

Análisis de regresión

Apunte del curso MAT266 escrito por

Fabián Ramírez

Autor:

Eduardo Valenzuela

Índice general

1. Regresión Lineal Simple	3
1. Motivación	3
1.1. Estimación de σ_u^2	7
2. Estimaciones intervalares de los parametros del modelo $(\beta_0, \beta_1, \sigma_u^2)$	8
2.1. Intervalo de confianza para la varianza σ_u^2	10
2.2. Prueba de hipótesis	10
2.3. Tabla ANOVA , R^2	10
2. Regresión lineal múltiple	11
1. Reescritura del problema	11
2. Recuerdo	16
2.1. Propiedades de $\hat{\beta}$	19
2.2. Necesitamos un estimador de σ_u^2	20
2.3. Prueba t para la significancia de un coeficiente específico	25
3. Suma de cuadrados extra	29
4. Predicciones en un modelo de regresión	32
5. Mínimos cuadrados ponderados	34
3. Correlaciones	37
1. Introducción	37

Regresión Lineal Simple

1. Motivación

Cuando se desea estudiar el comportamiento de una variable aleatoria Y , generalmente tendremos a nuestra disposición 'datos' de Y , es decir, tendremos una 'muestra aleatoria' es decir:

$$Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n$$

Donde Y_i son independientes e idénticamente distribuida.

La mejor forma de explicar el comportamiento de Y sería conociendo su función de distribución F_Y lo que no siempre es factible a partir de una muestra finita.

Otra forma de explicar Y , es usando una medida de localización y una de medida de dispersión, típicamente la media y la varianza de Y . Con esto podemos darnos una idea de cual la región (intervalo) en donde se concentra la 'mayoría' de los valores de Y .

Si queremos mejorar la forma de explicar el comportamiento de Y , debería buscar una o mas variables X_1, \dots, X_n que incidieran en el comportamiento de Y . Estas variables se denominan 'variables explicativas' o 'variables regresoras' e Y se denomina 'variable de respuesta' o 'explicada'.

Consideremos la situación con una variable, en donde $k = 1$, o sea, una sola variable explicativa. En esta situación lo ideal sería conocer la función de distribución condicional de Y dado $X = x$ es decir conocer $F_{Y|X=x}$ que claramente no siempre es factible obtener a partir de los datos de X y de Y , lo que nos lleva a mirar:

$$\mathbb{E}[Y|X=x] \text{ y } \mathbb{V}[Y|X=x]$$

O sea la localización y la dispersión condicionada a que $X = x$. Notemos que $\mathbb{E}[Y|X=x]$ produce una 'función' que depende del valor x de X y que entrega el valor esperando de Y cuando x se fijo en el punto x . Esta función se denomina 'Función de regresión' de Y sobre X , y se simboliza usualmente por:

$$\phi(x) = \mathbb{E}[Y|X=x]$$

Claramente esta función puede tener una estructura 'complicada' (podría ser cualquier cosa). Con el fin de poder estimar esta función a partir de una muestra de datos de X e Y , es común asumir una estructura 'lineal', es decir:

$$\phi(x) = \mathbb{E}[Y|X=x] = \beta_0 + \beta_1 x$$

Observación Notemos que la linealidad se refiere a β_0 y β_1 y no a x .

Por lo tanto el problema es determinar los parámetros β_0 y β_1 . Nuestros datos son de la forma (x_i, y_i) $i = 1, \dots, n$. Con ellos debemos poder 'estimar' estos parámetros β_0 y β_1 . Para esto es frecuente usar el método de 'mínimos cuadrados'. Necesitamos definir los errores asociados a la representación mediante la función de regresión.

Como Y es una variable aleatoria entonces el valor y_i no necesariamente coincidirá con $\mathbb{E}[Y|X=x_i]$ y esta diferencia:

$$u_i = y_i - \mathbb{E}[Y|X=x_i]$$

Es el error de aproximación de la variable, es decir:

$$\underbrace{y_i}_{\text{y}_i} = \mathbb{E}[Y|X=x_i] + u_i$$

Por tanto tenemos que $y_i = \beta_0 + \beta_1 x_i + u_i$. El criterio de mínimos cuadrados consiste en buscar valores $\widehat{\beta_0}$ y $\widehat{\beta_1}$ que minimiza la suma de cuadrados de los errores, vale decir, que minimicen la función:

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

Dado que la función g es cuadrática y convexa en β_0 y β_1 (diferenciable), la solución es trivialmente dada por $\widehat{\beta_0}$ y $\widehat{\beta_1}$ que cumplen:

1. $\nabla g(\beta_0, \beta_1)|_{(\beta_0, \beta_1)} = (\widehat{\beta_0}, \widehat{\beta_1}) = 0$
2. $Hf g(\beta_0, \beta_1)|_{(\beta_0, \beta_1)} = (\widehat{\beta_0}, \widehat{\beta_1})$ es definida positiva

Cuyas soluciones son:

$$\left\{ \begin{array}{l} \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} \\ \widehat{\beta}_1 = \frac{\sum_{i=0}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{array} \right.$$

Por otra parte

$$Hf(\beta_0, \beta_1) |_{(\beta_0, \beta_1)} = \begin{pmatrix} 2n & 2\sum_{i=1}^n x_i \\ 2\sum_{i=1}^n x_i & 2\sum_{i=1}^n x_i^2 \end{pmatrix}$$

Se ve que $2n > 0$ y $4n\sum_{i=1}^n x_i^2 - 4(\sum_{i=1}^n x_i)^2 > 0$ por tanto $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son los estimadores mínimo cuadráticos de β_0 y β_1 con estos estimadores obtenemos la estimación mínimo cuadrática de la recta de regresión o de la esperanza condicional que es:

$$\widehat{\phi}(x) = \mathbb{E}[\widehat{Y|X=x}] = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

Lo que nos lleva a la estimación de la variable de respuesta cuando $X = x$ en la forma:

$$\widehat{Y|X=x} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

Usualmente se escribe

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

Y se denomina **recta de regresión ajustada**, **recta de mínimos cuadrados**. Y con esto se define la **recta de regresión** $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ es una estimación puntual de $\mathbb{E}[Y|X=x_i]$ es decir:

$$\widehat{y}_i = \mathbb{E}[\widehat{Y|X=x_i}]$$

Notemos que el error u_i en

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

No es observable, en consecuencia el valor que produce la variable aleatoria $Y|X=x_i$ cuando X se fija en x_i **no es observable**. En consecuencia, si queremos obtener una idea de que valores generan $Y|X=x_i$, lo que se usa es determinar un intervalo de confianza en torno a $\mathbb{E}[\widehat{Y|X=x_i}]$. Notemos que $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son estimaciones puntuales de β_0 y β_1 respectivamente. Queremos determinar como cambian los valores de $\widehat{\beta}_0$ y $\widehat{\beta}_1$ en las distintas muestras o sea, cual es la distribución muestral. Para esto, se requieren hacer algunos supuestos.

En el modelo de regresión lineal simple:

$$y_i = \beta_0 + \beta_1 x_i + u_i, i = 1, \dots, n$$

Asumiremos

1. $\mathbb{E}[u_i] = 0, i = 1, \dots, n$

2. $\mathbb{V}[u_i] = \sigma_u^2, i = 1, \dots, n$

3. Se asume que la covarianza $\text{cov}(u_i, u_j) = 0; j \neq i$

Para algunos casos para hacer inferencia necesitamos el supuesto de normalidad; es decir necesitamos que $u_i \sim \mathcal{N}(0, \sigma_u^2)$; por tanto $u_i \sim^{IID} \mathcal{N}(0, \sigma_u^2)$.

Notemos que los estimadores:

$$\left\{ \begin{array}{l} \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} \\ \widehat{\beta}_1 = \frac{\sum_{i=0}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{array} \right.$$

pueden reescribirse como:

$$\begin{aligned} \widehat{\beta}_1 &= \frac{1}{\sum(x_i - \bar{x})^2} \left\{ \sum(x_i - \bar{x})(y_i - \bar{y}) \right\} \\ &= \sum_i \frac{x_i - \bar{x}}{\sum_j (x_j - \bar{x})^2} y_i \end{aligned}$$

y además

$$\widehat{\beta}_0 = \sum_i \frac{1}{n} y_i - \bar{x} \sum_i \frac{x_i - \bar{x}}{\sum_j (x_j - \bar{x})^2} y_i$$

Es decir ambos estimadores $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son combinaciones lineales de los $y_i, i = 1, \dots, n$, pero como $y_i = \beta_0 + \beta_1 x_i + u_i$ con $u_i \sim \mathcal{N}(0, \sigma_u^2)$. y $\beta_0 + \beta_1 x_i$ es determinista. Por lo tanto

$$y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma_u^2)$$

Obs: Se asume que y_i significa lo mismo que Y_i en este contexto.

