Capítulo 3

TEORÍA GENERAL DEL MODELO LINEAL

[1]

[1] Este capítulo forma parte del apunte para el curso MA44D dictado por la profesora Nancy Lacourly con profesores auxiliares Félix Carrasco y Claudio Pareja.

3.1. Introducción.

Un problema muy frecuente en estadística consiste en buscar y estimar interdependencias entre variables. En efecto, cuando un par de variables aleatorias (X,Y) no son independientes, el conocimiento del valor por X cambia nuestro incertidumbre con respecto a la realización de Y, esto se tiene debido a que la distribución de Y dado $X = x$ tiene una varianza que en promedio es menor que la varianza marginal de Y, en efecto:

Proposición 3.1 *Sean X e Y dos variables aleatorias, luego*

$$\mathbb{V}(Y) = \mathbb{E}_X\{\mathbb{V}(Y|X)\} + \mathbb{V}_X\{\mathbb{E}(Y|X)\}$$

Dem: Por definición de varianza tenemos que:

$$\begin{aligned}
\mathbb{V}(Y) &= \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 && (3.1)\\
&= \mathbb{E}_X\{\mathbb{E}(Y^2|X)\} - \mathbb{E}_X\{\mathbb{E}(Y|X)\}^2 && (3.2)\\
&= \mathbb{E}_X\{\mathbb{V}(Y|X) + \mathbb{E}(Y|X)^2\} - \mathbb{E}_X\{\mathbb{E}(Y|X)\}^2 && (3.3)\\
&= \mathbb{E}_X\{\mathbb{V}(Y|X)\} + (\mathbb{E}_X\{\mathbb{E}(Y|X)^2\} - \mathbb{E}_X\{\mathbb{E}(Y|X)\}^2) && (3.4)\\
&= \mathbb{E}_X\{\mathbb{V}(Y|X)\} + \mathbb{V}_X\{\mathbb{E}(Y|X)\} && (3.5)
\end{aligned}$$

donde:

- 3.1 es por definición de varianza.
- 3.2 se tiene por definición de esperanza condicional.
- 3.3 se utiliza la definición de varianza condicional.
- 3.4 es por que la esperanza de una suma es la suma de las esperanzas y se agrupan términos.
- 3.5 se utiliza la definición de varianza.

luego se deduce que $\mathbb{V}(Y) = \mathbb{E}_X\{\mathbb{V}(Y|X)\} + \mathbb{V}_X\{\mathbb{E}(Y|X)\}$.

■

De esta igualdad tenemos que $\mathbb{E}_X\{\mathbb{V}(Y|X))\} \leq \mathbb{V}(Y)$, que es un resultado promedio, lo cual no impide que para algunos valores de X, $\mathbb{V}(Y|X)$ sea mayor que $\mathbb{V}(Y)$.

Cuando se puede aceptar que el fenómeno aleatorio representado por una variable o un vector X puede servir para predecir aquel representado por Y, se busca una fórmula de predicción. Algunas relaciones son fáciles de plantear y verificar, como las relaciones planteadas a partir de leyes físicas o mecánicas, pero cuando la aleatoriedad juega un papel importante, el estudio se hace más difícil.

Se busca descubrir como un conjunto de variables $X_1, X_2, \ldots X_p$ influye sobre una o varias otras variables Y. Para este propósito, se busca una función f que permita reconstruir los valores obtenidos sobre una muestra de la variables respuesta Y:

$$Y = f(X_1, X_2, \ldots X_p\}.$$

Las variables $\{X_1, X_2, \ldots X_p\}$ se llaman **variables explicativas** [2] y la variables Y se llama **variable respuesta** [3].

Se darán algunos ejemplos, en que se ocupan estos modelos:

Ejemplo 3.1 La distancia que una partícula recorre en el tiempo t está dada por la fórmula:

$$d = \alpha + \beta t$$

en que β es la velocidad promedio y α la posición de la partícula en el tiempo inicial $t = 0$. Si α y β son desconocidos, observando la distancia d en dos épocas distintas, la solución del sistema de las dos ecuaciones lineales obtenidas permite obtener α y β. Sin embargo es difícil obtener en general una distancia sin error de medición. Por lo cual se observa una variable aleatoria: $Y = d + \varepsilon$ en vez de d, en que ε es de tipo aleatorio[4]. En este caso no basta tener dos ecuaciones, sino que diferentes valores de la distancia para distintos tiempos. Los métodos estadísticos basados en la aleatoriedad del error permiten estimar a α, β y d sobre la base de una relación funcional de tipo lineal.

Ejemplo 3.2 Si consideramos el peso P y la talla T de las mujeres chilenas adultas, está claro que no existe una relación funcional entre P y T, pero existe una **tendencia**. Considerando que P y T son variables aleatorias de distribución conjunta normal bivariada:

$$\begin{aligned} P &= f(T) + \varepsilon \\ con \quad f(T) &= E(P|T) \end{aligned}$$

en que ε refleja la variabilidad del peso P entre las chilenas de la misma talla con respecto a la media. El tipo de funcional f en este caso no es evidente, pero plantear en principio un modelo lineal puede servir al menos para descartarlo.

Ejemplo 3.3 Para decidir la construcción de la nueva central eléctrica, ENDESA busca prever el consumo total de electricidad en Chile después del año 2002. Se construye un modelo que liga el consumo de electricidad con variables económicas, demográficas y metereológicas, y este modelo estima en base a datos obtenidos en el pasado. Se aplica entonces el modelo para predecir el consumo de electricidad según ciertas evoluciones económicas, metereológicas y demográficas. En este caso resulta aún más complejo determinar cual es la relación funcional que existe entre las variables explicativas y la variable respuesta.

Ejemplo 3.4 Para establecer una determinada publicidad en la televisión, se cuantifica el efecto de variables culturales y socio-económicas en la audiencia de los diferentes programas. Sobre la base de una encuesta telespectadores se construye un modelo que determina los efectos de las variables culturales y socio-económicas en la audiencia.

Ejemplo 3.5 Ajuste polinomial. El modelo lineal puede ser generalizado tomando funciones de las variables explicativas y/o de la variable a explicar. Es el caso cuando se tiene una variables respuesta Y a partir de una sola variable X en un modelo polinomial: $Y = a_0 + a_1 X^1 + a_2 X^2 \cdots + a_p X^p$ en donde X^j corresponde a la potencia j de X. Note que es fácil argumentar que este ejemplo corresponde a un modelo lineal, definiendo como variable $X_j = X^j$.

Ejemplo 3.6 Se quiere estimar la constante g de la gravitación. Se toman los tiempos de caída t de un objeto desde la altura h dada del suelo: $d = \frac{1}{2} g t^2$.

[2] También conocidas como variables independientes o variables exógenas

[3] Conocida también como variable a explicar o variables dependiente o variable endógena

[4] Conocido como "ruido blanco", en el cual se suma a cada valor una cantidad con distribución normal de media el valor real y una varianza que se asume conocida

Observamos en los distintos ejemplos que las variables pueden ser aleatorias o no, las relaciones lineales o no y que cuando no son lineales pueden eventualmente existir transformaciones de las variables que llevan a relaciones lineales.

Se presenta a modo de introducción un enfoque teórico de la regresión funcional, para luego presentar el caso lineal sobre valores muestrales con más detalle. Se usarán dos métodos de estimación para el caso lineal:

- El método matemático de ajuste de los mínimos cuadrados, que permite estimar los coeficientes del modelo lineal a partir de valores observados. En este caso no se toma en cuenta la aleatoriedad de las variables en la estimación del modelo.
- El método de máxima verosimilitud basado en un modelo probabilístico normal, que permite justificar el método de mínimos cuadrados y discutir las propiedades de los estimadores y la precisión del ajuste.

Finalmente se usará el modelo lineal para predecir. Se enfatizará en los aspectos geométricos del problema y como hacer una crítica de los supuestos probabilísticos usuales.

3.2. Modelo teórico condicional.

Proposición 3.2 *Sean la v.a. $Y \in \mathbb{R}$ y el vector aleatorio $X \in \mathbb{R}^p$. El mínimo de $\mathbb{E}\{(Y - f(X))^2\}$ se alcanza en $f(X) = \mathbb{E}(Y|X)$.*

Dem: Geométricamente, en el espacio de Hilbert $\mathrm{L}^2(\mathbb{R}^p)$ tomando como producto escalar

$$< U, V >= \mathbb{E}(UV^t)$$

$\mathbb{E}(Y|X)$ es la proyección ortogonal de Y sobre el subespacio $\mathrm{L}^2(X)$ generado por las funciones de X.

La función que vamos a minimizar se denomina **Error Cuadrático Medio (ECM)**, y el problema a resolver es:

$$\min_{g \in \mathrm{L}^2(X)} \mathbb{E}\{(Y - g(X))^2\}$$

Consideremos $f(X) = \mathbb{E}(Y|X)$y sea $g(X) \in \mathrm{L}^2(X)$, se tiene:

$$\mathbb{E}\{(Y - g(X))^2\} = \mathbb{E}\{(Y - f(X))^2\} + 2\mathbb{E}\{(Y - f(X))(f(X) - g(X)) + \mathbb{E}\{(f(X) - g(X))^2\}$$

pero

$$\mathbb{E}\{(Y - f(X))(f(X) - g(X))\} = \mathbb{E}\{(f(X) - g(X))\mathbb{E}\{(Y - f(X))|X\}\}$$

y dado que $f(X) - g(X)$ es independiente de Y y $\mathbb{E}\{(Y - f(X)|X) = 0$ por definición de esperanza condicional, de donde podemos concluir que:

$$\begin{aligned} \mathbb{E}\{(Y - g(X))^2\} &= \mathbb{E}\{(Y - f(X))^2\} + \mathbb{E}\{(f(X) - g(X))^2\} \\ &\geq \mathbb{E}\{(Y - f(X))^2\} \end{aligned}$$

para toda $g(X) \in \mathrm{L}^2(X)$ por lo que $f(X) = \mathbb{E}(Y|X)$ resulta ser un mínimo, que es lo que queríamos demostrar.

■

Un índice para medir la calidad del modelo está dado por el coeficiente de correlación entre Y y $\mathbb{E}(Y|X)$ cuyo cuadrado es:

$$\eta^2_{Y|X} \equiv Cor^2(Y,\mathbb{E}(Y|X)) = \frac{\mathbb{V}\{\mathbb{E}(Y|X)\}}{\mathbb{V}(Y)} \tag{3.6}$$

$$= 1-\frac{\mathbb{V}(\varepsilon)}{\mathbb{V}(Y)} \tag{3.7}$$

donde $\varepsilon = Y-\mathbb{E}(Y|X)$, y entonces

$$\mathbb{V}(\varepsilon) = (1-\eta^2_{Y|X})\mathbb{V}(Y).$$

de aquí que si $|\eta| \approx 1$ se tiene que $\mathbb{E}(Y|X)$ resulta ser una buena aproximación de la variable aleatoria Y.

