PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE MATEMÁTICAS / DEPARTAMENTO DE ESTADÍSTICA
ELM2400 Métodos Estadísticos

Análisis de Regresión

Autora: Lorena Correa Arratia

Índice

Nota: Si encuentra algún error, envíalo a rigonzas@puc.cl

1. Introducción

Como hemos estudiado anteriormente, la covarianza y el coeficiente de correlación son medidas de asociación entre dos variables aleatorias.

Suponga dos variables aleatorias X e Y, el coeficiente de correlación lineal es definido por:

$$\rho_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}}$$

En general, disponemos de muestras de pares de datos $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ y se define el coeficiente de correlación muestral (estimador) como:

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)}} = \frac{S_{xy}}{S_{xx} \cdot S_{yy}}$$

Notemos que $-1 \leq \rho_{xy} \leq 1$, de la misma forma $-1 \leq r_{xy} \leq 1$.

Es recomendable graficar la información por ejemplo:

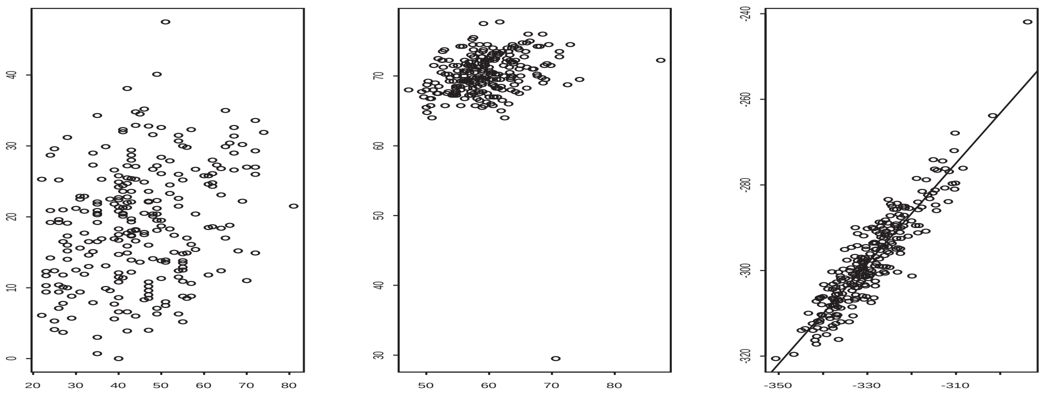

Figura 1: Asociaciones Nulas y Positiva

Observemos que la correlación es una medida de asociación lineal.

2. Regresión Lineal Simple

Con la regresión lineal se pretende ir más allá de medir el grado de asociación de dos variables aleatorias.

Concretamente se quiere:

1. Investigar la naturaleza de la relación.

2. Construir modelos que describan la relación de las variables.
3. Predecir el comportamiento de una de ellas a partir de valores de la otra.

Ejemplo

Supermercado	1	2	3	...	...	252
X: Número de empleados	10	17	17	...	...	48
Y: Ventas en cientos de miles de pesos	4	6	6	...	...	20

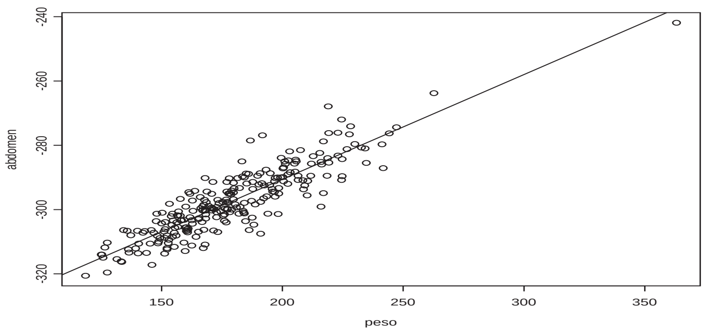

Figura 2: Numero de empleados v/s Ventas

El interés del gerente del supermercado esta en predecir las ventas a partir del numero de empleados contratados, con el propósito de determinar el numero óptimo para maximizar las ventas.

$$\underbrace{\text{Ventas}}_{\text{Variable respuesta}} = \underbrace{f(\text{Número de empleados})}_{\text{Variable independiente o explicatoria}}$$

Podemos plantear:

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Componente Deterministica}} + \underbrace{e_i}_{\text{Componente aleatoria}} \qquad , i = 1, 2, ..., n$$

Suponemos que $e_i \sim N(0, \sigma^2) \qquad , i = 1, 2, ..., n$.

