterización de errores de muestreo en censos de volumen y composición", M. Herz, J. Galárraga, M. Maldonado, XIV Congreso Argentino de Vialidad y Tránsito, Argentina 2005.

Ratificada la estructura del modelo general pasamos al análisis puntual para los coeficientes mensuales. Pero previo al análisis numérico revisemos algunos aspectos considerados en estudios similares a éste.

"...En el marco de los estudios de este informe, se hace necesaria la disponibilidad de análisis de los determinantes de la demanda de acuerdo a las estacionalidades...

En el modelo empleado la variable explicada son los vehículos circulantes y las variables explicativas:

- *Ingreso.* Esta variable se capta a través del Indice de Producción Industrial, utilizado especialmente por su gran fidelidad como aproximación al producto bruto y por su frecuencia. A través de esta variable se considera como impacta la actividad sobre el tránsito a estimar…

- *Población.* Es de singular importancia la inclusión de esta variable, en especial al tratarse de una estimación de transporte de pasajeros…

- *Peaje.* Se incluye como una variable adicional de precios…

Inicialmente se habían considerado variables adicionales, como las de matrículas educativas y personas ocupadas, pero estas dos variables resultaron no significativas al momento de realizar las estimaciones… además su importancia como determinantes era mas bien secundaria…"[50]

Si bien en este trabajo ya habíamos considerado la inclusión de las variables de borde, este párrafo expresa su empleo cuando se analizan las estacionalidades, tratadas aquí justamente mediante los coeficientes de corrección mensuales. Es por esto, para facilitar el análisis de las interrelaciones entre variables, que para esta parte del trabajo nos planteamos la idea desde el principio de llegar a los algoritmos buscados mediante la regresión múltiple.

Fijamos entonces nuestra variable dependiente Y para el valor del coeficiente buscado y una primera variable independiente X_1 para los meses del año, siendo:

- $X_1 = 1$ para enero
- $X_1 = 2$ para febrero
- $X_1 = 3$ para marzo
- $X_1 = 4$ para abril

[50] "Estimación econométrica del tránsito vehicular y de la demanda del servicio de transporte ferroviario y automotor en el Gran La Plata", J. Alonso, Expte. Muni. La Plata 78.449/01, Argentina 2001.

- X_1 = 5 para mayo
- X_1 = 6 para junio
- X_1 = 7 para julio
- X_1 = 8 para agosto
- X_1 = 9 para septiembre
- X_1 = 10 para octubre
- X_1 = 11 para noviembre
- X_1 = 12 para diciembre

Con estas dos variables podemos construir el gráfico de la Figura 3.37, para analizar la relación existente entre ambas en función de nuestros datos.

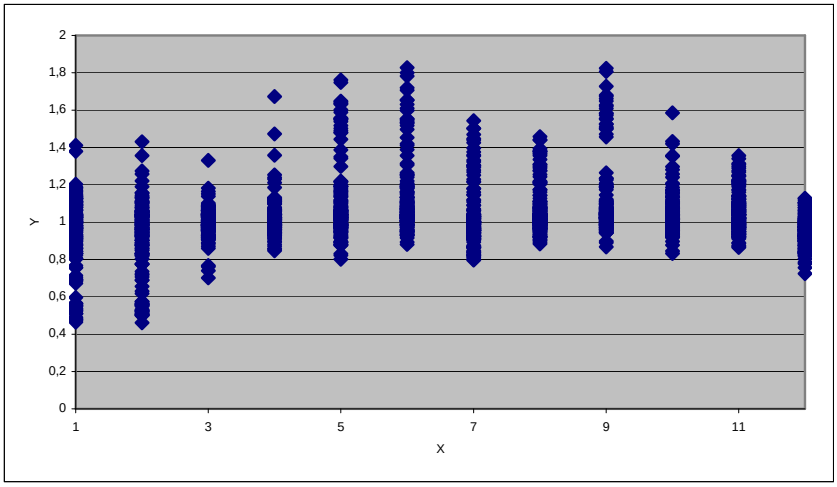

Fig. 3.37. Gráfico de coeficientes mensuales vs mes del año

Podemos deducir de esta gráfica que la nube de puntos podría ajustarse por una ecuación polinómica, pero es muy posible que el ajuste no sea bueno, debido a que la dispersión existente en cada valor de X_1 es alta. La inclusión de variables de entorno de la vía entonces puede que genere un modelo por regresión múltiple más ajustado.

Las nuevas variables a considerarse son:

- X_2 = urbanidad (X_2 = 0 rural, X_2 = 1 urbana)
- X_3 = uso (X_3 = 0 turística, X_3 = 1 comercial)
- X_4 = peaje (X_4 = 0 sin peaje, X_4 = 1 con peaje)
- X_5 = clasificación, expresado en % de autos más camionetas

Previo a los análisis de regresión identifiquemos si existe multicolinealidad entre las variables propuestas, para esto podemos analizar el gráfico de dispersión matricial o, lo que es análogo, las relaciones existente entre variables par a par. Para esto construimos los gráficos entre pares de variables y efectuamos los análisis correspondientes:

- X_1 vs. X_2 = Como X_1 representa los meses y contamos con series en forma equilibrada para ambiente urbano y rural, es de esperarse una gráfica como la obtenida en la Figura 3.38, en donde se evidencia la no existencia de una relación lineal.

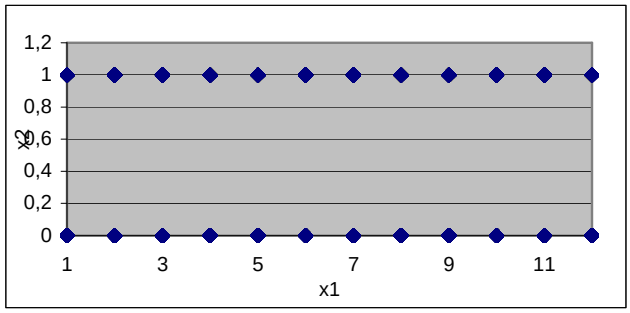

Fig. 3.38. Gráfico de X_1 vs. X_2

- X_1 vs. X_3 = Idem, pero comparadas las series en vías de uso comercial o turístico.
- X_1 vs. X_4 = Idem, pero comparadas las series en vías con o sin peaje.
- X_1 vs. X_5 = En este caso contamos con series para vías que presentan clasificaciones de autos y camionetas de entre un 30 a 100 %, habiendo barrido por lo tanto un muy alto espectro, sin presentarse por supuesto relación entre variables, tal como se ve en la Figura 3.39.

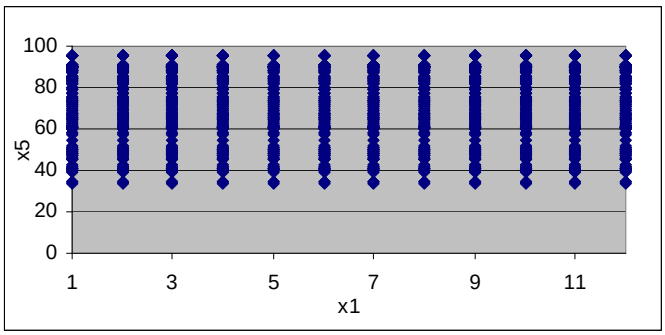

Fig. 3.39. Gráfico de X_1 vs. X_5

- X_2 vs. X_3 = Se trata de la comparación de dos variables binarias. Habría fuerte colinealidad si se agruparan los puntos exclusivamente en forma oblicua, es decir valores (0;0) con (1;1) o valores (0;1) con (1;0), o presentara pendiente muy pronunciada la correlación, lo cual no se registra, como se observa en la Figura 3.40.

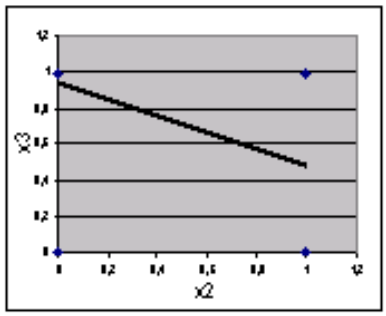

Fig. 3.40. Gráfico de X_2 vs. X_3

- X_2 vs. X_4 = Idem en la Figura 3.41.

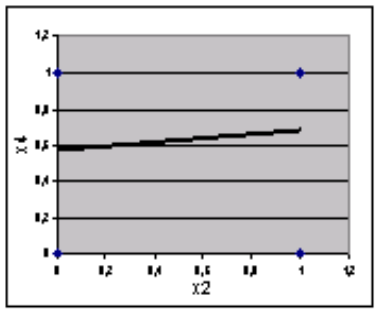

Fig. 3.41. Gráfico de X_2 vs. X_4

- X_3 vs. X_4 = Idem en la Figura 3.42.

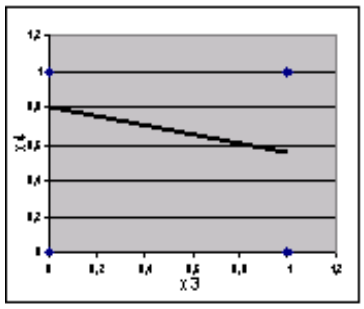

Fig. 3.42. Gráfico de X_3 vs. X_4

- X_5 vs. X_2 = Podemos observar en la Figura 3.43 como las series con las que se cuenta poseen sólo alta clasificación cuando la vía es urbana, lo que resulta en una forma de correlación.

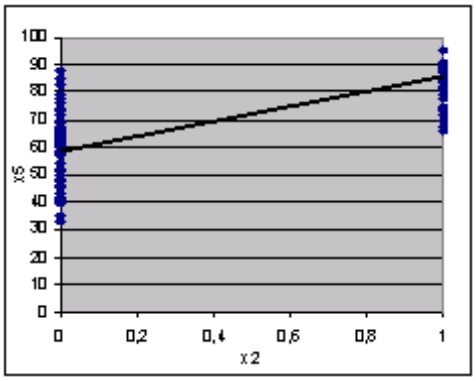

Fig. 3.43. Gráfico de X_2 vs. X_5

- X_5 vs. X_3 = En este caso observamos en la Figura 3.44 sólo altas clasificaciones en las vías turísticas, resultando en una correlación.

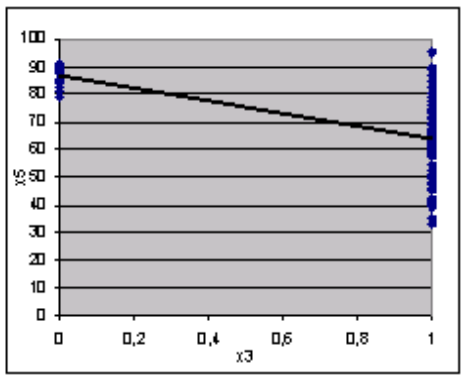

Fig. 3.44. Gráfico de X_3 vs. X_5

- X_5 vs. X_4 = No se observa en la Figura 3.45 una correlación entre la clasificación de la vía y la existencia de peaje.

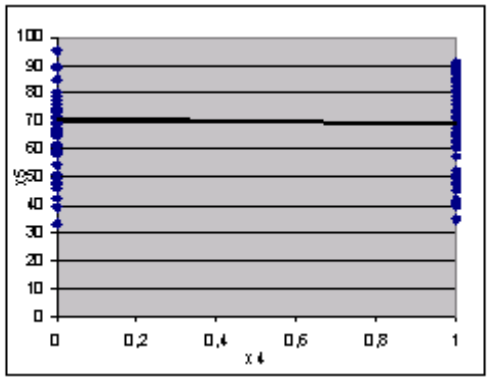

Fig. 3.45. Gráfico de X_3 vs. X_4

El análisis de las gráficas de relaciones entre variables, no hace más que ratificar lo que podemos deducir de un estudio lógico. Es decir, es de esperarse que en vías de uso turístico sea muy alto el porcentaje de vehículos livianos particulares en relación con vehículos pesados de transporte de carga (aunque existan vehículos pesados de transporte de pasajeros, que no aparentan resultar de importancia), como así también es de esperarse que en vías urbanas la presencia de vehículos pesados sea muy reducida (nuevamente la presencia de vehículos pesados de transporte de pasajeros no aparenta influir). Todo esto claro, para los datos con los cuales contamos en este estudio. Esto nos lleva a descartar el empleo de la variable de clasificación, ya que en cierta forma es explicada por la inclusión de las demás variables de entorno.

Como primera prueba analizamos entonces la regresión lineal múltiple con las variables seleccionadas, pero ésta nos da muy bajo ajuste:

- Función de regresión $Y = 1{,}02 + 0{,}01\ X_1 + 0{,}01\ X_2 - 0{,}08\ X_3 + 0{,}01\ X_4$
- Coeficiente de correlación múltiple 0,28
- Coeficiente R^2 de determinación 0,08
- Coeficiente R^2 ajustado 0,07
- Error típico 0,16

Del análisis de la gráfica residuo vs. X_1 de la Figura 3.46 podemos deducir que la relación lineal entre Y y X_1 no es la más adecuada, análisis fundamentado en que X_1 es la variable independiente de significancia (t = 7,09 *p-valor* 0,0000).

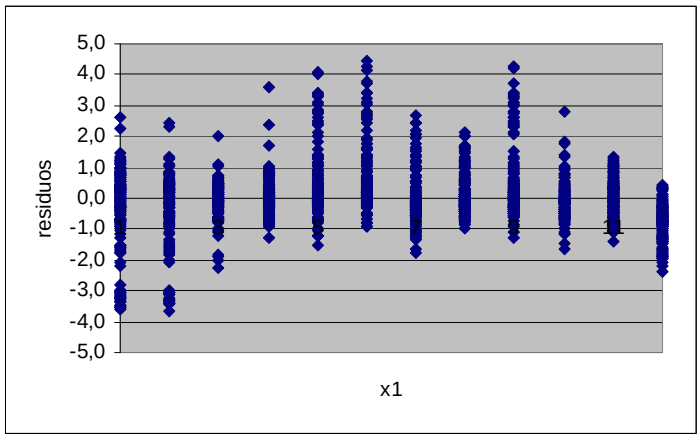

Fig. 3.46. Gráfico de X_1 vs residuos de la regresión múltiple simple

La otra variable de significancia es el uso (t = -7,07 *p-valor* 0,0000) frente a la urbanidad (t = 1,51 *p-valor* 0,13) y el peaje (t = 1,74 *p-valor* 0,0825).

También podemos en este punto ratificar la no existencia de multicolinealidad mediante la matriz de correlación de los estimadores de los coeficientes de la Tabla 3.1, ya que no hay correlación con valores absolutos superiores a 0,5 (no incluido el término constante).

	Cte.	PEAJE	URB	USO	MES
Cte.	1,0000	-0,4523	-0,5454	-0,7494	-0,4832
PEAJE	-0,4523	1,0000	0,0238	0,2004	0,0000
URB	-0,5454	0,0238	1,0000	0,5052	0,0000
USO	-0,7494	0,2004	0,5052	1,0000	0,0000
MES	-0,4832	0,0000	0,0000	0,0000	1,0000

Tabla 3.1. Matriz de correlación de los estimadores de los coeficientes

Realizamos ahora la regresión múltiple con el modelo polinómico de dos variables explicativas (X_1 y X_3) de grado dos y linealizando, el cual tiene la forma:

$$y_t = \alpha_0 + \alpha_1 x_{1t} + \alpha_2 x_{2t} + \alpha_{12} x_{1t} x_{2t} + \alpha_{11} x_{1t}^2 + \alpha_{22} x_{2t}^2 + \varepsilon_t, \quad (3.10)$$

Al intentar hallar la regresión múltiple el software falla, pues los datos de X_3^2 resultan una combinación lineal de las demás columnas, para solucionar esto debemos eliminar este término de la regresión. Los resultados que obtenemos entonces con el programa Statgraphics Plus son:

- La ecuación del modelo ajustado es

  Y = 0,756962 + 0,0960639 X_1 + 0,0803555 X_3 - 0,0259452 X_1X_3 - 0,0052679 X_1^2

- R^2 = 0,23

- R^2 *ajustado* = 0,23
- Error estándar de est. = 0,150163
- Error absoluto medio = 0,104626

Vemos que el ajuste mejora con respecto a la regresión lineal múltiple, lo cual es previsible cuando se observa la relación entre la variable de significancia X_1 e Y. Observamos también que mayor grado en el polinomio sería adecuado, ya que el gráfico de residuos vs. X_1 de la Figura 3.47 nuevamente no muestra una nube de puntos aleatoria.

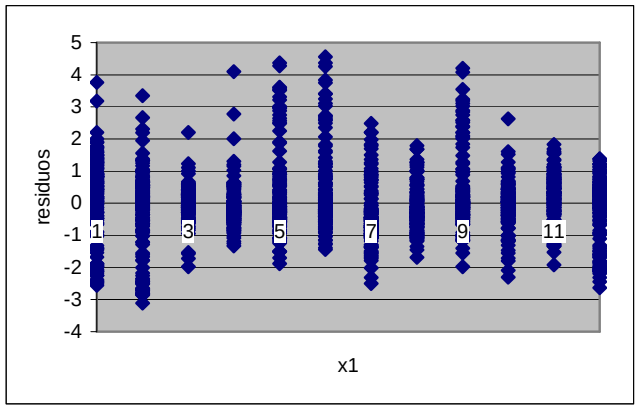

Fig. 3.47. Gráfico de X_1 vs residuos de la regresión múltiple de grado dos

En busca de mejores resultados, y en función del poder computacional del programa empleado, decidimos probar a continuación con un polinomio de grado tres de cuatro variables independientes (se agregan la urbanidad y peaje). Este polinomio presenta la forma:

$$Y = a\,X_1 + b\,X_2 + c\,X_3 + d\,X_4 + e\,X_1^2 + f\,X_2^2 + g\,X_3^2 + h\,X_4^2 + i\,X_1X_2 + j\,X_1X_3 + k\,X_1X_4 + l\,X_2X_3 + m\,X_2X_4 + n\,X_3X_4 + o\,X_1^3 + p\,X_2^3 + q\,X_3^3 + r\,X_4^3 + s\,X_1^2X_2 + t\,X_1^2X_3 + u\,X_1^2X_4 + v\,X_2^2X_1 + w\,X_2^2X_3 + x\,X_2^2X_4 + w\,X_3^2X_1 + y\,X_3^2X_2 + z\,X_3^2X_4 + aa\,X_4^2X_1 + ab\,X_4^2X_2 + ac\,X_4^2X_3 + ad\,X_1X_2X_3 + ae\,X_1X_2X_4 + af\,X_1X_3X_4 + ag\,X_2X_3X_4 \quad (3.11)$$

Aunque, claro está, la expresión se reduce cuando consideramos que por ser variables binarias tanto para X_2, X_3 y X_4 no tiene sentido considerar sus potencias al cuadrado o al cubo, pudiéndose escribir en los términos en los que aparecen como elevados a la primera potencia y permitiéndonos la simplificación del polinomio hasta llegar a la forma:

$$Y = a\,X_1 + b\,X_2 + c\,X_3 + d\,X_4 + e\,X_1^2 + i\,X_1X_2 + j\,X_1X_3 + k\,X_1X_4 + l\,X_2X_3 + m\,X_2X_4 + n\,X_3X_4 + o\,X_1^3 + s\,X_1^2X_2 + t\,X_1^2X_3 + u\,X_1^2X_4 + ad\,X_1X_2X_3 + ae\,X_1X_2X_4 + af\,X_1X_3X_4 + ag\,X_2X_3X_4 \quad (3.12)$$

Dado que algunos términos resultan una combinación lineal de otros o poseen un estadístico *F* bajo, deben ser descartados en la regresión. Los resultados que obtenemos finalmente son:

- Ecuación del modelo ajustado

  $$Y = 0{,}566 + 0{,}119\,X_1 - 0{,}01\,X_1^2 - 0{,}002\,X_1^2X_2 + 0{,}014\,X_1^2X_3 - 0{,}002\,X_1^2X_4 - 0{,}0002\,X_1^3 + 0{,}052\,X_1X_2 - 0{,}049\,X_1X_2X_3 + 0{,}017\,X_1X_2X_4 - 0{,}13\,X_1X_3 - 0{,}048\,X_1X_3X_4 + 0{,}065\,X_1X_4 + 0{,}065\,X_2 - 0{,}289\,X_2X_4 + 0{,}188\,X_2X_3X_4 + 0{,}408\,X_3$$

- $R^2 = 0{,}50$
- R^2ajustado = 0,50
- Error estándar de est. = 0,121258
- Error absoluto medio = 0,085315

El coeficiente de determinación toma valores buenos, sin llegar a los umbrales habituales de aceptación. En busca de mejoras realizamos el análisis de los residuos estándar para descartar datos atípicos. Una vez eliminados estos datos, llegamos a la siguiente ecuación de regresión:

- Ecuación del modelo ajustado

  $$Y = 0{,}479143985 + 0{,}136277392\,X_1 + 0{,}059669021\,X_2 + 0{,}523605787\,X_3 - 0{,}009715863\,X_1^2 + 0{,}034070315\,X_1\,X_2 - 0{,}152392231\,X_1\,X_3 + 0{,}045233251\,X_1\,X_4 - 0{,}000268142\,X_1^3 - 0{,}000651558\,X_1^2\,X_2 + 0{,}014428784\,X_1^2\,X_3 - 0{,}000729828\,X_1^2\,X_4 - 0{,}175791796\,X_2\,X_4 - 0{,}040418127\,X_1\,X_2\,X_3 + 0{,}010884546\,X_1\,X_2\,X_4 - 0{,}040714787\,X_1\,X_3\,X_4 + 0{,}114275601\,X_2\,X_3\,X_4$$

- $R^2 = 0{,}66$
- R^2ajustado = 0,65
- Error estándar de est. = 0,0745175
- Error absoluto medio = 0,0575518

Si bien el R^2 es menor a 0,7, umbral habitualmente empleado en análisis estadísticos, el valor de 0,66 alcanzado no es malo. Además podemos observar en la Figura 3.48 por fin una nube aleatoria de puntos en la gráfica de residuos vs. X_1, indicando la no necesidad de agregar un grado más a la ecuación, cosa que por otro lado resultaría poco práctico.

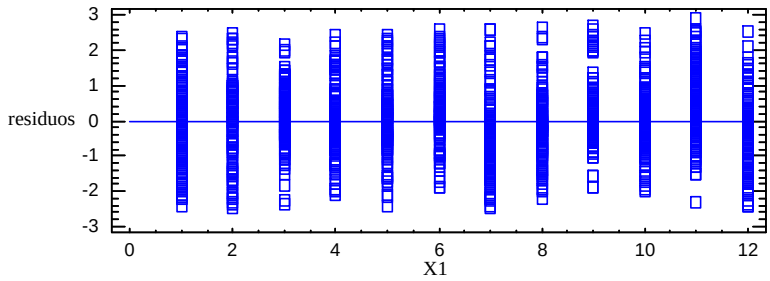

Fig. 3.48. Gráfico de X_1 vs residuos de la regresión múltiple de grado tres

Entre los residuos se observan valores fuera del umbral de valor absoluto 2, pero estos son muy reducidos en comparación con el resto de los datos, e incluso su distribución resulta marcadamente normal como se observa en la Figura 3.49.

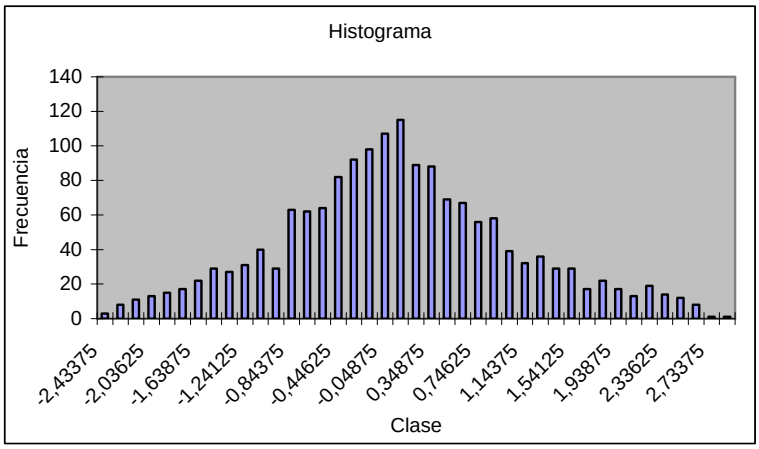

Fig. 3.49. Histograma de residuos de la regresión múltiple de grado tres

Esta normalidad se ratifica cuando efectuamos su análisis estadístico, resultando:

Media = -3,04137E-12

Varianza = 1,0

Desviación típica = 1,0

Asimetría tipificada = 2,07926

Curtosis tipificada = -0,251034

Cuando efectuamos el cálculo de los coeficientes en función de este modelo, para ser presentados en una tabla, observamos valores comparables a los incluidos en la base de datos sobre la cual trabajamos, salvo en el caso de vías de uso turístico, ambiente

rural y sin peaje, en donde los coeficientes obtenidos resultan muy pequeños. Al indagar por la causa de esta anomalía, descubrimos que para el estudio no se contaron con series de datos en vías de estas características, hecho que evidentemente ha influido en la obtención de un modelo no aplicable en estos casos. Por tal razón no debemos considerar como válidos a los coeficientes para esa combinación de variables de entorno, quedando excluido el caso de los resultados.

3.3. Resumen de resultados

Hemos generado los análisis volcados en los puntos anteriores, en donde se indican las regresiones efectuadas (con sus coeficientes y contrastes), las variables alternativas consideradas, los análisis de correlación, los análisis sobre los residuos y demás aspectos especificados en el párrafo anterior. Resta entonces sólo presentar en forma resumida la metodología desarrollada.

3.3.1. Pasos para la aplicación de los modelos

Paso 1: Obtención del TD_{real} sobre la vía, considerado desde las 0 horas hasta las 24 horas. Indicar día de la semana (*DS*), mes (*M*), uso de la vía (*C*), urbanidad (*U*) y existencia o no de peaje (*P*).

Paso 2: Establecer la tasa de crecimiento del tránsito estimada para la vía durante el año en estudio. Para esto realizar su estimación directa, o emplear el algoritmo o la Tabla 3.1 para su estimación mediante la variación del parque automotor durante el año en estudio y para la localidad en donde se encuentra el punto analizado:

$$TCT = 35{,}596896 - (243{,}628504 / VP) + (555{,}412790 / VP^2) - (585{,}523100 / VP^3) + (283{,}681553 / VP^4) - (51{,}088958 / VP^5) \quad (3.14)$$

Donde

TCT = Tasa crecimiento tránsito

VP = Variación parque automotor

VP	*TCT*
0,5	-10,1
1,0	-5,5
1,5	-4,2
2,0	-4,4
2,5	-3,7
3,0	-2,3
3,5	-0,5
4,0	1,3
4,5	3,1
5,0	4,8
5,5	6,4
6,0	7,9
6,5	9,3
7,0	10,5
7,5	11,7
8,0	12,7
8,5	13,7
9,0	14,6
9,5	15,5
10,0	16,2

Tabla 3.2. Tasa de Crecimiento de Tránsito en función del registro automotor

Paso 3: En función del día del año en la que se determina el TD_{real} y e la *TCT* obtenida, descontar la tendencia en forma proporcional para establecer un TD_0, con:

$$TD_0 = TD_{real} \times (1 - \frac{TCT}{100} \times \frac{DA}{365}) \quad (3.15)$$

Donde:

TD_0 = Tránsito diario sin tendencia

TD_{real} = Tránsito diario directamente establecido

TCT = Tasa de crecimiento del tránsito

DA = Día del año del dato (1 para el 1º de enero, ..., 365 para el 31º de diciembre)

Paso 4: Determinar los coeficientes diarios empleando, en función de las variables de entorno, los modelos o la Tabla 3.2:

Modelo para toda la clase de uso turístico (C = 0)

$CD = -0{,}043715\ DS^2 + 0{,}363511\ DS + 0{,}452025$ (3.16)

Modelo para clase de uso comercial (C = 1) y sin peaje (P = 0)

$CD = 1{,}299385 - 0{,}175416\ DS + 0{,}110582\ DS^2 \ln DS - 0{,}033388\ DS^3 + 0{,}001731\ e^{DS}$ (3.17)

Modelo para clase de uso comercial (C = 1) y con peaje (P = 1)

$$CD = 0{,}002781\ DS^5 - 0{,}053475\ DS^4 + 0{,}378762\ DS^3 - 1{,}184775\ DS^2 + 1{,}434157\ DS + 0{,}758143 \quad (3.18)$$

Donde:

CD = Coeficiente diario

DS = Día de la semana (1 para domingo, ..., 7 para sábado)

USO	PEAJE	COEFICIENTE DIARIO						
		DOM	LUN	MAR	MIE	JUE	VIE	SAB
Turístico	*con o sin peaje*	0,772	1,004	1,149	1,207	1,177	1,059	0,855
comercial	*sin peaje*	1,095	1,001	1,000	1,008	0,955	0,866	1,061
comercial	*con peaje*	1,336	1,151	0,969	0,937	0,924	0,845	1,005

Tabla 3.3. Coeficientes de corrección diarios

Paso 5: Determinar los coeficientes mensuales empleando, en función de las variables de entorno, el modelo o la Tabla 3.3:

$$CM = 0{,}479143985 + 0{,}136277392\ M + 0{,}059669021\ U + 0{,}523605787\ C - 0{,}009715863\ M^2 + 0{,}034070315\ M\ U - 0{,}152392231\ M\ C + 0{,}045233251\ M\ P - 0{,}000268142\ M^3 - 0{,}000651558\ M^2\ U + 0{,}014428784\ M^2\ C - 0{,}000729828\ M^2\ P - 0{,}175791796\ U^2\ P - 0{,}040418127\ M\ U\ C + 0{,}010884546\ M\ U\ P - 0{,}040714787\ M\ C\ P + 0{,}114275601\ U\ C\ P \quad (3.19)$$

Donde:

CM = Coeficiente mensual

M = Mes del año (1 para enero, ..., 12 para diciembre)

C = Clase de uso (0 para turístico y 1 para comercial)

U = Urbanidad (0 para ambiente rural y 1 para ambiente urbano)

P = Peaje (0 para sin peaje y 1 para con peaje)

El modelo no es aplicable en vías turísticas rurales y sin peajes, para las cuales se recomienda el empleo de la metodología clásica.