Y como se sabe que las combinaciones lineales de variables normales también son normales, se tiene que $\widehat{\beta}_0$ y $\widehat{\beta}_1$ se distribuyen normalmente, o sea:

$$\left\{ \begin{array}{l} \widehat{\beta}_0 \sim \mathcal{N}\left(\mathbb{E}[\widehat{\beta}_0], \mathbb{V}[\widehat{\beta}_0]\right) \\ \widehat{\beta}_1 \sim \mathcal{N}\left(\mathbb{E}[\widehat{\beta}_1], \mathbb{V}[\widehat{\beta}_1]\right) \end{array} \right.$$

Donde:

$$\left\{ \begin{array}{l} \mathbb{E}[\widehat{\beta}_1] = \sum_i \frac{x_i - \bar{x}}{\sum_j (x_j - \bar{x})^2} \color{red}{\beta_0 + \beta_1 x_i} = \beta_1 \\ \mathbb{E}[\widehat{\beta}_0] = \beta_0 \end{array} \right.$$

Es decir ambos estimadores son **insesgados**. Veremos después que la matriz importante en el análisis es

$$\begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}^{-1}$$

Observación Es recurrente utilizar una simbología que facilita la escritura de las formas y es:

Si tenemos los datos (x_i, y_i) para $i = 1, \dots, n$, representaremos por:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

definamos

- $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$
- $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$
- $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
- $S_x^2 = \frac{1}{n-1} S_{xx}$
- $S_y^2 = \frac{1}{n-1} S_{yy}$
- $C_{xy} = \frac{1}{n-1} S_{xy}$ estimación de la covarianza
- $r_{xy} = \frac{C_{xy}}{S_x S_y}$ estimación de correlación

Con esta notación las expresiones para varianzas y covarianzas de $\hat{\beta}_0$ y $\hat{\beta}_1$ quedan

$$\left\{ \begin{array}{l} \mathbb{V}[\hat{\beta}_1] = \frac{\sigma_u^2}{S_{xx}} \\ \mathbb{V}[\hat{\beta}_0] = \sigma_u^2 \left(\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}} \right) \\ \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}\sigma_u^2}{S_{xx}} \end{array} \right.$$

1.1. Estimación de σ_u^2

Como σ_u^2 es la varianza del término del error $u_i; i = 1, \dots, n$ estos errores representan si bien es cierto no son observables podemos estimar mediante los residuos definidos como:

$$e_i = y_i - \hat{y}_i = \hat{u}_i$$

y representan la parte residual que quedó en las observaciones y_i y no pudo ser explicada por la esperanza condicional estimada

$$\hat{y}_i = \mathbb{E}[\widehat{Y}|X=x_i]$$

Usando estos residuos que se puede ver que cumplen:

1. $\sum_{i=1}^n e_i = 0$

2. $\sum_{i=1}^n e_i x_i = 0$

Esto sugiere como estimador de la varianza usaremos:

$$S_e^2 = \frac{1}{n-2} \sum_{i=1}^n (e_i - \bar{e})^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$$

Además se tiene que S_e^2 cumple que

$$\frac{(n-2)S_e^2}{\sigma_u^2} \sim X_{(n-2)}^2$$

donde $\sum_{i=1}^n e_i^2$ es la suma de cuadrados residual (o del error).

Teorema 1.1: *enemos que las distribuciones muestrales de los estimadores $\hat{\beta}_0, \hat{\beta}_1$ y S_e^2 y son:*

- $\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma_u^2 \left\{ \frac{1}{n} - \frac{\bar{x}}{S_{xx}} \right\} \right)$

- $\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \sigma_u^2 \frac{1}{S_{xx}}\right)$

- $\frac{(n-2)S_e^2}{\sigma_u^2} \sim X_{(n-2)}^2$

2. Estimaciones intervalares de los parámetros del modelo ($\beta_0, \beta_1, \sigma_u^2$)

Para construir un intervalo de confianza para un parámetro θ , necesitamos una 'cantidad pivotal' Q , osea:

1. Una función $Q(x_1, \dots, x_n; \theta)$ de los datos y del parámetro θ
2. la distribución de $Q(x_1, \dots, x_n; \theta)$ no debe depender de θ

Veamos primero β_0 . Tenemos

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

depende de los datos y se tiene que

$$\frac{\hat{\beta}_0 - \beta_0}{\sigma_u \sqrt{\frac{1}{n} - \frac{\bar{x}}{S_{xx}}}}$$

Es 'casi' una cantidad pivotal para β_0 , pero depende de σ_u^2 que **no** se conoce. Por lo tanto debemos estimar previamente σ_u o bien σ_u^2 . **Para eso usaremos como cantidad pivotal:**

$$Q(x_1, \dots, x_n, y_1, \dots, y_n; \beta_0) = \frac{\widehat{\beta}_0 - \beta_0}{\textcolor{red}{S_e} \sqrt{\frac{1}{n} - \frac{\bar{x}}{S_{xx}}}}$$

O es mas conocido escrito como:

$$Q = \frac{\widehat{\beta}_0 - \beta_0}{S_{\widehat{\beta}_0}}$$

Donde $S_{\widehat{\beta}_0}$: Error estándar de $\widehat{\beta}_0$ y $S_{\widehat{\beta}_0}^2$: es la varianza estimada de $\widehat{\beta}_0$

$$S_{\widehat{\beta}_0}^2 = S_e^2 \left\{ \frac{1}{n} - \frac{\bar{x}}{S_{xx}} \right\}$$

Nuestra cantidad pivotal es:

$$Q = \frac{\widehat{\beta}_0 - \beta_0}{\sqrt{S_{\widehat{\beta}_0}^2}} = \frac{\frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\mathbb{V}[\widehat{\beta}_0]}}}{\sqrt{\frac{(n-2)S_{\widehat{\beta}_0}^2}{\mathbb{V}[\widehat{\beta}_0]}}} \sqrt{\frac{n-2}{(n-2)}}$$

Donde el numerador distribuye normal 0,1 y el denominador es la raíz cuadrada de una $X_{(n-2)}^2$ dividido en sus grados de libertad, por lo tanto:

$$Q \sim t(n-2)$$

una t de student con $n-2$ grados de libertad. Para construir un intervalo de confianza para β_0 con confianza γ , buscamos T_1 y T_2 tales que:

$$\mathbb{P}[T_1 \leq \beta_0 \leq T_2] = \gamma$$

Como sabemos que $Q \sim t(n-2)$ existen q_1 y q_2 tales que:

$$\mathbb{P}[q_1 \leq Q \leq q_2] = \gamma$$

Claramente $q_1 = -\tau_{\frac{1+\gamma}{2}(n-2)}$, $q_2 = \tau_{\frac{1+\gamma}{2}(n-2)}$ por lo tanto:

$$\mathbb{P}\left[-\tau_{\frac{1+\gamma}{2}(n-2)} \leq \frac{\widehat{\beta}_0 - \beta_0}{S_{\widehat{\beta}_0}} \leq \tau_{\frac{1+\gamma}{2}(n-2)}\right]$$

Despejando β_0 obtenemos:

$$\mathbb{P}\left[\widehat{\beta}_0 - S_{\widehat{\beta}_0} \tau_{\frac{1+\gamma}{2}(n-2)} \leq \beta_0 \leq \widehat{\beta}_0 + S_{\widehat{\beta}_0} \tau_{\frac{1+\gamma}{2}(n-2)}\right]$$

Por lo tanto el intervalo de confianza γ es:

$$\left[\widehat{\beta}_0 - S_{\widehat{\beta}_0} \tau_{\frac{1+\gamma}{2}(n-2)}; \widehat{\beta}_0 + S_{\widehat{\beta}_0} \tau_{\frac{1+\gamma}{2}(n-2)} \right]$$

Análogamente el intervalo de confianza para β_1 nos da:

$$\left[\widehat{\beta}_1 - S_{\widehat{\beta}_1} \tau_{\frac{1+\gamma}{2}(n-2)}; \widehat{\beta}_1 + S_{\widehat{\beta}_1} \tau_{\frac{1+\gamma}{2}(n-2)} \right]$$

2.1. Intervalo de confianza para la varianza σ_u^2

Sabemos que $S_e^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$ cumple que:

$$\frac{(n-2)S_e^2}{\sigma_u^2} \sim X_{(n-2)}^2$$

Es decir que Q igual a lo de arriba es una cantidad pivotal. Donde los puntos de corte son $X_{\frac{1-\gamma}{2}(r)}^2$

Por lo tanto:

$$\mathbb{P} \left[X_{\frac{1-\gamma}{2}(n-2)}^2 \leq \frac{(n-2)S_e^2}{\sigma_u^2} \leq X_{\frac{1+\gamma}{2}(n-2)}^2 \right] = \gamma$$

Por tanto:

$$\mathbb{P} \left[\frac{(n-2)S_e^2}{X_{\frac{1+\gamma}{2}(n-2)}^2} \leq \sigma_u^2 \leq \frac{(n-2)S_e^2}{X_{\frac{1-\gamma}{2}(n-2)}^2} \right] = \gamma$$

2.2. Prueba de hipótesis

Falta esta clase :c

2.3. Tabla ANOVA , R^2

También falta pero tengo un apunte de esta parte.

Regresión lineal múltiple

1. Reescritura del problema

Queremos extender los procedimientos desarrollados en regresión lineal simple a situaciones en donde se dispone de **mas** de una variable explicativa. Para esto revisaremos el caso de regresión simple, pero desde otra punto de vista.

El modelo era:

$$y_i = \beta_0 + \beta_1 x_i + u_i, i = 1, \dots, n$$

con $u_i \sim N(0, \sigma_u^2)$ I.I.D.

Al obtener los datos $(x_1, y_1), \dots, (x_n, y_n)$ y representarlos gráficamente se tiene una recta que ajusta los puntos.