Entonces de lo anterior, se tiene que

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2) \qquad , i = 1, 2..., n$$

2.1. Pasos de un modelo de Regresión

(a) Formular un modelo para $E(Y)$
(b) Testear las variables incorporadas
(c) Predicción

2.2. Estimación de los Parámetros de la Regresión

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad , i = 1, 2, ..., n$$

Donde $\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores de β_0 y β_1, obtenidos a través del método de mínimos cuadrados. El método de mínimos cuadrados consiste en minimizar la suma de cuadrados del error, es decir, minimizar:

$$S^2 = \sum_{i=1}^{n} (Y_i - E(Y_i))^2 = \sum_{i=1}^{n} e_i^2$$

En nuestro caso:

$$\begin{aligned} S^2 = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i)))^2 &= \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i))^2 \\ \longrightarrow \frac{\partial S^2}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \\ \longrightarrow \frac{\partial S^2}{\partial \beta_1} &= -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i = 0 \end{aligned}$$

Despejando $\hat{\beta}_0$ y $\hat{\beta}_1$ se tiene que:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Notacionalmente:

$$\sum_{i=1}^{n} x_i^2 \quad : \quad \text{Suma de Cuadrados de } x \text{ no corregida}$$

$$\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \quad : \quad \text{Suma de Cuadrados de } x \text{ corregida}$$

$$
\begin{aligned}
S_{xy} &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \\
S_{xx} &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \\
S_{yy} &= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \\
\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}}
\end{aligned}
$$