URB	USO	PEAJE	COEFICIENTE MENSUAL											
			ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
Rural	*Turíst*	*sin*	*caso no aplicable*											
Rural	*Turíst*	*con*	0,650	0,798	0,922	1,021	1,092	1,134	1,146	1,125	1,071	0,982	0,855	0,690
Rural	*Comer*	*sin*	0,991	0,987	0,990	0,997	1,006	1,018	1,029	1,038	1,044	1,045	1,039	1,025
Rural	*Comer*	*con*	0,995	0,993	0,997	1,003	1,011	1,019	1,025	1,028	1,026	1,017	1,000	0,974
Urbano	*Turíst*	*sin*	0,699	0,836	0,949	1,037	1,098	1,130	1,131	1,101	1,037	0,937	0,801	0,627
Urbano	*Turíst*	*con*	0,578	0,769	0,935	1,074	1,184	1,264	1,313	1,327	1,307	1,250	1,154	1,019
Urbano	*Comer*	*sin*	1,044	1,032	1,024	1,020	1,018	1,016	1,012	1,005	0,994	0,976	0,950	0,914
Urbano	*Comer*	*con*	0,997	0,998	1,002	1,009	1,015	1,021	1,023	1,020	1,012	0,995	0,969	0,933

Tabla 3.4. Coeficientes de corrección mensuales

Paso 6: Calcular *TMDA* mediante:

$$TMDA = TD_0 \times CD \times CM \times \left(1 + \frac{TCT}{100} \times \frac{1}{2}\right) \qquad (3.20)$$

En caso de contarse con más datos de tránsitos diarios, aplicar la metodología y calcular la estadística de los resultados obtenidos para convalidar o no la media de los mismos mediante la normalidad de los resultados.

Capítulo 4 - Validación y discusión

4.1. Validación de los modelos

Una vez obtenido un modelo es necesario analizar su validez, es decir, comprobar que sea confiable para el fin por el cual ha sido desarrollado.

"...Los modelos de regresión pueden ser *validados* en otro conjunto de datos de similares características -extraídos de la misma población-, con el fin de evaluar su fiabilidad..."[51]

Esta fiabilidad se pone de manifiesto en un modelo que ha sido desarrollado para determinación del *TMDA*, justamente cuando se analizan los valores de este parámetro que se obtienen mediante su empleo en diversas situaciones.

"...Las muestras de datos sujetas a las mismas técnicas de análisis permiten generalizar el comportamiento de la población... no obstante se debe analizar la variabilidad obtenida para así estar seguros, con cierto nivel de confiabilidad, que se puede aplicar a otro número de casos no incluidos, y que forman parte de la población...

En el análisis de volúmenes de tránsito, la media poblacional o tránsito promedio diario anual, *TPDA*, se estima con la media muestral...

Estadísticamente de ha demostrado que las medias de diferentes muestras, tomadas de la misma población, se distribuyen normalmente alrededor de la media poblacional...."[52]

Un modelo generado en función de muestras aproximadamente normales de series de tránsito y variables de entorno, como es el caso del nuestro, y alimentado con datos de tránsito diarios, debe dar como resultado un conjunto de datos con una

[51] "Construcción de modelos de regresión multivariantes", L. Molinero, Alce Ingeniería, España 2002.
[52] "Ingeniería de tránsito, fundamentos y aplicaciones", R. Cal y Mayor, J. Cárdenas, Alfaomega 7°ed., México 1995.

distribución simétrica, en donde la media sea el *TMDA* y un menor desvío estándar signifique mayor adaptación a la realidad (validez del modelo).

Basados en esta línea de pensamiento nos proponemos para analizar la validez del modelo obtenido, efectuar su aplicación en tramos de vía con demanda y condiciones de entorno conocidas, y comparar los valores obtenidos con los resultantes de la aplicación del método clásico, en función de los coeficientes relevados en una vía de la zona que sirve a similares itinerarios de tránsito.

Por esto analizamos los siguientes casos, que nos dan una combinación de las condiciones de entorno:

- Calle 28 entre 489 y 490 de La Plata. Vía urbana, de tránsito comercial y sin peaje.
- Ruta Nacional Nº9 entre Córdoba y Jesús María. Vía rural, de tránsito comercial y con peaje.
- Ruta Nacional Nº20 entre Córdoba y Carlos Paz. Vía rural, de tránsito turístico y con peaje.
- Autopista Buenos Aires – La Plata, en su tramo por Dock Sud. Vía urbana, de tránsito comercial y con peaje.

4.1.1. Primer caso de validación

La vía que seleccionamos para este primer caso es la Calle 28 de la ciudad de La Plata, la cual sirve de acceso y egreso a la ciudad desde sus localidades satélites de City Bell y Villa Elisa.

Para conocer la demanda real sobre la vía hemos colocado durante todo el año 2004 el contador automático que se observa en la Figura 4.1, ubicado en Calle 28 entre las calles 489 y 490. La vía posee en este tramo entorno urbano, no cuenta con peaje y resulta de uso mayoritariamente comercial.

Fig. 4.1. Contador automático de tránsito empleado en el estudio

El año 2004 fue año bisiesto, por lo cual obtuvimos 366 datos de tránsito diario. Con estos registros podemos confeccionar la curva de *TD* (Tránsito Diario medido) vs. día del año que observamos en la Figura 4.2, junto con su línea de tendencia positiva. Resulta muy importante destacar que esta serie no ha sido incluida entre las series empleadas en las regresiones, es decir que los modelos desarrollados no están influenciados por la misma.

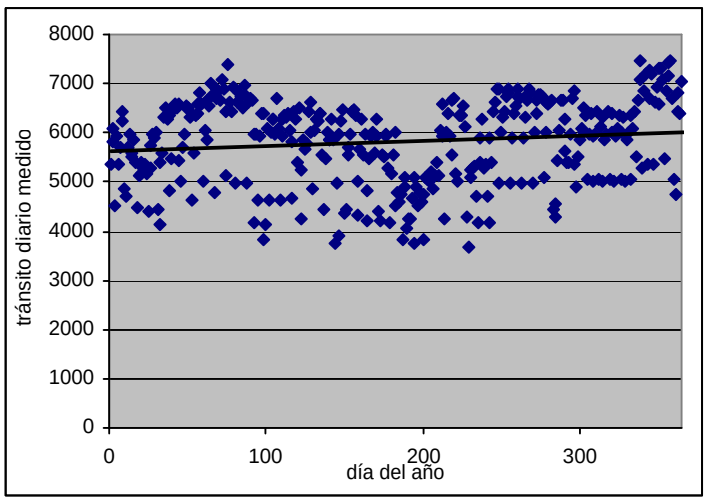

Fig. 4.2. Valores de *TD* durante el año 2004 para primer caso de validación

Comenzamos el estudio de validación analizando estadísticamente la muestra, obteniendo como resultado:

Frecuencia = 366

Media = 5811,42

Varianza = 709963,0

Desviación típica = 842,593

Mínimo = 3686,0

Máximo = 7478,0

Rango = 3792,0

Asimetría tipificada = -3,35455

Curtosis tipificada = -2,45725

Estos datos se complementan con la gráfica correspondiente de la Figura 4.3 que también nos permiten observar esta tendencia a la normalidad.

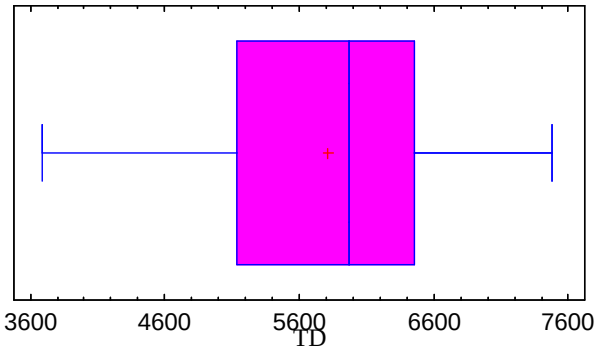

Fig. 4.3. Gráfico de caja y bigotes para *TD* en primer caso de validación

4.1.1.1. Determinación del *TMDA* mediante la metodología clásica

Realizamos la aplicación sobre esta vía según los lineamientos ya explicados en este documento. Para eso en primer lugar consideramos que no existen datos de conteos previos sobre la misma, razón por la cual debemos analizar la existencia de estos sobre vías cercanas con similares características.

Como dijimos la Calle 28 en el tramo en estudio sirve mayoritariamente al tránsito que ingresa y egresa a la ciudad de La Plata, desde las localidades satélites a ésta de City Bell y Villa Elisa. En forma paralela a esta vía, se encuentra la RN Nº1, conocida como el Camino Gral. Belgrano, y la RP Nº14, conocida como el Camino Centenario, que se observan en la Figura 4.4.

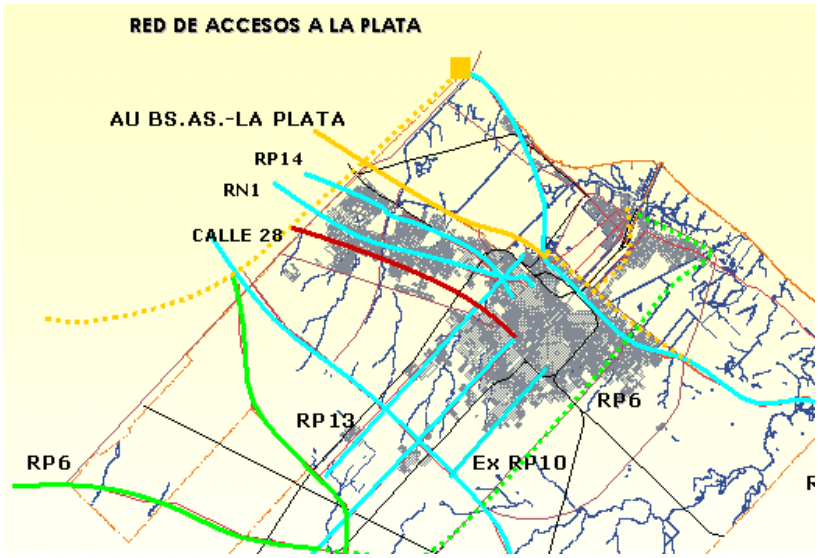

Fig. 4.4. Vías de acceso a la ciudad de La Plata

Hasta el año 2002 estas vías servían también al tránsito desde y hacia la Capital Federal y el Gran Buenos Aires, pero desde la inauguración del último tramo de la Autopista Buenos Aires – La Plata en ese año, éste ha sido trasladado en su mayoría, sirviendo ahora prioritariamente ambas arterias a itinerarios similares a los de la vía en estudio.

Mas allá de esta similitud de características (que podría resultar en cierto modo discutible), la única posibilidad de conseguir datos de tránsito en esa zona puede darse sobre los caminos citados (lo cual no es discutible). A lo que debe agregarse que en concordancia con el tramo en estudio ambos circulan por zonas urbanas, sin peajes y sirviendo mayoritariamente a tránsito comercial, al igual que la Calle 28.

Todo esto nos lleva en la aplicación de la metodología clásica al aceptar series de algunas de estas vías para ser aplicadas en nuestro estudio.

Cuando analizamos nuestra base de datos, vemos que contamos sólo con datos para el Camino Centenario durante el ciclo 2003, suministrados por la Dirección de Vialidad de la Provincia de Buenos Aires. Con estos datos puede elaborarse la Tabla 4.1 correspondiente.

COEFICIENTE MENSUAL											
ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
1,093	1,034	0,979	0,990	0,981	1,006	1,042	1,017	0,994	0,963	0,967	0,951

COEFICIENTE DIARIO						
DOM	LUN	MAR	MIE	JUE	VIE	SAB
1,269	1,042	0,948	0,972	0,941	0,816	1,136

Tabla 4.1. Coeficientes mensuales y diarios sobre el Camino Centenario

Sabemos además que durante ese ciclo se registró sobre la vía un crecimiento de tránsito del 2,7 %.

Con estos coeficientes y los datos de tránsito diario considerados sin tendencia, realizamos el cálculo del *TMDA* día a día.

Como la aplicación de la metodología no explica como estimar la tasa de crecimiento por la cual se deben afectar los cálculos, lo más coherente es suponer una tasa idéntica a la del ciclo anterior (salvo que existan serios indicios de que esto es inadecuado), es decir 2,7 %, tal cual lo realizamos en estos cálculos. Recordemos que la inclusión en este caso de este término es necesario porque en la obtención de estos coeficientes realizamos el descuento del crecimiento.

La estadística obtenida con esta aplicación resulta:

Frecuencia = 366

Media = 5918,83

Varianza = 633590,0

Desviación típica = 795,984

Mínimo = 3397,0

Máximo = 8066,0

Rango = 4669,0

Asimetría tipificada = -1,09412

Curtosis tipificada = 2,97803

La estadística se complementa con los gráficos de la Figura 4.5, 4.6 y 4.7 correspondientes.

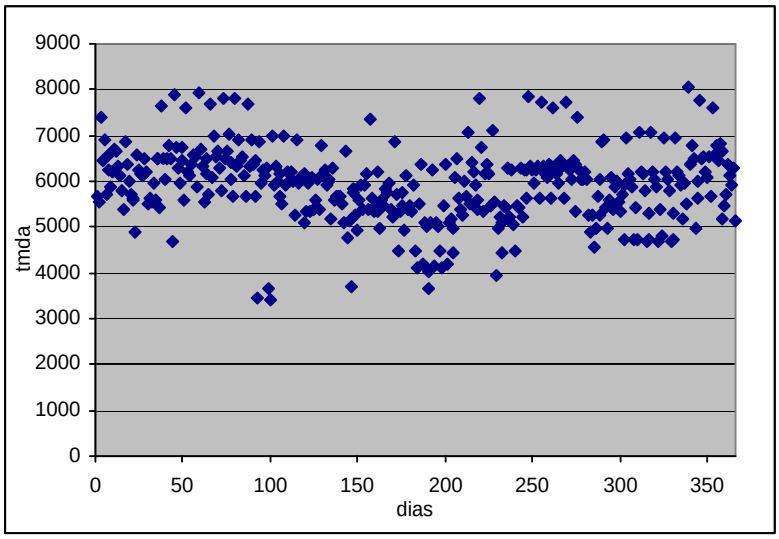

Fig. 4.5. Nube de resultados por metodología clásica, en primer caso de validación

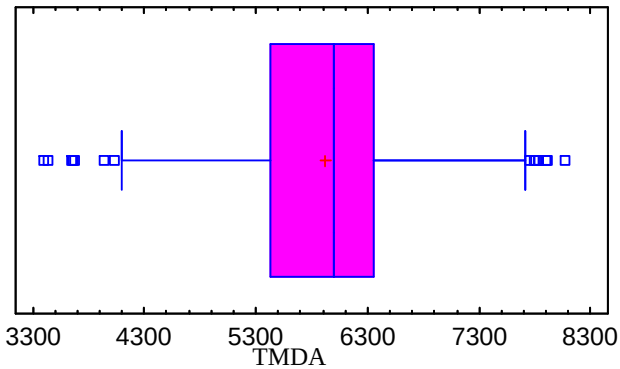

Fig. 4.6. Gráfico de caja y bigotes para resultados por metodología clásica en primer caso de validación

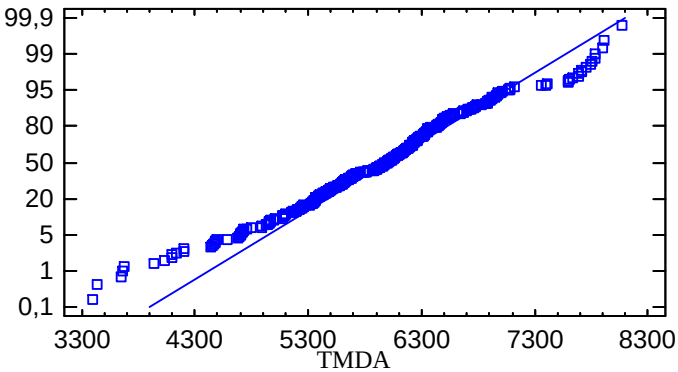

Fig. 4.7. Gráfico de probabilidad normal para resultados por metodología clásica en primer caso de validación

4.1.1.2. Determinación del *TMDA* mediante la metodología desarrollada

Para realizar el cálculo del *TMDA* día a día debemos en primer lugar estimar la tasa de crecimiento del tránsito mediante la variable variación del parque automotor. El registro de automotores para la localidad de La Plata ascendía en el año 2003 a 268.977 veh, registrándose para el 2004 unos 280.433 veh, por lo tanto la variación en el registro asciende a 4,3 %. Con este valor ingresamos a la Tabla 3.1 y determinamos que el incremento de tránsito correspondiente es de 2,3 %.
Luego, determinamos los coeficientes diarios y los volcamos en la Tabla 4.2.

COEFICIENTE DIARIO						
DOMINGO	LUNES	MARTES	MIERCOLES	JUEVES	VIERNES	SABADO
1,095	1,001	1,000	1,008	0,955	0,866	1,061

Tabla 4.2. Coeficientes diarios según metodología desarrollada, en primer caso de validación

Seguidamente obtenemos los coeficientes mensuales de la Tabla 4.3.

COEFICIENTE MENSUAL											
ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
1,044	1,032	1,024	1,020	1,018	1,016	1,012	1,005	0,994	0,976	0,950	0,914

Tabla 4.3. Coeficientes mensuales según metodología desarrollada, en primer caso de validación

Con estos valores podemos calcular el *TMDA* día a día, presentándose la siguiente estadística:

Frecuencia = 366

Media = 5811,15

Varianza = 588680,0

Desviación típica = 767,255

Mínimo = 3701,0

Máximo = 7701,0

Rango = 4000,0

Asimetría tipificada = -2,06178

Curtosis tipificada = 0,161674

Con la cual elaboramos las gráficas de la Figura 4.8, 4.9 y 4.10.

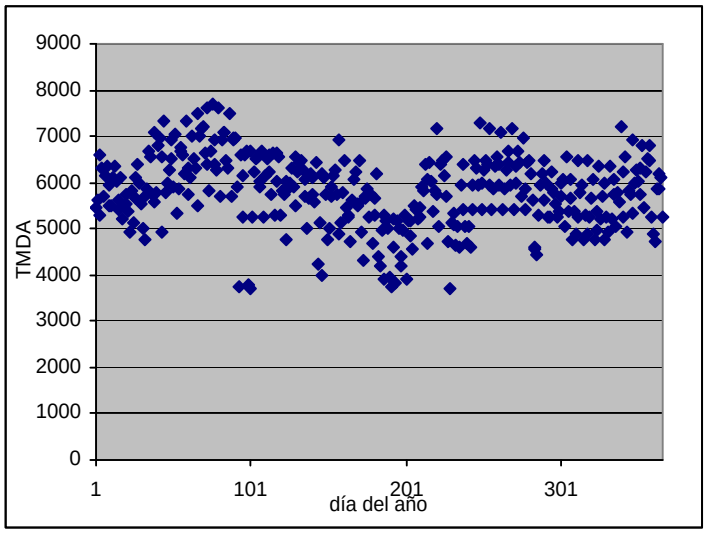

Fig. 4.8. Valores de *TMDA* por metodología desarrollada, en primer caso de validación

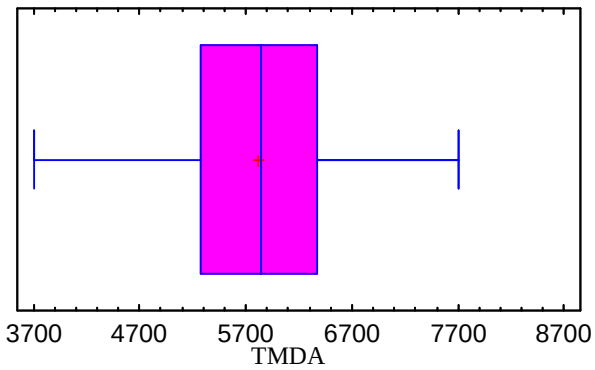

Fig. 4.9. Gráfico de caja y bigotes para resultados por metodología desarrollada, en primer caso de validación

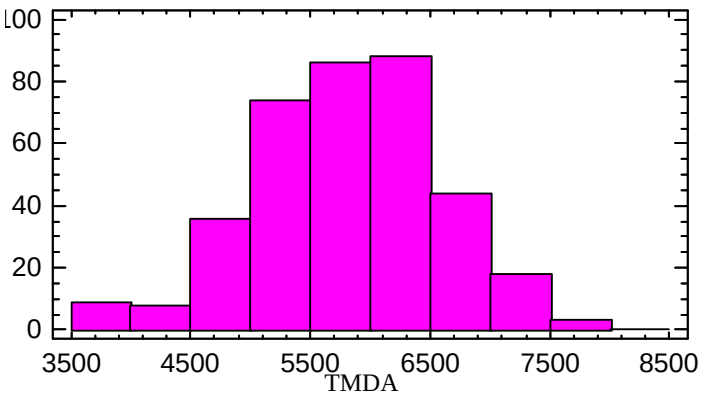

Fig. 4.10. Histograma de los resultados por metodología desarrollada, en primer caso de validación

4.1.1.3. Evaluación de resultados

El resumen de resultado puede observarse en la Tabla 4.4.

PARÁMETRO	**POR METODOLOGIA CLASICA**	**POR METODOLOGIA DESARROLLADA**
MEDIA	5919	5811
DESVIACION TIPICA	796	767
RANGO	4669	4000
ASIMETRIA TIPIFICADA	-1,09	-2,06
CURTOSIS TIPIFICADA	2,98	0,16

Tabla 4.4. Resumen de resultados para el primer caso

Como podemos observar en este primer caso de validación los resultados obtenidos son muy buenos. Sabíamos que el *TMDA* real asciende a 5811 veh/día, y determinamos con el método clásico una media de 5919 veh/día con una distribución prácticamente normal, es decir obtenemos un resultado general sólo un 1,8 % por encima del valor real y con un desvío estándar de 796 veh/día, o sea que con aproximadamente un 70 % de los datos obtenemos un entorno de ± 13,7 % del valor real (dado que está probado que en el intervalo de la media ± el desvío estándar se encuentran aproximadamente el 68 % de los datos en una distribución normal[53]). Todo esto no hace más que ratificar la decisión que tomamos de aceptar las series del Camino Centenario para su aplicación sobre la Calle 28.

Por su parte, los resultados obtenidos con la metodología desarrollada, siempre para este caso en particular, son aun mejores, ya que la media del *TMDA* calculado de 5811 veh/día (exactamente el valor real) con una distribución normal, resultando el desvío estándar de 767 veh/día, o sea que con aproximadamente un 70 % de los datos obtenemos un entorno de ± 13,2 % el valor real.

4.1.2. Segundo caso de validación

Analizamos en este caso la Ruta Nacional Nº 9 en su tramo entre las ciudades de Córdoba y Jesús María, conocida como la RN9norte, que podemos observar en la Figura 4.11.

[53] "La distribución normal", S. Pertegas Días, S. Pita Fernández, Fisterra, España 2001.

Fig. 4.11. Red de Accesos a Córdoba

Para esta vía obtenemos la serie de transito diarios completa para el año 2001, la cual observamos junto con su línea de tendencia negativa en la Figura 4.12.

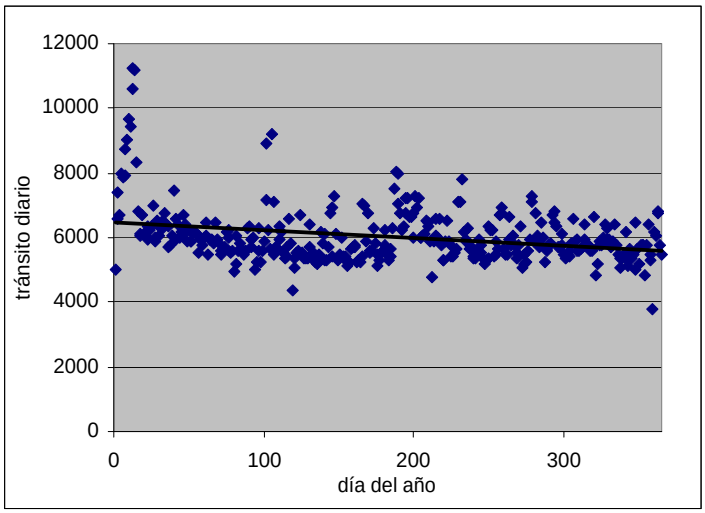

Fig. 4.12. Gráfico día del año vs tránsito diario medido, en segundo caso de validación

Estos valores nos permiten obtener la siguiente estadística:

Frecuencia = 365

Media = 6044,74

Varianza = 756350,0

Desviación típica = 869,684

Mínimo = 3807,0

Máximo = 11256,0

Rango = 7449,0

Asimetría tipificada = 19,4716

Curtosis tipificada = 39,5669

Valores con los que podemos construir el gráfico de la Figura 4.13.

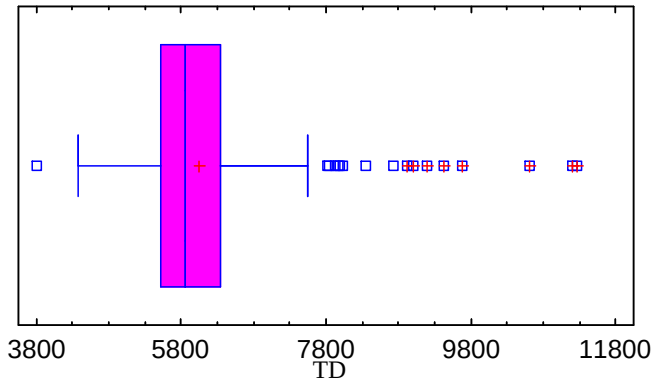

Fig. 4.13. Gráfico de caja y bigotes para los tránsitos medidos, en segundo caso de validación

4.1.2.1. Determinación de *TMDA* mediante la metodología clásica

Para este tramo de vía contamos con los factores de corrección denunciados en un trabajo técnico[54] sobre el propio punto en estudio (en realidad en el trabajo estos valores vienen expresados en porcentuales, los que hemos debido adaptar a la forma de coeficientes), los cuales son adoptados, aunque esto juegue como una ventaja comparativa en la aplicación de la metodología ya que no se trata de series sobre puntos cercanos que hubieran implicado la subjetividad de decidir sobre su empleo o no. Cabe aclarar que estos factores no consideran los descuentos por incrementos del tránsito, sino que están generados mediante los conceptos clásicos ya enunciados en este documento. Por tal razón su aplicación debe efectuarse en forma directa con los tránsitos diarios medidos, sin que sean necesarias consideraciones adicionales.

Los factores obtenidos del trabajo son los que se observan en la Tabla 4.5.

RN9 NORTE

COEFICIENTES MENSUALES											
ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
0,833	1,190	1,041	1,041	1,041	1,041	0,925	0,925	1,041	0,925	1,041	0,925

COEFICIENTES DIARIOS						
DOM	LUN	MAR	MIE	JUE	VIE	SAB
1,021	1,021	1,021	1,021	1,021	0,894	1,021

Tabla 4.5. Coeficientes para la metodología clásica, en segundo caso de validación

Con estos coeficientes calculamos día a día los *TMDA* que se observan en la Figura 4.14.

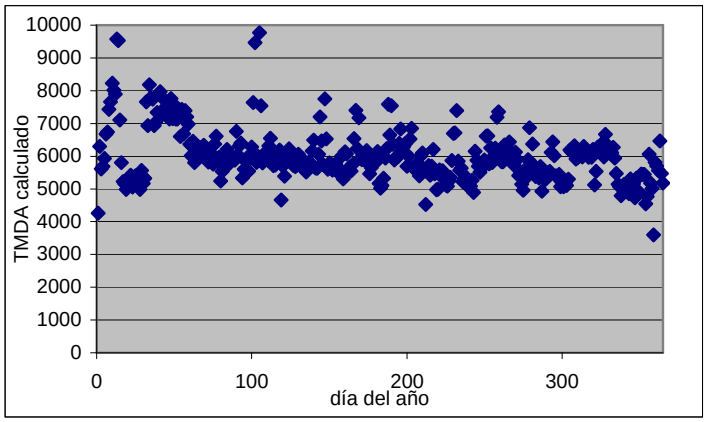

Fig. 4.14. *TMDA* por metodología clásica, en segundo caso de validación

Con estos valores obtenemos la siguiente estadística:

Frecuencia = 365

Media = 6002,98

Varianza = 672258,0

Desviación típica = 819,914

Mínimo = 3595,0

Máximo = 9771,0

Rango = 6176,0

Asimetría tipificada = 10,4034

Curtosis tipificada = 13,154

Con la que elaboramos el gráfico de la Figura 4.15.

[54] "Caracterización de errores de muestreo en censos de volumen y composición", M. Herz, J. Galárraga, M. Maldonado, XIV Congreso Argentino de Vialidad y Tránsito, Argentina 2005.