Reemplazando en el modelo da:

$$y_1 = \beta_0 + \beta_1 x_1 + u_1$$

$$y_2 = \beta_0 + \beta_1 x_2 + u_2$$

$\vdots = \vdots$

$$y_n = \beta_0 + \beta_1 x_n + u_n$$

O bien reescribiéndolo:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} b_0 \\ \vdots \\ b_0 \end{pmatrix} + \begin{pmatrix} b_1 x_1 \\ \vdots \\ b_1 x_n \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$$

Que se puede escribir como:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \beta_0 \mathbf{1} + \beta_1 (x_1, \dots, x_n)^T + (u_1, \dots, u_n)^T$$

Defino $\mathbf{x}_1 = (x_1, \dots, x_n)^T$, $\mathbf{1} = (1, \dots, 1)^T$, $\mathbf{Y} = (y_1, \dots, y_n)^T$ y $\mathbf{u} = (u_1, \dots, u_n)^T$ con $\beta_0, \beta_1 \in \mathbb{R}^2$, que cumple que:

$$\mathbf{Y} = \beta_0 \mathbf{1} + \beta_1 \mathbf{x}_1 + \mathbf{u}$$

o bien:

$$\mathbf{u} = \mathbf{Y} - (\beta_0 \mathbf{1} + \beta_1 \mathbf{x}_1)$$

Es decir cada elección de $(\beta_0, \beta_1) \in \mathbb{R}^2$ nos genera un vector en el subespacio generado por $\mathbf{1}$ y \mathbf{x}_1 y claramente la combinación lineal que este 'mas cerca' en el sentido de que $\|\mathbf{u}\|$ sea mínima, se obtendrá proyectando ortogonalmente \mathbf{Y} sobre este subespacio.

Si recordamos que para $u, v \in \mathbb{R}^n$ se tiene:

$$\|\mathbf{u}\|^2 = \sum_{i=1}^n u_i^2 = \mathbf{u}^T \mathbf{u}$$

Entonces el problema de encontrar la mejor aproximación para \mathbf{Y} desde el subespacio $\langle \mathbf{1}, \mathbf{x}_1 \rangle$ cumple $\mathbf{u} \perp \langle \mathbf{1}, \mathbf{x}_1 \rangle$, es decir que $\forall \beta_0, \beta_1 \in \mathbb{R}^2$ se tiene que:

$$\mathbf{u} \cdot (\beta_0 \mathbf{1} + \beta_1 \mathbf{x}_1) = 0$$

o bien:

$$\mathbf{u} \cdot (\mathbf{1} | \mathbf{x}_1) (\beta_0, \beta_1)^T = 0$$

Es decir se tiene que:

$$\mathbf{u}^T (\mathbf{1} | \mathbf{x}_1) (\beta_0, \beta_1)^T = 0$$

Denotemos por $\mathbf{X} = (\mathbf{1} | \mathbf{x}_1)$, $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ entonces tendremos que:

$$\{ \mathbf{Y} - (\beta_0 \mathbf{1} + \beta_1 \mathbf{x}_1) \}^T \mathbf{X} \boldsymbol{\beta} = 0$$

Entonces:

$$\{ \mathbf{Y} - \mathbf{X} \boldsymbol{\beta} \}^T \mathbf{X} \boldsymbol{\beta} = 0$$

Desarrollando:

$$\{ \mathbf{Y}^T \mathbf{X} - \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \} \boldsymbol{\beta} = 0 \quad \forall \boldsymbol{\beta} \in \mathbb{R}^2$$

Entonces:

$$\mathbf{X}^T \mathbf{Y} = \mathbf{X}^T \mathbf{X} \boldsymbol{\beta}$$

Por tanto si $\mathbf{X}^T \mathbf{X} \in \mathbb{R}^{2 \times 2}$ es invertible entonces:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Es decir los coeficientes que minimizan:

$$\|\mathbf{u}\|^2 = \sum_{i=1}^n u_i^2$$

Cumplen que:

$$\hat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)^T = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Siempre que $\mathbf{X}^T \mathbf{X}$ sea invertible.

Teorema 2.1: $\hat{\boldsymbol{\beta}}$ coincide con los valores de $\widehat{\beta}_0$ y $\widehat{\beta}_1$ del problema de regresión anterior.

Consideremos ahora un modelo con k variables explicativas x_1, \dots, x_k y supongamos que observamos estas k variables mas la variable de respuesta y en cada uno de los n individuos de la muestra, obteniendo así los datos multivariados:

$$(x_{i,1}, \dots, x_{i,k}, y_i)$$

Para $i = 1, \dots, n$ y planteando un modelo lineal:

$$\mathbb{E}[Y | X_1 = x_1, \dots, X_k = x_k] = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

y en consecuencia las observaciones de Y cumplirán:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + u_i, i = 1, \dots, n$$

Con los supuestos:

1. $\mathbb{E}[u_i] = 0, i = 1, \dots, n$

2. $\mathbb{V}[u_i] = \sigma_u^2, i = 1, \dots, n$

3. $\text{cov}(u_i, u_j) = \sigma_u^2 \delta_{i,j}$

4. $u_i \sim \mathcal{N}(0, \sigma_u^2)$

Si definimos los siguientes vectores:

- $Y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$

- $X_j = (x_{1,j}, x_{2,j}, \dots, x_{n,j})^T \in \mathbb{R}^n$ y $j \in \{0, \dots, k\}$

- $u = (u_1, \dots, u_n)^T \in \mathbb{R}^n$

- $\mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}^n$

- $\beta = (\beta_0, \dots, \beta_k) \in \mathbb{R}^{k+1}$

Y con esto tenemos que:

$$Y = \beta_0 \mathbf{1} + \beta_1 X_1 + \dots + \beta_k X_k + u$$

Definimos $X = (\mathbf{1} | X_1 | \dots | X_n) \in \mathbb{R}^{n \times k+1}$ entonces:

$$Y = X\beta + u$$

y

$$y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + u_i, i = 1, \dots, n$$

Con los supuestos:

1. $\mathbb{E}[u_i] = 0, i = 1, \dots, n$

2. $\mathbb{V}[u_i] = \sigma_u^2, i = 1, \dots, n$

3. $\text{cov}(u_i, u_j) = \sigma_u^2 \delta_{i,j}$

4. $u_i \sim \mathcal{N}(0, \sigma_u^2)$

Si definimos los siguientes vectores:

- $Y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$

- $X_j = (x_{1,j}, x_{2,j}, \dots, x_{n,j})^T \in \mathbb{R}^n$ y $j \in \{0, \dots, k\}$

- $u = (u_1, \dots, u_n)^T \in \mathbb{R}^n$

- $\mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}^n$

- $\beta = (\beta_0, \dots, \beta_k) \in \mathbb{R}^{k+1}$

Y con esto tenemos que:

$$Y = \beta_0 \mathbf{1} + \beta_1 X_1 + \dots + \beta_k X_k + u$$

Definimos $X = (\mathbf{1} | X_1 | \dots | X_n) \in \mathbb{R}^{n \times k+1}$ entonces:

$$Y = X\beta + u$$

Donde:

$$\begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \dots & x_{n,k} \end{pmatrix} \in \mathbb{R}^{n \times (k+1)}$$

Observación

1. La parte deterministica del modelo, o sea

$$\beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k}$$

Representa:

$$\mathbb{E}[Y(\nu.a.) | X_1 = x_{i,1}, \dots, X_k = x_{i,k}]$$

2. Esta esperanza condicional hemos supuesto que es 'lineal' con respecto a los coeficientes $\beta_0, \beta_1, \dots, \beta_k$ y no con respecto a los $x_{i,1}, \dots, x_{i,k}$.

3. En este sentido, son modelos de regresión lineal multiple:

- a) $y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$
- b) $y = \beta_0 + \beta_1 \sin(\omega x) + \beta_2 \cos(\omega x) + u$
- c) $\ln y = \beta_0 + \beta_1 e^{\alpha x} + u$

y no son lineales

- a) $y = \beta_0 \sin(\beta_1 x) + u$
- b) $y = \beta_0 + \beta_1 \beta_2 x + u$

4. En algunos casos no es lineal pero aplicando una transformación a los datos se puede convertir en lineal.

Ejemplo 2.2

Sea

$$y = \beta_0 a^{\beta_1 x} u$$

No es lineal pero aplicando logaritmo:

$$\ln y = \ln \beta_0 + \beta_1 x \ln a + \ln u$$

El cual es lineal.

Recordemos que si una variable aleatoria Y posee una distribución normal multivariada $Y \sim \mathcal{N}_n(\mu, \Sigma)$ con $\mu = \mathbb{E}[Y]$ y $\Sigma = \mathbb{V}[Y]$. Donde $\text{diag}(\Sigma) = (\sigma_1, \dots, \sigma_n)^T$ y se cumplen las propiedades:

1. $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$
2. $\alpha Y_i + \beta Y_j \sim \mathcal{N}_1(\alpha \mu_i + \beta \mu_j; \alpha^2 \sigma_i^2 + 2\alpha\beta\sigma_{i,j} + \beta^2 \sigma_j^2)$
3. $Y_j | Y_i = y_i \sim \mathcal{N}(\mu_j + \sigma_{j,i}^2(y_i - \mu_i); \sigma_j^2(1 - \rho_{j,i}))$, donde $\rho_{j,i}$ es la correlación entre Y_j e Y_i dada por:

$$\text{Corr}(Y_i, Y_j) = \text{cov}(Y_i, Y_j) / \sigma_j \sigma_i$$

Notemos que:

$$\mathbb{E}[Y_j | Y_i = y_i] = \mu_j + \frac{\sigma_{j,i}}{\sigma_i^2} (y_i - \mu_i) = \underbrace{\mu_j - \frac{\sigma_{j,i}}{\sigma_i^2} \mu_i}_{\beta_0} + \underbrace{\frac{\sigma_{j,i}}{\sigma_i^2} y_i}_{\beta_1}$$

Es la recta de regresión teórica o de poblaciones de Y_j con respecto a Y_i la cual será estimada mediante $\hat{\beta}_0 + \hat{\beta}_1 y_i$ mediante un modelo de regresión simple.