3. Precisión de la recta de Regresión

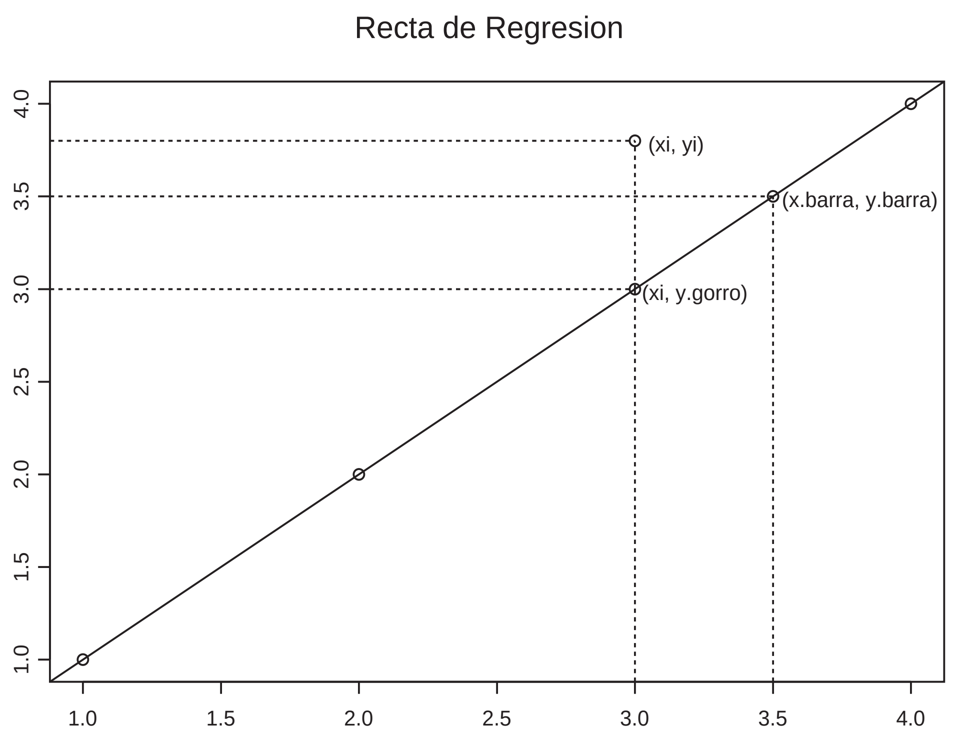

Figura 3: Precision de la recta de regresión

Entonces:
(1) Diferencia entre el valor observado y el valor ajustado $(y_i - \hat{y}_i)$
(2) Diferencia entre el valor ajustado y el valor medio $(\bar{y} - \hat{y}_i)$

(3) Diferencia entre el valor observado y el valor medio $(y_i - \bar{y})$

Consideremos la siguiente identidad:

$$
\begin{aligned}
(3) &= (1) + (2) \\
(y_i - \bar{y}) &= (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \qquad \backslash()^2 \\
(y_i - \bar{y})^2 &= (y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \qquad \backslash\sum \\
\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{SCT} &= \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{SCE} + \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{SCReg}
\end{aligned}
$$

Asi

$$
\begin{aligned}
\sum_{i=1}^{n} (y_i - \bar{y})^2 &: \text{ Suma de Cuadrados Total (SCT)} \\
\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 &: \text{ Suma de Cuadrados de la Regresión (SCReg)} \\
\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 &: \text{ Suma de Cuadrados del Error (SCE)}
\end{aligned}
$$

Entonces $SCT = SCReg + SCE$

Estas sumas de cuadrados nos permiten ver la çalidad”de la regresión. Buscaremos que SCR sea lo mas grande posible, o bien que la razón entre SCR y SCT sea lo mas cercano a uno.

El coeficiente de determinación se define como:

$$R^2 = \frac{SCR}{SCT} = \frac{SCT - SCE}{SCT} = 1 - \frac{SCE}{SCT}$$

R^2 se denomina coeficiente de determinación y mide la proporción de la varianza de la variable y explicada por las variables independientes en el modelo.

Si:

- $R^2 \approx 1$
  Indica que las variables independientes del modelo explican en gran parte las variaciones de la variable dependiente. Ósea, el modelo es bueno.

- $R^2 \approx 0$
  Indica que parte de la variación de la variable dependiente esta en los residuos, por tanto no es aplicada por las variables independientes. Ósea, el modelo es malo.

Se puede mostrar que R^2 es igual al cuadrado del coeficiente de correlación entre las observaciones y_i y los valores ajustados $\hat{y}_i$.

4. Tabla de Análisis de la Varianza (ANOVA)

Fuente	g.l.	SC	MC	Test F
Regresión	1	$\sum(y_i - \bar{y})^2$	$SCReg/1$	
Error	$n-2$	$\sum(y_i - \hat{y}_i)^2$	$SCE/(n-2)$	$F_c = MCReg/MCE$
Total	$n-1$	$\sum(\hat{y}_i - \bar{y})^2$		

4.1. Estimación por Máxima Verosimilitud

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2) \qquad , i = 1, 2, ..., n$$

Luego la verosimilitud de Y_i es:

$$\begin{aligned}
L(Y_i/\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} f(Y_i) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2} \\
&= (2\pi\sigma^2)^{-n/2} \cdot exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\Big)
\end{aligned}$$

$$\begin{aligned}
logL(Y_i/\beta_0, \beta_1, \sigma^2) &= -\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2 \\
\longrightarrow \frac{\partial \ell}{\partial \beta_0} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) \\
\longrightarrow \frac{\partial \ell}{\partial \beta_1} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)X_i
\end{aligned}$$

Igualando a cero:

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i &= 0 \\ \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i Y_i &= 0 \end{aligned}$$

Las ecuaciones de estimación coinciden con las del $EMCO$. Lo que implica que los EMV coinciden con los $EMCO$

