# REGRESION MULTIPLE Y OTRAS TÉCNICAS MULTIVARIADAS

## ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

Antonio Soriano Flores

[1] UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS
Y EN SISTEMAS

asoriano@sigma.iimas.unam.mx

7 de agosto de 2016

# Temario

# Objetivos y Evaluación

- **Objetivo**: Comprender el alcance del análisis de regresión así como los supuestos que debe de cumplir el modelo establecido. El alumno debe de ser capaz de ajustar un modelo de regresión, evaluando y mejorando el desempeño del mismo. El estudiante deberá aprender a usar algún paquete estadístico e interpretar los resultados obtenidos.
- **Importante**: Haber cursado Inferencia Estadística

# Modelos Matemáticos

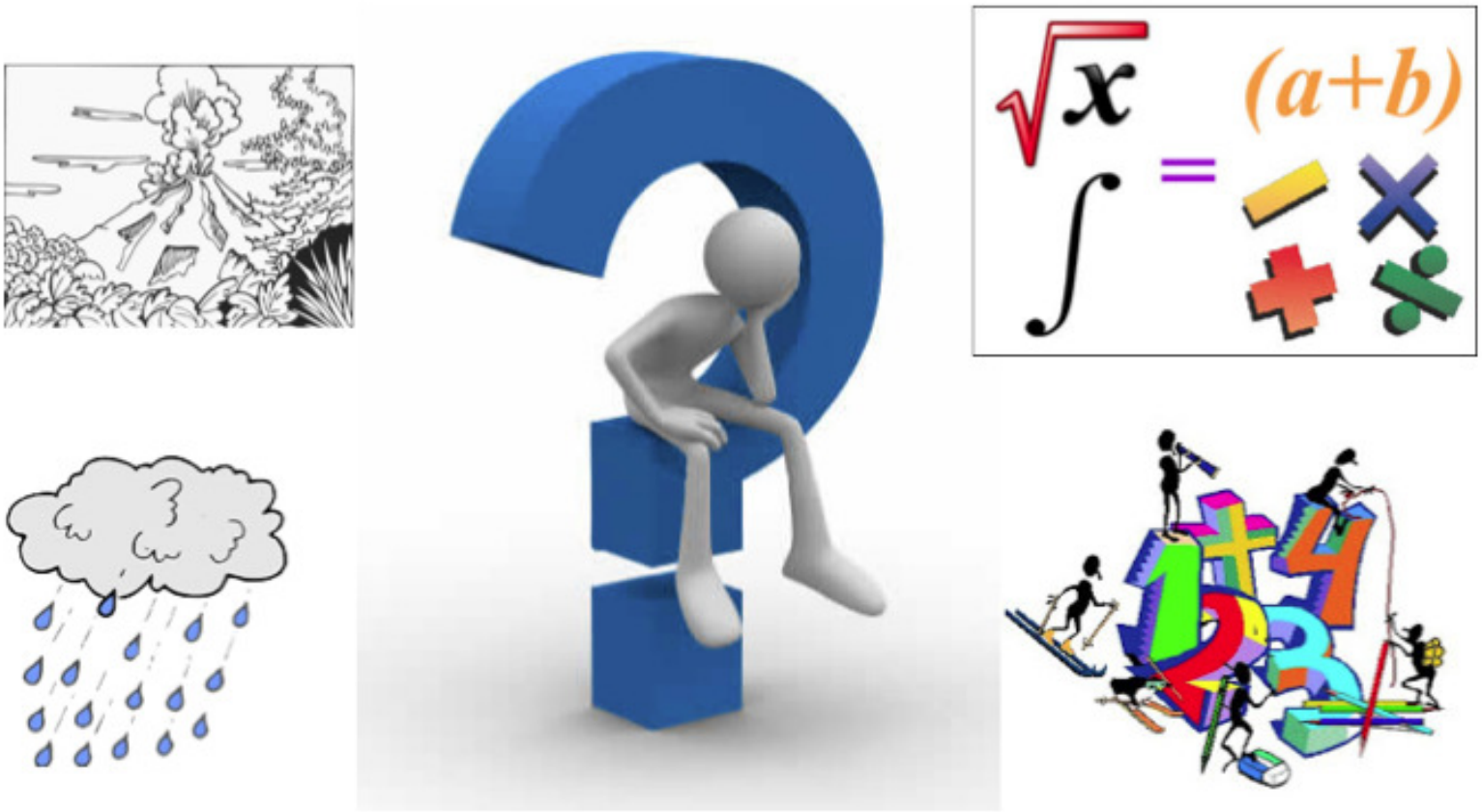



Figura: Modelación Matemática

# Modelos Matemáticos

Para poder modelar un fenómeno contamos con modelos deterministas y modelos probabilisticos (aleatorios)

- **Deterministas**: Las mismas entradas producirán invariablemente las mismas salidas, no contemplándose la existencia del azar ni el principio de incertidumbre.

$$F = ma; \quad e = mc^2; \quad d = \frac{1}{2}gt^2$$

F=Fuerza; e=Energia; d=Distancia
m=Masa; c=Velocidad de la Luz; t=Tiempo
a=Aceleración; g=Gravedad

# Modelos Matemáticos

- **Probabilistas**: Las mismas entradas no necesariamente producirán la mismas salidas pues contempla la interacción con variables que no podemos controlar (azar).
  Sea $X$ el número de águilas en 6 lanzamientos independientes de una moneda "honesta", se propone el modelo:

$$X \backsim Bin(6, 0.5); \quad \mathbb{E}(X) = 3; \quad \text{Var}(X) = 1.5$$

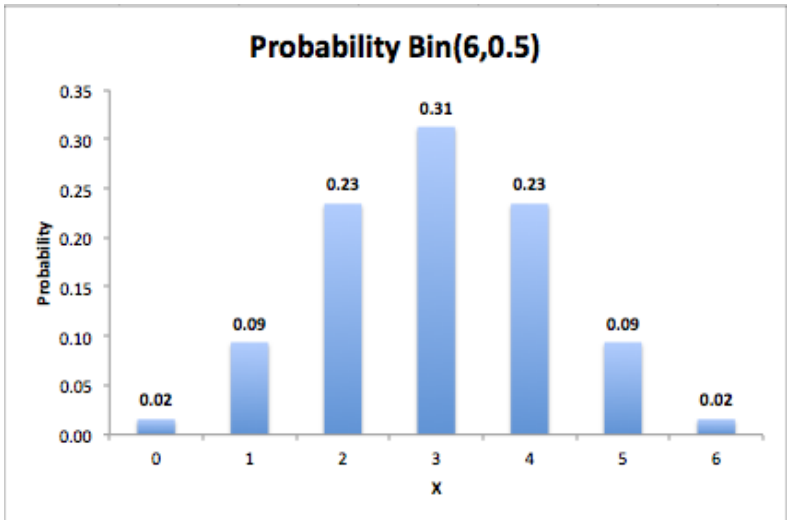



Figura: Binomial Density

# Modelos Matemáticos

En general un modelo busca encontrar una relación funcional entre ciertas variables independientes $(z_1, z_2, ...., z_q)$ y una variable respuesta $Y$ la cual queremos explicar o modelar.

$$Y = f(z_1, z_2, ...., z_q) \quad (Determinista)$$

$$Y = f(z_1, z_2, ...., z_q) + \varepsilon \quad (Probabilistico)$$

- ¿Qué función $f$ tomar?
- ¿El modelo es adecuado? ¿Existe el mejor modelo?
- ¿Qué supuestos debemos tomar en cuenta?
- ¿Las variables $z_i$ si explican a $Y$? ¿Está respaldado con buenos datos?

# Ejemplo

Se desea saber si la estatura máxima del padre influye en la estatura máxima del hijo, es decir, queremos modelar la estatura del hijo en función de la estatua del padre.

$$Y = f(z)$$

Donde $z$ = estatura del padre y $Y$ = estatura del hijo.
¿Exisite un modelo determinista?

