Capítulo 3

TEORÍA GENERAL DEL MODELO LINEAL

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: disminuye en general esta incertidumbre, ya que la distribución de Y dado $X = x$ tiene una varianza que en promedio es menor que la varianza marginal de Y:

$$Var(Y) = E_X\{Var(Y|X)\} + Var_X\{E(Y|X)\}$$

Demostración: Observamos en primer lugar que $E_X(E(Y|X)) = E(Y)$.

Consideramos

$$Var(Y) = \int (y - E(Y))^2 d\mathbb{P}^Y(y)$$

Como $y - E(Y) = y - E(Y|X) + E(Y|X) - E(Y)$, se tiene que

$$(y - E(X))^2 = (y - E(Y|X))^2 + (E(Y|X) - E(Y))^2 + 2(y - E(Y|X))(E(Y|X) - E(Y))$$

Además $d\mathbb{P}(x,y) = d\mathbb{P}^{Y|X}(y)d\mathbb{P}^X(x)$, en donde $\mathbb{P}^{Y|X}$ es la distribución condicional de Y dado X y $\mathbb{P}^X$ es la distribución marginal de X. Luego

$$Var(Y) = \int \left[\int (y - E(Y|x))^2 d\mathbb{P}^{Y|X}(y) \right] d\mathbb{P}^X(x) + \int \left[\int (E(Y|x) - E(Y))^2 d\mathbb{P}^{Y|X}(y) \right] d\mathbb{P}^X(x)$$

$$+2 \int (E(Y|x) - E(Y)) \left[\int (y - E(Y|x)) d\mathbb{P}^{Y|X}(y) \right] d\mathbb{P}^X(x)$$

Pero por definiciones y algunos desarrollos vemos que:

$$E_X(Var(Y|X)) = \int \left[\int (y - E(Y|x))^2 d\mathbb{P}^{Y|X}(y) \right] d\mathbb{P}^X(x)$$

$$Var_X(E(Y|X)) = \int \left[\int (E(Y|x) - E(Y))^2 d\mathbb{P}^{Y|X}(y) \right] d\mathbb{P}^X(x)$$

$$\int (y - E(Y|x)) d\mathbb{P}^{Y|X}(y) = 0$$

Luego $Var(Y) = E_X\{Var(Y|X)\} + Var_X\{E(Y|X)\}$.

Se deduce que $E_X\{Var(Y|X))\} \leq Var(Y)$. Es un resultado promedio, eso no impide que para algunos valores de X, $Var(Y|X)$ sea el mayor que $Var(Y)$.

Cuando se puede aceptar que el fenómeno aleatorio representado por una variable o un vector X puede servir para predecir aquel representado por Y, hay que buscar 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 aquí descubrir como un conjunto de variables $X_1, X_2, \cdot X_p$ 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 $\{X_1, X_2, \cdot X_p\}$ se llaman **variables explicativas o variables independientes o variables exógenas** y la variables Y se llama **variable a explicar o variable respuesta o variables dependiente o variable endógena**.

Daremos 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 partítula 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 (ruido blanco") es de tipo aleatorio. En ese caso no basta tener dos ecuaciones sino valores de la distancia para varios valores del tiempo. 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 ditribución conjunta normal bivariada:

$$P = f(T) + \varepsilon$$

$$\text{con} \quad f(T) = E(P|T)$$

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 no es evidente.

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.

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 polinómial. 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: $Y = a_0 + a_1X^2 + \cdots + a_pX^p$ en donde X^j es la potencia j de X.

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: $d = \frac{1}{2}gt^2$.

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 presentar después el caso lineal sobre valores muestrales.

Se usaran dos métodos de estimación:

- 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 cuadrádos y discutir las propiedades de los estimadores y la precisión del ajuste.

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

3.2. Modelo teórico condicional.

Proposición 3.1 Sean la v.a. $Y \in I\!R$ y el vector aleatorio $X \in I\!R^p$. El mínimo de $E\{(Y - f(X))^2\}$ se alcanza en $f(X) = E(Y|X)$.

Demostración: Geométricamente, en el espacio de Hilbert L^2_{p+1} de dimensión $p+1$ tomando como producto escalar

$$< U, V >= E(UV^t)$$

$E(Y|X)$ es la proyección ortogonal de Y sobre el subespacio L_X^2 generado por las funciones de X.

El criterio para minimizar es el **error cuadrático medio**

$$E\{(Y-g(X))^2\}$$

Si $f(X)=E(Y|X)$, entonces para toda función $g(X)$, se tiene:

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

En efecto

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

dado que $f(X)-g(X)$ es independiente de Y y $E\{(Y-f(X)|X)=0$ se obtiene el resultado.

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

$$\eta^2_{Y|X}=Cor^2(Y,E(Y|X))=\frac{Var\{E(Y|X)\}}{Var(Y)}=1-\frac{Var(\varepsilon)}{Var(Y)}$$

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

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

En efecto:

$$Cov(Y,E(Y|X))=E[(y-E(Y))(E(Y|X)-E(Y))]$$

Como $E(\varphi(X,Y)))=E_X\{E_{Y|X}(\varphi(X,Y)|X)\}$

$$Cov(Y,E(Y|X))=E_XE\{(Y-E(Y))(E(Y|X)-E(Y))|X\}$$

Ahora bien:

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

y

$$Cov(Y,E(Y|X))=E_X\{(E(Y|X)-E(Y))^2\}=Var(E(Y|X))$$

Finalmente

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

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

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

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

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 IR^{p+1}. Se plante el modelo lineal:

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

Consideraremos aquí el vector X como no aleatorio.

Denotamos $\forall i = 1,2,\ldots,n : x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})$ y hacemos los siguientes supuestos sobre los errores: $\varepsilon_i = y_i - E(Y) \sim N(0, \sigma^2)$, 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. Dos tipos de métodos de estimación se pueden usar aquí: 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. Escribiremos matricialmente el modelo aplicado a la muestra de observaciones.

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

Entonces el modelo se escribe

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

El criterio de los mínimos cuadrados consiste en buscar el punto del subespacio vectorial $W = Im(X)$ de IR^n generado por las columnas de la matriz X más cercano al punto Y. La solución es la proyección ortogonal

del punto Y sobre W y esta es obtiene de las **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 lineales independientes, es decir que forman una base del subespacio vectorial de W, o sea que la dimensión del rango de X es 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 sobre W, que es un operador lineal idempotente de orden 2 y simétrico, 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. El método no permite estimar a σ^2.