Por propiedad de los EMV sabemos que:

$$\hat{\beta}_i \sim N(\beta_i, Var(\hat{\beta}_i))$$

$$\begin{aligned} Var(\hat{\beta}_i) &= Var\Big(\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}\Big) \\ &= \frac{\sum (x_i - \bar{x})^2 Var(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\ &= \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \end{aligned}$$

Luego

$$\hat{\beta}_1 \sim N\Big(\beta_1, \frac{\sigma^2}{\sum (x_i - \bar{x})^2}\Big)$$

$$\begin{aligned} Var(\hat{\beta}_0) &= Var(\bar{y} - \hat{\beta}_1 \bar{x}) \\ &= Var(\bar{y}) + Var(\hat{\beta}_1 \bar{x}) - 2Cov(\bar{y}, \hat{\beta}_1 \bar{x}) \\ &= \frac{\sigma^2}{n} + \bar{x}^2 Var(\hat{\beta}_1) \\ &= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \\ &= \sigma^2 (\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}) \end{aligned}$$

Luego

$$\hat{\beta}_0 \sim N\Big(\beta_0, \sigma^2 (\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2})\Big)$$

Como $SCT = SCE + SCReg$

Entonces

$$\begin{aligned} E(SCReg) &= E(SCT) + E(SCE) \\ &= \sigma^2(n-1) - \sigma^2(n-2) \\ &= \sigma^2(n-1-n+2) \\ &= \sigma^2 \end{aligned}$$

Entonces

$$\frac{SCReg}{\sigma^2} \sim X^2_{(1)}$$

Con lo anterior podemos formar la tabla ANOVA.
Observemos que

$$\frac{SCReg/1}{SCE/(n-2)} \sim F_{1,n-2}$$

Estas distribuciones nos permitirán construir un Test de Hipótesis para verificar si existe o no regresión, es decir, si Y puede ser modelada a través de X, lo que implica docimar:

$$\begin{aligned} H_0 &: \beta_1 = 0 \Longleftrightarrow \text{No hay regresión} \\ H_1 &: \beta_1 \neq 0 \end{aligned}$$

Por otra parte, si solo estamos interesados en docimar:

$$\begin{aligned} H_0 &: \beta_i = 0 \\ H_1 &: \beta_i \neq 0 \qquad \text{, para cualquier } i = 0, 1 \end{aligned}$$

Se tiene que:

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{Var(\hat{\beta}_i)}} \sim t_{n-2}$$

5. Estimación de σ^2

$$\begin{aligned} logL(Y_i/\beta_0, \beta_1, \sigma^2) &= -\frac{n}{2} \cdot log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2} \cdot \frac{1}{2\pi\sigma^2} \cdot 2\pi + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \end{aligned}$$

Igualando a cero:

$$\hat{\sigma}^2 = \frac{\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n} = \frac{\sum \hat{e}_i^2}{n} = \frac{SCE}{n}$$

Observemos que

$$E(\hat{\sigma}^2) = \frac{1}{n}E(\sum \hat{e}_i^2) = \frac{1}{n}\sigma^2 \cdot E\Big(\underbrace{\frac{\sum \hat{e}_i^2}{\sigma^2}}_{X^2_{(n-2)}}\Big) = \frac{1}{n}\sigma^2(n-2)$$

Luego, $\hat{\sigma}^2$ no es un estimador insesgado, utilizaremos entonces como estimador de σ^2 a $\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{n-2}$

$$\hat{\sigma}^2 = \frac{SCE}{n-2}$$

Entonces

$$\begin{aligned} SCE/\sigma^2 &\sim X^2_{(n-2)} \\ SCT/\sigma^2 &\sim X^2_{(n-1)} \end{aligned}$$

6. Regresión Múltiple

En regresión múltiple tenemos varias variables explicativas, para modelar el comportamiento de una variable respuesta. En general, tenemos n observaciones de una variable respuesta Y, y para cada una de ellas, n observaciones correspondientes a cada una de las p variables explicatorias.