$$Y = f(z) + \varepsilon$$

# Ejemplo

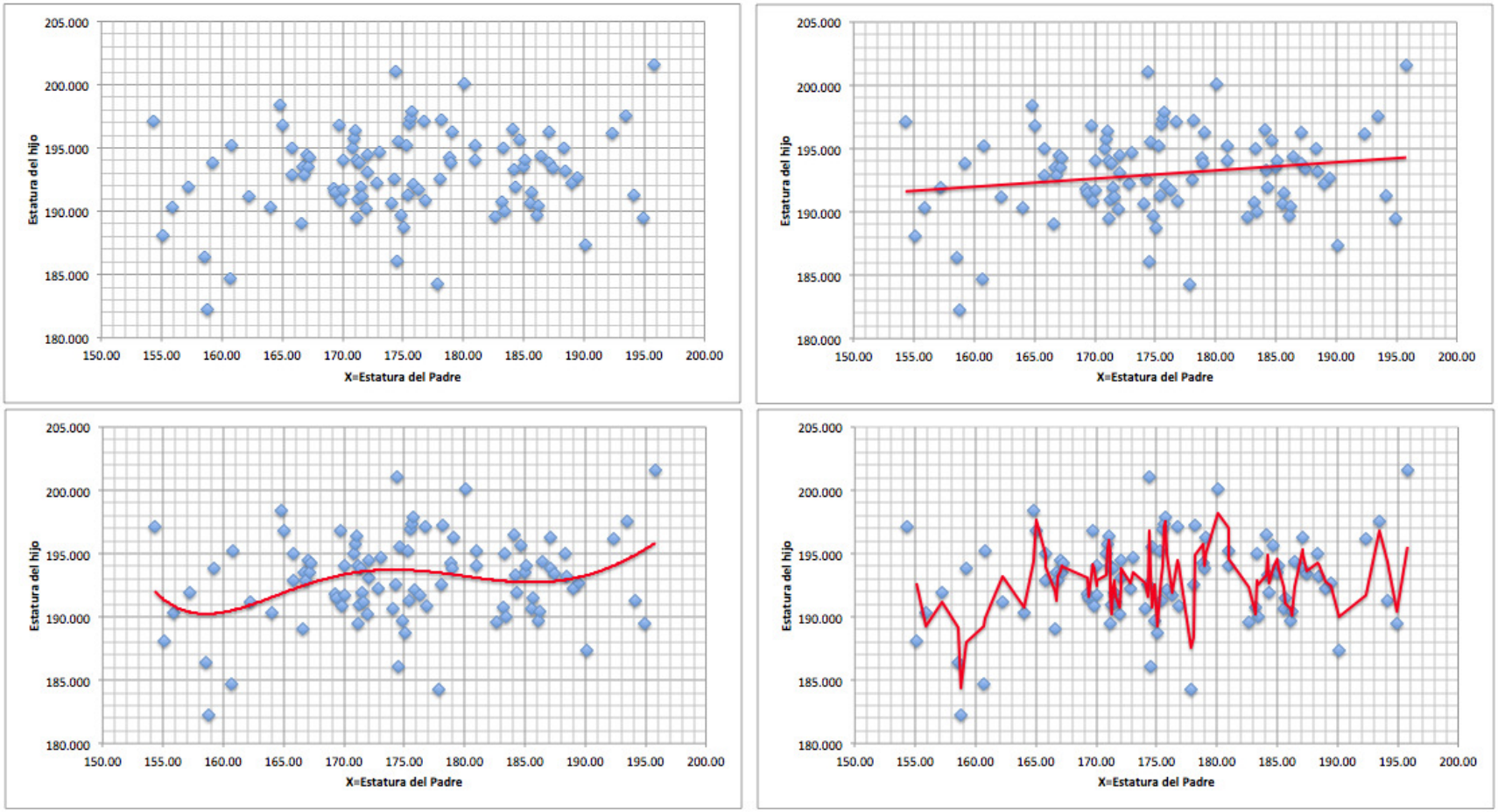
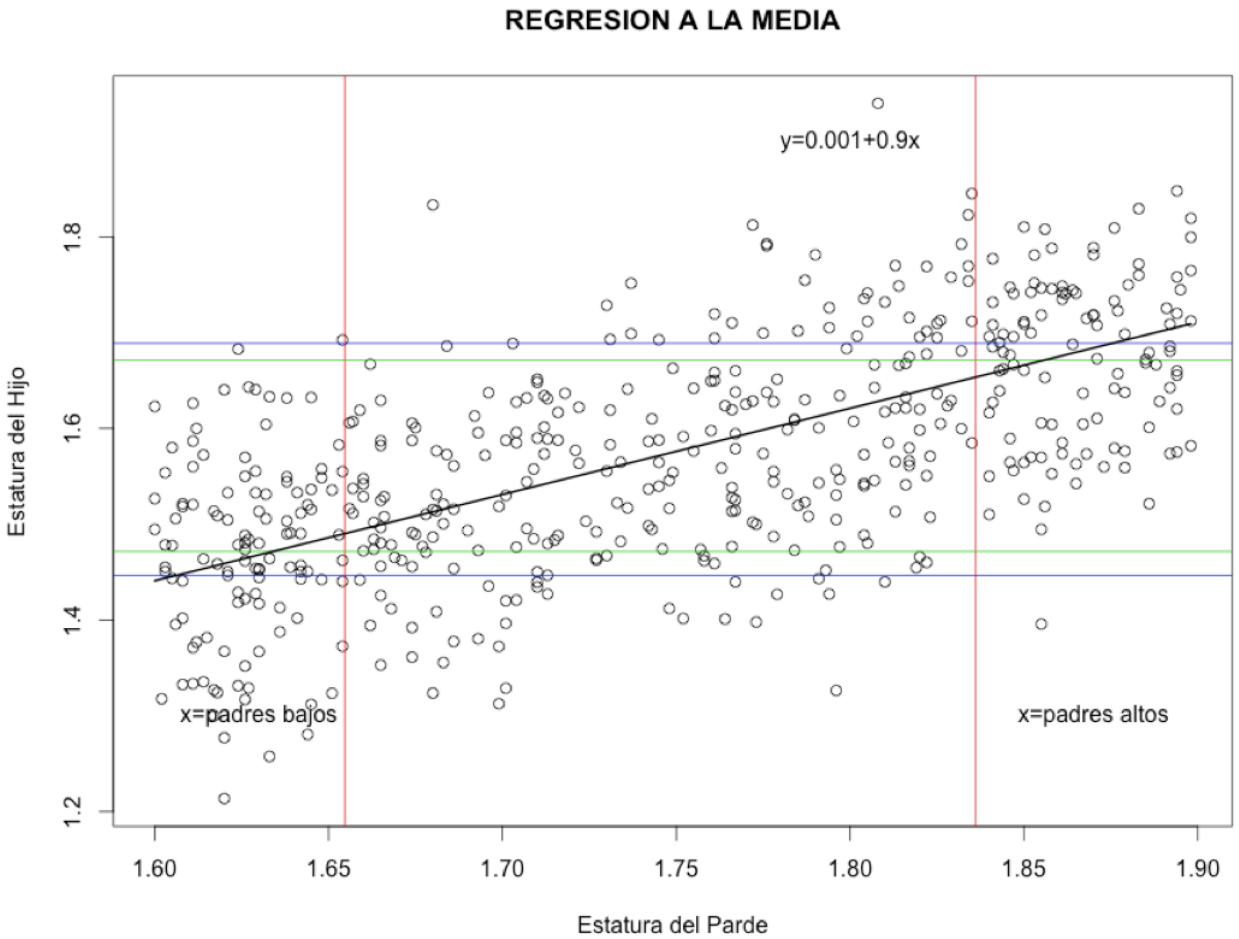



Figura: Estaturas

**Un poco de Historia....**

El término **regresión** fue acuñando por Francis Galton en el siglo XIX para describir un fenómeno biológico. Galton observó que las alturas de los descendientes de ancestros altos tienden a regresar hacia abajo, hacia un promedio normal (un fenómeno conocido como regresión hacia la media). (Programa 01 REGRESION A LA MEDIA.R)



Dada una variable respuesta Y y un conjunto de covariables $\underline{z} = (z_1, z_2, ...., z_q)$, surge de manera natural preguntarnos cuál deberá ser la relación funcional para modelar dicha relación.
Una forma de modelar podría ser:

$$\mathbb{E}(Y \mid \underline{z}) = \mu(\underline{z})$$

donde, en general, $\mu(.)$ es una función desconocida. En la práctica es común aproximar a $\mu(.)$ a traves de una función más simple:

$$\mu(\underline{z}) = \psi(\underline{z}; \underline{\beta})$$

donde $\underline{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$ denota a un vector de parámetros desconocidos.

La forma mas simple para modelar la relación es suponer una función lineal de $\underline{\beta}$ es decir:

$$\psi\left(\underline{z};\underline{\beta}\right) = \beta_0 + \beta_1 s_1\left(\underline{z}\right) + \ldots + \beta_k s_k\left(\underline{z}\right)$$

Donde $s_i : \mathbb{R}^q \to \mathbb{R}$ son funciones conocidas.
Finalmente esta función $\psi\left(\underline{z};\underline{\beta}\right)$ es tratada como si fuera la verdadera función de regresión $\mu\left(.\right)$, por lo que el problema se reduce a hacer inferencias sobre el valor del vector de parámetros $\underline{\beta}$.

$$\mathbb{E}(\,Y \mid \underline{z}) = \mu\left(\underline{z}\right) = \beta_0 + \beta_1 s_1\left(\underline{z}\right) + \ldots + \beta_k s_k\left(\underline{z}\right)$$

Muchas veces, en la vida real el modelo lineal no ajustará los datos por lo que tenemos que ver modelos no lineales. (NonLinear Regression)

# Ejemplos de modelos lineales

Modelos Lineales:

- $\psi\left(\underline{z};\underline{\beta}\right) = \beta_0 + \beta_1 z_1$
- $\psi\left(\underline{z};\underline{\beta}\right) = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p$
- $\psi\left(\underline{z};\underline{\beta}\right) = \beta_0 + \beta_1 z_1 + \beta_2 z_1^2 + \cdots + \beta_p z_1^p$
- $\psi\left(\underline{z};\underline{\beta}\right) = \beta_1 \sin(z_1) + \beta_2 \cos(z_2)$

Modelos no Lineales:

- $\psi\left(\underline{z};\underline{\beta}\right) = \beta_0 + \sin(\beta_1 z_1)$
- $\psi\left(\underline{z};\underline{\beta}\right) = \beta_0 + \cos(\beta_1 z_1)$
- $\psi\left(\underline{z};\underline{\beta}\right) = \tan(\beta_1 z_1) + \beta_2 z_2$

# Ejemplos de modelos linealizables

También existen en la literatura modelos denominados **linealizables**. Diremos que un modelo es linealizable si existe una función $g$ invertible tal que:

$$g\left(\psi\left(\underline{z};\underline{\beta}\right)\right) = \beta_0 + \beta_1 s_1\left(\underline{z}\right) + \ldots + \beta_k s_k\left(\underline{z}\right)$$

Por ejemplo:

- $\psi\left(\underline{z};\underline{\beta}\right) = \beta_0 \exp(\beta_1 z_1) \Rightarrow ln\left(\psi\left(\underline{z};\underline{\beta}\right)\right) = ln\left(\beta_0\right) + \beta_1 z_1 = \beta_0^* + \beta_1 z_1$
- $\psi\left(\underline{z};\underline{\beta}\right) = \log(\beta_0 + \beta_1 cos(z_1)) \Rightarrow exp\left(\psi\left(\underline{z};\underline{\beta}\right)\right) = \beta_0 + \beta_1 cos(z_1)$

Nuestro modelo lineal supone entonces que:

$$\mathbb{E}(Y \mid \underline{z}) = \mu(\underline{z}) = \beta_0 + \beta_1 s_1(\underline{z}) + \ldots + \beta_k s_k(\underline{z})$$

Esto lo podemos escribir en términos de v.a. como:

$$Y|\underline{z} \sim F\left(\mu(\underline{z}), \sigma^2\right) \quad (\sigma^2 > 0, desconocida)$$

