Capítulo 7

REGRESIÓN LINEAL

7.1. ¿PORQUE MODELAR?

Estudiamos en el capítulo anterior como detectar una asociación entre dos variables; generalmente los roles entre las variables no son simétricos - una variable puede influir sobre la otra y la recíproca no es necesariamente cierta - incluso más de una variable pueden intervenir en esta relación. En este caso nos interesaremos no solamente en evaluar la intensidad de la asociación, sino que también en describirla.

Algunas relaciones son conocidas y deterministas como ciertas leyes de la física o de la mecánica, y, dependen de constantes desconocidas que hay que determinar. Estas constantes pueden obtenerse a partir de experimentos que se utilizarán en el modelo ya planteado. El problema que surge entonces en la determinación de las constantes está en los errores de mediciones.

En otros problemas las relaciones no son conocidas y hay que determinar completamente el modelo. En ciencias sociales o economía, por ejemplo, los modelos no son deterministas y contienen una componente aleatoria, lo que dificulta la búsqueda de las relaciones. En este caso se busca descubrir como un conjunto de variables $X^1, X^2, ..., X^p$ influye sobre otra variable Y. Según el contexto, las variables X^j son llamadas **variables explicativas, variables independientes ó variables exógenas** y la variable Y es llamada **variable a explicar, variable repuesta, variable dependiente ó variable endógena**. Cuando las variables son cuantitativas, se busca una función real f que permita reconstituir los valores obtenidos sobre una muestra:

$$Y = f(X^1, X^2, ..., X^p)$$

Por una rázon histórica, este modelo se llama **regresión**. El mayor descubrimiento de Galton (párrafo **??**) fueron sus formulaciones sobre la regresión simple y su relación con la distribución normal bivariada. Hizo un estudio que mostró que la altura de los niños nacidos de padre altos tiende a retroceder o "regresar" hacia la altura promedio de la población. Por lo que utilizó entonces la palabra "regresión" para referirse a este fenómeno.

Ejemplo 7.1.1 La distancia d que una partícula recorre en un tiempo t esta dada por la formula:

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

en que β es la velocidad promedio y α es la posición de la partícula en $t = 0$. Si α y β son desconocidos, observando la distancia d en dos tiempos distintos, la solución del sistema de 2 ecuaciones lineales permite obtener α y β. Sin embargo es difícil obtener en general la distancia sin error de medición el que es de tipo aleatorio. Por lo cual se observa una variable aleatoria: $Y = d + \epsilon$ en vez de d, en donde ϵ es el error de medición. En este caso no basta tener dos ecuaciones, sino que observar los valores de la distancia recorrida en varios periodos de tiempo y métodos estadísticos basados en la aleatoriedad del error, los que permitirán estimar α, β y d sobre la base de una relación de tipo lineal.

Ejemplo 7.1.2 Si consideramos el peso y la talla de las mujeres chilenas, es obvio que no existe una relación lineal ni funcional entre la talla y el peso, pero parece existir una cierta *tendencia*. Considerando que el peso P y la talla T son variables aleatorias de distribución conjunta normal bivariada, se plantea el modelo lineal:

$$E(P|T) = \alpha + \beta T$$

en que α y β dependen de los parámetros de la distribución conjunta de P y T. El peso se escribe entonces:

$$P = \alpha + \beta T + \epsilon$$

en que ϵ refleja la variabilidad del peso P entre las chilenas de la misma talla con respecto a la media.

Ejemplo 7.1.3 Para decidir de la construción de una nueva central eléctrica, ENDESA busca estimar el consumo total de electricidad en Chile después del año 2002. Por lo tanto, se construye un modelo que liga el consumo de electricidad con variables económicas y demográficas, estimado a partir de datos de los años anteriores. Se aplica entonces el modelo para predecir el consumo de electricidad según ciertas evoluciones económicas y demográficas.

Ejemplo 7.1.4 Para establecer una determinada publicidad a la televisión, se cuantifica el efecto de variables culturales y socio-económicas sobre la audiencia de los diferentes programas.

Ejemplo 7.1.5 El modelo lineal puede ser generalizado tomando funciones de las variables explicativas y/o de la variable a explicar. En particular para un ajuste polinomial se tiene una variable Y y la variable X con algunas de sus potencias:

$$Y = a_o + a_1 X^1 + ... + a_p X^p$$

en donde X^j es la potencia j de X.