3.3.2. Solución de máxima verosimilitud

En el párrafo anterior, para estimar los coeficientes β_j se usó un criterio ma temático que permite ajustar un hiperplano afin de IR^{p+2} . Aquí usaremos el método de máxima verosimilitud para estimarlos. El modelo probabilístico se basa en los errores. El modelo

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

con $Y = E(Y) + \varepsilon = X\beta + \varepsilon$ en donde se supone $\varepsilon \sim N_n(0, \sigma^2 I_n)$. La función de verosimilitud utilizada es la densidad conjunta de los errores:

$$f(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\varepsilon^t\varepsilon\right)$$

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

Calculemos el estimador de máxima verosimilitud de β: $\dfrac{\partial \ln f}{\partial \beta} = 0 \Rightarrow \dfrac{\partial (Y - X\beta)^t (Y - X\beta)}{\partial \beta} = 0 \Rightarrow (X^t X)\hat{\beta} = X^T Y$ (Ecuaciones normales).

Calculemos el estimador de máxima verosimilitud de σ^2:

$$\frac{\partial \ln f}{\partial \sigma^2} = 0 \Rightarrow \sigma^2 = \frac{(Y - X\hat{\beta})^t (Y - X\hat{\beta})}{n}$$

y si $\hat{\varepsilon} = Y - X\hat{\beta}$, entonces

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

Es decir que la función de verosimilitud es máxima cuando se cumplen las ecuaciones normales: $(X^t X)\hat{\beta} = X^t Y$ y además $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2$ llamado la varianza residual dado que es la varianza empírica de los $\hat{\in}_i$; en efecto ya que $Y = X\hat{\beta} + \hat{\varepsilon}, \quad \hat{\varepsilon} \in Im(X)$ y

$$X\hat{\beta} \in (Im(X))^{\perp} \Rightarrow \hat{\varepsilon} \perp 1_n \Rightarrow \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = 0$$

El estimador de los mínimos cuadrados es igual entonces al estimador de máxima verosimilitud cuando se tiene el supuesto de normalidad $\varepsilon \sim N_n(0, \sigma^2 I_n)$.

3.4. Propiedades del estimador

Las propiedades del estimador $\hat{\beta}$ solución de las ecuaciones normales están ligadas a los supuestos hechos sobre los errores ε_i. Supondremos aquí que X es de rango completo $(p+1)$.

Propiedades:

- El estimador $\hat{\beta}$ es un estimador insesgado de β:

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

- El estimador $\hat{Y} = PY = X\hat{\beta}$ es un estimador insesgado de $X\beta$.
- $\varepsilon \sim N_n(0, \sigma^2 I_n) \Rightarrow \hat{Y} \sim N_n(X\beta, \sigma^2 P)$ donde P era el proyector ortogonal sobre W en IR^n.
- $\varepsilon \sim N_n(0, \sigma^2 I_n) \Rightarrow \hat{\varepsilon} \sim N_n(o, \sigma^2(I_n - P))$, con $\hat{\varepsilon}$ ortogonal a W o $\hat{\varepsilon}$ independiente de X.
- $\hat{\beta} \sim N_{p+1}(\beta, \sigma^2 (X^t X)^{-1})$.
- $\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2$ es un estimador sesgado para σ^2. En efecto:

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

luego $\tilde{\sigma}^2 = \dfrac{1}{n-p-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ es un estimador insesgado para σ^2. En efecto,

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

Luego

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

y

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

- $\hat{\beta}$ es independiente de $\sum_{i=1}^{n}\hat{\varepsilon}_i^2$.
- $\hat{\beta}$ 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.
- El estimador es óptimo con respecto a la varianza (ver el teorema de Gauss Markov a continuación).

Consideremos la siguiente definición:

Definición 3.1 Sean $A, B \in M_n(IR)$. 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 $E(\varepsilon) = 0$ y $E(\varepsilon\varepsilon^T) = \sigma^2 I_n$, entonces $\hat{\beta}$ tiene varianza 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 los estimadores insesgados de β.

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

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

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

$$Var(\beta^*) = Var(\hat{\beta}) + Var(DY) + 2Cov(\hat{\beta}, DY)$$

en donde

$$Cov(\hat{\beta}, DY) = E((\hat{\beta} - \beta)(DY)^t) = E(\hat{\beta}Y^tD^t) = E((X^tX)^{-1}X^tYY^tD^t) =$$

$$(X^tX)^{-1}X^tE(YY^t)D^t = (X^tX)^{-1}[Var(Y) + E(Y)E(Y)^t]D^t = 0$$

Finalmente $Var(\beta^*) = Var(\hat{\beta}) + \sigma^2 DD^t$ 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 la muestra multivariada para el parámetro β es igual a

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

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

Se obtiene facílmente una generalización de este teorema cuando $Var(\varepsilon) = \Gamma$, Γ que supondremos invertible. El estimador de mínima varianza es entonces:

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

Es decir que estamos proyectando ortogonalmente en el sentido de la métrica Γ^{-1}.

3.5. Calidad del modelo

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

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. Un índice que evita este problema está dado por:

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

que representa el cuadrado del coseno del ángulo del vector Y con el vector $\hat{Y}$ en IR^n (Figura **??**).

Se puede comparar las siguientes varianzas:

- 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: $\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$.

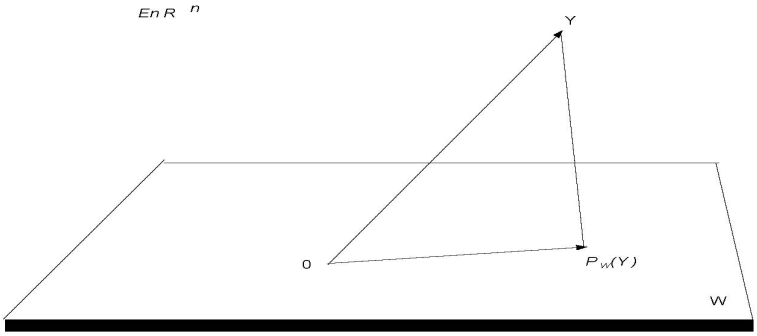


Figura 3.1: Proyección del vector Y en W

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

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

que compara la varianza explicada por el modelo con la varianza total. El coeficiente de correlación múltiple R es el coeficiente de correlación lineal entre Y e $\hat{Y}$. El valor de R está comprendido entre 0 y 1. Cuando $R=0$, el modelo obtenido es $\forall i : \hat{y}_i = \bar{y}$, 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, es decir que existe un modelo lineal que permite escribir las observaciones y_i exactamente como combinación de las variables explicativas. Cuando R es cercano a 1, el modelo es bueno siendo que los valores estimados $\hat{y}_i$ ajustan bien los valores observados y_i.

Para el caso general se tiene:

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

en donde 1_n es el valor de la bisectriz de IR^n de componentes todas iguales a 1.