Ahora imaginemos que recibiremos $n$ observaciones de este modelo para distintos niveles de las covariables $\underline{z}$, entonces podemos escribir

$$Y_i = \beta_0 + \beta_1 s_1(\underline{z}_i) + \ldots + \beta_k s_k(\underline{z}_i) + \varepsilon_i \quad \varepsilon_i \overset{iid}{\sim} F\left(0, \sigma^2\right) \tag{1}$$

Definamos:

$$x_{ij} = s_j\left(\underline{z}_i\right) \quad i \in \{1, \ldots, n\}; \quad j \in \{1, \ldots, k\}$$

Entonces el modelo (1) lo podemos escribir como:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i \quad \varepsilon_i \overset{iid}{\sim} F\left(0, \sigma^2\right) \tag{2}$$

**Ejemplo:**
Suponiendo que sólo tenemos una covariable $z \in \mathbb{R}$, y haciendo $s_j(z) = z^j$ entonces (2) toma la forma:

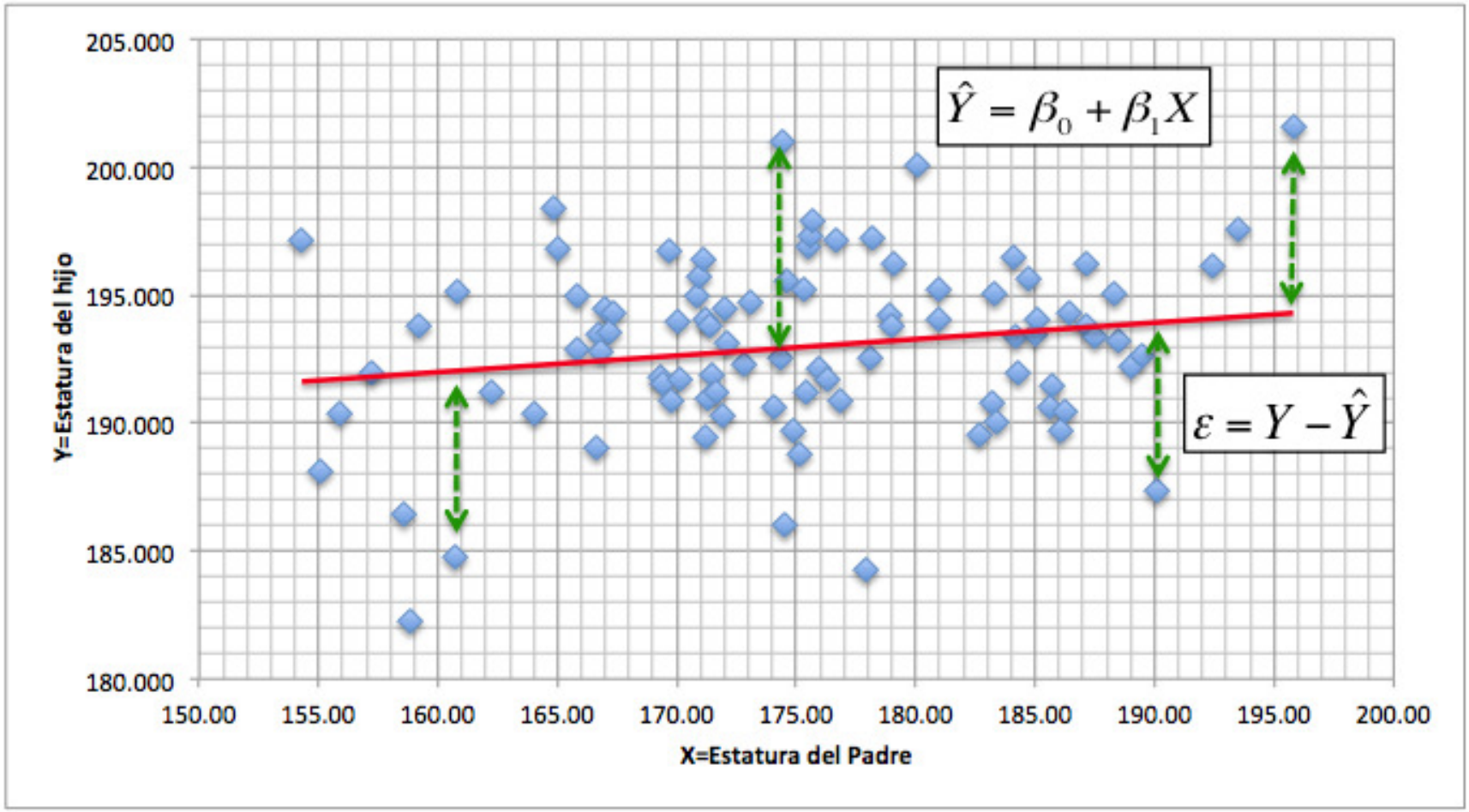
$$Y_i = \beta_0 + \beta_1 z_i + \ldots + \beta_k z_i^k + \varepsilon_i \quad \varepsilon_i \overset{iid}{\sim} F\left(0, \sigma^2\right) \tag{3}$$

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i \quad \varepsilon_i \overset{iid}{\sim} F\left(0, \sigma^2\right) \tag{4}$$

El modelo anterior pretende modelar el valor de $Y$ a través de una función polinomial de la covariable $z$.
Obs: Al modelo (2) se le conoce como Regresión Lineal Multiple mientras que cuando $k = 1$ lo denominamos el modelo lineal simple.

# Problema: ajuste por minimos cuadrados

Olvidemos un poco la teoría estadística y supongamos que tenemos el siguiente problema, se busca encontrar la ecuación de la recta $Y = \beta_0 + \beta_1 x$ que mas se ajuste a nuestros datos. (Encontrar $\beta_0$ y $\beta_1$ óptimos)



# Solución Analítica: Cálculo de Varias Variables

En términos matemáticos buscamos minimizar una función de dos variables $f : \mathbb{R}^2 \to \mathbb{R}$, con la siguiente forma:

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

Donde $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$ son los datos observados. Observe que estamos minimizando el cuadrado de la distancia entre el valor que arroja el modelo $\hat{Y}$ y el valor observado $Y$, es decir $(Y - \hat{Y})^2$. A los estimadores encontrados bajo esta metodología se les conoce como **Estimadores de Mínimos Cuadrados**

# Teoria de Calculo de várias variables.

**Teorema (Condiciones necesarias de extremo relativo)**

*Sea $f(x, y)$ una función real definida en conjunto abierto de $\mathbb{R}^2$ y $(a, b) \in \mathbb{R}^2$. Si $f$ admite derivadas parciales en $(a, b)$ y alcanza un máximo o mínimo relativo en dicho punto se verifica:*

$$\frac{\partial f}{\partial x}(a, b) = 0$$

$$\frac{\partial f}{\partial y}(a, b) = 0$$

- Las condiciones son necesarias pero no suficientes
- Las condiciones no tienen sentido para aquellos puntos en los que $f$ no posea derivadas parciales

# Teoria de Calculo de várias variables

## Definición (Puntos Críticos)

*Sea $f(x,y)$ una función real definida en un conjunto abierto de $\mathbb{R}^2$. Se llaman* ***puntos críticos*** *de $f$ a los puntos $(x,y)$ que sean solución del sistema de ecuaciones:*

$$\frac{\partial f}{\partial x}(x,y) = 0$$

$$\frac{\partial f}{\partial y}(x,y) = 0$$

# Teoria de Calculo de várias variables

## Teorema (Criterio de las derivadas parciales segundas)

*Sea $f(x,y)$ una función real con derivadas parciales primeras y segundas continuas en un conjunto abierto $D \subset \mathbb{R}^2$. Sea $(a,b) \in D$ un punto crítico de $f$ y*

$$\Delta = det \begin{bmatrix} \frac{\partial^2 f}{\partial^2 x} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial^2 y} \end{bmatrix}_{(x=a,y=b)}$$

*Entonces:*

- *Si $\Delta > 0$ y $\frac{\partial^2 f}{\partial^2 x}(a,b) > 0$, existe mínimo relativo en $(a,b)$*
- *Si $\Delta > 0$ y $\frac{\partial^2 f}{\partial^2 x}(a,b) < 0$, existe máximo relativo en $(a,b)$*
- *Si $\Delta < 0$, no existe ni máximo ni mínimo en $(a,b)$*
- *Si $\Delta = 0$, el criterio no da información*

# Solución Analítica ... (Continuación)

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

$$\left.\begin{array}{l} \frac{\partial f}{\partial \beta_0}(\beta_0, \beta_1) = 0 \\ \\ \frac{\partial f}{\partial \beta_1}(\beta_0, \beta_1) = 0 \end{array}\right\} \Rightarrow \left.\begin{array}{l} \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \\ \\ \sum_{i=1}^{n} y_i x_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \end{array}\right\}$$

$$det(H) = det \begin{bmatrix} 2n & 2\sum_{i=1}^{n} x_i \\ 2\sum_{i=1}^{n} x_i & 2\sum_{i=1}^{n} x_i^2 \end{bmatrix} = 4n \sum_{i=1}^{n} x_i^2 - 4 \left( \sum_{i=1}^{n} x_i \right)^2 > 0$$

Donde se concluye que el mínimo se obtiene en:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Con $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$ y $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

# Solución Analítica ... (Continuación)

**Observación Importante:** Los estimadores de Mínimos Cuadrados son combinaciones lineales de las $y_i$, es decir:

$$\hat{\beta}_1 = \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{S_{xx}} \right) y_i \qquad \hat{\beta}_0 = \sum_{i=1}^{n} \left( \frac{1}{n} - c_i \bar{x} \right) y_i$$

$$\hat{\beta}_1 = 0.06462304$$

$$\hat{\beta}_0 = 181.6538$$

Solución Numérica (Métodos Numéricos)
Ver codigo R
Función: lm y nlm

# Ejercicio:

Una empresa dedicada al venta de equipo deportivo desea modelar su ventas mensuales.

