Instituto Nacional de Estadística

OPOSICIONES AL CUERPO SUPERIOR DE ESTADÍSTICOS DEL ESTADO

BOE NÚM. 270, DE 12 DE OCTUBRE DE 2020, PÁG. 87165

Inferencia y Modelización Estadísticas

Grupo de Materias Comunes

INE
4 de julio de 2021

Índice general

Tema 8

Introducción a los modelos lineal y lineal generalizado. Tipos de modelos lineales. Modelos para datos experimentales y para datos observacionales. Componentes de un modelo lineal generalizado. Interpretación del término error en función del tipo de datos. Variables explicativas cuantitativas/cualitativas e interpretación de efectos. La esperanza condicionada y su modelización. Identificabilidad y estimabilidad.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía: A. Agresti (2015). *Foundations of Linear and Generalized Linear Models*. Wiley

Jeffrey M Wooldridge (2010). *Econometric analysis of cross section and panel data*. MIT press

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

8.1 Introducción a los modelos lineal y lineal generalizado

Los modelos lineales generales (GLM, del inglés *General Linear Model*) permiten construir una teoría de modelización unificada que engloba a la mayor parte de modelos de variables respuesta, tanto discretas como continuas.

El modelo lineal es un caso especial del modelo lineal generalizado. El primer paso consiste en definir modelos lineales generalizados y, al hacerlo, también se presenta el modelo lineal. Una de las ventajas de esta generalización de modelos lineales es que nos va a permitir modelar respuestas no normales (en el sentido de normalidad estadística), como datos categóricos y datos de recuento.

Especial importancia tienen los modelos que asumen una distribución binomial para la

variable respuesta. Esto tiene aplicaciones para datos binarios, como 'éxito' y 'fracaso' para posibles resultados en un ensayo médico, o 'estar a favor' y 'estar en contra' a posibles respuestas en una encuesta por muestreo.

Otra extensión de los GLM nos permitirá extender los modelos a respuestas multicategoría, asumiendo una distribución multinomial.

El modelo de regresión lineal ordinario utiliza la linealidad para describir la relación entre la media de la variable respuesta y un conjunto de variables explicativas, asumiendo la inferencia que la distribución de la respuesta es normal. Los GLM amplían los modelos de regresión lineal estándar para abarcar distribuciones de respuesta no normales y funciones no lineales de la media.

Una justificación para el hecho de haber introducido los GLM, es que anteriormente a su aparición, la tendencia era tener que transformar y para que tuviera aproximadamente una distribución condicionada normal con varianza constante. Por ejemplo, con datos de recuento que tienen una distribución de Poisson, la distribución está sesgada hacia la derecha con una varianza igual a la media, pero $\sqrt{y}$ tiene una distribución más cercana a la normal con una varianza aproximadamente igual a $\frac{1}{4}$. Para la mayoría de los datos, sin embargo, es difícil encontrar una transformación que proporcione tanto normalidad aproximada como varianza constante. La mejor transformación para lograr la normalidad suele diferir de la mejor transformación para lograr una varianza constante.

Con GLM, por el contrario, la elección de la función de enlace (*link function*) es independiente de la elección del componente aleatorio. Si una función de enlace es útil en el sentido de que un modelo lineal con las variables explicativas es plausible para ese enlace, no es necesario que también estabilice la varianza o produzca normalidad. Esto se debe a que el proceso de ajuste maximiza la probabilidad de la elección de la distribución de probabilidad para y, y esa elección no se limita a la normalidad.

Sea g una función, como por ejemplo la función logarítmica, que es una función de enlace en según el GLM planteado o una función transformada según el enfoque de datos transformados. Una ventaja de la formulación que permite el GLM es que los parámetros del modelo describen $g[E(y_i)]$, en lugar de $E[g(y_i)]$ como en el enfoque de datos transformados. Con el enfoque GLM, esos parámetros también describen los efectos de las variables explicativas en $E(y_i)$, después de aplicar la función inversa para g. Estos efectos suelen ser más relevantes que los efectos de las variables explicativas sobre $E[g(y_i)]$. Por ejemplo, si g es una función logarítmica, un GLM con $log[E(y_i)] = \beta_0 + \beta_1 x_{i1}$ se traduce en un modelo exponencial para la media, $E(yi) = exp(\beta_0 + \beta_1 x_{i1})$, pero el modelo de datos transformados $E[log(y_i)] = \beta_0 + \beta_1 x_{i1}$ no se traduce a información exacta sobre $E(y_i)$ o efecto de x_{i1} en $E(y_i)$. Además, a menudo la transformación preferida no se define en el límite del espacio muestral, como con la transformación logarítmica con un conteo o con proporción de cero.

8.2 Tipos de modelos lineales

La clase de los GLM incluye modelos para variables de respuesta continua. Los más importantes son los modelos lineales normales ordinarios. Tales modelos asumen una

distribución Normal[1] para el componente aleatorio,

$$y_i \sim N(\mu_i, \sigma); i = 1, \cdots, n$$

El parámetro natural para una distribución normal es la media. Entonces, la función de enlace canónica para un GLM Normal es el enlace de identidad (*indentity link*), y el GLM es simplemente un modelo lineal. En particular, los modelos estándar de regresión y análisis de varianza (ANOVA) son GLM que asumen un componente aleatorio Normal y utilizan como función de vínculo la identidad.

Para las respuestas binarias, los analistas suelen asumir una distribución binomial para el componente aleatorio de un GLM. A partir de la representación de su dispersión exponencial, el parámetro natural de la distribución binomial es el *log odds*, o comúnmente llamado modelo *logit*. La función de enlace canónico para los GLM binomiales es el *logit*, por lo que el modelo en sí se denomina regresión logística. Este es el modelo más importante para datos de respuesta binaria, y se ha venido utilizando para una amplia variedad de aplicaciones.

Los primeros usos fueron en estudios biomédicos, por ejemplo, para modelar los efectos del tabaquismo, el colesterol y la presión arterial sobre la presencia o ausencia de enfermedades cardíacas. Los últimos 25 años han tenido un uso sustancial en la investigación de las ciencias sociales para modelar opiniones (por ejemplo, favorecer u oponerse a la legalización del matrimonio entre personas del mismo sexo) y comportamientos, en aplicaciones de marketing para modelar las decisiones de los consumidores (una elección entre dos productos), y en finanzas, para modelar resultados relacionados con el crédito (por ejemplo, si la factura de una tarjeta de crédito se paga a tiempo).

Para las variables binarias, representamos por 1 y 0 los resultados 'éxito' y 'fracaso', y como 'favorecer' y 'oponerse' a una pregunta de una encuesta sobre la legalización del matrimonio entre personas del mismo sexo. Un ensayo de Bernoulli para la observación i tiene probabilidades

$$P(y_i = 1) = \pi_i$$

y

$$P(y_i = 0) = 1 - \pi_i$$

, para lo cual $\mu_i = \pi_i$. Este es el caso especial de la distribución binomial con el número de ensayos $n_i = 1$. El parámetro natural de la distribución binomial es $log(\frac{\mu_i}{(1-\mu_i)})$. Esta es la probabilidad de un ratio de la logística (*log odds*) de respuesta 1, el llamado *logit* de μ_i. El *logit* es la función de enlace canónica para componentes aleatorios binarios. Los GLM que utilizan el enlace *logit* tienen la forma:

$$log(\frac{\mu_i}{1-\mu_i}) = \sum_{j=1}^{p} \beta_j x_{ij}, i = 1, \ldots, n. \tag{8.1}$$

Son los denominados modelos de regresión logística o, a veces, simplemente modelos *logit*.

[1]Normal, comenzando en mayúsculas, hace referencia a lo largo del tema a que sigue una distribución estadística Normal

Los GLM para variables de respuesta con múltiples categorías asumen un componente aleatorio multinomial. Esto puede considerarse como una generalización de la regresión logística para variables de respuesta multinomial. Además, hay disponibles modelos separados para las variables de respuesta nominales y para las variables de respuesta ordinales.

En el caso de la respuesta nominal, el enfoque habitual es utilizar una ecuación logística binaria separada para cada par de categorías de respuesta. Un tipo importante de aplicación analiza los efectos de las variables explicativas en la elección de una persona entre un conjunto discreto de opciones, como la elección de comprar una determinada marca.

En el caso de un modelo para variables de respuesta ordinales, el enfoque es aplicar el logit o algún otro vínculo simultáneamente a todas las probabilidades de respuesta acumuladas, como para modelar si la importancia de la religión para una persona está por debajo o por encima de determinado valor en cierta escala ('sin importancia', 'poco importante', 'moderadamente importante', 'muy importante'). Una versión parsimoniosa del modelo utiliza los mismos parámetros de efecto para cada logit.

Denotamos c como el número de categorías de respuesta. Para el sujeto i, sea π_{ij} la probabilidad de respuesta en la categoría j, con $\sum_{j=1}^{c} \pi_{ij} = 1$. La elección de la categoría es el resultado de un único ensayo multinomial. Siendo $y_i = (y_{i1}, \ldots, y_{ic})$ el valor que representa la prueba multinomial para el sujeto $i, i = 1, \ldots, N$, donde $y_{ij} = 1$ cuando la respuesta corresponde a la categoría j y $y_{ij} = 0$ en caso contrario. Entonces, $\sum_j y_{ij} = 1$, y la distribución de probabilidad multinomial para ese sujeto es

$$p(y_{i1}, \ldots, y_{ic}) = \pi_{i1}^{y_{i1}} \ldots \pi_{ic}^{y_{ic}}$$

Muchas variables de respuesta tienen recuentos como posibles resultados. Ejemplos de ello son la cantidad de bebidas alcohólicas que tomó la semana anterior y la cantidad de dispositivos de su propiedad que pueden acceder a Internet (computadoras portátiles, teléfonos celulares inteligentes, tabletas, etc.). Los recuentos también ocurren como entradas en celdas de tablas de contingencia que clasifican de forma cruzada variables categóricas, como el número de personas en una encuesta que son mujeres, con estudios universitarios y están de acuerdo en que los seres humanos son responsables del cambio climático.

La distribución de probabilidad más simple para los datos de recuento es la Poisson. El modelo loglineal, que utiliza una función de enlace del tipo logarítmico para conectar la media con el predictor lineal, es el más común. El modelo se puede adaptar para modelar una tasa cuando el recuento se basa en un índice como el espacio o el tiempo.

Los modelos de Poisson y relacionados como los multinomiales pueden ser útiles para las tablas de contingencia al permitir analizar la independencia condicionada y la asociación para una variable de respuesta categórica multivariada. Para la distribución de Poisson, la varianza debe ser igual a la media, y los datos a menudo muestran una mayor variabilidad que esta.

La distribución más simple para datos de recuento, colocando su masa en el conjunto de valores enteros no negativos, es la de Poisson, ya que sus probabilidades dependen

de un solo parámetro, la media $\mu > 0$. La distribución de Poisson se usa a menudo para conteos de eventos que ocurren aleatoriamente en el tiempo o en el espacio a un ritmo particular, cuando los resultados en regiones o periodos de tiempo separados son independientes. Por ejemplo, un fabricante de teléfonos móviles podría encontrar que la Poisson describe razonablemente bien el número de reclamaciones de garantía recibidas cada semana.

La distribución de Poisson tiene el parámetro natural $log(\mu_i)$, por lo que la función de enlace canónico es la log, $\eta_i = log(\mu_i)$. El modelo que utiliza esta función de enlace es

$$log(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}, i = 1, \cdots, n. \tag{8.2}$$

Podemos ver en la tabla 8.1 algunos tipos de GLM.

Tabla 8.1: Modelos lineales generales importantes para el análisis estadístico

Componente aleatorio	Función de enganche	Modelo
Normal	Identidad	Regresión Análisis de varianza
Familia exponencial	Cualquiera	Modelo lineal general
Binomial	Logit	Regresión logística
Multinomial	Logits Generalizados	Respuesta multinomial
Poisson	Log	Loglinear

8.3 Modelos para datos experimentales y para datos observacionales

Datos no experimentales son datos sobre individuos, empresas o segmentos de la economía que no son obtenidos por medio de experimentos controlados. A los datos no experimentales en ocasiones también se les llama datos retrospectivos o datos observacionales, para subrayar el hecho de que el investigador es recolector pasivo de los datos.

En las ciencias naturales los datos experimentales suelen ser obtenidos en el laboratorio, pero en las ciencias sociales son mucho más difíciles de obtener. Aunque es posible idear experimentos sociales, suele ser imposible, prohibitivamente caro o moralmente indeseable realizar la clase de experimentos controlados que serían necesarios para abordar problemas económicos.

Precisamente, la econometría se ha desarrollado como una disciplina distinta de la estadística matemática ya que se centra en los problemas propios del análisis de datos

económicos de naturaleza no experimental. Estos datos no experimentales son utilizados, generalmente, para contrastar una teoría económica o una relación relevante para la toma de decisiones empresariales o para el análisis de políticas públicas.

Para verlo con mayor claridad, consideremos una situación en la que la utilidad del experimento aleatorizado controlado es clara. Consideremos el ejemplo que nos proporcionan los estudios de utilización generalizada de un fármaco (tratamiento) como medicamento. La evidencia obtenida a través de pruebas experimentales en pacientes proporciona evidencia estadística convincente para el uso generalizado del fármaco. Estas pruebas experimentales se articulan fácilmente siguiendo las pautas generales de un experimento aleatorizado controlado. El experimento consiste en que a unos pacientes seleccionados de forma aleatoria se les administra el fármaco y a otros se les proporciona un placebo. Las diferencias experimentales entre unos y otros conformarán los datos para posteriormente realizar un análisis causal en términos estadísticos y así llegar a una conclusión. Resulta difícil realizar este tipo de experimentos en economía: ¿dónde encontramos individuos a los que aleatoriamente se les ha administrado algún 'tratamiento'? Pese a la manifiesta dificultad, es posible que existan circunstancias externas que hagan que parezca como si algunos individuos hubieran sido tratados por azar (aleatoriamente). Este tipo de situaciones de 'como si' aparecen en economía con mucha más frecuencia, y son la base de lo que se denomina cuasiexperimento.

Uno de los campos, pero no el único, en el que los cuasiexperimentos han proliferado ha sido la evaluación de programas económicos y sociales. El objeto de esta área es evaluar el efecto de un programa, de una decisión política, o en general de alguna otra intervención (tratamiento). Por ejemplo, preguntas estudiadas por la literatura han sido: ¿cuál es el efecto sobre los salarios de acudir a un programa de formación laboral? ¿Qué efecto tiene sobre el empleo de trabajadores de baja cualificación un aumento del salario mínimo? ¿Cuál es el efecto sobre un colectivo de interés de un cambio en la cuantía del subsidio de desempleo o de la duración del mismo? En economía las unidades de análisis no solo son individuos. En general las unidades de análisis son sujetos económicos: individuos, hogares, mercados, empresas, provincias, regiones o países. Los tratamientos pueden ser muy variados. Por ejemplo, programas de asistencia en búsqueda de empleo, programas educativos, normativa legal, medicamentos farmacéuticos, exposición medioambiental, uso de tecnologías, etcétera.

Las técnicas propias de los experimentos aleatorizados controlados tienen ya bastante reconocimiento en la literatura estadística especializada en las mismas. Este tipo de literatura utiliza un lenguaje distinto del presentado hasta el momento y por tanto es necesario familiarizarse con el mismo. Debido a la proximidad conceptual entre el experimento aleatorizado controlado y el cuasiexperimento, las herramientas técnicas de los experimentos pueden adaptarse con ciertos ajustes, a los cuasiexperimentos. Se hace entonces necesario conocer este tipo de herramientas estadísticas, que en realidad se pueden reinterpretar, con ciertos matices, en los términos de las herramientas de regresión.

8.4 Componentes de un modelo lineal generalizado

El GLM tiene tres componentes principales:

- Componente aleatorio: especifica la variable respuesta y y su distribución de probabilidad. Las observaciones $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ en esa distribución se tratan como independientes.
- Predictor lineal: para un vector de parámetro $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ y una matriz $\boldsymbol{X}$ del modelo $n \times p$ que contiene valores de p variables explicativas para las n observaciones, el predictor lineal es $X\boldsymbol{\beta}$.
- Función de enlace: esta es una función g aplicada a cada componente de $E(\boldsymbol{y})$ que la relaciona con el predictor lineal,

$$g[E(\boldsymbol{y})] = X\boldsymbol{\beta}. \tag{8.3}$$

A continuación, se presenta cada componente del GLM en detalle.

8.4.1 Componente aleatorio del GLM

El componente aleatorio del GLM consiste en una variable de respuesta y con observaciones independientes $(y_1, \cdots, y_n)$ que tienen densidad de probabilidad o función de masa para una distribución en la familia exponencial. Una de las propiedades interesantes de esta familia de distribuciones consiste en que $\sum_i y_i$ es un estadístico suficiente para su parámetro, y las condiciones de regularidad se satisfacen para las derivaciones de propiedades como el rendimiento óptimo para muestras grandes de los estimadores de máxima verosimilitud.

Al restringir los GLM a distribuciones de familias exponenciales, obtenemos expresiones generales para las ecuaciones de verosimilitud del modelo, las distribuciones asintóticas de los estimadores para los parámetros del modelo y un algoritmo para ajustar los modelos. Por ahora, basta decir que las distribuciones más comúnmente utilizadas en Estadística, como la Normal, Binomial y Poisson, son distribuciones familiares exponenciales.

8.4.2 Predictor lineal del GLM

Para la observación $i, i = 1, \cdots, n$, se tiene que x_{ij} denota el valor de la variable explicativa $x_j, j = 1, \cdots, p$. Y sea $x_i = (x_{i1}, \cdots, x_{ip})$. Por lo general, establecemos $x_{i1} = 1$ o dejamos que la primera variable tenga un índice 0 con $xi0 = 1$, por lo que sirve como el coeficiente de un término de intersección en el modelo. El predictor lineal de un GLM relaciona los parámetros η_i pertenecientes a $E(y_i)$ con las variables explicativas $x_1, \cdots, x_p$ a través de una combinación lineal de ellos,

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}, i = 1, \cdots, n. \tag{8.4}$$

El denominar a $\sum^p = \beta_j x_{ij}$ predictor lineal, viene a reflejar que esta expresión es lineal en los parámetros. Las propias variables explicativas pueden ser funciones no lineales de variables subyacentes, como un término de interacción (por ejemplo, $x_{i3} = x_{i1}x_{i2}$)

o un término cuadrático (por ejemplo, $x_{i2} = x_{i1}^2$). En forma matricial, expresamos el predictor lineal como

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$$

, donde $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)^T$, $\boldsymbol{\beta}$ es el vector columna de los $p \times 1$ parámetros del modelo, y $\boldsymbol{X}$ es la matriz $n \times p$ de los valores de las variables explicativas x_{ij}. La matriz $\boldsymbol{X}$ se llama *matriz del modelo*. En estudios experimentales, a menudo también es llamada *matriz de diseño*. Ésta tiene n filas, una para cada observación, y p columnas, una para cada parámetro en $\boldsymbol{\beta}$. En la práctica generalmente $p \leq n$, pues el objetivo siguiendo el principio de parsimonia en el modelo, es resumir los datos utilizando un número considerablemente menor de parámetros.

El GLM trata y_i como aleatorio y el $\boldsymbol{x_i}$ como fijo. Debido a esto, el predictor lineal a veces se denomina componente sistemático. En la práctica, $\boldsymbol{x_i}$ es en sí mismo a menudo aleatorio, como en las encuestas por muestreo y otros estudios de observación.

8.4.3 Función de enganche de un GLM

El tercer componente de un GLM es the la función de enganche (*link function* en inglés), la cual conecta el componente aleatorio con el predictor lineas. Sea $\mu_i = E(yi), i = 1, \ldots, n$. El GLM conecta η_i a μ_i por medio de $\eta_i = g(\mu_i)$, donde la función de enganche $g(.)$ es monótona y diferenciable. De este modo, g conecta μ_i a las variables explicativas a través de la siguiente fórmula:

$$g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}, i = 1, \cdots, n. \tag{8.5}$$

En la representación de una distribución cualquiera perteneciente a la familia exponencial hay un parámetro que podemos considerar su parámetro natural. Este parámetro es la media en el caso de la distribución Normal, el logaritmo de las probabilidades en una distribución Binomial y el logaritmo de la media en una distribución de Poisson. La función de enlace g que transforma μ_i en el parámetro natural se llama enlace canónico.

La función de enlace $g(\mu_i) = \mu_i$ se llama función de enlace identidad (*identity link function*). En ella se da la igualdad $\eta_i = \mu_i$. Un GLM que utiliza la función de enlace identidad se denomina modelo lineal. Éste compara el predictor lineal con la media misma. Este GLM es

$$\mu_i = \sum_{j=1}^{p} \beta_j x_{ij}, i = 1, \cdots, n$$

La versión estándar de este modelo, al que nos referimos como modelo lineal ordinario, asume que las observaciones tienen una varianza constante, llamada *homocedasticidad*. Una forma alternativa de expresar el modelo lineal ordinario es

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

donde el término de error ϵ_i tiene $E(\epsilon_i) = 0$ y $var(\epsilon_i) = \sigma^2, i = 1, \cdots, n$. Esto es natural para el vínculo de identidad y para las respuestas Normales, pero no para la mayoría de los GLM.

8.5 Interpretación del término error en función del tipo de datos

Una forma alternativa de expresar el modelo lineal ordinario es

$$y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

para un término de error $\boldsymbol{\epsilon}$ tal que $E(\boldsymbol{\epsilon}) = \mathbf{0}$ y una matriz de covarianza $\boldsymbol{V} = var(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$. Sin embargo, una estructura aditiva tan simple para el término de error no es natural para muchos GLM, pero sí para los modelos Normales y las versiones de variables latentes de algunos otros modelos y sus extensiones con múltiples componentes de error. Para ser coherentes con las fórmulas del GLM, normalmente expresaremos los modelos lineales en términos de $E(\boldsymbol{y})$.

El modelo lineal mixto para y_{ij} es

$$E(y_{ij}|\boldsymbol{u_i}) = \boldsymbol{x_{ij}}\boldsymbol{\beta} + \boldsymbol{z_{ij}}\boldsymbol{u_i}$$

o, como también puede expresarse

$$y_{ij} = \boldsymbol{x_{ij}}\boldsymbol{\beta} + \boldsymbol{z_{ij}}\boldsymbol{u_i} + \epsilon_{ij}$$

donde $\boldsymbol{\beta}$ es un vector de efectos fijos $p \times 1$ y $\boldsymbol{u_i} \sim N(\mathbf{0}, \sum_{\boldsymbol{u}})$ es un vector de efectos fijos $q \times 1$. Por lo general, asumimos que $\epsilon_{ij} \sim N(0, \sigma^2)$, lo que produce el modelo lineal Normal mixto. El modelo básico asume que $\boldsymbol{u_i}$ y ϵ_{ij} son independientes entre grupos (es decir, sobre i) y entre sí. Para empezar, también asumimos que ϵ_{ij} son independientes dentro de los conglomerados -*clusters*- (es decir, sobre j para cada i). El modelo para y_{ij} se descompone en un término $\boldsymbol{x_{ij}}\boldsymbol{\beta}$ para la media, un término $\boldsymbol{z_{ij}}\boldsymbol{u_i}$ para la variabilidad entre conglomerados y un término ϵ_{ij} para la variabilidad dentro de los conglomerados. Para $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{id})^T$, el modelo tiene la forma

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{u_i} + \boldsymbol{\epsilon_i}$$

Los modelos lineales mixtos que asumen respuestas condicionalmente independientes en un conglomerado, dado un efecto aleatorio, a veces son inadecuados, especialmente para datos longitudinales o espaciales. En su lugar, podemos permitir que el término de error ϵ_i para un clúster tenga componentes correlacionados.

Los GLM para datos binarios y datos de recuento se expresan en términos de $E(y_{ij}|u_i)$ y no tienen un término de error separado, a menos que trabajemos con la versión del modelo para la variable latente.