Ejemplo 7.1.6 Se quiere estimar la constante g de la gravitación; para eso se toman los tiempos de caída t de un objeto desde una altura d dada del suelo.

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

Dados los errores de mediciones, varias observaciones son necesarias y se puede considerar este modelo como lineal tomando como variable t^2.

Nos limitaremos en este texto a los modelos lineales, es decir: la variable repuesta se escribe como combinación lineal de las variables explicativas.

Presentaremos dos métodos para estimar las constantes de un modelo lineal. Conside-raremos el problema como un problema de ajuste y se propondrán el método de los mímimos cuadrados, que permite estimar los coeficientes del modelo lineal a partir de valores observados y el modelo normal para los errores que permite estimar las constantes a partir del método de máxima verosimilitud lo que permite estudiar las propiedades de los estimadores de las constantes y dar una precisión del ajuste. Finalmente se usará el modelo para hacer predicciones.

7.2. LOS MÍNIMOS CUADRADOS

Sean $\{(y_i, x_i^1, x_i^2, ..., x_i^p) | i = 1, ..., n\}$ los valores obtenidos sobre una muestra $p+1$ dimensional de tamaño n. Se plantea el modelo lineal:

$$y_i = \beta_o + \beta_1 x_i^1 + ... + \beta_p x_i^p \quad \forall i$$

donde $\beta_o, \beta_1, ..., \beta_p$ son las constantes desconocidas o sea los parámetros del modelo.

Como generalmente no existen constantes que cumplan exactamente esta relación para todas las observaciones, se escribe:

$$y_i = \beta_o + \beta_1 x_i^1 + ... + \beta_p x_i^p + \epsilon_i \quad \forall i$$

en donde ϵ_i es el **error** para la observación i debido al modelo. Se buscar entonces minimizar una función de los errores, por ejemplo,

$$\sum_i \epsilon_i^2 \qquad \sum_i |\epsilon_i| \qquad \sum_i Max\{\epsilon_i\}$$

El criterio de los mínimos cuadrados toma la función cuadrática $\sum_i \epsilon_i^2$ cuya solución es fácil de obtener y que tiene una interpretación geométrica simple.

Escribamos matricialmente el modelo aplicado a la muestra de observaciones.

Sea $\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix}$, $X = \begin{pmatrix} 1 & x_1^1 & x_1^2 & ... & x_1^p \\ 1 & x_2^1 & x_2^2 & ... & x_2^p \\ & ... & ... & ... & \\ 1 & x_n^1 & x_n^2 & ... & x_n^p \end{pmatrix}$, $\underline{\beta} = \begin{pmatrix} \beta_o \\ \beta_1 \\ ... \\ \beta_p \end{pmatrix}$, $\underline{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ \epsilon_n \end{pmatrix}$

Entonces, el modelo se escribe:

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

El criterio de los mínimos cuadrados consiste entonces en buscar el punto del subespacio vectorial $W = Im(X)$ de $\mathbb{R}^n$ generado por las columnas de la matriz X lo más cercano al punto $\underline{y}$. La solución es la proyección ortogonal del punto $\underline{y}$ sobre W.

En efecto, $\sum_i \epsilon_i^2$ es igual a $\|\underline{\epsilon}\|^2$ el cuadrado de la norma del vector $\underline{\epsilon}$, es decir el cuadrado de la distancia entre los vectores $\underline{y}$ y $X\underline{\beta}$, siendo $X\underline{\beta}$ un vector del subespacio vectorial W. El vector de W solución es entonces la proyección ortogonal de Y sobre W. Si P es el operador lineal de proyección ortogonal sobre el subespacio vectorial W, entonces la solución es $X\underline{\hat{\beta}} = P\underline{y}$. La expresión matricial de P se puede obtener en función de la matriz X: El vector $\underline{y} - P\underline{y}$ es ortogonal a W o sea que $\underline{y} - X\underline{\hat{\beta}}$ es ortogonal a cada columna de X; si se denotan $X_o, X_1, ..., X_p$ las $p+1$ columnas de X, se expresa la ortogonalidad en termino de $p+1$ productos escalares:

$$< \underline{y} - X\underline{\beta}, X_j > \qquad (j = 0, 1, ...p)$$

Matricialmente se escribe: $X_j^t(\underline{y} - X\underline{\beta}) = 0$ $(\forall j)$, y juntando las $p+1$ ecuaciones se obtiene las **Ecuaciones Normales**:

$$X^t X\beta = X^t y$$

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

$$\underline{\hat{\beta}} = (X^t X)^{-1} X^t \underline{y}$$

Se puede obtener el resultado por derivación matricial también.