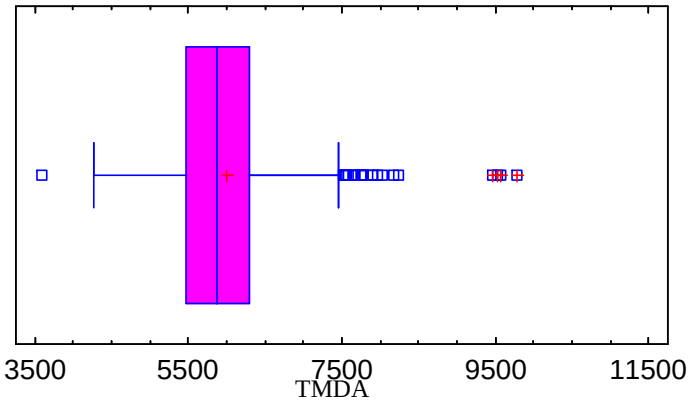

Fig. 4.15. Gráfico de caja y bigotes de *TMDA* por metodología clásica, en segundo caso de validación

4.1.2.2. Determinación del TMDA mediante la metodología desarrollada

Para poder aplicar la metodología debemos analizar las condiciones de borde de la vía, las cuales son un ambiente rural, con motivos de viaje mayoritariamente comercial y con peaje. Para estos datos, los coeficientes que corresponden por los modelos desarrollados son los de la Tabla 4.6.

COEFICIENTE MENSUAL											
ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
0,995	0,993	0,997	1,003	1,011	1,019	1,025	1,028	1,026	1,017	1,000	0,974

COEFICIENTE DIARIO						
DOM	LUN	MAR	MIE	JUE	VIE	SAB
1,336	1,151	0,969	0,937	0,924	0,845	1,005

Tabla 4.6. Coeficientes para metodología desarrollada, en segundo caso de validación

En cuanto a la tasa de crecimiento del tránsito, contamos con el dato de que en el año 2000 el parque automotor registrado en la ciudad de Córdoba asciende a 367.245 veh, mientras que en el año 2001 esta cifra llega a 376.743 veh, es decir que se registra un incremento en el parque del 2,6 %, a lo que corresponde de acuerdo al algoritmo obtenido una tasa de crecimiento del tránsito del –3,7 %, lo cual en cierta forma confirma la tendencia negativa de los tránsitos diarios medidos (Figura 4.12). Mediante los datos enunciados, obtenemos los *TMDA* que se observan en la Figura 4.16.

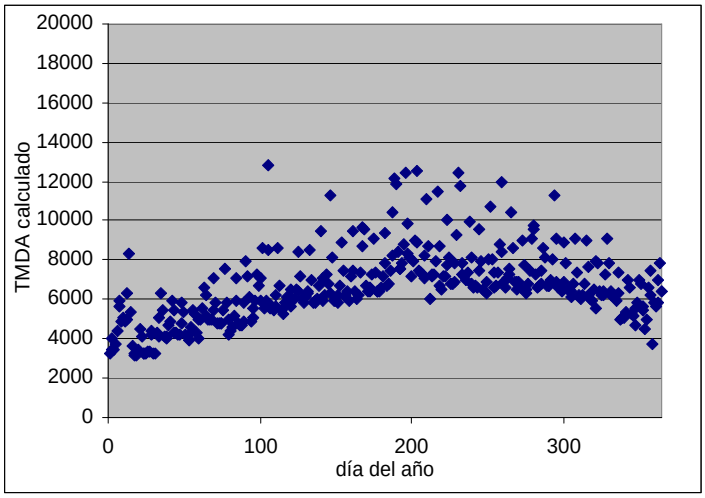

Fig. 4.16. *TMDA* por metodología desarrollada en segundo caso de validación

Valores con los cuales podemos establecer la siguiente estadística:

Frecuencia = 365

Media = 6615,38

Varianza = 629721,6

Desviación típica = 793,55

Mínimo = 3128,0

Máximo = 12836,0

Rango = 9708,0

Asimetría tipificada = 5,82467

Curtosis tipificada = 4,67669

Con la que podemos construir el gráfico de la Figura 4.17.

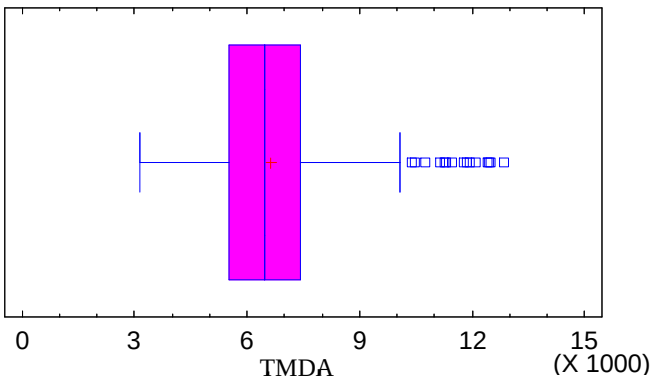

Fig. 4.17. Gráfico de caja y bigotes de *TMDA* por la metodología desarrollada, en segundo caso de validación

4.1.2.3. Evaluación de resultados

Con estos resultados podemos generar la Tabla 4.7 comparativa.

PARÁMETRO	**POR METODOLOGIA CLÁSICA**	**POR METODOLOGIA DESARROLLADA**
MEDIA	6003	6615
DESVIACION TIPICA	820	793
RANGO	6176	9708
ASIMETRIA TIPIFICADA	10,4	5,82
CURTOSIS TIPIFICADA	13,15	4,68

Tabla 4.7. Resumen de resultados para el segundo caso

Como vemos el $TMDA_{REAL}$ asciende a 6045 veh/día. Mediante el empleo de la metodología clásica, con los detalles particulares ya señalados para este caso, hemos obtenido un *TMDA* medio que difiere en un 0,7 % del real, pero con una distribución con asimetría tipificada de 10,4. Por su parte el *TMDA* medio obtenido por la metodología desarrollada difiere en 9,4 % del real, pero con una distribución con coeficiente de asimetría de 5,8.

Por lo expuesto podemos deducir que, si se calcula el *TMDA* con la metodología clásica y con menos datos (por ejemplo serie de 1, 3 o 7 días), siempre para este caso y con las salvedades ya enunciadas a su favor, no sería simple la obtención de un resultado ajustado por el alejamiento de los valores obtenidos de lo que es una distribución normal, más allá de que la media total se aproxima ajustadamente al valor real. En cambio si se realiza el cálculo con la metodología desarrollada las probabilidades de resultados adecuados es buena, aunque no se halla podido establecer una tasa de crecimiento tan cercana a la real.

4.1.3. Tercer caso de validación

Analizamos en este caso la Ruta Nacional Nº 20, en su tramo entre la ciudad de Córdoba y la localidad de Carlos Paz, que posee características de autopista y se observa en la Figura 4.18.

Fig. 4.18. Red de Accesos a Córdoba

Para esta vía obtenemos la serie de transito diarios completa para el año 2001, que se observa en la Figura 4.19, en donde también se ha incluido la línea de tendencia.

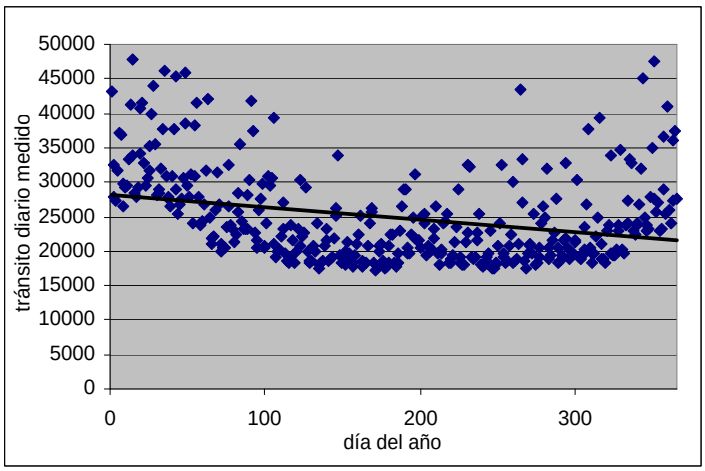

Fig. 4.19. Gráfico día del año vs tránsito diario medido, en tercer caso de validación

Estos valores nos permiten obtener la siguiente estadística:

Media = 24774,5

Varianza = 4,43674E7

Desviación típica = 6660,89

Mínimo = 17290,0

Máximo = 47868,0

Rango = 30578,0

Asimetría tipificada = 10,0659

Curtosis tipificada = 4,79866

Valores con los que podemos construir el gráfico de la Figura 4.20.

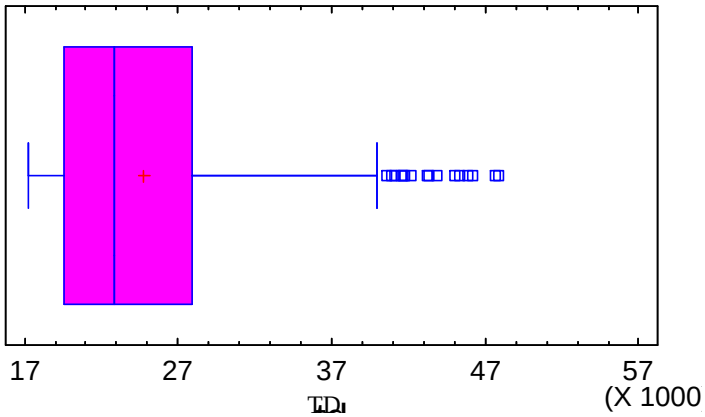

Fig. 4.20. Gráfico de caja y bigotes para los tránsitos medidos en tercer caso de validación

4.1.3.1. Determinación del *TMDA* mediante la metodología clásica

Nuevamente, para este tramo de vía contamos sólo con los factores de corrección denunciados en el trabajo técnico ya citado[55] sobre el propio punto en estudio, y expresados en valores porcentuales, los que hemos debido adaptar a la forma de coeficientes. Adoptamos estos valores, aunque no se trate de series sobre puntos cercanos que hubieran implicado la subjetividad de decidir sobre su empleo o no. Aclaramos de vuelta que estos factores no consideran los descuentos por incrementos del tránsito, sino que están generados mediante los conceptos clásicos ya enunciados en este documento. Por tal razón, su aplicación debe efectuarse en forma directa con los tránsitos diarios medidos, sin que sean necesarias consideraciones adicionales.

Los factores obtenidos del trabajo son los que se observan en la Tabla 4.8.

[55] "Caracterización de errores de muestreo en censos de volumen y composición", M. Herz, J. Galárraga, M. Maldonado, XIV Congreso Argentino de Vialidad y Tránsito, Argentina 2005.

RP20

COEFICIENTES MENSUALES											
ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
0,694	0,833	1,041	1,041	1,190	1,190	1,041	1,041	1,019	1,041	1,041	0,833

COEFICIENTES DIARIOS						
DOM	LUN	MAR	MIE	JUE	VIE	SAB
0,753	1,100	1,917	1,917	1,100	1,021	0,894

Tabla 4.8. Coeficientes para la metodología clásica, en tercer caso de validación

Con estos coeficientes calculamos día a día los *TMDA* que se observan en la Figura 4.21.

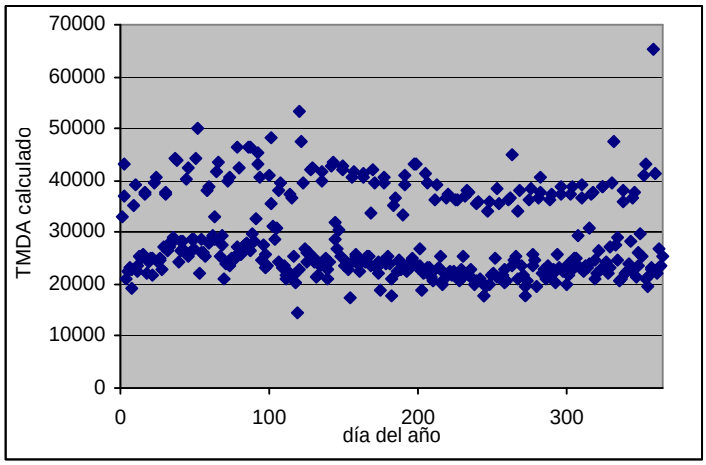

Fig. 4.21. *TMDA* por metodología clásica, en tercer caso de validación

Con estos valores obtenemos la siguiente estadística:

Frecuencia = 365

Media = 28914,0

Varianza = 6,6398E7

Desviación típica = 8148,5

Mínimo = 14415,0

Máximo = 65316,0

Rango = 50901,0

Asimetría tipificada = 7,24012

Curtosis tipificada = 0,747249

Con la que elaboramos el gráfico de la Figura 4.22.

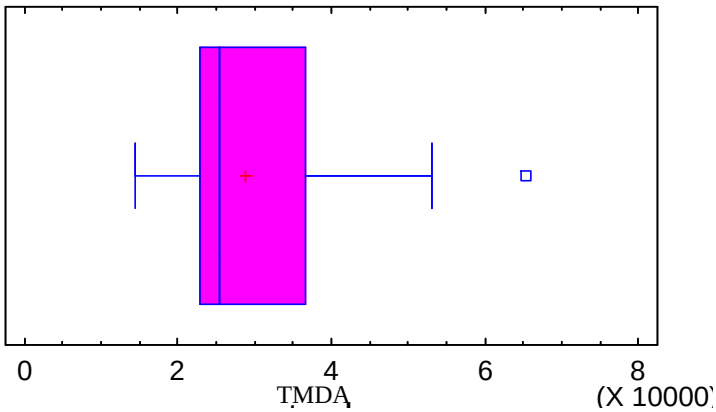

Fig. 4.22. Gráfico de caja y bigotes de *TMDA* por metodología clásica, en tercer caso de validación

4.1.3.2. Determinación del *TMDA* mediante la metodología desarrollada

Para poder aplicar la metodología debemos analizar las condiciones de borde de la vía, las cuales son un ambiente rural, con motivos de viaje mayoritariamente turístico y con peaje. Para estos datos, los coeficientes que corresponden por los modelos desarrollados son los de la Tabla 4.9.

RP20

COEFICIENTES MENSUALES											
ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
0,650	0,798	0,922	1,021	1,092	1,134	1,146	1,125	1,071	0,982	0,855	0,690

COEFICIENTES DIARIOS						
DOM	LUN	MAR	MIE	JUE	VIE	SAB
0,772	1,004	1,149	1,207	1,177	1,059	0,855

Tabla 4.9. Coeficientes para metodología desarrollada, en tercer caso de aplicación

En cuanto a la tasa de crecimiento del tránsito debemos realizar el mismo análisis que con la RN9norte, con el cual determinamos una tasa de crecimiento del tránsito del –3,7 %.

Mediante los datos enunciados, obtenemos los *TMDA* que se observan en la Figura 4.23.

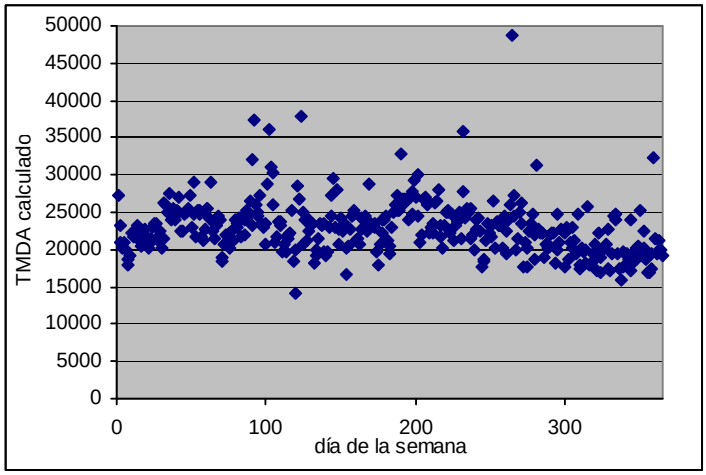

Fig. 4.23. TMDA por metodología desarrollada, en tercer caso de validación

Valores con los cuales podemos construir la siguiente estadística:

Frecuencia = 365

Media = 22866,1

Varianza = 1,31017E7

Desviación típica = 3619,63

Mínimo = 14123,0

Máximo = 48612,0

Rango = 34489,0

Asimetría tipificada = 3,4548

Curtosis tipificada = 2,3741

Con la que podemos construir el gráfico de la Figura 4.26.

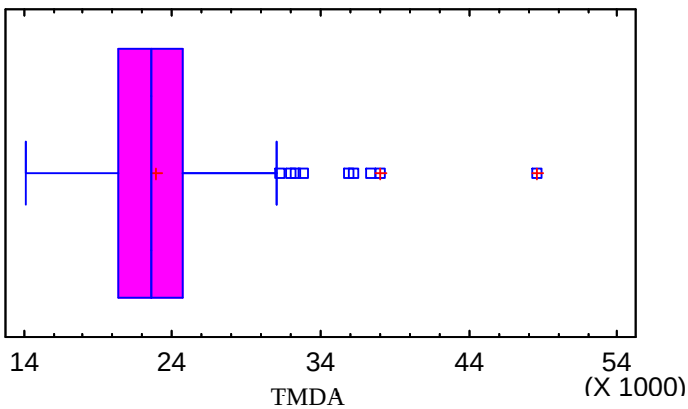

Fig. 4.24. Gráfico de caja y bigotes de *TMDA* por metodología desarrollada, en tercer caso de validación

4.1.3.3. Evaluación de resultados

Con estos resultados podemos confeccionar la Tabla 4.10.

PARÁMETRO	**POR METODOLOGIA CLASICA**	**POR METODOLOGIA DESARROLLADA**
MEDIA	28914	22866
DESVIACION TIPICA	8148	3620
RANGO	50901	34489
ASIMETRIA TIPIFICADA	7,24	3,45
CURTOSIS TIPIFICADA	0,75	2,37

Tabla 4.10. Resumen de resultados para el tercer caso

Como vemos mediante el empleo de la metodología clásica, con los detalles particulares ya señalados para este caso, hemos obtenido un *TMDA* medio que difiere en un 16,7 % del real, y con una distribución con asimetría tipificada de 7,2, lo que se suma a que la dispersión de los resultados es aun mayor a la de los propios datos de tránsitos diarios, ya que aproximadamente el 70 % de los datos dan como resultado un valor que difiere en ±28,2 % de la media. Por su parte el *TMDA* medio obtenido por la metodología desarrollada difiere sólo en 7,7 % del real, y con una distribución que se acerca mucho más a la normal, pudiéndose afirmar que aproximadamente un 70 % de los datos dan como resultado un valor que difiere en ±17,1 % de la media.

Por lo expuesto podemos deducir, que si se calcula el *TMDA* con la metodología clásica, siempre para este caso y con las salvedades ya enunciadas a su favor, existen menores probabilidades de aproximarse al valor real que con la aplicación de la metodología desarrollada.

4.1.4. Cuarto caso de validación

En este punto analizamos la Autopista Buenos Aires – La Plata en su tramo por la localidad de Dock Sud, en el partido de Avellaneda, que podemos observar en la Figura 4.25 y 4.26.

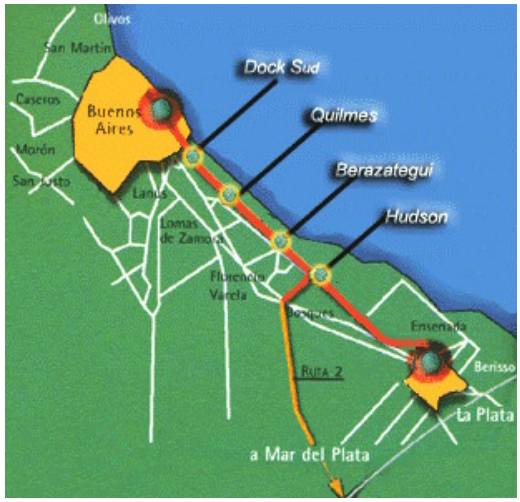

Fig. 4.25. Autopista Buenos Aires – La Plata

Fig. 4.26. Ubicación del tramo urbano en análisis, en cuarto caso de validación

Para esta vía contamos con datos de tránsito diario suministrado por un profesional particular que efectuó estudios sobre la vía durante una semana completa en el mes de octubre de 1999, los que se observan en la Figura 4.27.

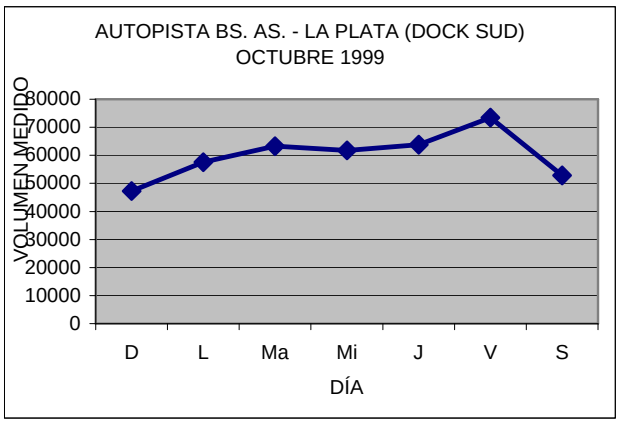

Fig. 4.27. Tránsito diario medido, en cuarto caso de validación

Además hemos recabado la información denunciada por la empresa COVIARES S.A., concesionaria de esta vía, de que en el tramo en estudio el $TMDA_{real}$ registrado durante 1999 asciende a 59045 veh/día, presentando un incremento del tránsito del 2,1 % en el ciclo.

Simultáneamente se cuenta con los factores de corrección obtenidos en 1998, ciclo inmediatamente anterior al analizado, sobre la misma vía pero en la localidad de Hudson, punto ubicado a pocos kilómetros del sector en estudio. Estos factores se observan en la Tabla 4.11.

PUESTO DE PEAJE HUDSON (1998)

COEFICIENTES MENSUALES											
ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
0,877	0,923	0,952	0,961	1,034	1,070	1,066	1,058	1,044	1,017	1,059	0,986

COEFICIENTES DIARIOS						
DOM	LUN	MAR	MIE	JUE	VIE	SAB
1,269	1,042	0,948	0,972	0,941	0,816	1,136

Tabla 4.11. Coeficientes para metodología clásica, en cuarto caso de validación

Con los valores disponibles procedemos a los cálculos del *TMDA*.

4.1.4.1. Determinación del *TMDA* sin considerar estacionalidades

Como primer paso podemos considerar la posibilidad de tomar a las mediciones de tránsito directamente como valores típicos (lo cual es comúnmente efectuado por profesionales no relacionados con la temática) y calcular su media sin considerar las estacionalidades. La estadística obtenida por este medio resulta:

Frecuencia = 7

Media = 59965,6

Varianza = 7,14682E7

Desviación típica = 8453,88

Mínimo = 47266,0

Máximo = 73468,0

Rango = 26202,0

Asimetría tipificada = 0,0523682

Curtosis tipificada = 0,155275

Con la cual se puede construir el gráfico de distribución de la Figura 4.28.

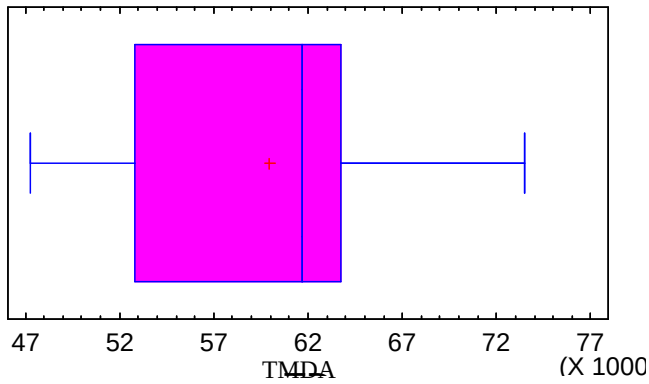

Fig. 4.28. Gráfico de caja y bigotes de *TMDA* directo, en cuarto caso de validación

4.1.4.2. Determinación del *TMDA* mediante la metodología clásica

Para esto empleamos los datos disponibles para la localidad de Hudson, considerando que el crecimiento de tránsito ha sido determinado por otros medios, resultando el 2,1 % producido en la realidad. Recordemos que este término debe incluirse en el cálculo por ser los coeficientes declarados obtenidos mediante el descuento del incremento del tránsito en el año de su medición.

Mediante esta metodología obtenemos los resultados que se presentan en la Figura 4.29.

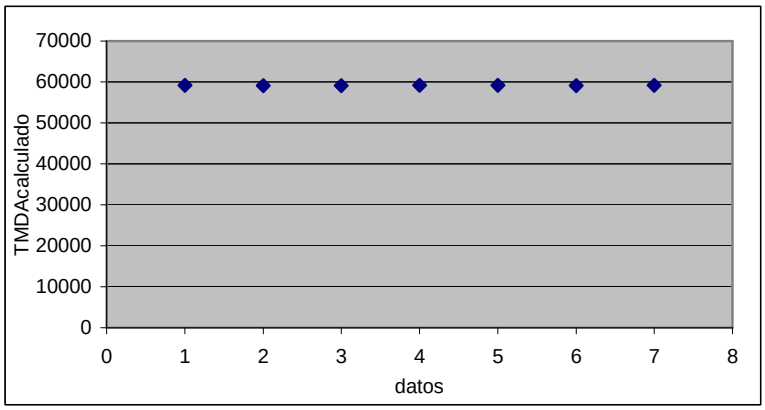

Fig. 4.29. *TMDA* por metodología clásica, en cuarto caso de validación

De los cuales puede obtenerse la siguiente estadística:

Frecuencia = 7

Media = 59154,9

Varianza = 362,476

Desviación típica = 19,0388

Mínimo = 59125,0

Máximo = 59180,0

Rango = 55,0

Asimetría tipificada = -0,485968

Curtosis tipificada = -0,355497

Que nos permite construir el gráfico de la Figura 4.30.

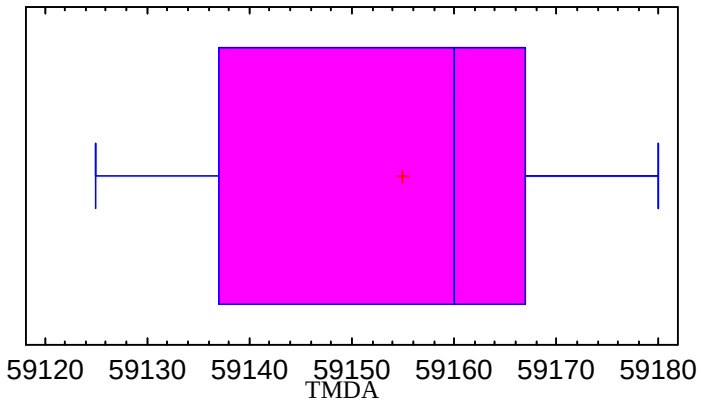

Fig. 4.30. Gráfico de caja y bigotes de *TMDA* por metodología clásica, en cuarto caso de validación

4.1.4.3. Determinación del *TMDA* mediante la metodología desarrollada

Para posibilitar el empleo de los modelos desarrollados debemos considerar que la vía en análisis se ubica en un sector urbano, posee peaje y, aunque presenta una componente muy fuerte de vehículos con viajes turísticos entre Buenos Aires y la Costa Atlántica, sirve mayoritariamente a viajes comerciales origen-destino entre Buenos Aires (y zona de influencia) y La Plata (y zona de influencia). Estas condiciones de borde llevan al empleo de los coeficientes de la Tabla 4.12, obtenidos mediante los modelos desarrollados.

COEFICIENTES POR MODELO DESARROLADOS

COEFICIENTES MENSUALES											
ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
0,997	0,998	1,002	1,009	1,015	1,021	1,023	1,020	1,012	0,995	0,969	0,933

COEFICIENTES DIARIOS						
DOM	LUN	MAR	MIE	JUE	VIE	SAB
1,336	1,151	0,969	0,937	0,924	0,845	1,005

Tabla 4.12. Coeficientes para metodología desarrollada, en cuarto caso de validación

Nos resta ahora calcular el incremento del tránsito producido durante el ciclo, pero como sólo contamos con registros de los vehículos en el partido de Avellaneda para el año 1999, sin obtenerse datos para 1998, no podemos emplear el algoritmo desarrollado para este término. Por tal razón, y según lo recomienda nuestra metodología, debemos obtener el incremento del tránsito de una fuente externa. En forma análoga a la que empleamos con la metodología clásica consideramos que este valor asciende al 2,1 %.

Al aplicar la metodología obtenemos los resultados que se observan en la Figura 4.31.

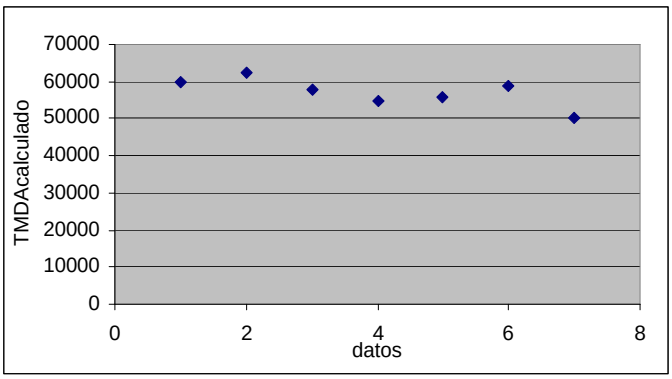

Fig. 4.31. *TMDA* por metodología desarrollada en cuarto caso de validación

Con estos resultados podemos generar la siguiente estadística:

Frecuencia = 7

Media = 57050,4

Varianza = 1,59875E7

Desviación típica = 3998,43

Mínimo = 50179,0

Máximo = 62600,0

Rango = 12421,0

Asimetría tipificada = -0,578556

Curtosis tipificada = 0,307131

La cual nos permite confeccionar el gráfico de la Figura 4.32.