- Ajuste un modelo lineal de la forma

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

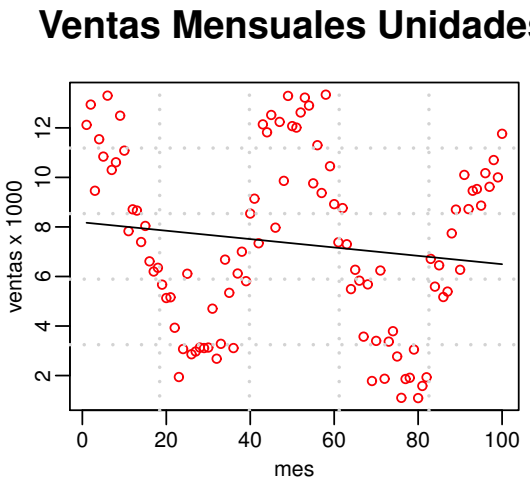
- Se sabe que la venta de artículos deportivos tiene una periodicidad de 2 años (48 meses) por lo que se propone ajustar un modelo de la forma de la forma:

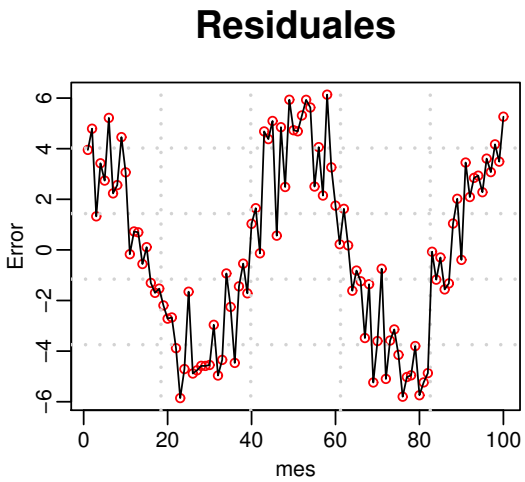
$$Y = \beta_0 + \beta_1 sin\left(\frac{2\pi}{48}x\right) + \beta_2 cos\left(\frac{2\pi}{48}x\right) + \varepsilon$$

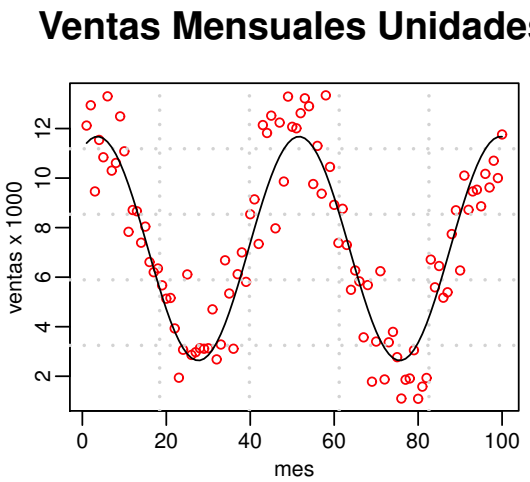
- Suponga ahora que no sabemos la periodicidad de los datos exactamente, solo sabemos que es aproximadamente cada 48 meses, ajuste el modelo "no"lineal de la forma:

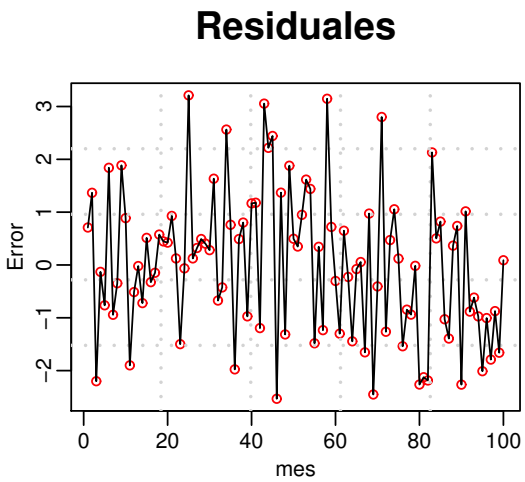
$$Y = \beta_0 + \beta_1 sin\left(\frac{2\pi}{\beta_3}x\right) + \beta_2 cos\left(\frac{2\pi}{\beta_4}x\right) + \varepsilon$$

# Ejercicio:

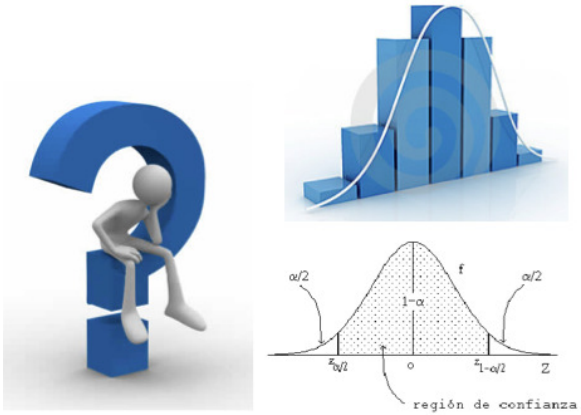

Ventas Mensuales Unidades
ventas x 1000
mes
2
4
6
8
10
12
0
20
40
60
80
100


Residuales
Error
mes
−6
−4
−2
0
2
4
6
0
20
40
60
80
100


Ventas Mensuales Unidades
ventas x 1000
mes
2
4
6
8
10
12
0
20
40
60
80
100


Residuales
Error
mes
−2
−1
0
1
2
3
0
20
40
60
80
100

# ¿Dónde entra la estadística?:




- ¿La estatura del padre influye realmente sobre la estatura del hijo?

$$H_0 : \beta_1 = 0 \quad H_0 \neq \beta_1$$

- ¿Cuál será la venta estimada en unidades para el siguiente periodo?

$$\mathbb{P}\left(\hat{y}_{n+1} \in (LI, LU)\right) = 0.95$$

- ¿El modelo está ajustando bien?

# Modelo Lineal General:

Recordemos que en el modelo lineal general se tiene un componente aleatorio con distribución $F(\mu = 0, \sigma)$, es decir:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i \quad \varepsilon_i \overset{iid}{\sim} F\left(0, \sigma^2\right)$$
$$\varepsilon \sim F(0, \sigma^2) \quad \mathbb{E}(\varepsilon) = 0 \quad \text{Var}(\varepsilon) = \sigma^2$$

Consecuencias:

- La variable respuesta $Y$ es una variable aleatoria pues es función de $\varepsilon$
- $$Y|\underline{z} = Y|\underline{x} \sim F\left(\mu\left(\underline{x}\right), \sigma^2\right)$$

# Modelo Lineal Simple:

Trabajemos el modelo lineal simple. Bajo los supuestos anteriores

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$
$$\mathbb{E}(Y) = \beta_0 + \beta_1 x_1 \quad \text{Var}(Y) = \sigma^2 \quad Y \sim F(\beta_0 + \beta_1 x_1, \sigma^2)$$

Suponga ahora que se observará una muestra de este modelo para distintos niveles de la covariable $x_1$, es decir se cuenta con observaciones $(y_1, x_{11}), \ldots, (y_n, x_{n1})$ de tal manera que:

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$$

Adicionemos una hipótesis al modelo:

- El componente aleatorio de una observación $\varepsilon_i$ *no* está correlacionado con el componente aleatorio de otra observación, es decir:

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$$

# Modelo Lineal Simple:

Tenemos los ingredientes necesarios para hacer inferencia:

- Una distribución parámetrica $Y_i \sim F(\beta_0 + \beta_1 x_{i1}, \sigma^2)$
- Parámetros desconocidos $\left(\beta_0, \beta_1, \sigma^2\right)$
- Una muestra de observaciones (no es m.a. bajo el enfoque de la inferencia estadística pues no tenemos variables idénticamente distribuidas) de la cual queremos obtener información para estimar los parámetros desconocidos

¿Cómo construir estimadores?

- Máxima Verosimilitud (Requiere saber la forma de la distribución F)
- Momentos (Requiere muestra aleatoria para poder igualar momentos)
- **Mínimos Cuadrados** (No requiere saber la forma de la distribución F)

Obs: Para fines de notación asumimos $x_{i1} = x_i$

# Modelo Lineal Simple (Estimación por minimos cuadrados):

La idea es encontrar $\hat{\beta}_0$ y $\hat{\beta}_1$ tal que la suma de los cuadrados de las diferencias de los valores ajustados $\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i$ y los valores observados $y_i$ sea mínimo, es decir:

$$\min_{\beta_0,\beta_1} \left\{ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right\}$$

Al utilizar técnicas de cálculo para varias variables y resolver las ecuaciones normales obtenemos que los estimadores son:

$$\left.\begin{array}{l} \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i x_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \end{array}\right\}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

# Propiedades de los Estimadores por Mínimos Cuadrados

- Los estimadores son combinaciones lineales de las v.a. $y_i's$ pues:

$$\hat{\beta}_1 = \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{S_{xx}} \right) y_i \qquad \hat{\beta}_0 = \sum_{i=1}^{n} \left( \frac{1}{n} - \left( \frac{x_i - \bar{x}}{S_{xx}} \right) \bar{x} \right) y_i$$

- Son insesgados es decir:

$$\mathbb{E}\left(\hat{\beta}_1\right) = \beta_1 \qquad \mathbb{E}\left(\hat{\beta}_0\right) = \beta_0$$

- La varianza de los estimadores es:

$$\text{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}} \qquad \text{Var}\left(\hat{\beta}_0\right) = \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \sigma^2$$

# Teorema de Gauss-Markov

En el modelo lineal simple $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ con $i = 1 \dots n$ y bajo las hipótesis siguientes:

- $\varepsilon_i$ es variable aleatoria con distribución $F$ para toda $i$
- $\mathbb{E}(\varepsilon_i) = 0$
- $\text{Var}(\varepsilon_i) = \sigma^2$ (homocedasticidad)
- $Cov\left(\varepsilon_i, \varepsilon_j\right) = 0 \quad i \neq j$
- $S_{xx} \neq 0$

Entonces, los estimadores de mínimos cuadrados $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ cumplen con lo siguiente:

- Son combinaciones lineales de las observaciones $y_i$
- Son insesgados
- **Son los mejores estimadores lineales (BLUES) - (MELI's)**. Es decir, si damos cualquier otro par de estimadores lineales $\left(\tilde{\beta}_0, \tilde{\beta}_1\right)$ entonces $\text{Var}\left(\hat{\beta}_0\right) \leq \text{Var}\left(\tilde{\beta}_0\right)$ y $\text{Var}\left(\hat{\beta}_1\right) \leq \text{Var}\left(\tilde{\beta}_1\right)$

# Estimación máximo verosímil:

A los supuestos que hemos venido manejando adicionaremos que además $\underline{\varepsilon} \sim N_n(\underline{0}, \sigma\mathbf{I})$. Consecuencias:

- Como $cov(\varepsilon_i, \varepsilon_j) = 0$ entonces $\varepsilon_i$ es independiente de $\varepsilon_j$
- Como $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ entonces $y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$
- $y_1, \ldots, y_n$ son independientes , pero no son idénticamente distribuidos.