8.6 Variables explicativas cuantitativas/cualitativas e interpretación de efectos

Hasta ahora hemos aprendido que un GLM consta de un componente aleatorio que identifica la variable de respuesta y su distribución, un predictor lineal que especifica

las variables explicativas, y una función de enlace que las conecta. Ahora echamos un vistazo más de cerca a la forma del predictor lineal.

En un GLM las variables explicativas pueden ser:

- cuantitativas, como en los modelos de regresión lineal simple.
- factores cualitativos, como en los modelos de análisis de varianza (ANOVA).
- mixtas, como un término de interacción que es el producto de una variable explicativa cuantitativa y un factor cualitativo.

Por ejemplo, supongamos que la observación i mide el ingreso anual de un individuo y_i, el número de años de experiencia laboral x_{i1} y el género $x_{i2}(1 = mujer, 0 = hombre)$. El modelo lineal con predictor lineal

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} \tag{8.6}$$

siendo x_{i1} cuantitativa, x_{i2} cualitativa y $x_{i3} = x_{i1}x_{i2}$ mixta para un término de interacción. Como ilustra la Figura 8.1, este modelo corresponde a líneas rectas $\mu_i = \beta_0 + \beta_1 x_{i1}$ para hombres y $\mu_i = (\beta_0 + \beta_2) + (\beta 1 + \beta 3)x_{i1}$ para mujeres. Con un término de interacción que relaciona dos variables, el efecto de una variable cambia según el nivel del otro. Por ejemplo, con este modelo, el efecto de la experiencia laboral sobre el ingreso anual medio tiene una pendiente de β_1 para los hombres y $\beta_1 + \beta_3$ para las mujeres. El caso especial, $\beta_3 = 0$, de falta de interacción, corresponde a líneas paralelas que relacionan el ingreso medio con la experiencia laboral de mujeres y hombres. El caso especial adicional que también tiene $\beta_2 = 0$ corresponde a líneas idénticas para mujeres y hombres. Cuando usamos el modelo para comparar los ingresos medios de mujeres y hombres mientras se tiene en cuenta el número de años de experiencia laboral como una covariable, se denomina modelo de análisis de covarianza.

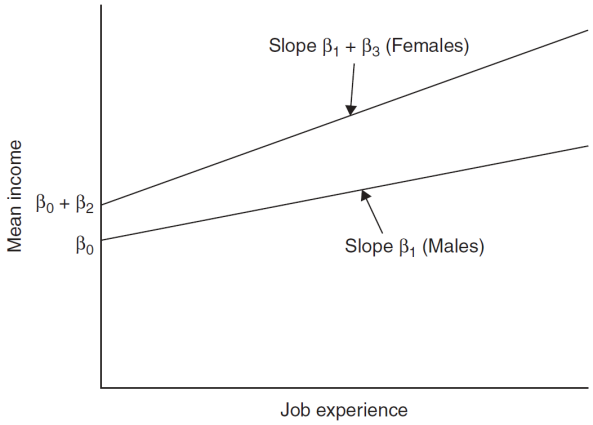

Figura 8.1: Representación de predictor lineal con variables explicativas cuantitativas y cualitativas

Una variable explicativa cuantitativa x está representada por un solo término βx en el predictor lineal y una sola columna en la matriz del modelo $\boldsymbol{X}$. Una variable explicativa cualitativa que tiene c categorías se puede representar mediante $c - 1$ variables

indicadoras y términos en el modelo lineal y $c-1$ columnas en la matriz del modelo $\boldsymbol{X}$.

Ejemplo 1. El software R utiliza por defecto la parametrización 'primera categoría de referencia' (*first-category-baseline*), que construye indicadores para las categorías $2, \ldots, c$. Los coeficientes de sus parámetros proporcionan contrastes con la categoría 1. Por ejemplo, supongamos que el estatus racial-étnico es una variable explicativa con $c = 3$ categorías ('negro', 'hispano' y 'blanco'). Un modelo que relacione el ingreso medio con el estatus racial-étnico podría ser

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

con $x_{i1} = 1$ para hispanos y 0 en caso contrario, $x_{i2} = 1$ para blancos y 0 en caso contrario, y $x_{i1} = x_{i2} = 0$ para negros. Entonces β_1 es la diferencia entre el ingreso medio de los hispanos y el ingreso medio de los negros, β_2 es la diferencia entre el ingreso medio de los blancos y el ingreso medio de los negros, y $\beta_1 - \beta_2$ es la diferencia entre el ingreso medio de los hispanos y el ingreso medio de los blancos. Algún otro software, como SAS, utiliza una parametrización predeterminada alternativa de 'última categoría de referencia' (*last-category-baseline*), que construye indicadores para las categorías $1, \ldots, c-1$. De este modo, sus parámetros proporcionan contrastes con la categoría c. Todas estas opciones posibles son equivalentes, ya que presentan el mismo ajuste del modelo.

La notación abreviada puede representar términos (variables y sus coeficientes) mediante símbolos utilizados para predictores lineales. Un efecto cuantitativo βx se denota con X, y un efecto cualitativo se denota con una letra cerca del comienzo del alfabeto, como A o B. Una interacción se representa por un producto de dichos términos, como $A \cdot B$ o $A \cdot X$. El período representa la formación de vectores de productos por componentes de las columnas constituyentes a partir de la matriz del modelo. El operador de cruce A^*B denota $A + B + A \cdot B$. El anidamiento de las categorías de B dentro de las categorías de A (por ejemplo, el factor A son los estados y el factor B son los condados dentro de esos estados) está representado por $A/B = A + AB$, o algunas veces por $A + B(A)$. Un término de intersección está representado por 1, pero generalmente se asume que está en el modelo a menos que se especifique lo contrario. La tabla 8.2 ilustra algunos tipos simples de predictores lineales y enumera los nombres de modelos lineales normales que igualan la media de la distribución de respuesta a ese predictor lineal.

Tabla 8.2: Tipos de predictores lineales para un modelo lineal Normal

Predictor lineal	Nombre del modelo
$X_1 + X_2 + X_3 + \ldots$	*Regresión múltiple*
A	*ANOVA de una vía*
$A + B$	*ANOVA de dos vías, sin interacción*
$A + B + AB$	*ANOVA de dos vías, con interacción*
$A + X$ or $A + X + AX$	*Análisis de covarianza*

8.6.1 Variables de intervalo, nominales y ordinales

Se dice que las variables cuantitativas se miden en una escala de intervalo, porque los intervalos numéricos separan los niveles de la escala. A veces se les llama variables de intervalo. Una variable cualitativa, representada en un modelo por un conjunto de variables indicadoras, tiene categorías que se tratan como desordenadas. Esta variable categórica se llama variable nominal. Por el contrario, una variable categórica cuyas categorías tienen un orden natural se denomina ordinal.

Ejemplo 2. La educación obtenida se puede medir con las categorías ('preparatoria', 'bachillerato', 'bachillerato' y 'posgrado'). Las variables explicativas ordinales pueden tratarse como cualitativas ignorando el orden y utilizando un conjunto de variables indicadoras. Alternativamente, pueden tratarse como cuantitativas asignando puntuaciones monótonas a las categorías y utilizando un solo término βx en el predictor lineal. Esto se hace a menudo cuando esperamos que $E(y)$ aumente progresivamente, o disminuya progresivamente, a medida que avanzamos en orden a través de esas categorías ordenadas.

8.6.2 Interpretación de los efectos

¿Cómo interpretamos los coeficientes β en los predictores lineales de GLM? Supongamos que la variable de respuesta es la puntuación de un estudiante universitario en la prueba de rendimiento en matemáticas y_i, y ajustamos el modelo lineal con $x_{i1} =$ *el número de años de educación matemática del estudiante* como variable explicativa, $\mu_i = \beta_0 + \beta_1 x_{i1}$. Dado que β_1 es la pendiente de una línea recta, podríamos decir: 'Si el modelo se mantiene, un aumento de un año en la educación matemática corresponde a un cambio de β_1 en la puntuación esperada de la prueba de rendimiento matemático'. Sin embargo, esto puede sugerir la conclusión causal inapropiada de que si un estudiante alcanza otro año de educación matemática, se espera que su puntuación en la prueba de rendimiento matemático cambie en β_1. Para llegar a tal conclusión de manera válida, necesitaríamos realizar un experimento que agregue un año de educación matemática para cada estudiante y luego observe los resultados. De lo contrario, una puntuación media más alta en la prueba a un nivel de educación matemática superior (si $\beta_1 > 0$) podría reflejar, al menos en parte, la correlación de varias otras variables con la puntuación de la prueba y el nivel de educación matemática, como los niveles educativos alcanzados por los padres, el CI del estudiante, el número de años de cursos de ciencias, etc.

Una interpretación adecuada sería la que sigue: si el modelo se mantiene, cuando comparamos la subpoblación de estudiantes que tienen una cierta cantidad de años de educación matemática con la subpoblación que tiene un año menos de educación matemática, la diferencia en las medias de sus calificaciones en las pruebas de rendimiento matemático es β_1.

Ahora, supongamos que se agrega al modelo $x_{i2} =$ *edad del estudiante* y $x_{i3} =$ *número de años de educación matemática de la madre*,

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

Dado que $\beta_1 = \frac{\partial \mu_i}{\partial x_{i1}}$, podríamos decir, 'el efecto de la cantidad de años de educación matemática en la puntuación promedio de la prueba de rendimiento matemático es igual a β_1, ajustando por la edad del estudiante y la educación matemática de la madre.'

El control de variables es posible en experimentos diseñados. Pero es antinatural y posiblemente inconsistente con los datos de muchos estudios observacionales imaginar el aumento de una variable explicativa mientras se mantienen todas las demás fijas. Por ejemplo, es probable que x_1 y x_2 estén correlacionadas positivamente, por lo que los aumentos en x_1 naturalmente tenderán a ocurrir con aumentos en x_2. En algunos conjuntos de datos es posible que ni siquiera se observe un rango de 1 unidad en una variable explicativa cuando las demás variables explicativas se mantengan constantes.

Una mejor interpretación sería la siguiente: 'La diferencia entre la puntuación promedio de la prueba de rendimiento matemático de una subpoblación de estudiantes que tienen una cierta cantidad de años de educación matemática y una subpoblación que tiene un año menos es igual a β_1, cuando ambas subpoblaciones tienen el mismo valor para $\beta_2 x_{i2} + \beta_3 x_{i3}$'. De forma más concisa, podríamos decir: 'El efecto del número de años de educación matemática en la puntuación media de la prueba de rendimiento matemático es igual a β_1, ajustando según la edad del estudiante y la educación matemática de la madre'. Cuando el modelo también tiene un factor cualitativo, como $x_{i4} =$ género (1 = mujer, 0 = hombre), entonces β_4 es la diferencia entre las puntuaciones medias de las pruebas de rendimiento en matemáticas para mujeres y hombres estudiantes, ajustando las otras variables explicativas del modelo.

El efecto β_1 en la ecuación con una única variable explicativa generalmente no es el mismo que β_1 en la ecuación con múltiples variables explicativas. Esto es debido a factores como la confusión. El efecto de x_1 en $E(y)$ generalmente será diferente si ignoramos otras variables que si ajustamos según ellas, especialmente en estudios observacionales que contienen 'variables acechantes' que están asociadas tanto con y como con x_1. Para resaltar tal distinción, a veces es útil usar una notación diferente para el modelo con múltiples variables explicativas.

$$\mu_i = \beta_0 + \beta y_{1 \cdot 23} x_{i1} + \beta y_{2 \cdot 13} x_{i2} + \beta y_{3 \cdot 12} x_{i3},$$

donde $\beta y_{j \cdot k\ell}$ denota el efecto de x_j sobre y tras haber ajustado por x_k y x_ℓ.

8.7 La esperanza condicionada y su modelización

La esperanza condicionada juega un papel clave en el análisis econométrico moderno. Aunque no siempre se establece explícitamente, el objetivo de la mayoría de los estudios econométricos aplicados consiste en estimar o probar hipótesis sobre el valor esperado de una variable, llamada variable explicada, variable dependiente o variable respuesta, y usualmente denotada y, condicionada ésta a un conjunto de variables explicativas, independientes, regresoras, variables de control o covariables, generalmente denotadas $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$.

Una parte sustancial de la investigación en metodología econométrica puede interpretarse como encontrar formas de estimar las esperanzas condicionadas en los numerosos

entornos en los que pueden surgir dentro de las aplicaciones económicas. De hecho, la mayoría de las veces nos interesan las esperanzas condicionadas que nos permiten inferir causalidad de una o más variables explicativas a la variable de respuesta.

Podemos estar interesados en el efecto de una variable w sobre el valor esperado de y, manteniendo fijo un vector de controles $\boldsymbol{c}$. La esperanza condicionada de interés es $E(y|w, \boldsymbol{c})$, que llamaremos esperanza condicionada estructural. Si podemos recopilar datos sobre y, w y $\boldsymbol{c}$ en una muestra aleatoria de la población subyacente de interés, entonces será bastante sencillo estimar $E(y|w, \boldsymbol{c})$, especialmente si estamos dispuestos a hacer una suposición sobre su forma funcional, en cuyo caso el efecto de w sobre $E(y|w, \boldsymbol{c})$, manteniendo $\boldsymbol{c}$ fijo, se puede estimar fácilmente.

Desafortunadamente, a menudo surgen complicaciones en la recopilación y análisis de datos económicos debido a la naturaleza no experimental de la economía. Las observaciones sobre variables económicas pueden contener errores de medición o, a veces, se las considera correctamente como el resultado de un proceso simultáneo. En otras ocasiones no podemos obtener una muestra aleatoria de la población, lo que nos impedirá estimar la esperanza condicionada. Quizás el problema más frecuente es que algunas variables que nos gustaría controlar (elementos de $\boldsymbol{c}$) no se pueden observar. En cada uno de estos casos existe una esperanza condicionada de interés, pero generalmente involucra variables para las cuales el econometrista no puede recolectar datos o requiere un experimento que no se puede realizar.

Bajo supuestos adicionales, a veces podremos recuperar la esperanza condicionada de interés, incluso si no podemos observar todos los controles deseados, o si solo observamos los resultados de equilibrio de las variables.

A continuación, se ofrece una descripción general de las características importantes del operador de la esperanza condicionada.

Definición 1

Sea y una variable aleatoria, y siendo $\mathbf{x} \equiv (x_1, x_2, \ldots, x_k)$ un vector aleatorio $1 \times k$ de variables explicativas. Si

$$E(|y|) < \infty$$

entonces, existe una función

$$\mu : \boldsymbol{R_k} \to \boldsymbol{R}$$

tal que

$$E(y|x_1, x_2, \ldots, x_k) = \mu(x_1, x_2, \ldots, x_k) \tag{8.7}$$

Esto se puede expresar como $E(y|\boldsymbol{x}) = \mu(\boldsymbol{x})$. La función $\mu(x)$ determina cómo cambia el valor promedio de y a medida que cambian los elementos de $\boldsymbol{x}$. Por ejemplo, si y es *salario* y $\boldsymbol{x}$ contiene varias características individuales, como *educación*, *experiencia* e *IQ*, entonces $E(salario|educ, exp, IQ)$ es el valor promedio del salario para los valores dados de *educ*, *exp* y *IQ*. Técnicamente, deberíamos distinguir $E(y|\boldsymbol{x})$ —que es una variable aleatoria porque x es un vector aleatorio definido en la población— de la esperanza condicionada cuando x toma un valor particular de x_0 : $E(y|\boldsymbol{x} = \boldsymbol{x_0})$. Hacer

esta distinción puede ser engorroso, y en la mayoría de los casos no es tan relevante, de modo que trataremos de evitarla. Al discutir las características probabilísticas de $E(y|\boldsymbol{x})$, $\boldsymbol{x}$ se ve necesariamente como una variable aleatoria.

Como $E(y|\boldsymbol{x})$ es una esperanza, se puede obtener de la densidad condicionada de y dada $\boldsymbol{x}$ mediante integración, suma o una combinación de las dos (dependiendo de la naturaleza de y). De ello se deduce que el operador de la esperanza condicionada tiene varias propiedades adicionales que son consecuencia de la aleatoriedad de $\mu(\boldsymbol{x})$.

En econometría muy habitualmente se especifica que un modelo para una esperanza condicionada depende de un conjunto finito de parámetros, lo que da un *modelo paramétrico* de $E(y|\boldsymbol{x})$. Esto reduce considerablemente la lista de posibles candidatos para $\mu(\boldsymbol{x})$.

Ejemplo 3. Para $K = 2$ variables explicativas, se consideran los siguientes ejemplos de esperanzas condicionadas:

$$E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{8.8}$$

$$E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 \tag{8.9}$$

$$E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \tag{8.10}$$

$$E(y|x_1, x_2) = exp[\beta_0 + \beta_1 log(x_1) + \beta_2 x_2], y \geq 0, x_1 > 0 \tag{8.11}$$

El modelo de la ecuación (8.8) es lineal en las variables explicativas x_1 y x_2. La ecuación (8.9) es un ejemplo de esperanza condicionada no lineal en x_2, aunque es lineal en x_1. Desde una perspectiva estadística las ecuaciones (8.8) y (8.9) pueden tratarse en el mismo marco porque son lineales en los parámetros b_j. El hecho de que la ecuación (8.9) no sea lineal en $\boldsymbol{x}$ tiene implicaciones importantes para interpretar los b_j, pero no para estimarlos. La ecuación (8.10) cae en esta misma clase: no es lineal en $\boldsymbol{x} = (x_1, x_2)$ pero lineal en b_j. La ecuación (8.11) difiere fundamentalmente de los tres primeros ejemplos en que es una función no lineal de los parámetros b_j, así como de x_j. Además, la no linealidad en los parámetros tiene implicaciones para estimar b_j.

8.7.1 Efectos parciales y elasticidades

Si y y $\boldsymbol{x}$ están relacionadas de manera determinista, digamos mediante $y = f(\boldsymbol{x})$, entonces a menudo estamos interesados en cómo y cambia cuando los elementos de $\boldsymbol{x}$ cambian. En un entorno estocástico no podemos asumir que $y = f(\boldsymbol{x})$ para alguna función conocida y vector observable $\boldsymbol{x}$ porque siempre hay factores no observados que afectan a y. No obstante, podemos definir los efectos parciales de x_j sobre la esperanza condicionada $E(y|\boldsymbol{x})$. Suponiendo que $\mu()$ es apropiadamente diferenciable y x_j es una variable continua, la derivada parcial $\partial\mu(x)\backslash\partial x_j$ nos permite aproximar el

cambio marginal en $E(y|\boldsymbol{x})$ cuando x_j aumenta en una pequeña cantidad, manteniendo $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots x_k$ constante:

$$\Delta E(y|\boldsymbol{x}) \approx \frac{\partial \mu(\boldsymbol{x})}{x_j} \cdot \Delta x_j \tag{8.12}$$

La derivada parcial de $E(y|\boldsymbol{x})$ con respecto a x_j se denomina generalmente efecto parcial de x_j sobre $E(y|\boldsymbol{x})$ (o, siendo ligeramente imprecisos, 'el efecto parcial de x_j sobre y'). La interpretación de las magnitudes de los coeficientes en modelos paramétricos generalmente proviene de la aproximación en la ecuación (8.12). Si x_j es una variable discreta (por ejemplo una variable binaria), los efectos parciales se calculan comparando $E(y|\boldsymbol{x})$ en diferentes configuraciones de x_j (por ejemplo, cero y uno cuando x_j es binario), manteniendo fijas otras variables.

Ejemplo 4. En la ecuación (8.8) tenemos

$$\frac{\partial E(y|x)}{\partial x_1} = \beta_1, \frac{parcial E(y|x)}{\partial x_2} = \beta_2$$

Como era de esperar, los efectos parciales en este modelo son constantes. En la ecuación (8.9),

$$\frac{\partial E(y|x)}{\partial x_1} = \beta_1, \frac{\partial E(y|x)}{\partial x_2} = \beta_2 + 2\beta_3 x_2$$

para que el efecto parcial de x_1 sea constante pero el efecto parcial de x_2 depende del nivel de x_2. En la ecuación (8.10),

$$\frac{\partial E(y|x)}{\partial x_1} = \beta_1 + \beta_3 x_2, \frac{\partial E(y|x)}{\partial x_2} = \beta_2 + \beta_3 x_1$$

para que el efecto parcial de x_1 dependa de x_2, y viceversa. En la ecuación (8.11),

$$\frac{\partial E(y|x)}{\partial x_1} = exp(.)(\beta_1/x_1), \frac{\partial E(y|x)}{\partial x_2} = exp(.)\beta_2$$

donde $exp(.)$ denota la función $E(y|x)$ en la ecuación (8.11). En este caso, los efectos parciales de x_1 y x_2 ambos dependen de $\boldsymbol{x} = (x_1, x_2)$.

A veces nos va a interesar una función particular de un efecto parcial, como en las denominadas elasticidades.

Definición 2

En el caso de $y = f(x)$, definimos una elasticidad de y con respecto de x_j como

$$\frac{\partial y}{\partial x_j} \cdot \frac{\partial x_j}{\partial y} = \frac{\partial f(x)}{\partial x_j} \cdot \frac{\partial x_j}{\partial f(x)} \tag{8.13}$$

Expresión válida, de nuevo, asumiendo que x_j sea continuo. El lado derecho de la ecuación (8.13) muestra que la elasticidad es función de x. Cuando y y x sean aleatorios, tiene sentido usar el lado derecho de la ecuación (8.13), pero donde $f(\boldsymbol{x})$ es la media condicionada, $\mu(x)$. Por lo tanto, la elasticidad (parcial) de $E(y|\boldsymbol{x})$ con respecto a x_j, manteniendo $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$ constante, es

$$\frac{\partial E(y|x)}{\partial x_j} \cdot \frac{\partial x_j}{\partial E(y|x)} = \frac{\partial \mu(x)}{\partial x_j} \cdot \frac{\partial x_j}{\partial \mu(x)} \tag{8.14}$$

Si $E(y|\boldsymbol{x}) > 0$ y $x_j > 0$ (como suele ser habitual), la ecuación (8.14) es idéntica a

$$\frac{log\left[\partial E(y|\boldsymbol{x})\right]}{log(\partial x_j)} \tag{8.15}$$

Esta última expresión da a la elasticidad su interpretación como el cambio porcentual aproximado en $E(y|x_j)$ cuando x_j aumenta en $1\,\%$.

Ejemplo 5. (continuación) En las ecuaciones (8.8) a (8.11), la mayoría de las elasticidades no son constantes. Por ejemplo, en (8.8), la elasticidad de $E(y|x_j)$ con respecto a x_1 es

$$\frac{\beta x_1}{\beta_0 + \beta_1 x_1 + \beta_2 x_2}$$

expresión que claramente depende de x_1 y x_2. Sin embargo, en la ecuación (8.11) la elasticidad con respecto a x_1 es constante e igual a β_1.

¿Cómo se compara la ecuación (8.15) con la definición de elasticidad de un modelo lineal con los logaritmos naturales? Si $y > 0$ y $x_j > 0$, podríamos definir la elasticidad como

$$\frac{\partial E[log(y)|\boldsymbol{x}]}{\partial log(x_j)} \tag{8.16}$$

Esta es la definición natural en un modelo como $log(y) = g(\boldsymbol{x}) + u$, donde $g(\boldsymbol{x})$ es alguna función de $\boldsymbol{x}$ y u es una perturbación no observada con media condicionada cero en $\boldsymbol{x}$. ¿Cómo se comparan las ecuaciones (8.14) y (8.15)? Generalmente, son diferentes (ya que el valor esperado del logaritmo y el logaritmo del valor esperado pueden ser muy diferentes). Si u es independiente de $\boldsymbol{x}$, entonces son iguales, porque entonces

$$E(y|\boldsymbol{x}) = \delta \cdot E(g(\boldsymbol{x}))$$

donde $\delta \equiv E(exp(u))$ (Si u y $\boldsymbol{x}$ son independientes, $exp(u)$ y $exp[g(\boldsymbol{x})]$.)