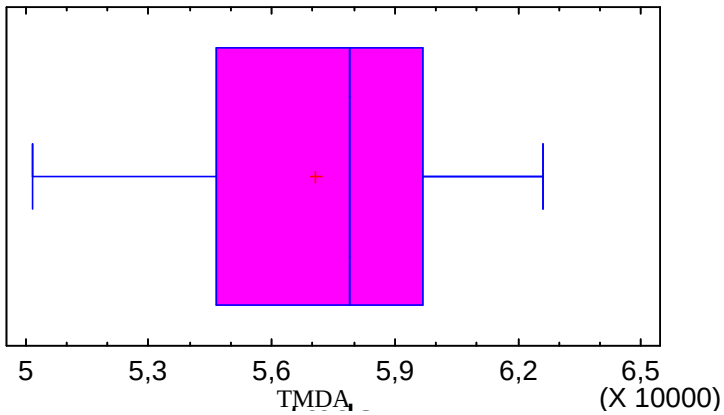

Fig. 4.32. Gráfico de caja y bigotes de *TMDA* por metodología desarrollada, en cuarto caso de validación

4.1.4.4. Evaluación de resultados

Con estos resultados puede confeccionarse la Tabla 4.13.

PARÁMETRO	**SIN CONSIDERAR ESTACIONALIDAD**	**POR METODOLOGIA CLÁSICA**	**POR METODOLOGIA DESARROLLADA**
MEDIA	59966	59155	57050
DESVIACION TIPICA	8454	19	3998
RANGO	26202	55	12421
ASIMETRÍA TIPIFICADA	0,05	-0,49	-0,58
CURTOSIS TIPIFICADA	0,15	-0,35	0,31

Tabla 4.13. Resumen de resultados para el cuarto caso

Observamos, si tomamos como cierto el *TMDA* denunciado por la empresa concesionaria en el punto en estudio de 59045 veh/día, que para este caso en particular se cumple:

- La obtención directa del *TMDA* por el promedio de los valores medidos sobre la vía, y sin considerar estacionalidades e incrementos de tránsito, difiere sólo un 1,6 % del valor real, con una dispersión que implica que aproximadamente el 70 % de los datos se encuentra a ±14,1 % de la media.
- La obtención del *TMDA* por el método clásico arroja valores que difieren sólo un 0,2 % del valor real, con una dispersión que implica que aproximadamente el 70 % de los datos se encuentra a ±0,03 % de la media.
- La obtención del *TMDA* por el modelo desarrollado, difiere sólo un 3,4 % del valor real, con una dispersión que implica que aproximadamente el 70 % de los datos se encuentra a ±7,0 % de la media.

De todo esto podemos concluir que, en este caso en particular, el calcular el *TMDA* directamente con la media de los valores medidos sobre la vía puede darnos resultados ajustados, pero con cierta dispersión, aunque es de esperarse que en meses con tránsitos más alejados de la media los resultados obtenidos posean mucho menor justeza y confiabilidad. Para el caso de empleo de la metodología clásica se obtienen resultados muy buenos, ratificando la elección de la serie empleada, pero sin dejarse de lado la necesidad de recopilación de antecedentes y de la interpretación de la validez de estos por parte de un profesional relacionado con la temática. Finalmente, para el caso de empleo del modelo desarrollado observamos que, si bien los resultados no llegan al nivel de aproximación de la metodología clásica, estos se acercan mucho a los valores reales, con una dispersión admisible y nuevamente sin la necesidad de recolectar series históricas de puntos cercanos, aunque cabe recordar que en esta oportunidad no pudo emplearse el algoritmo para obtención de la tasa de crecimiento del tránsito.

4.2. Discusión de la metodología de estudio empleada

Como dijimos oportunamente, la obtención de los coeficientes por regresión matemática es una de las formas posibles de trabajo, y es la que hemos elegido para este estudio. Pero seguramente pueden emplearse otras herramientas, las cuales no dejan de ser otras líneas de trabajo tan valederas como la nuestra.

Seguramente la forma alternativa más simple de obtención de los coeficientes es la del cálculo de cada uno de ellos como media de los coeficientes relevados.
A continuación analizamos los resultados que se obtienen mediante esta metodología y los comparamos con los obtenidos por regresión, incluyendo algunas observaciones adicionales que nos parecen de interés.

4.2.1. Obtención de los coeficientes por valores medios

Para la obtención de los coeficientes mediante los valores medios necesitamos previamente dividir las nubes de puntos en tantas categorías como sean necesarias. Esta división podría realizarse en forma visual, determinando las subdivisiones de las nubes de puntos en función de la dispersión de valores observadas. Pero para que el análisis comparativo entre metodologías sea valedero, decidimos tomar en este punto las divisiones y clasificaciones obtenidas mediante las regresiones ya efectuadas. Cabe recordar que cuando realizamos el estudio por medio de regresiones, pudimos descartar algunos datos atípicos en función de sus residuos, en cambio en el caso del cálculo de coeficientes por medio de los valores medios, esto deberíamos efectuarlo observando por ejemplo los gráficos de caja y bigotes que se obtienen, eliminando datos outliers y volviendo a calcular la estadística. Esta tarea resulta laboriosa y escapa a la finalidad del estudio, razón por la cual decidimos realizar directamente los cálculos de las medias con todos los valores disponibles pero considerando esta particularidad en la metodología empleada a la hora de las conclusiones.

Según las clasificaciones establecidas tenemos nubes de puntos para:

- Coeficientes diarios de vías de uso turístico
- Coeficientes diarios de vías de uso comercial y sin peaje
- Coeficientes diarios de vías de uso comercial y con peaje
- Coeficientes mensuales de vías de uso turístico, en ambiente rural y con peaje
- Coeficientes mensuales de vías de uso turístico, en ambiente urbano y sin peaje
- Coeficientes mensuales de vías de uso turístico, en ambiente urbano y con peaje
- Coeficientes mensuales de vías de uso comercial, en ambiente rural y sin peaje
- Coeficientes mensuales de vías de uso comercial, en ambiente rural y con peaje
- Coeficientes mensuales de vías de uso comercial, en ambiente urbano y sin peaje

- Coeficientes mensuales de vías de uso comercial, en ambiente urbano y con peaje

Pasemos a la determinación de los valores medios en cada uno de los casos de cada uno de los coeficientes, estableciendo simultáneamente sus intervalos de confianza del 95 %. Para esto último debemos emplear la fórmula:

$$\overline{X} \pm 1{,}96 \cdot \frac{S}{\sqrt{n}} \quad (4.1)$$

Donde:

$\overline{X}$ = media aritmética

S = desvío estándar

n = número de muestras

4.2.2. Análisis comparativo para los coeficientes diarios

Los coeficientes diarios que obtenemos como media y los valores para un intervalo de confianza del 95 % se resumen en la Tabla 4.14.

USO	PEAJE	COEFICIENTE DIARIO													
		DOM		LUN		MAR		MIE		JUE		VIE		SAB	
		COEF	+/-	COEF	+/-	COEF	+/-	COEF	+/-	COEF	+/-	COEF	+/-	COEF	+/-
turístico	*c/s peaje*	0,827	**0,121**	1,081	**0,093**	1,190	**0,011**	1,189	**0,012**	1,151	**0,084**	1,025	**0,019**	0,888	**0,04**
comercial	*sin peaje*	1,121	**0,072**	1,000	**0,035**	0,999	**0,025**	0,984	**0,043**	0,960	**0,036**	0,891	**0,051**	1,043	**0,06**
comercial	*con peaje*	1,335	**0,109**	1,154	**0,108**	0,960	**0,037**	0,949	**0,034**	0,914	**0,040**	0,847	**0,065**	1,002	**0,1**

Tabla 4.14. Coeficientes diarios e intervalos de confianza obtenidos por valores medios

La estadística de estos últimos valores para el intervalo de confianza del 95 % resulta:

Frecuencia = 21

Media = 0,0568571

Varianza = 0,00116343

Desviación típica = 0,0341091

Mínimo = 0,011

Máximo = 0,121

Rango = 0,11

Asimetría tipificada = 1,0122

Curtosis tipificada = -0,899133

Y su gráfica de distribución es la que se observa en la Figura 4.33.

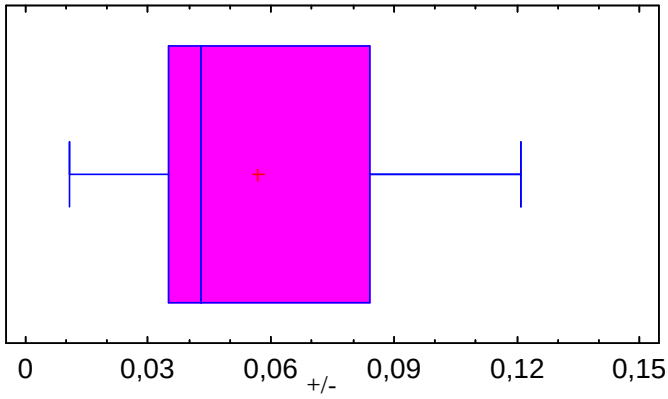

Fig. 4.33. Gráfico de caja y bigotes para los intervalos de confianza de los coeficientes diarios por valores medios

Este mismo análisis extendido a los coeficientes diarios obtenidos por regresión nos lleva a la Tabla 4.15.

USO	PEAJE	COEFICIENTE DIARIO													
		DOM		LUN		MAR		MIE		JUE		VIE		SAB	
		COEF	+/-	COEF	+/-	COEF	+/-	COEF	+/-	COEF	+/-	COEF	+/-	COEF	+/-
turístico	*c/s peaje*	0,772	**0,034**	1,004	**0,022**	1,149	**0,024**	1,207	**0,026**	1,177	**0,024**	1,059	**0,024**	0,855	**0,04**
comercial	*sin peaje*	1,095	**0,031**	1,001	**0,024**	1,000	**0,021**	1,008	**0,022**	0,955	**0,022**	0,866	**0,031**	1,061	**0,03**
comercial	*con peaje*	1,336	**0,057**	1,151	**0,056**	0,969	**0,050**	0,937	**0,043**	0,924	**0,050**	0,845	**0,056**	1,005	**0,06**

Tabla 4.15. Coeficientes diarios e intervalos de confianza obtenidos por regresión

Con la correspondiente estadística para los valores del intervalo de confianza del 95 %:

Frecuencia = 21

Media = 0,0354286

Varianza = 0,000184657

Desviación típica = 0,0135889

Mínimo = 0,021

Máximo = 0,057

Rango = 0,036

Asimetría tipificada = 1,10476

Curtosis tipificada = -1,23832

Estadística que nos lleva al gráfico de la Figura 4.34.

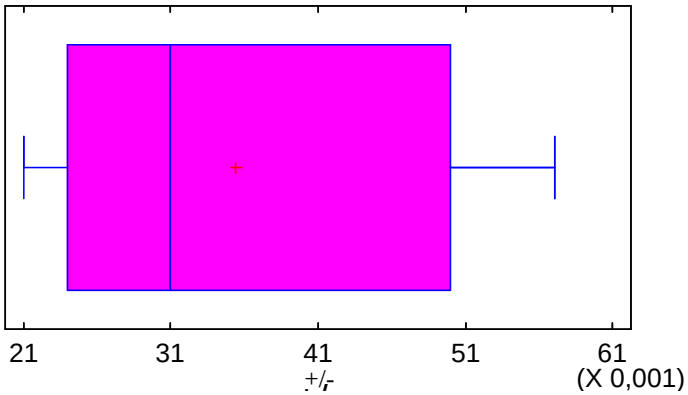

Fig. 4.34. Gráfico de caja y bigotes para los intervalos de confianza de los coeficientes diarios por regresión

Como podemos observar, aunque para algunos puntos presenta mejores valores una metodología y para el resto la otra, puede deducirse en forma general que para los coeficientes obtenidos por cálculo de las medias tenemos un valor promedio para el intervalo de confianza del 95 % de ±0,057 y un desvío estándar de 0,034. Por su parte, los valores para los coeficientes obtenidos por regresión presentan un promedio de ±0,035 y un desvío estándar de 0,013. Resulta claro entonces que, para la muestra de datos considerada, es más efectiva la obtención de los coeficientes por regresión matemática que por cálculo de los valores medios, recordando que en este caso pudieron descartarse puntos por su residuo y en el otro no.

4.2.2. Análisis comparativo para los coeficientes mensuales

Los coeficientes mensuales obtenidos como media y los valores para un intervalo de confianza del 95 % correspondientes para la muestra analizada se resumen en la Tabla 4.16.

							comercial							
	rural		urbano				rural				urbano			
	c/peaje		s/peaje		c/peaje		s/peaje		c/peaje		s/peaje		c/peaje	
	COEF	+/-	COEF	+/-	COEF	+/-	COEF	+/-	COEF	+/-	COEF	+/-	COEF	+/-
ENE	0,669	0,068	1,099	0,180	1,011	0,027	0,717	0,047	0,551	0,033	1,066	0,066	0,991	0,062
FEB	0,753	0,023	1,082	0,171	0,986	0,026	0,894	0,045	0,762	0,036	1,083	0,043	0,982	0,046
MAR	0,984	0,028	1,048	0,062	1,007	0,017	1,007	0,018	0,937	0,044	1,000	0,020	1,024	0,046
ABR	1,008	0,033	0,970	0,058	1,012	0,024	0,993	0,012	1,123	0,072	0,999	0,020	1,020	0,021
MAY	1,047	0,013	0,935	0,082	0,999	0,025	1,076	0,026	1,290	0,077	0,998	0,023	1,043	0,044
JUN	1,066	0,027	1,007	0,088	1,043	0,021	1,105	0,029	1,335	0,073	1,008	0,015	1,033	0,043
JUL	1,035	0,022	0,940	0,100	0,962	0,023	1,042	0,017	1,331	0,050	0,976	0,023	1,050	0,040
AGO	1,023	0,032	1,023	0,047	0,997	0,017	1,058	0,022	1,291	0,049	0,990	0,022	1,020	0,028
SEP	1,033	0,015	1,055	0,116	1,038	0,015	1,084	0,022	1,338	0,082	0,994	0,026	0,982	0,025
OCT	1,003	0,026	0,944	0,069	1,026	0,018	1,042	0,018	1,202	0,058	0,969	0,025	0,999	0,025
NOV	0,892	0,051	1,020	0,069	1,020	0,020	0,879	0,025	1,207	0,040	0,975	0,033	0,987	0,027
DIC	0,717	0,057	0,931	0,072	0,971	0,019	0,754	0,022	0,988	0,034	0,956	0,032	0,952	0,051

Tabla 4.16. Coeficientes mensuales e intervalos de confianza obtenidos por valores medios

Con la siguiente estadística para los valores para un intervalo de confianza del 95 %:

Frecuencia = 84

Media = 0,041881

Varianza = 0,000926106

Desviación típica = 0,030432

Mínimo = 0,012

Máximo = 0,18

Rango = 0,168

Asimetría tipificada = 8,94198

Curtosis tipificada = 14,1965

La que nos permite confeccionar el gráfico de distribución de la Figura 4.35.

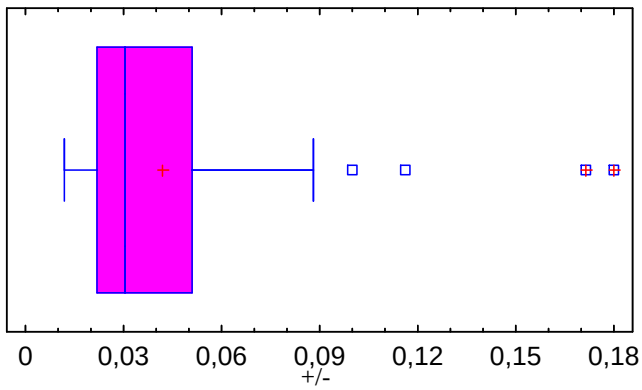

Fig. 4.35. Gráfico de caja y bigotes para los intervalos de confianza de los coeficientes mensuales por valores medios

Por su parte, para un intervalo de confianza del 95 % en los coeficientes mensuales obtenidos mediante los modelos de regresión desarrollados, tenemos los valores de la Tabla 4.17.

							comercial							
	rural		urbano				rural				urbano			
	c/peaje		s/peaje		c/peaje		s/peaje		c/peaje		s/peaje		c/peaje	
	COEF	+/-	COEF	+/-	COEF	+/-	COEF	+/-	COEF	+/-	COEF	+/-	COEF	+/-
ENE	0,650	0,025	0,991	0,013	0,995	0,013	0,699	0,042	0,578	0,025	1,044	0,024	0,997	0,022
FEB	0,798	0,032	0,987	0,009	0,993	0,009	0,836	0,033	0,769	0,018	1,032	0,019	0,998	0,018
MAR	0,922	0,026	0,990	0,010	0,997	0,009	0,949	0,027	0,935	0,016	1,024	0,016	1,002	0,015
ABR	1,021	0,023	0,997	0,010	1,003	0,009	1,037	0,024	1,074	0,015	1,020	0,015	1,009	0,014
MAY	1,092	0,021	1,006	0,010	1,011	0,009	1,098	0,022	1,184	0,015	1,018	0,014	1,015	0,013
JUN	1,134	0,021	1,018	0,010	1,019	0,008	1,130	0,022	1,264	0,016	1,016	0,013	1,021	0,012
JUL	1,146	0,022	1,029	0,009	1,025	0,008	1,131	0,022	1,313	0,016	1,012	0,013	1,023	0,012
AGO	1,125	0,024	1,038	0,009	1,028	0,009	1,101	0,022	1,327	0,015	1,005	0,014	1,020	0,012
SEP	1,071	0,026	1,044	0,010	1,026	0,009	1,037	0,024	1,307	0,015	0,994	0,015	1,012	0,014
OCT	0,982	0,030	1,045	0,010	1,017	0,009	0,937	0,027	1,250	0,015	0,976	0,017	0,995	0,015
NOV	0,855	0,036	1,039	0,012	1,000	0,011	0,801	0,031	1,154	0,018	0,950	0,020	0,969	0,018
DIC	0,690	0,044	1,025	0,017	0,974	0,015	0,627	0,038	1,019	0,025	0,914	0,024	0,933	0,022

Tabla 4.17. Coeficientes mensuales e intervalos de confianza obtenidos por regresión

Con estos valores construimos la siguiente estadística:

Frecuencia = 84

Media = 0,0179881

Varianza = 0,0000647107

Desviación típica = 0,0080443

Mínimo = 0,008

Máximo = 0,044

Rango = 0,036

Asimetría tipificada = 4,07555

Curtosis tipificada = 2,01718

Que nos permite confeccionar el gráfico de distribución de la Figura 4.36.

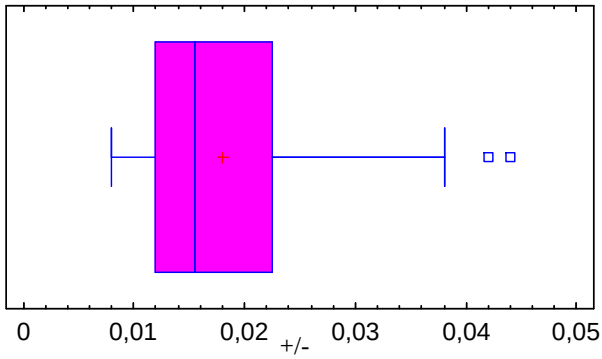

Fig. 4.36. Gráfico de caja y bigotes para los intervalos de confianza de los coeficientes mensuales por regresión

Nuevamente podemos observar que aunque para algunos puntos presenta mejores valores una metodología y para el resto la otra, puede asegurarse en forma general que los coeficientes obtenidos por cálculo de las medias conllevan un valor promedio para el intervalo de confianza del 95 % de ±0,042 y un desvío estándar de 0,030. Por su parte, los valores para los coeficientes obtenidos por regresión presentan un promedio de ±0,018 y un desvío estándar de 0,008. Resulta claro entonces que, para la muestra de datos considerada, resulta más efectiva la obtención de los coeficientes por regresión matemática, recordando que para estos la realización del estudio permitió el descarte de algunos datos atípicos.

Todo lo expuesto nos lleva a considerar que, si bien no es posible realizar una comparación directa entre ambas metodologías de obtención de los coeficientes, por no haberse considerado la eliminación de los datos outliers en el cálculo de valores medios, cuando se comparan sus intervalos de confianza al 95 %, para los datos con los cuales son obtenidos en cada caso, el modelo de regresión matemática resulta más adecuado que el de cálculo por las medias. Por lo tanto es razonable pensar que mediante la técnica de regresión pueden obtenerse coeficientes de corrección al menos tan confiables como cuando se obtienen por valores medios.

Capítulo 5 – Conclusiones y recomendaciones

5.1. Conclusiones

Las siguientes conclusiones se desprenden de los análisis realizados en cada uno de los capítulos de este trabajo y tienen validez para el área en estudio aquí establecida y en función de los datos relevados.

5.1.1. Respecto a la problemática detectada y marco teórico para su resolución

- La obtención de valores de *TMDA* confiables, como dato fundamental para la aplicación de técnicas resolutivas en problemáticas viales, es una falencia latente a nivel nacional y regional, sobre todo cuando se trata de analizar puntos fuera de grandes ciudades o de la red primaria de carreteras.
- La metodología clásica para su cálculo mediante conteos esporádicos implica la intervención de profesionales con formación afín cuando se desea alcanzar esta confiabilidad, los cuales no siempre están disponibles ni son justificables.
- Es necesario por lo tanto contar con metodologías objetivas alternativas, fundamentadas en las series históricas disponibles, y aplicables en zonas relativamente homogéneas en función de sus características socioeconómicas.
- La obtención de los modelos incluidos en estas metodologías puede generarse por medio de técnicas de regresión que combinen los datos de tránsito con estas variables socioeconómicas.

5.1.2. Respecto a la obtención de datos

- Las series de tránsito existentes en el área en estudio son abundantes, pero se encuentran expresadas en forma muy variada, lo que hace necesario su adaptación para la aplicación de técnicas comparativas.
- Los datos socioeconómicos también son abundantes, pero para el nivel de desagregación establecido en el estudio se reduce sustancialmente la disponibilidad de datos, resultando acotadas las variables empleables.
- La consulta a instituciones y profesionales particulares, para ambas tipologías de datos, permite generar una base de datos voluminosa para el análisis de la región conformada por las provincias de Buenos Aires, Córdoba, Santa Fe, Entre Ríos y La Pampa, y para el periodo de estudio que va desde 1993 a 2003.

5.1.3. Respecto al empleo de los datos

- Es posible incluir una nueva forma de conceptualización del tránsito como una serie de tiempo, que se ajusta más a las técnicas de análisis estadísticos que la forma clásica comúnmente aceptada y que se adapta mejor a las pronunciadas fluctuaciones características de nuestra economía nacional.
- El empleo de adecuadas clasificaciones y divisiones de los datos, permite el desarrollo de modelos que se ajusten a los umbrales de confiabilidad comúnmente exigidos en estudios de estas características.
- La variación del parque automotor año a año, en las diversas localidades que componen el área en estudio, posee una relación estadísticamente acorde a las variaciones en los volúmenes de tránsito de las vías que cruzan dichas localidades.
- Las variaciones de los volúmenes de tránsito durante la semana en vías turísticas resultan relativamente homogéneas. En cambio las variaciones registradas en vías de uso comercial predominante requieren la consideración adicional de la existencia de cobro de peaje sobre las mismas para la obtención grupos de datos homogéneos.
- Las variaciones de los volúmenes de tránsito mensuales a lo largo del año sobre una vía, pueden explicarse por medio de la consideración de su entorno, de la finalidad de su empleo y de la existencia o no sobre la misma de cobro

de peaje, salvo en el caso de las vías turísticas rurales sin peajes, para las cuales los modelos generados no tienen validez, por haberse carecido de datos sobre éstas en el desarrollo de los mismos. Por esto, en caso de analizarse vías de esta tipología deberán emplearse metodologías alternativas a la desarrollada.

- Para las variaciones mensuales resulta superflua la consideración de la clasificación del tránsito, ya que ésta se ve explicada por el uso de la vía y si ésta se encuentra en ambiente rural o urbano.

5.1.4. Respecto a la validación de la metodología desarrollada

- De acuerdo a los casos analizados puede asegurarse que la aplicación de la metodología desarrollada, salvo para la tipología de vía en que no es utilizable, da como resultados valores tan o más confiables que los obtenidos mediante la metodología clásica, guardando además la mayor sencillez y objetividad en su empleo establecidas como parámetros de diseño.
- Estos resultados además se ajustan adecuadamente a los valores reales, lo cual ratifica la forma general elegida para el modelo. Esto implica, en función del análisis, que no es necesario la inclusión de un factor adicional que distinga de la ubicación de la semana dentro del mes. De todos modos, en caso de que las condiciones particulares de la vía indiquen la conveniencia de inclusión de tal parámetro, éste debe ser independiente del mes, por lo que bastará con analizar que valor debe tomar el mismo en función de las condiciones de entorno de la vía, por ejemplo.

5.1.5. Respecto a la discusión por la metodología de estudio

- La metodología de análisis estadístico empleada para el desarrollo de los modelos es al menos tan buena, sino más, que su principal técnica estadística alternativa. Es decir, que los coeficientes de corrección para el cálculo del *TMDA* obtenidos por regresión son al menos tan confiables como los obtenidos por cálculo de valores medios de los coeficientes.

Todo lo expuesto lleva a la conclusión final de que si la metodología desarrollada se aplica en forma coherente, en vías ubicadas dentro del área en estudio, para los casos en los cuales los modelos tienen validez, los resultados de *TMDA* obtenidos poseen un buen nivel de confiabilidad.

5.2. Recomendaciones

Las recomendaciones que surgen del estudio son:

- Tender a un procedimiento único para que todas las instituciones que de alguna manera se encuentren relacionadas con el levantamiento de datos de tránsito presenten valores comparables, tanto en lo que hace a su volumen, clasificación e identificaciones de sus condiciones de entorno.
- Adoptar la forma conceptual propuesta para la consideración del tránsito con un crecimiento del mismo producido en forma proporcional a lo largo del año, como una manera de acercarse más al análisis estadístico de las series de tiempo. Esto permite además la obtención de series sobre un mismo punto comparables año a año, con mayor grado de independencia de las condiciones económicas coyunturales.
- Aplicar la metodología de análisis en otras zonas homogéneas del país, para tender a la obtención de modelos o cuadros que permitan la cobertura total del territorio nacional.
- Continuar con el empleo experimental de la metodología desarrollada en forma conjunta con la metodología clásica para confirmar la validación de los modelos o detectar la necesidad de ajustes a los mismos.

Anexo A

a.1. Reseña teórica 1

- *Método de máxima verosimilitud.*

Conocida una muestra de tamaño *n*, $\{(x_i, y_i) : i = 1, \ldots, n\}$, de la hipótesis de normalidad se sigue que la densidad condicionada en y_i es

$$f(y_i/x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i - (\alpha_0 + \alpha_1 x_i))^2}{\sigma^2}\right), \quad i = 1, \ldots, n, \tag{a.1}$$

y, por tanto, la función de densidad conjunta de la muestra es,

$$f\left(\vec{\mathbf{Y}}/\alpha_0, \alpha_1, \sigma^2\right) = \prod_{i=1}^{n} f(y_i/x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \alpha_0 - \alpha_1 x_i)^2\right). \tag{a.2}$$

Una vez tomada la muestra y, por tanto, que se conocen los valores de $\{(x_i, y_i)\}_{i=1}^{n}$, se define la función de verosimilitud asociada a la muestra como sigue

$$l\left(\alpha_0, \alpha_1, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \alpha_0 - \alpha_1 x_i)^2\right), \tag{a.3}$$

esta función (con variables α_0, α_1 y σ^2) mide la *verosimilitud* de los posibles valores de estas variables en base a la muestra recogida.

El método de máxima verosimilitud se basa en calcular los valores de α_0, α_1 y σ^2 que maximizan la función y, por tanto, hacen máxima la probabilidad de ocurrencia de la muestra obtenida. Por ser la función de verosimilitud una función creciente, el problema es más sencillo si se toman logaritmos y se maximiza la función resultante, denominada *función soporte*,

$$\begin{aligned} L\left(\alpha_0, \alpha_1, \sigma^2\right) &= \ln l\left(\alpha_0, \alpha_1, \sigma^2\right) = \\ &-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\alpha_0 + \alpha_1 x_i))^2. \end{aligned} \tag{a.4}$$

Maximizando la anterior se obtienen los siguientes *estimadores máximo verosímiles***,**

$$\hat{\alpha}_{0,MV} = \bar{y} - \hat{\alpha}_{1,MV}\,\bar{x}$$

$$\hat{\alpha}_{1,MV} = \frac{s_{XY}}{s_x^2}$$

$$\hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (\hat{\alpha}_{0,MV} + \hat{\alpha}_{1,MV}\, x_i)\right)^2 \qquad \text{(a.5)}$$

donde se ha denotado $\bar{x}$ e $\bar{y}$ a las medias muestrales de *X* e *Y*, respectivamente; s_x^2 es la varianza muestral de *X* y s_{XY} es la covarianza muestral entre *X* e *Y*. Estos valores se calculan de la siguiente forma:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2,$$

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \bar{x}\bar{y}. \qquad \text{(a.6)}$$

- *Método de mínimos cuadrados.*

A partir de los estimadores: $\hat{\alpha}_0$ y $\hat{\alpha}_1$, se pueden calcular las *predicciones* para las observaciones muestrales, dadas por,

$$\hat{Y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 x_i, \qquad i = 1, 2, \ldots, n, \quad \text{(a.7)}$$

o, en forma matricial,

$$\hat{\mathbf{Y}} = \hat{\alpha}_0 \vec{\mathbf{1}} + \hat{\alpha}_1 \vec{\mathbf{X}}, \quad \text{(a.8)}$$

donde $\hat{\mathbf{Y}}^t = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$. Ahora se definen los *residuos* como

$e_i = y_i - \hat{y}_i$, $i = 1,2,...,n$,

Residuo = Valor observado – Valor previsto

en forma matricial,

$$\vec{\mathbf{e}} = \vec{\mathbf{Y}} - \hat{\mathbf{Y}}, \quad \text{con} \quad \vec{\mathbf{e}}^t = (e_1, \ldots, e_n). \quad \text{(a.9)}$$

Los estimadores por mínimos cuadrados se obtienen minimizando la suma de los cuadrados de los residuos, o sea minimizando la siguiente función,

$$\Psi(\alpha_0, \alpha_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - (\alpha_0 + \alpha_1 x_i)\right)^2, \qquad \text{(a.10)}$$

derivando e igualando a cero se obtienen las siguientes ecuaciones, denominadas *ecuaciones canónicas***,**

$$\left\{ \begin{array}{lll} \sum_{i=1}^{n}\left(y_i - (\alpha_0 + \alpha_1 x_i)\right) & = \sum_{i=1}^{n} e_i & = 0 \\ \sum_{i=1}^{n}\left(y_i - (\alpha_0 + \alpha_1 x_i)\right) x_i & = \sum_{i=1}^{n} e_i x_i & = 0 \end{array} \right\} \Rightarrow$$

$$\left\{ \begin{array}{lll} \sum_{i=1}^{n} y_i & = & \hat{\alpha}_0 n + \hat{\alpha}_1 \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i & = & \hat{\alpha}_0 \sum_{i=1}^{n} x_i + \hat{\alpha}_1 \sum_{i=1}^{n} x_i^2 \end{array} \right\} \Rightarrow$$

$$\left\{ \begin{array}{rcl} \bar{y} & = & \hat{\alpha}_0 + \hat{\alpha}_1 \bar{x} \\ \overline{xy} & = & \hat{\alpha}_0 \bar{x} + \hat{\alpha}_1 \overline{x^2} \end{array} \right\} \quad \text{(a.11)}$$

De donde se deducen los siguientes *estimadores mínimo cuadráticos* de los parámetros de la recta de regresión

$$\hat{\alpha}_{0,mc} = \bar{y} - \hat{\alpha}_{1,mc}\, \bar{x}$$
$$\hat{\alpha}_{1,mc} = \frac{s_{XY}}{s_x^2}. \quad \text{(a.12)}$$

Se observa que los estimadores por máxima verosimilitud y los estimadores mínimo cuadráticos de α_0 y α_1 son iguales. Esto es debido a la hipótesis de normalidad, asegurar que $\hat{\alpha}_0 = \hat{\alpha}_{0,MV} = \hat{\alpha}_{0,mc}$ y $\hat{\alpha}_1 = \hat{\alpha}_{1,MV} = \hat{\alpha}_{1,mc}$.

a.2. Reseña teórica 2

Problemas en el ajuste de modelos:

- En la Figura a.1 la nube de puntos muestrales bidimensionales parece ajustarse bien a una recta.