En efecto, veamos que se satisface la relación 3.6, por definicion de covarianza tenemos:

$$\mathbb{C}ov(Y,\mathbb{E}(Y|X)) = \mathbb{E}[(y-\mathbb{E}(Y))(\mathbb{E}(Y|X)-\mathbb{E}(Y))]$$

como $\mathbb{E}(\varphi(X,Y))) = \mathbb{E}_X\{\mathbb{E}_{Y|X}(\varphi(X,Y)|X)\}$ se tiene:

$$\mathbb{C}ov(Y,\mathbb{E}(Y|X)) = \mathbb{E}_X\{\mathbb{E}\{(Y-\mathbb{E}(Y))(\mathbb{E}(Y|X)-\mathbb{E}(Y))|X\}\}$$

Ahora bien:

$$\mathbb{E}\{(Y-\mathbb{E}(Y))(\mathbb{E}(Y|X)-\mathbb{E}(Y))|X\} = (\mathbb{E}(Y|X)-\mathbb{E}(Y))(\mathbb{E}(Y|X)-\mathbb{E}(Y)) = (\mathbb{E}(Y|X)-\mathbb{E}(Y))^2$$

y

$$\mathbb{C}ov(Y,\mathbb{E}(Y|X)) = \mathbb{E}_X\{(\mathbb{E}(Y|X)-\mathbb{E}(Y))^2\} = \mathbb{V}(\mathbb{E}(Y|X))$$

Finalmente

$$Cor^2(Y,\mathbb{E}(Y|X)) = \frac{\mathbb{V}(\mathbb{E}(Y|X))^2}{\mathbb{V}(Y)\mathbb{V}(\mathbb{E}(Y|X))} = \frac{\mathbb{V}(\mathbb{E}(Y|X))}{\mathbb{V}(Y)}$$

En el caso lineal $f(X) = \mathbb{E}(Y|X) = \beta^T X$ y $\mathbb{E}(\varepsilon) = 0$.

Minimizar $\mathbb{V}(\varepsilon)$ equivale a tomar $\mathbb{C}ov(\varepsilon,X) = 0$. Luego $\mathbb{C}ov(Y,X) = \beta\mathbb{V}(X)$ en donde $\beta = (\mathbb{V}(X))^{-1}\mathbb{C}ov(Y,X)$:

$$\mathbb{V}(Y) = \mathbb{V}\{\mathbb{E}(Y|X)\} + \mathbb{V}(\varepsilon).$$

3.3. Estimación de los parámetros del modelo lineal

Sean $\{(y_i,x_{i,1},x_{i,2},\ldots,x_{i,p})|i=1,2,\ldots,n\}$ los valores obtenidos sobre una muestra aleatoria simple de tamaño n del vector $(Y,X_1,X_2,\ldots,X_p)$ de $\mathbb{R}^{p+1}$. Se plantea el modelo lineal:

$$\mathbb{E}(Y|X) = (x_{i,1},x_{i,2},\ldots,x_{i,p})) = \beta_0+\beta_1 x_{i,1}+\beta_2 x_{i,2}+\cdots+\beta_p x_{i,p}.$$

consideraremos aquí el vector X como no aleatorio.

Se denota $x_i = (x_{i,1},x_{i,2},\ldots,x_{i,p})\ \ \forall i = 1,2,\ldots,n$ y se supone sobre los errores: $\varepsilon_i = y_i-\mathbb{E}(Y) \sim N(0,\sigma^2)$, independientes entre si e independientes de los x_i. Tenemos entonces $p+2$ parámetros a estimar, que son $\beta_0,\beta_1,\ldots,\beta_p$ y σ^2. Se muestran dos métodos de estimación: el método de ajuste de los mínimos cuadrados y el método de máxima verosimilitud.

3.3.1. Solución de los mínimos cuadrados

Se busca minimizar una función de los errores, como por ejemplo:

$$\sum_{i=1}^{p} \varepsilon_i^2, \qquad \sum_{i=1}^{n} |\varepsilon_i|, \qquad \max_i \{\varepsilon_i\}$$

el criterio de los mínimos cuadrados toma como función $\sum_{i=1}^{n} \varepsilon_i^2$ cuya solución es fácil de obtener y que tiene una interpretación geométrica simple. Primero escribiremos matricialmente el modelo aplicado a la muestra de observaciones.

$$\text{Sea} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

entonces el modelo se escribe

$$Y = X\beta + \varepsilon.$$

El criterio de los mínimos cuadrados consiste en buscar la proyección ortogonal sobre el subespacio vectorial generado por las columnas de la matriz X, ie $W = Im(X) \subset \mathbb{R}^n$, la cual se obtiene de imponer que la derivada de $\varepsilon = Y - X\beta$ con respecto a X sea igual a cero, pues buscamos que la distancia entre entre el hiperplano y la variable Y sea mínima, con esto se obtienen las denominadas **ecuaciones normales** con la métrica usual:

$$X^t X\beta = X^t Y$$

Este sistema de ecuaciones lineales tiene una solución única cuando las columnas de X son linealmente independientes, es decir que forman una base del subespacio vectorial W, o equivalentemente que el rango de X sea igual a $p+1$. En este caso la solución de los mínimos cuadrados es igual a:

$$\hat{\beta} = (X^t X)^{-1} X^t Y.$$

Se deduce que el operador de proyección ortogonal P sobre W, es un operador lineal idempotente de orden 2 y simétrico (¿por qué?), y se escribe matricialmente como:

$$P = X(X^t X)^{-1} X^t$$

Si el rango de X es inferior a $p+1$, basta encontrar una base de W entre las columnas de X, y reemplazar X por X_1 la matriz formada por estas columnas linealmente independientes. Se observará que si bien $\hat{\beta}$ no es necesariamente único, $Y = X\beta = PY$ y $\hat{\varepsilon} = Y - X\hat{\beta} = (I-P)Y$ lo son (ver 3.7). Si bien la manera de entender y plantear este problema resulta ser bastante sencillo, tiene un problema puesto que **no es posible obtener una estimación de la varianza de los errores** σ^2.

3.3.2. Solución de máxima verosimilitud

El método anterior permite estimar los coeficientes β_j usando un criterio matemático que permite ajustar el mejor hiperplano afín de $\mathbb{R}^{p+2}$ utilizando un criterio para minimizar el error. En esta sección usaremos el método de máxima verosimilitud que a diferencia del anterior, involucra la aleatoriedad de los errores en la resolución. El modelo es:

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = X\beta$$

con $Y = \mathbb{E}(Y) + \varepsilon = X\beta + \varepsilon$ y se supone que $\varepsilon \sim N_n(0, \sigma^2 I_n)$. La función de verosimilitud, recordando que es la densidad conjunta de los errores, queda con la forma:

$$\begin{aligned} f(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\varepsilon^t\varepsilon\right) \\ f(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n; \beta, \sigma^2) &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)^t(Y - X\beta)\right) \end{aligned}$$

pues $\varepsilon = Y - \mathbb{E}(Y)$. Calculemos el estimador de máxima verosimilitud de β, tomando primero logaritmo sobre la función de verosimilitud:

$$\begin{aligned} \frac{\partial \ln f}{\partial \beta} = 0 &\Rightarrow \frac{\partial (Y - X\beta)^t (Y - X\beta)}{\partial \beta} = 0 \\ &\Rightarrow (X^t X)\hat{\beta} = X^T Y \end{aligned}$$

que corresponden a las **Ecuaciones Normales**. Como uno de los párametros de la función de verosimilitud es la varianza σ^2, planteamos de igual forma la ecuación para encontrar su estimador $\hat{\sigma}^2$:

$$\begin{aligned} \frac{\partial \ln f}{\partial \sigma^2} = 0 &\Rightarrow \hat{\sigma}^2 = \frac{(Y - X\hat{\beta})^t (Y - X\hat{\beta})}{n} \\ &\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 \end{aligned}$$

puesto que $\hat{\varepsilon} = Y - X\hat{\beta}$. De aquí podemos concluir que la función de verosimilitud es máxima cuando se cumplen las siguientes condiciones:

$$\begin{aligned} (X^t X)\hat{\beta} &= X^t Y \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 \end{aligned}$$

Esta última expresión es denominada varianza residual y corresponde a la varianza empírica de los ε_i.

Así el estimador de los mínimos cuadrados de β es igual al estimador de máxima verosimilitud cuando se supone que $\varepsilon \sim N_n(0, \sigma^2 I_n)$.

3.4. Propiedades del estimador

Mostramos en esta sección una serie de propiedades que tiene el estimador $\hat{\beta}$, solución de las ecuaciones normales, que están ligadas a los supuestos hechos sobre los errores ε_i.

Proposición 3.3 (Propiedades) *Supongamos que X es de rango completo, luego se tienen las siguientes propiedades:*

1. *El estimador $\hat{\beta}$ es un estimador insesgado de β:*

$$\mathbb{E}(\varepsilon) = 0 \Rightarrow \mathbb{E}(Y) = X\beta \Rightarrow \mathbb{E}(\hat{\beta}) = \beta$$

2. *El estimador $\hat{Y} = PY = X\hat{\beta}$ es un estimador insesgado de $X\beta$.*

3. *$\varepsilon \sim N_n(0, \sigma^2 I_n) \Rightarrow \hat{Y} \sim N_n(X\beta, \sigma^2 P)$ donde P era el proyector ortogonal sobre W en $\mathbb{R}^n$.*

4. $\varepsilon \sim N_n(0,\sigma^2 I_n) \Rightarrow \hat{\varepsilon} \sim N_n(o,\sigma^2(I_n - P))$, *con* $\hat{\varepsilon}$ *ortogonal a* W *o* $\hat{\varepsilon}$ *independiente de* X.

5. $\hat{\beta} \sim N_{p+1}(\beta,\sigma^2(X^tX)^{-1})$.

6. $\dfrac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ *es un estimador sesgado para* σ^2.

7. $\hat{\beta}$ *es independiente de* $\sum_{i=1}^{n}\hat{\varepsilon}_i^2$.

8. $\hat{\beta}$ *es un estimador consistente para* β *y* $\tilde{\sigma}^2 = \dfrac{1}{n-p-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ *es consistente para* σ^2.

Dem: Se presenta solo la demostración de la propiedad 6, las demás quedan como ejercicio para el lector.

El estimador es sesgado puesto que:

$$\mathbb{E}\left(\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \middle| X\right) = (n-p-1)\sigma^2 \tag{3.8}$$

luego

$$\tilde{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

resulta ser un estimador insesgado para σ^2. Ahora bien la expresión 3.8 proviene del hecho que

$$\hat{\varepsilon} = (I_n - P)Y = (I_n - P)X\beta + (I_n - P)\varepsilon = (I_n - P)\varepsilon.$$

Luego

$$\hat{\varepsilon}^t\hat{\varepsilon} = \varepsilon^t(I_n - P)\varepsilon = Traza((I_n - P)\varepsilon\varepsilon^t)$$

y

$$\mathbb{E}(\hat{\varepsilon}^t\hat{\varepsilon}) = \sigma^2 Traza((I_n - P)) = (n-p-1)\sigma^2$$

que era lo que queríamos demostrar.

■

Uno de los resultados importantes en esta sección es la optimalidad que tiene el estimador $\hat{\beta}$ bajo ciertas condiciones, que se traduce en el *Teorema de Gauss-Markov*. Primero introducimos una noción para comparar matrices:

Definición 3.1 *Sean* $A, B \in M_n(\mathbb{R})$. *Se dice que* $A \leq B$ *si y solamente si* $B = A + C$, *en donde* C *es semi-definida positiva*

Teorema 3.1 (Teorema de Gauss-Markov) *Si* $\mathbb{E}(\varepsilon) = 0$ *y* $\mathbb{E}(\varepsilon\varepsilon^T) = \sigma^2 I_n$, *entonces* $\hat{\beta}$ *tiene varianza mínima entre los estimadores insesgados de* β, **lineales en** Y. *Además si* $\varepsilon \sim N_n(0,\sigma^2 I_n)$, *entonces* $\hat{\beta}$ *tiene mínima varianza entre* **todos** *los estimadores insesgados de* β.