Si 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 significa que los valores de las p variables explicativas no influyen en los valores de Y. Como $\hat{\varepsilon} \sim N_n(0, \sigma^2(I_n - P))$

e $\hat{Y} \sim N_n(X\beta, \sigma^2 P)$, si r es el rango de la matriz X, se tiene:

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

Como $\hat{Y}|_{H_0} \sim N_n(\beta_1 1_n, \sigma^2 P) \iff \hat{\beta}_0 = \bar{y}$, se tiene:

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

Además $\frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sigma^2}$ 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:

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

en donde $F_{r-1,n-r}$ sigue una distribución de Fisher a $r-1$ y $n-r$ grados de libertad. Se puede expresar F en función del coeficiente de correlación múltiple R:

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

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 está definida por

$$I\!P(F_{r-1,n-r} > c_\alpha) = \alpha.$$

Se rechaza H_0, por lo tanto se declara el modelo globalmente significativo cuando se encuentra un valor F en la muestra mayor que c_α.

En la práctica, se define la **probabilidad crítica** o ***p*-valor** que es el valor p_c tal que $I\!P(F_{r-1,n-r} > F) = p_c$. Si el valor de la probabilidad crítica p_c es alta, no se rechaza H_0, es decir que se declara el modelo como poco significativo.

3.5.2. Medición del efecto de cada variable en el modelo

Cuando las variables explicativas son independientes, el efecto asociado a l variable X_j se mide con $X_j\hat{\beta}_j$. Se observará que el modelo lineal es invariante por el cambio de escalas de mediciones.

Consideramos la hipótesis nula $H_0 : \beta_j = 0$. Como $\hat{\beta}_j \sim N(\beta_j, \sigma_j^2)$ en donde $\sigma_j^2 = Var(\hat{\beta}_j)$ ($\sigma_j^2 = \sigma^2 (X^t X)^{-1}_{j,j}$ en el caso del modelo con rango completo), $\frac{\hat{\beta}_j - \beta_j}{\sigma_j} \sim N(0,1)$. Por otra parte, como $\frac{(n-r)\hat{\sigma}}{\sigma^2} \sim \chi^2_{n-r}$, se deduce que

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

Bajo la hipótesis nula $H_0 : \beta_j = 0$,

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

Si la probabilidad crítica o P-valor $\mathbb{P}\left(|t_{n-r}| > \frac{\hat{\beta}_j}{\hat{\sigma}_j}\right) = p_c$ 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**.

3.5.3. Coeficiente de correlación parcial

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

$$X = \alpha Z + \vartheta$$

$$Y = \gamma Z + \eta.$$

Una vez eliminada la influencia de la variable Z sobre las variables X e Y se mide solamente a partir de los restos:

$$X - \alpha Z = \vartheta$$

$$Y - \gamma Z = \eta.$$

Definición 3.2 El coeficiente de correlación parcial entre X e Y bajo Z constante es el coeficiente de correlación entre los errores ϑ y γ:

$$\rho(X, Y|Z) = Corr(\vartheta, \gamma)$$

Se observa que si X y Z son muy correlacionados entonces la correlación parcial entre X e Y es muy pequeña. En efecto X aporta casi ninguna información nueva sobre Y (o vice-versa) cuando Z es conocida.

Se usa el gráfico de los errores para medir los efectos y el tipo de efecto (lineal o no). Del gráfico3.2(a) podemos decir que la variable X_2 no tiene efecto sobre la variable Y en presencia de la variable X_1. Pero en

el gráfico3.2(b) la variable X_2 aporta a la explicación de la variable Y aún si la variable X_1 es presente en el modelo.

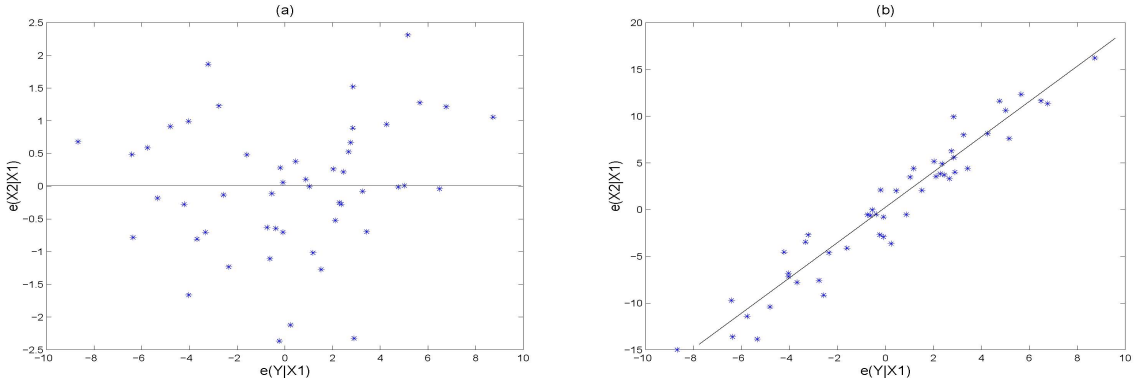


Figura 3.2: Interpretación de los errores y del coeficiente de correlación parcial

Se puede generalizar a más de 2 variables Z_j, $j = 1,2,\ldots,q$. Si

$$X = \sum_{j=1}^{q} \alpha_j Z_j + \vartheta \quad Y = \sum_{j=1}^{q} Z_j + \gamma$$

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

$$\rho(X,Y|Z_1,Z_2,\ldots,Z_q) = Corr(\vartheta,\gamma).$$

Si las variables Z_j no tienen efecto sobre X e Y, es decir que las correlaciones $Corr(X,Z_j)$ y $Corr(Y,Z_j)$ son todas nulas, entonces $\rho(X,Y|Z_1,Z_2,\ldots,Z_q) = Corr(X,Y)$.

Se generaliza también la matriz de correlación parcial con mas de dos variables. Definimos para eso la matriz de varianza-covarianza del vector X dado el vector Z fijo:

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

Se tiene una intertretación geométrica del coeficiente parcial $\rho(X,Y|Z)$ mediante los triángulos esféricos: El ángulo (A) del triángulo esférico(ABC) está definido por el ángulo entre las dos tangentes en A a los lados del triángulo esférico (Gráfico3.3). El ángulo (A) es entonces igual a la proyección del ángulo entre OX y OY sobre el plano ortogonal a OZ. Los ángulos siendo relacionados a los arcos, se tiene:

$$\cos(A) = \frac{\cos(a) - \cos(b)\cos(c)}{\sin(b)\sin(c)}.$$

Luego:

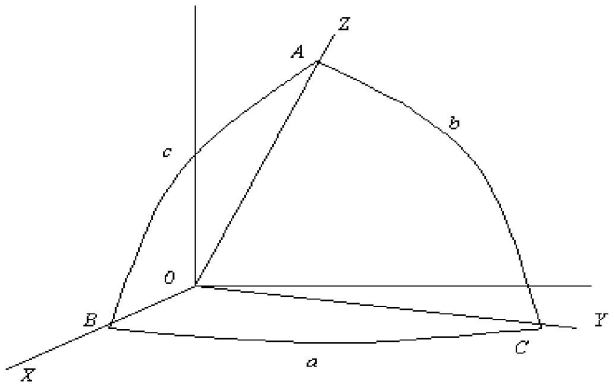


Figura 3.3: Representación esférica del coeficiente de correlación parcial

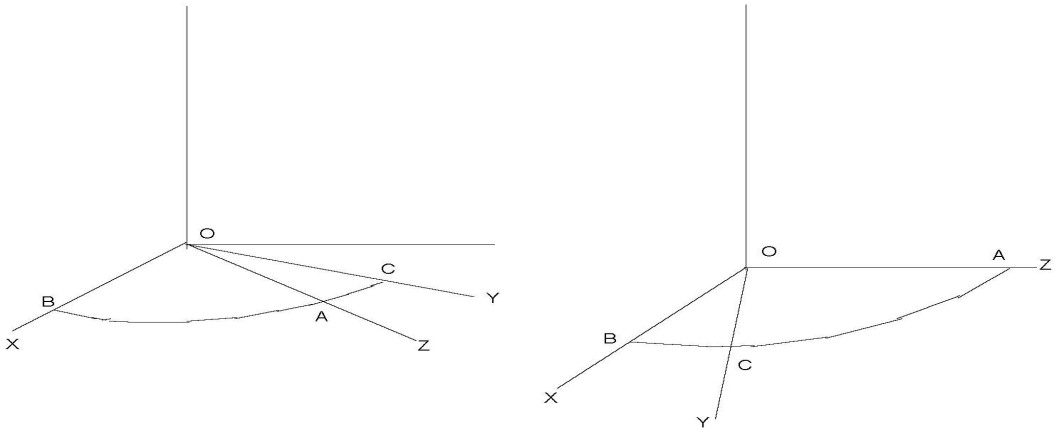


Figura 3.4: (a) $\rho(X,Y|Z) = -1$ (b) $Corr(X,Z) = 0, Corr(Y,Z) = 0,01, \rho(X,Y|Z) = 1$.

$$\boxed{\rho(X,Y|Z) = \frac{Corr(X,Y) - Corr(X,Z)Corr(Y,Z)}{\sqrt{1-Corr^2(X,Z)}\sqrt{1-Corr^2(Y,Z)}}}$$

En la figura3.4a, el ángulo $A = \frac{\pi}{2}$, el coeficiente de correlación parcial es -1. Pero puede haber un efecto escondido de la variable X sobre la variable Z como se ilustra en la figura3.4b: el coeficiente de correlación múltiple de X e Y sobre Z es igual a 1, a pesar que el coeficiente de correlación entre X y Z es nulo y el coeficiente de correlación entre Y y Z es muy pequeño aquel entre X e Y cercano a 1. El coeficiente de correlación parcial es igual a 1 también.

3.5.4. Efecto de un grupo de variables

Vimos que el efecto global de todas las variables explicativas y los efectos individuales. Veremos aquí 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 la hipótesis nula $H_0 : \beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_k} = 0$ contra $H_1 : E(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.

Sean $X_{j_{k+1}}, X_{j_{k+2}}, \ldots, X_{j_p}$ el restante de las P variables. Bajo H_0, el modelo se escribe: $Y = \gamma_0 + \gamma_{j_{k+1}} X_{j_{k+1}} + \cdots + \gamma_{j_p} X_{j_p} + \varepsilon_o$. Se tiene la varianza residual bajo H_1 menor que la varianza residual bajo H_0:

$$\sum_i \hat{\varepsilon}_i^2 \leq \sum_i \hat{\varepsilon}_{oi}^2$$

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

Bajo la hipótesis H_0

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

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

Considerando otra forma de escribir el problema. Sea la hipótesis nula $H_0 : E(Y) = X_0\beta \in W_0$, con X_0 de rango s, contra $H_1 = X\beta \in W$.

La hipótesis H_0 equivale a $(X - X_0)\beta = 0$ lo que corresponde a $k = p - s + 1$ ecuaciones independientes $\underbrace{D}_{k\times(p+1)} \beta = 0$, en que D es de rango k. Para que el test tenga sentido, $D\beta$ tiene que ser estimable, es decir que el estimador $D\beta$ no debe depender de una solución particular $\hat{\beta}$ de las ecuaciones normales.

Sean $\hat{Y}$ e Y^* las proyecciones Y sobre W y W_0 respectivamente y $E(Y) = \mu_0$ bajo H_0 y $E(Y) = \mu$ bajo H_1.

$$\|Y - \mu_0\|^2 = \|Y - Y^* + Y^* - \mu_0\|^2 = \|Y - Y^*\|^2 + \|Y^* - \mu_0\|^2$$

$$\|Y - \mu\|^2 = \|Y - \hat{Y}\|^2 + \|\hat{Y} - \mu\|^2$$

Sean $S^2 = \dfrac{\|Y - Y^*\|^2}{\|Y - \hat{Y}\|^2}$ y $R^2 = \dfrac{\|\hat{Y} - Y^*\|^2}{\|Y - \hat{Y}\|^2}$. Bajo H_0, se tiene $\dfrac{n-p-1}{k} R^2 \sim F_{k,n-r}$. La región crítica es de la forma $\dfrac{n-r}{k} R^2 > C$.

Se puede plantear el test de razón de verosimilitudes también: $\Lambda = \dfrac{\max_{H_0} L}{\max L}$. La región crítica se escribe $S > C'$ Este test coincide con el test F.

Se observará que $\dfrac{\|Y - Y^*\|^2}{n-s}$ y $\dfrac{\|\hat{Y} - Y^*\|^2}{k}$ son ambos estimadores insesgados de σ^2 bajo H_0.

Cuando la varianza σ^2 es conocida, la razón de verosimilitudes es igual a:

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

La región crítica del test se escribe entonces $\|\hat{Y} - Y^*\|^2 > \sigma^2\chi^2_k$.

Se puede construir un test a partir de $D\hat{\beta} \sim N(D\beta, \sigma^2\Gamma)$ cuando $\varepsilon \sim \mathcal{N}(0, \sigma^2\Gamma)$. Bajo H_0, $\dfrac{\hat{\beta}^t D^t \Gamma^{-1} D\hat{\beta}}{\sigma^2} \sim \chi^2_k$. Pero este test no equivale en general al test de razón de verosimilitudes basado en $\|\hat{Y} - Y^*\|^2$.

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. $A\beta$ tiene que ser estimable, es decir no debe depender de una solución de las ecuaciones normales. Se supondrá aquí un modelo de rango completo.

Sea $\hat{\beta} = (X^tX)^{-1}X^tY$ 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:

$$Q = (Y - X\beta)^t(Y - X\beta) + 2\lambda(A\beta - c)$$

$$\frac{\partial Q}{\partial \beta} = 0 \Rightarrow X^tX\hat{\beta}_0 = X^tY + A^t\Lambda \Rightarrow \hat{\beta}_0 = (X^tX)^{-1}(X^tY + A^t\lambda) = \hat{\beta} + (X^tX)^{-1}A^t\lambda.$$

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

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

Sean P_0 y P los proyectores asociados respectivamente a $X\hat{\beta}_0$ y $X\hat{\beta}$, es decir tales que $P_0Y = X\hat{\beta}_0$ y $Py = X\hat{\beta}$. Entonces

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

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 $T \geq V$,consideramos $U = T - V$ que compararemos a V.