Volviendo a la normal multivariada $Y \sim \mathcal{N}(\mu, \Sigma)$ se tienen:

1. $X = AY + b$ con $A \in \mathbb{R}^{n \times n}$ y $b \in \mathbb{R}^{n \times 1}$ se tiene que $X \sim \mathcal{N}(A\mu + b; A\Sigma A^T)$

2. Si $Y = \begin{pmatrix} Y^{(1)} \\ Y^{(2)} \end{pmatrix}$ con $Y^{(1)} \in \mathbb{R}^{p \times 1}$ e $Y^{(2)} \in \mathbb{R}^{q \times 1}$ con $p + q = n$, de la misma manera $\mu = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}$, y finalmente:

$$\Sigma = \begin{pmatrix} \Sigma^{(1)} & \vdots & \Sigma^{(1,2)} \\ \dots & & \dots \\ \Sigma^{(2,1)} & \vdots & \Sigma^{(2)} \end{pmatrix}$$

En donde:

- $\mu^{(i)} = \mathbb{E}[Y^{(i)}]$
- $\Sigma^{(1)} = \mathbb{V}[Y^{(1)}] \in \mathbb{R}^{p \times p}$, $\Sigma^{(2)} = \mathbb{V}[Y^{(2)}] \in \mathbb{R}^{q \times q}$
- $\Sigma^{(1,2)} = \text{cov}(Y^{(1)}, Y^{(2)}) \in \mathbb{R}^{p \times q}$ y $(\Sigma^{(1,2)})^T = \Sigma^{(2,1)}$

Por lo tanto $Y^{(1)}$ es independiente con $Y^{(2)}$ si y solo si $\Sigma^{(1,2)} = \mathbf{0}$ por lo tanto si $Y^{(2)} | Y^{(1)} = y_1^{(1)}$ cumple que:

$$Y \sim \mathcal{N}_q \left(\mu^{(2)} + \Sigma^{(2,1)} [\Sigma^{(1)}]^{-1} (y_1^{(1)} - \mu^{(1)}) ; \Sigma^{(2)} - \Sigma^{(2,1)} [\Sigma^{(1)}]^{-1} \Sigma^{(1,2)} \right)$$

Volviendo al modelo lineal:

$$Y = X\beta + u$$

con $u \sim \mathcal{N}_n(\mathbf{0}, \sigma_u^2 Id)$, como u es normal tenemos que Y también es normal, por lo tanto:

$$\mathbb{E}[Y] = X\beta \quad \wedge \quad \mathbb{V}[Y] = \mathbb{V}[u] = \sigma_u^2 Id$$

Por lo tanto $Y \sim \mathcal{N}_n(X\beta, \sigma_u^2 Id)$

2. Recuerdo

Consideraremos una variable respuesta Y y k variables explicativas, x_1, \dots, x_n , queremos encontrar un modelo del tipo:

$$y_i = \beta_0 + \dots + \beta_k x_{i,k} + u_i$$

En donde $i = 1, \dots, n$.

$x_{i,j}$: corresponde a la observación de la j -ésima variable $j = 0, \dots, k$ en el individuo i .

y_i : es el valor de la variable de respuesta en el i -ésimo individuo.

β_j : son los coeficientes del modelo $j = 0, \dots, k$

u_i : error correspondiente a la i -ésima observación. Notemos que los datos son de la forma:

$$(y_i; x_{i,1}, \dots, x_{i,k}); i = 1, \dots, n$$

Definamos:

$$Y = (y_1, \dots, y_n)^T; \beta = (\beta_0, \dots, \beta_k)^T; U = (u_1, \dots, u_n)^T; x = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,k} \\ & \dots & & \\ 1 & x_{n,1} & \dots & x_{n,k} \end{pmatrix}$$

Con $Y, U \in \mathbb{R}^n$, $\beta \in \mathbb{R}^{k+1}$ y $X \in \mathbb{R}^{n \times (k+1)}$.

Usando esta simbología, las expresiones:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{1,1} + \dots + \beta_k x_{1,k} + u_1 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n,1} + \dots + \beta_k x_{n,k} + u_n \end{aligned}$$

Pueden reescribirse como:

$$Y = X\beta + U$$

En donde al vector U se le imponen las siguientes condiciones:

1. $\mathbb{E}[u_i] = 0$ con $i = 1, \dots, n$
2. $\mathbb{V}[u_i] = \sigma_u^2$ con $i = 1, \dots, n$
3. $\text{cov}(u_i, u_j) = \sigma_u^2 \cdot \mathbb{1}_{i=j}$ con $i, j = 1, \dots, n$

O bien matricialmente $u \sim \mathcal{N}_n(0; \sigma_u^2 \cdot Id)$.

Observación: Para encontrar un estimador $\hat{\beta}$ de β no se requieren estas hipótesis, sin embargo para obtener las propiedades de $\hat{\beta}$ y poder efectuar inferencias relativas al parámetro β se necesitan estas hipótesis. Por lo tanto nuestro modelo queda

$$Y = X\beta + U$$

Para encontrar un estimador de β es frecuente utilizar el criterio de mínimos cuadrados, que busca entonces un $\hat{\beta}$ tal que:

$$\sum_{i=1}^n u_i^2$$

sea mínima. Es decir buscamos un $\hat{\beta}$ en \mathbb{R}^{k+1} tal que la función:

$$g(\beta) = \sum_{i=1}^n u_i^2$$

sea mínima. O bien como $u = Y - X\beta$:

$$g(\beta) = \sum_{i=1}^n u_i^2 = \|u\|^2 = \|Y - X\beta\|^2$$

Notemos que esto equivale:

$$g(\beta) = (Y - X\beta)^T (Y - X\beta)$$

Es decir, si miramos esto geométricamente, tenemos:

(mirar gráfico del video)

O sea buscamos $\hat{\beta} \in \mathbb{R}^{k+1}$ que produzca $X\hat{\beta}$ 'mas cercano' a Y , lo que se logra considerando 'la proyección ortogonal' de Y sobre el sub-espacio S generado por X . Por lo tanto

$$Y - X\hat{\beta} \perp X\beta, \forall \beta \in \mathbb{R}^{k+1}$$

de donde entonces resulta que:

$$\langle Y - X\hat{\beta}, X\beta \rangle = 0 \quad \forall \beta \in \mathbb{R}^{k+1}$$

Equivalentemente:

$$\langle X\beta, Y - X\hat{\beta} \rangle = 0$$

o sea

$$(X\beta)^T (Y - X\hat{\beta}) = 0 \implies \beta^T X^T (Y - X\hat{\beta}) = 0 \implies \beta^T X^T Y - \beta^T X^T X\hat{\beta} = 0$$

Luego:

$$\beta^T (X^T Y - X^T X\hat{\beta}) = 0$$

Por lo tanto $X^T Y - X^T X\hat{\beta} = 0$ de donde:

$$X^T X\hat{\beta} = X^T Y$$

A las cuales se le conocen como **ecuaciones normales**, con $k+1$ ecuaciones con $k+1$ incógnitas, si la matriz $X^T X$ es invertible entonces el estimador de mínimos cuadrados de β esta dado por:

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

Observación:

(ver imagen del vídeo)

2.1. Propiedades de $\hat{\beta}$

Sabemos que $U \sim \mathcal{N}_n(0, \sigma_u^2 \cdot Id)$ y que $\hat{\beta} = (X^T X)^{-1} X^T Y$ además $Y = X\beta + u$, por lo tanto:

$$Y \sim N_n(\mathbb{E}[Y], \mathbb{V}[Y])$$

donde $\mathbb{E}[Y] = \mathbb{E}[X\beta + U] = X\beta$, y la varianza de Y se tiene que $\mathbb{V}[Y] = \mathbb{V}[X\beta + U] = \mathbb{V}[U] = \sigma_u^2 Id$, por tanto:

$$Y \sim \mathcal{N}_n(X\beta, \sigma_u^2 Id)$$

de aquí:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \sim \mathcal{N}_{k+1}(\mathbb{E}[\hat{\beta}], \mathbb{V}[\hat{\beta}])$$

Donde $\mathbb{E}[\hat{\beta}] = \mathbb{E}[(X^T X)^{-1} X^T Y] = (X^T X)^{-1} X^T \mathbb{E}[Y] = (X^T X)^{-1} (X^T X)\beta = \beta$ por lo tanto es insesgado.

Por otro lado la varianza es:

$$\begin{aligned} \mathbb{V}[\hat{\beta}] &= \mathbb{V}[(X^T X)^{-1} X^T Y] \\ &= (X^T X)^{-1} X^T \mathbb{V}[Y] (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T (\sigma_u^2 Id) X (X^T X)^{-1} \\ &= \sigma_u^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\ &= \sigma_u^2 (X^T X)^{-1} \end{aligned}$$

Por lo tanto $\hat{\beta} \sim \mathcal{N}_{k+1}(\beta, \sigma_u^2 (X^T X)^{-1})$.