Dem: Si entre los estimadores insesgados de β y lineales en Y, $\hat{\beta}$ tiene la varianza más pequeña, hay que mostrar que:

$$\forall \beta^* = GY : \mathbb{E}(\beta^*) = \beta \Rightarrow \mathbb{V}(\hat{\beta}) \leq \mathbb{V}(\beta^*).$$

Sea $\hat{\beta} = AY$ una solución de las ecuaciones normales, entonces $\beta^* = \hat{\beta} + DY$, en que $D = G - A$.

Como los dos estimadores son insesgados, $\mathbb{E}(\beta^* - \hat{\beta}) = \mathbb{E}(DY) = 0$ y como $Y = X\beta + \varepsilon$ entonces $DX = 0$. Calculemos la varianza de β^*:

$$\mathbb{V}(\beta^*) = \mathbb{V}(\hat{\beta}) + \mathbb{V}(DY) + 2\mathbb{C}ov(\hat{\beta}, DY)$$

en donde

$$\begin{aligned}
\mathbb{C}ov(\hat{\beta}, DY) = \mathbb{E}(\hat{\beta}(DY)^t) &= \mathbb{E}(\hat{\beta}Y^tD^t) \\
&= \mathbb{E}((X^tX)^{-1}X^tYY^tD^t) \\
&= (X^tX)^{-1}X^t\mathbb{E}(YY^t)D^t \\
&= (X^tX)^{-1}[\mathbb{V}(Y) + \mathbb{E}(Y)\mathbb{E}(Y)^t]D^t \textit{ revisar} \\
&= 0
\end{aligned}$$

Finalmente $\mathbb{V}(\beta^*) = \mathbb{V}(\hat{\beta}) + \sigma^2 DD^t$ en donde DD^t es semi-definida positiva.

Si además los errores siguen una distribución normal, el estimador $\hat{\beta}$ es de mínima varianza entre todos los estimadores insesgados de β. En efecto la cantidad de información de Fisher de la muestra multivariada para el parámetro β es:

$$I_n(\beta) = \frac{1}{\sigma^2}X^tX$$

y el estimador $\hat{\beta}$ tiene una matriz de varianza igual a $\sigma^2(X^tX)^{-1}$. Luego se obtiene la igualdad en la desigualdad de Cramer-Rao, lo que concluye la última aseveración.

■

Se obtiene facílmente una generalización de este teorema cuando $\mathbb{V}(\varepsilon) = \Gamma$, donde se supone que Γ es invertible. El estimador de mínima varianza es entonces:

$$\hat{\beta} = (X^t\Gamma^{-1}X)^{-1}X^t\Gamma^{-1}Y$$

que se interpreta como la proyección ortogonal en el sentido de la métrica Γ^{-1} (Vea esto como un ejercicio propuesto).

3.5. Calidad del modelo

Para ver si el modelo es válido, se deben realizar varios estudios, entre los que se cuentan: la verificación de los supuestos sobre los errores, la forma y significación de las dependencias y el aporte de cada variable explicativa. Lo que se hará estudiando, mediante gráficos, índices y test de hipótesis, además no solamente se debe verificar la calidad del modelo global y el aporte individual de cada variable explicativa, sino que también el aporte de un grupo de m variables explicativas cualquiera.

3.5.1. Calidad global del modelo.

Los residuos $\hat{\varepsilon}_i$ dan la calidad del ajuste para cada observación de la muestra. Pero es una medida individual que depende de la unidad de medición, para evitar este problema un índice adecuado es :

$$\frac{\sum_{i=1}^{n}\hat{y}_i^2}{\sum_{i=1}^{n}y_i^2}$$

que representa el cuadrado del coseno del ángulo del vector Y con el vector $\hat{Y}$ en $\mathbb{R}^n$ (Figura 3.1).

Existen tres varianzas que son candidatas para comparar y que colocamos a continuación:

- Varianza residual: $\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$.
- Varianza explicada por el modelo: $\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$.
- Varianza total: $\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$.

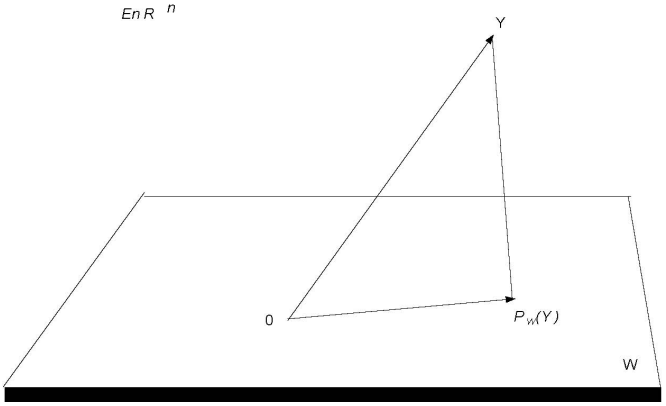


Figura 3.1: Proyección del vector Y en W

Un índice estadísticamente más interesante es el **coeficiente de correlación múltiple** R o su cuadrado, el **coeficiente de determinación**:

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

que compara la varianza explicada por el modelo con la varianza total. El coeficiente de correlación múltiple R también denominado coeficiente de correlación lineal entre Y e $\hat{Y}$[5] satisface, de su definición, que su valor está comprendido entre 0 y 1, así cuando $R=0$, se obtiene que $\hat{y}_i=\bar{y}\ \ \forall i$ ie, la mejor estimación para los valores y_i corresponde a la media muestral de los valores y_i y en consecuencia las variables no explican nada en el modelo. En cambio cuando R es igual a 1, el vector Y pertenece al subespacio vectorial $W=im(X)$, es decir que existe un modelo lineal que permite escribir las observaciones y_i exactamente como combinación de las variables explicativas. Como este ultimo caso corresponde a un valor extremadamente favorable y que en general no se tiene, se utiliza como referencia un valor de R cercano a 1, para decir que el modelo es bueno y que los valores estimados $\hat{y}_i$ ajustan bien a los valores observados y_i.

Para el caso general se tiene:

$$Corr(Y,\hat{Y}) = \max_{Z=X\beta} Corr(Y,Z) = \frac{\|\hat{Y}-\bar{y}1_n\|}{\|Y-\bar{y}1_n\|}$$

[5]Esto pues estamos planteando una relación de tipo lineal para Y

en donde 1_n es el valor de la bisectriz de $\mathbb{R}^n$ de componentes todas iguales a 1.

Ahora bien tal como se estudio en el capítulo pasado, la idea es poder comprobar las hipótesis que hemos supuesto sobre modelo completo. De esta manera se plantea la hipótesis global:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \iff H_0 : E(Y) = \beta_0 1_n$$

esta hipótesis se interpreta como que los valores de las p variables explicativas no influyen en los valores de Y. Nuestro objetivo ahora es descubrir cuales son los estadísticos que utilizaremos para decidir el test. Recordando que $\hat{\varepsilon} \sim N_n(0, \sigma^2(I_n - P))$ e $\hat{Y} \sim N_n(X\beta, \sigma^2 P)$, y suponiendo que r es el rango de la matriz X, se tiene:

$$\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\sigma^2} = \frac{(n-r)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-r}.$$

Notemos que bajo la hipótesis H_0 se tiene que $\hat{Y}|_{H_0} \sim N_n(\beta_1 1_n, \sigma^2 P)$ que es equivalente a que $\hat{\beta}_0 = \bar{y}$, de donde se puede deducir que:

$$\sum_{i=1}^{n} \left(\frac{y_i - \beta_0}{\sigma} \right)^2 = \sum_{i=1}^{n} \left(\frac{\hat{y}_i - \bar{y}}{\sigma} \right)^2 \sim \chi^2_{r-1}$$

Además $\frac{1}{\sigma^2} \sum_{i=1}^{n} \hat{y}_i^2$ y $\sum_{i=1}^{n} \left(\frac{\hat{y}_i - \bar{y}}{\sigma} \right)^2$ son independientes. Se tiene entonces que bajo la hipótesis nula H_0:

$$F \equiv \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2/(r-1)}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2/(n-r)} \sim F_{r-1,n-r}$$

en donde $F_{r-1,n-r}$ es la denominación usual para una distribución de Fisher a $r-1$ y $n-r$ grados de libertad. De la definición de F es posible deducir una expresión para este en función del coeficiente de correlación múltiple R, la cual enunciamos a continuación:

$$F = \frac{(n-r)R^2}{(r-1)(1-R^2)}.$$

De los desarrollos hechos la región crítica para la hipótesis nula $H_0 : E(Y|X) = \beta_0 1_n$ contra la hipótesis alternativa $H_1 : E(Y|X) = X\beta$ con un nivel de significación α queda definida por

$$\begin{aligned} \Omega_r &= \{F > c_\alpha\} \text{ donde } c_\alpha \text{ satisface} \\ \mathbb{P}(\text{Rechazar } H_0 | H_0 \text{ es cierto}) &= \mathbb{P}(F_{r-1,n-r} > c_\alpha) = \alpha \end{aligned}$$

Cuando se rechaza H_0, es decir cuando se encuentra un valor de F en la muestra mayor que c_α, se declara el modelo globalmente significativo.

En la práctica, se define la **Probabilidad crítica** o ***p*-valor** que corresponde al valor p_c que satisface $\mathbb{P}(F_{r-1,n-r} > F) = p_c$. Si el valor de la probabilidad crítica p_c es bajo con respecto a algún valor de referencia α dado, entonces se rechaza H_0, es decir se declara el modelo como significativo y alguna de las variables X_j si incide en el valor de Y. Caso contrario por supuesto no se rechaza H_0 pues puede que ninguna de las variables incida en el valor de Y.

3.5.2. Medición del efecto de cada variable en el modelo

En la sección anterior revisamos cual era el efecto global de las variables del modelo, lo cual nos impide distinguir del aporte significativo que tenga cada una de ellas, para esto se realizan test de hipótesis asociados con cada una de ellas.

Para realizar estos test suponemos que las variables explicativas **son independientes**, además recordemos que el efecto asociado a la variable X_j se mide con el término$X_j\hat{\beta}_j$. Gracias a estos test se observará que el modelo lineal es invariante por el cambio de escalas de mediciones.

Consideramos las hipótesis:

$$\begin{aligned} H_0 : \beta_j &= 0 \\ H_1 : \beta_j &\neq 0 \end{aligned}$$

Recordando que:

$$\begin{aligned} \hat{\beta}_j &\sim N(\beta_j, \sigma_j^2) \quad \text{con} \\ \sigma_j^2 &= Var(\hat{\beta}_j) = \sigma^2 (X^t X)^{-1}_{j,j} \end{aligned}$$

sujeto a considerar un modelo con matriz X de rango completo. Luego se tiene que:

$$\frac{\hat{\beta}_j - \beta_j}{\sigma_j} \sim N(0,1)$$

Por otra parte, como:

$$\frac{(n-r)\hat{\sigma}}{\sigma^2} \sim \chi^2_{n-r} \Longrightarrow \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \sim t_{n-r}.$$

y bajo la hipótesis nula $H_0 : \beta_j = 0$ se tiene que:

$$\frac{\hat{\beta_j}}{\hat{\sigma}_j} \sim t_{n-r}.$$

donde t_{n-r} es la denominación usual de una distribución *t- student* a $n-r$ grados de libertad. Note que gracias a esto es posible deducir intervalos de confianza para cada uno de los $\hat{\beta}_j$.

Luego si la probabilidad crítica o P-valor definida como:

$$\mathbb{P}\left(|t_{n-r}| > \frac{\hat{\beta}_j}{\hat{\sigma}_j} \right) = p_c$$

es grande, no se rechaza H_0 y si es pequeña se rechaza H_0, lo que en este caso muestra un efecto significativo de la variables X_j sobre Y.