La verosimilitud:

$$\begin{aligned} \mathscr{L}\left(\beta_0, \beta_1, \sigma^2; \underline{x}, \underline{y}\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\frac{n}{2}} \exp\left(-\sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right) \end{aligned}$$

$$\log\left(\mathscr{L}\left(\beta_0,\beta_1,\sigma^2;\underline{x},\underline{y}\right)\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)^2$$

Al maximizar la log-verosimilitud usando técnicas de cálculo de varias variables

$$\left(\frac{\partial\log\left(\mathscr{L}\right)}{\partial\beta_0},\frac{\partial\log\left(\mathscr{L}\right)}{\partial\beta_1},\frac{\partial\log\left(\mathscr{L}\right)}{\partial\sigma^2}\right) = (0,0,0)$$

Nos lleva a resolver las siguientes ecuaciones:

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1\sum_{i=1}^{n} x_i \tag{5}$$

$$\sum_{i=1}^{n} y_i x_i = \beta_0\sum_{i=1}^{n} x_i + \beta_1\sum_{i=1}^{n} x_i^2 \tag{6}$$

$$\frac{n}{2\sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)^2 \tag{7}$$

# Estimación máximo verosímil:

Las ecuaciones (5) y (6) son las ecuaciones normales de donde concluimos que el estimador máximo verosímil es el mismo que ya habíamos obtenido.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Sin embargo la ecuación (7) ahora nos permite encontrar un estimador máximo verosímil para $\sigma^2$, como estamos resolviendo un sistema de ecuaciones basta sustituir en (7) los valores obtenidos para $\beta_0$ y $\beta_1$ es decir:

$$\hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(e_i)^2$$

Donde $e_i$ son conocidos como los residuales del modelo y juegan un papel importante en los análisis de validación del modelo.

# Estimación máximo verosímil:

Entonces los estimadores máximos verosímiles bajo el supuesto de normalidad son

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}; \quad \hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2;$$

Ya sabemos que para el caso de $\hat{\beta}_1$ y $\hat{\beta}_0$ se cumple la propiedad de insesgamiento, sin embargo en el caso del estimador para $\sigma^2$ esto no ocurre pues:

$$\mathbb{E}\left(\hat{\sigma}^2_{MV}\right) = \frac{(n-2)}{n}\sigma^2$$

Por lo que comúnmente se suele usar al estimador insesgado $\hat{\sigma}^2$ definido como:

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

Obs: No se demuestra formalmente en este curso pero se puede probar que:

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

Obs: Si damos por hecho lo anterior es fácil probar el insesgamiento de $\hat{\sigma}^2$ y el sesgo de $\hat{\sigma}^2_{MV}$. Basta recordar que si $X \sim \chi^2_n$ entonces $\mathbb{E}(X) = n$ y $\text{Var}(X) = 2n$.
Entonces:

$$\mathbb{E}\left(\hat{\sigma}^2_{MV}\right) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right) = \frac{\sigma^2}{n}\mathbb{E}\left(\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sigma^2}\right) = \frac{(n-2)}{n}\sigma^2$$

$$\mathbb{E}\left(\hat{\sigma}^2\right) = \frac{1}{n-2}\mathbb{E}\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right) = \frac{\sigma^2}{n-2}\mathbb{E}\left(\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sigma^2}\right) = \sigma^2$$

# Pruebas de hipótesis:

El objetivo es contrastar hipótesis de la forma:

$$H_0 : \beta_1 = b_1 \quad vs \quad H_1 : \beta_1 \neq b_1; \quad H_0 : \beta_1 \leq b_1 \quad vs \quad H_1 : \beta_1 > b_1$$

$$H_0 : \beta_1 \geq b_1 \quad vs \quad H_1 : \beta_1 < b_1$$

Sacando provecho de la normalidad y de que los estimadores máximo verosímiles son combinaciones lineales de las observaciones $y_i$ entonces:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$

Recordando que:

**Theorem (Distribución $\tau$)**

*Sea* $X \sim N(0,1); \quad Y \sim \chi^2_n$ *entonces* $T = \frac{X}{\sqrt{\frac{Y}{n}}} \sim \tau_{(n)}$

# Pruebas de hipótesis:

Como $\frac{(n-2)}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-2}$ y $\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$ entonces:

$$t = \frac{\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}}}{\sqrt{\frac{(n-2)}{(n-2)\sigma^2}\hat{\sigma}^2}} \sim t_{(n-2)}$$

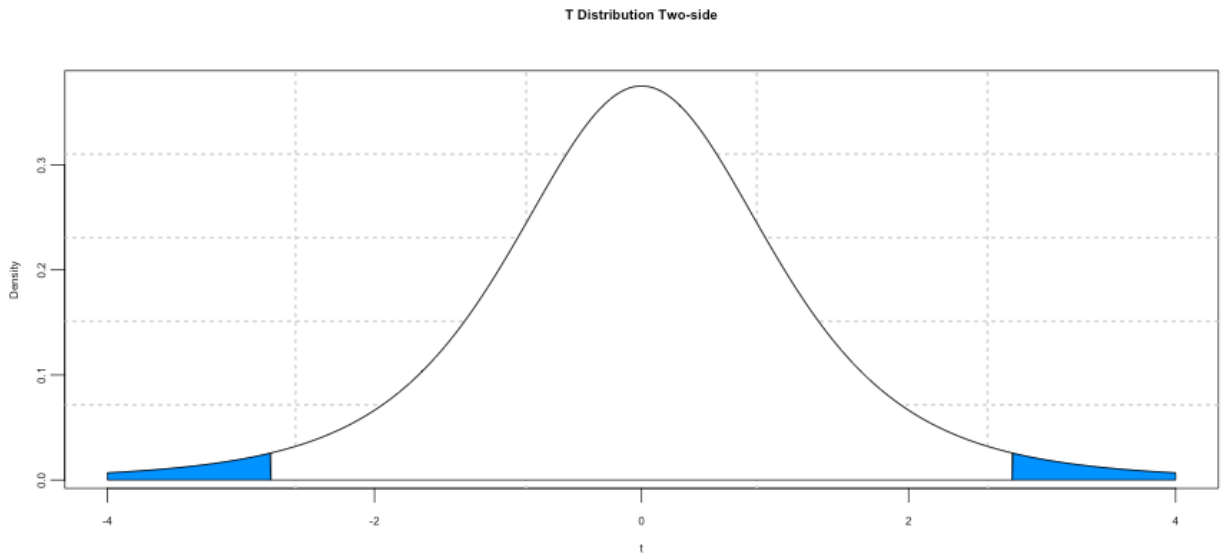
$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{n-2}$$

Obs: $t$ así definida *no* es una estadística (Pues depende de $\beta_1$ que es desconocida) sin embargo al fijar $\beta_1$ en una prueba de hipótesis ya puede ser utilizada para construir la region de rechazo.

# Pruebas de hipótesis:

$$H_0 : \beta_1 = b_1 \quad vs \quad H_1 : \beta_1 \neq b_1$$

La regla es rechazar $H_0$ cuando $|t| \geq \tau_{n-2}^{1-\alpha/2}$, donde $t = \frac{\hat{\beta}_1 - b_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}}$ ;



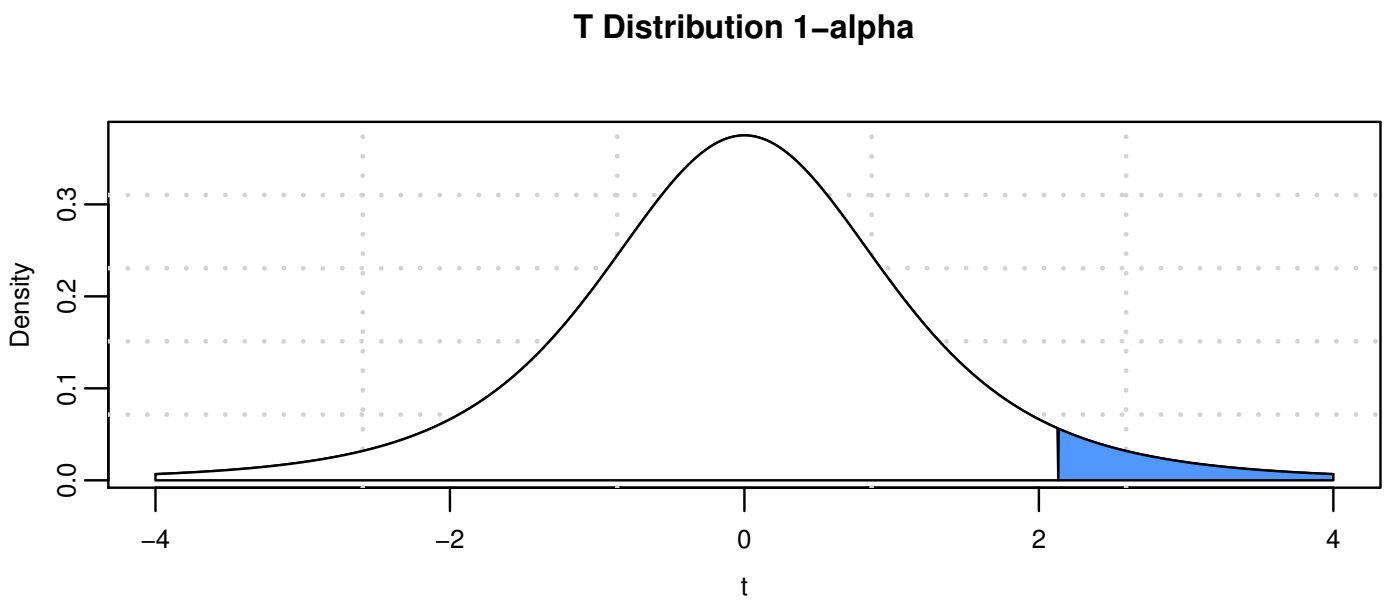

Importante: Esta es la prueba de hipótesis mas importante dentro del análisis de regresión lineal simple pues cuando hacemos $b_1 = 0$ nos ayuda a determinar si la pendiente de la recta ajustada es *estadísticamente* diferente de cero lo que se traduce en verificar si hay un efecto de la variable $X_1$ en la variable $Y$.