Proposición 3.2 La diferencia de las varianzas residuales con y sin restricción es:

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

y bajo la hipótesis nula $\dfrac{U}{\sigma^2} \sim \chi^2_k$.

Demostración:

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

Como $P_0Y = PY + X(X^tX)^{-1}A^t[A(X^tX)^{-1}A^t]^{-1}(c - A\hat{\beta})$ y $U = Y^t(P - P_0)^t(P - P_0)Y \Rightarrow U = (A\hat{\beta} - c)^t[A(X^tX)^{-1}A^t]^{-1}(A\hat{\beta} -$

c).

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 $V = \sum_i \hat{\varepsilon}_i^2$ (ver ejercicio 1.7b), el estadístico del test es:

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

3.5.6. 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

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

considerándolos 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 (3.5) 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 tres partes: la normalidad, la independencia y la igualdad de las varianzas de los errores.

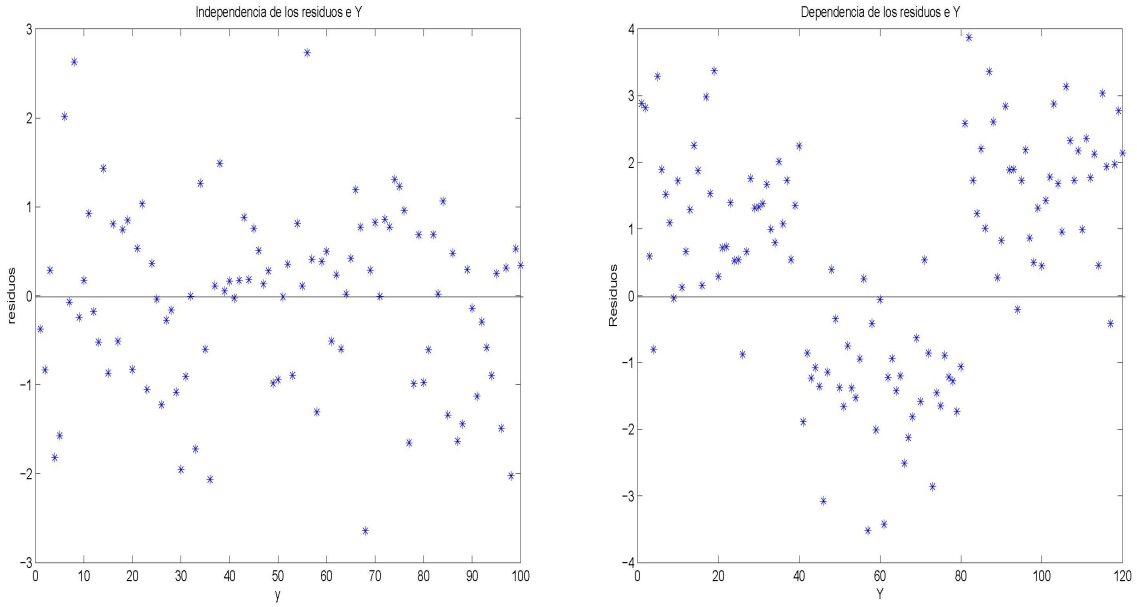


Figura 3.5: Gráficos de residuos

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 $u_i = \frac{\varepsilon_i}{\sigma}$, $H_o : u_i \sim N(0,1)$. Esto sugiere de 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 3.6).

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 basada o bien otras alternativas según la causa de la no normalidad (no simetría, observaciones atípicas, etc..

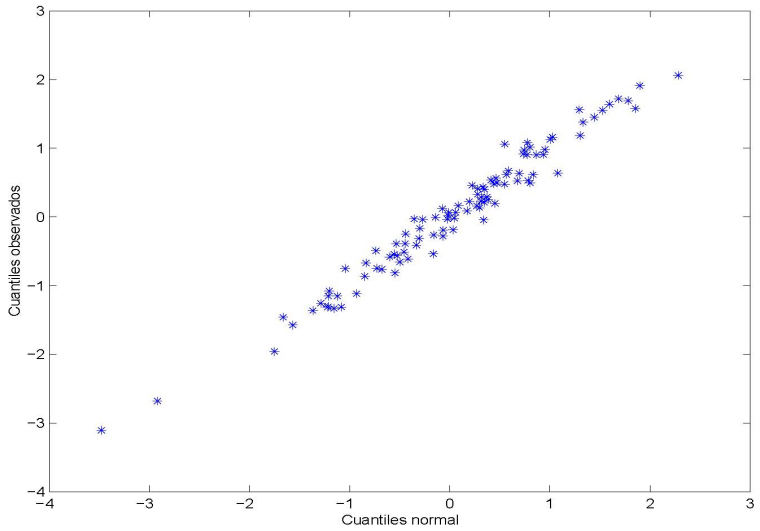


Figura 3.6: Recta de Henri

Test de Durbin y Watson

Test para la igualdad de las varianzas

3.6. 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:

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

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:

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

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

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

Un problema distinto es de estimar un intervalo para y_0. Hablamos de un intervalo para la predicción. En

este caso hay que tomar en cuenta de la varianza aleatoria y_0:

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

La varianza de $\hat{\varepsilon}_0$ es igual a: $\sigma^2 + \hat{\sigma}^2 x_0^t (X^t X)^{-1} x_0$, dado que $\hat{y}_0$. 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:

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

3.7. 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 $A = (X^t X)^-$ no se obtiene necesariamente un estimador insesgado de β. En efecto, si b es una solución de las ecuaciones normales:

$$(X^t X)b = X^t Y$$

entonces $b = AX^t Y \Rightarrow E(b) = AX^t E(Y) = AX^t X\beta$. Si $H = AX^t X$, entonces $E(b) = H\beta \Rightarrow H\beta \neq \beta$ en general: b es un estimador insesgado de $H\beta$ y no de β.

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

Veremos las relaciones que existen entre estos métodos.

3.7.1. El modelo reducido.

Sea X de rango r ($< p+1$), entonces $U \in M_{r,p+1-r}$ tal que $X = (X_1|X_2)$ con X_1 de rango completo r

$X_2 = X_1 U$. Entonces, si $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$:

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

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:

$$\begin{cases} \hat{\beta}^+ = \hat{\beta}_1 + U\hat{\beta}_2 = (X_1^t X_1)^{-1} X_1^t Y \\ E(\hat{\beta}^+) = \beta^+ \\ Var(\beta^+) = \sigma^2 (X_1^t X_1)^{-1} \end{cases}$$

3.7.2. Funciones vectoriales estimables

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

Cuando $G = IR$, se habla de función estimable.