Estos test individuales sobre los efectos tienen validez cuando las variables explicativas son relativamente independientes. Cuando no es el caso, es decir cuando una variable X_j puede tener un efecto sobre Y distinto cuando se combina con otras variables, hay entonces que eliminar los efectos de las otras variables. Para eso se puede usar el **coeficiente de correlación parcial** que estudiaremos en la próxima sección.

3.5.3. Coeficiente de correlación parcial

El efecto de una variable X_1 sobre la variable Y puede estar afectado por una tercera variable X_2 cuando X_2 tiene efecto sobre X_1 también. El estudio se basa entonces en las dos relaciones del tipo lineal:

$$\begin{aligned} X_1 &= \alpha X_2 + \vartheta \\ Y &= \gamma X_2 + \eta \end{aligned}$$

Donde ϑ y η son los errores del modelo. La influencia de la variable X_2 sobre las variables X_1 e Y se mide solamente a partir de los restos:

$$\begin{aligned} X_1 - \alpha X_2 &= \vartheta \\ Y - \gamma X_2 &= \eta \end{aligned}$$

Definición 3.2 *El coeficiente de correlación parcial entre X_1 e Y bajo X_2 constante es el coeficiente de correlación entre los errores ϑ y η:*

$$\rho(X_1, Y|X_2) = Corr(\vartheta, \eta)$$

Con esto, si X_1 y X_2 son muy correlacionados entonces la correlación parcial entre X_1 e Y es muy pequeña. En efecto X_1 aporta casi ninguna información nueva sobre Y (o vice-versa) cuando X_2 es conocida.

Se puede usar un gráfico de los errores para medir los efectos y el tipo de efecto (lineal o no): en el gráfico (3.2) se presentan los errores de la regresión de X_1 tomando como variable explicativa X_2, versus, la regresión de Y tomando X_2. En (a) podemos decir que la variable X_2 no tiene efecto sobre la variable Y en presencia de la variable X_1; pero en (b) la variable X_2 aporta a la explicación de la variable Y aún si la variable X_1 es presente en el modelo.

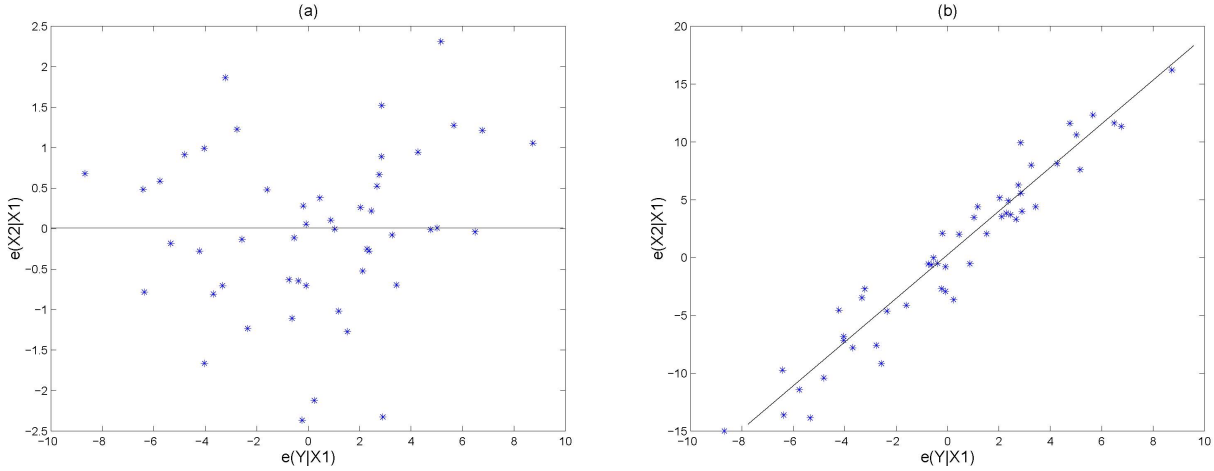


Figura 3.2: Interpretación de los errores y del coeficiente de correlación parcial

Se puede generalizar a más de 2 variables X_j, $j = 2,\ 3,\ \ldots,\ q$. Si

$$X = \sum_{j=2}^{q} \alpha_j X_j + \vartheta \quad Y = \sum_{j=2}^{q} \gamma X_j + \eta$$

entonces se define el coeficiente de correlación parcial entre X_1 e Y, dadas las variables X_j, por:

$$\rho(X_1, Y|X_2, X_3,\ \ldots,\ X_q) = Corr(\vartheta, \eta).$$

Si las variables X_j no tienen efecto sobre X_1 e Y, es decir que las correlaciones $Corr(X,X_j)$ y $Corr(Y,X_j)$ son todas nulas, entonces $\rho(X,Y|X_2,X_3,\ \ldots,\ X_q) = Corr(X,Y)$.

Se tiene una intertretación geométrica del coeficiente de correlacion parcial; pensemos en sólo tres variables, sean r_{X_1} y r_Y los residuos de las regresiones de X_1 e Y sobre X_2, ie r es el vector de las diferencias entre los valores a estimar y su estimadores. Un cálculo simple mostrará que r_{X_1} y r_Y viven en el plano perpendicular a X_2, entonces el coeficiente de correlación es el coseno del ángulo entre ellos en ese plano.

En la misma dirección que el coeficiente de correlación parcial definimos la matriz de varianza-covarianza del vector X dado el vector Z cuando el vector conjunto es una normal:

$$\begin{pmatrix} X \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Z \end{pmatrix}, \begin{pmatrix} \Gamma_{XX} & \Gamma_{XZ} \\ \Gamma_{XZ} & \Gamma_{ZZ} \end{pmatrix} \right)$$

Luego

$$\mathbb{V}(X|Z) = \Gamma_{XX} - \Gamma_{XZ}\Gamma_{ZZ}^{-1}\Gamma_{ZX}$$

En general ρ y V no son conocidos, por lo cual se estiman con la matriz de varianza covarianza muestral S, dando (considerando la misma descompocisión que en Γ):

$$\begin{aligned} \hat{\mathbb{V}}(X|Z) &= S_{X\bullet Z} = S_{XX} - S_{XZ}S_{ZZ}^{-1}S_{ZX} \\ \hat{\rho}(X_1,Y|X_2) &= \frac{Corr(X_1,Y) - Corr(X_1,X_2)Corr(Y,X_2)}{\sqrt{1-Corr(X_1,X_2)^2}\sqrt{1-Corr(Y,X_2)^2}} \end{aligned}$$

Se usa el siguiente estadístico para estudiar la hipotesis nula $H_0 : \rho(X,Y|X_2,X_3,\ \ldots,\ X_q) = 0$:

$$t = \hat{\rho}(X_1,Y|X_2)\sqrt{\frac{n-2-k}{1-\hat{\rho}(X_1,Y|X_2)}} \sim t_{n-2-k}$$

Bajo los supuestos de normalidad esta distribución es exacta, si no, es sólo es una aproximación para grandes muestras. Con este estadístico rechazamos H_0 si $|t| > C$ con C tal que $\mathbb{P}(t_{n-2-k} > C) = \alpha/2$, donde α es el nivel de significancia.

3.5.4. Efecto de un grupo de variables

Vimos como estudiar el efecto global de todas las variables explicativas y los efectos individuales, aquí veremos el efecto de un grupo de k variables, sean $X_{j_1}, X_{j_2}, \ldots, X_{j_k}$ ($k \leq p$), entre las p variables. El efecto de estas variables se mide considerando las hipótesis:

$$\begin{aligned} H_0 &: \beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_k} = 0 \\ H_1 &: \mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \end{aligned}$$

Si $X_{j_{k+1}}, X_{j_{k+2}}, \ldots, X_{j_p}$ son las restantes variables explicativas, bajo H_0, el modelo se escribe: $Y = \gamma_0 + \gamma_{j_{k+1}} X_{j_{k+1}} + \cdots + \gamma_{j_p} X_{j_p} + \varepsilon_o$. Notar que la varianza residual bajo H_1, $\sum_i \hat{\varepsilon}_i^2$, menor que la varianza residual bajo H_0, $\sum_i \hat{\varepsilon}_{oi}^2$.

Se puede estudiar el cociente de éstas varianzas $(\sum_i \hat{\varepsilon}_{oi}^2)/(\sum_i \hat{\varepsilon}_i^2)$ o su complemento $(\sum_i \hat{y}_{oi}^2)/(\sum_i \hat{\varepsilon}_i^2)$ en donde $\hat{y_{oi}}$ son las componentes del estimador $\mathbb{E}(Y|X)$ bajo H_0.

Bajo la hipótesis H_0 se tiene que

$$Q = \frac{\sum_i (\hat{y}_i - \hat{y}_{oi})^2/k}{\sum_i \hat{\varepsilon}_i^2/(n-r)} \sim F_{k,n-r}.$$

Lo que conduce a un test de región crítica de la forma $Q \geq c_\alpha$.

Se puede considerar otra forma de escribir el problema. Sea la hipótesis nula $H_0 : \mathbb{E}(Y) = X_0\beta \in W_0$, con X_0 de rango s, contra $H_1 : \mathbb{E}(Y) = X\beta \in W$. La hipótesis H_0 equivale a $(X - X_0)\beta = 0$ lo que corresponde a $k = p - s + 1$ ecuaciones independientes $D\beta = 0$, en que D es de $x \times p + 1$ con rango k. Para que el test tenga sentido, $D\beta$ no debe depender de una solución particular $\hat{\beta}$ de las ecuaciones normales.

Sean $\hat{Y}$ e Y_0 las proyecciones Y sobre W y W_0, respectivamente, además sean μ y μ_0 las medias bajo H_1 y H_0. Se define $S^2 = \frac{\|Y - Y^*\|^2}{\|Y - \hat{Y}\|^2}$ y $R^2 = \frac{\|\hat{Y} - Y^*\|^2}{\|Y - \hat{Y}\|^2}$, sigue que, bajo H_0, se tiene $\frac{n-p-1}{k} R^2 \sim F_{k,n-r}$. Con lo que la región crítica es de la forma $\frac{n-r}{k} R^2 > C$ con C tal que $\mathbb{P}(\frac{n-r}{k} R^2 > C) = \alpha$ (α el nivel de significacia).

Se puede plantear el test de razón de verosimilitudes también: $\Lambda = \frac{\max_{H_0} L}{\max L}$. Demostraremos que la región crítica se escribe $S > C'$ y que éste test coincide con el test F. Cuando la varianza σ^2 es conocida, la razón de verosimilitudes es igual a:

$$\Lambda = \frac{\max\limits_{H_0} L}{\max L} = \exp\left\{ -\frac{1}{2\sigma^2} \|\hat{Y} - Y_0\|^2 \right\}.$$

Luego la región crítica del test se escribe usando el hecho que $\|\hat{Y} - Y_0\|^2 > \sigma^2 \chi^2_k$, al tomar σ^2 desconocida, caemos en el caso anterior, ie, usamos un estadístico basado en una F de Fisher.

3.5.5. Caso de una hipótesis lineal general

Sea la hipótesis nula $H_0 : A\beta = c$ contra la hipótesis alternativa $H_1 : A\beta \neq c$, en donde $A \in M_{k,p+1}$ es conocida y de rango k. Se supondrá aquí un modelo de rango completo.