Obs	Variable Respuesta	Variables Explicatorias
	Y	$X_1X_2\ldots X_p$
1	Y_1	$X_{11}X_{12}\ldots X_{1p}$
2	Y_2	$X_{21}X_{22}\ldots X_{2p}$
$\vdots$	$\vdots$	$\vdots$
n	Y_n	$X_{n1}X_{n2}\ldots X_{np}$

Donde x_{ij} corresponde a la $i-esima$ observación de la $j-esima$ variable explicatoria.

$$Y = \beta_0 + \beta_1X_1 + ... + \beta_pX_p + \varepsilon$$

Considerando las n observaciones, se tiene que:

$$Y_i = \beta_0 + \beta_1X_{i1} + ... + \beta_pX_{ip} + \varepsilon_i \qquad , i = 1, 2, ..., n$$

Al igual que antes $\beta_0, \beta_1, ..., \beta_p$ son parámetros desconocidos y ε es el error.
Matricialmente se tiene que:

$$Y_{n\times 1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \varepsilon_{n\times 1} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \qquad X_{n\times p} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}}_{\text{Matriz de Diseño}}$$

El modelo de regresión múltiple se escribe como:

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

Supuestos del modelo:

1. $E(\varepsilon) = 0 \longleftrightarrow E(Y) = X\beta$
2. $Var(\varepsilon) = \sigma^2 I$
3. $\varepsilon \sim N_n(0, \sigma^2 I)$

6.1. Estimación de Mínimos Cuadrados

Se quiere minimizar

$$S^2 = \sum_{i=1}^{n} e_i^2 = \varepsilon^T \varepsilon$$

$$\begin{aligned} SCE = \varepsilon^T\varepsilon &= (Y - \hat{Y})^T(Y - \hat{Y}) = (Y^T - \hat{Y}^T)(Y - \hat{Y}) \\ &= (Y^T - \hat{\beta}^T X^T)(Y - X\hat{\beta}) \\ &= Y^T Y - Y^T X\hat{\beta} - \hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta} \end{aligned}$$

Así $SCE = Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta}$.

Derivando con respecto a $\hat{\beta}$

$$\begin{aligned} \frac{\partial SCE}{\partial \hat{\beta}} &= -2X^T Y + 2X^T X\hat{\beta} = 0 \\ & \quad (X^T X)\hat{\beta} = X^T Y \qquad (*) \\ \longrightarrow & \quad \hat{\beta} = (X^T X)^{-1} X^T Y \end{aligned}$$

Si existe $(X^T X)^{-1}$ el sistema (*) tiene solución. Para esto necesitamos que la matriz $(X^T X)$ sea de rango completo.

Propiedades de $\hat{\beta}$:

1. Insesgado: $E(\hat{\beta}) = \beta$
2. $var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$

TAREA

1. Demuestre lo anterior.

2. Para

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{pmatrix}$$

Encuentre

$$\hat{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad \text{y } Var(\hat{\beta})$$

6.2. Estimación de las Medias (Valores Ajustados) y Predicción

Consideremos el problema de estimar la media de una observación y_i . La media de $\hat{y}_i$ esta dada por:

$$E(y_i) = u_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} = x_i^T \beta$$

Un estimador natural para u_i es

$$\hat{u}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_p x_{pi} = x_i^T \hat{\beta}$$

Se puede mostrar que este es un estimador insesgado y tiene varianza

$$Var(\hat{u}_i) = Var(x_i^T \hat{\beta}) = \sigma^2 x_i^T (x^T x)^{-1} x_i$$

La estimación de u_i, también es posible para una observación fuera de la muestra, por ejemplo y^*. Así

$$\begin{aligned} E(y^*) &= u^* = x^{T*} \beta \\ \hat{u}^* &= x^{T*} \hat{\beta} \\ Var(\hat{u}^*) &= \sigma^2 x^{T*} (x^T x)^{-1} x^* \end{aligned}$$

Así el error estándar del valor medio estimado será:

$$S.E.(\hat{u}^*) = \sigma \left(x^{T*} (x^T x)^{-1} x^* \right)^{1/2}$$

A $\hat{u}^*$ le llamaremos la predicción de y^*.