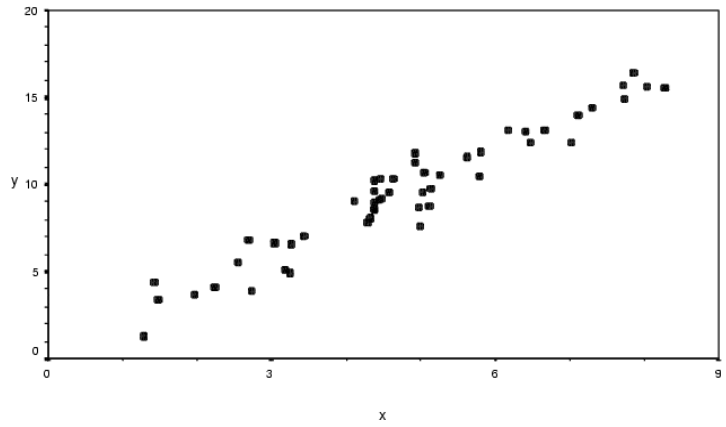

Fig. a.1. Nube de puntos que ajusta bien a la recta

- En la Figura a.2 el ajuste lineal no parece adecuado para esta muestra.

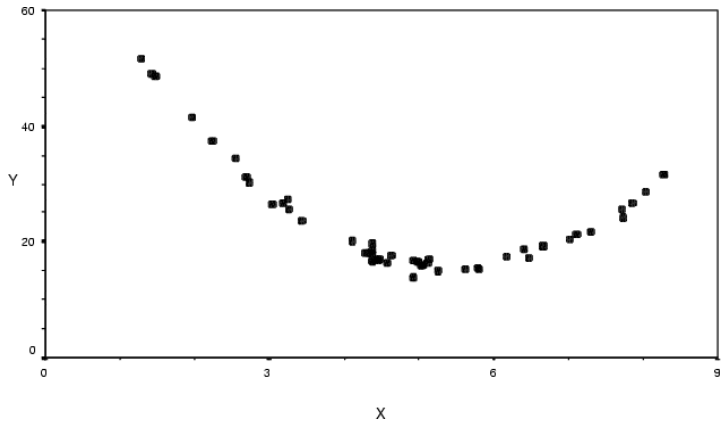

Fig. a.2. Nube de puntos para la cual el ajuste lineal no resulta adecuado

- En la Figura a.3 no existe relación lineal entre las dos variables.

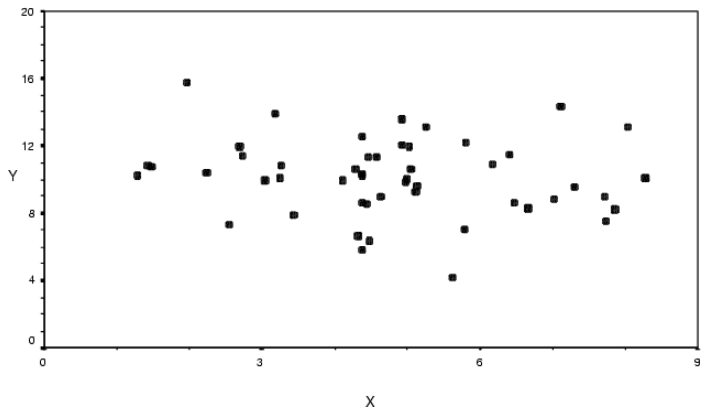

Fig. a.3. Nube de puntos sin relación lineal entre variables

- En la Figura a.4 hay claros indicios de heterocedasticidad.

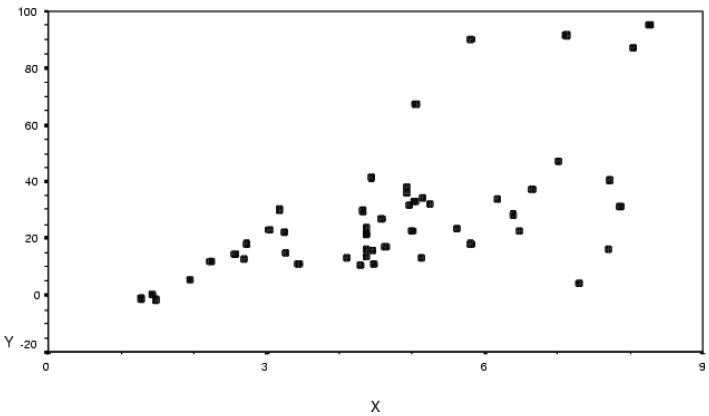

Fig. a.4. Nube de puntos con claros indicios de heterocedasticidad

- En la Figura a.5 existen puntos atípicos que probablemente influyan en la estimación de la recta ajustada.

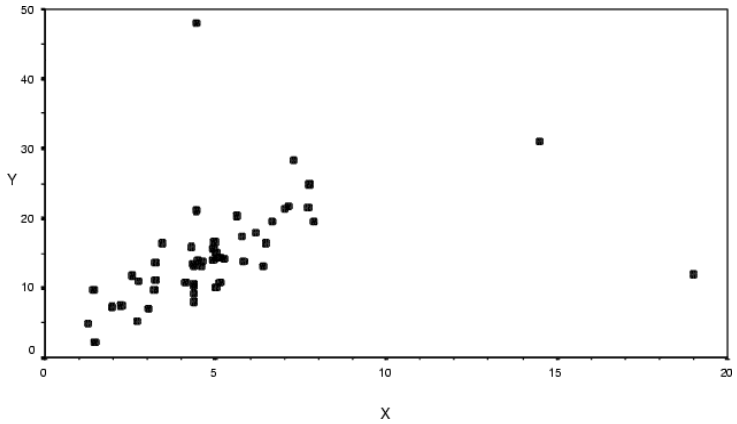

Fig. a.5. Nube de puntos con datos atípicos

- En la Figura a.6 existe una variable regresora binaria que se debe de incluir en el modelo de regresión.

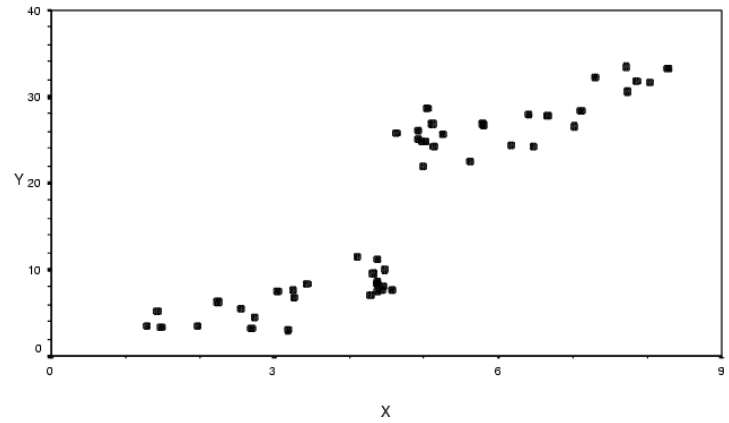

Fig. a.6. Nube de puntos con posibilidad de inclusión de variable binaria

a.3. Reseña teórica 3

Transformaciones para la regresión:

- En la Figura a.7 se observa el modelo $Y = \exp(\alpha_0 + \alpha_1 x)$

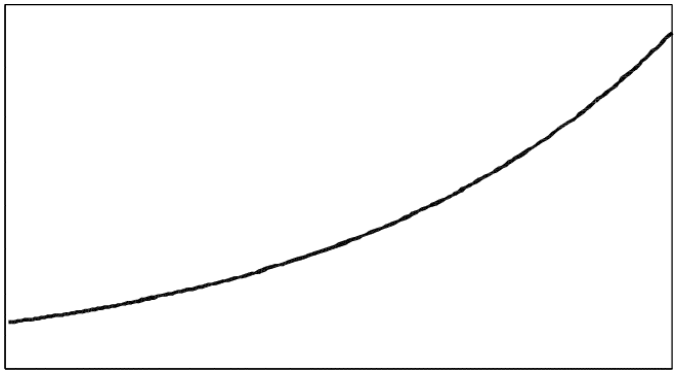

Fig. a.7. Modelo $Y = \exp(\alpha_0 + \alpha_1 x)$

- En la Figura a.8 se observa el modelo $Y = 1/(\alpha_0 + \alpha_1 X)$

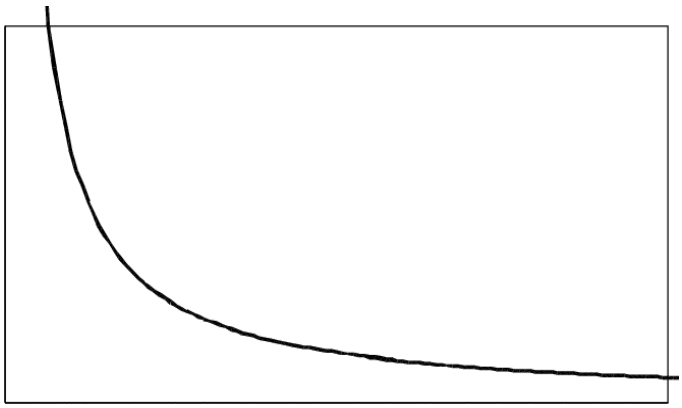

Fig. a.8. Modelo $Y = 1/(\alpha_0 + \alpha_1 X)$

- En la Figura a.9 se observa el modelo $Y = \alpha_0 + \alpha_1 \lg X$

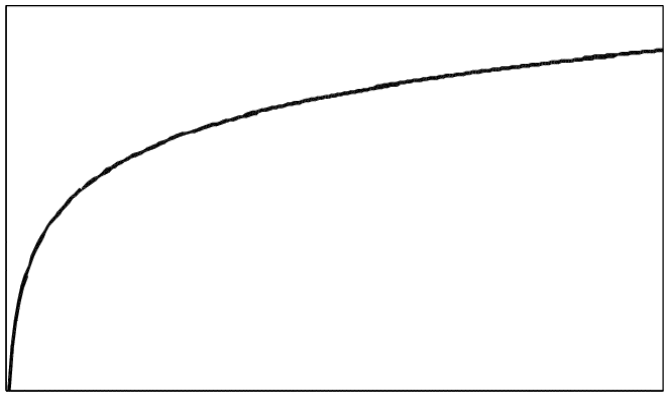

Fig. a.9. Modelo $Y = \alpha_0 + \alpha_1 \lg X$

- En la Figura a.10 se observa el modelo $Y = \alpha_0 X^{\alpha}{}_1$

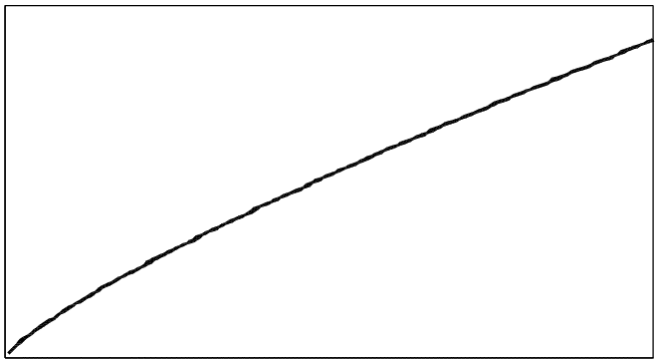

Fig. a.10. Modelo $Y = \alpha_0 X^{\alpha}{}_1$

- En la Figura a.11 se observa el modelo $Y = \alpha_0 X^{-\alpha}{}_1$

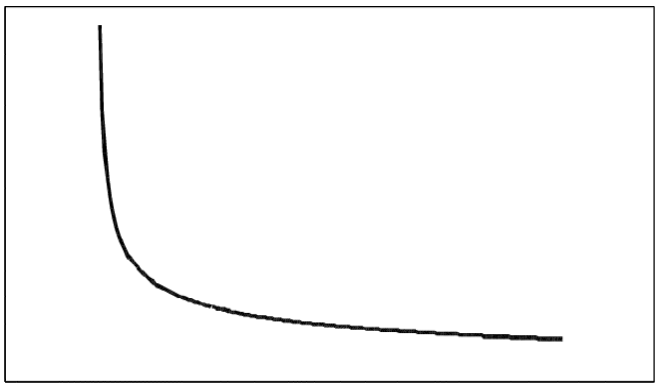

Fig. a.11. Modelo $Y = \alpha_0 X^{-\alpha}{}_1$

- En la Figura a.12 se observa el modelo $Y = \exp X$

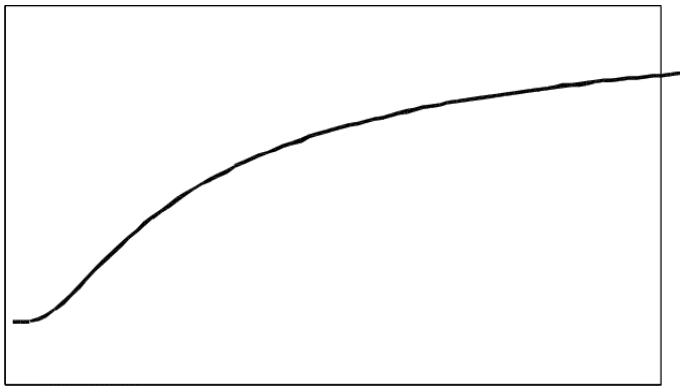

Fig. a.12. Modelo $Y = \exp X$

a.4. Reseña Teórica 4

Análisis para el cumplimiento de las hipótesis básicas en modelos de regresión simple:

- La hipótesis de normalidad: Una hipótesis básica es que los errores del modelo siguen una distribución normal y para ello se debe de contrastar la hipótesis de que los residuos $e_i=y_i - \hat{y}_i$, $i=1,...,n$, provienen de una distribución normal. Pero dado que $Var(e_i) = \sigma^2(e_i) = \sigma^2(1-h_{ii})$, $i=1,...,n$, es preferible trabajar con los residuos estandarizados o estudentizados que tienen la misma varianza (próxima a 1).

Para estudiar la normalidad de los residuos estandarizados se pueden utilizar las técnicas de:

- Gráficos: el gráfico de cajas, el histograma, la estimación no paramétrica de la función de densidad, el gráfico de simetría y los gráfico *p-p* y *q-q*.
- Contrastes de normalidad: contraste de asimetría y curtosis, contraste chi-cuadrado, contraste de Kolmogoroff-Smirnoff-Lilliefors.

Si la hipótesis de normalidad no se verifica, se afecta al modelo en:

- Los estimadores por mínimos-cuadrados de los parámetros del modelo no coinciden con los máximo-verosímiles. Los estimadores obtenidos son centrados pero no eficientes.
- Los contrastes de significación (de la *F* y de la *t*) dejan de ser válidos y los intervalos de confianza obtenidos para los parámetros del modelo no son correctos. A pesar de ello, si el tamaño muestral es razonablemente grande, por el Teorema Central del Límite, los contrastes e intervalos de confianza obtenidos son una buena aproximación de los reales.

Es muy importante conocer la causa por la que no se cumple la hipótesis de normalidad ya que esta información puede ayudar a corregir el modelo de regresión ajustado. Entre otros motivos, la falta de normalidad puede ser debida a un conjunto pequeño de observaciones atípicas que originan apuntamiento o a la existencia de una variable cualitativa oculta que hace que la distribución sea multimodal. En estos casos se puede mejorar el modelo corrigiendo estos problemas. En otras ocasiones la

falta de normalidad es debida a una fuerte asimetría de la distribución que, en muchos casos, va acompañada de otros problemas como falta de linealidad o heterocedasticidad. Entonces lo recomendable es transformar la variable respuesta que normalmente arregla ambos problemas. La familia de transformaciones de Box-Cox es la que normalmente se utiliza.

- La hipótesis de homocedasticidad: Implica que $Var(\varepsilon_i)=\sigma^2=cte$, se detecta fácilmente en el gráfico de residuos (e_{ij}) frente a las predicciones ($\hat{y}_i$) o, equivalentemente, en el gráfico de los residuos (e_{ij}) frente a la variable regresora (x_i).

Un modelo bastante frecuente de heterocedasticidad es el siguiente

$$Y_i = \alpha_0 + \alpha_1 x_i + g(x_i)\,\varepsilon_i, \quad i = 1, \ldots, n, \quad \text{(a.13)}$$

con error $\xi_i = g(x_i)\varepsilon_i$, la varianza del error es

$$Var(\xi_i) = Var(g(x_i)\,\varepsilon_i) = (g(x_i))^2\,\sigma^2, \quad \text{(a.14)}$$

si g no es constante, el modelo es heterocedástico y el caso más frecuente es el siguiente

$$g(x_i) = k x_i^{\nu}. \quad \text{(a.15)}$$

En este caso se puede transformar el modelo para obtener un modelo homocedástico.

Si ν=1, la desviación típica de los errores crece linealmente con la variable regresora, la transformación adecuada es multiplicar todo el modelo por $1/X$, obteniendo

$$\begin{aligned} \frac{Y_i}{x_i} &= \alpha_0 \frac{1}{x_i} + \alpha_1 + \frac{\xi_i}{x_i}, \quad i = 1, \ldots, n. \\ \Rightarrow \quad \frac{Y_i}{x_i} &= \alpha_0 \frac{1}{x_i} + \alpha_1 + \frac{k x_i \varepsilon_i}{x_i}, \quad i = 1, \ldots, n. \end{aligned} \quad \text{(a.16)}$$

Este modelo puede escribirse como un modelo lineal simple homocedástico,

$$\tilde{Y}_i = \tilde{\alpha}_0 + \tilde{\alpha}_1 \tilde{x}_i + \tilde{\varepsilon}_i, \quad i = 1, \ldots, n, \quad \text{(a.17)}$$

donde se denota $\tilde{Y}_i=Y_i x_i$, $\tilde{\alpha}_0=\alpha_1$, $\tilde{\alpha}_1=\alpha_0$, $\tilde{x}_i=1\text{-}x_i$ y errores $\varepsilon_i=k\varepsilon_i$ con varianza $Var(\tilde{\varepsilon}_i)=k^2\sigma^2=cte$.

En algunos casos transformando solamente la variable respuesta se consigue homocedasticidad y se resuelven otros posibles problemas como falta de simetría y de normalidad. Nuevamente, la familia de transformaciones de Box-Cox es útil para este propósito y la sencilla transformación λ=0 (tomar logaritmos en la variable respuesta) es suficiente para obtener homocedasticidad.

Una alternativa para estimar el parámetro λ que se puede utilizar en la transformación de Box-Cox es la siguiente:

- Ordenar las predicciones de menor a mayor ($\hat{y}_{(i)}$).
- Hacer grupos (normalmente de tamaño entre 5 y 11) de los respectivos residuos manteniendo ese orden.
- Calcular en cada grupo la media de las predicciones $(\bar{y}_{k.})$ y la desviación típica de los residuos (s_k) con k = 1,2,...,m, donde m es el número de grupos utilizado.
- Dibujar la gráfica de pares $(\bar{y}_{k.}, s_k)$.
- Ajustar a esta nube de puntos la curva $s_k = \theta \bar{y}_{k.}^{\nu}$.

Si v=0, hay homocedasticidad y no es necesario hacer ninguna transformación.

Si $v \neq 0$ se transforma la variable respuesta según la transformación de Box-Cox con parámetro λ= 1-v.

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & \text{si} \quad \lambda \neq 0 \\ \lg(y_i) & \text{si} \quad \lambda = 0 \end{cases} \qquad i = 1, 2, \ldots, n.$$

- La hipótesis de independencia: La hipótesis de que las observaciones muestrales son independientes es una hipótesis básica en el estudio de los modelos de regresión lineal. Con ello se entiende que los errores $\{\varepsilon_i\}_{i=1}^{n}$ son variables aleatorias independientes.

La falta de independencia, se produce fundamentalmente cuando se trabaja con variables aleatorias que se observan a lo largo del tiempo, es decir, cuando se trabaja con series temporales. Por ello, una primera medida para tratar de evitar la dependencia de las observaciones consiste en aleatorizar la recogida muestral.

El que no se cumpla la hipótesis de independencia afecta gravemente a los resultados del modelo de regresión, se obtienen estimadores de los parámetros y predicciones ineficientes y los intervalos de confianza y contrastes que se deducen de la tabla ANOVA no son válidos. Esto es debido a que se utiliza el resultado de que "la varianza de la suma de variables independientes es igual a la suma de las varianzas de cada variable". Propiedad que no se cumple para variables dependientes.

Si no se cumple la hipótesis de independencia se tienen dos alternativas. La primera, se basa en transformar los datos para obtener observaciones incorreladas (independientes, bajo hipótesis de normalidad) y luego aplicar las técnicas de regresión estudiadas (mínimos cuadrados), este método es un caso particular de la denominada técnica de mínimos cuadrados generalizados, que se puede aplicar en situaciones muy precisas y, por tanto, su utilización es un tanto restringida. La

segunda, se basa en aplicar métodos estadísticos diseñados para el estudio con observaciones dependientes como son los métodos de series de tiempo y los modelos de regresión dinámica.

La dependencia entre las observaciones surge la mayoría de las veces porque los datos son recogidos a lo largo del tiempo, y los gráficos y contrastes expuestos son válidos para detectarla.

- Gráficos para detectar dependencia son: el gráfico de los residuos frente al índice (tiempo), (t, e_t), el gráfico de los residuos e_{t+1} frente a e_t y el correlograma.
- Contrastes para detectar dependencias son: los contrastes basados en rachas, contrastes sobre las autocorrelaciones, el contraste de Ljung-Box.

Dentro de los contrastes de autocorrelaciones para modelos de regresión, el contraste de Durbin-Watson es muy utilizado.

a.5. Reseña teórica 5

¿Cómo detectar la multicolinealidad? La multicolinealidad indica que existe una fuerte correlación entre las variables regresoras, por lo tanto para detectarla se debe estudiar:

- Gráfico de dispersión matricial. El gráfico de dispersión matricial de las regresoras permite tener una idea acerca de la posible relación lineal entre dos regresoras.
- La matriz de correlaciones de las variables regresoras, **R.** La existencia de algún valor alto fuera de la diagonal de esta matriz ($r_{i,j}$, $i \neq j$, es próximo a ± 1) , indica que existe una fuerte relación lineal entre las variables regresoras x_i y x_j.

  Pero esto no es suficiente ya que la matriz **R** no detecta fuertes relaciones de una variable regresora con un conjunto de variables regresoras.

  Por ejemplo, considérese un conjunto de k (k grande) regresoras, donde las variables $x_1, x_2, \ldots, x_{k-1}$ son independientes pero la variable x_k está relacionada con las otras por la siguiente relación exacta

$$x_k = \frac{1}{k-1}\sum_{i=1}^{k-1} x_i. \quad \text{(a.18)}$$

Éste es un caso extremo de multicolinealidad y no se puede calcular $\hat{\alpha}$ ya que $rang(\mathbf{X}) = k$. Pero si k es grande todos los términos de **R** son pequeños, $r_{i,j} = 0$, si $i \neq j$, $i,j = 1,...,k-1$ y $r_{i,k} \simeq 0$, $i = 1,...,k-1$.

- Los elementos de la diagonal de la matriz $\mathbf{R}^{-1}$. Ya que se verifica que el i-ésimo elemento de esta matriz es

$$diag_{(i)}\mathbf{R}^{-1} = FIV(i) = \frac{1}{1 - r^2_{i\cdot resto}}, \qquad i = 1, \ldots, k, \quad \text{(a.19)}$$

por tanto si $FIV(i)$ es un valor muy alto, existe multicolinealidad causada por la variable x_i. Por ejemplo

si $diag_{(i)}\mathbf{R}^{-1} = FIV(i) > 10 \Rightarrow r_{i\,resto}{}^2 > 0{,}9$.

Como consecuencia se debería eliminar la variable explicativa x_i del modelo de regresión.

El inconveniente de este método es que la matriz $\mathbf{R}^{-1}$ se calcula con poca precisión (depende mucho de la muestra) cuando la matriz **R** es casi singular (su determinante es próximo a cero).

- Calcular los autovalores de la matriz **R**. Si las variables regresoras son ortogonales, todos los autovalores de **R** son iguales a uno, pero si hay multicolinealidad, al menos uno de los autovalores de **R** es próximo a cero, la variable regresora asociada a ese autovalor será la que es aproximadamente una combinación lineal de las otras variables regresoras.

  Para medir si un autovalor es próximo a cero o, equivalentemente, para medir la multicolinealidad asociada a la matriz **R** se utiliza el índice de condicionamiento (IC) de la matriz **R** que es una buena medida de la singularidad de esta matriz. La definición del índice de condicionamiento es la siguiente,

$$IC(\mathbf{R}) = \left(\frac{máx\ autovalor\ de\ R}{mín\ autovalor\ de\ R}\right)^{1/2} > 1 \quad \text{(a.20)}$$

  A modo indicativo se puede utilizar el siguiente criterio:

  - Si $10 \leq IC(\mathbf{R})$ no hay multicolinealidad.
  - Si $10 \leq IC(\mathbf{R}) \leq 30$, hay moderada multicolinealidad.
  - Si $IC(\mathbf{R}) > 30$, hay alta multicolinealidad.

a.6. Reseña Teórica 6

Con los residuos estandarizados o estudentizados se pueden construir los siguientes gráficos de interés:

- El gráfico de dispersión matricial de la Figura a.13, de todas las variables del modelo (respuesta y regresoras). En el estudio de un modelo de regresión lineal múltiple es el primer gráfico que se debe observar. Proporciona una primera idea de la existencia de relación lineal o de otro tipo entre la respuesta y las regresoras y también da una idea de posibles relaciones lineales entre las variables regresoras, lo que crea problemas de multicolinealidad.

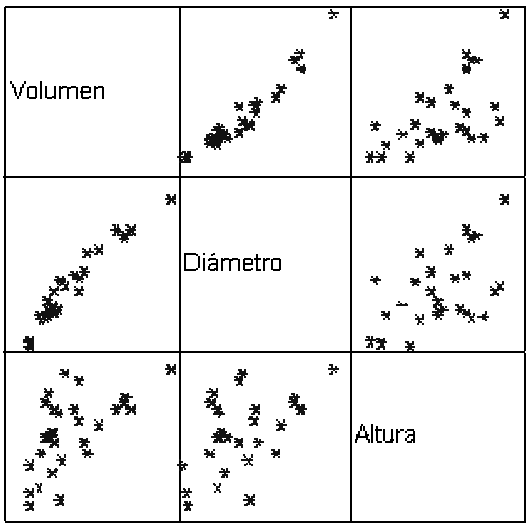

Fig. a.13. Gráfico de dispersión matricial

- El histograma de los residuos, que sirve para observar la existencia de normalidad, simetría y detectar observaciones atípicas.
- El gráfico probabilístico de normalidad (*p-p* y *q–q*) y el gráfico de simetría, que permite contrastar la normalidad (simetría) de la distribución de los residuos.

- El gráfico de residuos (e_i) frente a las predicciones $(\hat{y}_i)$, que permite detectar diferentes problemas:

  - *Heterocedasticidad*, la varianza no es constante y se deben de transformar los datos (la variable *Y*) o aplicar mínimos cuadrados ponderados.

  - *Error en el análisis*, se ha realizado mal el ajuste y se verifica que los residuos negativos se corresponden con los valores pequeños $\hat{y}_i$ y los errores positivos se corresponden con los valores grandes de $\hat{y}_i$, o al revés.