Sea $\hat{\beta} = (X^t X)^{-1} X^t Y$ el estimador de máxima verosimilitud sin restricción y $\hat{\beta}_0$ el estimador bajo $H_0 : A\beta = c$. Se obtiene $\hat{\beta}_0$ usando los multiplicadores de Lagrange:

$$\begin{array}{rcrcl} & & Q & = & (Y - X\beta)^t (Y - X\beta) + 2\lambda(A\beta - c) \\ \frac{\partial Q}{\partial \beta} = 0 & \Rightarrow & X^t X \hat{\beta}_0 & = & X^t Y + A^t \lambda \\ & \Rightarrow & \hat{\beta}_0 & = & (X^t X)^{-1}(X^t Y + A^t \lambda) \\ & & & = & \hat{\beta} + (X^t X)^{-1} A^t \lambda. \end{array}$$

Utilizando la restricción $A\hat{\beta}_0 = c$, obtenemos que $\lambda = [A(X^t X)^{-1} A^t]^{-1}(c - A\hat{\beta})$

$$\Rightarrow \hat{\beta}_0 = \hat{\beta} + (X^t X)^{-1} A^t [A(X^t X)^{-1} A^t]^{-1}(c - A\hat{\beta})$$

Sean P_0 y P los proyectores asociados respectivamente a $X\hat{\beta}_0$ y $X\hat{\beta}$, es decir tales que $P_0 Y = X\hat{\beta}_0$ y $Py = X\hat{\beta}$. Entonces

$$P_0 Y = PY + X(X^t X)^{-1} A^t [A(X^t X)^{-1} A^t]^{-1}(c - A\hat{\beta}).$$

Sea la varianza residual del modelo sin restricción: $V = (Y - X\hat{\beta})^t (Y - X\hat{\beta})$ y la varianza residual bajo H_0 : $T = (Y - X\hat{\beta}_0)^t (Y - X\hat{\beta}_0)$. Como $T \geq V$, consideraremos $U = T - V$ para compararlo a V.

Proposición 3.4 *La diferencia de las varianzas residuales con y sin restricción es:*

$$U = (A\hat{\beta} - c)^t [A(X^tX)^{-1}A^t]^{-1}(A\hat{\beta} - c)$$

Además, bajo la hipótesis nula $\frac{U}{\sigma^2} \sim \chi^2_k$.

Dem:

$$U := (Y - X\hat{\beta}_0)^t(Y - X\hat{\beta}_0) - (Y - X\hat{\beta})^t(Y - X\hat{\beta}) = Y^t(P - P_0)Y.$$

Reemplazando $P_0Y = PY + X(X^tX)^{-1}A^t[A(X^tX)^{-1}A^t]^{-1}(c - A\hat{\beta})$ sigue el resultado.

Por otro lado como A es de rango igual a k, $A\hat{\beta} \sim N_k(A\beta, \sigma^2 A(X^tX)^{-1}A^t)$, luego $\frac{U}{\sigma^2} \sim \chi^2_k$.

■

Como $\hat{\beta}$ es independiente de $V = \sum_i \hat{\varepsilon}_i^2$, usamos como el estadístico del test:

$$\frac{U/k}{V/(n-p)} \sim F_{k,n-p}.$$

3.6. Análisis de los residuos

Se supone que el efecto de numerosas causas no identificadas está contenido en los errores, lo que se traduce como una perturbación aleatoria. De aquí los supuestos sobre los errores, que condicionan las propiedades del estimador. Es importante entonces comprobar si los supuestos se cumplen.

La mejor forma de chequear si los errores son aleatorios de medias nulas, independientes y de la misma varianza, consiste en estudiar los residuos

$$\forall i = 1, 2, \ldots, n : \hat{\varepsilon}_i = y_i - \sum_j \hat{\beta}_j x_{i,j}$$

considerandolos como muestra i.i.d. de una distribución normal.

Se puede usar el gráfico $(Y_i, \hat{\varepsilon}_i)$, que debería mostrar ninguna tendencia de los puntos, o bien construir test de hipótesis sobre los errores. En el gráfico de la izquierda de la figura 3.3 se puede ver los residuos aleatorios independientes de Y, lo que no es el caso de los residuos del gráfico de la derecha.

Si el supuesto que los errores son $N(0, \sigma^2)$ no se cumple, tenemos que estudiar el efecto que esto tiene sobre la estimación de los parámetros y sobre los tests de hipótesis, además tenemos que detectar si este supuesto es cierto o no y corregir eventualmente la estimación de los parámetros y tests.

Vimos donde interviene el supuesto de normalidad en la estimación de los parámetros del modelo y en los tests de hipótesis para verificar la significación de las variables en el modelo. Este tema se relaciona con el concepto de la *robustez* (ver MILLER R.G. (1986), Beyond ANOVA, Basics of Applied Statistics). La teoría de estimación y de test de hipótesis se basa en supuestos sobre la distribución de población. Por lo tanto si estos supuestos son inexactos, la estimación o la conclusión del test sera distorsionada. Se buscan entonces métodos que sean lo menos sensibles a la inexactitud de los supuestos. Se habla entonces de robustez del método.

Se divide el estudio en dos partes: la normalidad y la igualdad de las varianzas de los errores.

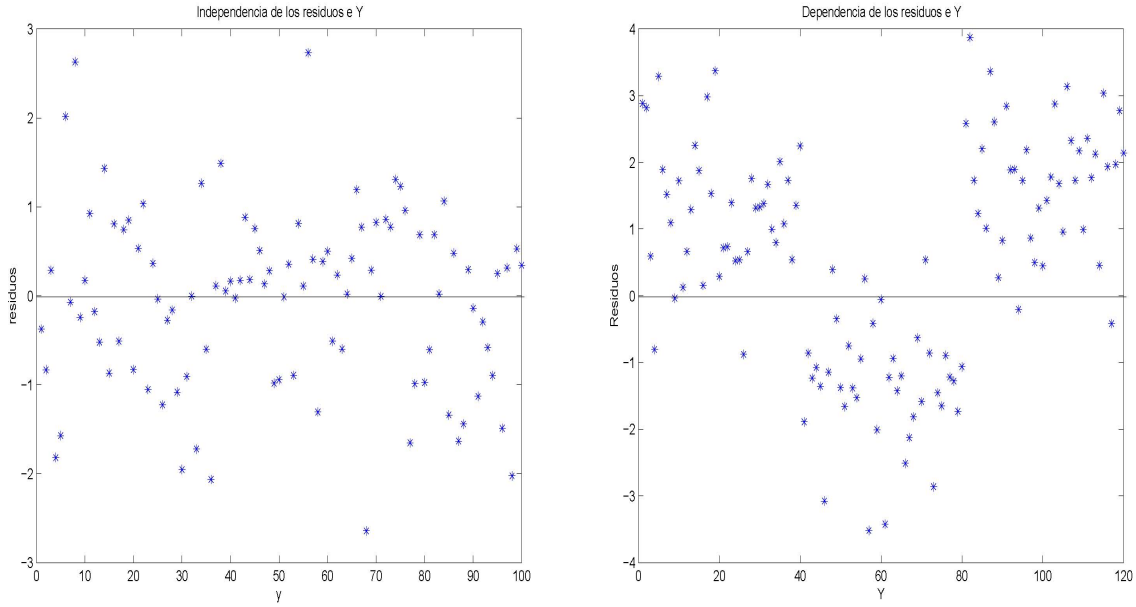


Figura 3.3: Gráficos de residuos

3.6.1. Estudio de la normalidad de los errores

Si no se cumple la normalidad de los errores, los efectos sobre la estimación o tests relativos a los parámetros son pequeños, pero son más importantes sobre los tests relativos a coeficiente de correlación. El problema es más agudo en presencia de observaciones atípicas.

Tenemos entonces que verificar la hipótesis nula $H_o : \varepsilon_i \sim N(0, \sigma^2)$ o sea si $u_i = \varepsilon_i/\sigma$, $H_0 : u_i \sim N(0,1)$. Esto sugiere comparar la función de distribución empírica F_n de los residuos normalizados con la función de distribución de la $N(0,1)$.

Sea F la función de distribución de la $N(0,1)$ (que es invertible), entonces si los u_i provienen de $N(0,1)$, $F^{-1}(F_n(u_i)) \approx u_i$. Consideramos entonces los estadísticos de orden de los u_i, que son los residuos normalizados ordenados de menor a mayor: sea $u_{(1)} \leq u_{(2)} \leq ... \leq u_{(n)}$. La función de distribución empírica es entonces:

$$F_n(u) = \frac{card\{u_{(i)} \leq u\}}{n}$$

Se define los cuantiles empíricos $q_i = F^{-1}(F_n(u_{(i)})$. Notemos que $F_n(u_{(i)}) = F_n(\varepsilon_{(i)})$.

Si F_n se parece a F, los puntos (u_i, q_i) deberían ser colineales (sobre la primera bisectriz). Este gráfico se llama *probit* o *recta de Henri* (gráfico 3.4).

Si los puntos en ell gráfico probit aparecen como no lineal, se rechaza la normalidad de los errores y se puede corregir utilizando la regresión no paramétrica o bien otras alternativas según la causa de la no normalidad: no simetría, observaciones atípicas, etc..

3.6.2. Test para la igualdad de las varianzas

Queremos probar el supuesto de igualdad de varianzas u homocestacidad. En rigor queremos probar la validez de la hipótesis nula $H_0 : \mathbb{V}(\varepsilon|X_1, \ldots, X_p) = \sigma^2$, como $\mathbb{E}(\varepsilon) = 0$, podemos escribir la hipotesis nula

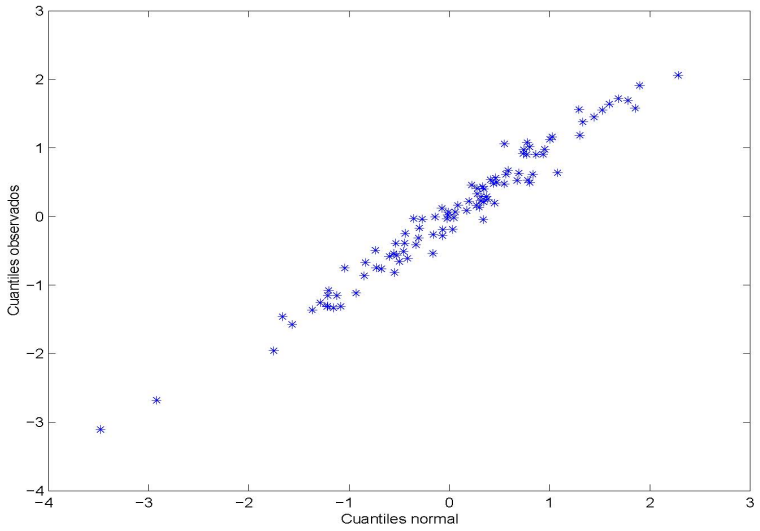


Figura 3.4: Recta de Henri

como:

$$\mathbb{E}(\varepsilon^2|X_1,\dots,X_p) = \mathbb{E}(\varepsilon^2) = \sigma^2$$

En este contexto, tenemos que la esperanza condicional de los errores sobre las variables corresponde a:

$$\varepsilon^2 = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_p X_p + u_i$$

Luego se puede pensar en lo anterior como una regresión para ε^2 usando las variables explicativas. H_0 implica que todos los coeficientes son iguales a 0, y esto equivale a hacer un test usual F sobre la importancia de la regresión de ε^2 sobre las variables X (ver 3.5.1). Como ε no es conocido, se usan los estimados durante la regresión original, $\hat{\varepsilon}$.