Se puede observar que una predicción coincide con la estimación de la media correspondiente.

La varianza del error de predicción la podemos escribir como:

$$\begin{aligned} Var(y^* - \hat{y}^*) &= Var(y^*) + Var(\hat{y}^*) - 2Cov(y^*, \hat{y}^*) \\ &= \sigma^2 + \sigma^2 x^{T*} (x^T x)^{-1} x^* - 0 \\ &= \sigma^2 (1 + x^{T*} (x^T x)^{-1} x^*) \end{aligned}$$

Observemos que:

$$Cov(y^*, \hat{y}^*) = Cov(x^{T*}\beta + \varepsilon^*, x^{T*}\hat{\beta}) = Cov(\varepsilon^*, x^{T*}\hat{\beta}) = 0$$

Pues la observación y^* es independiente de las utilizadas para estimar β.
El error estándar del error de predicción es:

$$S.E.(y^* - \hat{y}^*) = \sigma\big(1 + x^{T*}(x^T x)^{-1}x^*\big)^{1/2}$$

6.3. Tabla ANOVA

Fuente	**SC**	**g.l.**	**SE**	F_c
Regresión	$(\hat{y}-\bar{y})^T(\hat{y}-\bar{y})$	p	$\sigma^2 p$	
Error	$(y-\hat{y})^T(y-\hat{y})$	$n-(p+1)$	$\sigma^2(n-(p+1))$	$\frac{SCReg/p}{SCE/(n-p-1)}$
Total	$(y-\bar{y})^T(y-\bar{y})$	$n-1$	$\sigma^2(n-1)$	

$$\begin{aligned} SCE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - ... - \hat{\beta}_p x_{pi})^2 = \sigma^2 X^2_{(n-(p+1))} \\ SCT = \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sigma^2 X^2_{(n-1)} \end{aligned}$$

Luego

$$\begin{aligned} SCR &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = SCT - SCE \\ \longrightarrow g.l.(SCR) &= g.l.(SCT) - g.l.(SCE) \\ &= n - 1 - (n - (p+1)) \\ &= p \end{aligned}$$

6.4. Restricciones Lineales

En aplicaciones es usual el querer estimar un modelo de regresión bajo restricciones lineales en los coeficientes β_j, $j = 0, ..., p$ y desarrollar un Test de hipótesis para determinar la validez de las restricciones.

El Test t solo se podrá aplicar cuando el número de restricciones sea igual a uno. Para un número mayor se requiere desarrollar otro procedimiento.

Ejemplo
Suponga el modelo de regresión múltiple

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \qquad , i = 1, 2, ..., n$$

Con las restricciones

$$\begin{aligned} \beta_1 - 2\beta_2 &= 1 \\ \beta_0 &= 3 \end{aligned}$$

Despejando en función de los parámetros independientes se tiene:

$$\begin{aligned} y_i &= 3 + (1 + 2\beta_2)x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \\ y_i &= 3 + x_{1i} + 2\beta_2 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \\ y_i - 3 - x_{1i} &= (2x_{1i} + x_{2i})\beta_2 + \beta_3 x_{3i} + \varepsilon_i \\ y_i^* &= \beta_2 x_{2i}^* + \beta_3 x_{3i} + \varepsilon_i \end{aligned}$$

Así podemos estimar β_2 y β_3 de este nuevo modelo y los otros parámetros a partir de las restricciones.
Consideremos el problema de hacer un Test de las restricciones lineales planteadas. Supongamos que se tienen m restricciones. Entonces se puede demostrar que bajo la hipótesis nula, que dice que las restricciones son verdaderas.

$$F_c = \frac{\Big(SCE(cr) - SCE(sr)\Big)/m}{SCE(sr)/(m - p - 1)}$$

Donde

$SCE(cr)$ es la suma de cuadrados del Error del modelo con restricción.