Notemos que la matriz $(X^T X)^{-1}$ no es simple de expresar en función de los valores $x_{i,j}$, por lo que es común reemplazarla por la matriz C, es decir:

$$C = (X^T X)^{-1} = c_{i,j}$$

Claramente $C \in \mathbb{R}^{(k+1) \times (k+1)}$ esto nos permite obtener las distribuciones de los estimadores de cada coeficiente.

Teorema 2.3 (Propiedad): Se tiene que:

$$\hat{\beta}_j \sim \mathcal{N}_1(\beta_j, \sigma_u^2 c_{j+1,j+1}); j = 0, \dots, k$$

2.2. Necesitamos un estimador de σ_u^2

Usando el estimador $\hat{\beta}$ se define el **modelo ajustado** o los valores ajustados por:

$$\hat{Y} = X\hat{\beta}$$

con estos valores ajustados, se define el vector de residuos:

$$e = Y - \hat{Y} = Y - X\hat{\beta}$$

El vector de residuos es un estimador del vector de errores U . Por lo tanto e nos permitirá estimar la varianza de U .

Reescribamos e en la forma siguiente:

$$\begin{aligned} e &= Y - X\hat{\beta} \\ &= Y - X(X^T X)^{-1} X^T Y \\ &= \{Id - X(X^T X)^{-1} X^T\} Y \end{aligned}$$

Notemos que la matriz $X(X^T X)^{-1} X^T$ se presenta también en:

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

Usualmente esta matriz se simboliza por H o sea:

$$H = X(X^T X)^{-1} X^T$$

y se denomina **matriz de proyección**. Por tanto tenemos que:

$$\hat{Y} = X\hat{\beta} = HY$$

Teorema 2.4 (Propiedad): *La matriz H cumple con ser idempotente:*

$$H^2 = H$$

y también $H^T = H$

Demostración 2.1 Ver Clase 2

Observación Notemos que podemos obtener otra proyección sobre el espacio ortogonal a $\{X\beta, \beta \in \mathbb{R}^{k+1}\}$ esta dada por:

$$KY = Y - \hat{Y} = e$$

donde:

$$KY = Y - \hat{Y} = Y - HY = (Id - H)Y$$

Por lo tanto $K = Id - H$

Teorema 2.5 (Propiedad): La matriz K cumple que:

- $K^T = K$
- $K^2 = K$
- $HK = HK = 0$

Demostración 2.2 Ver Clase 2

Observación Las matrices H y K son matrices de proyecciones ortogonales. Ellas permiten 'descomponer' el vector ' Y ' en componentes ortogonales.

Como $e = Y - \hat{Y} = KY = (Id - H)Y$ Podemos obtener las propiedades del vector de residuos.

Teorema 2.6 (Propiedad): e cumple que:

- $\mathbb{E}[e] = 0$
- $\mathbb{V}[e] = \sigma_u^2 K$
- $e \sim \mathcal{N}_n(0, \sigma_u^2 K)$
- $X^T e = 0$

Demostración 2.3 Ver Clase 2

Observación Notemos que $X^T e = 0$ produce las siguientes restricciones lineales:

$$X^T e = \begin{pmatrix} 1 & \dots & 1 \\ x_{1,1} & & x_{n,1} \\ \vdots & & \vdots \\ x_{1,k} & \dots & x_{n,k} \end{pmatrix} \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n e_i \\ \sum_{i=1}^n x_{i,1} e_i \\ \vdots \\ \sum_{i=1}^n x_{i,k} e_i \end{pmatrix}$$

Es decir que::

$$\begin{pmatrix} \sum_{i=1}^n e_i \\ \sum_{i=1}^n x_{i,1} e_i \\ \vdots \\ \sum_{i=1}^n x_{i,k} e_i \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Es decir tengo $k + 1$ restricciones. En otras palabras los residuos e_1, \dots, e_n ya no poseen una dimensionalidad igual a n sino que mas bien igual a $n - (k + 1)$ lo que se denomina sus **grados de libertad**. Con los residuos podemos contruir un estimador de la varianza σ_u^2 de la componente de error en la forma:

$$S_e^2 = \frac{1}{n - (k + 1)} \sum_{i=1}^n e_i^2$$

Observación Se puede mostrar que:

$$\frac{(n - k - 1)S_e^2}{\sigma_u^2} \sim \chi_{(n-k-1)}^2$$

Esta distribución nos proporciona una cantidad pivotal para construir intervalos de confianza para σ_u^2 de la siguiente forma:

$$\left[\frac{(n - k - 1)S_e^2}{\chi_{\frac{1+\gamma}{2}}^2(n - k - 1)}, \frac{(n - k - 1)S_e^2}{\chi_{\frac{1-\gamma}{2}}^2(n - k - 1)} \right]$$

Hasta el momento entonces hemos visto que el estimador mínimo cuadrático $\hat{\beta}$ es lineal en las observaciones y es insesgado. ¿Será $\hat{\beta}$ el mejor estimador con estas propiedades?

Teorema 2.7 (Gauss-Markov): *El estimador mínimo-cuadrático $\hat{\beta}$ de β en el modelo $Y = X\beta + U$ es el estimador lineal insesgado que posee la 'menor' varianza entre todos los estimadores lineales e insesgados de β*

Idea de la demostración una idea de la demostración la puedes encontrar en [el siguiente enlace](#). Queremos ahora analizar un método para evaluar si el modelo ajustado es útil en un sentido global. Para esto consideraremos las sumas de cuadrados que aparecen al hacer el ajuste. Recordemos que:

- $SCT = \sum_{i=1}^n (y_i - \bar{y})^2$
- $SCR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$
- $SCE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$

Veamos que

$$SCE = \sum_{i=1}^n (y_i - \hat{y}_i)r = \sum_{i=1}^n e_i^2$$

Sabemos que $e = KY$ por lo tanto

$$SCE = (KY)^T(KY)$$

Teorema 2.8: *Se puede demostrar que:*

$$\begin{aligned} SCE &= U^T K U \\ &= Y^T Y - \hat{\beta}^T X^T Y \end{aligned}$$

Demostración 2.4 [Ver el siguiente enlace](#)

Por otra parte la suma total de cuadrados $SCT = \sum_{i=1}^n (y_i - \bar{y})^2$ queda igual a:

$$\begin{aligned} SCT &= \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \sum_{i=1}^n \{y_i^2 - 2\bar{y}y_i + \bar{y}^2\} \\ &= \sum_{i=1}^n y_i^2 - 2\bar{y} \sum_{i=1}^n y_i + \sum_{i=1}^n \bar{y}^2 \\ &= \mathbf{Y}^T \mathbf{Y} - n\bar{y}^2 \end{aligned}$$

Finalmente la suma de cuadrados de la regresión es:

$$SCR = \sum_{i=1}^n \{\hat{y}_i - \bar{\hat{y}}\}^2$$

Previamente veamos:

$$\begin{aligned} \bar{\hat{y}} &= \frac{1}{n} \sum_{i=1}^n \hat{y}_i \\ &= \frac{1}{n} \{y_i - e_i\} \\ &= \frac{1}{n} \sum_{i=1}^n y_i - \underbrace{\frac{1}{n} \sum_{i=1}^n e_i}_0 \\ &= \bar{y} \end{aligned}$$

Por lo tanto $\bar{\hat{y}} = \bar{y}$ luego:

$$\begin{aligned} SCR &= \sum_{i=1}^n \{\hat{y}_i - \bar{\hat{y}}\}^2 \\ &= \sum_{i=1}^n \{\hat{y}_i - \bar{y}\}^2 \\ &= \sum_{i=1}^n \hat{y}_i^2 - n\bar{y}^2 \\ &= \mathbf{\hat{Y}}^T \mathbf{\hat{Y}} - n\bar{y}^2 \end{aligned}$$

Por lo tanto $SCR = \mathbf{\hat{Y}}^T \mathbf{\hat{Y}} - n\bar{y}^2$ pero $\hat{y} = X\hat{\beta}$ entonces:

$$\begin{aligned} SCR &= \mathbf{\hat{Y}}^T \mathbf{\hat{Y}} - n\bar{y}^2 \\ &= (X\hat{\beta})^T (X\hat{\beta}) - n\bar{y}^2 \\ &= \hat{\beta}^T X^T X \hat{\beta} - n\bar{y}^2 \\ &= \hat{\beta}^T X^T X (X^T X)^{-1} X^T Y - n\bar{y}^2 \\ &= \hat{\beta}^T X^T Y - n\bar{y}^2 \end{aligned}$$

Por lo tanto $SCR = \hat{\beta}^T X^T Y = n\bar{y}^2$. En resumen:

- $SCE = Y^T Y - \hat{\beta}^T X^T Y$
- $SCR = \hat{\beta}^T X^T Y - n\bar{y}^2$
- $SCT = Y^T Y - n\bar{y}^2$

por lo tanto:

$$SCT = SCR + SCE$$

Observación Esta descomposición nos permite cuantificar que parte de SCT logra explicarse con el modelo de regresión ajustado. Es más frecuente indicar la 'proporción' de SCT explicada; para esto se define **el coeficiente de determinación** R^2 , mediante:

$$R^2 = \frac{SCR}{SCT}$$

Que corresponde a la **proporción explicada**. Claramente $0 \leq R^2 \leq 1$ y un valor cercano a 1 representa un 'buen' ajuste.