Ejemplo 6. Si

$$log(y) = \beta_0 + \beta_1 log(x_1) + \beta_2 x_2 + u$$

donde u tiene media cero y es independiente de (x_1, x_2), entonces la elasticidad de y con respecto a x_1 es β_1 usando cualquier definición de elasticidad. Si $E(u|\boldsymbol{x}) = 0$ pero u y $\boldsymbol{x}$ no son independientes, las definiciones son generalmente diferentes.

8.7.2 El término error en la esperanza condicionada

Cuando y es una variable aleatoria que nos gustaría explicar en términos de variables observables x, es útil descomponer y como

$$y = E(y|\boldsymbol{x}) + u \tag{8.17}$$

$$E(u|\boldsymbol{x}) = 0 \tag{8.18}$$

En otras palabras, las ecuaciones (8.17) y (8.18) son definitorias: siempre podemos escribir y como su esperanza condicionada, $E(y|\boldsymbol{x})$, más un término de error o perturbación que tiene una media condicionada de cero.

El hecho de que $E(u|\boldsymbol{x}) = 0$ tiene las siguientes implicaciones importantes:

- $E(u) = 0$;
- u no está correlacionado con ninguna función de $x_1, x_2, \ldots, x_K$ y, en particular, u no está correlacionado separadamente con $x_1, x_2, \ldots, x_K$. Que u tenga esperanza condicionada cero se obtiene como un caso especial de la *ley de esperanzas iteradas* (LIE). Intuitivamente, es bastante razonable que $E(u|\boldsymbol{x}) = 0$ implique $E(u) = 0$. La segunda implicación es menos obvia, pero muy importante. El hecho de que u no esté correlacionado con ninguna función de $\boldsymbol{x}$ es mucho más fuerte que simplemente decir que u no está correlacionado con $x_1, \ldots, x_K$

Si la ecuación (8.8) se mantiene, entonces se puede escribir

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, E(u|x_1, x_2) = 0 \tag{8.19}$$

y entonces

$$E(u) = 0, Cov(x_1, u) = 0, Cov(x_2, u) = 0 \tag{8.20}$$

Pero podemos ir más allá, pues en la ecuación (8.19), u tampoco está correlacionado con cualquier otra función que podamos pensar, como $x_1^2, x_2^2, x_1 x_2, exp(x_1)$ y $log(x_2^2 + 1)$. Este hecho asegura que hemos tenido en cuenta completamente los efectos de x_1 y x_2 sobre el valor esperado de y. Otra forma de enunciar este punto es que tenemos la forma funcional de $E(y|\boldsymbol{x})$ debidamente especificada.

Ejemplo 7. Supongamos que los precios de la vivienda están determinados por el modelo simple

$$hprice = \beta_0 + \beta_1 sqrft + \beta_2 distance + u$$

donde *sqrft* son los metros cuadrados de la casa, y emphdistancia es la distancia entre la casa y un incinerador de la ciudad. Para que β_2 represente

$$\frac{\partial E(hprice|sqrft, distancia)}{\partial distancia}$$

debemos asumir que $E(u|sqrft, distancia) = 0$

8.7.3 Algunas propiedades de la esperanza condicionada

Una de las herramientas más útiles para manipular las esperanzas condicionadas es la ley de las esperanzas iteradas. Supongamos que $\boldsymbol{w}$ es un vector aleatorio e y es una variable aleatoria. Sea $\boldsymbol{x}$ un vector aleatorio función de $\boldsymbol{w}$, digamos $\boldsymbol{x} = f(\boldsymbol{w})$ (el vector $\boldsymbol{x}$ podría ser simplemente un subconjunto de $\boldsymbol{w}$). Esta definición va a implicar que si conocemos el resultado de $\boldsymbol{w}$, entonces conocemos el resultado de $\boldsymbol{x}$. La definición más general de la ley de las esperanzas iterada que necesitaremos es

$$E(y|\boldsymbol{x}) = E[E(y|\boldsymbol{w})|\boldsymbol{x}] \tag{8.21}$$

En otras palabras, si se formula

$$\mu_1(\boldsymbol{w}) \equiv E(y|\boldsymbol{w})$$

y

$$\mu_2(\boldsymbol{x}) \equiv E(y|\boldsymbol{x})$$

se puede obtener $\mu_2(\boldsymbol{x})$ calculando el valor esperado de $\mu_2(\boldsymbol{w})$ dado $\boldsymbol{x}$: $\mu_1(\boldsymbol{x}) = E(\mu_1(\boldsymbol{w})|\boldsymbol{x})$. Hay un resultado similar al del la ecuación (8.21), pero que resulta más sencillo de comprobar:

$$E(y|\boldsymbol{x}) = E[E(y|\boldsymbol{x})|\boldsymbol{w}] \tag{8.22}$$

Se puede observar cómo las posiciones de $\boldsymbol{x}$ y $\boldsymbol{w}$ se han cambiado en el lado derecho de la ecuación (8.22) en comparación con la ecuación (8.21). El resultado en la ecuación (8.22) se sigue fácilmente del aspecto condicionado de la esperanza: dado que $\boldsymbol{x}$ es una función de $\boldsymbol{w}$, saber $\boldsymbol{w}$ implica conocer que $\boldsymbol{x}$; dado que $\mu_2 = E(y|\boldsymbol{x})$ es una función de $\boldsymbol{x}$, el valor esperado de $\mu_2(\boldsymbol{x})$ dado $\boldsymbol{w}$ es precisamente $\mu_2(\boldsymbol{x})$.

Para muchas aplicaciones econométricas, es útil pensar en $\mu_1(\boldsymbol{x}, \boldsymbol{z}) = E(y|\boldsymbol{x}, \boldsymbol{z})$ como una esperanza condicionada estructural, pero donde $\boldsymbol{z}$ no se observa. Si el interés está en $E(y|\boldsymbol{x}, \boldsymbol{z})$, entonces queremos que los efectos de x_j mantengan los otros elementos de $\boldsymbol{x}$ y $\boldsymbol{z}$ fijos. Si no se observa $\boldsymbol{z}$, no podemos estimar $E(y|\boldsymbol{x}, \boldsymbol{z})$ directamente. Sin embargo, dado que se observan y y $\boldsymbol{x}$, generalmente podemos estimar $E(y|\boldsymbol{x})$. La pregunta, entonces, es si podemos relacionar $E(y|\boldsymbol{x})$ con la esperanza original de interés (ésta es una versión del 'problema de identificación' en econometría.) La LIE proporciona una forma conveniente de relacionar las dos esperanzas.

Obtener $E[\mu_1(\boldsymbol{x}, \boldsymbol{z})|x]$ generalmente requiere integrar (o sumar) $\mu_1(\boldsymbol{x}, \boldsymbol{z})$ contra la densidad condicionada de z dada x, pero en muchos casos la forma de $E(y|\boldsymbol{x}, \boldsymbol{z})$ es lo

suficientemente simple como para no requerir una integración explícita. Por ejemplo, supongamos que comenzamos con el modelo

$$E(y|x_1, x_2, z) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 z \tag{8.23}$$

pero donde z no es observada. Según la LIE, y la linealidad del operador de la esperanza condicionada

$$E(y|x_1, x_2) = E(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 z|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 E(z|x_1, x_2) \tag{8.24}$$

Si se hace una asunción sobre $E(z|x_1, x_2)$, por ejemplo, que sea lineal tanto en x_1 como en x_2,

$$E(z|x_1, x_2) = \delta_0 + \delta_1 x_1 + \delta_2 x_2 \tag{8.25}$$

entonces se puede insertarlo en la cuación (8.24) y reorganizar del siguiente modo:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3(\delta_0 + \delta_1 x_1 + \delta_2 x_2)$$

$$(\beta_3\delta_0 + \beta_0) + (\beta_1 + \beta_3\delta_1)x_1 + (\beta_2 + \beta_3\delta_2)x_2$$

Esta última expresión es $E(y|x_1, x_2)$; dados nuestros supuestos, será necesariamente lineal en (x_1, x_2).

La forma general de la LIE va a tener otras implicaciones útiles. Supongamos que para alguna función (vectorial) $f(x)$ y una función real $g(.)$, $E(y|\boldsymbol{x}) = g[f(\boldsymbol{x})]$. Entonces

$$E[y|f(\boldsymbol{x})] = E(y|\boldsymbol{x}) = g[f(\boldsymbol{x})] \tag{8.26}$$

También necesitamos saber cómo se relaciona la noción de independencia estadística con la de esperanza condicionada. Si u es una variable aleatoria independiente asociada al vector aleatorio $\boldsymbol{x}$, entonces $E(u|\boldsymbol{x}) = E(u)$, de modo que si $E(u) = 0$ y u y x son independientes, entonces $E(u|\boldsymbol{x}) = 0$. Lo contrario de esto no es cierto: $E(u|\boldsymbol{x}) = E(u)$ no implica independencia estadística entre u y $\boldsymbol{x}$ (al igual que la correlación cero entre u y $\boldsymbol{x}$ no implica independencia).

8.8 Identificabilidad y estimabilidad

Cuando la matriz del modelo no es de rango completo, $\boldsymbol{\beta}$ no es identificable.

Definición 3

Definición. Para un GLM con predictor lineal $X\boldsymbol{\beta}$, el vector de parámetro $\boldsymbol{\beta}$ es identificable siempre que $\boldsymbol{\beta}^* \neq \boldsymbol{\beta}$, luego $X\boldsymbol{\beta}^* \neq X\boldsymbol{\beta}$.

De manera equivalente, $\boldsymbol{\beta}$ es identificable si $X\boldsymbol{\beta}^* = X\boldsymbol{\beta}$ implica que $\boldsymbol{\beta}^* = \boldsymbol{\beta}$, entonces esta definición nos dice que si sabemos que $g(\boldsymbol{\mu}) = X\boldsymbol{\beta}$ (y por lo tanto, si sabemos que $\boldsymbol{\mu}$ satisface el modelo), entonces también podemos determinar $\boldsymbol{\beta}$. Para la parametrización que se acaba de dar para el diseño unidireccional, $\boldsymbol{\beta}$ no es identificable, porque

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_c)^T$$

y

$$\boldsymbol{\beta}^* = (\beta_0 + d, \beta_1 - d, \cdots, \beta_c - d)^T$$

no tienen diferentes valores de predictores lineales. En tales casos, podemos obtener identificabilidad y eliminar el efecto de perturbación entre los parámetros redefiniendo el predictor lineal con menos parámetros. Entonces, diferentes valores de $\boldsymbol{\beta}$ tienen diferentes valores de predictores lineales $X\boldsymbol{\beta}$, y la estimación de $\boldsymbol{\beta}$ es posible.

Una definición un poco más general de identificabilidad se refiere en cambio a combinaciones lineales de parámetros

$$\boldsymbol{\ell}^T\boldsymbol{\beta}$$

. Y se establece que

$$\boldsymbol{\ell}^T\boldsymbol{\beta}$$

es identificable siempre que

$$\boldsymbol{\ell}^T\boldsymbol{\beta}^* \neq \boldsymbol{\ell}^T\boldsymbol{\beta}$$

luego

$$X\boldsymbol{\beta}^* \neq X\boldsymbol{\beta}$$

Esta definición permite que un subconjunto de los términos en $\boldsymbol{\beta}$ sea identificable, en lugar de tratar el $\boldsymbol{\beta}$ completo como identificable o no identificable. Por ejemplo, supongamos que ampliamos el modelo como un diseño unidireccional para incluir una variable explicativa cuantitativa que toma el valor x_{ij} para la observación j en el grupo i, lo que produce el análisis del modelo de covarianza

$$g(\mu_{ij}) = \beta_0 + \beta_i + \gamma x_{ij}$$

Entonces, sin una restricción en β_i o β_0, de acuerdo con esta definición β_i y β_0 no son identificables, pero γ sí lo es. Aquí, tomando $\boldsymbol{\ell}^T\boldsymbol{\beta} = \gamma$, diferentes valores de $\boldsymbol{\ell}^T\boldsymbol{\beta}$ producen diferentes valores de $X\boldsymbol{\beta}$.

En un modelo lineal, el adjetivo estimable se va a referir a ciertas cantidades que pueden estimarse de manera insesgada.

Definición 4

En un modelo lineal $E(\boldsymbol{y}) = X\boldsymbol{\beta}$, la cantidad $\boldsymbol{\ell}^T\boldsymbol{\beta}$ es estimable si existen unos coeficientes tales que $E(\boldsymbol{a}^T\boldsymbol{y}) = \boldsymbol{\ell}^T\boldsymbol{\beta}$.

Es decir, alguna combinación lineal de las observaciones estima $\boldsymbol{\ell}^T\boldsymbol{\beta}$ sin sesgo.

Cuando $\boldsymbol{X}$ tiene rango completo, $\boldsymbol{\beta}$ es identificable, y entonces todas las combinaciones lineales $\boldsymbol{\ell}^T\boldsymbol{\beta}$ son estimables. Las estimaciones no dependen de qué restricciones empleamos para obtener la identificabilidad. Cuando $\boldsymbol{X}$ no tiene rango completo, $\boldsymbol{\beta}$ no es identificable. También en ese caso, para la definición más general de identificabilidad en términos de combinaciones lineales $\boldsymbol{\ell}^T\boldsymbol{\beta}$, al menos un componente de $\boldsymbol{\beta}$ no es identificable. De hecho, para esa definición, $\boldsymbol{\ell}^T\boldsymbol{\beta}$ es estimable si y solo si es identificable. Entonces, las cantidades estimables son simplemente las funciones lineales de $\boldsymbol{\beta}$ que son identificables.

La no identificabilidad de $\boldsymbol{\beta}$ es irrelevante siempre que nos centremos en $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ y otras características estimables. En particular, cuando $\boldsymbol{\ell}^T\boldsymbol{\beta}$ es estimable, los valores de $\boldsymbol{\ell}^T\boldsymbol{\beta}$ son los mismos para todas las soluciones $\boldsymbol{\beta}$ de las ecuaciones de verosimilitud. Entonces, ¿cuál es el conjunto de combinaciones lineales $\boldsymbol{\ell}^T\boldsymbol{\beta}$ que son estimables? Dado que $E(\boldsymbol{a}^T\boldsymbol{y}) = \boldsymbol{\ell}^T\boldsymbol{\beta}$ con $\boldsymbol{\ell}^T = \boldsymbol{a}^T\boldsymbol{X}$, el espacio lineal de tales $p \times 1$ vectores $\boldsymbol{\ell}$ es precisamente el conjunto de combinaciones lineales de filas de $\boldsymbol{X}$; es decir, es el espacio de filas de la matriz modelo $\boldsymbol{X}$, que es equivalentemente $C(\boldsymbol{X}^T)$. Esto no es sorprendente, ya que cada media es el producto interno de una fila de $\boldsymbol{X}$ con $\boldsymbol{\beta}$.

Bibliografía

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley.

Wooldridge, Jeffrey M (2010). *Econometric analysis of cross section and panel data*. MIT press.

Tema 9

Introducción. Ajuste del modelo de mínimos cuadrados. Proyecciones de datos sobre el modelo de espacios. Resumen de la variabilidad en un modelo lineal. Residuos, apalancamiento (leverage) e influencia. Optimalidad de mínimos cuadrados y de mínimos cuadrados generalizados

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía:

A. Agresti (2015). *Foundations of Linear and Generalized Linear Models*. Wiley

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

9.1 Introducción

Tanto este tema como el siguiente consideran el ajuste e inferencia del modelo lineal ordinario. Para n observaciones independientes $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ con $\mu_i = E(y_i)$ y $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$, se denota la matriz de covarianzas por

$$\boldsymbol{V} = var(\boldsymbol{y}) = E[(y - \mu)(y - \mu)^T]$$

Sea $\boldsymbol{X} = (x_{ij})$ el modelo $n \times p$ matriz, donde x_{ij} es el valor de la variable explicativa j para la observación i. En este capítulo aprenderemos sobre el ajuste del modelo cuando $\mu = \boldsymbol{X}\boldsymbol{\beta}$ con $\boldsymbol{V} = \sigma^2\boldsymbol{I}$, donde $\boldsymbol{\beta}$ es un vector de parámetro $p \times 1$ con $p \leq n$ y $\boldsymbol{I}$ es la matriz de identidad $n \times n$. La matriz de covarianzas es una matriz diagonal con valor común σ^2 para la varianza. Con el supuesto adicional de un componente aleatorio normal, este es el modelo lineal normal, que es un modelo lineal generalizado (GLM) con función de enlace de identidad. Agregaremos el supuesto de normalidad en el próximo capítulo. Sin embargo, aquí obtendremos muchos resultados sobre el ajuste de modelos lineales y la comparación de modelos que no requieren supuestos distributivos.

Una forma alternativa de expresar el modelo lineal ordinario es

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

para un término de error $\boldsymbol{\epsilon}$ que tiene $E(\boldsymbol{\epsilon}) = \mathbf{0}$ y una matriz de covarianzas $\boldsymbol{V} = var(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$. Sin embargo, una estructura aditiva tan simple para el término de error no es natural para la mayoría de los GLM, excepto para los modelos normales y las versiones de variables latentes de algunos otros modelos y sus extensiones con múltiples componentes de error. Para ser coherentes con las fórmulas GLM, normalmente expresaremos modelos lineales en términos de $E(\boldsymbol{y})$.

9.2 Ajuste del modelo de mínimos cuadrados

Habiendo formado una matriz de modelo $\boldsymbol{X}$ y observado $\boldsymbol{y}$, ¿cómo obtenemos estimaciones de los parámetros $\widehat{\beta}$ y los valores ajustados $\widehat{\mu} = \boldsymbol{X}\widehat{\beta}$ que mejor satisfacen el modelo lineal? El enfoque estándar utiliza el método de mínimos cuadrados. Esto determina el valor de $\widehat{\boldsymbol{\mu}}$ que minimiza

$$||\boldsymbol{y} - \widehat{\boldsymbol{\mu}}||^2 = \sum_i (y_i - \widehat{\mu}_i)^2 = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\widehat{\beta}_j x_{ij})^2$$

De modo que los valores ajustados $\widehat{\mu}$ son los siguientes

$$||\boldsymbol{y} - \widehat{\boldsymbol{\mu}}|| \leq ||\boldsymbol{y} - \boldsymbol{\mu}||$$

para todo $\boldsymbol{\mu} \in C(\boldsymbol{X})$.

El uso de mínimos cuadrados corresponde a la máxima probabilidad cuando agregamos un supuesto de normalidad al modelo. El logaritmo de las probabilidades para observaciones independientes $y_i \sim N(\mu_i, \sigma^2), i = 1, \ldots, n$, es (en términos de μ_i)

$$log[\prod_{i=1}^{n}(\frac{1}{\sqrt{2\pi}\sigma}e^{-(y_i-\mu_i)^2/2\sigma^2})] = constant - [\sum_{i=1}^{n}(y_i - \mu_i)^2]/2\sigma^2$$

Para maximizar la función de probabilidad logarítmica, debemos minimizar $\sum_i (y_i - \mu_i)^2$.

9.2.1 Las ecuaciones Normales

La expresión

$$L(\boldsymbol{\beta}) = \sum_i (y_i - \mu_i)^2 = \sum_i (y_i - \sum_j \beta_j x_{ij})^2$$

es cuadrática en β_j, por lo que podemos minimizarlo equiparando

$$\frac{\partial L}{\partial \beta_j} = 0, j = 1, \cdots, p.$$

Estas derivadas parciales producen las ecuaciones:

$$\sum_i (y_i - \mu_i) x_{ij} = 0, j = 1, \cdots, p.$$

Así que la estimación de mínimos cuadrados satisface

$$\sum_i^n y_i x_{ij} = \sum_i^n \widehat{\mu}_i x_{ij}, j = 1, \cdots, p \tag{9.1}$$

Y estas ecuaciones son las llamadas *ecuaciones normales* [1]. Estas ecuaciones se situan a un nivel más general que los mínimos cuadrados. Estas son las ecuaciones de verosimilitud para GLM que utilizan la función de vínculo canónico, como en el modelo lineal Normal, el modelo de regresión logística Binomial y el modelo loglineal de Poisson. El uso de álgebra matricial proporciona una expresión económica para la solución de estas ecuaciones en términos del vector de parámetros del modelo β para el modelo lineal $\mu = X\beta$. En forma de matriz,

$$L(\boldsymbol{\beta}) = ||\boldsymbol{y} - \boldsymbol{X\beta}||^2 = (\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta}) = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{2y}^T\boldsymbol{X\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta}$$

Usamos entonces los resultados para derivadas matriciales en los que

$$\partial(\boldsymbol{a}^T\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \boldsymbol{a}$$

$$\partial(\boldsymbol{\beta}^T\boldsymbol{A\beta})/\partial\boldsymbol{\beta} = (\boldsymbol{A} + \boldsymbol{A^T})\boldsymbol{\beta}$$

que equivale a $2\boldsymbol{A\beta}$ para $\boldsymbol{A}$ simétrica. Entonces, $\partial L(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X\beta})$. En términos de $\widehat{\boldsymbol{\beta}}$, las ecuaciones normales (9.1) son

$$\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}. \tag{9.2}$$

Supongamos que $\boldsymbol{X}$ tiene rango completo p. Entonces, la matriz $p \times p$ $(\boldsymbol{X}^T\boldsymbol{X})$ también tiene rango p y no es singular, su inversa existe y el estimador de mínimos cuadrados de $\boldsymbol{\beta}$ es

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{9.3}$$

Dado que $\partial^2 L(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^2 = 2\boldsymbol{X}^T\boldsymbol{X}$ es positivo definido, el mínimo de $L(\boldsymbol{\beta})$ ocurre en $\boldsymbol{\beta}\widehat{\boldsymbol{\beta}}$.

Los valores ajustados $\widehat{\boldsymbol{\mu}}$ son una transformación lineal de $\boldsymbol{y}$,

$$\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

[1] En este contexto 'normales 'no quieres decir de la distribución Normal sino en el sentido de vector normal u ortogonal.

La matriz de influencia (también *hat matrix* en inglés) es una matriz $n \times n$, $\boldsymbol{H} = \boldsymbol{X}(X^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ que transforma linealmente $\boldsymbol{y}$ en $\widehat{\boldsymbol{\mu}} = \boldsymbol{H}\boldsymbol{y}$. La matriz $\boldsymbol{H}$ es una matriz de proyección, que proyecta $\boldsymbol{y}$ en $\widehat{\boldsymbol{\mu}}$ en el espacio modelo $(\boldsymbol{X})$.

Es preciso recordar que para una matriz de constantes $\boldsymbol{A}$, $E(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A}E(\boldsymbol{y})$ y $var(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A}var(\boldsymbol{y})\boldsymbol{A}^T$. Entonces, la media y la varianza del estimador de mínimos cuadrados son

$$E(\widehat{\boldsymbol{\beta}}) = E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta} \tag{9.4a}$$

$$var(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\sigma^2\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{9.4b}$$

Para el modelo lineal ordinario con componente aleatorio normal, dado que $\widehat{\boldsymbol{\beta}}$ es una función lineal de $\boldsymbol{y}$, $\widehat{\boldsymbol{\beta}}$ tiene una distribución normal con estos dos momentos.

A continuación, se ilustran los mínimos cuadrados usando el modelo lineal con una sola variable explicativa para una sola respuesta, es decir, el 'modelo lineal bivariado '

$$E(y_i) = \beta_0 + \beta_1 x_i$$

From (9.1) with $x_{i1} = 1$ and $x_{i2} = x_i$, the normal equations are

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = \beta_0(\sum_{i=1}^{n} x_i) + \beta_1 \sum_{i=1}^{n} x_i^2$$

Mediante la sencilla solución que plantean estas dos ecuaciones, se puede verificar que las estimaciones de mínimos cuadrados son

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{9.5a}$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x} \tag{9.5b}$$

De la solución para β_0, la ecuación ajustada por mínimos cuadrados $\widehat{\mu}_i = \beta_0 + \beta_1 x_i$ satisface $\overline{y} = \beta_0 + \beta_1\overline{x}$; pasando por el centro de gravedad de los datos, es decir, el punto $(\overline{x}, \overline{y})$. El resultado análogo es válido para el modelo lineal con múltiples variables explicativas, siendo el punto $(\overline{x}_1, \ldots, \overline{x}_p, \overline{y})$.