  - *El modelo es inadecuado* por falta de linealidad y se deben de transformar los datos o introducir nuevas variables que pueden ser cuadrados de las existentes o productos de las mismas. O bien se deben introducir nuevas variables explicativas.

  - *Existencia de observaciones atípicas* o puntos extremos.

  - Tener en cuenta que se debe utilizar el gráfico de residuos (e_i) frente a las predicciones $(\hat{y}_i)$ en lugar del gráfico de residuos (e_i) frente a las observaciones (y_i) porque las variables $\vec{e}$ e $\vec{Y}$ están correladas, mientras que las variables $\vec{\mathbf{e}}$ e $\hat{\mathbf{Y}}$ no lo están.

- El gráfico de residuos (e_i) frente a una variable explicativa $(x_{i,j})$ de la Figura a.14, permite deducir si la existencia de heterocedasticidad o la falta de linealidad en el modelo son debidas a la variable explicativa representada.

  Gráficos de este tipo son los representados en las figuras. En la primera de ellas se observa que la relación con la variable x_j no es lineal y, probablemente, un ajuste cuadrático sea adecuado, también se tendrían dudas acerca de la homocedasticidad del modelo.

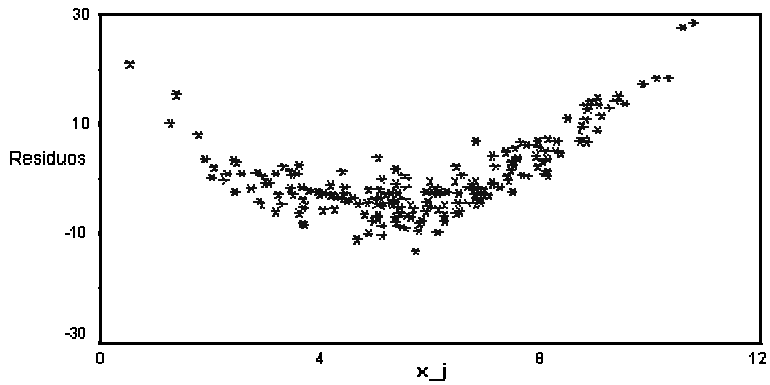

Fig. a.14. Gráfico de residuos frente a una variable explicativa

En la Figura a.15, se observa que el modelo es heterocedástico y la causa de este problema puede ser la variable explicativa x_j. Por ello, la solución se basa en transformar el modelo teniendo en cuenta este hecho.

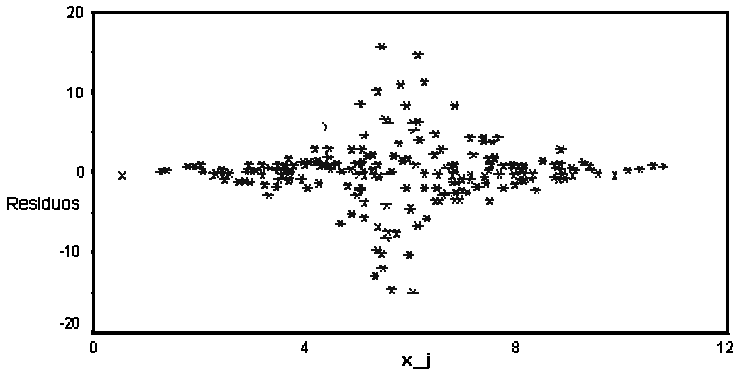

Fig. a.15. Modelo heterocedástico

- El gráfico de residuos (e_i) frente a una variable omitida de la Figura a.16, permite valorar si esta variable influye en el modelo y por lo tanto se debe incluir como una nueva variable regresora.

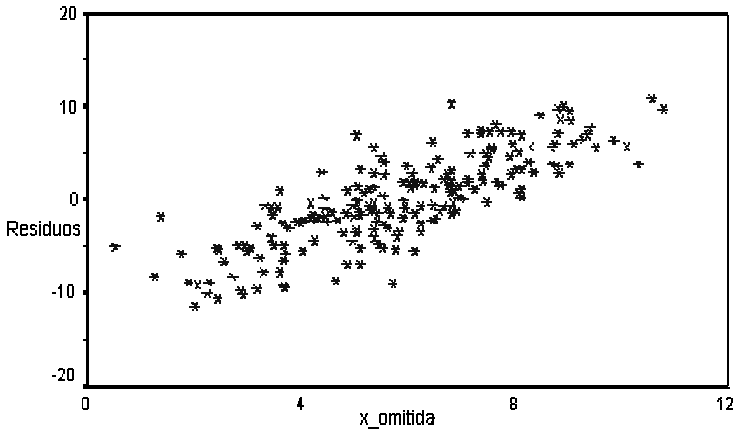

Fig. a.16. Gráfico de residuos frente a una variable omitida

En la figura de residuos frente a una variable omitida (x_{omit}) se observa que existe una relación lineal con esta variable y por tanto se mejora el ajuste si se incluye la variable x_{omit}.

Una situación frecuente se produce cuando se tienen observaciones de diferentes poblaciones y se debe de incluir una variable de clasificación en el

modelo de regresión. Esto se puede observar en el gráfico de residuos frente a predicciones de la Figura a.17.

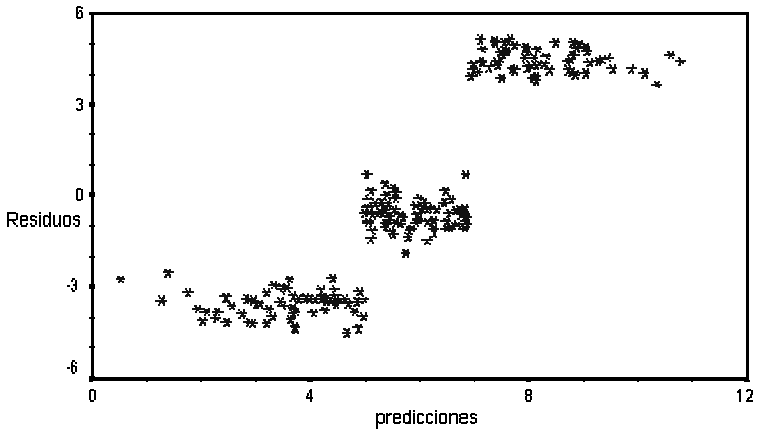

Fig. a.17. Gráfico de residuos frente a las predicciones

El gráfico de los residuos frente a la variable de clasificación omitida se presenta en la Figura a.18.

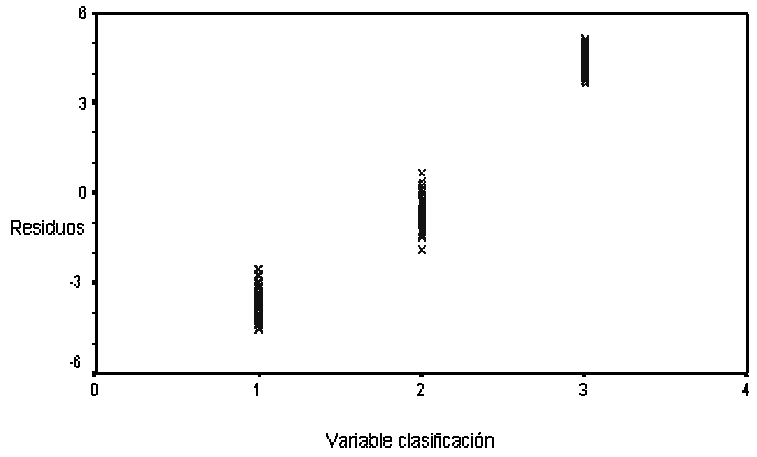

Fig. a.18. Gráfico de residuos frente a una variable de clasificación omitida

- El gráfico parcial de residuos, es útil para valorar la influencia real de una variable regresora, esto es, conocer la información nueva que aporta la variable regresora en estudio y que no aportan las otras variables regresoras. Según el paquete estadístico que se utilice los gráficos parciales de residuos se pueden construir de diferentes formas.

Tipo 1. Si se tienen *k* variables regresoras $\{x_1, x_2, \ldots, x_k\}$ y se desea obtener el gráfico parcial de residuos respecto a la variable x_k, se procede de la siguiente forma:

- Se calcula el modelo de regresión respecto a las restantes (*k*-1) variables regresoras,

$$\vec{\mathbf{y}} = \hat{\alpha}_0^*\vec{\mathbf{1}} + \hat{\alpha}_1^*\vec{\mathbf{x}}_1 + \hat{\alpha}_2^*\vec{\mathbf{x}}_2 + \ldots + \hat{\alpha}_{k-1}^*\vec{\mathbf{x}}_{k-1} + \vec{\mathbf{e}}_k^* \quad \text{(a.21)}$$

- Se calculan los residuos

$$\vec{\mathbf{e}}_k^* = \vec{\mathbf{y}} - \hat{\mathbf{y}}^* = \vec{\mathbf{y}} - \left(\hat{\alpha}_0^*\vec{\mathbf{1}} + \hat{\alpha}_1^*\vec{\mathbf{x}}_1 + \hat{\alpha}_2^*\vec{\mathbf{x}}_2 + \ldots + \hat{\alpha}_{k-1}^*\vec{\mathbf{x}}_{k-1}\right), \quad \text{(a.22)}$$

que son la parte de *Y* no explicada por las variables $x_1, x_2, \ldots, x_{k-1}$.

- Por tanto, la gráfica de los residuos "parciales" e_k^* frente a la variable x_k permite valorar la importancia real de esta variable.

Tipo 2. Un gráfico muy parecido y más fácil de calcular se obtiene de la siguiente forma. Calcular

$$\tilde{\mathbf{e}}_k^* = \vec{\mathbf{e}} + \hat{\alpha}_k\vec{\mathbf{x}}_k = (\vec{\mathbf{y}} - \hat{\mathbf{y}}) + \hat{\alpha}_k\vec{\mathbf{x}}_k$$

$$\tilde{\mathbf{e}} = \vec{\mathbf{y}} - \left(\hat{\alpha}_0\vec{\mathbf{1}} + \hat{\alpha}_1\vec{\mathbf{x}}_1 + \hat{\alpha}_2\vec{\mathbf{x}}_2 + \ldots + \hat{\alpha}_{k-1}\vec{\mathbf{x}}_{k-1}\right) \quad \text{(a.23)}$$

Se obtiene un nuevo gráfico parcial representando los residuos "parciales" $\tilde{e}_k^*$ frente a la variable x_k.

Si la variable x_k es ortogonal a las restantes variables explicativas los estimadores $\hat{\alpha}_i^*$ y $\hat{\alpha}_i$, $i = 1,\ldots,k-1$, serán muy próximos y, por tanto, también lo son los vectores e_k^* y $\tilde{e}_k^*$. Lo que hace que los dos gráficos de residuos parciales sean casi iguales en este caso.

Tipo 3. Otro gráfico parcial de interés que proporcionan algunos paquetes estadísticos es el siguiente (se quiere calcular el gráfico parcial respecto a x_k):

Se calculan los modelos de regresión de las variables *Y* y x_k respecto a las restantes (*k*-1) variables regresoras,

$$y = \hat{\alpha}_0^* + \hat{\alpha}_1^*x_1 + \hat{\alpha}_2^*x_2 + \ldots + \hat{\alpha}_{k-1}^*x_{k-1} + e_k^*$$

$$x_k = \gamma_0 + \gamma_1x_1 + \gamma_2x_2 + \ldots + \gamma_{k-1}x_{k-1} + e_{\gamma,k} \quad \text{(a.24)}$$

Se representa el gráfico de residuos de e_k^* frente a los residuos $e_{\gamma,k}$. Esto es, el gráfico de los pares $(e_{\gamma,k}, e_k^*)$. Este gráfico da una idea de la relación entre la variable *Y* y la variable x_k una vez que se ha eliminado la influencia de las otras variables regresoras.

- El gráfico de residuos (e_i) frente al índice (tiempo=***i***), proporciona información acerca de la hipótesis de independencia de los residuos. En este gráfico se pueden observar algunas características que indican falta de independencia, tales como una correlación positiva o negativa, la existencia de tendencias, saltos estructurales, rachas,....,etc.

  En este gráfico también se puede observar si existe una relación lineal con el índice y éste debe de incluirse en el modelo de regresión como variable explicativa.

Como ya se indicó anteriormente al realizar estos gráficos, una alta variabilidad en los residuos (σ^2 grande) puede "esconder" una pauta de comportamiento de los mismos y, en este caso, es conveniente "filtrar" o "suavizar" los residuos utilizando la técnica de "medias móviles" o "medianas móviles" u otro análogo. Así el filtro de "medianas móviles de orden tres" a partir de los residuos originales $\{e_t\}_{t=1}^{n}$ genera una nueva sucesión de residuos $\{\check{e}_t\}_{t=2}^{n-1}$ más "suave".

$$e_t \longrightarrow filtro\ m.m.(3) \longrightarrow \check{e}_t = mediana\{e_{t-1}, e_t, e_{t+1}\} \quad \text{(a.25)}$$

para t = 2,...,n- 1. Si se considera que la sucesión resultante no está suficientemente suavizada se puede repetir el procedimiento de suavización.

a.7. Reseña teórica 7

Repasemos las características más importantes de las hipótesis del modelo de regresión lineal múltiple.

- *Hipótesis de normalidad.* La hipótesis de normalidad afirma que los errores del modelo ε siguen una distribución normal. Esta hipótesis se contrasta a partir de los residuos estandarizados $\{r_i\}_{i=1}^{n}$. Se pueden aplicar los contrastes y gráficos de normalidad.

- Gráficos para observar la normalidad son: el histograma, estimador núcleo de la densidad de Rosenblatt-Parzen, gráfico *p-p* y gráfico *q-q*.
- Contrastes de normalidad son: contraste de asimetría y curtosis, contraste chi-cuadrado, contraste de Kolmogorov-Smirnov-Liliefors.

En relación con la utilización de los residuos para contrastar la normalidad, debe de tenerse en cuenta que de la relación $\vec{\mathbf{e}} = (\mathbf{I} - \mathbf{H})\vec{\varepsilon}$ se sigue que

$$e_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j, \quad i = 1, \ldots, n. \tag{a.26}$$

Por tanto, si ε_i es pequeño, el término dominante en la relación anterior es el sumatorio que por el Teorema Central del Límite es aproximadamente normal. Entonces puede ocurrir que los e_i sean aproximadamente normales aunque los ε_i no lo sean. En cualquier caso, si n es grande en relación con k+1 se pueden utilizar los residuos estandarizados r_i para contrastar la hipótesis de normalidad.

La falta de normalidad influye en el modelo en:

- Los estimadores mínimo-cuadráticos no son eficientes (de mínima varianza).
- Los intervalos de confianza de los parámetros del modelo y los contrastes de significación son solamente aproximados y no exactos.

Causas que dan origen a la falta de normalidad son las siguientes:

- *Existen observaciones heterogéneas.* En este caso se debe averiguar la causa que origina estas observaciones: errores en la recogida de datos; el modelo especificado no es correcto porque se han omitido variables regresoras (por ejemplo, no se ha tenido en cuenta una variable de clasificación cuando las observaciones proceden de diferentes poblaciones).

  Se debe hacer un estudio de influencia de las observaciones atípicas para averiguar el grado de influencia en la estimación del modelo. Si esta influencia es muy grande puede ser conveniente recurrir a procedimientos de estimación robusta en el cálculo del modelo.
- *Existe asimetría en la distribución.* En este caso suele ser conveniente transformar la variable respuesta (transformación de Box-Cox). Este problema suele estar relacionado con otros problemas como falta de linealidad o heterocedasticidad, la solución de transformar las observaciones pueden resolverlos conjuntamente.
- Si la hipótesis de normalidad no se verifica y las soluciones anteriores no son válidas se pueden obtener intervalos de confianza de los parámetros por métodos diferentes de los expuestos en los que se tiene en cuenta la distribución específica de los errores.

- *Hipótesis de homocedasticidad.* Una hipótesis del modelo de regresión es la homocedasticidad y todo lo comentado sobre este problema en el modelo de regresión lineal simple sigue siendo válido en el modelo de regresión lineal múltiple.

La falta de homocedasticidad influye en el modelo de regresión lineal, los estimadores mínimo-cuadráticos siguen siendo centrados pero no son eficientes y las fórmulas de las varianzas de los estimadores de los parámetros no son correctas. Por tanto no pueden aplicarse los contrastes de significación.

La heterocedasticidad se detecta en los gráficos de residuos:

- De forma general, en el gráfico de residuos (e_i) frente a las predicciones $(\hat{y}_i)$.
- En el gráfico de residuos (e_i) frente a una variable explicativa (x_{ij}) si se sospecha que la heterocedasticidad es debida a la variable explicativa x_j.
- Si los gráficos anteriores son dudosos se pueden hacer grupos de los residuos ordenados de menor a mayor según las predicciones $(\hat{y}_i)$ y en cada grupo calcular la media de las predicciones $(\bar{y}_{k\cdot})$ y la desviación típica de los residuos $(\hat{s}_k)$. Si hay homocedasticidad, la nube de puntos $(\bar{y}_{k\cdot}, \hat{s}_k)$ se ajusta a una recta horizontal, en caso contrario, es necesario transformar los datos.
- Existen contrastes específicos para contrastar la homocedasticidad.

Para resolver este problema las alternativas que hay son las siguientes:

- Transformar los datos. En muchos casos es suficiente con tomar logaritmos en la variable respuesta (o de forma más compleja, aplicar la transformación de Box-Cox). Por otra parte, el problema puede estar ligado a otros problemas como falta de normalidad, falta de linealidad que, normalmente, también se resuelven al hacer la transformación.
- Si la heterocedasticidad es debida a una variable regresora (por ejemplo x_k) y la varianza aumenta linealmente con la variable x_k, $Var(\varepsilon_i) = kx_{ik}$. Entonces se obtiene homocedasticidad haciendo la siguiente transformación del modelo de regresión

$$\frac{y}{\sqrt{x_k}} = \hat{\alpha}_0 \frac{1}{\sqrt{x_k}} + \hat{\alpha}_1 \frac{x_1}{\sqrt{x_k}} + \hat{\alpha}_2 \frac{x_2}{\sqrt{x_k}} + \ldots + \hat{\alpha}_{k-1} \frac{x_{k-1}}{\sqrt{x_k}} + \hat{\alpha}_k \frac{1}{\sqrt{x_k}} + \frac{\varepsilon}{\sqrt{x_k}} \quad \text{(a.27)}$$

Si la que varía linealmente con x_k es la desviación típica, la transformación a realizar sería la siguiente

$$\frac{y}{x_k} = \hat{\alpha}_0 \frac{1}{x_k} + \hat{\alpha}_1 \frac{x_1}{x_k} + \hat{\alpha}_2 \frac{x_2}{x_k} + \ldots + \hat{\alpha}_{k-1} \frac{x_{k-1}}{x_k} + \hat{\alpha}_k + \frac{\varepsilon}{x_k} \quad \text{(a.28)}$$

Las transformaciones anteriores son casos particulares del método de mínimos cuadrados ponderados, método muy utilizado para obtener estimadores de los parámetros en situaciones de heterocedasticidad.

- Las transformaciones anteriores son casos particulares del denominado *método de mínimos cuadrados ponderados.* El método se basa en calcular los estimadores de los parámetros del modelo como los valores que minimizan la siguiente función de los residuos

$$\Psi(\vec{\alpha}) = \sum_{i=1}^{n} (y_i - \vec{\mathbf{x}}_i.\vec{\alpha})^2 \, \omega(e_i) \qquad \text{(a.29)}$$

donde $\omega(e_i)$ es una función peso que toma valores altos si la varianza de e_i es pequeña y toma valores bajos si la varianza de e_i es grande.

El método de mínimos cuadrados ponderados es un caso particular del *método de mínimos cuadrados generalizados*.

- *Hipótesis de independencia.* La independencia de los errores es una hipótesis básica en el estudio de un modelo de regresión lineal.

La falta de cumplimiento de la hipótesis de independencia tiene efectos graves sobre los resultados del estudio. Influye en:

- Los estimadores $\hat{\alpha}$ son centrados pero ineficientes (no son de varianza mínima).
- El estimador $\hat{s}_R^2$ normalmente subestima el parámetro σ^2, lo que hace que los contrastes de significación (contrastes individuales de la *t*) no sean válidos y tienden a detectar relaciones inexistentes, denominadas *relaciones espurias*, que son relaciones falsas entre variables independientes que siguen una evolución análoga en el tiempo y tienen un R^2 alto.
- Las predicciones son ineficientes.
- La falta de independencia se suele dar situaciones en que las observaciones son recogidas secuencialmente en el tiempo. Esto ocurre en el estudio de muchas variables económicas, sociales y demográficas. En este caso la variable "tiempo" puede ser una variable regresora.

Se detecta la falta de independencia en:

- Los siguientes gráficos: el gráfico de residuos (e_t) frente al índice (o tiempo), (*t*); el gráfico de (e_t) frente a (e_{t-1}); el gráfico de la función de autocorrelación simple de los residuos (fas).

- Los siguientes contrastes de independencia: el contraste de Durbin-Watson sobre el primer coeficiente de correlación; el contraste de Ljung-Box sobre las autocorrelaciones que se consideren significativas.

Si existe dependencia entre las observaciones la metodología descrita para estudiar los modelos de regresión lineal general por mínimos cuadrados ordinarios no es válida y, en la mayoría de las situaciones, deben utilizarse técnicas de series de tiempo y regresión dinámica.

En algunas situaciones se pueden estimar los parámetros del modelo de regresión por el método de *mínimos cuadrados generalizados.*

"...Es importante realizar un análisis de influencia para conocer las observaciones muestrales que tienen una mayor influencia en el modelo y las observaciones atípicas o heterogéneas que no se ajustan al modelo..." [56]

Este estudio es análogo al desarrollado para el modelo de regresión lineal simple y que en esta sección se generaliza al caso múltiple.

- *Influencia a priori. Valor de influencia.* Para el estudio de la influencia de una observación en el cálculo del modelo de regresión se debe tener en cuenta la siguiente ecuación

$$\hat{y}_t = \hat{\alpha}_0 + \hat{\alpha}_1 x_{t1} + \ldots + \hat{\alpha}_k x_{tk} = \sum_{i=1}^{n} h_{ti}\, y_i, \quad t = 1, \ldots, n, \tag{a.30}$$

donde h_{ti} son unos pesos que en el modelo de regresión lineal simple (k=1) tienen la forma

$$h_{ti} = \frac{1}{n} + \frac{(\bar{x}_t - \bar{x})(\bar{x}_i - \bar{x})}{\sum_{i=1}^{n} (\bar{x}_i - \bar{x})^2}, \quad t, i = 1, \ldots, n. \tag{a.31}$$

La ecuación en forma matricial es

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\alpha} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\vec{\mathbf{Y}} = \mathbf{H}\vec{\mathbf{Y}}, \tag{a.32}$$

siendo $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ la matriz de proyección ortogonal en el espacio generado por las variables regresoras. $H = \{h_{t,i}\}_{t,i=1}^{n}$ es una matriz cuadrada y simétrica.

De esto se deduce que la predicción de una observación $\hat{y}_t$ es una combinación lineal de los valores de la variable respuesta ($\vec{Y} = (y_1, \ldots, y_n)$), con pesos h_{ti} que se obtienen a partir de los valores de la matriz de diseño **X**. La influencia de la observación $(\vec{\mathbf{x}}_i, y_i)$ en el cálculo de $\hat{y}_t$ viene dado por:

- El valor de y_i.
- El valor de h_{ti}.

Por tanto, el valor h_{ti} mide, al menos parcialmente, la influencia "a priori" de la observación *i*-ésima en el cálculo de la predicción $\hat{y}_t$. Los elementos de la diagonal de la matriz **H**, h_{ii}, i=1,...,n, miden la influencia de la observación *i*-ésima en el cálculo de $\hat{y}_i$. Su expresión viene dada por

$$h_{ii} = \vec{\mathbf{x}}_{i\cdot}^{t} \left(\mathbf{X}^t\mathbf{X}\right)^{-1} \vec{\mathbf{x}}_{i\cdot}, \qquad i = 1, 2, \ldots, n, \quad \text{(a.33)}$$

donde $\vec{\mathbf{x}}_i{}^{\cdot t}$ es la fila *i*-ésima de la matriz **X** (datos de la observación *i*-ésima).

En particular, en el modelo de regresión lineal simple (k=1) se verifica

$$h_{ii} = \frac{1}{n}\left(1 + \frac{(\bar{x}_i - \bar{x})^2}{s_X^2}\right), \quad i = 1, \ldots, n, \quad \text{(a.34)}$$

En resumen, la influencia a priori de las observaciones viene dada por los elementos de la diagonal de **H**, h_{ii}, i=1,2,...,n, el valor h_{ii} mide la distancia del punto $\vec{\mathbf{x}}_i$ al centro $\bar{\mathbf{x}}$, y se le denomina *valor de influencia a priori* (en inglés *"leverage"*). Observaciones con valor de influencia alto son observaciones que "a priori" influyen en el cálculo del modelo y observaciones con valor de influencia bajo "a priori" influyen poco.

Para saber si un h_{ii} es un valor grande o no se debe de tener en cuenta que si no hay filas repetidas en la matriz de diseño **X** se verifica que:

$$\frac{1}{n} \le h_{ii} \le 1 \qquad i = 1, 2, \ldots, n$$

$$\sum_{i=1}^{n} h_{ii} = traza\,(H) = k + 1 \quad \text{(a.35)}$$

Por tanto $E(h_{tt}) = (k+1)/n$. Y se puede considerar que una observación tiene un valor de influencia grande si se verifica que

$$h_{tt} > 2\,\frac{k+1}{n}. \quad \text{(a.36)}$$

Otro criterio se basa en calcular la varianza de los h_{ii}

$$s_h^2 = \frac{1}{n}\sum_{i=1}^{n}\left(h_{ii} - \frac{k+1}{n}\right)^2, \quad \text{(a.37)}$$

y considerar que una observación tiene un valor de influencia grande si

$$h_{ii} > E\,(h_{ii}) + 3s_h = \frac{k+1}{n} + 3\,s_h. \quad \text{(a.38)}$$

---

[56] "Modelos Estadísticos aplicados", J. Vilar Fernández, Universidade da Coruña, España 2003

El valor de influencia de las observaciones muestrales es un valor comprendido entre 1/*n* y 1, siendo los casos extremos los siguientes:

- $\vec{\mathbf{x}}_i = \bar{\mathbf{x}}$, entonces $h_{ii} = 1/n$.
- Considérese la muestra $\{(x^*, y_1), (x^*, y_2), \ldots, (x^*, y_{n-1}), (x_n, y_n)\}$ de un modelo de regresión lineal simple, entonces $h_{ii} = 1/(n-1)$, $i = 1,\ldots,n-1$, puntos en los que $x_i = x^*$, y $h_{nn} = 1$, el mayor valor que puede tomar. En este caso la recta de regresión pasa por los puntos $(x^*, \bar{y}_{(n-1)})$ y (x_n, y_n), siendo $\bar{y}_{(n-1)} = 1/(n-1) \sum_{i=1}^{n-1} y_i$.

Unas pocas observaciones con valor de influencia a priori grande pueden producir multicolinealidad entre dos o más variables regresoras. Esto puede observarse claramente en la Figura a.19, donde se representa el gráfico de dos variables regresoras x_1 y x_2 en el que la mayoría de las observaciones están agrupadas en una nube pero hay dos observaciones con un alto valor de influencia a priori, estas dos observaciones producen una alta correlación (R=0,632) entre las dos variables regresoras. Si se eliminan las dos observaciones influyentes de la muestra la correlación es casi nula (R=0,079), las variables son incorreladas.

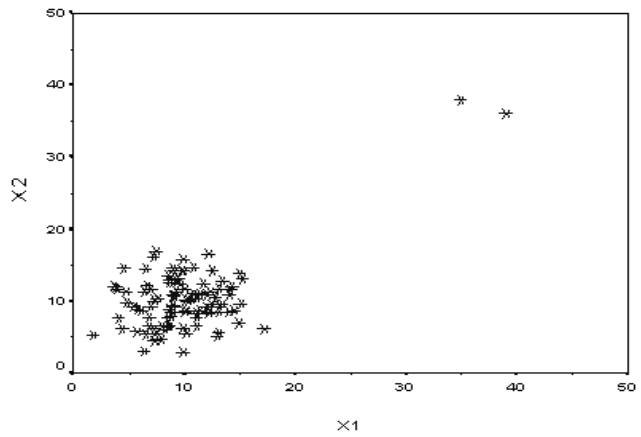

Fig. a.19. Gráfico entre las variables X_1 y X_2

Muchos paquetes estadísticos proporcionan la distancia de Mahalanobis de los puntos muestrales $\vec{\mathbf{x}}_{i.} = (x_{i1}, x_{i2}, \ldots, x_{ik})$, $i = 1,2,\ldots,n$, al punto medio de la nube de las variables regresoras $\bar{\mathbf{x}}=(\bar{x}_{.1}, \bar{x}_{.2}, \ldots, \bar{x}_{.k})$, donde $\bar{x}_j$, $j = 1,2,\ldots,k$ es la media de los datos de la variable x_j. Esta distancia viene definida como sigue

$$d_M^2(\vec{\mathbf{x}}_{i.};\bar{\mathbf{x}}) = (\vec{\mathbf{x}}_{i.} - \bar{\mathbf{x}})\,\mathbf{S}^{-1}\,(\vec{\mathbf{x}}_{i.} - \bar{\mathbf{x}})^t \quad \text{(a.39)}$$

siendo *S* la matriz de varianzas-covarianzas del vector de variables $(x_1, x_2, \ldots, x_k)$.
La distancia de Mahalanobis es una distancia estadística que generaliza la distancia euclídea entre dos vectores en la que se tiene en cuenta la dispersión de las variables y su dependencia. Un valor alto de la distancia de Mahalanobis indica que el punto se aleja del centro de la nube y, por tanto, es una posible observación influyente a priori.