# Pruebas de hipótesis:

$$H_0 : \beta_1 \leq b_1 \quad vs \quad H_1 : \beta_1 > b_1$$

La regla es rechazar $H_0$ cuando $t \geq \tau_{n-2}^{1-\alpha}$, donde

$$t = \frac{\hat{\beta}_1 - b_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \quad ;$$

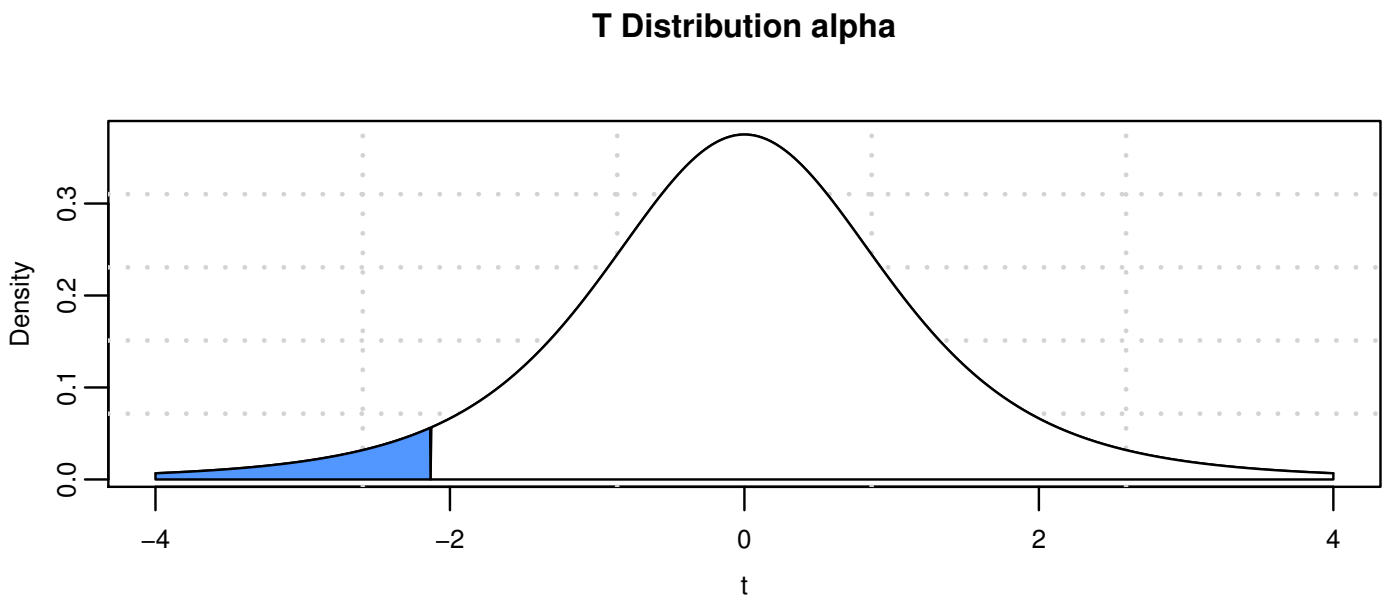


# Pruebas de hipótesis:

$$H_0 : \beta_1 \geq b_1 \quad vs \quad H_1 : \beta_1 < b_1$$

La regla es rechazar $H_0$ cuando $t \leq \tau^{\alpha}_{n-2}$, donde

$$t = \frac{\hat{\beta}_1 - b_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \quad ;$$



# Intervalo de Confianza:

La estadística $t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}}$ es una cantidad pivotal y por tanto puede ser utilizada para construir intervalos de confianza:

$$\mathbb{P}\left(-\tau_{n-2}^{1-\alpha/2} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \leq \tau_{n-2}^{1-\alpha/2}\right) = 1 - \alpha$$

$$\mathbb{P}\left(\hat{\beta}_1 - \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\tau_{n-2}^{1-\alpha/2} \leq \beta_1 \leq \hat{\beta}_1 + \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\tau_{n-2}^{1-\alpha/2}\right) = 1 - \alpha$$

Por lo tanto un intervalo de confianza al $(1-\alpha)$ % es:

$$\left(\hat{\beta}_1 - \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\tau_{n-2}^{1-\alpha/2}, \hat{\beta}_1 + \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\tau_{n-2}^{1-\alpha/2}\right)$$

Donde $\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$ se le conoce como error estándar del estimador.

Notemos que la region de rechazo construida para la prueba $H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$ fue obtenida sacando provecho de las caraterísticas distribucionales que se dieron al suponer $\varepsilon_i \sim N(0, \sigma^2)$. Ahora utilizaremos el método del cociente de verosimilitudes generalizados para encontrar la region de rechazo.
Recordemos que el método indica que tenemos que rechazar $H_0$ si:

$$\frac{\sup \mathscr{L}(\Theta_{H_0})}{\sup \mathscr{L}(\Theta)} < \lambda; \quad \lambda \leq 1$$

Donde:

$$\Theta_{H_0} := \left\{ \left(\beta_0, \beta_1, \sigma^2\right) : \beta_0 \in \mathbb{R}, \beta_1 = 0, 0 < \sigma^2 < \infty \right\}$$

$$\Theta := \left\{ \left(\beta_0, \beta_1, \sigma^2\right) : \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}, 0 < \sigma^2 < \infty \right\}$$

Sabemos que $\sup \mathscr{L}(\Theta)$ se obtiene evaluando en los estimadores máximo verosímiles, es decir :

$$\sup \mathscr{L}(\Theta) = \mathscr{L}\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2_{MV}\right) = \left(2\pi\hat{\sigma}^2_{MV}\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{2\hat{\sigma}^2_{MV}}\right)$$

Recordando que:

$$\hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\sup \mathscr{L}(\Theta) = \left(2\pi\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)^{-\frac{n}{2}} \exp\left(-\frac{n}{2}\right)$$

Ahora nos enfocaremos a encontrar $\sup \mathscr{L}(\Theta_{H_0})$, como suponemos que $H_0$ es cierta entonces $\beta_1 = 0$ por lo tanto

$$\mathscr{L}(\Theta_{H_0}) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0)^2\right)$$

Maximizando respecto a $\beta_0$ y $\sigma^2$ (Hint: Hay que aplicar logaritmo y luego maximizar) obtenemos:

$$\hat{\beta}_{0_{H_0}} = \bar{y} \qquad \hat{\sigma}^2_{H_0} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})$$

Entonces:

$$\sup \mathscr{L}(\Theta_{H_0}) = \left(2\pi\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^{-\frac{n}{2}} \exp\left(-\frac{n}{2}\right)$$

Aplicando la siguiente descomposición a la suma de cuadrados:

$$\underbrace{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}_{SCT}=\underbrace{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}_{SCR}+\underbrace{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}_{SCE}$$

Entonces el cociente de verosimilitudes generalizado queda como:

$$\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}\right)^2}{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2+\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}<\lambda^*$$

Lo que equivale a rechazar $H_0$ si:

$$\frac{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}>k^* \tag{8}$$

En la tarea ustedes tienen que probar que bajo el modelo de regresión lineal simple se cumple que:

$$\sum_{i=1}^{n} (\hat{y} - \bar{y})^2 = \hat{\beta}_1^2 S_{xx}$$

Entonces el cociente de verosimilitudes generalizado que obtuvimos en (8), nos dice que tenemos que rechazar $H_0$ si:

$$\frac{\hat{\beta}_1^2 S_{xx}}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} > k^* \Leftrightarrow \frac{\hat{\beta}_1^2}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{S_{xx}}} > k^* \Leftrightarrow \frac{\hat{\beta}_1^2}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{S_{xx}(n-2)}} > k^*(n-2)$$

Sacando raíz cuadrada

$$\frac{\left|\hat{\beta}_1\right|}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} > k^*(n-2) = t \quad \textit{Llegamos a la misma prueba !!!} \tag{9}$$

# La Prueba F:

La prueba $t-student$ es la mas usada para llevar acabo la significancia de los parámetros en la regresión de forma independiente. Existe otra forma de llevar a cabo la prueba de hipótesis utilizando la distribución $F$ de Snedecor. La distribución $F$ es una distribución de probabilidad continua y tiene dos parámetros asociados. Su densidad está dada por:

$$f(x;n_1,n_2)=\frac{1}{Beta\left(\frac{n_1}{2},\frac{n_2}{2}\right)}\left(\frac{n_1x}{n_1x+n_2}\right)^{\frac{n_1}{2}}\left(1-\frac{n_1x}{n_1x+n_2}\right)^{\frac{n_2}{2}}x^{-1}$$

**Teorema**

*Sea* $X\sim\chi_{n_1}$ *y* $Y\sim\chi_{n_2}$ *independientes entonces:*

$$F=\frac{\frac{X}{n_1}}{\frac{Y}{n_2}}=\frac{n_2X}{n_1Y}\sim\mathscr{F}_{(n_1,n_2)}$$

**Teorema**

*Sea* $t\sim T_n$ *entonces:*

$$t^2\sim\mathscr{F}_{(1,n)}$$

Asumiendo que :

$$\frac{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}{\sigma^2}\sim\chi_{(1)}$$

$$\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{\sigma^2}\sim\chi_{(n-2)}$$