Notemos que si tenemos una muestra de tamaño n y aumentamos el numero de variables explicativas 'independientes' k hasta que

$$n = k + 1$$

en tal caso la matriz $X \in \mathbb{R}^{n \times (k+1)}$ es cuadrada e invertible. Por lo tanto

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

Por tanto $e = Y - \hat{Y} = Y - X(X^{-1} Y) = 0$ Por lo tanto $SCE = e^T e = 0$ y $SCT = SCR + 0$ entonces:

$$R^2 = 1$$

Por otra parte R^2 es una función creciente del número de variables k , en consecuencia, no es capaz de detectar cuando el aporte a SCR que hace una determinada variable es útil o no.

Para solucionar este problema se define el **coeficiente de determinación ajustado** o corregido mediante lo siguiente, pero antes notemos que:

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

con esto se define

$$R_a^2 = 1 - \frac{\frac{SCE}{n-k-1}}{\frac{SCT}{n-1}}$$

Ya que se puede ver que las sumas de cuadrados poseen las siguientes distribuciones muéstrales:

- $\frac{n-k-1}{\sigma_u^2} SCE \sim \chi^2_{n-k-1}$

Fuente	Suma de cuadrados	Grados de libertad	Suma de cuadrados media	Razón F
Regresión	SCR	k	$MCR = \frac{SCR}{k}$	$F^* = \frac{MCR}{MCE}$
Error	SCE	n-k-1	$MCE = \frac{SCE}{n-k-1} = S_e^2$	
Total	SCT	n-1	$MCT = \frac{SCT}{n-1} = S_Y^2$	

- $\frac{k}{\sigma_u^2} SCR \sim \chi_k^2$
- $\frac{n-1}{\sigma_u^2} SCT \sim \chi_{n-1}^2$

Esto permite definir la **razón F** en la forma:

$$F = \frac{\frac{SCR}{k}}{\frac{SCE}{n-k-1}} \sim F_{k, n-k-1}$$

La cual nos facilita la prueba de hipótesis siguiente:

$$\begin{cases} H_0 : \beta_1 = \dots = \beta_k = 0 \\ H_1 : \text{Algún o algunos coeficientes } \beta_j; j = 1, \dots, k \text{ son no nulos} \end{cases}$$

El criterio de decisión es:

$$\text{Si } F^* = \frac{\frac{SCR}{k}}{\frac{SCE}{n-k-1}} > F_{1-\alpha}(k, n-k-1) \text{ entonces se rechaza } H_0 \text{ al nivel } \alpha$$

Observación Esta prueba de hipótesis nos permite evaluar la significancia **global** del modelo, es decir, si la incorporación de estas variables logra explicar el comportamiento medio de la respuesta o no. Usualmente la información necesaria para realizar esta prueba se representa por medio de una **tabla de análisis de varianza**

Observación En el caso en que se rechace la hipótesis nula significa que todos o algunos son no nulos por lo tanto se requiere efectuar pruebas de significancia individuales.

2.3. Prueba t para la significancia de un coeficiente específico

Queremos probar la hipótesis

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0, \quad j = 1, \dots, k \end{cases}$$

Suponiendo dado α el nivel de significancia, tenemos:

$$\hat{\beta}_j \sim \mathcal{N}_1(\beta_j; \sigma_u^2 C_{j+1,j+1})$$

En donde $C = (C_{i,j}) = (X^T X)^{-1}$. Además σ_u^2 lo estimamos mediante $S_e^2 = MCE = \frac{SCE}{n - k - 1}$. De aquí obtenemos:

$$\frac{\hat{\beta}_j - \beta_j}{\sigma_u^2 \sqrt{C_{j+1,j+1}}} \sim \mathcal{N}(0, 1)$$

y

$$\frac{\hat{\beta}_j - \beta_j}{S_e \sqrt{C_{j+1,j+1}}} \sim \text{t}(n - k - 1)$$

Con esto tenemos que el criterio de decisión queda como:

$$\begin{aligned} \text{Si } \left| \frac{\hat{\beta}_j - \beta_j}{S_e \sqrt{C_{j+1,j+1}}} \right| > t_{1-\frac{\alpha}{2}}(n - k - 1) \text{ entonces se rechaza} \\ \left\{ \begin{array}{l} H_0: \beta_j = 0 \\ V/S \\ H_1: \beta_j \neq 0 \end{array} \right. \\ \text{al nivel } \alpha \end{aligned}$$

Observación Al concluir que $\beta_j = 0$, antes de evaluar la significancia de otros coeficientes, se debe reestimar el modelo excluyendo la variable X_j . En algunas ocasiones se requiere realizar pruebas de hipótesis como las siguientes:

$$\left\{ \begin{array}{l} H_0: \left\{ \begin{array}{l} \beta_2 = 3\beta_1 - \beta_3 \\ \beta_5 = \beta_2 + \beta_1 \end{array} \right. \\ H_1: \text{distintos} \end{array} \right.$$

Esto lo podemos analizar de dos formas:

- Resolver el sistema de ecuaciones lineales planteado en H_0 para así eliminar algunas de las variables, quedando un modelo con menos variables. En nuestro caso nos queda:

$$\left\{ \begin{array}{l} \beta_2 = 3\beta_1 - \beta_3 \\ \beta_5 = (3\beta_1 - \beta_3) + \beta_1 = 4\beta_1 - \beta_3 \end{array} \right.$$

Por lo tanto:

$$\begin{aligned} y_i &= \beta_0 + x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5} + u_i \\ &= \beta_0 + \beta_1 x_{i,1} + (3\beta_1 - \beta_3)x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + (4\beta_1 - \beta_3)x_{i,5} + u_i \\ &= \beta_0 + \beta_1(x_{i,1} + 3x_{i,2} + 4x_{i,5}) + \beta_3(-x_{i,2} + x_{i,3} - x_{i,5}) + \beta_4 x_{i,4} + u_i \end{aligned}$$

Entonces reescribiendo los parámetros:

$$y_i = \gamma_0 + \gamma_1 z_{i,1} + \gamma_2 z_{i,2} + \gamma_3 z_{i,3} + u_i$$

2. Expresando matricialmente la hipótesis nula, sea $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^T$ y:

$$R = \underbrace{\begin{pmatrix} 0 & 3 & -1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & -1 \end{pmatrix}}_{2 \times 6} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 \\ 0 \end{pmatrix}}_r$$

Es decir:

$$\begin{cases} H_0 : R\beta = r \\ H_1 : R\beta \neq r \end{cases}$$

Con $r = (0, 0)^T$ (podría ser distinto al origen), si la hipótesis nula involucrara ecuaciones no homogéneas), en general, tendriamos:

$$\begin{cases} H_0 : R\beta = r; R \sim 1 \times (k+1), \beta \sim (k+1) \times 1, r \sim q \times 1 \\ H_1 : R\beta \neq r \end{cases}$$

Asumiendo que el rango de R es q (Las relaciones lineales en H_0 son L.I.).

Sabemos que:

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

y que $\hat{\beta} \sim \mathcal{N}_{k+1}(\beta, \sigma_u^2 (X^T X)^{-1})$ por lo tanto:

$$R\hat{\beta} \sim \mathcal{N}(R\beta; \sigma_u^2 (R(X^T X)^{-1} R^T))$$

Entonces $R\hat{\beta} - R\beta \sim \mathcal{N}_q(0; \sigma_u^2 (R(X^T X)^{-1} R^T))$, también sabemos que el estimador de σ_u^2 es $S_e^2 = \frac{(n-k-1)}{\sigma_u^2} SCE \underset{H_0}{\sim} \chi^2(n-k-1)$ que sigue dicha distribución bajo H_0 verdadera. Entonces:

$$R\hat{\beta} - r \sim \mathcal{N}_q(0; \sigma_u^2 R(X^T X)^{-1} R^T) \implies (R\hat{\beta} - r)^T \{ \sigma_u^2 R(X^T X)^{-1} R^T \}^{-1} (R\hat{\beta} - r) \sim \chi^2(q)$$

Por lo tanto:

$$F^* = \frac{\frac{(R\hat{\beta} - r)^T \{ \sigma_u^2 R(X^T X)^{-1} R^T \}^{-1} (R\hat{\beta} - r)}{q}}{\frac{n-k-1}{\sigma_u^2} SCE} \sim F(q, n-k-1)$$

Por lo tanto si $F^* > F_{1-\alpha}(q, n-k-1)$ se rechaza $H_0 : R\beta = r$ en favor de $H_1 : R\beta \neq r$ al nivel α

Observación Notemos que en el caso que no podamos rechazar H_0 , debemos asumir (mientras no aparezca nueva información) que la relación $R\beta = r$ es cierta y en consecuencia, tendremos que estimar β incorporando esta relación, como lo hicimos al estimar β_2 y β_5 en el ejemplo visto. Otra forma es efectuar la estimación de β , pero incorporando la restricción, o sea:

$$\begin{cases} \min_{\beta} (Y - X\beta)^T (Y - X\beta) \\ s.a. \quad R\beta = r \end{cases}$$

Para esto, definamos el **Lagrangiano**:

$$L(\beta, \lambda) = (Y - X\beta)^T (Y - X\beta) - 2\lambda^T (R\beta - r)$$

con $\lambda \in \mathbb{R}^q$ o bien:

$$\begin{aligned} L(\beta, \lambda) &= Y^T Y - \beta^T X^T Y - Y^T X\beta + \beta^T X^T X\beta - 2\lambda^T (R\beta - r) \\ &= Y^T Y - 2Y^T X\beta + \beta^T X^T X\beta - 2\lambda^T (R\beta - r) \end{aligned}$$

debemos buscar un $\widehat{\beta}_R$ que produzca $\nabla L(\widehat{\beta}_R, \lambda) = 0$