Observamos que el estimador $\underline{\hat{\beta}}$ de β es lineal en Y.

El operador de proyección ortogonal sobre W se escribe matricialmente como:

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

Este operador lineal P es idempotente de orden 2 $(P^2 = P)$ y simétrico $(P^t = P)$.

Si la matriz X es de rango incompleto (rango inferior a $p+1$), basta encontrar una base de W entre las columnas de X, y reemplazar X por la matriz formada de estas columnas linealmente independientes.

7.3. MÁXIMA VEROSIMILITUD

En el párrafo anterior, se uso un criterio matemático para estimar los coeficientes β_j. Aquí usaremos un modelo probabilístico y el método de máxima verosimilitud para estimarlos. El modelo consiste en la esperanza condicional de la variable repuesta Y dadas las variables explicativas $X^1, X^2, ..., X^p$:

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

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:

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

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

El estimador de máxima verosimilitud de β verifica las Ecuaciones Normales:

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

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 $\varepsilon \sim N_n(0, \sigma^2 I_n)$.

7.4. PROPIEDADES DE LOS ESTIMADORES

Las propiedades del estimador $\underline{\hat{\beta}}$ estan ligadas a los supuestos hechos sobre los errores ϵ_i. Supondremos aquí que X es de rango $p+1$ o sea $\underline{\hat{\beta}} = (X^t X)^{-1} X^t \underline{y}$.

- El estimador es insesgado: $E(\underline{\epsilon}) = \underline{0} \Longrightarrow E(\underline{\hat{\beta}}) = \underline{\beta}$
- El estimador es consistente.

- El estimador tiene mínima varianza: Teorema de GAUSS MARKOV:

Teorema 7.4.1 *Si $E(\underline{\epsilon}) = \underline{0}$ y $E(\underline{\epsilon}\underline{\epsilon}^t) = \sigma^2 I_n$, entonces toda combinación lineal $a^t\underline{\hat{\beta}}$ de $\underline{\hat{\beta}}$ tiene mínima varianza entre los estimadores insesgados lineales en $\underline{y}$ de $a^t\underline{\beta}$. Además si $\varepsilon \sim N_n(0, \sigma^2 I_n)$, entonces $\hat{\beta}$ tiene mínima varianza entre todos los estimadores insesgados de β.*

Demostración Hay que comparar las varianzas de $a^t\underline{\hat{\beta}}$ y $a^t\underline{\beta^*}$ en que $\underline{\beta^*}$ es un estimador insesgado de la forma $C\underline{y}$.

$\underline{\beta^*} = \underline{\hat{\beta}} + D\underline{y}$, en que $D = C - (X^tX)^{-1}X^t$.

Como los dos estimadores son insesgados, $E(Dy) = 0$ y luego $DX = 0$.

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

$Cov(\underline{\hat{\beta}}, D\underline{y}) = \sigma^2(X^tX)^{-1}X^tD^t = 0$

$Var(\underline{\beta^*}) = Var(\underline{\hat{\beta}}) + \sigma^2 DD^t$

Luego, $Var(a^t\underline{\beta^*}) = a^tVar(\underline{\hat{\beta}})a + \sigma^2 a^t DD^t a$

Como $\sigma^2 a^t DD^t a > 0$, $Var(\underline{\beta^*}) > Var(\underline{\hat{\beta}})$.

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

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

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. ■

- La estimación de σ^2 obtenida por máxima verosimilitud es sesgada:
$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i^1 - ... - \hat{\beta}_p x_i^p$. Entonces, si $Q = I - P \Rightarrow \underline{\hat{\varepsilon}} = Q\underline{y} = Q\underline{\epsilon} \Rightarrow \sum \hat{\varepsilon}_i^2 = \underline{\hat{\varepsilon}}^t\underline{\hat{\varepsilon}} = \underline{\epsilon}^tQ^tQ\underline{\epsilon} = \underline{\epsilon}^tQ\underline{\epsilon} = Traza(Q\underline{\epsilon}\underline{\epsilon}^t)$ Luego $E(\underline{\hat{\varepsilon}}^t\underline{\hat{\varepsilon}}) = Traza(QE(\underline{\epsilon}\underline{\epsilon}^t)) = \sigma^2 Traza(Q)$