Entonces de la ecuación (8) el cociente de verosimilitudes generalizado nos dice que tenemos que rechazar cuando:

$$\frac{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}>k^*\Leftrightarrow\frac{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2/\sigma^2}{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2/\sigma^2}>k^*\Leftrightarrow$$

$$\frac{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2/\sigma^2}{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2/\left(\sigma^2(n-2)\right)}>k^*(n-2)=K$$

Finalmente el cociente de verosimilitudes nos lleva a rechazar $H_0$ si:

$$F=\frac{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2/(n-2)}>\mathscr{F}_{(1,n-2)}^{(1-\alpha)}$$

# Analisis de Varianza o ANOVA

Como mencionamos una de las pruebas mas importantes en el análisis de regresión simple es verificar que el coeficiente asociado a la pendiente de la recta es significativo, es decir, $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$. Esto lo podemos verificar en el caso de la regresión simple de dos maneras, utilizando la prueba $t$ o bien la prueba $F$. Un método de verificar esta hipótesis de manera mas formal es utilizando la tabla ANOVA la cual se basa en el estadístico $F$ que ya hemos visto. La parte fundamental del ANOVA se basa en la siguiente descomposición de la suma de cuadrados:

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

# Analisis de Varianza o ANOVA

Se puede verificar via formas cuadráticas que (mas adelante):

$$\underbrace{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sigma^2}}_{\frac{SCT}{\sigma^2}} \sim \chi_{(n-1)} \quad \text{Suma de Cuadrados del Modelo Reducido}$$

$$\underbrace{\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sigma^2}}_{\frac{SCR}{\sigma^2}} \sim \chi_{(1)} \quad \text{Suma de Cuadrados explicada por la Modelo Completo}$$

$$\underbrace{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sigma^2}}_{\frac{SCE}{\sigma^2}} \sim \chi_{(n-2)} \quad \text{Suma de Cuadrados no explicada por el Modelo Compl}$$

# Analisis de Varianza o ANOVA

Con lo anterior construimos la siguiente tabla:

| Fnte. de Var. | S. C. | G. Lib | S.C.M. | F |
|---|---|---|---|---|
| Regresión (SCR) | $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ | 1 | $\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{1}$ | $\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}}$ |
| Residuales (SCE) | $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ | $n-2$ | $\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$ | |
| Total (SCT) | $\sum_{i=1}^{n}(y_i-\bar{y})^2$ | $n-1$ | $\frac{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}{n-1}$ | |

Esta tabla resume mucha información sobre el modelo lineal y por lo general siempre aparece en los paquetes de computo estadístico. Aparte de que la tabla nos brinda un estadístico para contratastar la hipótesis de la significancía de la pendiente de la recta ajustada, también nos sirve para tener una medida del ajuste de los datos al modelo lineal

# Coeficiente de Determinación

El coeficiente de determinación es un número entre 0 y 1 que comúnmente arrojan los paquetes estadísticos y lo denotan como $R^2$ el cual sirve como una medida del ajuste del modelo. Este valor puede ser obtenido de la siguiente manera utilizando la Tabla ANOVA:

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

La interpretación es la siguiente: Recordemos que $SCT$ mide la variabilidad total para el caso en que no ajustamos un modelo con pendiente es decir $(\beta_1 = 0)$ luego de esta variabilidad total que tienen nuestros datos podemos preguntarnos cuanta fue explicada por la Regresion con una pendiente $\beta_1 \neq 0$ (modelo completo). Dado que $SCT = SCR + SCE$ y como $SCE$ mide la variabilidad que no pude ser explicado por el modelo completo, entonces $SCR$ se considera como la variabilidad que logramos explicar al ajustar un modelo completo, luego entonces $\frac{SCR}{SCT}$ medirá el porcentaje de variabilidad explicada por nuestro modelo. Debe de observarse ademas que cuando $SCE = 0$ entonces $SCT = SCR$ y por tanto tenemos un ajuste perfecto de nuestros datos a la recta ajustada y por tanto $R^2 = 1$

# Coeficiente de Determinación vs Coeficiente de Correlación de Pearson

En estadística descriptiva es común hablar del **Coeficiente de Correlación de Pearson** el cual es una medida de correlación entre dos variables cuantitativas, de manera menos formal podemos definirlo como índicie que pude utilizarse para medir el grado de relación de dos variables. El coeficiente se define como:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Este coeficiente es tal que $-1 \leq r \leq 1$ de tal forma que cuando $r = 1$ existe una correlación positiva entre las variables y si $r = -1$ entonces las variables están correlacionadas negativamente. En el contexto de Regresión y usando nuestra notación tenemos que:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Partiendo de esta igualdad demostrarán que:

$$r^2 = R^2$$

# Análisis para $\beta_0$

Toda la teoría que desarrollamos hasta ahora fue para analizar el comportamiento de $\beta_1$, se puede demostrar haciendo un análisis similar que para $\beta_0$ se tiene que:

| Prueba | Región de Rechazo |
|---|---|
| $H_0 : \beta_0 \leq b_0$ vs $H_1 : \beta_0 > b_0$ | $\frac{\hat{\beta}_0 - b_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} > \tau_{n-2}^{1-\alpha}$ |
| $H_0 : \beta_0 \geq b_0$ vs $H_1 : \beta_0 < b_0$ | $\frac{\hat{\beta}_0 - b_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} < \tau_{n-2}^{\alpha}$ |
| $H_0 : \beta_0 = b_0$ vs $H_1 : \beta_0 \neq b_0$ | $\frac{\lvert\hat{\beta}_0 - b_0\rvert}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} > \tau_{n-2}^{1-\alpha/2}$ |

Intervalo de confianza para $\beta_0$ es:

$$\left(\hat{\beta}_0 - \tau_{n-2}^{(1-\alpha/2)}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}, \hat{\beta}_0 + \tau_{n-2}^{(1-\alpha/2)}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}\right)$$

# Análisis para $\sigma^2$

Dentro del modelo lineal simple, la varianza de los errores $\sigma^2$ es un parámetro desconocido y por tanto también se puede llevar a cabo inferencia

Tomando en cuenta nuevamente que:

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

Se prueba que:

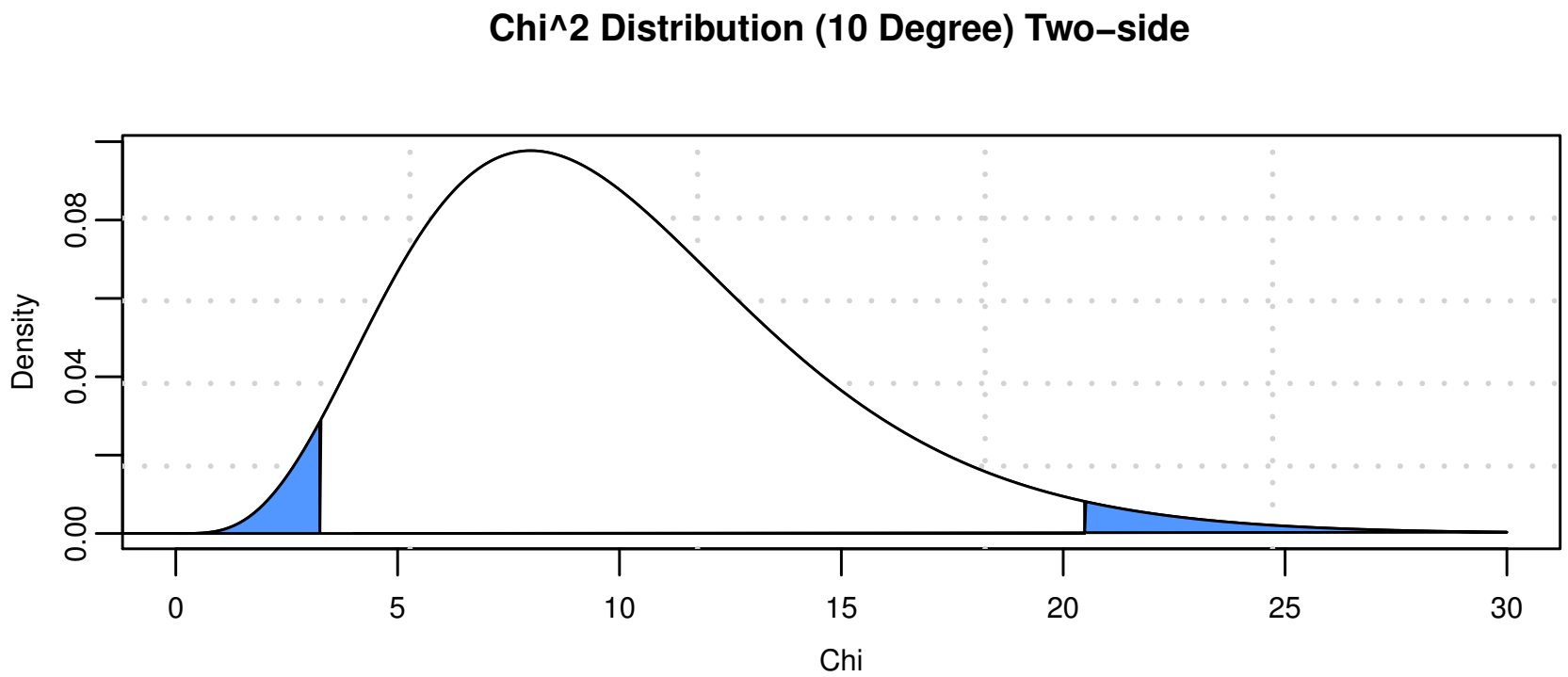
| Prueba | Región de Rechazo |
|---|---|
| $H_0 : \sigma^2 \leq s$ vs $H_1 : \sigma^2 > s$ | $\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{s} > \chi^{2\ (1-\alpha)}_{n-2}$ |
| $H_0 : \sigma^2 \geq s$ vs $H_1 : \sigma^2 < s$ | $\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{s} < \chi^{2\ (\alpha)}_{n-2}$ |
| $H_0 : \sigma^2 = s$ vs $H_1 : \sigma^2 \neq s$ | $\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{s} > \chi^{2\ (1-\alpha/2)}_{n-2}$ o $\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{s} < \chi^{2\ (\alpha/2)}_{n-2}$ |

Intervalo de confianza para $\sigma^2$ es:

$$\left( \frac{1}{\chi^{2\ (1-\alpha/2)}_{n-2}} \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 , \frac{1}{\chi^{2\ (\alpha/2)}_{n-2}} \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \right)^*$$

# Observación: Longitud del intervalo

Un problema que tiene esta construcción de intervalo de confianza es que no es de longitud mínima, esto se debe a que la densidad $\chi^2$ no es simétrica.