3.7. Caso de modelo de rango incompleto.

Vimos que en el caso de una matriz X de rango r menor que $p+1$, la solución de las ecuaciones normales no es única, se habla en este caso de **modelo de rango incompleto**. Construyendo una solución de las ecuaciones normales a partir de una inversa generalizada $A = (X^tX)^-$ no se obtiene necesariamente un estimador insesgado de β. En efecto, si $\hat{\beta}$ es una solución de las ecuaciones normales, ie $(X^tX)\hat{\beta} = X^tY$, entonces $\hat{\beta} = AX^tY$ y sique que $\mathbb{E}(\hat{\beta}) = AX^t\mathbb{E}(Y) = AX^tX\beta$.

Sin embargo, $\hat{Y} = X\hat{\beta} = (XAX^t)Y$ es único, dado que XAX^t no depende de la inversa generalizada A. Luego $\mathbb{E}(\hat{Y}) = \mathbb{E}(X\hat{\beta}) = (XAX^t)X\beta = X\beta$. Los vectores $\hat{Y}$ de las predicciones y $\hat{\varepsilon}$ de los residuos son invariantes e insesgados y $\hat{\sigma}^2 = (Y^t(1-XAX^t)Y)/(n-r)$, el estimador de σ^2, lo es también.

Se presentan tres enfoques para estudiar estos modelos de rango incompleto

- mediante un modelo reducido;
- a partir de funciones estimables;
- mediante restricciones identificables sobre los coeficientes del modelo.

También se verán las relaciones que existen entre estos métodos.

3.7.1. El modelo reducido.

Sea X de rango r ($r < p+1$) y $U \in M_{r,p+1-r}$ tal que si $X = (X_1|X_2)$ con X_1 de rango completo r, $X_2 = X_1U$. Entonces, si $\beta = (\beta_1\ \beta_2)^t$:

$$X\beta = X_1\beta_1 + X_2\beta_2 = X_1(\beta_1 + U\beta_2) = X_1\beta^+$$

El modelo se escribe entonces: $Y = X\beta + \varepsilon = X_1\beta^+ + \varepsilon$, que es un modelo de rango completo sobre X_1 equivalente al modelo de rango incompleto:

$$\left\{ \begin{array}{rcl} \hat{\beta}^+ & = & \hat{\beta}_1 + U\hat{\beta}_2 = (X_1^tX_1)^{-1}X_1^tY \\ \mathbb{E}(\hat{\beta}^+) & = & \beta^+ \\ \mathbb{V}(\beta^+) & = & \sigma^2(X_1^tX_1)^{-1} \end{array} \right.$$

3.7.2. Funciones vectoriales estimables

Definición 3.3 *Sea $\mathbb{R}^{p+1}$ el espacio vectorial de los parámetros β y $G = R^k$. Una aplicación lineal $H : \mathbb{R}^{p+1} \to \mathbb{R}^k$ es* **estimable** *si existe una aplicación lineal $L : \mathbb{R}^n \to G$ tal que $\mathbb{E}(LY) = H\beta$.*

Veamos a continuación algunas condiciones para que H sea estimable.

Teorema 3.2 *Una condición necesaria y suficiente para que $H : E \to G$ sea estimable es que existe $L \in l(\mathbb{R}^n, \mathbb{R}^k)$ (ie $L \in M_{k,n}$) tal que $LX\beta = H\beta$.*

Dem:

$$\begin{array}{rcl} H \text{ es estimable} & \iff & \exists L \in l(\mathbb{R}^n, \mathbb{R}^k) \text{ tq } E(LY) = H\beta \\ & \iff & \mathbb{E}(L(X\beta + \varepsilon)) = H\beta \\ & \iff & LX\beta = H\beta \end{array}$$

■

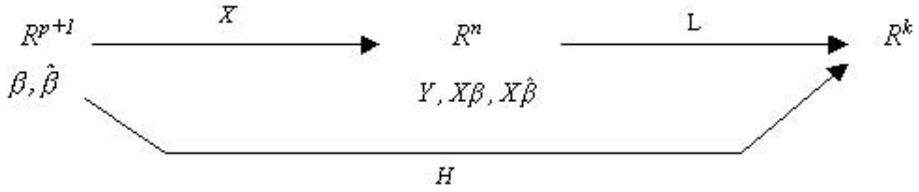


Figura 3.5: Definición de función estimable

Teorema 3.3 *Una condición necesaria y suficiente para que H sea estimable es que $Ker(X) \subseteq Ker(H)$.*

Dem:
($\Rightarrow$) Si H es estimable, $\exists L \in l(\mathbb{R}^n, \mathbb{R}^k)$ tq $LX = H$; luego si $X\beta = 0 \Rightarrow H\beta = 0$, ie $Ker(X) \subseteq Ker(H)$.

($\Leftarrow$) Si $Ker(X) \subseteq Ker(H)$, se tiene que si $X\beta = 0 \Rightarrow Hb = 0 \Rightarrow \forall L \in M_{k,n} : LX\beta = H\beta = 0$.
Tomar $F \subset \mathbb{R}^{p+1}$ de manera que $\mathbb{R}^{p+1} = Ker(X) \oplus F$, así X es un isomorfismo sobre $W := Im(X) = X(\mathbb{R}^{p+1}) = X(F) \subset \mathbb{R}^n$. Entonces a todo $Y \in W$ corresponde un solo $\beta \in F$ tal que $Y = X\beta$.
Con lo anterior podemos definir L para $Y = Y_X + Y_2 \in \mathbb{R}^n$, con $Y_X \in W$ (único), como $LY = H\beta$, donde β el único elemento asociado a Y_X del cual se habló, luego, $LX\beta = H\beta$. Se deduce entonces que H es estimable.

■

Corolario 3.1 *Del teorema anterior siguen los siguientes resultados.*

1. *Sea* $\hat{\beta}$ *una solución de las ecuaciones normales. Si* H *es estimable, entonces* $H\hat{\beta}$ *no depende de la solución particular* $\hat{\beta}$ *elegida. Se tiene además* $H\hat{\beta}$ *es insesgado para* $H\beta$.

2. *Si se busca un estimador insesgado de* β*, el modelo tiene que ser de rango completo.*

3. *En un modelo de rango completo, toda función vectorial de* X *es estimable.*

Dem:

1. En efecto, $\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1$ en que $\hat{\beta}_0 \in Ker(X)$, y $\hat{\beta}_1 \in F$. Luego $\hat{\beta}_0 \in Ker(H)$, con lo que $H\hat{\beta} = H\hat{\beta}_0 + H\hat{\beta}_1 = H\hat{\beta}_1$ que es invariante ya que la descomposición es única. Además $LX\hat{\beta} = H\hat{\beta}$ no depende de L.

2. En este caso H es la identidad en $\mathbb{R}^{p+1}$, luego $Ker(H) = \{0\} \Rightarrow Ker(X) = \{0\}$, ie X debe tener rango completo.

3. Directo del punto anterior.

■

Aplicación al modelo de rango incompleto.

Sea X de rango r y $X = (X_1|X_2)$ en que X_1 es de rango completo, r, y $X_2 = X_1U$. Sea $\beta = (\beta_1\ \beta_2)^t$ la descomposición de β tal que $X\beta = X_1\beta_1 + X_2\beta_2$. Sea $\beta_1 \in L_1$ de dimensión r y $\beta_2 \in L_2$ de dimensión $p+1-r$, entonces $\beta^+ = \beta_1 + U\beta_2 \in L_1$.

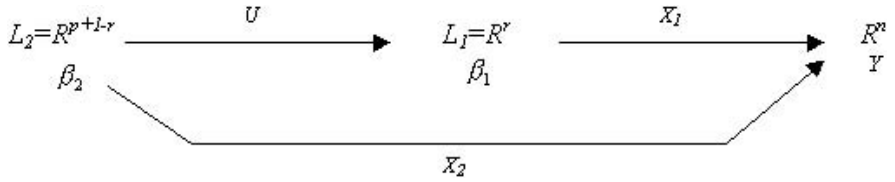


Figura 3.6: Definición de U, X_1, X_2, β_1, β_2.

Teorema 3.4 *Una condición necesaria y suficiente para que* $H \in l(\mathbb{R}^{p+1}, \mathbb{R}^k)$ *sea estimable es que* $H_2 = H_1U$*, en donde* H_1 *y* H_2 *son las restricciones de* H *a* L_1 *y* L_2.

Dem:

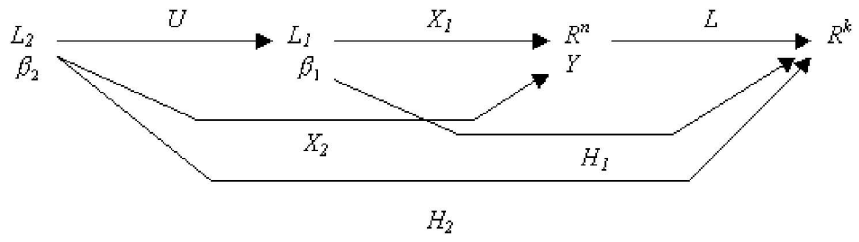


Figura 3.7: Demostración teorema 3.4

Si H es estimable, existe L tal que $LX = H$ y además $LX_1 = H_1$ y $LX_2 = H_2$. Sigue que

$$LX\beta = L(X_1\beta_1 + X_2\beta_2) = LX_1\beta_1 + LX_2\beta_2 = H_1\beta_1 + H_2\beta_2$$

Pero $LX_2 = LX_1U$, por lo tanto $H_2 = H_1U$.

Recíprocamente, si $H_2 = H_1U$, mostramos que se puede construir L tal que $LX = H$. Observemos que basta construir L sobre $Im(X)$.

X_1 es de rango completo, luego $\forall\ Y \in Im(X) \quad \exists!\ \beta_1 \in L_1$ tq $Y = X_1\beta_1$. Definimos L tal que $LY = LX_1b_1 := H_1\beta_1$, con lo que, con notaciones obvias:

$$\forall\beta \in \mathbb{R}^{p+1} : LX\beta = L(X_1\beta_1 + X_2\beta_2) = L(X_1\beta_1 + X_1U\beta_2) = H_1(\beta_1 + U\beta_2) = H\beta$$

Luego H es estimable.

■

El siguiente teorema nos muestra la optimalidad del estimador para el caso en que se tienen funciones estimables, tal como se hizo en el modelo de rango completo donde estudiamos el *Teorema de Gauss-Markov*

Teorema 3.5 *Si H es una función vectorial estimable, el único estimador lineal insesgado de mínima varianza de $H\beta$ es $H\hat{\beta}$ en donde $\hat{\beta}$ es cualquier solución de las ecuaciones normales.*

Dem:

$$\begin{aligned} H\beta &= H_1\beta_1 + H_2\beta_2 = H_1\beta^+ \\ \Rightarrow H\hat{\beta} &= H_1\hat{\beta}^+ \\ \Rightarrow \mathbb{V}(H\hat{\beta}) &= \mathbb{V}(H_1\hat{\beta}^+) = \sigma^2 H_1(X_1^tX_1)^{-1}H_1^t \end{aligned}$$

Y esto no depende de la partición X en X_1 y X_2.

■

Estudio imponiendo restricciones.

Si el rango de X es igual a $r < p+1$, se tiene $p+1-r$ grados de indeterminación en la elección de $\hat{\beta}$. Se puede levantar esta indeterminación imponiendo $p+1-r$ restricciones lineales independientes sobre $\hat{\beta}$, de manera que conociendo $\hat{Y}$, se obtenga un único estimador β^* de β tal que $\hat{Y} = X\beta^*$.

Las restricciones son de la forma $K\beta^* = 0$ con $K \in M_{p+1-r,p+1}$, K es de rango $s = p+1-r$.

Se tiene entonces que estimar β con β^* tal que

$$\hat{Y} = X\beta^* \text{ con la restricción } K\beta^* = 0. \qquad (1)$$

Veamos que esta condición nos asegura de obtener la unicidad con cualquier K de rango s (seguimos usando la misma notación que el punto anterior).