Si en el ajuste del modelo de regresión lineal no se utiliza la observación *i*-ésima, el vector de predicciones es

$$\hat{\mathbf{Y}}_{(i)} = \mathbf{X}\left(\mathbf{X}_{(i)}^t\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_{(i)}^t\vec{\mathbf{Y}}_{(i)}, \quad \text{(a.40)}$$

donde el subíndice (*i*) indica que no se utiliza la observación *i*-ésima.
Se define el *residuo eliminado* $e^{(i)}$ como el residuo obtenido utilizando la predicción calculada a partir de la muestra excepto la *i*-ésima observación, $\hat{y}_i^{(i)}$. Esto es,

$$e_{(i)} = y_i - \hat{y}_{i(i)}, \quad i = 1, \ldots, n. \quad \text{(a.41)}$$

Teniendo en cuenta la siguiente relación entre los residuos ordinarios y los eliminados

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}, \quad i = 1, \ldots, n, \quad \text{(a.42)}$$

se puede deducir un nuevo criterio para distinguir a las observaciones influyentes a priori: si la observación *i*-ésima influye mucho (h_{ii} es grande) los residuos ordinarios y los residuos eliminados son distintos, por el contrario, si el valor de influencia es pequeño ($h_{ii} \approx 0$) los dos residuos (ordinario y eliminado) son parecidos.

La identificación de las observaciones influyentes a posteriori es de mayor interés. Una observación influyente a posteriori es aquella $(\vec{\mathbf{x}}_{i.}, y_i)$ cuya inclusión en el ajuste modifica sustancialmente la estimación del modelo. En este caso, se están considerando los datos de las variables regresoras y de la variable respuesta.
El problema básico es determinar la influencia del dato $(\vec{\mathbf{x}}_{i.}, y_i)$ en el ajuste del modelo de regresión lineal múltiple. O, equivalentemente, se desea realizar el siguiente contraste estadístico (C_I):

- H_0 : El modelo ajustado con toda la muestra es igual al modelo ajustado con la muestra excepto el dato $(\vec{\mathbf{x}}_{i.}, y_i)$.

- H_1 : El modelo ajustado con toda la muestra es distinto al modelo ajustado con la muestra excepto el dato $(\vec{\mathbf{x}}_{i.}, y_i)$.

Si la observación $(\vec{\mathbf{x}}_{i.}, y_i)$ es influyente en el modelo de regresión se observa en

- la estimación de los parámetros del modelo de regresión $(\vec{\alpha})$: $\hat{\alpha}$
- el vector de predicción de las observaciones: $\hat{Y}$
- la predicción de la respuesta en el punto *i*-ésimo: $\hat{y}_i$

Los estadísticos para resolver el contraste C_I se basan en calcular la distancia entre las estimaciones de cualquiera de los tres valores anteriores cuando se utiliza toda la muestra

$$\hat{\alpha} \quad - \quad \hat{Y} \quad - \quad \hat{y}_i \quad \text{(a.43)}$$

y las mismas estimaciones cuando se utiliza toda la muestra excepto el dato $(\vec{\mathbf{x}}_{i.}, y_i)$

$$\hat{\alpha}_{(i)} \quad - \quad \hat{Y}_{(i)} \quad - \quad \hat{y}_{i(i)} \quad \text{(a.44)}$$

Las tres distancias llevan al mismo estadístico, el *D*-estadístico de Cook, definido por

$$D(i) = \frac{\left(\hat{\alpha} - \hat{\alpha}_{(i)}\right)^t \mathbf{X}^t \mathbf{X} \left(\hat{\alpha} - \hat{\alpha}_{(i)}\right)}{(k+1)\,\hat{s}_R^2} = \frac{\left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)}\right)^t \left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)}\right)}{(k+1)\,\hat{s}_R^2} =$$

$$= \frac{\left(\hat{y}_i - \hat{y}_{i(i)}\right)^2}{(k+1)\,\hat{s}_R^2 h_{ii}} = \left(\frac{h_{ii}}{1-h_{ii}}\right) \frac{e_i^2}{(k+1)\,\hat{s}_R^2\,(1-h_{ii})} = \left(\frac{h_{ii}}{1-h_{ii}}\right) \frac{r_i^2}{k+1}, \quad \text{(a.45)}$$

siendo r_i el *i*-ésimo residuo estandarizado y *k* el número de variables regresoras. Bajo la hipótesis nula, la observación *i*-ésima no es una observación influyente a posteriori, se verifica que

$$D(i) \sim F_{k+1;n-(k+1)}. \quad \text{(a.46)}$$

La familia de estadísticos *DFFITS* relacionados con el ***D***-estadístico de Cook se definen como

$$DFFITS(i) = \frac{\left(\hat{y}_i - \hat{y}_{i(i)}\right)^2}{\hat{s}_{R,(i)}\sqrt{h_{ii}}} = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} \frac{e_i}{\hat{s}_{R,(i)}\sqrt{1-h_{ii}}} = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} t_i, \quad \text{(a.47)}$$

donde t_i es el residuo estudentizado. Belsey, Kuh y Welsch (1980) proponen utilizar como cota superior de este estadístico el valor 2 $(k/n)^{1/2}$. Esto es, la observación $(\vec{\mathbf{x}}_{i.}, y_i)$ es influyente a posteriori si

$$DFFITS(i) > 2\sqrt{\frac{k}{n}}. \quad \text{(a.48)}$$

a.8 Reseña teórica 8

Procedimientos para selección de variables regresoras:

- "Eliminación progresiva" ("Backward Stepwise Regression"). Este procedimiento parte del modelo de regresión con todas las variables regresoras y en cada etapa se elimina la variable menos influyente según el contraste individual de la *t* (o de la *F*) hasta una cierta regla de parada.

  El procedimiento de eliminación progresiva tiene los inconvenientes de necesitar mucha capacidad de cálculo si *k* es grande y llevar a problemas de multicolinealidad si las variables están relacionadas. Tiene la ventaja de no eliminar variables significativas.

- "Introducción progresiva" ("Fordward Stepwise Regression"). Este algoritmo funciona de forma inversa que el anterior, parte del modelo sin ninguna variable regresora y en cada etapa se introduce la más significativa hasta una cierta regla de parada.

  El procedimiento de introducción progresiva tiene la ventaja respecto al anterior de necesitar menos cálculo, pero presenta dos graves inconvenientes, el primero, que pueden aparecer errores de especificación porque las variables introducidas permanecen en el modelo aunque el algoritmo en pasos sucesivos introduzca nuevas variables que aportan la información de las primeras. Este algoritmo también falla si el contraste conjunto es significativo pero los individuales no lo son, ya que no introduce variables regresoras.

- "Regresión paso a paso" ("Stepwise Regression"). Este método es una combinación de los procedimientos anteriores, comienza como el de introducción progresiva, pero en cada etapa se plantea si todas las variables introducidas deben de permanecer. Termina el algoritmo cuando ninguna variable entra o sale del modelo.

  El algoritmo es el siguiente:

  Paso 1. Se elige un "criterio de entrada", t_{IN} y un "criterio de salida", t_{OUT}.

  Un "criterio de entrada" es un valor t_{IN} de una variable con distribución *t* tal que el intervalo $(-t_{IN}, t_{IN})$ es la región de aceptación de que una variable regresora no es significativa. Análogamente un "criterio de salida" es un valor de una variable t_{OUT} con distribución *t* tal que el intervalo $(-t_{OUT}, t_{OUT})$ es la

región de aceptación de que la variable regresora no es significativa (no entra en el modelo).

Se calculan los coeficientes de correlación lineal simple $r(Y, x_i)$, $i = 1,...,k$.

Supongamos que el mayor de ellos corresponde a la variable x_k, que será la candidata a entrar en el modelo.

Paso 2. Se obtiene la regresión de Y sobre x_k y se calcula el estadístico $\hat{t}_k$ para el coeficiente α_k

$$\hat{t}_k = \frac{\hat{\alpha}_k}{\hat{s}_R \sqrt{q_{kk}}}, \quad \text{(a.49)}$$

(Es equivalente hacerlo con los contrastes individuales de la F, que es lo que hacen la mayoría de los programas estadísticos, entonces el criterio de salida viene dado por un número F_{OUT} y la región de aceptación es $(0, F_{OUT})$, y el criterio de entrada sería un número F_{IN}.)

Paso 3. El valor $\hat{t}_k$ se compara con el valor t_{IN} elegido, de forma que:

- si $|\hat{t}_k| \geq t_{IN}$, entonces la variable x_k es significativa y se introduce en el modelo. Ir al Paso 4.

- si $|\hat{t}_k| < t_{IN}$, se acepta que la variable x_k no es significativa y no se introduce en el modelo. Se termina el algoritmo.

Paso 4. Una vez introducido x_k en el modelo se calculan las correlaciones parciales (eliminando la influencia de x_k): $r_{Y,i\cdot k}$, $i = 1,...,k - 1$. Se calcula la correlación parcial mayor que supongamos que es la correspondiente a la variable x_{k-1} : $r_{Y,k-1\cdot k}$,

Paso 5. Se calcula el modelo de regresión de Y respecto a x_k y x_{k-1}. Se calculan los estadísticos $\hat{t}_{k-1}$ y $\hat{t}_k$.

Paso 6. Se compara $\hat{t}_{k-1}$ con t_{IN} .

- si $|\hat{t}_{k-1}| \geq t_{IN}$, entonces la variable x_{k-1} es significativa y se introduce en el modelo. Ir al Paso 7.

- si $|\hat{t}_{k-1}| < t_{IN}$, se acepta que la variable x_{k-1} no es significativa y no se introduce en el modelo. Se termina el algoritmo.

Paso 7. Se decide si la variable x_k debe permanecer en el modelo. Para ello se compara $\hat{t}_k$ con t_{OUT}.

- si $|\hat{t}_k| < t_{OUT}$, se acepta que la variable x_k no es significativa y se elimina del modelo. Se vuelve al Paso 4, con x_{k-1} como variable regresora. Continúa el proceso.

- si $|t_k| \geq t_{OUT}$, entonces la variable x_k es significativa. Se vuelve al Paso 4, con x_{k-1} y x_k como variables regresoras. Continúa el proceso.

Muchos paquetes estadísticos tienen programado este algoritmo utilizando el contraste de la *F* en lugar del contraste de la *t* y, generalmente, utilizan que $F_{IN} = F_{OUT}$, esto es una elección del usuario pero no una condición para su utilización. Lo que si es necesario es que $F_{IN} \geq F_{OUT}$, para evitar que una variable que entra en una etapa salga en la siguiente.

El algoritmo paso a paso tiene las ventajas del algoritmo de introducción progresiva pero lo mejora al no mantener fijas en el modelo las variables que ya entraron en una etapa, evitando de esta forma problemas de multicolinealidad. En la práctica, es un algoritmo bastante utilizado que proporciona resultados razonables cuando se tiene un número grande de variables regresoras.

En todo caso, la utilización de estos algoritmos de manera automática es peligroso y una vez obtenido el modelo de regresión se debe chequear que se verifican las hipótesis del modelo así como tener en mente el problema de regresión que se está estudiando.

Cuando se está buscando el subconjunto de variables regresoras que deben entrar en el modelo de regresión, un estudio complementario a la utilización de los algoritmos descritos sería considerar todos los posibles subconjuntos. Si *k* es grande ($k>5$) hacer esto de una "forma directa" es muy caro computacionalmente. Pero puede hacerse "implicítamente", entendiendo con ello, que un sencillo estudio permita desechar un gran número de subconjuntos. Por ejemplo, si se utiliza como medida de bondad de ajuste el coeficiente de determinación, R^2, y a priori se dispone de diez variables regresoras, si algunas de ellas presentan un coeficiente de correlación lineal simple muy bajo con la variable *Y*, es probable que algunas de ellas se puedan desechar. Por ejemplo, si x_1 tiene un R^2 mayor que el modelo con las variables x_6 y x_7, entonces el subconjunto (x_1, x_i) tiene un R^2 mayor que el subconjunto (x_6, x_7). En base a esta idea está el "*branch-and-bound algorithm*" de Furnival y Wilson que obtiene buenos resultados para obtener implícitamente todas las posibles regresiones e identificar el

mejor subconjunto de un determinado número de variables regresoras según un criterio de ajuste prefijado.

Para decidir entre dos o más subconjuntos de variables regresoras en el estudio de un modelo de regresión múltiple es interesante disponer de medidas que midan la bondad del ajuste del modelo construido. Se supone que el número de variables explicativas que puede haber en el modelo es *k*, el número de observaciones es *n* y, si se ajusta un modelo de regresión lineal con *i* variables, el número de parámetros del modelo es *i*+1. Entonces se definen las siguientes medidas de bondad de ajuste:

- Coeficiente de determinación, R^2, definido como

$$R^2 = \frac{scR}{scG} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \quad \text{(a.50)}$$

  Este criterio aumenta al ir introduciendo nuevas variables en el modelo. Sea denota R_j^2, $j = 1,...,k$, el máximo valor posible de R^2 cuando en el modelo hay *j* variables explicativas, se verifica $R_{j-1}^2 \leq R_j^2$, (R_j^2 es monótona creciente) y las diferencias $R_j^2 - R_{j-1}^2$ decrecen. En base a esto, un criterio sencillo sería considerar un número pequeño Δy elegir el modelo con *j* más pequeño y tal que $R_k^2 - R_j^2 < \Delta$ (R_k^2 es el coeficiente de determinación del modelo con las *k* variables regresoras). Este criterio tiene el inconveniente de no tener en cuenta el número de variables regresoras. Tiende a sobreajustar y utilizar demasiadas variables regresoras.

- Coeficiente de determinación corregido, $\bar{R}^2$, esta medida de bondad de ajuste evita el problema de la medida anterior. Se define como

$$\bar{R}^2 = 1 - \frac{\hat{s}_R^2}{\hat{s}_Y^2} = 1 - (1 - R^2)\frac{n-1}{n-(k+1)}. \quad \text{(a.51)}$$

  Por tanto, $\bar{R}^2 \leq R^2$, y el coeficiente $\bar{R}^2$ tiene en cuenta el número de variables regresoras y no tiene porque crecer al introducir nuevas variables regresoras. Se denota $\bar{R}_j^2$ al mayor valor de $\bar{R}^2$ para el modelo de *j* variables, entonces un buen criterio sería elegir el subconjunto de *j* variables que maximiza este coeficiente, $\bar{R}_j^2$.

- Varianza residual, $\hat{s}_R^2$. Se ha definido $\hat{s}_R^2$ como

$$\hat{s}_R^2 = \frac{1}{n-(k+1)}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-(k+1)}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = scmR, \quad \text{(a.52)}$$

donde *scmR* (Mean Square Error) es la media de los errores al cuadrado. Un buen criterio de selección del subconjunto de variables es elegir el subconjunto de *j* variables que minimiza el valor $scmR_j$, siendo ésta la varianza residual obtenida con el modelo de *j* variables.

Teniendo en cuenta que

$$\bar{R}^2 = 1 - \frac{1}{\hat{\sigma}_Y^2}\, scmR, \quad \text{(a.53)}$$

se deduce que

$$\bar{R}_k^2 > \bar{R}_j^2 \Leftrightarrow scmR_k < scmR_j, \quad \text{(a.54)}$$

por tanto, el criterio de minimizar la varianza residual es equivalente al criterio de maximizar el coeficiente de determinación corregido.

- El estadístico C_p de Mallows. Los criterios anteriores se basan en el *scmR*, pero también es interesante tener en cuenta el sesgo en la selección del modelo ya que si se omite una variable regresora importante los estimadores de los coeficientes de regresión son sesgados y los criterios anteriores pueden elegir un modelo que tenga sesgo grande aunque su *scmR* sea pequeño. Un criterio que tenga en cuenta el sesgo ayudará a elegir el modelo adecuadamente. Con este objetivo surge el estadístico C_p de Mallows definido como,

$$C_p = p + (n-p)\left(\frac{\hat{s}_R^2(p) - \hat{s}_R^2}{\hat{s}_R^2}\right) \quad \text{(a.55)}$$

donde *p* es el número de parámetros del modelo (en un modelo de regresión lineal múltiple $p = j + 1$, con *j* el número de variables regresoras), $\hat{s}_R^2$ es la varianza del modelo con todas las variables y $\hat{s}_R^2(p)$ es la varianza residual al ajustar el modelo con $j=p-1$ variables regresoras.

Para interpretar este estadístico, se define el error cuadrático medio de predicción (*ECMP*) para los puntos observados cuando se utiliza un modelo con *p* parámetros como

$$ECMP_p = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{y}_{p,i} - m_{p,i})^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{y}_{p,i} - E(\hat{y}_{p,i}) + E(\hat{y}_{p,i}) - m_{p,i})^2$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n} Var(\hat{y}_{p,i}) + Sesgo^2(\hat{y}_{p,i}), \quad \text{(a.56)}$$

donde $\hat{y}_{p,i}$ es la predicción cuando se utiliza el modelo con *p* parámetros y $m_{p,i} = E(Y/\vec{x}_{p,i})$.

Siendo un buen criterio de selección del modelo el de elegir el modelo que tenga el $ECMP_p$ mínimo. Este criterio es equivalente a minimizar el estadístico C_p de Mallows.

Además puede probarse que en los modelos sin sesgo $C_p = p$. Por tanto, aquellos subconjuntos de j variables regresoras que tengan un $C_p \simeq p = j + 1$, son "buenos". Normalmente se construye una gráfica de C_p para los diferentes subconjuntos que se quieren analizar frente a p. Y se consideran buenos los subconjuntos que tienen C_p pequeño y además están por debajo de la diagonal $C_p = p$.

En la Figura a.20 se puede observar el C_p para dos subconjuntos de variables regresoras y se observa que el subconjunto A tiene un sesgo mucho mayor que el del subconjunto B, pero éste tiene menor C_p.

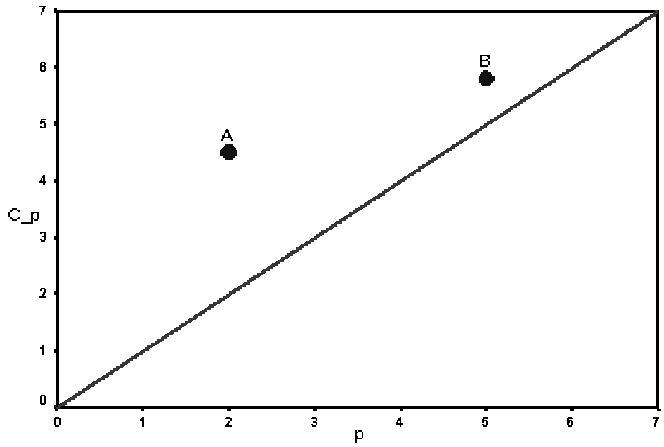

Fig. a.20. Gráfico de dos variables regresoras

a.9. Reseña teórica 9

Analicemos a continuación la estimación por mínimos cuadrados generalizados.

En un modelo de regresión lineal se supone que la matriz de varianzas-covarianzas de los errores es de la forma

$$E\left(\vec{\varepsilon}\vec{\varepsilon}^{t}\right) = \sigma^2 \mathbf{I}_n, \quad \text{(a.57)}$$

siendo $\mathbf{I}_n$ la matriz identidad de orden *n*. Si no se verifica la hipótesis de homocedasticidad, o la de independencia, o ambas, entonces la matriz de varianzas-covarianzas tiene la forma general

$$E\left(\vec{\varepsilon}\vec{\varepsilon}^t\right) = \sigma^2\Psi, \quad \text{(a.58)}$$

siendo Ψ una matriz simétrica, definida positiva de orden $n \times n$. En este caso, se puede calcular el estimador de $\vec{\alpha}$ por el método de mínimos cuadrados generalizados. Este método se desarrolla en dos etapas: en una primera etapa se transforma el modelo de regresión original

$$\vec{\mathbf{Y}} = \mathbf{X}\vec{\alpha} + \vec{\varepsilon}. \quad \text{(a.59)}$$

Para ello y por ser Ψ una matriz simétrica, definida positiva, existe una matriz cuadrada **P** tal que

$$\mathbf{P}\Psi\mathbf{P}^t = \mathbf{I}_n \Rightarrow \Psi = \mathbf{P}^{-1}\left(\mathbf{P}^t\right)^{-1} \Rightarrow \Psi^{-1} = \mathbf{P}^t\mathbf{P}, \quad \text{(a.60)}$$

esta matriz Ψ no tiene porque ser única, pero si existe. Multiplicando por **P** la ecuación de regresión se obtiene

$$\mathbf{P}\vec{\mathbf{Y}} = \mathbf{P}\mathbf{X}\vec{\alpha} + \mathbf{P}\vec{\varepsilon}. \quad \text{(a.61)}$$

Denominando $\vec{\mathbf{Y}}^* = \mathbf{P}\vec{\mathbf{Y}}$, $\mathbf{X}^* = \mathbf{P}\mathbf{X}$ y $\vec{\varepsilon}^* = \mathbf{P}\vec{\varepsilon}$, se obtiene la ecuación de regresión

$$\vec{\mathbf{Y}}^* = \mathbf{X}^*\vec{\alpha} + \vec{\varepsilon}^*, \quad \text{(a.62)}$$

y los errores del modelo verifican

$$E\left(\vec{\varepsilon}^*\,\vec{\varepsilon}^{*t}\right) = \mathbf{P}E\left(\vec{\varepsilon}\,\vec{\varepsilon}^t\right)\mathbf{P}^t = \sigma^2\mathbf{P}\Psi\mathbf{P}^t = \sigma^2\mathbf{I}_n, \quad \text{(a.63)}$$

por tanto los errores son incorrelados y homocedásticos. Ahora se puede aplicar el método de mínimos cuadrados ordinarios a estos datos transformados $(\mathbf{PX}, \mathbf{P}\vec{\mathbf{Y}})$ para obtener el estimador

$$\hat{\alpha}_G = \left((\mathbf{PX})^t\mathbf{PX}\right)^{-1}(\mathbf{PX})^t\mathbf{PY} = \left(\mathbf{X}^t\mathbf{P}^t\mathbf{PX}\right)^{-1}\mathbf{X}^t\mathbf{P}^t\mathbf{PY} = \left(\mathbf{X}^t\Psi^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\Psi^{-1}\mathbf{Y}. \quad \text{(a.64)}$$

Por el Teorema de Gauss-Markov, este estimador $\hat{\alpha}_G$ es el mejor estimador lineal insesgado. En la práctica, la matriz **P**, aunque existe, es desconocida y es necesario estimarla ($\hat{\mathbf{P}}$) a partir de las observaciones, obteniendo el estimador

$$\hat{\alpha}_F = \left(\mathbf{X}^t\hat{\mathbf{P}}^t\hat{\mathbf{P}}\mathbf{X}\right)^{-1}\mathbf{X}^t\hat{\mathbf{P}}^t\hat{\mathbf{P}}\mathbf{Y} = \left(\mathbf{X}^t\hat{\Psi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\hat{\Psi}^{-1}\mathbf{Y}. \quad \text{(a.65)}$$

A continuación se exponen dos situaciones comunes en las que se puede aplicar este método de estimación.

- *Heterocedasticidad* .Si las observaciones son independientes pero heterocedásticas entonces la matriz de varianzas-covarianzas viene dada por

$$E\left(\vec{\varepsilon}\vec{\varepsilon}^{t}\right)=\begin{pmatrix}\sigma_1^2 & 0 & \cdots & 0\\ 0 & \sigma_2^2 & \ddots & \vdots\\ \vdots & \ddots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma_n^2\end{pmatrix}$$

Y la matriz **P**

$$\mathbf{P}=\begin{pmatrix}\frac{1}{\sigma_1} & 0 & \cdots & 0\\ 0 & \frac{1}{\sigma_2} & \ddots & \vdots\\ \vdots & \ddots & \ddots & \vdots\\ 0 & 0 & \cdots & \frac{1}{\sigma_n}\end{pmatrix}$$

En este caso los datos transformados son

$$\vec{\mathbf{Y}}^{*}=\mathbf{P}\vec{\mathbf{Y}}=\begin{pmatrix}\frac{1}{\sigma_1} & 0 & \cdots & 0\\ 0 & \frac{1}{\sigma_2} & \ddots & \vdots\\ \vdots & \ddots & \ddots & \vdots\\ 0 & 0 & \cdots & \frac{1}{\sigma_n}\end{pmatrix}\begin{pmatrix}Y_1\\ \vdots\\ \vdots\\ Y_n\end{pmatrix}=\begin{pmatrix}\frac{Y_1}{\sigma_1}\\ \vdots\\ \vdots\\ \frac{Y_n}{\sigma_n}\end{pmatrix}$$

$$\mathbf{X}^{*}=\mathbf{P}\mathbf{X}=\begin{pmatrix}\frac{1}{\sigma_1} & 0 & \cdots & 0\\ 0 & \frac{1}{\sigma_2} & \ddots & \vdots\\ \vdots & \ddots & \ddots & \vdots\\ 0 & 0 & \cdots & \frac{1}{\sigma_n}\end{pmatrix}\begin{pmatrix}\vec{\mathbf{x}}_{1\cdot}\\ \vdots\\ \vdots\\ \vec{\mathbf{x}}_{n\cdot}\end{pmatrix}=\begin{pmatrix}\frac{\vec{\mathbf{x}}_{1\cdot}}{\sigma_1}\\ \vdots\\ \vdots\\ \frac{\vec{\mathbf{x}}_{n\cdot}}{\sigma_1}\end{pmatrix} \quad \text{(a.67)}$$

Esto equivale a trabajar con el modelo transformado

$$\frac{Y_i}{\sigma_i}=\frac{\vec{\mathbf{x}}_{i\cdot}}{\sigma_i}\,\vec{\alpha}+\frac{\varepsilon_i}{\sigma_i},\quad i=1,\ldots,n. \quad \text{(a.68)}$$

Sobre este modelo se aplica ahora el método de mínimos cuadrados ordinarios. En particular, si se trabaja con el modelo de regresión lineal se obtiene el siguiente estimador del coeficiente de regresión (α_1)

$$\hat{\alpha}_{1,G}=\frac{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\left(x_i-\bar{x}\right)^2}. \quad \text{(a.69)}$$

Este estimador se denomina estimador por mínimos cuadrados ponderados y es un caso particular del estimador por mínimos cuadrados generalizados. En la práctica,

para utilizar este estimador hay que calcular estimadores de los parámetros σ_1^2,...,σ_n^2 , lo que puede hacerse por uno de los siguientes métodos:

- Suponer que la varianza se ajusta a una función

$$\sigma_i^2 = g(\vec{\mathbf{x}}_{i\cdot}), \quad i = 1, \ldots, n. \qquad \text{(a.70)}$$

  y estimar la función *g*.

- Hacer grupos en las observaciones (en el orden en que se han recogido) normalmente del mismo tamaño *k* y suponer que en cada grupo la varianza es constante. Entonces se estima la varianza en cada grupo a partir de las observaciones del grupo. Una forma de conseguir esto es ajustar el modelo de regresión por mínimos cuadrados ordinarios a las observaciones originales y a partir de los residuos de este modelo obtener los estimadores de la varianza en cada grupo.

- Observaciones dependientes. Si las observaciones son homocedásticas pero dependientes entonces la matriz de varianzas-covarianzas es de la forma general

$$E\left(\vec{\varepsilon}\vec{\varepsilon}^t\right) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \ddots & \rho_{n-2} \\ \vdots & \ddots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{pmatrix}, \qquad \text{(a.71)}$$

En la mayoría de las situaciones la estructura de dependencia de los errores puede ajustarse a un modelo paramétrico. Un modelo sencillo y muy utilizado es el modelo *AR*(1), (modelo autorregresivo de orden uno). En este caso se verifica que los errores siguen la ecuación

$$\varepsilon_i = \rho\varepsilon_{i-1} + a_i, \quad i = 1, \ldots, n, \qquad \text{(a.72)}$$

siendo ρ la autocorrelación de orden 1 del proceso ε_t, por tanto, $|\rho| < 1$, y a_t es una sucesión de variables aleatorias independientes e igualmente distribuidas.

En este caso, la matriz de varianzas-covarianzas es

$$E\left(\vec{\varepsilon}\vec{\varepsilon}^t\right) = \sigma^2\Psi = \sigma^2 \frac{1}{1-\rho^2} \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \ddots & \rho^{n-2} \\ \vdots & \ddots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{pmatrix}, \qquad \text{(a.73)}$$

la matriz **P** de transformación es

$$\mathbf{P} = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{pmatrix},$$

y la matriz Ψ^{-1} es

$$\Psi^{-1} = \begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{pmatrix},$$

Utilizando esta matriz se obtiene el estimador por mínimos cuadrados generalizados

$$\hat{\alpha}_G = \left(\mathbf{X}^t \Psi^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^t \Psi^{-1} \mathbf{Y}. \quad \text{(a.74)}$$

Nuevamente, en la práctica, Ψ^{-1} es desconocido y se tiene que estimar. Por la forma de la matriz Ψ^{-1}, es suficiente con estimar el parámetro ρ y sustituir en la matriz. Para estimar ρ, puede utilizarse el siguiente procedimiento: ajustar a los datos el modelo de regresión lineal por mínimos cuadrados ordinarios y calcular los residuos mínimo cuadráticos

$$e_i = Y_i - \vec{\mathbf{x}}_i \hat{\alpha}_{MCO}, \quad i = 1, \ldots, n. \quad \text{(a.75)}$$

A partir de estos residuos se obtiene el siguiente estimador de ρ,

$$\hat{\rho} = \frac{\sum_{i=1}^{n-1} e_i e_{i+1}}{\sum_{i=1}^{n} e_i^2}, \quad \text{(a.76)}$$

sustituyendo ρ por $\hat{\rho}$ en la matriz Ψ^{-1} se obtiene la matriz estimada $\hat{\Psi}^{-1}$, a partir de la cual se obtiene el estimador

$$\hat{\alpha}_F = \left(\mathbf{X}^t \hat{\Psi}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^t \hat{\Psi}^{-1} \mathbf{Y}. \quad \text{(a.77)}$$

Siguiendo este procedimiento se puede obtener el siguiente estimador iterativo:

- Paso 1. Se utiliza el estimador $\hat{\alpha}_F$ para obtener nuevos residuos e_i'.
- Paso 2. De estos residuos se obtiene un nuevo estimador $\hat{\rho}'$.
- Paso 3. Utilizando $\hat{\rho}'$ se calcula un nuevo estimador $\hat{\alpha}_F'$.