Teorema 2.9: Sea $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ una función lineal de la forma $f(x) = Ax$ con $A \in \mathbb{R}^{m \times n}$ entonces:

1. $\frac{\partial (Ax)}{\partial x} = A$
2. $\frac{\partial (x^T Ax)}{\partial x} = -(A + A^T)x$

Usando el teorema tenemos que

ver clase 4

lo que da la solución:

$$\widehat{\beta}_R = \widehat{\beta} + (X^T X)^{-1} R^T \{R(X^T X)^{-1} R^T\}^{-1} R (\widehat{\beta}_R - \widehat{\beta})$$

Pero sabemos que $\widehat{\beta}_R$ cumple que $R\widehat{\beta}_R = r$ lo que nos da:

$$\widehat{\beta}_R = \widehat{\beta} + (X^T X)^{-1} R^T \{R(X^T X)^{-1} R^T\}^{-1} (r - R\widehat{\beta})$$

Teorema (Propiedades): $\widehat{\beta}_R$ cumple que:

1. $R\widehat{\beta}_R = r$
2. Como $\widehat{\beta}$ sigue la ley normal entonces $\widehat{\beta}_R$ sigue la ley normal.
3. $\widehat{\beta}_R$ es insesgado, si y sólo si $R\widehat{\beta} = r$
4. $\mathbb{V}[\widehat{\beta}_R]$ es 'menor' que $\mathbb{V}[\widehat{\beta}]$, pues: $\mathbb{V}[\widehat{\beta}] - \mathbb{V}[\widehat{\beta}_R]$ produce una matriz positiva semi-definida.

3. Suma de cuadrados extra

Un aspecto recurrente en el análisis de regresión, es decidir sobre la incorporación a un modelo de otro grupo de variables, para lo cual deberemos analizar si el aporte adicional que ellas hacen a la suma de cuadrados de la regresión es o no significativo.

Para esto consideremos:

$$\left\{ \begin{array}{l} \text{Modelo } I: y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + u_i \\ \text{Modelo } II: y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \underbrace{\beta_k x_{i,k} + \dots + \beta_{k+l} x_{i,k+l}}_{l \text{ variables adicionales}} + u_i \end{array} \right.$$

Definamos:

- $Y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$
- $\beta_I = (\beta_0, \dots, \beta_k)^T \in \mathbb{R}^{k+1}$
- $\beta_{II} = (\beta_0, \dots, \beta_k, \beta_{k+1}, \dots, \beta_{k+l})^T \in \mathbb{R}^{k+l+1}$
- $U = (u_1, \dots, u_n)^T \in \mathbb{R}^n$
- $X_I = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \dots & x_{n,k} \end{pmatrix} \in \mathbb{R}^{n \times (k+1)}$
- $X_{II} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,k} & x_{k+1} & \dots & x_{1,k+l} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \dots & x_{n,k} & x_{n,k+1} & \dots & x_{n,k+l} \end{pmatrix} \in \mathbb{R}^{n \times (k+l+1)}$

con esto los modelos quedan:

$$\left\{ \begin{array}{l} \text{Modelo } I: Y = X_I \beta_I + U \\ \text{Modelo } II: Y = X_{II} \beta_{II} + U \end{array} \right.$$

de donde los vectores de estimadores en ambos modelos son:

$$\left\{ \begin{array}{l} \widehat{\beta}_I = (X_I^T X_I)^{-1} X_I^T Y \\ \widehat{\beta}_{II} = (X_{II}^T X_{II})^{-1} X_{II}^T Y \end{array} \right.$$

Y ellos generan las siguientes sumas de cuadrados:

- Suma de cuadrados de regresión $\left\{ \begin{array}{l} SCR_I = \widehat{\beta}_I^T X_I^T Y \\ SCR_{II} = \widehat{\beta}_{II}^T X_{II}^T Y \end{array} \right.$

- Suma de cuadrados del error $\left\{ \begin{array}{l} SCE_I = Y^T Y - \widehat{\beta}_I^T X^T Y \\ SCE_{II} = Y^T Y - \widehat{\beta}_{II}^T X_{II}^T Y \end{array} \right.$
- Suma de cuadrados de total $\left\{ \begin{array}{l} SCT_I = SCR_I + SCE_I \\ SCT_{II} = SCR_{II} + SCE_{II} \end{array} \right.$

podemos visualizar esto en un diagrama, en la forma siguientes (ver [aquí](#))

En donde se ve que la región en rojo es la que debemos evaluar si es o no significativa.

Sabemos que si $U \sim \mathcal{N}_n(0, \sigma_u^2 Id_n)$ entonces:

- Error $\left\{ \begin{array}{l} \frac{n-k-1}{\sigma_u^2} SCE_I \sim \chi^2(n-k-1) \\ \frac{n-k-l-1}{\sigma_u^2} SCE_{II} \sim \chi^2(n-k-l-1) \end{array} \right.$
- Regresión $\left\{ \begin{array}{l} \frac{k}{\sigma_u^2} SCR_I \sim \chi^2(k) \\ \frac{k+l}{\sigma_u^2} SCR_{II} \sim \chi^2(k+l) \end{array} \right.$
- Total $\left\{ \frac{n-1}{\sigma_u^2} SCT \sim \chi^2(n-1) \right.$

Se define **la suma de cuadrados extra de regresión del modelo II dado el modelo I** por:

$$SCR(II/I) = SCR_{II} - SCR_I$$

se puede ver que:

$$\frac{l}{\sigma_u^2} SCR(II/I) \sim \chi^2(l)$$

Esto nos lleva a considerar la siguiente razón F^* :

$$F^* = \frac{\frac{SCR(II/I)}{l}}{\frac{SCE_{II}}{(n-k-l-1)}} \sim F(l, n-k-l-1)$$

Con lo cual podemos probar la hipótesis:

$$\left\{ \begin{array}{l} H_0 : \beta_{k+1} = \dots = \beta_{k+l} = 0 \\ H_1 : \text{Alguna(s) } \beta_j \neq 0 \text{ para algun(os) } j = k+1, \dots, k+l \end{array} \right.$$

Que dado un nivel de significancia α , genera el siguiente criterio de decisión:

Si $F^* > F_{1-\alpha}(l, n - k - l - 1)$ entonces se rechaza H_0 en favor de H_1 con un nivel de significancia α

Observación Equivalentemente, el valor-p de esta prueba es $\mathbb{P}[F > F^*]$ en donde F^* es el valor empírico obtenido de la razón y F es una variable aleatoria con distribución $F(l, n - k - l - 1)$. **Observación**

1. Típicamente la aplicación de esta prueba es en el caso $l = 1$, osea queremos analizar la significancia de la incorporación de **una** variable adicional en tal caso se tiene:

$$\text{Si } F^* > F_{1-\alpha}(1, n - k - 1 - 1) \text{ entonces rechazamos } \begin{cases} H_0 : \beta_{k+1} = 0 \\ H_1 : \beta_{k+1} \neq 0 \end{cases}$$

Pero notemos que $F(1, r) = t^2(r)$, por lo tanto este criterio queda de la forma equivalente:

$$\text{Si } |t^*| > t_{1-\frac{\alpha}{2}}(n - k - 1 - 1) \text{ entonces rechazamos } \begin{cases} H_0 : \beta_{k+1} = 0 \\ H_1 : \beta_{k+1} \neq 0 \end{cases} \text{ en donde } t^* = \sqrt{F^*}$$

2. Este criterio de **suma de cuadrados extra** es la base de los procedimientos **automáticos** de ajuste de modelos de regresión:

El primero, conocido como **forward**, va introduciendo variables una a una, según si el aporte que hacen es o no significativo.

Para analizar cual variable sera la primera en incorporarse se calculan las correlaciones de la respuesta con las variables explicativas Y se selecciona la que presente mayor correlación absoluta.

Para seleccionar la segunda variable a incorporar, se obtienen los residuos generados a partir del modelo con la primera variable elegida y se calculan las correlaciones de este vector de residuos, con las variables explicativas restantes eligiendo aquella que produzca la mayor correlación absoluta.

El segundo procedimiento, es partir con un modelo que contiene todas las variables explicativas que disponemos e ir **eliminando** las variables una a una, según si su aporte no es significativo, este esquema se conoce como **backward**.

La combinación de ambos métodos se le llama **stepwise**.

4. Predicciones en un modelo de regresión

Habiendo ya estimado los parámetros del modelo $Y = X\beta + U$ con $U \sim \mathcal{N}(0, \sigma_u^2 Id_n)$ y evaluando su bondad de ajuste, queremos usar este modelo ajustado $\hat{Y} = X\hat{\beta}$ para obtener **predicciones** de la variable respuesta. Distinguiremos dos tipos de predicciones:

1. Predicción de la respuesta esperada.
2. Predicción de **un** valor individual

Si tenemos k variables explicativas X_1, \dots, X_k y sera x_1^*, \dots, x_k^* los valores de ellas, para los cuales queremos obtener predicciones.

Sea $x^* = (1, x_1^*, x_2^*, \dots, x_k^*)$, por lo tanto queremos encontrar estimaciones de:

1. $\mathbb{E}[Y|X_1 = x_1^*, \dots, X_k = x_k^*]$
2. $Y|X_1 = x_1^*, \dots, X_k = x_k^*$ (una realización).