Teorema 3.6 *Considerando K_1 y K_2 las restricciones de K a L_1 y L_2, la condición necesaria y suficiente para que (1) tenga una solución única es que $K_2 - K_1U$ sea invertible.*

Dem:
La solución de (1) puede escribirse usando la partición $X = (X_1, X_2)$, $\hat{\beta} = X_1^t X_1)^{-1} X_1^t Y$:

$$\begin{cases} \hat{\beta}^+ &= \hat{\beta}_1 + U\hat{\beta}_2 \\ 0 &= K_1\hat{\beta}_1 + K_2\hat{\beta}_2 \end{cases} \Rightarrow \begin{cases} \hat{\beta}_1 &= \hat{\beta}^+ - U\hat{\beta}_2 \\ (K_2 - K_1U)\hat{\beta}_2 &= -K_1\hat{\beta}^+ \end{cases}$$

Este último sistema de ecuaciones tiene una solución única si y solo si $K_2 - K_1U$ es invertible.

■

Obs:

- K no puede ser estimable en este caso. Si lo fuera $K_2 = K_1U$ y $\hat{\beta}_2$ no es único.
- Si H es estimable, $H\hat{\beta}$ no depende del estimador $\hat{\beta}$ solución de las ecuaciones normales por lo tanto de las restricciones elegidas.

Dos maneras de encontrar la solución de (2):

1. Como $Kb = 0$, se puede escribir las ecuaciones normales de la forma:

   $$(X^tX + MK)\hat{\beta}_K = X^tY$$

   en donde M es una matriz tal que $X^tX + MK$ invertible. El problema es el de encontrar esta matriz M.

2. La otra manera , más operativa, consiste en construir el modelo aumentado:

   $$\begin{bmatrix} X^tX & K^t \\ K & 0 \end{bmatrix} \begin{pmatrix} \beta \\ 0 \end{pmatrix} = \begin{pmatrix} X^tY \\ 0 \end{pmatrix}.$$

   Si la matriz aumentada $A = \begin{pmatrix} X^tX & K^t \\ K & 0 \end{pmatrix}$ es invertible, al escribir su inversa como $A^{-1} = \begin{pmatrix} C & P^t \\ P & Q \end{pmatrix}$, entonces $\hat{\beta} = CX^tY$.

3.8. Predicción.

Si se tiene una nueva observación para la cual se conocen los valores de las variables explicativas, sean $x_{0,1}, x_{0,2}, \ldots, x_{0,p}$, pero se desconoce el valor Y_0 de la variables respuesta, se puede entonces usar el modelo para inferir un valor para Y_0 a través de su modelo esperado:

$$\mu_0 = \mathbb{E}(y_0) = x_0^t\beta$$

en que $x_0^t = (x_{0,1}\, x_{0,2} \ldots x_{0,p})$.

Si $\hat{\beta}$ es el estimador de β obtenido sobre las antiguas observaciones, se estima μ_0 dados los valores tomados por las variables explicativas por:

$$\hat{\mu}_0 = \mathbb{E}(y_0) = x_0^t\hat{\beta}.$$

Se puede calcular un intervalo de confianza para μ_0: la distribución de $\hat{y}_0$ es $N(\mu_0, \sigma^2 x_0^t(X^tX)^{-1}x_0)$, luego $\dfrac{\hat{y}_0 - \mu_0}{\hat{\sigma}\sqrt{x_0^t(X^tX)^{-1}x_0}} \sim t_{n-p-1}$. Se usa este estadístico para construir un intervalo de confianza de nivel $1-\alpha$ para μ_0:

$$\mathbb{P}\left(\hat{y}_0 - t_{n-p-1}^{\alpha/2}\hat{\sigma}\sqrt{x_0^t(X^tX)^{-1}x_0} \leq \mu_0 \leq \hat{y}_0 + t_{n-p-1}^{\alpha/2}\hat{\sigma}\sqrt{x_0^t(X^tx)^{-1}x_0}\right) = 1-\alpha$$

Un problema distinto es estimar un intervalo para y_0. Hablamos de un intervalo para la predicción. En este caso hay que tomar en cuenta la varianza de la variable aleatoria y_0:

$$y_0 = \hat{y}_0 + \varepsilon_0.$$

La varianza de ε_0 es igual a: $\sigma^2 + \sigma^2 x_0^t (X^t X)^{-1} x_0$. Un intervalo de predicción para y_0 se obtiene entonces a partir de $\dfrac{y_0 - \hat{y}_0}{\hat{\sigma}\sqrt{1 + (x_0^t (X^t X)^{-1} x_0)}} \sim t_{n-p-1}$

El intervalo es entonces definido por:

$$\mathbb{P}\left(\hat{y}_0 - t_{n-p-1}^{\alpha/2}\hat{\sigma}\sqrt{1 + x_0^t (X^t X)^{-1} x_0} \leq y_o \leq \hat{y}_0 + t_{n-p-1}^{\alpha/2}\hat{\sigma}\sqrt{1 + x_0^t (X^t X)^{-1} x_0}\right) = 1 - \alpha.$$

3.9. Intervalos simultanéos y regiones de confianza.

Vimos que los test de hipótesis sobre los parámetros individualmente no son adecuados en general. Por la misma razón, no se construye en general intervalo de confianza para cada parámetro por separado. Se propone construir regiones de confianza o intervalos simultáneos de confianza.

3.9.1. Regiones de confianza.

Vemos aquí intervalos o regiones de confianza para parámetros individualmente o para funciones de los parámetros.

Para cada parámetro β_j del modelo lineal, se puede construir un intervalos de confianza utilizado:

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \sim t_{n-r}$$

en donde $\hat{\sigma}_j^2$ es la estimación de $\mathbb{V}(\hat{\beta}_j) = \sigma^2 (X^t X)_{jj}^{-1}$; es decir $\hat{\sigma}_j^2 (X^t X)_{jj}^{-1}$. El intervalo de confianza de nivel de confianza igual a $1 - \alpha$ es:

$$\left[\hat{\beta}_j - t_{n-r}^{\alpha/2}\hat{\sigma}_j, \hat{\beta}_j + t_{n-r}^{\alpha/2}\hat{\sigma}_j\right]$$

Para una combinación lineal del vector β : $\dfrac{a^t\hat{\beta} - a^t\beta}{\hat{\sigma}\sqrt{a^t (X^t X)^{-1} a}} \sim t_{n-r}$, el intervalo de confianza queda:

$$\left[a^t\hat{\beta} - t_{n-r}^{\alpha/2}\hat{\sigma}\sqrt{a^t (X^t X)^{-1} a}, a^t\hat{\beta} + t_{n-r}^{\alpha/2}\hat{\sigma}\sqrt{a^t (X^t X)^{-1} a}\right]$$

Ejemplo 3.7 Sean $p = 3$, $n = 18$, $(X^t X)^{-1} = \frac{1}{n}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$, $2\hat{\sigma}^2 = n$ y $\hat{\beta} = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix}$.

Las varianzas de β_1 y β_2 son: $\hat{\sigma}_1^2 = 0{,}5$ y $\hat{\sigma}_2^2 = 1$. Los intervalos de confianza individuales con $1 - \alpha = 0{,}95$ para β_1 y β_2 son: $\beta_1 \in [-0{,}13; 4{,}13]$ y $\beta_2 \in [-1{,}13; 3{,}13]$.

Calculamos el intervalo para $\beta_1 - \beta_2$: $a^t = (\ 0 \quad 1 \quad -1\ )$, así $\mathbb{V}(a^t\hat{\beta}) = 3 \Rightarrow \beta_1 - \beta_2 \in [-2{,}691; 4{,}691]$.

Ahora para $\beta_1 + \beta_2$: $a^t = (\ 0 \quad 1 \quad -1\ )$; $\mathbb{V}(a^t\hat{\beta}) = 1 \Rightarrow \beta_1 + \beta_2 \in [0{,}891; 5{,}131]$.

En la figura 3.8a se representó los dos intervalos de confianza individuales para β_1 y β_2 y en la figura 3.8b, las regiones de confianza para $\beta_1 - \beta_2$ y $\beta_1 + \beta_2$.

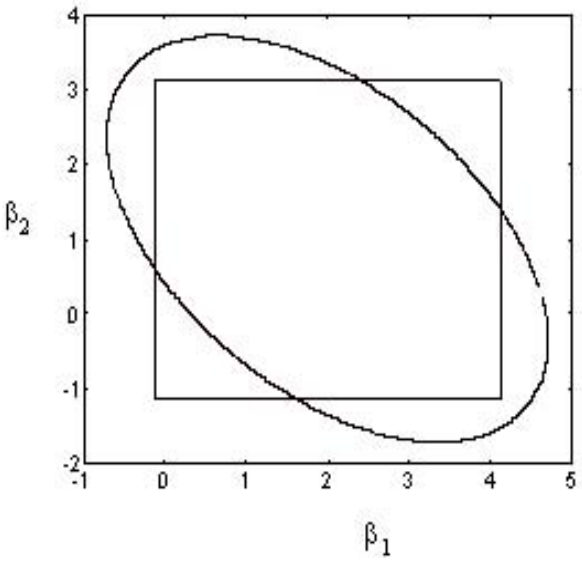
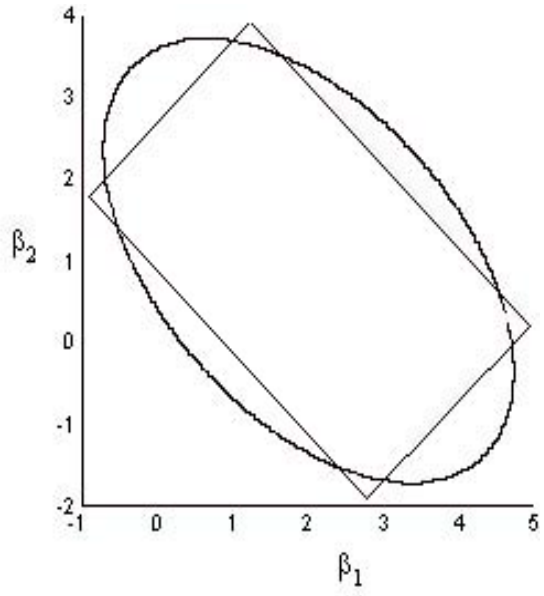


Figura 3.8: (a) Intervalo para β_1 y β_2



(b) Intervalo para $\beta_1 - \beta_2$ y $\beta_1 + \beta_2$

3.10. Ejercicios.

1. Cuatro médicos estudian los factores que explican porque hacen esperar a sus pacientes en la consulta. Toman una muestra de 200 pacientes y consideran el tiempo de espera de cada uno el día de la consulta, la suma de los atrasos de los médicos a la consulta este mismo día, el atraso del paciente a la consulta este día (todos estos tiempos en minutos) y el números de médicos que están al mismo tiempo es la consulta este día. Se encuentra un tiempo promedio de espera de 32 minutos con una desviación típica de 15 minutos. Se estudia el tiempo de espera en función de las otras variables mediante un modelo lineal cuyos resultados están dados a continuación:

Variable	**Coeficiente**	**Desv. típica**	***t*-Student**	$\mathbb{P}(\|X\| > t)$
Constante	22,00	4,42	4,98	0,00
Atraso médico	0,09	0,01	9,00	0,00
Atraso paciente	-0,02	0,05	0,40	0,66
Número de médicos	-1,61	0,82	1,96	0,05
Coef. determinación=0,72	F de Fisher=168	$\mathbb{P}(X > F) = 0,000$		

Cuadro 3.1: Resultados de la regresión

a) Interprete los resultados del modelo lineal. Comente su validez global y la influencia de cada variable sobre el tiempo de espera. Especifique los grados de libertad de las t de Student y la F de Fisher.

b) Muestre que se puede calcular la F de Fisher a partir del coeficiente de determinación. Si se introduce una variable explicativa suplementaria en el modelo, el coeficiente de determinación será más elevado?

c) Dé un intervalo de confianza a 95

d) Predecir el tiempo de espera, con un intervalo de confianza a 95que llega a la hora un día que el consultorio funciona con 4 médicos que tienen respectivamente 10, 30, 0, 60 minutos de atraso.