$Traza(Q) = Traza(I - X(X^tX)^{-1}X^t) = n - Traza(I_{p+1}) = n - p - 1$

Es decir que $E(\underline{\hat{\varepsilon}}^t\underline{\hat{\varepsilon}}) = (n - p - 1)\sigma^2$

Se obtiene entonces un estimador insesgado de σ^2 tomando:

$$\hat{\sigma}^2 = \frac{\underline{\hat{\varepsilon}}^t\underline{\hat{\varepsilon}}}{n - p - 1} = \frac{1}{n - p - 1}(\underline{y} - X\underline{\hat{\beta}})^t(\underline{y} - X\underline{\hat{\beta}})$$

7.5. INTERVALO DE CONFIANZA PARA LOS COEFICIENTES

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 $Var(\hat{\beta}_j) = \sigma^2(X^tX)^{-1}_{jj}$; es decir $\hat{\sigma}_j^2(X^tX)^{-1}_{jj}$. El intervalo de confianza de nivel de confianza igual a $1-\alpha$ es:

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

7.6. CALIDAD DEL MODELO

Para ver si el modelo es válido, hay que realizar varios estudios: 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, 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.

7.6.1. 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 $\mathbb{R}^n$ (Figura **??**).

Se pueden comparar las siguientes varianzas:

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

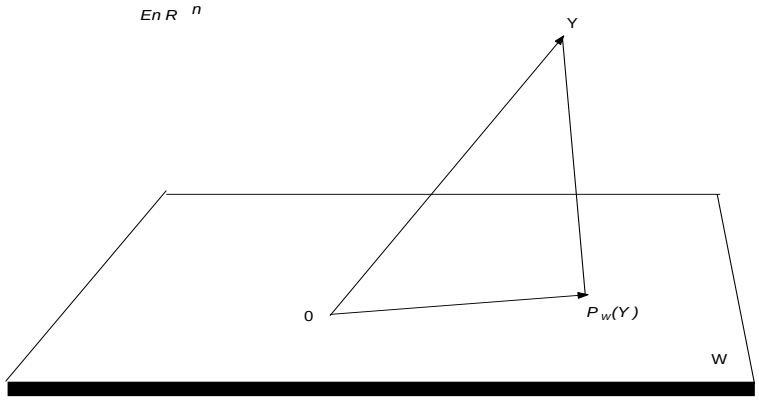


Figura 7.1: Proyección del vector Y en W

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

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 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: $\hat{y}_i = \bar{y}$ $(\forall i)$ ($\bar{y}$ es 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 $\mathbb{R}^n$ de componentes todas iguales a 1.

Si se plantea la hipótesis global $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \iff H_0 : E(y_i) = \beta_0$ $(\forall i)$, 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 $\hat{Y}|_{H_0} \sim N_n(\beta_0 1_n, \sigma^2 P) \iff \hat{\beta}_0 = \bar{y}$, se tiene:

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

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:

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

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.

7.6.2. Medición del efecto de cada variable en el modelo

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

Consideremos 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 tests individuales sobre los efectos tienen validez cuando las variables explicativas son relativamente independientes. Cuando esto ocurre, es decir, cuando una variable X_j puede tener un efecto sobre Y distinto combinado con otras variables, hay entonces que eliminar los efectos de las otras variables. Para eso se puede usar el **coeficiente de correlación parcial**.

7.6.3. 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 7.6.1 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 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, 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 más 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 (Gráfico **??**). 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:

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

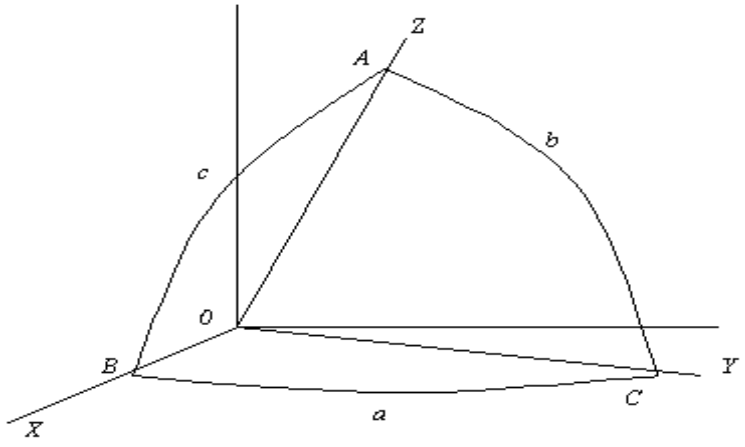


Figura 7.2: Representación esférica del coeficiente de correlación parcial

7.6.4. 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: $Y = \gamma_0 + \gamma_{j_{k+1}} X_{j_{k+1}} + \cdots + \gamma_{j_p} X_{j_p} + \varepsilon_0$. Se tiene la varianza residual bajo H_1 menor que la varianza residual bajo H_0:

$$\sum_i \hat{\varepsilon}_i^2 \leq \sum_i \hat{\varepsilon}_{0,i}^2$$

Se puede estudiar el cociente de las dos varianzas residuales $\dfrac{\sum\limits_i \hat{\varepsilon}_{0,i}^2}{\sum\limits_i \hat{\varepsilon}_i^2}$ o su complemento $\dfrac{\sum\limits_i \hat{y}_{0,i}^2}{\sum\limits_i \hat{\varepsilon}_i^2}$

en donde $\hat{y_{0,i}} = y_i - \hat{\varepsilon}_{0,i}^2$ son las componentes del estimador $E(Y|X)$ bajo H_0.

Bajo la hipótesis H_0

$$Q = \frac{\dfrac{\sum\limits_i (\hat{y}_i - \hat{y}_{0,i})^2}{k}}{\dfrac{\sum\limits_i \hat{\varepsilon}_i^2}{n-r}} \sim F_{k,n-r}.$$

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: $\Lambda = \dfrac{\max\limits_{H_0} L}{\max L}$. La región crítica se escribe $S > C'$ Este test coincide con el test F.

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

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

$$\Lambda = \frac{\underset{H_0}{\text{máx}}\, L}{\text{máx}\, L} = \exp\left\{-\frac{1}{2\sigma^2}\|\hat{Y}-y^*\|^2\right\}.$$

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 $D\hat{\beta} \sim N(D\beta, \sigma^2\Gamma)$, en que Γ depende solamente de D y X. Bajo H_0, $\frac{\hat{\beta}^t D^t \Gamma^{-1} D\hat{\beta}}{\sigma^2} \sim \chi^2_k$. Pero este test no equivale en general al test de razón de verosimilitudes basado en $\|\hat{Y}-Y^*\|^2$.

7.7. 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 7.7.1 *La diferencia de las varianzas residuales con y sin restircción es:*

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

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

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 $\dfrac{U}{\sigma^2} \sim \chi^2_k$.
■

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

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

7.8. 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

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

considerndolos 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 (gráfico 7.3) 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[9]).

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.

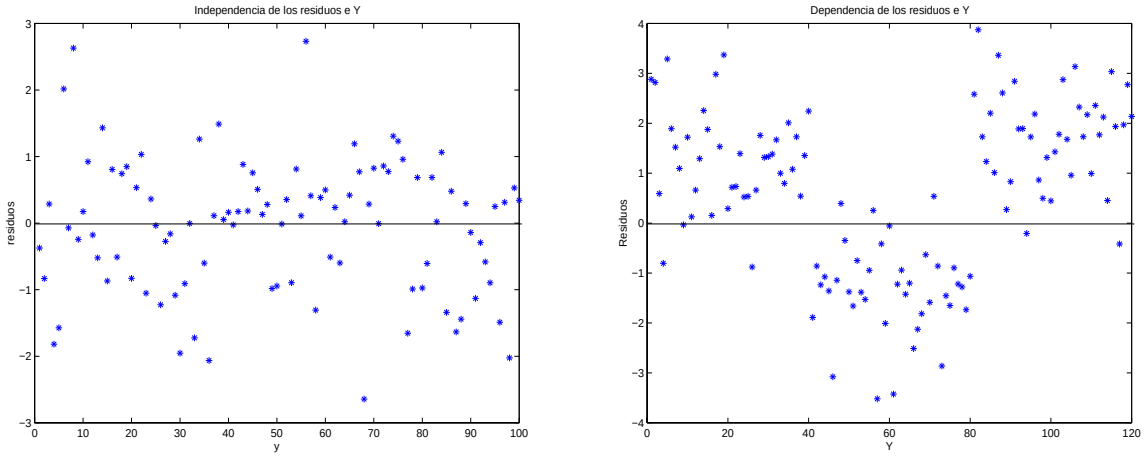


Figura 7.3: Gráficos de residuos

7.8.1. 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 pequeos, 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 hiptesis 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 $q_i = F^{-1}(F_n(u_{(i)}$. 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 7.4).