$SCE(sr)$ es la suma de cuadrados del Error del modelo sin restricción.

El estadígrafo F_c tiene distribución Fisher con m grados de libertad en el numerador y $n - p - 1$ en el denominador bajo H_0, es decir

$$F_c \sim F_{m,n-p-1}$$

Valores de F_c mayores que un valor critico hacen rechazar la hipótesis planteada por las restricciones.

Rechazar H_0 si

$$F_c > F_{m,n-p-1,\alpha}$$

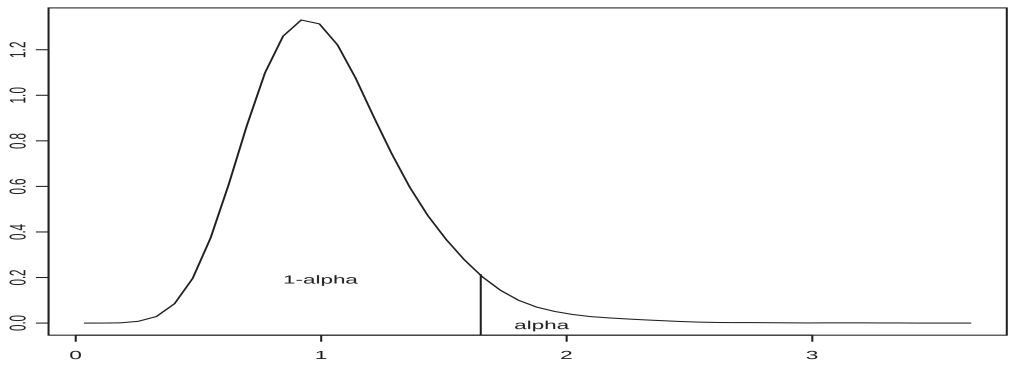

Figura 4: Región de rechazo

Una aplicación importante del Test F se conoce como el Test de significancía de la regresión.

Consideremos entonces el modelo que solo considera la constante β_0, y la hipótesis nula.

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = ... = \beta_p = 0 \\ H_1 &: \text{algún } \beta_i \neq 0 \text{ para } i = 1, 2, ..., p \end{aligned}$$

El modelo restringido corresponde a:

$$y_i = \beta_0 + \varepsilon_1 \qquad , i = 1, 2, ..., n$$

Y por lo tanto la suma de cuadrados de los residuos del modelo restringido es igual a:

$$SCE(cr) = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Pues

$$\begin{aligned} S^2 &= \sum_{i=1}^{n} (y_i - \beta_0)^2 \\ \longrightarrow \frac{\partial S^2}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_0) = 0 \\ \longrightarrow \hat{\beta}_0 &= \bar{y} \end{aligned}$$

Luego F_c queda de la forma:

$$F_c = \frac{\Big(\sum (y_i - \bar{y})^2 - SCE(sr) \Big) / p}{SCE(sr)/(n - p - 1)}$$

Bajo H_0, $F_c \sim F_{p,n-p-1}$

En general, si:

$$\begin{aligned} H_0 &: \lambda^T\beta = 0 \\ H_1 &: \lambda^T\beta \neq 0 \end{aligned}$$

Donde $\beta_{p+1\times 1}$ es el vector de parámetros y $\lambda_{p+1\times 1}$ es un vector de constantes conocidas

$$\begin{aligned} \lambda^T\beta &= \lambda_0\beta_0 + \lambda_1\beta_1 + ... + \lambda_p\beta_p \\ \lambda^T\hat{\beta} &= \lambda_0\hat{\beta}_0 + \lambda_1\hat{\beta}_1 + ... + \lambda_p\hat{\beta}_p \end{aligned}$$