Cuando $\boldsymbol{X}$ no tiene rango completo, tampoco lo tiene $(\boldsymbol{X}^T\boldsymbol{X})$ en las ecuaciones normales. Una solución $\widehat{\boldsymbol{\beta}}$ de las ecuaciones normales hace uso de la *inversa generalizada* de $(\boldsymbol{X}^T\boldsymbol{X})$, denotada por $(\boldsymbol{X}^T\boldsymbol{X})^-$. Conviene recordar aquí que para una matriz A, G es una inversa generalizada sí y solo sí $AGA = A$.

Las inversas generalizadas siempre existen, pero pueden no ser únicas. La estimación de mínimos cuadrados $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y}$ no es entonces única, lo que refleja que $\boldsymbol{\beta}$ no es identificable. Con $rank(\boldsymbol{X}) < p$, el espacio nulo $N(\boldsymbol{X})$ tiene elementos distintos de cero. Para cualquier solución $\widehat{\boldsymbol{\beta}}$ de las ecuaciones normales $\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ y cualquier elemento $\boldsymbol{\gamma} \in N(\boldsymbol{X}), \tilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} + \boldsymbol{\gamma}$ también es una solución. Esto se debe a que $\boldsymbol{X}\boldsymbol{\gamma} = \boldsymbol{0}$ y $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}(\widehat{\boldsymbol{\beta}} + \boldsymbol{\gamma}) = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta}$.

Además, si $\boldsymbol{\ell}^T\boldsymbol{\beta}$ es estimable, entonces $\boldsymbol{\ell}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}$ es lo mismo para todas las soluciones a las ecuaciones normales. Esto se debe a que $\boldsymbol{\ell}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}$ se puede expresar como $\boldsymbol{a}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ por unos $\boldsymbol{a}$, y los valores ajustados son idénticos para todos los $\widehat{\boldsymbol{\beta}}$.

9.2.2 Alternativas a los mínimos cuadrados

Al ajustar un modelo lineal, ¿por qué minimizar $\sum_i (y_i - \widehat{\mu})^2$ en lugar de alguna otra métrica, como $\sum_i |y_i - \widehat{\mu}|$? La razón es que minimizar una suma de cuadrados es matemáticamente y computacionalmente mucho más simple. Por esta razón, los mínimos cuadrados tienen una larga historia, que se remonta a un artículo publicado por el matemático francés Adrien-Marie Legendre (1805), aunque esto provocó la consabida reclamación del matemático alemán Carl Friedrich Gauss en 1809 por que según él ya lo venía utilizando desde 1795. Otra motivación es que corresponde a la máxima verosimilitud cuando agregamos el supuesto de normalidad. Y otra motivación más muestra que el estimador de mínimos cuadrados es el mejor en la clase de estimadores que son insesgados y lineales en los datos. Investigaciones recientes han desarrollado alternativas a los mínimos cuadrados que dan respuestas sensatas en situaciones en las que las estimaciones por este método son inestables por algún motivo.

Por ejemplo, la inestabilidad puede ser causada por un valor atípico grave, porque al minimizar una suma de desviaciones cuadradas, una sola observación puede tener una influencia sustancial. La inestabilidad también podría ser causada por una variable explicativa determinada linealmente (o casi) por las otras variables explicativas, una condición llamada *multicolinealidad*. Finalmente, la inestabilidad también puede ocurrir al usar mínimos cuadrados con conjuntos de datos que contienen un gran número de variables explicativas, a veces incluso con $p > n$.

Para resolver el problema de inestabilidad causado por un número muy elevado de variables cada vez es más común, sobretodo ahora dentro del marco del *Big data*, recurrir a los llamados *métodos de regularización*, los cuales agregan un término adicional a la función minimizada, como $\lambda \sum_j |\beta_j|$ o $\lambda \sum_j \beta_j^2$ para una constante λ. Entonces, la solución es un suavizado de las estimaciones de mínimos cuadrados que las puede reducir a cero. Esto es muy eficaz cuando tenemos una gran cantidad de variables explicativas pero esperamos que pocas de ellas tengan un efecto sustancialmente importante. A menos que n sea extremadamente grande, debido a la variabilidad muestral, las estimaciones de mínimos cuadrados ordinarios β_j tienden a ser mucho mayores en valor absoluto que los valores verdaderos β_j. La contracción hacia 0 provoca un sesgo en los estimadores, pero tiende a reducir la varianza sustancialmente, lo que hace que tiendan a acercarse más a β_j. Los métodos de regularización son cada vez más importantes a medida que involucran más aplicaciones de Big Data.

9.3 Proyecciones de datos sobre el modelo de espacios

Se ha mencionado que el ajuste de mínimos cuadrados $\widehat{\boldsymbol{\mu}}$ es una proyección de los datos $\boldsymbol{y}$ sobre el espacio modelo $\boldsymbol{C(X)}$, y la matriz $\boldsymbol{H}$ que proyecta $\boldsymbol{y}$ sobre $\widehat{\boldsymbol{\mu}}$ es una matriz de proyección. Ahora se explica con más precisión qué se entiende por proyección de un vector $\boldsymbol{y} \in \mathbf{R}^n$ sobre un subespacio vectorial como $\boldsymbol{C(X)}$.

Definición 5

Una matriz cuadrada $\boldsymbol{P}$ es una *matriz de proyección* sobre un subespacio vectorial $\boldsymbol{W}$ si

- para todo $\boldsymbol{y} \in \boldsymbol{W}, \boldsymbol{Py} = \boldsymbol{y}$.
- para todo $\boldsymbol{y} \in \boldsymbol{W}^{\perp}, \boldsymbol{Py} = \boldsymbol{0}$.

Para una matriz de proyección $\boldsymbol{P}$, como $\boldsymbol{Py}$ es una combinación lineal de las columnas de $\boldsymbol{P}$, el subespacio vectorial $\boldsymbol{W}$ en el que $\boldsymbol{P}$ se proyecta es el espacio columna $\boldsymbol{C(P)}$. La matriz de proyección $\boldsymbol{P}$ sobre $\boldsymbol{W}$ proyecta un vector $\boldsymbol{y} \in \mathbf{R}^n$ arbitrario en sus componentes $y_1 \in \boldsymbol{W}$ para la descomposición ortogonal única de $\boldsymbol{y}$ en $y_1 + y_2$ usando $\boldsymbol{W}$ y $\boldsymbol{W}^{\perp}$. A continuación, enumeramos esta y otras propiedades de una matriz de proyección:

- Si $\boldsymbol{y} = \boldsymbol{y_1} + \boldsymbol{y_2}$ con $\boldsymbol{y_1} \in \boldsymbol{W}$ y $\boldsymbol{y_2} \in \boldsymbol{W}^{\perp}$, entonces $\boldsymbol{Py} = \boldsymbol{P}(\boldsymbol{y_1} + \boldsymbol{y_2}) = \boldsymbol{Py_1} + \boldsymbol{Py_2} = \boldsymbol{y_1} + \boldsymbol{0} = \boldsymbol{y_1}$. Dado que la descomposición ortogonal es única, también lo es la proyección sobre $\boldsymbol{W}$.
- La matriz de proyección sobre un subespacio $\boldsymbol{W}$ es única. Para ver por qué, suponga que $\boldsymbol{P}^*$ es otro. Entonces, para la descomposición ortogonal $\boldsymbol{y} = \boldsymbol{y_1} + \boldsymbol{y_2}$ con $\boldsymbol{y_1} \in \boldsymbol{W}, \boldsymbol{P}^*\boldsymbol{y} = \boldsymbol{y_1} = \boldsymbol{Py}$ para todos $\boldsymbol{y}$. Por tanto, $\boldsymbol{P} = \boldsymbol{P}^*$. (Recuerde que si $Ay = By$ para todos los $\boldsymbol{y}$, entonces $\boldsymbol{A} = \boldsymbol{B}$).
- $\boldsymbol{I} - \boldsymbol{P}$ es la matriz de proyección sobre $\boldsymbol{W}^{\perp}$. Para un $\boldsymbol{y} = \boldsymbol{y_1} + \boldsymbol{y_2}$ arbitrario con $\boldsymbol{y_1} \in \boldsymbol{W}$ y $y_2 \in \boldsymbol{W}^{\perp}$, tenemos $\boldsymbol{Py} = \boldsymbol{y_1}$ y $(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{y_1} = \boldsymbol{y_2}$. Por tanto, $\boldsymbol{y} = \boldsymbol{Py} + (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}$ proporciona la descomposición ortogonal de y. Además, $\boldsymbol{P}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = 0$.
- $\boldsymbol{P}$ es una matriz de proyección si y solo si es simétrica e idempotente (es decir, $\boldsymbol{P}^2 = \boldsymbol{P}$).

Usaremos esta última propiedad a menudo, así que veamos por qué es verdad. Primero, supongamos que $\boldsymbol{P}$ es simétrico e idempotente y mostramos que esto implica que $\boldsymbol{P}$ es una matriz de proyección. Para cualquier $\boldsymbol{\nu} \in \boldsymbol{C(P)}$ (el subespacio en el que se proyecta $\boldsymbol{P}$), $\boldsymbol{\nu} = \boldsymbol{Pb}$ para algunos $\boldsymbol{b}$. Entonces, $\boldsymbol{P\nu} = \boldsymbol{P}(\boldsymbol{Pb}) = \boldsymbol{P2b} = \boldsymbol{Pb} = \boldsymbol{\nu}$. Para cualquier $\boldsymbol{\nu} \in \boldsymbol{C(P)}^{\perp}$, tenemos $\boldsymbol{PT\nu} = \boldsymbol{0}$, pero esto también es $\boldsymbol{P\nu}$ por la simetría de $\boldsymbol{P}$. Entonces, hemos demostrado que $\boldsymbol{P}$ es una matriz de proyección sobre $\boldsymbol{C(P)}$. En segundo lugar, para demostrar lo contrario, supongamos que $\boldsymbol{P}$ es la matriz de proyección sobre $\boldsymbol{C(P)}$ y mostraremos que esto implica que $\boldsymbol{P}$ es simétrico e idempotente. Para cualquier

$\boldsymbol{\nu} \in \boldsymbol{Rn}$, sea $\boldsymbol{\nu} = \nu_1 + \nu_2$ con $\nu_1 \in \boldsymbol{C(P)}$ y $\nu_2 \in \boldsymbol{C(P)}^{\perp}$. Como

$$\boldsymbol{P2\nu} = \boldsymbol{P}(\boldsymbol{P}(\nu_1 + \nu_2)) = \boldsymbol{P}\nu_1 = \nu_1 = \boldsymbol{P\nu},$$

tenemos $\boldsymbol{P2} = \boldsymbol{P}$. Para mostrar la simetría, sea $\boldsymbol{w} = w_1 + w_2$ cualquier otro vector en $\boldsymbol{Rn}$, con $w_1 \in \boldsymbol{C(P)} y w_2 \in \boldsymbol{C(P)}^{\perp}$. Dado que $\boldsymbol{I} - \boldsymbol{P}$ es la matriz de proyección sobre $\boldsymbol{C(P)}^{\perp}$,

$$\boldsymbol{wTPT(I - P)\nu} = \boldsymbol{wT}v_2 = 0.$$

Dado que esto es cierto para cualquier $\boldsymbol{\nu}$ y $\boldsymbol{w}$, tenemos $\boldsymbol{PT(I - P)} = \boldsymbol{0}$, o $\boldsymbol{PT} = \boldsymbol{PTP}$. Como $\boldsymbol{PTP}$ es simétrico, también lo es $\boldsymbol{PT}$ y, por tanto, $\boldsymbol{P}$.

Ahora se muestran dos propiedades útiles sobre los autovalores y el rango de una matriz de proyección.

- Los valores propios de cualquier matriz de proyección $\boldsymbol{P}$ son todos 0 y 1. Esto se deriva de las definiciones de una matriz de proyección y un valor propio.
- Para cualquier matriz de proyección $\boldsymbol{P}$, $rang(\boldsymbol{P}) = tr(\boldsymbol{P})$, que es la suma de sus elementos de la diagonal principal.

La primera propiedad se deriva de las definiciones de una matriz de proyección y un valor propio λ de $\boldsymbol{P}$ con vector propio $\boldsymbol{\nu}$, $\boldsymbol{P\nu} = \lambda\boldsymbol{\nu}$; pero $\boldsymbol{P\nu} = \boldsymbol{\nu}$ (si $\boldsymbol{\nu} \in W$) o $\boldsymbol{P} = 0$ (si $\boldsymbol{\nu} \in \boldsymbol{W}^{\perp}$), entonces $\lambda = 1$ ó 0. De hecho, esta es una propiedad de las matrices idempotentes simétricas.

Y la segunda propiedad se debe a que la traza de una matriz cuadrada es la suma de sus valores propios, y para las matrices simétricas el rango es el número de valores propios distintos de cero. Dado que los autovalores de $\boldsymbol{P}$ (que es simétrico) son todos 0 y 1, la suma de sus autovalores es igual al número de estos distintos de cero.

Una última propiedad útil viene de suponer que $\boldsymbol{P}$ es una matriz $n \times n$ simétrica tal que $\sum P_i = I$. Entonces, las siguientes tres condiciones son equivalentes:

- P_i es idempotente para cada i.
- $P_i P_j = 0$ para todo $i \neq j$.
- $\sum_i rg(P_i) = n$.

Una resultado que se va a emplear es el de que las matrices idempotentes simétricas (y por tanto también matrices de proyección) que satisfacen $\sum_i P_i = I$ también satisfacen $P_i P_j = 0$ para todo $i \neq j$. La prueba de esto es un subproducto de un resultado clave (el teorema de Cochran) sobre formas cuadráticas de la chi-cuadrado que son independientes entre sí.

Sea $\boldsymbol{PX}$ una matriz de proyección en el espacio modelo $C(\boldsymbol{X})$ correspondiente a una matriz modelo $\boldsymbol{X}$ para un modelo lineal. A continuación se presentan varias propiedades para este caso concreto.

- Si $\boldsymbol{X}$ tiene rango completo, $\boldsymbol{PX}$ es la matriz $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T}$.
- Si $\boldsymbol{y} \in C(\boldsymbol{X})$, entonces $\boldsymbol{y} = \boldsymbol{Xb}$ para algún $\boldsymbol{b}$. Entonces

$$\boldsymbol{Hy} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X}T\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X}T\boldsymbol{Xb} = \boldsymbol{Xb} = \boldsymbol{y}.$$

- Si $\boldsymbol{X}$ no tiene rango completo, entonces $\boldsymbol{PX} = \boldsymbol{X}(\boldsymbol{X^T X})^{-}\boldsymbol{X}T$. Además, $\boldsymbol{PX}$ es invariante a la elección del inverso generalizado $(\boldsymbol{X^T X})^{-}$.
- Si $\boldsymbol{X}$ y $\boldsymbol{W}$ son matrices modelo que satisfacen $C(\boldsymbol{X}) = C(\boldsymbol{W})$, entonces $\boldsymbol{PX} = \boldsymbol{PW}$.

Cuando el modelo a es un caso especial del modelo b, con matrices de proyección P_a y P_b para matrices modelo X_a y X_b, entonces tenemos que $P_aP_b = P_bP_a = P_a$ y $P_b - P_a$ también será una matriz de proyección.

Ejemplo 8. A continuación, se ilustra la geometría que subyace a las proyecciones de modelos lineales. Se hace esto para dos modelos simples, para los cuales se pueden representar fácilmente las proyecciones gráficamente. El primer modelo tiene una única variable explicativa cuantitativa

$$\mu_i = E(y_i) = \beta x_i, i = 1, \cdots, n$$

y no contiene constante. Su matriz modelo X es un vector $n \times 1$ $(x_1, \cdots, x_n)^T$. La Figura 9.1 muestra el modelo, los datos y el ajuste. Los valores de la variable respuesta $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ son un punto en $\mathbf{R}^n$. Los valores de la variable explicativa $\boldsymbol{X}$ son otro de esos puntos. El predictor lineal valora $\boldsymbol{X}\beta$ para todos los posibles valores reales de $\boldsymbol{\beta}$ y traza una recta en $\mathbf{R}^n$ que pasa por el origen. Este es el espacio modelo $C(\boldsymbol{X})$. El ajuste del modelo $\widehat{\boldsymbol{\mu}} = \boldsymbol{P_X y} = \widehat{\beta}\boldsymbol{X}$ es la proyección ortogonal de $\boldsymbol{y}$ sobre la recta del espacio modelo.

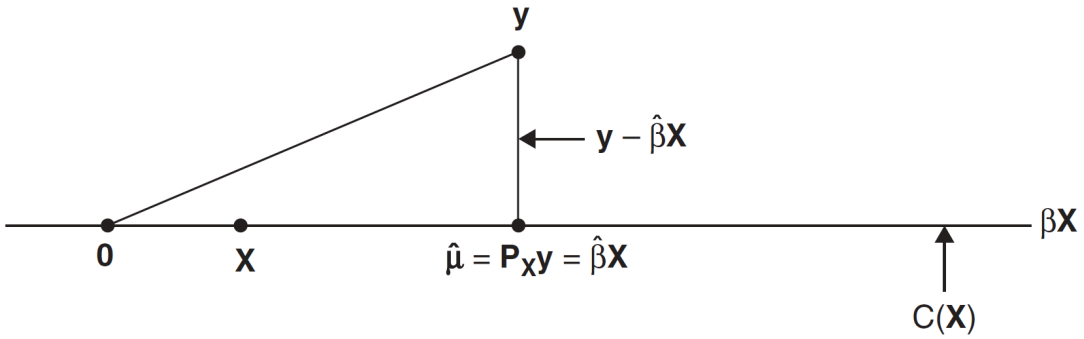

Figura 9.1: Esquema vectorial de un modelo lineal simple con un sólo predictor y sin constante. Se muestran las observaciones $\boldsymbol{y}$, la matriz del modelo $\boldsymbol{X}$ de los valores predichos, y el ajuste $\widehat{\boldsymbol{\mu}} = \boldsymbol{P_X y} = \widehat{\beta}\boldsymbol{X}$.

A continuación, se amplia el modelo para manejar dos variables explicativas cuantitativas. Se consideran los modelos

$$E(y_i) = \beta_{y1}x_{i1}, E(y_i) = \beta y_2 x_{i2}, E(y_i) = \beta y_{1\cdot 2}x_{i1} + \beta y_{2\cdot 1}x_{i2}$$

Se emplea la notación de Yule para reflejar que $\beta y_{1\cdot 2}$ y $\beta y_{2\cdot 1}$ normalmente difieren de βy_1 y βy_2. La Figura 9.2 muestra los datos y los tres ajustes del modelo. Cuando se evalúa para todos los $\beta y_{1\cdot 2}$ reales y $\beta y_{2\cdot 1}$, $\boldsymbol{\mu}$ traza un plano en $\mathbf{R}^n$ que pasa por el

origen. La proyección $P_{12}\boldsymbol{y} = \widehat{\beta} y_{1\cdot 2} X_1 + \widehat{\beta} y_{2\cdot 1} X_2$ da el ajuste de mínimos cuadrados usando ambos predictores juntos. La proyección $P_1\boldsymbol{y} = \widehat{\beta} y_1 X_1$ en el espacio modelo para $X_1 = (x_{11}, \cdots, x_{n1})^T$ da el ajuste de mínimos cuadrados cuando x_1 es el único predictor. La proyección $P_2\boldsymbol{y} = \widehat{\beta} y_2 X_2$ en el espacio modelo para $X_2 = (x_{12}, \cdots, x_{n2})^T$ da el ajuste de mínimos cuadrados cuando x_2 es el único predictor.

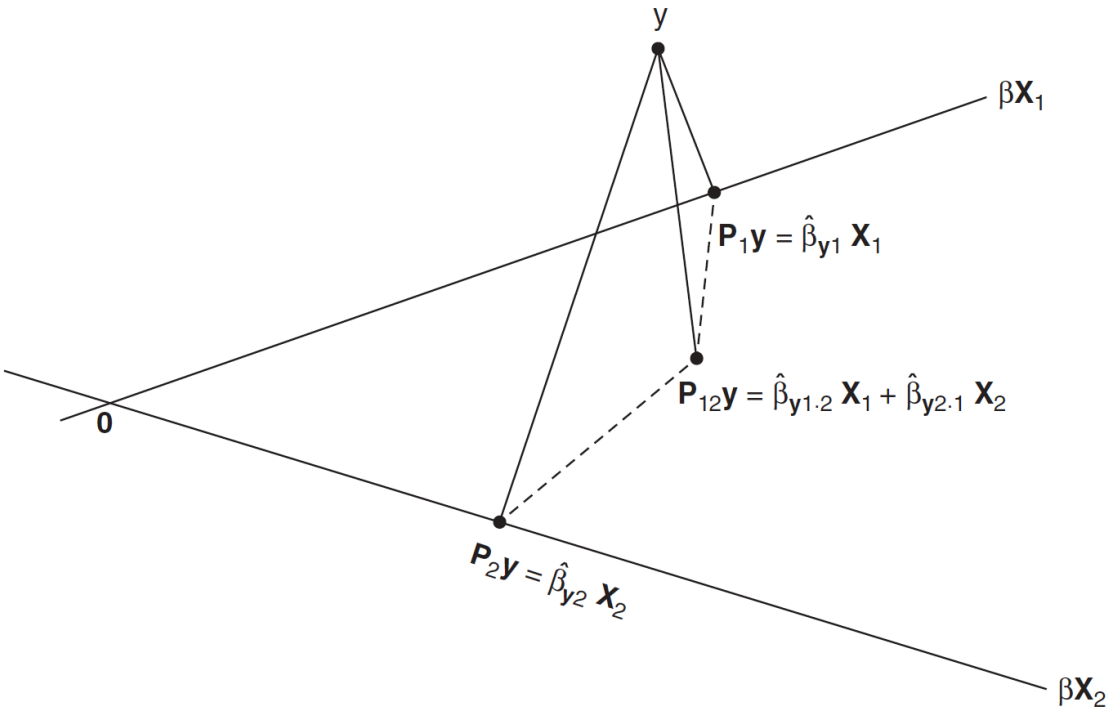

Figura 9.2: Esquema de un modelo lineal con dos variables predictoras, mostrando las observaciones $\boldsymbol{y}$, y las ajustadas $P_1 y = \widehat{\beta} y_1 X_1$, $P_2 y = \widehat{\beta} y_2 X_2$, y $P_{12} y = \widehat{\beta} y_{1\cdot 2} X_1 + \widehat{\beta} y_{2\cdot 1} X_2$.

9.4 Resumen de la variabilidad en un modelo lineal

Para un modelo lineal $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$ con matriz de modelo $\boldsymbol{X}$ y matriz de covarianzas $\boldsymbol{V} = \sigma^2 \boldsymbol{I}$, ya introdujimos la descomposición ortogonal 'datos = ajuste + residuos' por medio de la matriz de proyección $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ (es decir, la matriz H),

$$\boldsymbol{y} = \widehat{\boldsymbol{\mu}} + (y - \widehat{\boldsymbol{\mu}}) = \boldsymbol{P_X}\boldsymbol{y} + (\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{y}.$$

Esto representa la ortogonalidad de los valores ajustados $\widehat{\boldsymbol{\mu}}$ y los residuos brutos $\boldsymbol{e} = (\boldsymbol{y} - \widehat{\boldsymbol{\mu}})$. Hemos utilizado $\boldsymbol{P_X}\boldsymbol{y} = \widehat{\boldsymbol{\mu}}$ para estimar $\boldsymbol{\mu}$. La otra parte de esta descomposición, $(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{y} = (\boldsymbol{y} - \widehat{\boldsymbol{\mu}})$, cae en el espacio de error $C(X)^{\perp}$ ortogonal al espacio modelo $C(\boldsymbol{X})$. Entonces se va a usar para estimar la varianza σ^2 de la distribución condicionada de cada y_i, respecto a los valores de las variables explicativa. Esta varianza a veces se denomina varianza del error, a partir de la representación del modelo como $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$ con $var(\epsilon) = \sigma^2 \boldsymbol{I}$.