2. Suponga que tenemos un modelo lineal $Y = X\beta + \varepsilon$ con $\varepsilon \sim N_n(0, \sigma^2 I_n)$, $\beta \in \mathbb{R}^{p+1}$, $X \in M_{n,p+1}(\mathbb{R})$.

   a) Escribamos X como: $X = (X_1, X_2)$, con X_1 y X_2 submatrices de X tales que $X_1^t X_2 = 0$ (la matriz nula). El modelo inicial $Y = X\beta + \varepsilon$ se escribe $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$ con $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$. Si $\hat{\alpha}_1$ es el estimador de máxima verosimilitud de α_1 en el modelo $Y = X_1\alpha_1 + \varepsilon$ y $\hat{\alpha}_2$ es el estimador de máxima verosimilitud de β es igual a $\begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix}$.

   (Indicación: se usará el siguiente resultado: si $A \in M_{n,n}(\mathbb{R})$ es una matriz diagonal por bloque, i.e. $A^{-1} = \begin{bmatrix} A_1^{-1} & 0 \\ 0 & A_2^{-1} \end{bmatrix}$, con las submatrices A_1 y A_2 invertibles , entonces A es invertible, y $A^{-1} = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$).

   b) Si $X_1^t X_2 \neq 0$ y si se toma $\hat{\beta} = \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix}$ como estimador de β, que propiedad pierde $\hat{\beta}$ bajjo el supuesto usual $\mathbb{E}(\varepsilon) = 0$.

3. Consideremos tres variables Y, X, Z observadas sobre una muestra de tamaño $n = 40$, Muestra $= \{(y_i, x_i, z_i) \quad tq \quad i = 1, \ldots, 40\}$. Se busca explicar Y linealmente a partir de X y Z.

   a) Se representan los resultados de modelo lineal: $y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1, \ldots, 40$: Interprete estos

Variable	Medias	Desv. típica
Y	11,68	3,46
X	5,854	2,74

Constante	Estimación	Dev. típica estimación	t-Student	$\mathbb{P}(\lvert X\rvert > t)$
α	7,06	1,03	6,84	0,00
β	0,79	0,16	4,94	0,00
Coef. determinación=0,39		F de Fisher=24,44		$\mathbb{P}(X > F) = 0,000$

Cuadro 3.2: Resultados de la regresión

   resultados y efectúe el test de hipótesis $H_0 : \beta = 0$.

   b) Dé una estimación insesgada para σ^2 la varianza de los errores de este modelo.

   c) Comente el gráfico de los residuos en función de los y_i.

   d) Se tiene una nueva observación que toma sobre la variable X el valor $x_0 = 6,50$. Dé una estimación $\hat{y}_0$ del valor y_0 que toma sobre la variable Y.

   e) Se presentan los resultados del modelo lineal: $y_i = \delta + \gamma z_i + \varepsilon_i$:
   Se tiene $\sum_i x_i z_i = 0$ y $\sum_i z_i = 0$.

   Muestre que si $X_1 = (1_n | X)$ es una matriz formada del vector de unos y del vector de los x_i y $X_2 Z$ el vector formado de los z_i, se tiene $X_1^t X_2 = 0$. Usando los resultados del ejercicio 2 deduzca las estimaciones de los parámetros del modelo $y_i = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$.

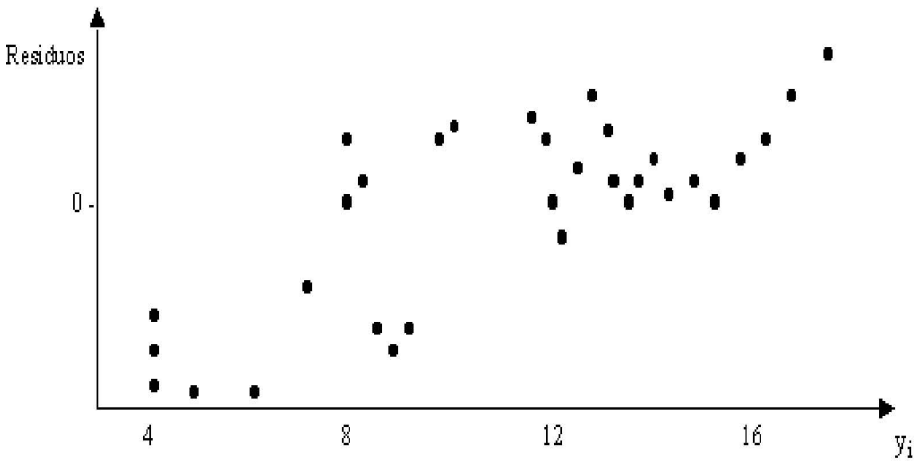
4. Se requiere ajustar una función escalón $y = f(t)$ con f constante en los intervalos en que $j = 0, \ldots, K$ y $a_0 < a_1 < \ldots < a_K$. Para ello se observan datos $\{(t_i, y_i) \quad i = 1, \ldots, n\}$. Se asume que los y_i son mutuamente independientes y que la distribución de los y_i es $N(f(t_i), \sigma^2)$.

   a) Formule el problema anterior como modelo lineal.

Variable	Medias	Desv. típica
Y	11,68	3,46
Z	0,00	2,65

Constante	Estimación	Dev. típica estimación	t-Student	$\mathbb{P}(\|X\| > t)$
δ	11,68	0,36	32,54	0,00
γ	1,00	0,14	7,27	0,00
Coef. determinación=0,58		F de Fisher=52,78		$\mathbb{P}(X > F) = 0,000$

Cuadro 3.3: Resultados de la regresión



b) Obtenga la función ajustada por mínimos cuadrados.

c) Concluya un intervalo de confianza para $\int_{a_0}^{a_K} f(t)dt$.

5. Sea $Y \in \mathbb{R}^n$ un vector aleatorio con $\mathbb{E}(Y) = \mu$ y $\mathbb{V}(Y) = \sigma^2 I_n$. Se considera el modelo lineal $Y = X\beta + \varepsilon$, en que $X \in M_{n,p}$ es de rango completo. Llamaremos W al subespacio de $\mathbb{R}^n$ conjunto imagen de X e $\hat{Y}$ al estimador de mínimos cuadrados de $\mu = \mathbb{E}(Y)$.

   a) Sea $a \in W$ y Δ_a la recta generada por a. Se define $H_0 = \{z \in W \quad tq \quad a^t z = 0\}$ el suplemento ortogonal de Δ_a en W. Se tiene entonces la descomposición en suma directa ortogonal de W: $W = H_a \oplus \Delta_a$. Muestre que el estimador de mínimos cuadrados Y^* de μ en H_a se escribe como: $Y^* = \hat{Y} - \left(\frac{a^t \hat{Y}}{a^t a}\right) a$.

   b) Si $b \in \mathbb{R}^n$, muestre que $Var(b^t Y^*) = Var(b^t \hat{Y}) - \sigma^2 \frac{(b^t b)^2}{a^t a}$.

   c) Suponiendo que los errores son normales, dé la distribución de $\frac{\sum_i \varepsilon_i^{*2}}{\sigma^2}$, en que $\varepsilon_i^* = Y_i - Y_i^*$.

   d) Se considera el caso particular $a = I_n$. Dé la distribbución de $\frac{\sum_i Y_i^{*2}/p}{\sum_i \varepsilon_i^{*2}/(n-p)}$. Muestre que si las variables son centradas, $\hat{Y} = Y^*$.

6. **Teorema de Gauss-Markov generalizado**

Si $\mathbb{V}(Y) = \Gamma$, Γ invertible, entonces el estimador $\hat{\beta}$ insesgado de mínima varianza entre los estimadores lineales insesgados de β es aquel que minimiza $\|Y - X\beta\|^2_{\Gamma^{-1}}$.

a) Encuentre el estimador de máxima verosimilitud de β y Γ.

b) Demuestre el teorema.

c) Si el rango de X es igual a r, muestre que la norma del vector de residuos de un modelo lineal

$$\|Y - \hat{Y}\|^2_{\Gamma^{-1}} \sim \chi^2_{n-r}$$

en donde $\hat{Y}$ la proyección Γ^{-1}-ortogonal de Y sobre $Im(X)$.

7. Sea el modelo lineal: $y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i$, $i = 1, 2, \ldots, n$. Matricialmente $Y = X\beta + \varepsilon$, con $rng(X) = p + 1$, $\mathbb{E}(\varepsilon) = 0$, $\mathbb{V}(\varepsilon) = \sigma^2 I_n$.

   a) Se escribe $X^t X = \begin{bmatrix} n & a^t \\ a & V \end{bmatrix}$. Dé las expresiones de a y V. Muestre que V es definida positiva. Muestre que a es un vector nulo cuando las variables explicativas están centradas $\left(\forall j : \sum_{i=1}^{n} x_{i,j} = 0 \right)$. Relacione los valores propios de V con los de V^{-1}.

   b) Muestre que $\sum_j \mathbb{V}(\hat{\beta}_j)$ sujeto a $\forall j : \sum_{i=1}^{n} x_{i,j} = 0$ y $\forall j : \sum_{i=1}^{n} x_{i,j}^2 = c$ (c es una constante positiva) alcanza su mínimo cuando $X^t X$ es diagonal.

   c) En qué difieren de las propiedades optimales obtenidas en el teorema de Gauss-Markov?

   d) Se supone que $X^t X$ es diagonal con $\forall j : \sum_{i=1}^{n} x_{i,j} = 0$ y $\forall j : \sum_{i=1}^{n} x_{i,j}^2 = c$. Deducir las expresiones de $\hat{\beta}$, $\mathbb{V}(\hat{\beta})$, $\hat{Y}$. Exprese el coeficiente de correlación múltiple R^2 en función de los coeficientes de correlación lineal de Y con las variables explicativas X.

8. Sea el modelo lineal $Y = X\beta + \varepsilon$, con X de rango completo pero $X^t X$ no diagonal.

   a) Dé la expresión de una predicción de la variable respuesta Y y un intervalo de confianza asociado.

   b) Se hace un cambio de base de las columnas de X, sea Z la matriz de las nuevas columnas, de manera que $Im(X) = Im(Z)$ y que $Z^t Z$ sea diagonal. Muestre que el cambio de variables explicativas no cambia las predicciones de Y. Deduzca la expresión del intervalo de confianza en función de Z.

9. Concluye el test de razón de verosimilitudes para la hipótesis nula $H_0 : A\beta = c$ para los supuestos usuales. Muestre que es equivalente al test F de Fisher dado en 2.5.5.

10. Sea el modelo lineal $Y = X\beta + \varepsilon$ con $\varepsilon \sim N_n(0, \sigma^2 I_n)$, $X \in \mathcal{M}_{np}$ de rango incompleto. Sea $A\beta$ una función vectorial estimable de β, con $A \in \mathcal{M}_{sp}$ de rango completo s. Muestre que $A(X^t X)^- A^t$ es invertible.