Con las hipótesis habituales $U \sim \mathcal{N}_n(0, \sigma_u^2 Id_n)$ y $X^T X$ invertible tenemos $\hat{\beta} = (X^T X)^{-1} X^T Y$, con esto obtenemos:

1. Estimación puntual de $\mathbb{E}[Y|X_1 = x_1^*, \dots, X_k = x_k^*]$ esta dada por:

$$\mathbb{E}[\widehat{Y|X^* = x^*}] = (x^*)^T \hat{\beta}$$

en donde $X^* = (1, X_1, \dots, X_k)^T$ o bien mas explícitamente:

$$\mathbb{E}[Y|X_1 = \widehat{x_1^*}, \dots, X_k = x_k^*] = (x^*)^T \hat{\beta}$$

Por otra parte para entregar una estimación intervalar necesitamos conocer la esperanza y la varianza de este estimador.

Observación Notemos que el modelo ajustado $\hat{Y} = X\hat{\beta}$ produce las filas $(\hat{y}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \dots + \hat{\beta}_k x_{i,k}; i = 1, \dots, n$, si ahora consideramos los valores x_1^*, \dots, x_k^* para las variables, tendríamos el valor **predicho**

$$\begin{aligned}\hat{y}_i^* &= \hat{\beta}_1 + \hat{\beta}_1 x_1^* + \dots + \hat{\beta}_k x_k^* \\ &= (1, x_1^*, \dots, x_k^*)(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)^T \\ &= (x^*)^T \hat{\beta}\end{aligned}$$

Como $\hat{\beta}$ es normal claramente el estimador $\mathbb{E}[\widehat{Y|X^* = x^*}]$ también lo es, por tanto:

$$\mathbb{E}[(x^*)^T \hat{\beta}] = (x^*)^T \mathbb{E}[\hat{\beta}] = (x^*)^T \beta$$

Notemos que en $Y = X\beta + U$ si consideramos $Y|X^* = x^*$ tendremos:

$$Y|X^* = x^* = (x^*)^T \beta + u$$

cuya esperanza es:

$$\mathbb{E}[Y|X^* = x^*] = \mathbb{E}[(x^*)^T \beta + u] = (x^*)^T \beta$$

Por otro lado notemos que:

$$\begin{aligned} \mathbb{V}[(x^*)^T \hat{\beta}] &= (x^*)^T \mathbb{V}[\hat{\beta}](x^*) \\ &= (x^*)^T \{\sigma_u^2 (X^T X)^{-1}\}(x^*) \\ &= \sigma_u^2 (x^*)^T (X^T X)^{-1} (x^*) \end{aligned}$$

Con esto y usando que:

$$S_e^2 = \frac{1}{n-k-1} \sum_{i=1}^n e_i^2 = \frac{Y^T Y - \hat{\beta}^T X^T Y}{n-k-1}$$

es un estimador de σ_u^2 , podemos construir una estimación intervalar para $\mathbb{E}[Y|X^* = x^*]$ en la forma:

$$\left[(x^*)^T \hat{\beta} \pm S_e \sqrt{(x^*)^T (X^T X)^{-1} (x^*)} t_{\frac{n-k-1}{2}} \right]$$

2. Para obtener la predicción de un valor individual necesitamos conocer como se distribuye el error de predicción o sea:

$$\underbrace{e|X^* = x^*}_{\text{Residuo}} = Y|X^* = x^* - \widehat{Y|X^* = x^*}$$

Es decir

$$e|_{X^* = x} = \{(x^*)^T \beta + u|_{X^* = x}\} - (x^*)^T \hat{\beta}$$

Por lo tanto la varianza es:

$$\mathbb{V}[e|X^* = x^*] = \sigma_u^2 \left\{ 1 + (x^*)^T (X^T X)^{-1} (x^*) \right\}$$

Para ver desarrollo ver clase 6.

Con esto obtenemos que la varianza de $Y|X^* = x$ es:

$$\mathbb{V}[Y|X^* = x^*] = \sigma_u^2 \left\{ 1 + (x^*)^T (X^T X)^{-1} (x^*) \right\}$$

de donde la estimación intervalar para $Y^*|X^* = x^*$ es:

$$\left[(x^*)^T \hat{\beta} \pm S_e \sqrt{1 + (x^*)^T (X^T X)^{-1} (x^*)} t_{\frac{n-k-1}{2}} \right]$$

5. Mínimos cuadrados ponderados

Una de las hipótesis que hemos asumido es que $\mathbb{V}[u_1] = \dots = \mathbb{V}[u_n] = \sigma_u^2$ (**homoscedasticidad**) en muchos casos reales, esta hipótesis no se cumple, por lo que debemos entender el modelo para permitir el análisis de tales situaciones.

Supongamos que las varianzas son distintas $\mathbb{V}[u_i] = \sigma_i^2; i = 1, \dots, n$ y con esperanza 0 todas, por lo tanto:

$$U \sim N_n(0, \Sigma); \Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$$

Que las varianzas sean distintas se le conoce como **heteroscedasticidad**.

En el modelo $Y = X\beta + U$ consideremos una matriz $A \sim \mathbb{R}^{n \times n}$ y multipliquemos el modelo por A es decir $AY = AX\beta + AU$ y definamos $V = AU$.

¿Cómo seleccionar A para que $\mathbb{V}[V] = \sigma_v^2 Id_n$?

Claramente $\mathbb{E}[V] = \mathbb{E}[AU] = \mathbf{0}$ luego

$$\mathbb{V}[V] = \sigma_v^2 Id_n \implies A\Sigma A^T = \sigma_v^2 Id_n$$

Luego $A = \sigma_v \begin{pmatrix} \sigma_1^{-1} & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & \sigma_n^{-1} \end{pmatrix}$, si $AY = AX\beta + Au$ escribimos $Z = AY, W = AX, V = AU$ nos queda un nuevo modelo:

$$Z = W\beta + V$$

con $V \sim \mathcal{N}_n(0, \sigma_v^2 Id_n)$, por lo tanto el estimador de $\hat{\beta}$ queda como:

$$\hat{\beta} = (W^T W)^{-1} W^T Z$$

y como $\Sigma^{-1} = A^T A$ se tiene que:

$$\hat{\beta}_p = \{X^T \Sigma^{-1} X\}^{-1} X^T \Sigma^{-1} Y$$

Es el estimador de mínimos cuadrados ponderado de β .

Se habla de mínimos cuadrados ponderados pues $\hat{\beta}_p$ minimiza $\sum_{i=1}^n \frac{1}{\sigma_i^2} u_i^2$. Notemos que:

1. $\hat{\beta}_p$ es normal

2. $\mathbb{E}[\hat{\beta}_p] = \beta$, por tanto es insesgado.

3. $\mathbb{V}[\hat{\beta}_p] = (X^T \Sigma^{-1} X)^{-1}$

Puede ver el argumento de estar propiedades en [clase 7. Observación](#) En el caso general de que la matriz de varianza de U : $\Sigma = \mathbb{V}[U]$ no sea diagonal, o sea, existen asociaciones no nulas entre

distintos u_i ; $i = 1, \dots, n$ como Σ es semi-definida positiva y simétrica, existe una matriz simétrica e invertible P tal que

$$\Sigma = P^T P \text{ (cholesky)}$$

Y entonces $\widehat{\beta}_p = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$ puede obtenerse en función de P , en este caso se habla de **mínimos cuadrados generalizados**

Correlaciones

1. Introducción

El análisis de correlaciones pretende detectar, clasificar y cuantificar el grado de asociación lineal que existe entre:

1. Una variable respuesta Y y una variable explicativa X , esta se denomina **coeficiente de correlación simple**, que se denota por $r_{Y;X}$
2. Una variable respuesta Y y varias variables explicativas ' X_1, \dots, X_k' , esta se denomina **coeficiente de correlación múltiple** y se denota por $r_{Y;X_1, \dots, X_n}$
3. Una variable de respuesta Y y algunas variables explicativas X_1, \dots, X_k eliminando previamente el efecto de asociación que existe de otras variables explicativas X_{k+1}, \dots, X_{k+l} , esto se denomina **coeficiente de correlación parcial entre 'Y' y ' X_{k+1}, \dots, X_{k+l} '** y se denota por $r_{Y;X_1, \dots, X_k, X_{k+1}, \dots, X_{k+l}}$
4. Varias variables de respuesta Y_1, \dots, Y_r y varias variables explicativas X_1, \dots, X_k , esto se denomina **coeficiente de correlación canónico**

Observación

1. Todos los coeficientes fluctúan entre -1 y 1 .

2. Los cuadrados de estos coeficientes se denominan **coeficiente de determinación, simple, múltiple, parcial y canónico** y se simbolizan por:

- $R_{Y;X}^2$
- $R_{Y;X_1,\dots,X_k}^2$
- $R_{Y;X_1,\dots,X_k,X_{k+1},\dots,X_{k+l}}^2$
- $R_{Y_1,\dots,Y_r;X_1,\dots,X_K}^2$

Estas correlaciones pueden visualizarse geométricamente como se ve en clase 7.

Se puede ver que cumple la siguiente relación:

Teorema: Se tiene que:

$$1 - R_{Y;x_1, X_2}^2 = (1 - R_Y^2; X_2 . X_1)(1 - R_{Y;X_1}^2)$$

Observación Notemos la analogía con:

$$\mathbb{P}[C \cap (A \cap B)] = \mathbb{P}[C \cap B / A] \mathbb{P}[A]$$