Para obtener un estimador insesgado de σ^2, se va a aplicar el resultado sobre $E(\boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y})$, para una matriz $n \times n$ A. Como $E(\boldsymbol{y} - \boldsymbol{\mu}) = 0$,

$$E[(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{\mu})] = E(\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}) - \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}.$$

Y usando la propiedad commutativa de la traza de una matriz se tiene que:

$$E[(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{\mu})] = Etr[(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{\mu})] = Etr[\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})^T] =$$
$$= tr\boldsymbol{A}E[(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})^T] = tr(\boldsymbol{A}\boldsymbol{V}).$$

A lo que sigue que,

$$E(\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}) = tr(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}. \tag{9.6}$$

Para un modelo lineal con matriz de modelo de rango completo $\boldsymbol{X}$ y matriz de proyección $\boldsymbol{P_X}$, ahora aplicamos este resultado con $\boldsymbol{A} = (\boldsymbol{I} - \boldsymbol{P_X})$ y $\boldsymbol{V} = \sigma^2\boldsymbol{I}$ para la matriz de identidad $n \times n$, $\boldsymbol{I}$. El rango de $\boldsymbol{X}$, que también es el rango de $\boldsymbol{P_X}$, es el número de parámetros del modelo p. Tenemos

$$E[\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{y}] = tr[(\boldsymbol{I}-\boldsymbol{P_X})\sigma^2\boldsymbol{I}] + \boldsymbol{\mu}^T(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{\mu} = \sigma^2 tr(\boldsymbol{I}-\boldsymbol{P_X})$$

como $(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{\mu} = \boldsymbol{\mu} - \boldsymbol{\mu} = 0$. Entonces

$$tr(\boldsymbol{P_X}) = tr[\boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}^T] = tr[\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^T)^{-1}] = tr(\boldsymbol{I_p})$$

donde I_p es la matriz identidad $p \times p$,

Lo que, tras otras pocas operaciones (ver Agresti 2015 para más detalles en la demostración), nos conduce a que el estimador insesgado de la varianza del error σ^2 en un modelo lineal con una matriz modelo de rango máximo es

$$s^2 = \frac{\sum_{i=1}^n (y_i - \widehat{\mu}_i)^2}{n-p}$$

el cual es un promedio de los residuos al cuadrado. Aquí, el promedio se toma con respecto a la dimensión del espacio del error en el que residen estos componentes residuales. Cuando $\boldsymbol{X}$ presenta menor rango al completo o máximo $r < p$, el mismo argumento se mantiene para $tr(\boldsymbol{P_X}) = r$. Entonces, s^2 tiene denominador $n - r$. La estimación s^2 se llama *error cuadrático medio*, donde *error* = *residuo*, y por tanto también se puede expresar cuadrado medio residual. Por ejemplo, en el caso del modelo nulo, el numerador de s^2 es $\sum_i^n (y_i - \overline{y})^2$ y el rango de $\boldsymbol{X} = 1_n$ es 1. Y un estimador insesgado de σ^2 es

$$s^2 = \frac{\sum_{i=1}^n (y_i - \overline{y}_i)^2}{n-1}$$

Esta es la varianza muestral, empleada usualmente como estimador de la varianza marginal de y.

9.4.1 Descomposición de la Suma Total de Cuadrados

La suma de cuadrados se designa SCE , y es la 'suma de errores cuadrados', también se conoce como *suma de cuadrados de los residuos* (*Sum of Squared Errors*). La descomposición ortogonal de los datos resulta $\boldsymbol{y} = \boldsymbol{P_x y} + (\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{y}$, y expresa la observación i como $y_i = \mu_i + (y_i - \widehat{\mu}_i)$. La que corrigiendo por la media muestral, queda así:

$$(y_i - \overline{y}) = (\widehat{\mu}_i - \overline{y}) + (y_i - \widehat{\mu}_i)$$

Usar $(y_i - \overline{y})$ como la observación corresponde a ajustar y_i por medio de la inclusión de un término de intersección antes de investigar los efectos de las variables explicativas. (para el modelo nulo $E(y_i) = \beta$, se demostró que $\widehat{\mu}_i = \overline{y}$.) Esta descomposición ortogonal en el componente en el espacio modelo y el componente en el espacio de error produce la descomposición por suma de cuadrados:

$$\sum_i (y_i - \overline{y})^2 = \sum_i (\widehat{\mu}_i - \overline{y})^2 + \sum_i (y_i - \widehat{\mu}_i)^2$$

Descomposición que se abrevia del siguiente modo:

$$SCT = SCR + SCE$$

para la suma total (corregida) de cuadrados SCT, la suma de cuadrados debida al modelo de regresión SCR y la suma cuadrática de los errores SCE. La SCT resume la variación total en los datos después de ajustar el modelo que contiene solo una intersección. El componente SCE representa la variación en y 'no explicada' por el modelo completo, es decir, un resumen del error de predicción restante después de ajustar ese modelo. El componente SCR representa la variación en y 'explicada' por el modelo completo, es decir, la reducción en la variación de SCT a SCE resultante de agregar variables explicativas a un modelo que contiene solo un término de intersección.

9.4.2 ¿Cómo afecta a la SCE y la SCR la inclusión de variables al modelo?

Cuando agregamos una variable explicativa a un modelo, SCE no puede aumentar, pues podríamos (en el peor de los casos) obtener el mismo valor de SCE estableciendo $\widehat{\beta}_j = 0$ para la nueva variable. Entonces, SCE es monótono decreciente a medida que crece el conjunto de variables explicativas. Dado que SCT depende solo de $\{y_i\}$ y es idéntico para cada modelo ajustado a un conjunto de datos en particular, $SCR = SCT - SCE$ es monótono y aumenta a medida que se agregan variables.

Denotemos por $SCR(x_1, x_2)$ a la suma de cuadrados de regresión para un modelo con dos variables explicativas, y que $SCR(x_1)$ y $SCR(x_2)$ denoten a los dos modelos que tienen solo una de esas variables explicativas (más, en cada caso, la intersección). Podemos dividir $SCR(x_1, x_2)$ en $SCR(x_1)$ y la variabilidad adicional explicada agregando x_2 al modelo. Y además, se denota esta variabilidad adicional explicada por x_2, una vez se ha ajustado por x_1, por $SCR(x_2|x_1)$. Es decir,

$$SCR(x_1, x_2) = SCR(x_1) + SCR(x_2|x_1).$$

O equivalentemente, $SCR(x_2|x_1)$ es lo que decrece la SCE al añadir x_2 al modelo.

Si $\widehat{\mu}_{i1}$ denota los valores ajustados cuando x_1 es la única variable explicativa, y $\widehat{\mu}_{i12}$ denota los valores ajustados cuando tanto x_1 como x_2 son variables explicativas. Entonces, a partir de la descomposición ortogonal $(\widehat{\mu}_{i12} - \overline{y}) = (\widehat{\mu}_{i1} - \overline{y}) + (\widehat{\mu}_{i12} - \widehat{\mu}_{i1})$, se tiene

$$SCR(x_2|x_1) = \sum_{i=1}^{n} (\widehat{\mu}_{i12} - \widehat{\mu}_{i1})^2.$$

A continuación, se considera el caso general con p variables explicativas, $x_1, x_2, \cdots, x_p$, y una intersección o valor centrado de y. Al ingresar estas variables en secuencia en el modelo, obtenemos la suma de regresión de cuadrados e incrementos sucesivos

$$SCR(x_1), SCR(x_2|x_1), SCR(x_3|x_1, x_2), \cdots, SCR(x_p|x_1, x_2, \cdots, x_{p-1})$$

Estos componentes se conocen como sumas secuenciales de cuadrados. Suman la suma de cuadrados de regresión para el modelo completo, denotado por $SCR(x_1, \cdots, x_p)$. La suma secuencial de cuadrados correspondientes a la adición de un término al modelo puede depender en gran medida de qué otras variables ya están en el modelo. Esto es debido a las correlaciones existentes entre los predictores. Por ejemplo, $SCR(x_p)$ a menudo tiende a ser mucho más grande que $SCR(x_p|x_1, \cdots, x_{p-1})$ cuando x_p está altamente correlacionado con los otros predictores, como ocurre en muchos estudios observacionales.

Un conjunto de incrementos alternativo a las sumas de cuadrados de regresión, llamado *sumas parciales de cuadrados*, usa el mismo conjunto de p variables explicativas para cada uno:

$$SCR(x_1|x_2, \cdots, x_p), SCR(x_2|x_1, x_3, \cdots, x_p), \cdots, SCR(x_p|x_1, \cdots, x_{p-1})$$

Cada uno de estos representa la variabilidad adicional explicada al agregar una variable explicativa particular al modelo, cuando todas las demás variables explicativas ya están incluidas en el modelo. Estos valores SCR parciales pueden diferir de todos los valores SCR secuenciales correspondientes $SCR(x_1), SCR(x_2|x_1), \ldots, SCR(x_p|x_1, x_2, \ldots, x_{p-1})$, excepto el final.

9.4.3 R-cuadrado y la correlación múltiple

Para un conjunto de datos y un valor de SCT en particular, cuanto mayor sea el valor de SCR en relación con SCE, más efectivas serán las variables explicativas para predecir la variable respuesta. Un resumen de este poder predictivo va a ser recogido por la siguiente medida.

Definición 6

$$R^2 = \frac{SCR}{SCT} = \frac{SCT - SCE}{SCT} = \frac{\sum_i (y_i - \overline{y})^2 - \sum_i (y_i - \widehat{\mu}_i)^2}{\sum_i (y_i - \overline{y})^2} \tag{9.7}$$

R^2 mide la reducción proporcional del error, y sus valores oscilan entre 0 y 1. También se le suele denominar coeficiente de determinación, o $\boldsymbol{R - cuadrado}$.

Esta medida está relacionada con la correlación:

$$corr(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}) = \sqrt{SCR/SCT} = +\sqrt{R^2}$$

A esta raíz cuadrada positiva de R^2 se le llama correlación múltiple. Hay que tener en cuenta que $0 \leq R \leq 1$. Con una sola variable explicativa, $R = |corr(\boldsymbol{x}, \boldsymbol{y})|$.

Cuando se agregan variables explicativas a un modelo, dado que SCE no puede aumentar, R y R^2 son monótonos crecientes. Cuando n es pequeño y un modelo tiene varias variables explicativas, R^2 tiende a sobreestimar el valor poblacional correspondiente. Por esto se introduce la siguiente medida que va a compensar o ajustar el valor de R^2.

Definición 7

El **R-cuadrado ajustado** está diseñado para reducir este sesgo. Se define como la reducción proporcional de la varianza basada en las estimaciones de varianza insesgadas, s_y^2 para la distribución marginal y s^2 para las distribuciones condicionales;

$$R^2 ajustado = 1 - \frac{n-1}{n-p}(1 - R^2) \tag{9.8}$$

El R^2 ajustado es ligeramente menor que la versión normal de R^2, y no es necesariamente monótono creciente según se introducen más variables explicativas al modelo.

9.5 Residuos, apalancamiento (leverage) e influencia

Dado que los residuos del ajuste del modelo lineal están en el espacio de error, ortogonal al espacio del modelo, contienen la información de los datos que el modelo no explica. Por lo tanto, son útiles para investigar la falta de ajuste de un modelo. En este apartado se analizan más de cerca los residuos, incluidos sus momentos y las formas de graficarlos. Todo esto va a ayudar a verificar la calidad o 'bondad' del modelo. También se presenta el concepto de 'influencia' que tiene cada observación en el ajuste de mínimos cuadrados, utilizando los valores residuales y de 'apalancamiento' de la matriz de influencia H.

Tal y como se ha visto, la ecuación normal correspondiente al término β_0, que es la constante del modelo, es $\sum_i y_i = \sum_i \widehat{\mu}_i$. Y de este modo, $\sum_i e_i = \sum_i (y_i - \widehat{\mu}_i) = 0$, y los residuos tienen una media muestral de 0.

$$E(\boldsymbol{e}) = E(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) = \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}E(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}.$$

Además, al ser ortogonales $\boldsymbol{e}$ y $\widehat{\boldsymbol{\mu}}$ esto implica que $corr(\boldsymbol{e}, \widehat{\boldsymbol{\mu}}) = 0$.

9.5.1 Gráficas de residuos

Como $corr(\boldsymbol{e}, \widehat{\boldsymbol{\mu}}) = 0$, la recta de mínimos cuadrados ajustada a una gráfica de dispersión de los elementos de $\boldsymbol{e} = (\boldsymbol{y} - \widehat{\boldsymbol{\mu}})$ versus los elementos correspondientes de $\widehat{\boldsymbol{\mu}}$ tiene pendiente 0. Un diagrama de dispersión de los residuos frente a los valores ajustados ayuda a identificar patrones de falta de ajuste de un modelo. Algunos ejemplos son la varianza no constante, a veces denominada heterocedasticidad, y la no linealidad. Asimismo, dado que los residuos también son ortogonales a $C(\boldsymbol{X})$, se pueden graficar contra cada variable explicativa para detectar la falta de ajuste.

La Figura 9.3 muestra cómo una gráfica de $\boldsymbol{e}$ contra $\widehat{\boldsymbol{\mu}}$ tiende a verse si (a) el modelo lineal se cumple, (b) la varianza es constante (homocedasticidad), pero la media de y es una función cuadrática en lugar de lineal del predictor, y (c) el predictor de tendencia lineal es correcto, pero la varianza aumenta drásticamente a medida que aumenta la media. Lógicamente, en la práctica, las gráficas no tienen una apariencia tan 'limpia' y clara, pero estos ejemplos 'ideales' ilustran cómo las gráficas pueden resaltar lo inedacuado de un modelo.

Para el modelo lineal normal, la distribución condicional de $\boldsymbol{y}$, dadas las variables explicativas, es normal. Esto implica que los residuos, que son lineales en $\boldsymbol{y}$, también tienen distribuciones normales. Un histograma de los residuos proporciona información sobre la distribución condicional real. Otra comprobación del supuesto de normalidad se hace por medio de una gráfica de valores residuales, los cuales se ordenan frente a los valores ordenados esperados en una distribución $N(0, 1)$. Este gráfico se denomina $\boldsymbol{Q} - \boldsymbol{Q}$ (del inglés *quantile VS quantile plot*).

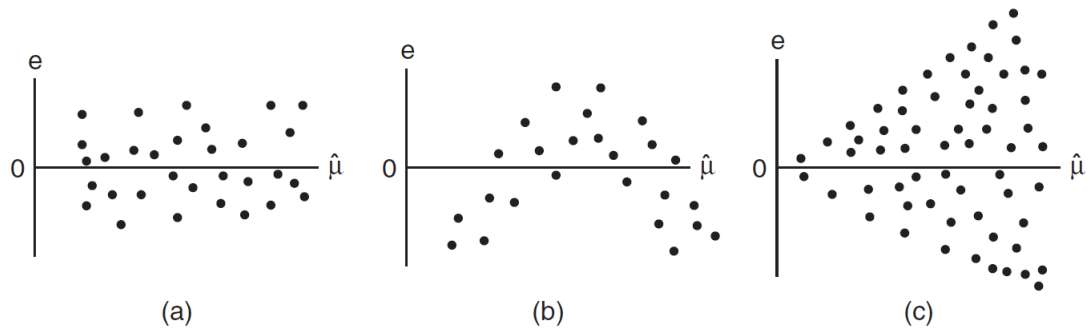

Figura 9.3: Los residuos graficados juntos a los valores ajustados del modelo reflejan (a) el modelo es adecuado, (b) relación cuadrática en lugar de lineal, y (c) varianza no constante.

Ejemplo 9. A continuación, se muestran unos ejemplos de estos gráficos de residuos. Para estos datos, el histograma de la Figura 9.4 sugiere que la distribución condicionada

de $\boldsymbol{y}$ tiene forma de montículo, pero posiblemente sesgada hacia la derecha. Además, la observación 64 tiene un residuo estandarizado negativo relativamente grande de $-4,2$. La gráfica Q–Q también muestra evidencia de sesgo hacia la derecha, porque los cuantiles teóricos positivos grandes tienen cuantiles muestrales que son más grandes en valor absoluto, mientras que los cuantiles teóricos negativos grandes tienen cuantiles muestrales que son más pequeños en valor absoluto (excepto el valor atípico). Sin embargo, es difícil juzgar bien la forma, a menos que n sea suficientemente grande y la tasa de error real para la inferencia estadística bilateral sobre los parámetros β_j en el modelo lineal sea robusta a las violaciones del supuesto de normalidad. La insuficiencia de la inferencia estadística y las consiguientes conclusiones sustantivas suelen verse más afectadas por un predictor lineal inadecuado (es decir, que carece de una interacción importante) y por problemas prácticos de muestreo (es decir, datos faltantes, errores de medición) que por la no normalidad de la respuesta. Con residuos claramente atípicos, se puede transformar $\boldsymbol{y}$ para mejorar la normalidad. Pero el predictor lineal puede entonces describir peor la relación, y los efectos sobre $E[g(\boldsymbol{y})]$ son de menos interés que los efectos sobre $E(\boldsymbol{y})$. Por lo tanto, se recomiendan estos gráficos principalmente para ayudar a detectar observaciones inusuales que podrían derivar en conclusiones erróneas.

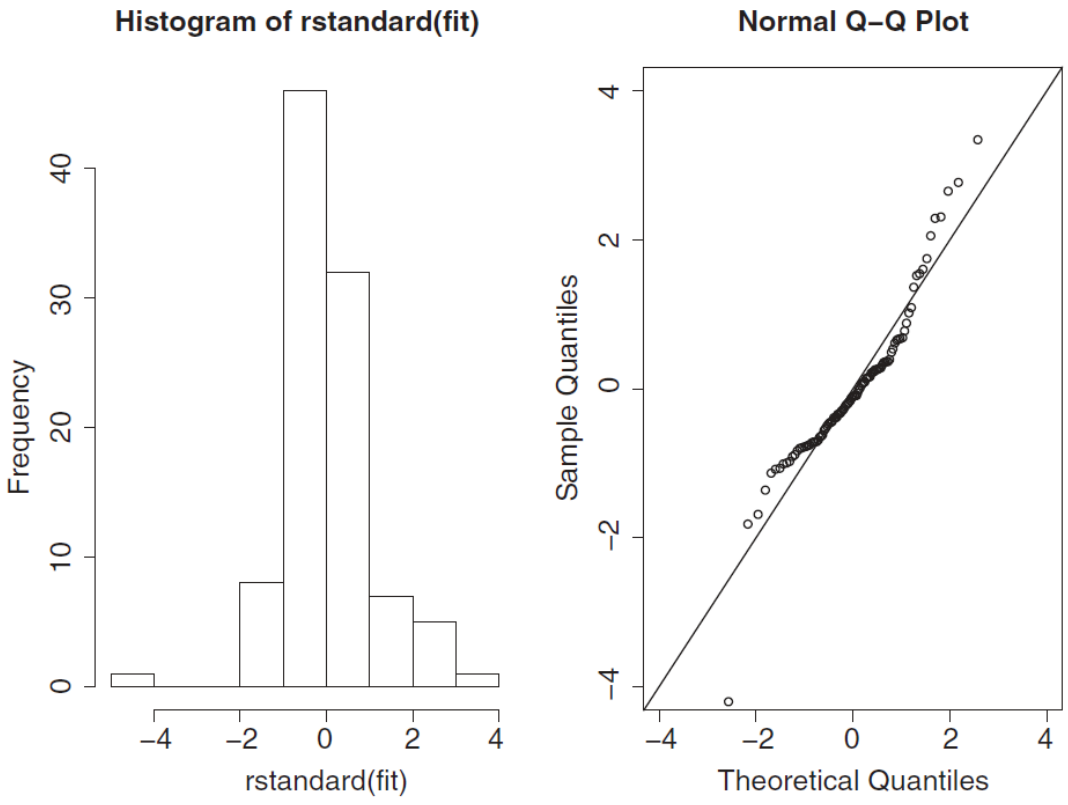

Figura 9.4: Histograma y gráfico Q–Q de residuos estandarizados, para el modelo lineal normal que predice el precio de venta usando el tamaño y el ser nuevo como variables explicativas.

Para investigar la adecuación del predictor lineal, graficamos los residuos contra los valores ajustados (ver Figura 9.5) y contra el tamaño. Si se cumple el modelo lineal normal, una gráfica de los residuos contra los valores ajustados o los valores de las variables explicativas debe mostrar un patrón aleatorio de aproximadamente 0 con una variabilidad relativamente constante. La Figura 9.5 también destaca la inusual observación 64, pero a modo general no parece indicar falta de ajuste. Existe una sugerencia de que los residuos pueden tender a ser mayores en valor absoluto a valores más altos de la respuesta. En lugar de una varianza constante, parece plausible que la varianza sea mayor a precios de venta medios más altos.

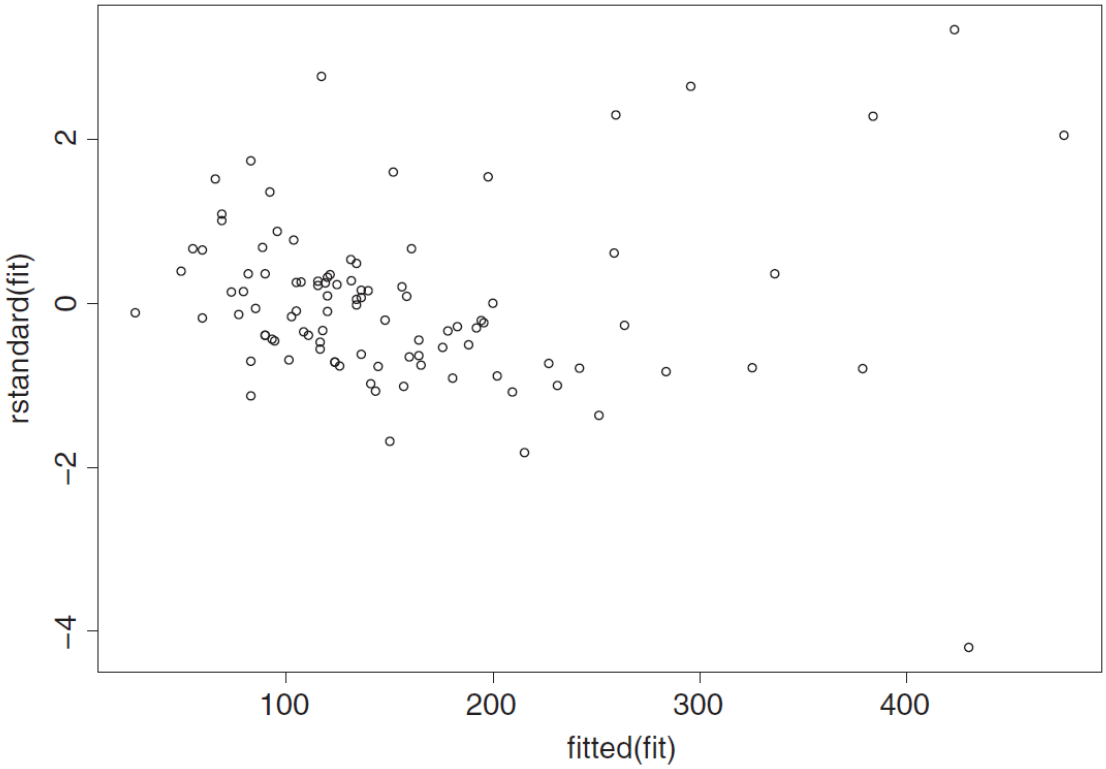

Figura 9.5: Gráfico de residuos estandarizados *versus* valores ajustados, para el modelo lineal que predice el precio de venta usando el tamaño y el ser nuevo como variables explicativas.

9.5.2 Residuos estandarizados

Para el modelo lineal ordinario $var(\widehat{\mu}_i) = \sigma^2 h_{ii}$, donde $\{h_{ii}\}$ son los elementos diagonales principales de $\boldsymbol{H}$. Además, los residuos están correlacionados y su varianza no necesita ser constante, con

$$var(e_i) = var(y_i - \widehat{\mu}_i) = \sigma^2(1 - h_{ii})$$

Y, dado que las varianzas son no negativas, $0 \leq h_{ii} \leq 1$. En la práctica σ es desconocido, por lo que lo reemplazamos por su estimación: s.