Como $\lambda\hat{\beta}$ es una combinación lineal de variables aleatorias normales entonces también es Normal, así

$$\lambda^T\hat{\beta} \sim N(\lambda^T\beta, \sigma^2\lambda^T(x^Tx)^{-1}\lambda)$$
$$\frac{\lambda^T\hat{\beta} - \lambda^T\beta}{\sqrt{\sigma^2\lambda^T(x^Tx)^{-1}\lambda}} \sim N(0,1)$$

Luego, rechazo H_0 si

$$\mid \frac{\lambda^T\hat{\beta} - \lambda^T\beta}{\sqrt{\sigma^2\lambda^T(x^Tx)^{-1}\lambda}} \mid > z_{1-\alpha/2}$$

Cuando σ^2 es conocido, rechazo H_0 si

$$\mid \frac{\lambda^T\hat{\beta} - \lambda^T\beta}{\sqrt{\sigma^2\lambda^T(x^Tx)^{-1}\lambda}} \mid > t_{n-p-1,1-\alpha/2}$$

Cuando σ^2 es desconocido, con

$$\hat{\sigma}^2 = \frac{SCE}{n-p-1}$$

7. Análisis de Residuos

El residuo ε_i, se define como $\varepsilon_i = y_i - \hat{y}_i, \forall i = 1,2,...,n$ y en cierto sentido es una estimación de:

$$\varepsilon_i = y_i - E(y_i)$$

El modelo de regresión supone que:

1. $E(\varepsilon_i) = 0$
2. $Var(\varepsilon_i) = \sigma^2$

3. $Cov(\varepsilon_i, \varepsilon_j) = 0 \qquad , \forall i \neq j$
4. ε_i tiene distribución Normal

Notacionalmente $\varepsilon_i \sim N(0, \sigma^2)$

En el modelo matricial $Y = X\beta + \varepsilon$

Luego

$$\begin{aligned} \varepsilon &= y - \hat{y} \\ &= y - x\hat{\beta} \\ &= y - x(x^T x)^{-1} x^T y \\ &= (I - x(x^T x)^{-1} x^T) y \\ &= My \end{aligned}$$

La matriz M es conocida como Matriz de Proyección, tiene algunas propiedades como por ejemplo ser idempotente.

La Matriz de Varianzas - Covarianzas del vector ε la denotaremos por Σ_ε y esta dada por:

$$\Sigma_\varepsilon = M \Sigma_y M^T = \sigma^2 M M^T = \sigma^2 M$$

- Se llama residuo estandarizado a

  $$r_i = \frac{\varepsilon_i}{\sqrt{\hat{\sigma}^2 (1 - h_{ii})}}$$

  Con $H = (x(x^T x)^{-1} x^T)$ y h_{ii} es el elemento (i, i) de la matriz H.

- Observaciones

  Muchos programas entregan los valores h_{ii} o bien los residuos estandarizados (MINITAB, SPLUS, SAS)

  Los r_i no tienen distribución Normal, ni t-Student pero es frecuente encontrar aproximaciones a la Normal Estándar.

- Normalidad

  1. Histograma de los Residuos
  2. Test de Normalidad para los residuos (Bondad de Ajuste, QQPlot, Shapiro-Wilk, Kolmogorov-Smirnov, etc.)

  Se debe testear que $\varepsilon_i \sim N(0, \sigma^2)$ o bien que los residuos estandarizados $r_i, \forall i = 1, 2, ..., n$

  $$r_i \sim N(0, 1)$$

- Asociación

  Como $\varepsilon_i \sim N(0, \sigma^2)$ o bien $r_i \sim N(0, 1)$ y el supuesto es que estos son independientes se debe probar que:

$$Corr(\varepsilon_i, \varepsilon_j) = 0 \qquad , \forall i \neq j$$

  1. Calcular $\rho = Corr(\varepsilon_i, \varepsilon_{i-1})$ y testear

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

  2. Test de Durbin - Watson

$$DW = \frac{\sum (\hat{\varepsilon}_{i+1} - \hat{\varepsilon}_i)^2}{\sum \hat{\varepsilon}_i^2}$$

  Existen tablas con los valores críticos.