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

7.9. 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

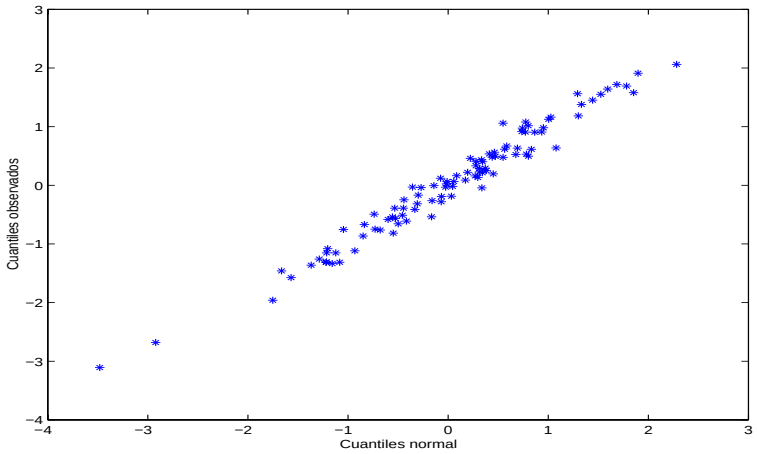


Figura 7.4: Recta de Henri

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:

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

7.10. EJERCICIOS

1. Cuatro médicos estudian los factores que explican la espera de los 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	$I\!P(X > F) = 0,000$		

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

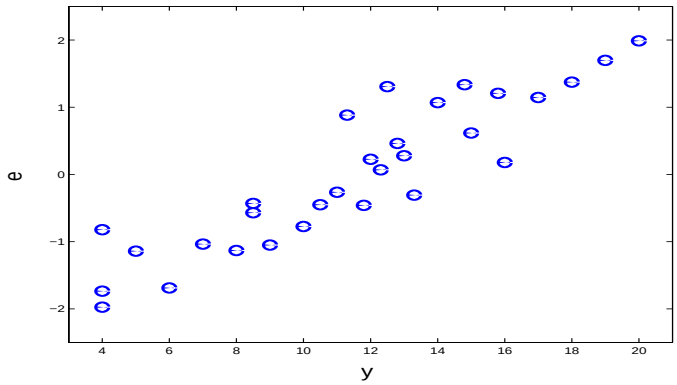
2. Consideramos el modelo lineal $Y = X\beta + \varepsilon$ con $\varepsilon \sim N_n(0, \sigma^2 I_n)$, $\beta \varepsilon I\!R^{p+1}$, $X \varepsilon M_{n,p+1}(I\!R)$.

a) Escribamos X como: $X = (X_1 \quad X_2)$, 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 $\hat{\beta}$ bajo 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.



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

Variable	Medias	Desv. típica	Estimación	Dev. típ. estimación	t-Student	$\mathbb{P}(\|X\| > t)$
Y	11,68	3,46				
Constante			7,06	1,03	6,84	0,00
X	5,854	2,74	0,79	0,16	4,94	0,00

Coef. determinación=0,39 F de Fisher=24,44 $\mathbb{P}(X > F) = 0,000$

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	Estimación	Dev. típ. estimación.	t-Student	$\mathbb{P}(\|X\| > t)$
Y	11,68	3,46				
Y	11,68	3,46				
Constante			11,68	0,36	32,54	0,00
Z	0,00	2,65	1,00	0,14	7,27	0,00

Coef. determinación=0,58 F de Fisher=52,78 $\mathbb{P}(X > F) = 0,000$

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 $Y \in I\!R^n$ 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 $I\!R^n$ 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 $b \in I\!R^n$, 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 $Var(Y) = \Gamma$,
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 $X^t X = \begin{bmatrix} n & a^t \\ a & V \end{bmatrix}$. 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 $\sum_j Var(\hat{\beta})$ 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 $X^t X$ 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.