Definición 8

El **residuo estandarizado** es una versión estandarizada de $e_i = (y_i - \widehat{\mu}_i)$ que divide a éste por $\sigma\sqrt{1-h_{ii}}$ (o s al ser la desviación teórica desconocida) y de este modo logra una desviación estándar de 1. Por tanto, este va a describir el número de desviaciones típicas que $(y_i - \widehat{\mu}_i)$ se desvía de 0.

$$r_i = \frac{y_i - \widehat{\mu}_i}{s\sqrt{1-h_{ii}}} \tag{9.9}$$

Si el modelo lineal normal realmente se mantiene, casi todos los residuos deberían estar entre -3 y $+3$.

Un residuo ligeramente diferente al estandarizado, el llamado **residuo estudentizado**, estima σ en la expresión para $var(y_i - \widehat{\mu}_i)$ y está basado en el ajuste del modelo con las $n-1$ observaciones, y tras haber excluido el efecto de la propia observación i. De este modo se va a lograr que esta estimación sea independiente de la observación i.

9.5.3 Apalancamiento *-leverage-* e influencia

Definición 9

El elemento h_{ii} de la matriz $\boldsymbol{H}$, del cual depende $var(e_i)$, se llama **el apalancamiento *-leverage-* de la observación** i.

Dado que $var(\widehat{\mu}_i) = \sigma^2 h_{ii}$ con $0 \leq h_{ii} \leq 1$, el apalancamiento determina la precisión con la que $\widehat{\mu}_i$ estima μ_i. Para h_{ii} grandes cercanos a 1, $var(\widehat{\mu}_i) \approx var(y_i)$ y $var(e_i) \approx 0$. En este caso, y_i puede tener una gran influencia sobre $\widehat{\mu}_i$. En el caso extremo $h_{ii} = 1, var(e_i) = 0$ y $\widehat{\mu}_i = y_i$. Por el contrario, cuando h_{ii} está cerca de 0 y $var(\widehat{\mu}_i)$ es relativamente pequeño, esto sugiere que $\widehat{\mu}_i$ se está basando en las contribuciones de muchas observaciones.

Aquí, cabe preguntarse: ¿qué aspecto tienen estos apalancamientos? Para el modelo lineal bivariado $E(yi) = \beta_0 + \beta x_i$, el apalancamiento para la observación i es

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{k=1}^{n}(x_k - \overline{x})^2}$$

Los n apalancamientos tienen una media de $\frac{2}{n}$, y tienden a ser más pequeños con conjuntos de datos más grandes. Considerando múltiples variables, el apalancamiento aumenta a medida que $\boldsymbol{x_i}$ está más lejos de $\overline{\boldsymbol{x}}$. Con p variables explicativas, incluida la intersección, los apalancamientos tienen una media de $\frac{p}{n}$. Las observaciones con apalancamientos relativamente grandes, digamos que con valores alrededor de $\frac{3p}{n}$, pueden influir en el proceso de ajuste.

Una observación que presenta un apalancamiento pequeño no influye fuertemente en $\widehat{\mu}_i$ y $\widehat{\beta}_j$, incluso si es un valor atípico en la dirección y. Un punto con un apalancamiento extremadamente grande puede ser influyente, pero no tiene por qué serlo. Es influyente cuando la observación sea un valor atípico (*outlier*) que caiga lejos de la línea de mínimos

cuadrados que resulta cuando se usan solo las $n-1$ observaciones restantes. Este caso se aprecia en el primer panel de la Figura 9.6. Por el contrario, cuando la observación tiene un gran apalancamiento, pero es consistente con la tendencia mostrada por las otras observaciones, no es influyente. Esto se refleja en el segundo panel de la Figura 9.6. Para ser influyente, un punto debe tener tanto un gran apalancamiento como un gran residuo estandarizado.

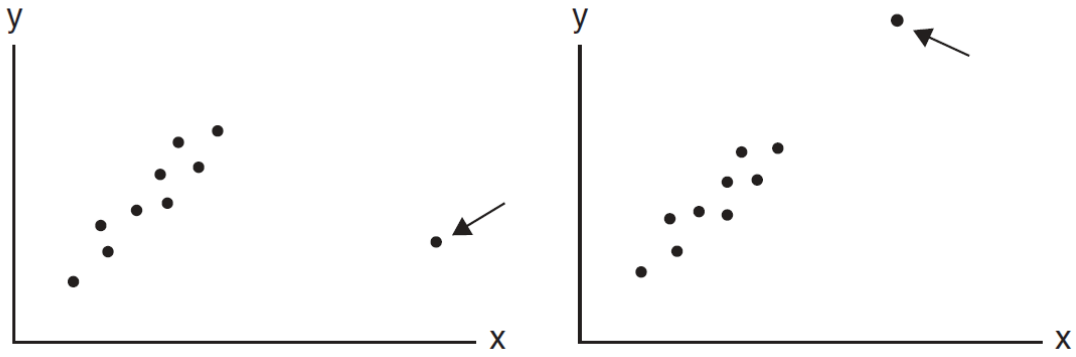

Figura 9.6: Puntos con un apalancamiento alto en un modelo lineal ajustado pueden ser influyentes (primer panel) o no (segundo panel).

Las medidas de resumen que describen la influencia de una observación combinan información de los apalancamientos y los residuos. Para cualquier medida de influencia de este tipo, valores mayores corresponden a una mayor influencia.

Definición 10

La **distancia de Cook** (Cook 1977) es una medida de la importancia de un punto que se basa en el cambio en $\widehat{\boldsymbol{\beta}}$ cuando la observación se elimina del conjunto de datos.

Sea $\widehat{\boldsymbol{\beta}}(i)$ la estimación de mínimos cuadrados de $\boldsymbol{\beta}$ para las $n-1$ observaciones que restan al excluir la observación i, entonces, la distancia de Cook para la observación i es

$$D_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^T [\widehat{var}(\widehat{\boldsymbol{\beta}})]^{-1} (\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p} = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^T (\boldsymbol{X}^T \boldsymbol{X}) (\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{ps^2}$$

La incorporación de la varianza estimada de $\widehat{\boldsymbol{\beta}}$ hace que la medida esté libre de las unidades de medida y aproximadamente libre del tamaño de la muestra. Un D_i relativamente grande, generalmente del orden de 1, ocurre cuando tanto el residuo estandarizado como el apalancamiento son relativamente grandes.

Una medida con un propósito similar, DFFIT, describe el cambio en $\widehat{\mu}_i$ debido a la eliminación de la observación i. Una versión estandarizada (DFFITS) es igual al residuo studentizado multiplicado por el 'factor de apalancamiento' $\sqrt{h_{ii}/(1-h_{ii})}$. Otra

medida, en este caso específica de una variable, es DFBETA (y su versión estandarizada DFBETAS). Ésta se basa en el cambio en $\widehat{\beta}_j$ cuando la observación se elimina del conjunto de datos. Cada observación tiene una DFBETA separada para cada $\widehat{\beta}_j$.

9.6 Optimalidad de mínimos cuadrados y de mínimos cuadrados generalizados

En este tema, hasa ahora se han usado el método de mínimos cuadrados para estimar los parámetros del modelo lineal ordinario, para lo que se asume independencia entre observaciones y varianza constante. A continuación, se muestra un criterio por el cual dichos estimadores son óptimos. Luego se generalizan los mínimos cuadrados para permitir que las observaciones estén correlacionadas y tengan una varianza no constante. Para el modelo lineal ordinario, los mínimos cuadrados proporcionan el mejor estimador posible de los parámetros del modelo, pero esto es cierto bajo cierta restricción que señala el siguiente teorema:

Teorema 1

Teorema de Gauss–Markov: Supongamos que $E(\boldsymbol{y}) = \boldsymbol{X\beta}$, donde $\boldsymbol{X}$ tiene rango completo, y $var(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$. El estimador de mínimos cuadrados $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ es el *mejor estimador lineal insesgado* (BLUE en inglés) de $\boldsymbol{\beta}$, en el siguiente sentido: Para cualquier $\boldsymbol{a}^T\boldsymbol{\beta}$, de los estimadores que son lineales en $\boldsymbol{y}$ e insesgados, $\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}$ tiene mínima varianza.

Para probar esto, expresemos $a^T\widehat{\boldsymbol{\beta}}$ en su forma lineal en y como

$$a^T\widehat{\boldsymbol{\beta}} = a^T(X^TX)^{-1}X^Ty = c^Ty$$

donde $c^T = a^T(X^TX)^{-1}X^T$. Supongamos que b^Ty es un estimador lineal alternativo $a^T\boldsymbol{\beta}$ que es insesgado. Entonces

$$E(b-c)^Ty = E(b^Ty) - E(c^Ty) = a^T\boldsymbol{\beta} - a^T\boldsymbol{\beta} = 0$$

para todo $\boldsymbol{\beta}$. Pero esto equivale también a $(b-c)^TX\boldsymbol{\beta} = [\boldsymbol{\beta}^TX^T(b-c)]^T$ para todo $\boldsymbol{\beta}$. Entonces, $X^T(b-c) = 0$. Así que, $(b-c)$ está en el espacio residual $C(X) \perp= N(X^T)$ para el modelo. Y

$$var(b^Ty) = var[cTy + (b-c)Ty] = var(cTy) + ||b-c||2\sigma^2 + 2cov[c^Ty, (b-c)Ty]$$

Pero al ser $X^T(b-c) = 0$,

$$cov[c^Ty, (b-c)^Ty] = c^Tvar(y)(b-c) = \sigma^2a^T(X^TX)^{-1}X^T(b-c) = 0$$

Entonces, $var(b^Ty) \geq var(c^Ty) = var(a^T\widehat{\boldsymbol{\beta}})$, con igualdad sí y solo sí $b = c$.

Si se tiene 1 en la posición j y 0 en otra parte, entonces, el teorema de Gauss-Markov implica que, para todo $j, var(\widehat{\beta}_j)$ toma el valor mínimo de todos los estimadores lineales

insesgados de β_j. A primera vista, el teorema de Gauss-Markov es muy potente, sin embargo, la restricción a estimadores lineales e insesgados es severa.

El modelo lineal ordinario, para el cual $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$ con $var(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$, asume que las observaciones de la variable respuesta tienen varianzas idénticas y no están correlacionadas. En la práctica, esto a menudo no va a ser admisible. Con los datos de recuento, por ejemplo, la varianza suele ser mayor cuanto mayor sea la media. O con las observaciones de las series temporales, pues los datos cercanos en el tiempo a menudo están altamente correlacionadas. Con los datos de encuestas, los diseños muestrales suelen ser más complejos que el muestreo aleatorio simple, y los analistas ponderan las observaciones para que reciban la influencia adecuada.

Un modelo lineal con una estructura más general para la matriz de covarianzas es $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$ con $var(\boldsymbol{y}) = \sigma^2 V$ donde V no necesita ser la matriz identidad.

Definición 11

El estimador $\widehat{\boldsymbol{\beta}}_{GLS}$ recibe el nombre de **estimador de mínimos cuadrados generalizado $\boldsymbol{\beta}$**. Y su expresión es:

$$\widehat{\boldsymbol{\beta}}_{GLS} = [(\boldsymbol{X}^*)^T\boldsymbol{X}^*]^{-1}(\boldsymbol{X}^*)^T\boldsymbol{y}^* = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}. \tag{9.10}$$

Cuando $\boldsymbol{V}$ es diagonal y $var(y_i) = \sigma^2/w_i$ para un peso positivo conocido w_i, como en el caso de un diseño muestral que de más peso a unas observaciones que a otras, $\widehat{\boldsymbol{\beta}}_{GLS}$ también se conoce como **estimador de mínimos cuadrados ponderados**.

Bibliografía

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley.

Cook, R Dennis (1977). "Detection of influential observation in linear regression". En: *Technometrics* 19.1, págs. 15-18.

Tema 10

Modelos lineales: Inferencia Estadística. Teoría de la distribución para variables normales y no normales. Tests de significación para modelos lineales normales y no normales. Intervalos de confianza e intervalos de predicción para modelos lineales normales y no normales. Comparaciones múltiples: Bonferroni, Tukey y métodos FDR.

Este tema está elaborado usando la siguiente bibliografía.

A. Agresti (2015). *Foundations of Linear and Generalized Linear Models*. Wiley

Jeffrey M Wooldridge (2006). *Introducción a la econometría. Un enfoque moderno: un enfoque moderno*. Editorial Paraninfo

En este tema se verán los principios básicos de inferencia estadística sobre los parámetros del *modelo lineal normal* (MLN), considerando que $\{y_i\}$ tienen distribuciones Normales. Asemás, se verán los principios básicos de inferencia estadística sobre los *modelos no normales* (MNN).

Primero, se revisa la teoría de distribución relevante para modelos lineales normales y no normales. Las formas cuadráticas incorporan variables de respuesta distribuidas Normalmente y matrices de proyección que generan distribuciones Chi-cuadrado. Uno de esos resultados, el Teorema de Cochran, es la base de las pruebas de significación sobre $\boldsymbol{\beta}$ en el MLN. La ecuación (10.2) muestra cómo los tests usan las formas cuadráticas de la Chi-cuadrado para construir estadísticos de test que tienen distribuciones F. Un resultado general útil sobre la comparación de dos modelos anidados también se deriva como un test de razón de verosimilitud. La ecuación (10.3) presenta intervalos de confianza para elementos de $\boldsymbol{\beta}$ y respuestas esperadas, así como intervalos de predicción para observaciones futuras. La ecuación (10.4) presenta métodos para realizar múltiples inferencias con una tasa de error global fija, así como múltiples métodos de comparación

para construir intervalos de confianza simultáneos para las diferencias entre todos los pares de un conjunto de medias. Sin el supuesto de Normalidad, caso de No Normalidad, los métodos de inferencia exactos se aplican al modelo lineal ordinario de una manera aproximada para grandes valores de n.

10.1 Teoría de la distribución para variables normales y no normales

10.1.1 Teoría de la distribución para variables normales

La inferencia estadística para MLNs utiliza distribuciones muestrales, derivadas de formas cuadráticas con variables aleatorias Normales multivariantes.

Por tanto, se revisa la distribución Normal multivariante y las distribuciones muestrales relacionadas.

Definición 12

Dada $N(\boldsymbol{\mu}, V)$ que denota la distribución Normal multivariante con media $\boldsymbol{\mu}$ y matriz de covarianza V. Si $y = (y_1, \cdots, y_n)^T$ tiene esta distribución y V es definida positiva, entonces, la función de densidad (de probabilidad) (pdf) es

$$f(\boldsymbol{y}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{V}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right]$$

donde $|\boldsymbol{V}|$ denota al determinante de $\boldsymbol{V}$.

A continuación se indican algunas propiedades cuando cuando $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$

1. Si $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}$, entonces $\boldsymbol{x} \sim N\left(\boldsymbol{A}\mu + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^{\mathrm{T}}\right)$
2. Supongamos que $\boldsymbol{y}$ se divide como

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix}, \text{ con } \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \text{ y } \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}$$

La distribución marginal de $\boldsymbol{y}_a$ es $N\left(\boldsymbol{\mu}_a, \boldsymbol{V}_{aa}\right), a = 1, 2$. La distribución condicional

$$(\boldsymbol{y}_1 \mid \boldsymbol{y}_2) \sim N\left[\boldsymbol{\mu}_1 + \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\left(\boldsymbol{y}_2 - \boldsymbol{\mu}_2\right), \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}\right]$$

Además, $\boldsymbol{y}_1$ y $\boldsymbol{y}_2$ son independientes si y sólo si $\boldsymbol{V}_{12} = \boldsymbol{0}$.

- De la propiedad anterior, si $\boldsymbol{V} = \sigma^2\boldsymbol{I}$, entonces $y_i \sim N\left(\mu_i, \sigma^2\right)$ y $\{y_i\}$ son independientes.

El MLN asume que $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ con $\boldsymbol{V} = \sigma^2\boldsymbol{I}$. El estimador de mínimos cuadrados $\widehat{\beta}$ y los residuos e también tienen distribuciones Normales multivariantes, ya que son funciones lineales de $\boldsymbol{y}$, pero sus elementos suelen estar correlacionados. Este estimador $\widehat{\boldsymbol{\beta}}$ es también el estimador de máxima verosimilitud (ML) bajo el supuesto de Normalidad.

Sea χ^2_p una distribución Chi-cuadrado con p grados de libertad (df) (*degrees of freedom*). Una variable aleatoria Chi-cuadrada es no negativa con media = df y varianza = 2(df). Su distribución[1] está sesgada hacia la derecha pero adquiere más forma de campana a medida que aumentan su grados de libertad (df).

Recordar que cuando $y_1, \cdots, y_p$ son variables aleatorias independientes Normales estándar, $\sum_{i=1}^{p} y_i^2 \sim \chi^2_p$. En particular, si $y \sim N(0,1)$, entonces $y^2 \sim \chi^2_1$. De manera más general,

- Si una variable aleatoria p-dimensional $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ con $\boldsymbol{V}$ no singular y de rango p, entonces

$$x = (\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \sim \chi^2_p$$

- Si $z \sim N(0,1)$ y $x \sim \chi^2_p$, con x y z independientes, entonces

$$\frac{z}{\sqrt{x/p}} \sim t_p$$

  la distribución t con $df = p$

La distribución t es simétrica alrededor del 0 con varianza $= df/(df-2)$ cuando $df > 2$. El término x/p en el denominador es una media de p variables aleatorias $N(0,1)$ independientes cuadráticas, por lo que $p \to \infty$ converge en probabilidad a su valor esperado de 1 . Por lo tanto, la distribución t converge a una distribución $N(0,1)$ a medida que su df aumenta.

Esta es una forma clásica en la que se produce la distribución t para respuestas independientes $y_1, \ldots, y_n$ de una distribución $N(\mu, \sigma^2)$ con media muestral $\bar{y}$ y varianza muestral s^2 : Para probar $H_0 : \mu = \mu_0$, el test estadístico $z = \sqrt{n}(\bar{y} - \mu_0)/\sigma$ tiene la distribución $N(0,1)$ nula. Además, s^2/σ^2 es una χ^2_{n-1} vaciación de $x = (n-1)s^2/\sigma^2$ dividido por su df. Dado que $\bar{y}$ y s^2 son independientes para observaciones independientes de una distribución normal, por debajo de H_0

$$t = \frac{z}{\sqrt{x/(n-1)}} = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

Valores mayores de $|t|$ proporcionan evidencia más fuerte contra H_0.

- Si $x \sim \chi^2_p$ y $y \sim \chi^2_q$, con x e y independientes, entonces

$$\frac{x/p}{y/q} \sim F_{p,q}$$

[1] Su función de densidad es un caso especial de la función de densidad de la distribución Gamma con parámetro $k = df/2$.

la distribución F con $df_1 = p$ y $df_2 = q$.

Una variable aleatoria F toma valores no negativos, tal que, cuando $df_2 > 2$, tiene media = $df_2/(df_2 - 2)$, aproximadamente 1 para valores grandes de df_2. Usaremos esta distribución para probar las hipótesis en ANOVA y regresión, tomando una razón de cuadrados medios independientes. Para una variable t aleatoria, t^2 tiene distribución F con $df_1 = 1$ y df_2 igual a los df de esta variable t.

En las pruebas de significación, para analizar el comportamiento de los estadísticos de contraste, cuando las hipótesis nulas son falsas, utilizamos distribuciones muestrales no centradas que se dan bajo los valores de los parámetros de la hipótesis alternativa. Dichas distribuciones determinan la potencia de un test (es decir, la probabilidad de rechazar H_0), que se puede analizar en función del valor real del parámetro. Cuando las observaciones tienen una distribución Normal multivariante, las distribuciones de muestreo en tales casos, no nulos, contienen las que se resumen como casos especiales.

Sea $\chi^2_{p,\lambda}$ que denota una distribución Chi-cuadrada no centralizada con $df = p$ y con parámetros de no centralidad λ. Esta es la distribución de $x = \sum_{i=1}^{p} y_i^2$ en la que $\{y_i\}$ son independientes con $y_i \sim N(\mu_i, 1)$ y $\lambda = \sum_{i=1}^{p} \mu_i^2$. Para esta distribución[2], $E(x) = p + \lambda$ y $\mathrm{var}(x) = 2(p + 2\lambda)$. La distribución Chi-cuadrada ordinaria (centralizada) es el caso especial con $\lambda = 0$.

- Si una variable aleatoria p-dimensional $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ con $\boldsymbol{V}$ no singular, de rango p, entonces

$$x = \boldsymbol{y}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{y} \sim \chi^2_{p,\lambda} \quad \text{with} \quad \lambda = \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{\mu}$$

  La construcción de la Chi-cuadrada no centralizada a partir de la suma de variables aleatorias $N(\mu_i, 1)$ cuadráticas independientes se obtiene cuando $\boldsymbol{V} = \boldsymbol{I}$.

- Si $z \sim N(\mu, 1)$ y $x \sim \chi^2_p$, con x y z independientes, entonces

$$t = \frac{z}{\sqrt{x/p}} \sim t_{p,\mu}$$

  la distribución t no centralizada con $df = p$ y no centralidad μ.

  La distribución t no centralizada es unimodal, pero sesgada en la dirección del signo de $\mu = E(z)$. Cuando $p > 1$ y $\mu \neq 0$, su media $E(t) \approx [1 - 3/(4p - 1)]^{-1}\mu$, la cual está cerca de μ pero un poco más grande en valor absoluto. Para p grande, la distribución de t es aproximadamente la distribución $N(\mu, 1)$.

[2] Aquí hay una forma alternativa de definir la no centralidad: Sea $z \sim \mathrm{Poisson}(\phi)$ y $(x \mid z) \sim \chi^2_{p+2z}$. Entonces incondicionalmente $x \sim \chi^2_{p,\phi}$. Esta no centralidad ϕ se relaciona con la no centralidad λ que definimos por $\phi = \lambda/2$

- If $x \sim \chi_p, \lambda^2$ y $\sim \chi^2_q$, con x e y independientes, entonces

$$\frac{x/p}{y/q} \sim F_{p,q,\lambda}$$

la función no centralizada F con $df_1 = p, df_2 = q$, y no centralidad λ.

Cochran 1934 mostró[3] el siguiente resultado, que también proporciona una interpretación de los grados de libertad:

Definición 13

Forma cuadrática normal con matriz de proyección y distribución Chi-cuadrado.

Supongamos $y \sim N(\boldsymbol{\mu}, \sigma^2 I)$, y P simétrica. Entonces

$\frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{P}(\boldsymbol{y} - \boldsymbol{\mu}) \sim \chi^2_r \Leftrightarrow \boldsymbol{P}$ matriz de proyeccións de rango r.

Por tanto, los grados de libertad representan la dimensión del subespacio vectorial al que se proyecta P. Sobre el resultado anterior Cochran 1934, formuló el siguiente Teorema:

Teorema 2

Teorema de Cochran[a]:

Supongamos n observaciones $y \sim N(\boldsymbol{\mu}, \sigma^2 I)$ and $P_1, \cdots, P_k$ son matrices de proyección con $\sum_i P_i = I$. Entonces, $\{y^T P_i y\}$ son independientes y $\left(\frac{1}{\sigma^2}\right) \boldsymbol{y}^{\mathrm{T}} \boldsymbol{P_i} \boldsymbol{y} \sim \chi^2_{r_i,\lambda_i}$ donde $\lambda_i = \frac{1}{\sigma^2} \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{P_i} \boldsymbol{\mu}, \quad i = 1, \ldots, k$, con $\sum_i r_i = n$.