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 $L \in l(IR^n, IR^k)$ (o $L \in M_{k,n}$) tal que $LX = H$.

Demostración: Si H es estimable $\iff \exists L \in l(IR^n, IR^k)$ tal que $E(LY) = H\beta \iff E(L(X\beta+\varepsilon)) = H\beta \iff LX = H$.

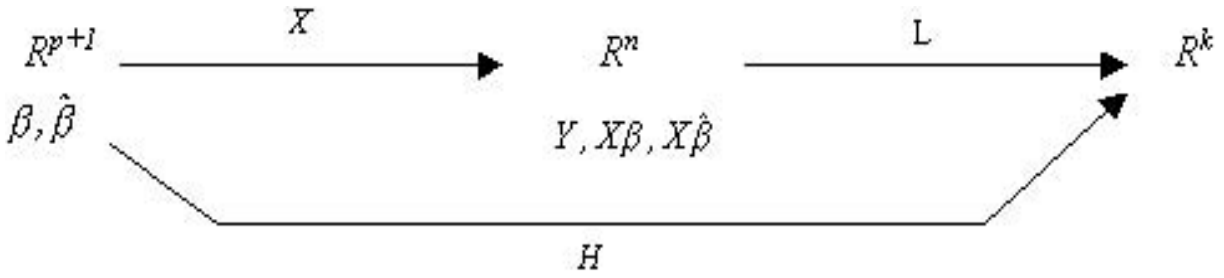


Figura 3.7: Esquema de funciones

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

Demostración: ($\Rightarrow$) Si H es estimable $\Rightarrow \exists L \in l(IR^n, IR^k) \quad tq \quad LX = H$; si $Xb = 0 \Rightarrow Hb = 0$, luego $Ker(X) \subseteq Ker(H)$.

($\Leftarrow$) Si $Ker(X) \subseteq Ker(H)$, luego si $Xb = 0 \Rightarrow Hb = 0 \Rightarrow \forall L \in M_{k,n} : LXb = Hb = 0$. Sea $IR^{p+1} = Ker(X) \oplus F$, entonces X es un isomorfismo sobre $W = Im(X) = X(IR^{p+1}) = X(F)$. Entonces a todo $Y \in W$ corresponde

a un solo $b \in F$ tal que $Y = Xb$. Si se toma $LY = Hb$, lo que define L de manera única, se tiene $\forall Y \in W$: $LY = Hb$, es decir que $\forall b \in F : LXb = Hb$. Se deduce entonces que H es estimable.

Consecuencias del teorema:

- Sea b una solución de las ecuaciones normales. Si H es estimable, entonces Hb no depende de la solución particular b elegida. En efecto, $b = b_0 + b_1$ en que $b_0 \in Ker(X)$, b_1 único. Luego $b_0 \in Ker(H)$ y $Hb = Hb_0 + Hb_1 = Hb_1$ que es invariante. Además $LXb = Hb$ no depende de L. Además Hb es insesgado para $H\beta$.
- Si se busca un estimador insesgado de β, como se tiene $q = p+1$, H es la identidad en IR^{p+1} y como $Ker(H) = \{0\} \Rightarrow Ker(X) = \{0\}$. El modelo tiene que ser de rango completo.
- En conclusión, en un modelo de rango completo, toda función vectorial de X es estimable.

3.7.3. 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 $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$ 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 $\beta^+ = \beta_1 + U\beta_2 \in L_1$.

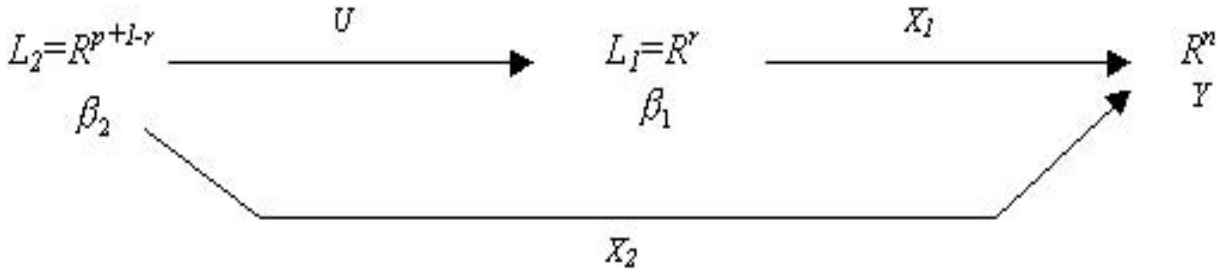


Figura 3.8:

Teorema 3.4 Una condición necesaria y suficiente para que $H \in l(IR^{p+1}, IR^k)$ 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.

Demostración:

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

$LX\beta = L(X_1\beta_1 + X_2\beta_2) = LX_1\beta_1 + LX_2\beta_2 = H_1\beta_1 + H_2\beta_2$. 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 $Im(X)$. Además X_1 es de rango completo, luego $\forall Y \in Im(X) \quad \exists! b_1 \in L_1 \quad tq \quad Y = X_1b_1$. Existe L tal que $LY = LX_1b_1$. Entonces $H_2 = LX_1U = LX_2$.

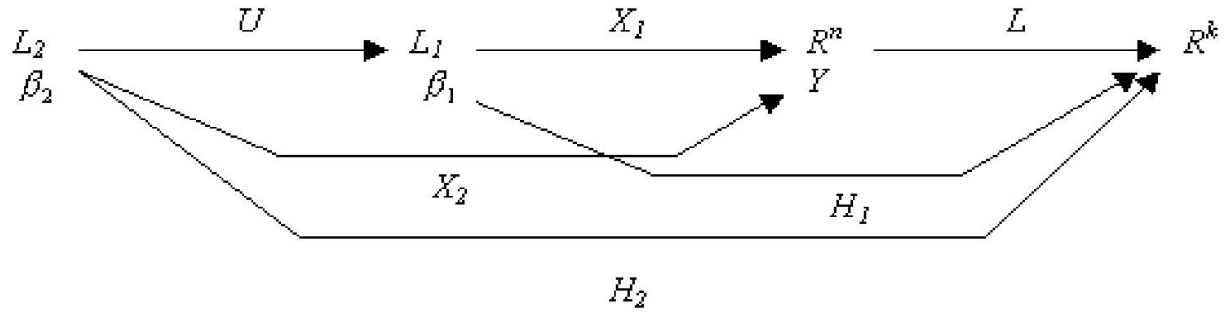


Figura 3.9:

Finalmente $\forall\beta \in IR^{p+1} : LX\beta = L(X_1\beta_1 + X_2\beta_2) = H_1\beta_1 + H_2\beta_2 = H\beta$. Luego H es estimable.