Se continúa el proceso de forma iterativa (volver al Paso 1) hasta obtener la convergencia del estimador $\hat{\alpha}_F$ (*estimador iterativo de Cochran y Orcutt* (1949)).

En este problema también se pueden considerar otros estimadores del parámetro ρ o modelos de dependencia más complejos que dependen de un número mayor de parámetros.

Otra forma de análisis es con la estimación robusta.

Cuando existe evidencia existen una o varias observaciones heterogéneas que influyen en la estimación del modelo, la regresión robusta es una alternativa a la regresión por mínimos cuadrados ordinarios. La idea básica es calcular el estimador $\hat{\alpha}_R$ que minimiza la siguiente función

$$\Psi(\vec{\alpha}) = \sum_{i=1}^{n} \omega(e_i)\, e_i^2, \qquad \text{(a.78)}$$

donde $\omega(\cdot)$ es una función de ponderación que se introduce para reducir (e incluso eliminar) el efecto de los residuos altos. Por tanto se definen los pesos $\omega(e_i)$ de forma que tomen valores pequeños en los residuos e_i "grandes". Para aplicar esta definición es necesario conocer los residuos e_i. Este razonamiento conduce al siguiente algoritmo iterativo análogo al descrito para el método de mínimos cuadrados generalizados:

- Etapa 1. Calcular un estimador inicial (por ejemplo, el estimador por mínimos cuadrados ordinarios) $\hat{\alpha}^{(0)} = \hat{\alpha}_{MCO}$ de los parámetros del modelo, a partir del cual se obtienen los residuos iniciales, $e_i^{(0)}$

$$e_i^{(0)} = Y_i - \vec{\mathbf{x}}_i.\hat{\alpha}_{(0)}, \quad i = 1, \ldots, n. \qquad \text{(a.79)}$$

- Etapa 2. Se define una función de ponderación "razonable". Por ejemplo, la función de Huber de la Figura a.21.

$$\omega_i\left(e_i^{(0)}\right) = \left\{ \begin{array}{ll} 1/2 & \text{si } \left|r_i^{(0)}\right| < C \\ \left|\dfrac{C}{r_i^{(0)}}\right| - \dfrac{1}{2}\left|\dfrac{CC}{r_i^{(0)}}\right|^2 & \text{si } \left|r_i^{(0)}\right| > C \end{array} \right\} \qquad \text{(a.80)}$$

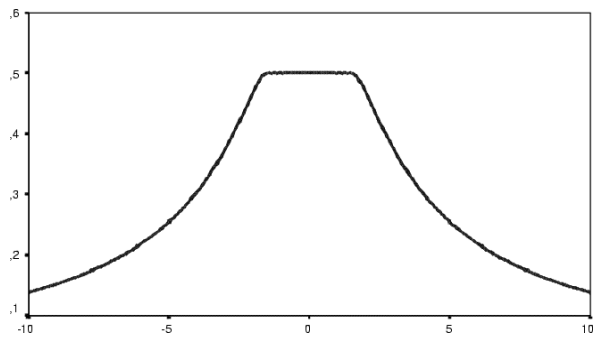

Fig. a.21. Función de Huber

donde $r_i^{(0)}$ es el residuo estandarizado asociado a $e_i^{(0)}$ y C es una constante. Si C toma valores pequeños (inferior a 1,5) entonces las observaciones con residuos relativamente grandes influyen poco en la estimación del modelo.

- Etapa 3. Se calcula el valor de $\vec{\alpha}$ que minimiza la función

$$\Psi(\vec{\alpha}) = \sum_{i=1}^{n} \omega\left(e_i^{(0)}\right) e_i^2. \quad \text{(a.81)}$$

A este vector se le denomina $\hat{\alpha}(1)'$.

En el modelo de regresión lineal simple, el estimador que se obtiene para el coeficiente de regresión lineal es

$$\hat{\alpha}_{1,(1)} = \frac{\sum_{i=1}^{n} \omega\left(e_i^{(0)}\right)(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} \omega\left(e_i^{(0)}\right)(x_i - \bar{x})^2} \quad \text{(a.82)}$$

- Etapa 4. Con los nuevos estimadores $\vec{\alpha}(1)$ se obtienen unos nuevos residuos $e_t^{(1)}$ y se continúa el proceso en la Etapa 2 hasta obtener la convergencia de las estimaciones que según Huber (1981) se consigue de forma rápida en la mayoría de las situaciones.

Una tercera forma es la de estimación polinómica.

En algunas situaciones para obtener un buen ajuste del modelo de regresión es necesario utilizar términos polinómicos. Si se trabaja con una única variable explicativa, el modelo polinómico de grado p obtenido a partir de la muestra $\{(x_i;y_i) : i = 1, \ldots, n\}$ tiene la forma

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \ldots + \alpha_p x_i^p + \varepsilon_i, \quad i = 1, \ldots, n. \quad \text{(a.83)}$$

Para el estudio de este modelo se utiliza la teoría de la regresión lineal múltiple, y basta con utilizar la relación

$$x^j = x_j, \qquad j = 1, \ldots, p, \quad \text{(a.84)}$$

esto es, la variable explicativa j-ésima es x^j.

Al ajustar un modelo polinómico se deben tener en cuenta los siguientes puntos:

- Si se utiliza un grado p muy alto se puede conseguir un ajuste muy bueno e incluso un ajuste exacto si $p = n$-1. Pero esta formulación no es correcta ya que se estaría ajustando el error del modelo. Por ello se recomienda utilizar un valor de p bajo ($p \leq 3$) y, en la mayoría de las situaciones, es suficiente

utilizar p=2. Es necesario hacer un análisis de los residuos para determinar si el ajuste es adecuado y se satisfacen las hipótesis básicas.

- Dado que las variables x_i y x_j (respectivamente x^i y x^j) son dependientes pueden surgir problemas de multicolinealidad. Para disminuir los efectos de este problema es conveniente trabajar con las variables centradas y, por tanto, utilizar el siguiente modelo de regresión

$$y_i = \sum_{j=0}^{p} \alpha_j \left(x_i - \bar{x}\right)^j + \varepsilon_i, \quad i = 1, \ldots, n. \tag{a.85}$$

- Si en el gráfico de la nube de puntos se observa que hay indicios de periodicidad (configuración cíclica) puede ser conveniente utilizar términos trigonométricos y ajustar un modelo de la forma

$$y_i = \sum_{j=0}^{p} \alpha_j x_i^j + \sum_{h=1}^{\lambda} \left(\gamma_h \sin\left(hx_i\right) + \delta_h \cos\left(hx_i\right)\right) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{a.86}$$

donde p y λ son valores a determinar. Una ventaja de los términos trigonométricos es que $\sin(hx_i)$ y $\cos(hx_i)$ son ortogonales si los x_i están equiespaciados.

El modelo polinómico con dos variables explicativas de grado dos tiene la forma

$$y_t = \alpha_0 + \alpha_1 x_{1t} + \alpha_2 x_{2t} + \alpha_{12} x_{1t} x_{2t} + \alpha_{11} x_{1t}^2 + \alpha_{22} x_{2t}^2 + \varepsilon_t, \quad i = 1, \ldots, n, \tag{a.87}$$

donde además de los términos cuadráticos hay un término de interacción de las dos variables explicativas $(x_{1t}x_{2t})$. Este modelo se conoce con el nombre de superficie respuesta y es muy utilizado en diseño de experimentos y control de calidad industrial.

Anexo B

b.1. Ejemplo 1

Se exponen varias nubes de observaciones y el ajuste lineal obtenido.

- En la Figura b.1 existe una dependencia funcional lineal, las observaciones están sobre la recta de regresión. $r = R^2 = 1$, recta de regresión: $y = x$.

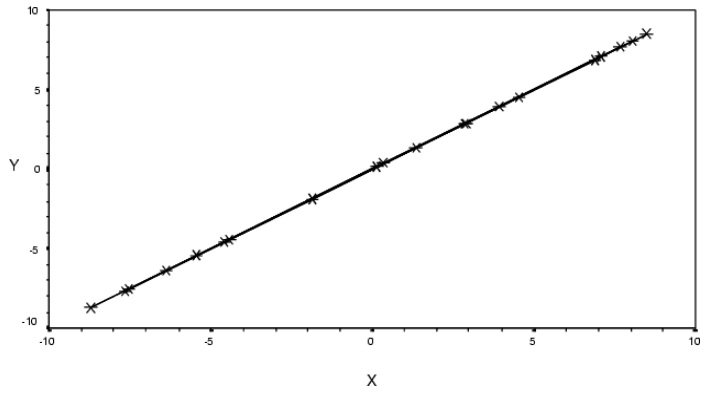

Fig. b.1. Existencia de dependencia funcional lineal

- En la Figura b.2 la relación lineal entre las variables es muy pequeña y no parece que exista otro tipo de relación entre ellas, la nube de puntos indica que las variables son "casi" independientes.

  $r = 0{,}192$, $R2 = 0{,}037$, recta de regresión: $y = 6{,}317 + 0{,}086x$.

  Contraste de regresión: $_R = 0{,}687 \in F_{1,18} \Rightarrow$ *p-valor* = 0,418. Se acepta la no influencia de la variable regresora en *Y*.

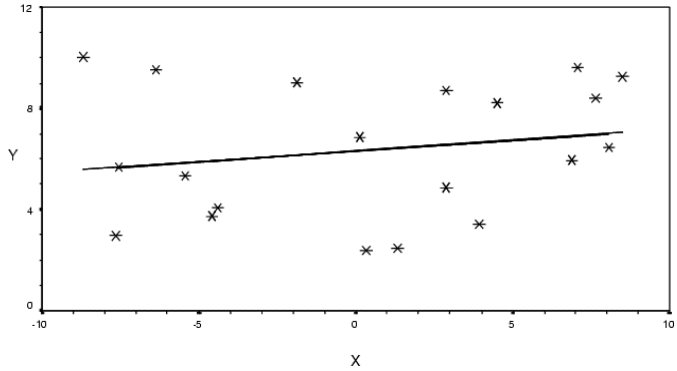

Fig. b.2. Relación lineal entre variables pequeña

- En la Figura b.3 existe una dependencia funcional entre las observaciones pero no de tipo lineal, por tanto la correlación es muy pequeña

  r = 0,391, $R2$ = 0,153, recta de regresión: y = 32,534 – 1,889x.

  Contraste de regresión: $_R$ = 3,252 $\in F_{1,18} \Rightarrow$ *p-valor* = 0,088. Se acepta que no existe relación lineal con α = 0,05. En base a la figura se debe de hacer un ajuste del tipo parabólico $Y = \alpha_0 + \alpha_1 x + \alpha_2 x2$.

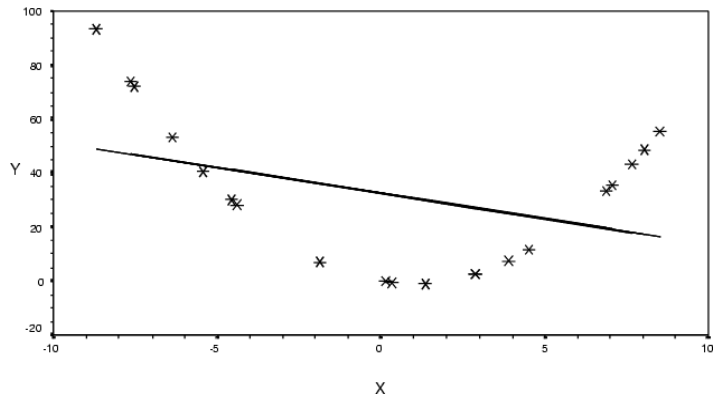

Fig. b.3. Dependencia entre variables no lineal

- En la Figura b.4 la nube de datos se ajusta razonablemente a una recta con pendiente positiva.

  r = 0,641, $R2$ = 0,410, recta de regresión: y = (-3, 963) + (-1,749)x.

  Contraste de regresión: $_R$ = 12,522 $\in F_{1,18} \Rightarrow$ *p-valor* = 0,002. Se rechaza la no influencia lineal de la variable x.

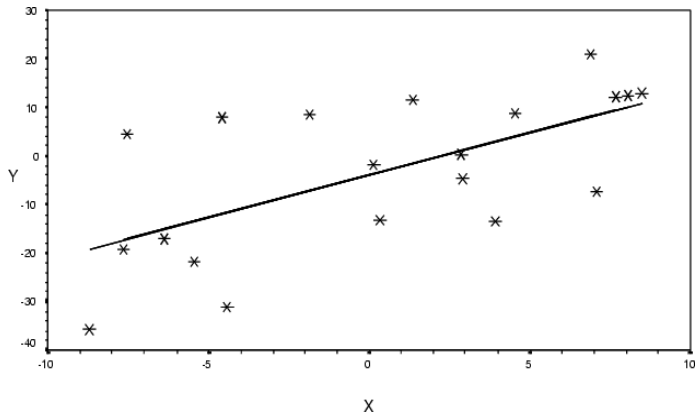

Fig. b.4. Ajuste razonable a una recta

- En la Figura b.5 existe una fuerte dependencia lineal negativa entre las dos variables y la correlación es muy alta (próxima a 1).
  r = 0,924, $R2$ = 0,846, recta de regresión: y = -2,528 – 2,267x
  Contraste de regresión: $_R$ = 105,193 $\in F_{1,18} \Rightarrow$ *p-valor* = 0,000. Se acepta la existencia de una relación lineal.

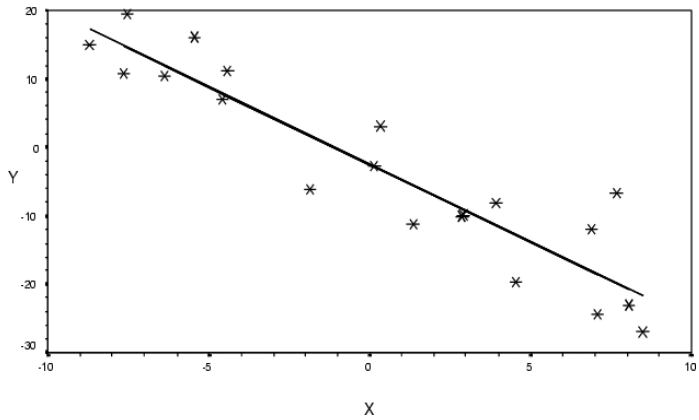

Fig. b.5. Fuerte dependencia lineal negativa

b.2. Ejemplo 2

El siguiente ejemplo gráfico puede ayudar a comprender el concepto de influencia de una observación. Considérese una muestra de 17 datos bidimensionales que siguen claramente el modelo de regresión lineal y tres datos adicionales (denotados A, B y C) que se separan claramente de la nube, tal como se ve en la Figura b.6.

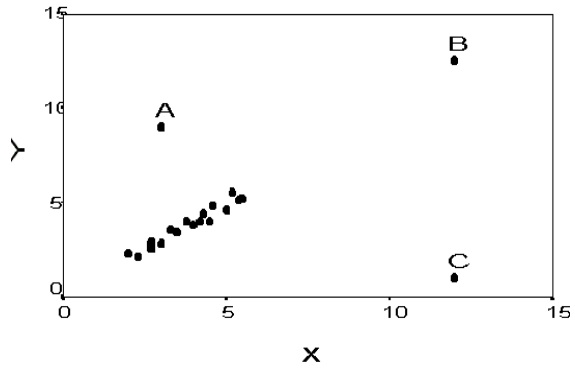

Fig. b.6. Nube con tres observaciones extremas (outliers).

A partir de esta muestra se calcula la recta de regresión de cuatro formas: primero, utilizando los 17 puntos y, en segundo lugar, utilizando los 17 puntos y uno de los tres puntos extremos. Los resultados obtenidos se presentan en la Tabla b.1.

	Recta de regresión	***R*2**	***R***
Sin valores extremos (17 ptos.)	$y = 0{,}242 + 0{,}923x$	0,945	0,972
Con A (18 ptos.)	$y = 1{,}534 + 0{,}672x$	0,212	0,460
Con B (18 ptos.)	$y = -0{,}177 + 1{,}034x$	0,986	0,993
Con C (18 ptos.)	$y = 3{,}876 - 0{,}048x$	0,008	0,087

Tabla b.1. Recta de regresión con puntos extremos

La gráfica de la nube de puntos, la recta calculada a partir de la muestra de 17 puntos y la recta calculada de los 17 puntos y la observación adicional (A, B o C) se representan en la Figura b.7, b.8 y b.9.

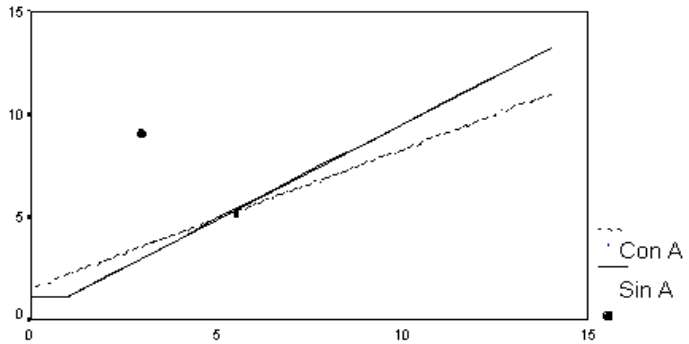

Fig. b.7. Influencia del punto A.

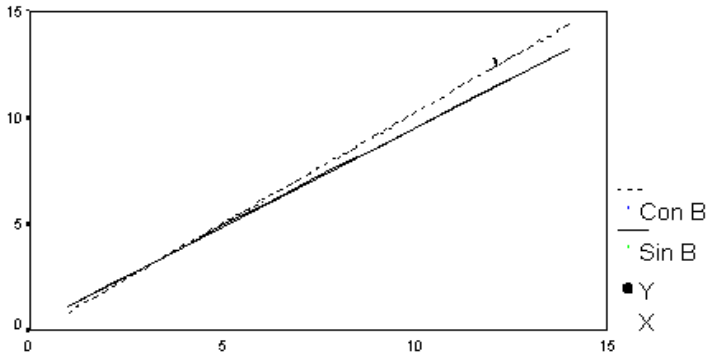

Fig. b.8. Influencia del punto B.

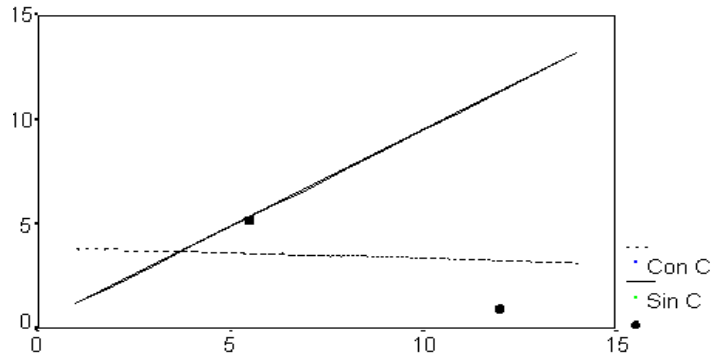

Fig. b.9. Influencia del punto C.

De esta tabla y gráficos se deduce:

- *El punto A,* no es un punto heterogéneo respecto a las *x* (x_A está cerca del centro $\bar{x}$) pero sí respecto a las *y*. Esto hace que sea un punto influyente en la estimación de la recta, ya que según se utilice o no el punto A en la estimación de la recta, ésta cambia de forma sustancial. Se dice que el punto A no es influyente "a priori" pero si es influyente a "posteriori". También se observa que y_A se separa claramente de la recta ajustada, calculada a partir de la muestra con A, por tanto, el punto A es atípico.
- *El punto B,* es un punto influyente "a priori" porque x_B está separado de $\bar{x}$, pero no influye (utilizarlo o no) en el cálculo de la recta de regresión, por tanto, el punto B no es influyente "a posteriori". Y como y_B está próximo a la recta ajustada $(\hat{y}_B)$ no es atípico.
- *El punto C,* es un punto influyente "a priori" e influyente "a posteriori", porque es un punto heterogéneo respecto a las *x* y a las *y*. Además se observa

que su influencia es muy grande, si se utiliza o no el punto C en el cálculo de la recta de regresión el resultado cambia totalmente. Por otra parte, y_C no se separa mucho de su predicción $(\hat{y}_C)$ cuando se utiliza la muestra con el punto C y, probablemente, no sea un dato atípico.

b.3. Ejemplo 3

En los siguientes ejemplos gráficos:

- Caso 1. Al omitir la variable atributo, la relación lineal obtenida entre la variable de interés y la variable explicativa no es correcta, como se ve en la Figura b.10.

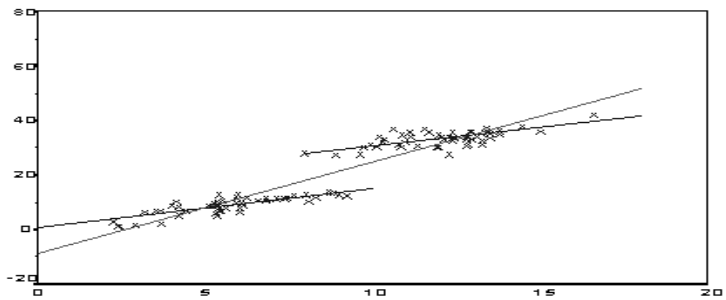

Fig. b.10. Efecto de omitir un atributo

- Caso 2. Al omitir la variable atributo en la Figura b.11, se oculta una relación lineal que si existe.

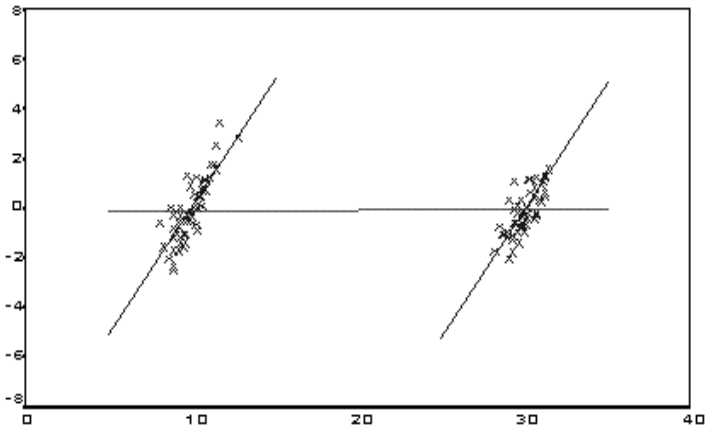

Fig. b.11. Efecto al omitir un atributo

- Caso 3. Al omitir la variable atributo en la Figura b.12, aparece una relación lineal que no existe.

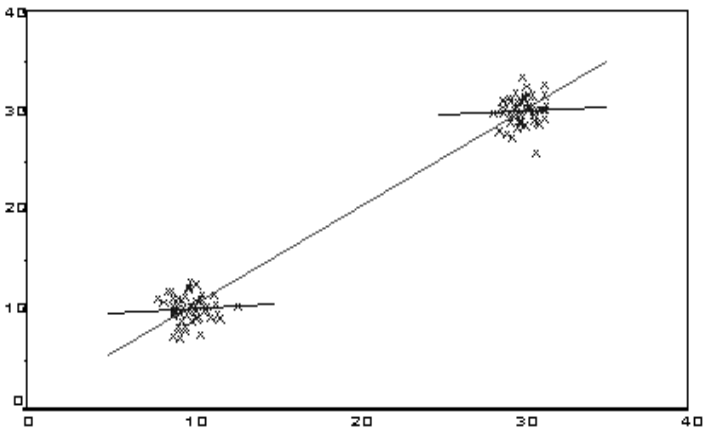

Fig. b.12. Efecto al omitir un atributo

Bibliografía

- Alonso J. (2001), “Estimación econométrica del tránsito vehicular y de la demanda del servicio de transporte ferroviario y automotor en el Gran La Plata”, Expte. Muni. La Plata 78.449/01, Argentina.
- Arranz P., Marhuenda F., Masciarelli E. (2005), “Estimación de cambios en el volumen de tránsito a causa del cobro de peaje en rutas de acceso a Córdoba”, XIV Congreso Argentino de Vialidad y Tránsito, Argentina.
- Arranz P., Masciarelli E., Marhuenda F. (2004), “Estudio econométrico y pronóstico del tránsito que pasa por casillas de peaje en concesiones viales de Argentina”, ISIT Universidad Nacional de Córdoba, Argentina.
- Asociación Argentina de Carreteras (2003), “Análisis de las concesiones de los corredores viales nacionales 1990-2003”, Memoria, Argentina.
- Auditoria General de la Nación (1995), “Estadísticas básicas de los corredores viales otorgados en concesión”, Documento Técnico Nº4, Argentina.
- AUFE (2004), “Informe Autopistas concesionadas”, Anuario, Argentina.
- AUOESTE (2004), “Estado contable anual”, Grupo Concesionario del Oeste, Argentina.
- AUSOL (2004), “Informe Económico Financiero”, Autopistas del Sol, Argentina.
- Banco Mundial (2002), “Ciudades en movimiento”, TWU-44.
- Barón López J. (1998), “Bioestadística: Métodos y Aplicaciones”, Universidad de Málaga, España.
- Cal y Mayor (2004), “Manual de planeación y diseño para la administración del tránsito y transporte”, Alcaldía Mayor de Bogotá, Colombia.
- Cal y Mayor R., Cárdenas J. (1995), “Ingeniería de tránsito, fundamentos y aplicaciones”, Alfaomega, México.

- Cevallos E. (2005), “Estudios económicos de las obras viales que conforman el programa norte grande, provincia de Tucumán”, DVPT Estudio I.EE. 156-8-1-(A), Argentina.
- CIMOP (2003), “Una visión estratégica del Transporte en la Argentina”, Argentina.
- Dirección de Señalización Luminosa (2002), “Metodología para el cálculo del Indice de Tránsito”, GCBA, Argentina.
- División Tránsito de la Dirección Nacional de Vialidad (2000), “Tránsito medio diario anual 98/99”, Argentina.
- EMVI (2005), “Regresión lineal”, Universidad de Málaga, España.
- Federal Highway Administration (1976), “Guide for manual of instructions for traffic surveys”, EEUU.
- Fernández Morales A., Lacomba Arias B. (2004), “Estadística Básica Aplicada”, Ágora Universidad, España.
- García R. (2001), “Curso básico de Statgraphics Plus 5.0”, SLADI Universidad Complutense de Madrid, España.
- Girardotti L. (2003), “Planeamiento del transporte”, Fac. de Ing. UBA, Argentina.
- Graham-Rowe D. (2005), “Smart traffic forecast offers seven-day predictions”, NewScientist, EEUU.
- Hara J. (1998), “Transportation system análisis and software application”, University of Osaka Prefecture, Japón.
- Hay W. (1998), “Ingeniería de transporte”, Limusa, México.
- Herz M., Galárraga J., Maldonado M. (2005), “Caracterización de errores de muestreo en censos de volumen y composición”, XIV Congreso Argentino de Vialidad y Tránsito, Argentina.
- Instituto Superior de Ingeniería de Transporte (1996), “Censos y proyecciones de tránsito de la red de accesos a Córdoba”, Universidad Nacional de Córdoba, Argentina.
- Instituto Superior de Ingeniería de Transporte (1996), “Red de Acceso a Córdoba; Capacidad y Nivel de Servicio para el tránsito actual y su predicción”, Universidad Nacional de Córdoba, Argentina.

- Khisty J. (1996), “An introduction of transportation engineering”, University of British Columbia, Canada.
- Leiva F. (2002), “El túnel subfluvial Paraná-Santa Fe, 30 años al servicio del tránsito”, Ente Interprovincial Túnel Subfluvial, Argentina.
- Lima Coimbra R. (2003), “La Región Pampeana”, UNCPBA, Argentina.
- Mix Ingeniería (2000), “Flujo vehicular en Bahía Blanca”, Documento Técnico, Argentina.
- Molinero L. (2002), “Construcción de modelos de regresión multivariantes”, Alce Ingeniería, España.
- Navin F. (1993), “The science, engineering and practice of land transport”, University of British Columbia, Canada.
- OCCOVI (2004), “Control de Gestión”, Informe Técnico, Argentina.
- Ortúzar, J. de D. (2000), “Modelos de demanda de transporte”, Universidad Católica de Chile, Alfaomega, Chile.
- Papacostas C. (1987), “Fundamentals of transportation engineering”, Prentice-Hall, EEUU.
- Pertegas Días S., Pita Fernández S. (2001), “La distribución normal”, Fisterra, España.
- Russel R., Taylor B. (2003), “Operations management. Focusing on quality and competitiveness”, Prentice Hall, EEUU.
- Sociedad Argentina de Ingeniería de Tránsito (1989), “2º Reunión de la Ingeniería de Tránsito”, Equitel S.A., Argentina.
- Spiegel M. (1988), “Estadística”, Mc Graw Hill, EEUU.
- Transportation Research Board (2000), “Highway Capacity Manual 2000”, National Research Council, EEUU.
- Vilar Fernández J. (2003), “Modelos Estadísticos aplicados”, Universidade da Coruña, España.
- Wahr C. (2003), “Vialidad II”, Universidad Técnica Federico Santa María, Chile.