[a] La demostración del Teorema de Cochran puede verse en Agresti 2015

Si reemplazamos y por $(y - \mu)$ en las formas cuadráticas, obtenemos distribuciones centradas Chi-cuadrado $(\lambda_i = 0)$. Este resultado es la base de las pruebas de significancia para parámetros en modelos lineales normales. La prueba del resultado de independencia muestra que todos los pares de matrices de proyección en esta descomposición satisfacen $\boldsymbol{P_i P_j} = 0$

La prueba de este Teorema está basada en uno de Monahan 2008, págs. 113-114. Primero observamos que si $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ and $\boldsymbol{P}$ es una matriz de proyección de rango r, entonces $\left(\frac{1}{\sigma^2}\right) \boldsymbol{y}^{\mathrm{T}} \boldsymbol{P} \boldsymbol{y} \sim \chi^2_{r,\lambda}$ con $\lambda = \frac{1}{\sigma^2} \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{P} \boldsymbol{\mu}$. Dado $\boldsymbol{P}$ simétrica e idempotente con

[3] Del resultado de Cochran I, ya que una matriz simétrica cuyos valores propios son 0 y 1 es idempotente.

rango r, sus autovalores son $1(r$ veces $)$ y $0(n-r$ veces $)$. Por la descomposión espectral de una matriz simétrica, podemos expresar $\boldsymbol{P}=\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathrm{T}}$, donde $\boldsymbol{\Lambda}$ es una matriz diagonal de $(1,1,\ldots,1,0,\ldots,0)$, los autovalores de $\boldsymbol{P}$, y $\boldsymbol{Q}$ es una matriz ortogonal con columnas que son autovectores de $\boldsymbol{P}$. Sea $z=\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{y}/\sigma$. Entonces, $z\sim N\left(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{\mu}/\sigma,\boldsymbol{I}\right)$, y $\left(\frac{1}{\sigma^2}\right)\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{y}=\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\Lambda}z=\sum_{i=1}^{r}z_i^2$. Dado que cada z_i es Normal con desviación estándar $1,\sum_{i=1}^{r}z_i^2$ tiene una distribución Chi-cuadrada no centralizada con $df=r$ y parámetro de no centralidad

$$\sum_{i=1}^{r}\left[E\left(z_i\right)\right]^2=\left[E(\boldsymbol{\Lambda}z)\right]^{\mathrm{T}}\left[E(\boldsymbol{\Lambda}z)\right]=\left(\frac{1}{\sigma^2}\right)\left[\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{\mu}\right]^{\mathrm{T}}\left[\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{\mu}\right]=$$

$$=\left(\frac{1}{\sigma^2}\right)\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{\mu}=\left(\frac{1}{\sigma^2}\right)\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{\mu}$$

Ahora consideramos k formas cuadráticas con k matrices de proyección que son una descomposición de $\boldsymbol{I}$, la $n\times n$ matriz identidad. El rango de una matriz de proyección es su traza, entonces $\sum_i r_i=\sum_i \operatorname{traza}\left(\boldsymbol{P}_i\right)=\operatorname{traza}\left(\sum_i \boldsymbol{P}_i\right)=\operatorname{traza}(\boldsymbol{I})=n$. Aplicamos la descomposición espectral a cada matriz de proyección, con $\boldsymbol{P}_i=\boldsymbol{Q}_i\boldsymbol{\Lambda}_i\boldsymbol{Q}_i^{\mathrm{T}}$, donde $\boldsymbol{\Lambda}_i$ es una matriz diagonal de $(1,1,\ldots,1,0,\ldots,0)$ con r_i entradas que son 1. Por la forma de $\boldsymbol{\Lambda}_i$, ésta es idéntica a $\boldsymbol{P}_i=\tilde{\boldsymbol{Q}}_i\boldsymbol{I}_{r_i}\tilde{\boldsymbol{Q}}_i^{\mathrm{T}}=\tilde{\boldsymbol{Q}}_i\tilde{\boldsymbol{Q}}_i^{\mathrm{T}}$, donde $\tilde{\boldsymbol{Q}}_i$ es una $n\times r_i$ matriz de las primeras r_i columnas de $\boldsymbol{Q}_i$. Tenga en cuenta que $\tilde{\boldsymbol{Q}}_i^{\mathrm{T}}\tilde{\boldsymbol{Q}}_i=\boldsymbol{I}_{r_i}$. Apilamos los $\left\{\tilde{\boldsymbol{Q}}_i\right\}$ juntos tal que

$$\boldsymbol{Q}=\left[\begin{array}{ccccc}\tilde{\boldsymbol{Q}}_1 & :\tilde{\boldsymbol{Q}}_2: & \cdots & :\tilde{\boldsymbol{Q}}_k\end{array}\right]$$

para las que

$$QQ^{\mathrm{T}}=\tilde{\boldsymbol{Q}}_1\tilde{\boldsymbol{Q}}_1^{\mathrm{T}}+\cdots+\tilde{\boldsymbol{Q}}_k\tilde{\boldsymbol{Q}}_k^{\mathrm{T}}=\boldsymbol{P}_1+\cdots+\boldsymbol{P}_k=\boldsymbol{I}_n$$

Por lo tanto, $\boldsymbol{Q}$ es una matriz ortogonal $n\times n$ y también $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}=\boldsymbol{I}_n$ y $\tilde{\boldsymbol{Q}}_i^{\mathrm{T}}\tilde{\boldsymbol{Q}}_j=\mathbf{0}$ para $i\neq j$. Luego $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{y}\sim N\left(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{\mu},\sigma^2\boldsymbol{I}\right)$, y sus componentes $\left\{\tilde{\boldsymbol{Q}}_i^{\mathrm{T}}\boldsymbol{y}\right\}$ son independientes, como son $\left\{\left\|\tilde{\boldsymbol{Q}}_i^{\mathrm{T}}\boldsymbol{y}\right\|^2=\boldsymbol{y}^{\mathrm{T}}\tilde{\boldsymbol{Q}}_i\tilde{\boldsymbol{Q}}_i^{\mathrm{T}}\boldsymbol{y}=\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{y}\right\}$. Notar[4] también que para $i\neq j,\boldsymbol{P}_i\boldsymbol{P}_j=\tilde{\boldsymbol{Q}}_i\tilde{\boldsymbol{Q}}_i^{\mathrm{T}}\tilde{\boldsymbol{Q}}_j\tilde{\boldsymbol{Q}}_j^{\mathrm{T}}=\mathbf{0}$

10.1.2 Teoría de la distribución para variables no normales

La **distribución Chi-cuadrada** se obtiene directamente a partir de variables normales estándar, independientes. Sean $Z_i,i=1,2,\ldots,n$, variables aleatorias independientes,

[4]El resultado de que $P_iP_j=0$ es también un caso especial del resultado más fuerte sobre la descomposición de matrices de proyección

cada una con una distribución Normal estándar. Se define una nueva variable como la suma de los cuadrados de las Z:

$$X = \sum_{i=1}^{n} Z_i^2 \tag{10.1}$$

Entonces, X tiene lo que se conoce como una distribución Chi-cuadrada (ó Ji-cuadrada) con n grados de libertad (gl) (ó degrees freedom (df)). Esto se expresa $X \sim \chi_n^2$. En una distribución Chi-cuadrada, los df corresponden a la cantidad de términos en la suma de la ecuación (10.1).

En la Figura 10.1 se presentan las fdp (funciones de densidad) de distribuciones Chi-cuadrada correspondientes a diversos grados de libertad. De acuerdo con la ecuación (10.1), está claro que las variables aleatorias Chi-cuadradas son siempre no negativas, y que, a diferencia de la distribución Normal, la distribución Chi-cuadrada no es simétrica respecto a ningún punto. Se puede demostrar que es si $X \sim \chi_n^2$, entonces el valor esperado de X es n [el número de términos en (10.1)] y la varianza de X es $2n$.

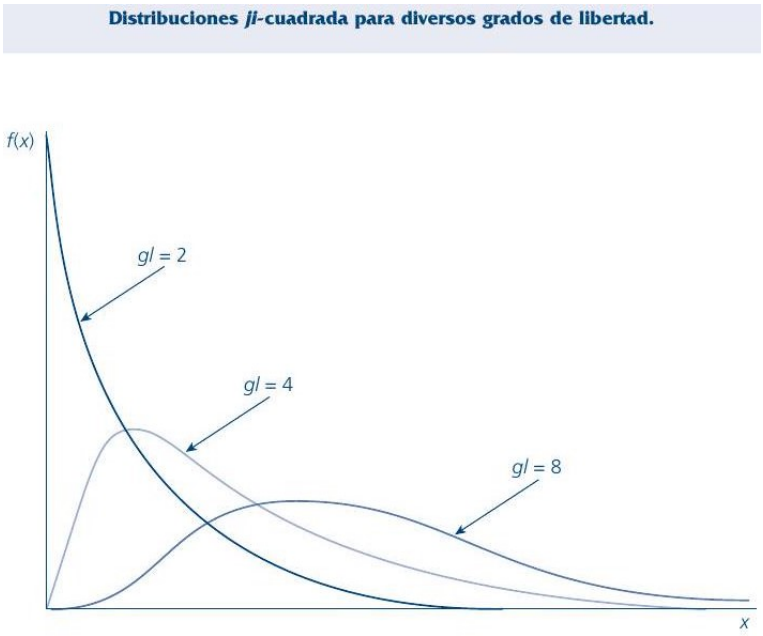

Figura 10.1: Distribuciones Chi-cuadrado para diferentes grados de libertad.

La **distribución** t es el caballo de batalla de la estadística clásica y del análisis de regresión múltiple. Una distribución t se obtiene a partir de una variable aleatoria Normal estándar y una variable aleatoria Chi-cuadrada.

Sea Z una variable aleatoria que tiene una distribución Normal estándar y sea X una variable aleatoria que tiene una distribución Chi-cuadrada con n grados de libertad.

Suponga, además, que Z y X son independientes. Entonces, la variable aleatoria

$$T = \frac{Z}{\sqrt{X/n}} \tag{10.2}$$

tiene una distribución t con n grados de libertad. Esto se denotará $T \sim t_n$. Las distribuciones t obtienen sus grados de libertad de la variable aleatoria Chi-cuadrada en el denominador de la ecuación (10.2).

La fdp de la distribución t tienen una forma similar a la de la distribución Normal estándar, sólo que es más dispersa, y por tanto tiene áreas mayores en las colas. El valor esperado de una variable aleatoria con distribución t es cero (estrictamente hablando, el valor esperado existe sólo para $n > 1$) y la varianza es $n/(n-2)$ para $n > 2$. (Para $n \leq 2$ no existe la varianza debido a que la distribución es demasiado dispersa.) En la Figura se grafican distribuciones t correspondientes a diversos grados de libertad. A medida que los grados de libertad aumentan, las distribuciones t se aproximan a la distribución Normal estándar.

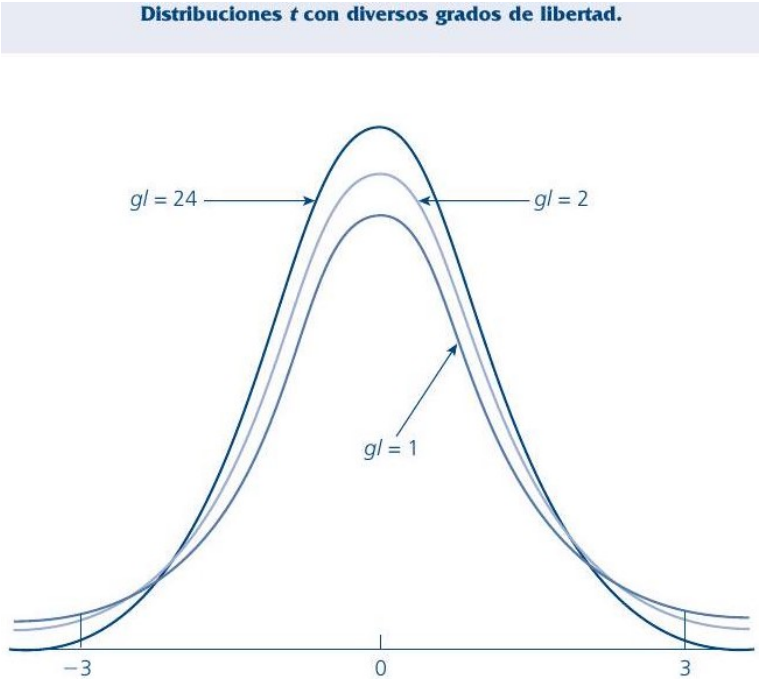

Figura 10.2: Distribuciones t para diferentes grados de libertad.

Otra distribución importante es la **distribución** F. Ésta se usará en especial para probar hipótesis en el contexto del análisis de regresión múltiple.

Para definir una variable aleatoria F, sean $X_1 \sim \chi^2_{k_1} y X_2 \sim \chi^2_{k_2} y$ suponemos que X_1 y X_2 son independientes. Entonces, la variable aleatoria

$$F = \frac{(X_1/k_1)}{(X_2/k_2)} \tag{10.3}$$

tiene una distribución F con (k_1, k_2) grados de libertad. Esto se denota como $F \sim F_{k_1,k_2}$. En la Figura 10.3 se muestran las fdp de distribuciones F con diversos grados de libertad.

En F_{k_1,k_2} el orden de los grados de libertad es crítico. El entero k_1 son los grados de libertad en el numerador, debido a que corresponde a la variable Chi-cuadrada del numerador. De igual manera, el entero k_2 son los grados de libertad en el denominador, debido a que corresponde a la variable Chi-cuadrada del denominador. Aquí hay que tener cuidado, pues la ecuación (10.3) también puede expresarse como $(X_1 k_2) \,/\, (X_2 k_1)$, de manera que k_1 aparece en el denominador. Sólo hay que recordar que los df del numerador es el entero asociado con la variable Chi-cuadrada en el numerador de la ecuación (10.3) y de manera similar para los df del denominador.

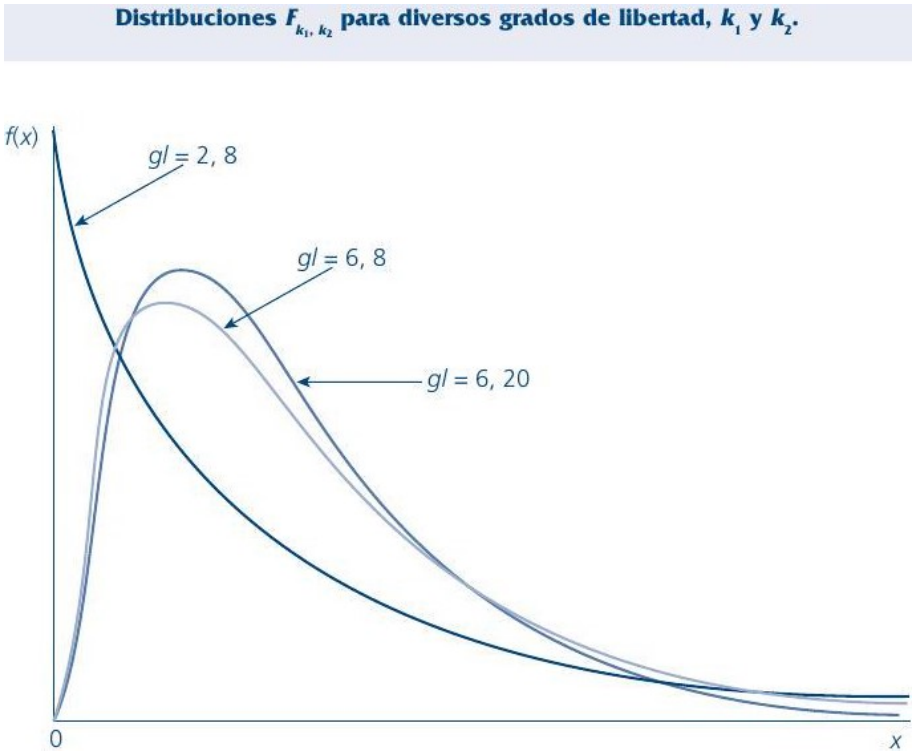

Figura 10.3: Distribuciones F para diferentes grados de libertad.

10.2 Tests de significación para modelos lineales normales y no normales

10.2.1 Tests de significación para modelos lineales normales

El teorema de Cochran nos da la posibilidad de derivar tests de significación fundamentales para el MLN. El primer test estadístico es el siguiente

$$F = \frac{\sum_{i=1}^{c} n_i \left(\bar{y}_i - \bar{y}\right)^2 / (c-1)}{\sum_{i=1}^{c} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2 / (n-c)} \sim F_{c-1,n-c,\lambda}.$$

Bajo la hipótesis $H_0, \lambda = 0$, y el estadístico F del test, tiene una distribución F con $df_1 = c - 1$ y $df_2 = n - c$. Valores más grandes de F son más contradictorios con H_0, luego el P-valor es la probabilidad de la cola de la derecha de esa distribución, que está por encima del valor observado del test estadístico, F_{obs}.

Esta prueba de significación para el diseño unilateral se conoce como análisis de varianza (unilateral), debido a R. A. Fisher (1925). La siguiente Tabla es la bien conocida Tabla ANOVA, cuya forma se muestra en la siguiente Tabla 10.1, y que también incluye el P-valor, $P_{H_0}(F > F_{obs})$.

Source	*df*	Sum of Squares	Mean Square	F_{obs}
Mean	1	$n\bar{y}^2$		
Group	$c-1$	$\sum_i n_i \left(\bar{y}_i - \bar{y}\right)^2$	$\frac{\sum_i n_i (\bar{y}_i - \bar{y})^2}{c-1}$	$\frac{\sum_i n_i (\bar{y}_i - \bar{y})^2/(c-1)}{\sum_i \sum_j (y_{ij} - \bar{y}_i)^2/(n-c)}$
Error	$n-c$	$\sum_i \sum_j \left(y_{ij} - \bar{y}_i\right)^2$	$\frac{\sum_i \sum_j (y_{ij} - \bar{y}_i)^2}{n-c}$	
Total	n	$\sum_{i=1}^{c} \sum_{j=1}^{n_i} y_{ij}^2$		

Tabla 10.1: Tabla ANOVA completa para el MLN para diseño unilateral

En la práctica, casi todas las hipótesis probadas sobre los efectos en modelos lineales se pueden expresar como $H_0 : \Lambda\boldsymbol{\beta} = 0$ para una matriz $\ell \times p$ de constantes $\boldsymbol{\Lambda}$ y un vector de cantidades estimables $\boldsymbol{\Lambda\beta}$. Un caso especial es el ejemplo considerado de $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$ para comparar un modelo completo con el modelo nulo. Otro ejemplo es una prueba para un contraste o un conjunto de contrastes, como $H_0 : \beta_j - \beta_k = 0$ para comparar las medias j y k en un diseño unidilateral. La forma $H_0 : \boldsymbol{\Lambda\beta} = 0$ es denominada *hipótesis lineal general*.

Una inferencia común para el modelo lineal es contrastar $H_0 : \beta_j = 0$ tal que se puede descartar una sola variable explicativa en el modelo. Este es el caso especial de $H_0 : \boldsymbol{\Lambda\beta} = 0$ que sustituye $\boldsymbol{\Lambda}$ por un vector fila $\boldsymbol{\lambda}$ con un múltiplo de 1 β_j y 0 en otros casos. Dado que el denominador del estadístico F del test para comparar dos modelos anidados es $s2$ (el error medio cuadrado) para el modelo completo, el test estadístido F entonces se simplifica a

$$F = \frac{(\mathrm{SSE}_0 - \mathrm{SSE}_1)/1}{\mathrm{SSE}_1/(n-p)} = \frac{(\lambda\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\left[\lambda\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\lambda^{\mathrm{T}}\right]^{-1}\lambda\widehat{\boldsymbol{\beta}}}{s^2} = \frac{\widehat{\beta}_j^2}{(SE_j)^2},$$

donde SE_j denota al error estándar de $\widehat{\beta}_j$, cuyo cuadrado es s^2 veces el elemento de la correspondiente fila y columna de $(XTX)^{-1}$. Este test estadístico tiene como grados de libertad, $df_1 = 1$ y $df_2 = n - p$.

10.2.2 Tests de significación / asintóticos para modelos lineales no normales

Si el tamaño muestral es lo bastante grande para invocar el teorema del límite central, la mecánica del test de hipótesis para las medias poblacionales es idéntica, sin importar si la distribución poblacional es Normal o no. La justificación teórica proviene del hecho que, bajo la hipótesis nula

$$T = \sqrt{n}(\overline{Y} - \mu_0)/S \to Normal(0, 1).$$

Por tanto, con n grande, se puede comparar el estadístico t con los valores críticos de una distribución Normal estándar. Debido a que la distribución t_{n-1} converge a la distribución Normal estándar a medida que n aumenta, t y los valores críticos Normales estándar estarán muy cerca para una n extremadamente grande.

Debido a que la teoría asintótica está basada en una n que crece de manera ilimitada, ésta no puede indicar si son mejores los valores críticos Normales estándares o los t. Para los valores moderados de n, por ejemplo, entre 30 y 60, es tradicional emplear la distribución t debido a que se sabe que es lo correcto para las poblaciones Normales. Para $n > 120$, la elección entre t y las distribuciones Normales estándar suele ser irrelevante debido a que los valores críticos prácticamente son los mismos.

Causado porque los valores críticos elegidos utilizan la distribución t o la Normal estándar son sólo aproximadamente válidos para poblaciones no Normales, los niveles de significancia elegidos son sólo aproximados; por tanto, para poblaciones no Normales, los niveles de significación son realmente niveles de significación asintóticos. De este modelo, si se elige un nivel de significación de $5\,\%$, pero la población es no Normal, entonces el nivel de significación real será mayor o menor que $5\,\%$ (y no se puede saber cuál es el caso). Cuando el tamaño muestral es grande, el nivel de significación real será muy cercano a $5\,\%$. En términos prácticos, la distinción no es importante, por tanto, se elimina el calificativo 'asintótico'.

Ejemplo 10. **Discriminación racial en las contrataciones.** En el estudio del *Urban Institute* sobre discriminación en la contratación, el interés principal estribaba en probar

$H_0 : \mu = 0$ frente a $H_1 : \mu < 0$, donde $\mu = \theta_B - \theta_W$ es la diferencia de probabilidades de que las personas negras recibieran ofertas de trabajo con relación a los blancos. Recordemos que μ es la media poblacional de la variable $Y = B - W$, donde B y W son los indicadores binarios.

Mediante las $n = 241$ comparaciones pareadas, se obtuvo $\overline{y} = -0,133$ y $s\sqrt{n} = 0,482/241 \approx 0,031$. El estadístico t para probar $H_0 : \mu = 0$ es $t = -0,133/0,031 \approx 4,29$. Conviene saber que la distribución Normal estándar es, para fines prácticos, indistinguible de la distribución t a partir de cierto número de grados de libertad. Vamos a considerar que con 240 grados de libertad se alcanza dicho límite. El valor de $4,29$ está tan lejano en el extremo izquierdo de la distribución que se rechaza H_0 a cualquier nivel de significación razonable. De hecho, el valor crítico de $0,005$ (medio punto porcentual para una prueba de una cola) es de aproximadamente $2,58$. Un valor t de $4,29$ es una evidencia muy sólida contra H_0 y en favor de H_1. Por tanto, se concluye que existe discriminación en la contratación.

10.3 Intervalos de confianza e intervalos de predicción para modelos lineales normales y no normales

Aprendemos más de la construcción de intervalos de confianza para los valores de los parámetros, que de los tests de significación. Un intervalo de confianza nos muestra el rango completo de valores plausibles para un parámetro, en lugar de centrarse simplemente en si un valor particular es plausible.

10.3.1 Intervalos de confianza e intervalos de predicción para modelos lineales normales

Para construir un intervalo de confianza para un parámetro β_j en un MLN, construimos y luego invertimos un test t de $H_0 : \beta_j = \beta_{j0}$ sobre los valores potenciales para for β_j. El test estadístico es,

$$t = \frac{\widehat{\beta}_j - \beta_{j0}}{SE_j}$$

Dado que s^2 es una función de los residuos, $\widehat{\boldsymbol{\beta}}$ y s^2 son independientes, y también lo son el numerador y el denominador del estadístico t, como se requiere para obtener una distribución t.