Consecuencias: el teorema de Gauss-Markov, que vimos para el caso de modelo de rango completo, puede aplicarse al caso de un modelo de rango incompleto para estimadores de funciones estimables:

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 cualquiera solución de las ecuaciones normales.

Demostración:

$$H\beta = H_1\beta_1 + H_2\beta_2 = H_1\beta^+$$

$$H\hat{\beta} = H_1\hat{\beta}^+$$

$$Var(H\hat{\beta}) = Var(H_1\hat{\beta}^+) = \sigma^2 H_1(X_1^tX_1)^{-1}H_1^t$$

que no depende de la partición X en X_1 y X_2.

3.7.4. 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.

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.

Demostración: La ecuación (1) puede escribirse usando la partición $X = (X_1, X_2)$;

$$\begin{cases} \hat{\beta}^+ = \hat{\beta}_1 + U\hat{\beta}_2 = (X_1^t X_1)^{-1} X_1^t Y \\ K_1\hat{\beta}_1 + K_2\hat{\beta}_2 = 0 \end{cases}$$

$$\begin{cases} \hat{\beta}_1 = \hat{\beta}^+ - U\hat{\beta}_2 \\ (K_2 - K_1U)\hat{\beta}_2 = -K_1\hat{\beta}^+ \end{cases} \qquad (2)$$

este sistema de ecuaciones (2) tiene una solución única si y solo si $K_2 - K_1U$ es invertible.

Notas:

- 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, $H\beta^*$ no depende del estimador β^* solución de las ecuaciones normales por lo tanto de las restricciones elegidas.

  Dos maneras de encontrar la solución de (2):

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

  $$(X^tX + MK)\beta_K = X^tY$$

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

- 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, su inversa se escribe: $A^{-1} = \begin{pmatrix} C & P^t \\ P & Q \end{pmatrix}$, entonces $\beta^* = CX^tY$.

3.8. Intervalos 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.

3.8.1. 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 $Var(\hat{\beta}_j) = \sigma^2 (X^t X)^{-1}_{j,j}$; es decir $\hat{\sigma}_j^2 (X^t X)^{-1}_{j,j}$. 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 : $\dfrac{a^t\hat{\beta} - a^t\beta}{\hat{\sigma}\sqrt{a^t(X^tX)^{-1}a}} \sim t_{n-r}$, luego el intervalo de confianza es:

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

Para un vector $A\beta \in IR^k$ con $A \in M_{k,p+1}$, sabemos de 2.5.5 que

$$(\beta - \hat{\beta})^t A^t [A(X^tX)^{-1}A^t]^{-1} A(\beta - \hat{\beta}) \sim \sigma^2 \chi_k^2$$

$$\text{y} \quad \frac{1}{\sigma^2}\sum_i \hat{\varepsilon}_i^2 \sim \chi_{n-r}^2$$

son independientes.

Luego

$$(\beta - \hat{\beta})^t A^t [A(X^tX)^{-1}A^t]^{-1} A(\beta - \hat{\beta}) \leq \frac{k}{n-r}\hat{\sigma}^2 F_{k,n-r}^{\alpha}$$

define una región de confianza elipsoidal para $A\beta$.

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

Las varianzas de β_1 y β_2 son: $\hat{\sigma}_1^2 = 1$ 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 $\beta_2 \in [-1,13; 3,13]$.

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

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

En la figura3.10a se representó los dos intervalos de confianza individuales para β_1 y β_2 y en la figura3.10b, las regiones de confianza para $\beta_1 - \beta_2$ y $\beta_1 + \beta_2$.

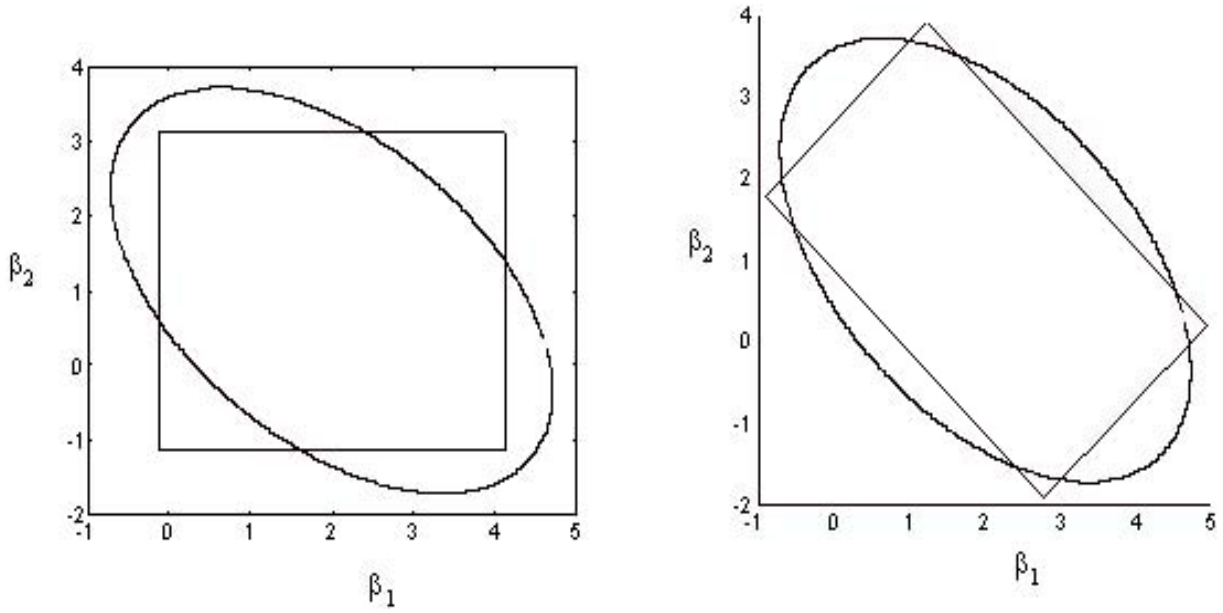


Figura 3.10: (a) Intervalo para β_1 y β_2 (b) Intervalo para $\beta_1 - \beta_2$ y $\beta_1 + \beta_2$

3.8.2. Intervalos simultáneos de confianza.

Vimos que par un vector $A\beta \in IR^k$ con $A \in M_{k,p+1}$, la región de confianza elipsoidal es tal que

$$I\!P((\beta - \hat{\beta})^t A^t [A(X^t X)^{-1} A^t]^{-1} A(\beta - \hat{\beta}) \leq \frac{k}{n-r} \hat{\sigma}^2 F^{\alpha}_{k,n-r}) = 1 - \alpha.$$

Ahora bien, si representamos esta región, por ejemplo, con los k intervalos asociados a los $a_j^t\beta$, $j = 1, 2, \ldots k$ es donde a_j^t es la fila j de A, tomando la intersección de los k intervalos:

$$\left[a_j^t \hat{\beta} - \sqrt{\frac{k}{n-r} \hat{\sigma} F^{\alpha}_{n-r} Var(a_j^t \hat{\beta})}, a_j^t \hat{\beta} + \sqrt{\frac{k}{n-r} \hat{\sigma} F^{\alpha}_{n-r} Var(a_j^t \hat{\beta})} \right]$$

obtenemos una región más amplia que la definida por el elipsoide. En efecto:

Proposición 3.3 Si C es invertible, entonces $\sup_{u \neq 0} \left(\frac{(u^t b)^2}{u^t C u} \right) = b^t C^{-1} b$.

Demostración:

$$\|v - \alpha h\|^2 = \alpha^2 \|h\|^2 - 2\alpha h^t v + \|v\|^2 = \left(\alpha \|h\| - \frac{h^t v}{\|h\|} \right)^2 + \|v\|^2 - \frac{(h^t v)^2}{\|h\|^2} \geq 0.$$

Para $h \neq 0 \Rightarrow \|v\|^2\|h\|^2 \geq (h^t v)^2$. Luego $\frac{(h^t v)^2}{\|v\|^2} \leq \|h\|^2 \iff \sup_{v\neq 0} \frac{(h^t v)^2}{\|v\|^2} = \|h\|^2$. Tomando $v = C^{1/2}u$ y $h = C^{-1/2}b$ obtenemos $\frac{(b^t u)^2}{u^t C u} \leq b^t C^{-1} b$.

Aplicando este resultado a la región de confianza de $A\beta$:

$$I\!P\left((\beta-\hat{\beta})^t A^t [A(X^tX)^{-1}A^t]^{-1} A(\beta-\hat{\beta}) \leq \frac{k}{n-r}\hat{\sigma}^2 F^{\alpha}_{k,n-r}\right) = 1-\alpha.$$

y tomando $\gamma = A\beta$, $C = [A(X^tX)^{-1}A^t]$ y $q = \frac{k}{n-r}\hat{\sigma}^2 F^{\alpha}_{k,n-r}$, obtenemos

$$I\!P((\hat{\gamma}-\gamma)^t C^{-1}(\hat{\gamma}-\gamma) \leq q) = 1-\alpha \Rightarrow I\!P\left(\forall u \neq 0 : \frac{(u^t(\hat{\gamma}-\gamma)^2)}{u^t C u} \leq q\right) = 1-\alpha.$$

Ahora bien, cuando se quiere un intervalo para $A\beta$, es equivalente a pedir k intervalos I_j al mismo tiempo para los $a_j^t\beta$, $j = 1,2,\ldots,k$ en donde a_j^t es la fila j de A. De lo anterior deducimos que el elipsoide obtenido para $A\beta$ es más que lo que pedimos que es para $\cap_j I_j$:

$$I\!P(\cap I_j) \geq 1-\alpha.$$

Para $A = I$, Scheffé propone proyectar el elipsoide asociado a β sobre los ejes de coordenadas. En general puede ser demasiado pesimista dado que

$$I\!P(\cup I_j) \geq 1-\alpha.$$

Bonferroni propone simplemente que cada I_j sea tal que $I\!P(I_j) = 1-\frac{\alpha}{k}$ $(j = 1,2,\ldots,k)$. Aquí también se tiene que $I\!P(\cup I_j) \geq 1-\alpha$. En efecto

$$I\!P(\cup I_j) = 1 - I\!P\left(\overline{\cup I_j}\right) = 1 - I\!P\left(\cap \overline{I_j}\right) \geq 1 - \sum_j I\!P\left(\overline{I_j}\right) = 1 - \sum_j \alpha_j = 1-\alpha.$$

3.9. 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	$I\!P(\|X\| > t)$
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 $I\!P(X > F) = 0,000$				

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 $\varepsilon \sim N_n(0, \sigma^2 I_n)$, $\beta \varepsilon I\!R^{p+1}$, $X \varepsilon M_{n,p+1}(I\!R)$.

   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 $A \in M_{n,n}(I\!R)$ 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 $E(\varepsilon) = 0$.

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

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

Constante	Estimación	Dev. típica estimación	t-Student	$I\!P(\|X\| > t)$
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	$I\!P(X > F) = 0,000$

Cuadro 3.2: Resultados de la regresión

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

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

Constante	Estimación	Dev. típica estimación	t-Student	$I\!P(\|X\| > t)$
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	$I\!P(X > F) = 0,000$

Cuadro 3.3: Resultados de la regresión

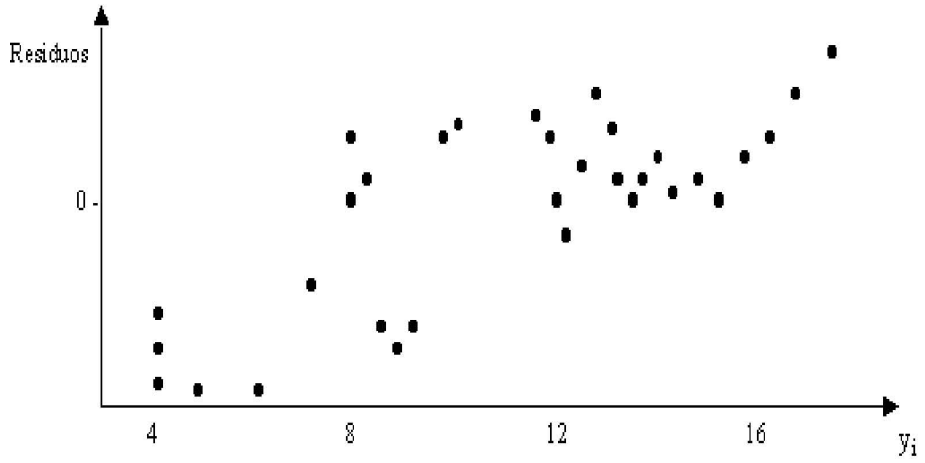
Se tiene $\sum_i x_i z_i = 0$ y $\sum_i z_i = 0$.

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.

   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 $Y \in IR^n$ un vector aleatorio con $E(Y) = \mu$ y $Var(Y) = \sigma^2 I_n$. 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 IR^n conjunto imagen de X e $\hat{Y}$ al estimador de mínimos cuadrados de $\mu = E(Y)$.

   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 $b \in IR^n$, 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 $Var(Y) = \Gamma$, 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 $rango(X) = p+1$, $E(\varepsilon) = 0$, $Var(\varepsilon)\sigma^2 I_n$.

   a) Se escribe $X^tX = \begin{bmatrix} n & a^t \\ a & V \end{bmatrix}$. 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 $\sum_j Var(\hat{\beta})$ 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^tX es diagonal.

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

   d) Se supone que X^tX 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 $\hat{\beta}$, $Var(\hat{\beta})$, $\hat{Y}$. 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^tX 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^tZ 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 $A(X^tX)^{-}A^t$ es invertible.