El intervalo de confianza $100(1-\alpha)\,\%$ para β_j es el conjunto de todos los valores de β_{j0} para los que el test tiene un P-valor $> \alpha$, es decir, para los que $|t| < t_{\alpha/2,n-p}$, es el cuantil $1 - \alpha/2$ de la distribución t , con $df = n - p$. Por ejemplo, el intervalo de confianza al 95 % es

$$\widehat{\beta}_j \pm t_{0,025,n-p}\,(SE_j)$$

Podemos construir un intervalo de confianza para $E(y) = x_0\boldsymbol{\beta}$. Hacemos esto construyendo y luego invirtiendo un test t sobre valores para ese predictor lineal.
Sea $\widehat{\mu} = \boldsymbol{x}_0\widehat{\boldsymbol{\beta}}$. Entonces

$$\operatorname{var}(\widehat{\mu}) = \operatorname{var}\left(\boldsymbol{x}_0\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{x}_0 \operatorname{var}(\widehat{\boldsymbol{\beta}})\boldsymbol{x}_0^{\mathrm{T}} = \sigma^2\boldsymbol{x}_0\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0^{\mathrm{T}}$$

Dado que $\boldsymbol{x}_0\widehat{\boldsymbol{\beta}}$ es una función lineal de y, entonces tiene una distribución Normal. Por lo tanto,

$$z = \frac{\boldsymbol{x}_0\widehat{\boldsymbol{\beta}} - \boldsymbol{x}_0\boldsymbol{\beta}}{\sigma\sqrt{\boldsymbol{x}_0\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0^{\mathrm{T}}}} \sim N(0,1)$$

y

$$t = \frac{\boldsymbol{x}_0\widehat{\boldsymbol{\beta}} - \boldsymbol{x}_0\boldsymbol{\beta}}{s\sqrt{\boldsymbol{x}_0\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0^{\mathrm{T}}}} = \frac{\boldsymbol{x}_0\widehat{\boldsymbol{\beta}} - \boldsymbol{x}_0\boldsymbol{\beta}}{\sigma\sqrt{\boldsymbol{x}_0\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0^{\mathrm{T}}}} \Big/ \sqrt{\frac{s^2}{\sigma^2}} \sim t_{n-p}$$

De ello se deduce que un intervalo de confianza de $100(1-\alpha)\,\%$ para $E(y) = x_0\beta$ es

$$\boldsymbol{x}_0\widehat{\boldsymbol{\beta}} \pm t_{\alpha/2,n-p}s\sqrt{\boldsymbol{x}_0\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0^{\mathrm{T}}} \tag{10.4}$$

Cuando x_0 es el valor de la variable explicativa x_i para una observavión particular, el término debajo de la raíz cuadrada es el apalancamiento h_{ii} de la matriz Hat del modelo.

La construcción de este intervalo se extiende directamente a los intervalos de confianza para combinaciones lineales $\ell\beta$. Un ejemplo es un contraste de los parámetros, como $\beta_j - \beta_k$ para un par de niveles de un factor.

Para un valor particular x_0, ¿cómo podemos formar un intervalo que es muy probable que contenga una observación futura y en ese valor? Esto es más desafiante que construir un intervalo de confianza para la respuesta esperada. Con una gran cantidad de datos, podemos hacer inferencias precisas sobre la predicción media, pero no sobre una única observación futura, de manera precisa.

El MLN establece que un valor futuro y satisface.

$$y = \boldsymbol{x}_0\boldsymbol{\beta} + \epsilon, \quad \text{donde} \quad \epsilon \sim N\left(0, \sigma^2\right)$$

Del ajuste del modelo, la predicción a futuro del valor y es $\widehat{\mu} = \boldsymbol{x}_0\widehat{\boldsymbol{\beta}}$. Entonces, el valor futuro de y tambien satisface

$$y = \boldsymbol{x}_0\widehat{\boldsymbol{\beta}} + e, \quad \text{donde} \quad e = y - \widehat{\mu}$$

que es el residuo (error) para esta observación. Dado que el futuro y es independiente de las observaciones $y_1, \ldots, y_n$ utilizadas para determinar $\widehat{\boldsymbol{\beta}}$ y entonces $\widehat{\mu}$

$$\operatorname{var}(e) = \operatorname{var}(y - \widehat{\mu}) = \operatorname{var}(y) + \operatorname{var}(\widehat{\mu}) = \sigma^2 \left[1 + \boldsymbol{x}_0 \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{x}_0^{\mathrm{T}}\right]$$

Resulta que

$$\frac{y - \widehat{\mu}}{\sigma\sqrt{1 + x_0 \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{x}_0^{\mathrm{T}}}} \sim N(0,1) \quad \text{y} \quad \frac{y - \widehat{\mu}}{s\sqrt{1 + \boldsymbol{x}_0 \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{x}_0^{\mathrm{T}}}} \sim t_{n-p}$$

Al invertir esto se obtiene un intervalo de predicción $100(1-\alpha)\,\%$ para la observación y futura,

$$\widehat{\mu} \pm t_{\alpha/2,n-p} s\sqrt{1 + \boldsymbol{x}_0 \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{x}_0^{\mathrm{T}}} \tag{10.5}$$

Ejemplo 11. Ilustramos el intervalo de confianza para la media y el intervalo de predicción para una observación futura con el modelo lineal bivariado,

$$E(y_i) = \beta_0 + \beta_1 x_i$$

Para una observación futura y su predicción independiente $\widehat{\mu}$

$$\operatorname{var}(y - \widehat{\mu}) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]$$

A medida que n aumenta, $\operatorname{var}(\widehat{\mu})$ decrece hacia 0, pero $\operatorname{var}(y - \widehat{\mu})$ tiene σ^2 de límite inferior. Incluso si podemos estimar casi perfectamente la línea de regresión, estamos limitados en la precisión con la que podemos predecir cualquier observación futura.

La Figura 10.4 traza el intervalo de confianza y el intervalo de predicción, en función de x_0. A medida que n aumenta, la amplitud de un intervalo de confianza para la media en cualquier x_0 decrece hacia 0, pero la amplitud del intervalo de predicción al $95\,\%$ decrece hacia $2(1{,}96)\sigma$.

Interpretar un intervalo de predicción es incómodo. Con $\alpha = 0,05$, nos gustaría decir que, condicionado a los datos observados y al ajuste del modelo, tenemos un 95 % de confianza que el valor futuro de y estará en el intervalo; es decir, cerca del 95 % de un gran número de observaciones futuras caerían en el intervalo.

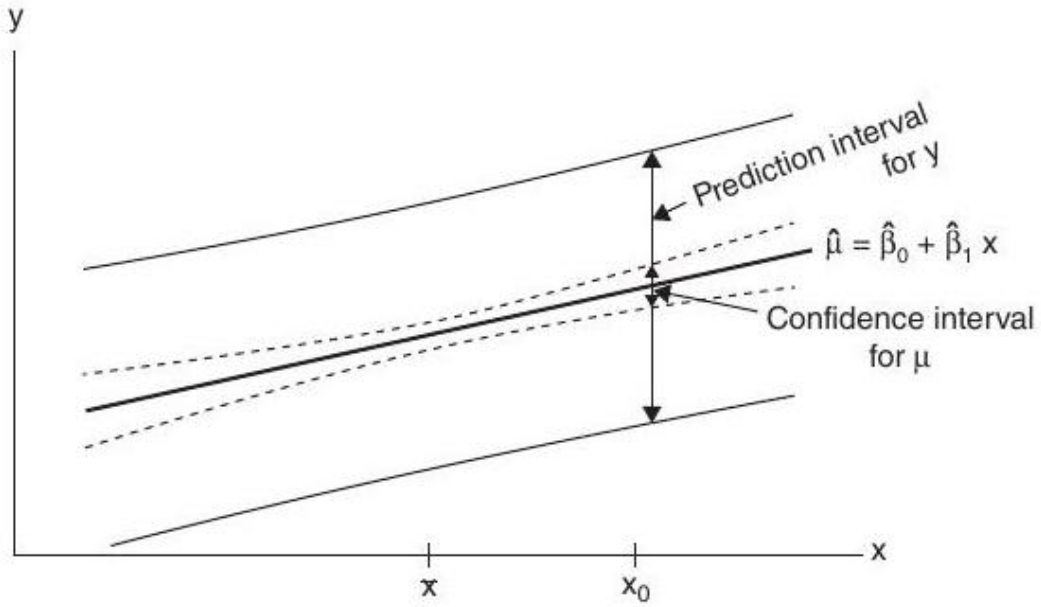

Figura 10.4: Intervalo de confianza para la media $E(y) = \beta_0 + \beta_1 x_0$ e intervalo de predicciones para una futura observación y para varios valores x_0

Sin embargo, las distribuciones de probabilidad en la derivación de la (10.3), relativa al *Intervalo de predicción para un valor futuro* y, tratan a $\widehat{\mu}$ así como al futuro valor y como aleatorios, mientras que en la práctica usamos el intervalo tras observar los datos y, por lo tanto $\widehat{\mu}$. La probabilidad condicional de que el intervalo de predicción incluya a un futuro valor y, dado $\widehat{\mu}$, no es 0,95. A partir del razonamiento que llevó a la ecuación (10.5), antes de recopilar cualquier dato, para encontrar los $\widehat{\mu}$ (y s) y luego, al futuro y,

$$P\left[|y - \widehat{\mu}|/s\sqrt{1 + x_0 \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{x}_0^{\mathrm{T}}} \leq t_{0,025,n-p}\right] = 0,95$$

Una vez que observamos los datos y encontramos $\widehat{\mu}$ y s, esta probabilidad (con y como la única parte aleatoria) no es igual a $0,95$, depende sobre dónde cayó $\widehat{\mu}$. No es necesario que esté cerca de $0,95$ a menos que $\mathrm{var}(\widehat{\mu})$ sea insignificante en comparación con (y). El $95\,\%$ de confianza para un intervalo de confianza significa lo siguiente: si al usar repetidamente este método con muchos conjuntos de datos de observaciones independientes, que satisfacen el modelo (es decir, para construir tanto la ecuación ajustada, como este intervalo), cada vez que hagamos una observación futura, a largo plazo $95\,\%$ del tiempo, el intervalo construido contendá la observación futura.

En la práctica, debemos tener una fe considerable en el modelo antes de construir intervalos de predicción. Incluso si no creemos realmente en el modelo (en la práctica, la situación habitual), un intervalo de confianza para $E(y) = x_0\beta$ en varios valores de x_0 es útil para describir el ajuste del modelo en la población de interés. Pero, si el modelo falla, ya sea en su descripción de la media poblacional en función de las variables explicativas,

o en sus supuestos de Normalidad con varianza constante, entonces el porcentaje real de muchas observaciones futuras que caen dentro de los límites del $95\,\%$ de los intervalos de predicción, pueden ser bastante diferentes de $95\,\%$.

10.3.2 Intervalos de confianza asintóticos e intervalos de predicción para modelos lineales no normales

En algunas aplicaciones, la población es claramente no Normal. Un caso sobresaliente es la distribución de Bernoulli, donde la variable aleatoria asume sólo los valores cero y uno. En otros casos, la población no Normal no tiene distribución estándar. Esto no importa, siempre y cuando el tamaño de la muestra sea lo bastante grande para el teorema central del límite para dar una buena aproximación de la distribución de la media muestral $\overline{Y}$. Para una n, grande, un intervalo de confianza aproximado al $95\,\%$ es

$$\overline{y} \pm 1,96 s/\sqrt{n} \tag{10.6}$$

donde el valor $1,96$ es el $97,5$-ésimo percentil en la distribución normal estándar. En términos mecánicos, calcular un intervalo de confianza aproximado no difiere del caso Normal. Una pequeña diferencia consiste en que el número que multiplica el error estándar proviene de la distribución Normal estándar, y no de la distribución t debido a que se están usando asintóticas. Debido a que la distribución t se aproxima a la Normal estándar a medida que los df aumentan, la ecuación

$$\overline{y} \pm c_{\alpha/2} s/\sqrt{n}$$

también se legitima perfectamente como una aproximación al intervalo de $95\,\%$; algunos prefieren ésta a (10.6), pues la primera es exacta para poblaciones Normales.

Ejemplo 12. **Discriminación racial en las contrataciones** El *Urban Institute* realizó un estudio en 1988 en Washington, D.C. para examinar el grado de discriminación racial en las contrataciones. Se entrevistó a cinco pares de personas para varios puestos. En cada par, una persona era blanca y la otra negra, las cuales ofrecieron currículos que indicaba que eran virtualmente iguales en términos de experiencia, educación y otros factores que determinaba la calificación para el trabajo. La idea era hacer que los individuos fueran tan similares como fuera posible con excepción de la raza. Las personas de cada par pasaban por una entrevista para solicitar el mismo trabajo y los investigadores anotaron qué solicitante recibió cada oferta de trabajo.

Este es un ejemplo de un análisis de pares igualados, donde cada prueba consiste en datos de dos personas (o dos empresas, ciudades, etcétera) que se consideran similares en muchos aspectos pero diferentes en una característica importante. Sea θ_B la probabilidad de que la persona negra obtenga el trabajo y θ_W la probabilidad de que la persona blanca lo obtenga. Lo que aquí interesa es la diferencia, $\theta_B - \theta_W$.

Sea B_j una variable de Bernoulli igual a uno si la persona negra obtiene una oferta de trabajo de su empleador i, y cero de otra manera. Asimismo, $W_i = 1$ si la persona blanca obtiene una oferta de trabajo del empleador i, y cero en caso contrario. Al reunir los resultados de los cinco pares de personas, hubo un total de $n = 241$ pruebas (pares de entrevistas con empleadores). Los estimadores insesgados de θ_B y θ_W son $\overline{B}$ y $\overline{W}$, las fracciones de entrevistas en las cuales personas blancas y negras recibieron ofertas de trabajo, respectivamente.

Para calcular un intervalo de confianza para la media de una población, definimos una nueva variable $Y_i = B_i - W_i$. Ahora, Y_i puede tomar tres valores: -1 si la persona negra no obtiene el trabajo, pero la persona blanca sí, 0 si ninguna de las personas obtiene el trabajo o si las dos lo obtienen, y 1 si la persona negra obtiene el trabajo y la persona blanca no. Entonces, $\mu \equiv E(Y_i) = E(B_i) - E(W_i) = \theta_B$ - θ_W.

La distribución de Y_i sin lugar a dudas es no Normal; es discreta y sólo toma tres valores. Sin embargo, un intervalo de confianza aproximado para $\theta_B - \theta_W$ se puede obtener mediante métodos muestrales grandes.

Mediante los 241 puntos de datos observados $\overline{b} = 0,224$ y $\overline{w} = 0,357$, así que $\overline{y} = 0,224 - 0,357 = -0,133$ Por tanto, $22,4\,\%$ de los solicitantes negros recibieron oferta de trabajo, mientras que $35,7\,\%$ de los solicitantes blancos recibieron oferta de trabajo.

Esto es a primera vista una evidencia de la discriminación en contra de las personas negras, pero es posible saber mucho más al calcular un intervalo de confianza para μ. Para calcular un intervalo de confianza al $95\,\%$, se necesita la desviación estándar muestral. En este caso, ésta resulta ser $s = 0,482$ y, mediante (10.6), se obtiene un IC a $95\,\%$ para $\mu = \theta_B - \theta_W$ cuando $-0,133 \pm 1,96(0,482/\sqrt{241}) = -0,133 \pm 0,031 = [-0,164, -0,102]$. El IC aproximado al $99\,\%$ es $-0,133 \pm 2,58(,482/\sqrt{241}) = [-0,213, -0,053]$ Naturalmente, este contiene a un rango más amplio de valores que el IC al $95\,\%$. Pero incluso el IC al $99\,\%$ no contiene al valor cero. Por tanto, se tiene un alto grado de confianza en que la diferencia poblacional $\theta_B - \theta_W$ no sea cero.

10.4 Comparaciones múltiples: Bonferroni, Tukey y métodos FDR.

El uso de un modelo para comparar muchos grupos, o para evaluar la importancia de muchas variables explicativas potenciales en un modelo, puede implicar un gran número de estimaciones. Por ejemplo, en un diseño unilateral, comparar cada par de grupos c implica c(c - 1)/2 estimaciones, lo cual es considerable cuando c en sí mismo es un valor grande. Incluso si cada estimación tiene una pequeña probabilidad de error, la probabilidad de que al menos una estimación sea errónea puede ser sustancial. En esos casos, podemos construir las estimaciones de modo que la probabilidad de error se aplique a toda la familia de estimaciones, en lugar de a cada una de ellas. Por ejemplo, al construir intervalos de confianza para comparaciones de medias por pares, podemos proporcionar una *confianza familiar* del 95 % de que todo el conjunto de intervalos con-

tiene simultáneamente las diferencias reales.

Una forma típica de realizar múltiples estimaciones, mientras se controla la tasa de error global, se basa en una simple desigualdad mostrada por el matemático británico Boole 1854, en un impresionante tratado del cual varios capítulos presentaban leyes de probabilidad.

Definición 14

Desigualdad de Boole: Sean $E_1, E_2, \cdots, E_t$, t eventos en un espacio muestral. Entonces, la probabilidad de que ocurra al menos uno de estos eventos tiene el límite superior

$$P\left(\cup_j E_j\right) \leq \sum_{j=1}^{t} P\left(E_j\right)$$

La prueba de esto es simple. Se puede construir un diagrama de Venn para ilustrarla. Sea

$$B_1 = E_1, B_2 = E_1^c \cap E_2, B_3 = E_1^c \cap E_2^c \cap E_3, \ldots$$

Entonces, $\cup_j B_j = U_j E_j$ y $B_j \subset E_j$, pero los $\{B_j\}$ son disjuntos, y por tanto $P\left(\cup_j B_j\right) = \sum_j P\left(B_j\right)$. Luego,

$$P\left(\cup_j E_j\right) = P\left(\cup_j B_j\right) = \sum_{j=1}^{t} P\left(B_j\right) \leq \sum_{j=1}^{t} P\left(E_j\right)$$

En el contexto de intervalos de confianza múltiples, sea E_j (para $j = 1, \ldots, t$) que denota al evento de que el intervalo j sea erróneo, no contiene el valor relevante del parámetro. Si cada intervalo tiene un coeficiente de confianza $(1 - \alpha/t)$, entonces la probabilidad (a priori) de que al menos uno de los intervalos t sea erróneo está acotada por arriba por $t(\alpha/t) = \alpha$. Entonces, el coeficiente de confianza familiar para el conjunto de los intervalos t está acotado por abajo por $1 - \alpha$.

Por ejemplo, para el diseño unilateral con $c = 5$ significa que, si usamos el nivel de confianza $99\,\%$ para cada una de las 10 comparaciones por pares, el nivel de confianza general es al menos $90\,\%$. Este método para construir intervalos de confianza simultáneos se denomina *método de Bonferroni*. Se basa simplemente en la desigualdad de Boole, pero el nombre se refiere al probabilista / matemático italiano Carlo Bonferroni, quien en 1936 amplió la desigualdad de Boole de varias formas.

Una ventaja del método Bonferroni es su generalidad. Se aplica a cualquier inferencia basada en probabilidades para cualquier distribución, no solo a los intervalos de confianza para un modelo lineal normal. Una desventaja es que el método es $conservador$: si queremos una confianza general del $90\,\%$ (digamos), el método asegura que el nivel de confianza real sea al menos así de alto. Como consecuencia, los intervalos son más amplios que los que producirían exactamente ese nivel de confianza.

El siguiente método que discutimos es más limitado, y está diseñado específicamente para comparar medias en modelos lineales normales balanceados, pero no tiene esta desventaja.

En 1953, el gran estadístico John Tukey propuso un método para comparar simultáneamente las medias de varias distribuciones Normales, al utilizar una distribución de probabilidad para el rango de observaciones de una distribución Normal, que aplica a diseños balanceados, como son los diseños unilaterales y bilaterales con tamaños muestrales iguales.

Definición 15

Supongamos que $\{y_i\}$ son independientes, con $y_i \sim N(\mu, \sigma^2), i = 1, \ldots, c$. Let s^2 siendo una estimación independiente de σ^2 con $vs^2/\sigma^2 \sim \chi^2_v$. Entonces,

$$Q = \frac{\text{máx}_i\, y_i - \text{mín}_i\, y_i}{s}$$

tiene la *distribución de rango estudentizada* con parámetros c y ν. Denotamos a la distribución por $Q_{c,v}$ y sus cuantiles $1-\alpha$ por $Q_{1-\alpha,c,v}$.

Para ilustrar cómo el método de Tukey usa la distribución de rango estudentizado, consideramos el diseño unilateral balanceado para el MLN. Las medias muestrales $\bar{y}_1, \ldots, \bar{y}_c$ tienen tamaño muestral $n_i = n$. Sea $N = \sum_i n_i = cn$. Sea $s^2 = \sum_{i=1}^{c} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 / (N - c)$ que denota la estimación de la varianza combinada del ANOVA unilateral (es decir, el error cuadrado medio). Entonces cada $\sqrt{n}(\bar{y}_i - \mu_i)$ tiene una distribución $N(0, \sigma^2)$, y por tanto

$$\sqrt{n}\left[\max_i (\bar{y}_i - \mu_i) - \min_i (\bar{y}_i - \mu_i)\right] / s \sim Q_{c,N-c}$$

A priori, la probabilidad es $(1-\alpha)$ por lo que su estadístico es menor que $Q_{1-\alpha,c,N-c}$. Cuando el estadístico está acotado por arriba por $Q_{1-\alpha,c,N-c}$, entonces

$$\text{Todo } |(\bar{y}_i - \mu_i) - (\bar{y}_j - \mu_j)| < Q_{1-\alpha,c,N-c}(s/\sqrt{n})$$

y entonces $(\mu_i - \mu_j)$ cae dentro $Q_{1-\alpha,c,N-c}(s/\sqrt{n})$ con $(\bar{y}_i - \bar{y}_j)$ para todos los pares. Por

tanto, podemos construir intervalos de confianza familiares para los pares $\{\mu_i - \mu_j\}$ usando simultáneamente para todo i y j,

$$(\bar{y}_i - \bar{y}_j) \pm Q_{1-\alpha,c,N-c}\left(\frac{s}{\sqrt{n}}\right)$$

El coeficiente de confianza para la familia de todo $t = c(c-1)/2$, tal que sus comparaciones son igual a $1 - \alpha$. Una diferencia $|\bar{y}_i - \bar{y}_j|$ que excede $Q_{1-\alpha,c,N-c}(s/\sqrt{n})$ se considera estadísticamente significativa, como el intervalo para $(\mu_i - \mu_j)$ no contiene $0.$, el correspondiente margen de error usando el método de Bonferroni es $t_{\alpha/c(c-1),N-c}s\sqrt{2/n}$

Para entenderlo, supongamos que planeamos construir intervalos de confianza al $95\,\%$ para 45 pares de medias para $c = 10$ grupos, y tenemos $n = 20$ observaciones para cada grupo, y una desviación típica de $s = 15.$ El margen de error para cada comparación es $Q_{0,95,10,190}(15/\sqrt{20}) = 15,19$ por el método de Tukey, y $t_{0,05/2(45),190}(15\sqrt{2/20}) = 15,71$ por el método de Bonferroni. Los cuantiles Q y t usados aquí se pueden obtener fácilmente con algún software estadístico:

```
>qtukey(0,95, 10, 190); qt(1 - 0,05/(2*45), 190)
[1]  4,27912
[1]  3,311379
```

El método de Tukey se aplica exactamente a este caso balanceado, para el cual las medias muestrales tienen varianzas iguales. Una versión generalizada se aplica de manera ligeramente conservadora para casos no balanceados.

Bibliografía

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley.

Boole, George (1854). *An investigation of the laws of thought: on which are founded the mathematical theories of logic and probabilities*. Vol. 2. Walton y Maberly.

Cochran, William G (1934). "The distribution of quadratic forms in a normal system, with applications to the analysis of covariance". En: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 30. 2. Cambridge University Press, págs. 178-191.

Monahan, John F (2008). *A primer on linear models*. CRC Press.

Wooldridge, Jeffrey M (2006). *Introducción a la econometría. Un enfoque moderno: un enfoque moderno*. Editorial Paraninfo.