# Intervalo de Confianza para la Respuesta Media

Uno de los usos mas importantes del análisis de regresión es estimar la respuesta media para un valor particular de la variable explicativa $X_1 = x_*$, es decir, estimar $\mathbb{E}(Y \mid X_1 = x_*)$. Sea $x_*$ cualquier valor de la variable explicativa **(por lo generar pedimos que $x_*$ esté dentro del rango de las datos originales de $X_1$)**, en nuestro modelo sabemos que:

$$\mathbb{E}(Y \mid X_1 = x_*) = \beta_0 + \beta_1 x_* \tag{10}$$

Como $\beta_0$ y $\beta_1$ son desconocidos se sigue que el valor para la respuesta media es desconocido $\mathbb{E}(Y \mid X_1 = x_*)$ por lo que procedemos a estimarlo. La forma mas natural de estimar este valor es **enchufar** los estimadores que ya tenemos para $\beta_0$ , $\beta_1$ en la expresión (10) obteniendo el siguiente estimador:

$$\widehat{\mathbb{E}(Y \mid X_1 = x_*)} = \hat{\beta}_0 + \hat{\beta}_1 x_*$$

Surge entonces la necesidad de verificar propiedades de nuestro estimador:

- ¿Es insegado?
- ¿Cuál es su varianza?
- ¿Qué distribución sigue?

# Intervalo de Confianza para la Respuesta Media

Primero recordemos que $\hat{\beta}_0$ y $\hat{\beta}_1$ son combinaciones lineales de las variables $y_i's$ luego entonces $\mathbb{E}(Y \mid X_1 = x_*)$ también es combinación lineal de las $y_i's$. En efecto:

$$\begin{aligned}
\widehat{\mathbb{E}(Y \mid X_1 = x_*)} &= \hat{\beta}_0 + \hat{\beta}_1 x_* \\
&= \sum_{i=1}^{n} \left( \frac{1}{n} - \left( \frac{x_i - \bar{x}}{S_{xx}} \right) \bar{x} \right) y_i + \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{S_{xx}} \right) y_i x_* \\
&= \sum_{i=1}^{n} \left( \frac{1}{n} - \left( \frac{x_i - \bar{x}}{S_{xx}} \right) \bar{x} + \left( \frac{x_i - \bar{x}}{S_{xx}} \right) x_* \right) y_i \\
&= \sum_{i=1}^{n} \left( \frac{1}{n} + \left( \frac{x_i - \bar{x}}{S_{xx}} \right) (x_* - \bar{x}) \right) y_i
\end{aligned}$$

Luego entonces, concluimos que que al ser las variables $y_i's$ normales e independientes se sigue que $\widehat{\mathbb{E}(Y \mid X_1 = x_*)}$ tiene un a distribución Normal, luego de la última ecuación y tras unos pasos algebraicos obtenemos que la varianza del estimador es:

$$\text{Var}(\mathbb{E}(Y \mid X_1 = x_*)) = \sigma^2 \left( \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}} \right)$$

# Intervalo de Confianza para la Respuesta Media

Luego, para obtener la esperanza del estimador solo basta recordar que los estimadores $\hat{\beta}_0$, $\hat{\beta}_1$ son insesgado, entonces:

$$\mathbb{E}\Big(\mathbb{E}(\widehat{Y \mid X_1 = x_*})\Big) = \mathbb{E}\Big(\hat{\beta}_0 + \hat{\beta}_1 x_*\Big) = \beta_0 + \beta_1 x_*$$

Finalmente hemos probado que:

$$\mathbb{E}(\widehat{Y \mid X_1 = x_*}) \sim N\left(\beta_0 + \beta_1 x_*, \sigma^2\left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}\right)\right)$$

Luego entonces recordando que $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$ y que $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi_{(n-2)}$ entonces podemos probar que:

$$\frac{\Big(\mathbb{E}(\widehat{Y \mid X_1 = x_*}) - \mathbb{E}(Y \mid X_1 = x_*)\Big)}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}\right)}} \sim t_{(n-2)}$$

La anterior es una cantidad pivotal y por tanto puede ser utilizada para crear intervalos de confianza.

# Intervalo de Confianza para la Respuesta Media

El Intervalo al $(1-\alpha)\,\%$ de confianza para la respuesta media es:

$$\left(\mathbb{E}(\widehat{Y \mid x_*}) - \tau_{(n-2)}^{(1-\frac{\alpha}{2})}\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{(x_*-\bar{x})^2}{S_{xx}}\right)}, \mathbb{E}(\widehat{Y \mid x_*}) + \tau_{(n-2)}^{(1-\frac{\alpha}{2})}\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{(x_*-\bar{x})^2}{S_{xx}}\right)}\right)$$

Observaciones:

- La longitud del intervalo se minimiza si $S_{xx}$ que mide la variabilidad de las variables independientes es grande
- Otra forma de minimizar la longitud del intervalo es poner $x_* = \bar{x}$
- Conforme nos alejamos de $\bar{x}$ la longitud del intervalo es mas grande.

# Intervalo de Predicción para nuevas Observacioes

**Definición (Intervalo de Predicción)**

*Un intervalo de predicción al* $100(1-\alpha)\,\%$ *para una variable aleatoria no observada* $Y$ *basado en variables no observadas* $\boldsymbol{X}$*, es un intervalo aleatorio:*

$$[L(\boldsymbol{X}), U(\boldsymbol{X})]$$

*Que tiene la propiedad que;*

$$\mathbb{P}(L(\boldsymbol{X}) \leq Y \leq U(\boldsymbol{X})) = 1-\alpha$$

# Intervalo de Predicción para nuevas Observacioes

Ahora supongamos que no nos interesa estimar a $\mathbb{E}(Y \mid X_1 = x_*)$ sino que queremos estimar el valor directamente de la respuesta $Y$ cuando $X$ toma cierto valor $x_0$.
Queremos una estimación para la variable respuesta $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$.
Una estimación natural dadas las hipótesis que hemos venido manejando es

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Notemos que como $y_0$ es una nueva observación entonces $y_0$ es independiente de las observaciones anteriores $(y_1, \ldots, y_n)$ y como los estimadores $\hat{\beta}_0$,$\hat{\beta}_1$ son combinaciones lineales de las observaciones $(y_1, \ldots, y_n)$ , se sigue entonces que $y_0$ es independiente de $\hat{y}_0$, por lo tanto $(y_0 - \hat{y}_0)$ sigue una distribución normal. Luego como:

$$\begin{aligned}
\mathbb{E}((y_0 - \hat{y}_0)) &= \mathbb{E}\left(\beta_0 + \beta_1 x_0 + \varepsilon_0 - \hat{\beta}_0 + \hat{\beta}_1 x_0\right) = 0 \\
\mathrm{Var}((y_0 - \hat{y}_0)) &= \mathrm{Var}(y_0) + \mathrm{Var}(\hat{y}_0) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right) \\
&= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)
\end{aligned}$$

Finalmente hemos probado que:

$$(y_0 - \hat{y}_0) \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$$

Nuevamente recordando que $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$ y que $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi_{(n-2)}$ entonces podemos probar que:

$$\frac{(y_0 - \hat{y}_0)}{\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{(n-2)}$$

# Intervalo de Predicción para nuevas observaciones

El Intervalo de predicción al $(1-\alpha)\,\%$ para la respuesta cuando al variable independiente es una nueva observación:

$$\left(\hat{y}_0 - \tau_{(n-2)}^{(1-\frac{\alpha}{2})}\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}, \hat{y}_0 + \tau_{(n-2)}^{(1-\frac{\alpha}{2})}\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}\right)$$

Observaciones:

- Este intervalo es similar al encontrado anteriormente solo que agrega mas variabilidad debido a que se esta evaluando el modelo en una nueva observación que no se tenia en muestra.
- La longitud del intervalo se minimiza si $S_{xx}$ que mide la variabilidad de las variables independientes es grande
- Otra forma de minimizar la longitud del intervalo es poner $x_0 = \bar{x}$.
- Conforme nos alejamos de $\bar{x}$ la longitud del intervalo es mas grande, es decir se tiene mayor incertidumbre.
- Teóricamente con este intervalo podemos inferir el valor de la variable respuesta para cualquier valor de la variable independiente, sin embargo siempre debemos tomar en cuenta el contexto del fenómeno que estamos estudiando.

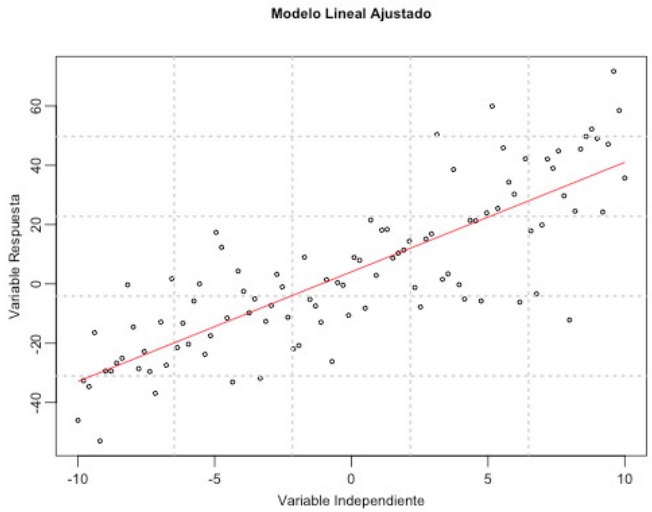
# Validación del Modelo

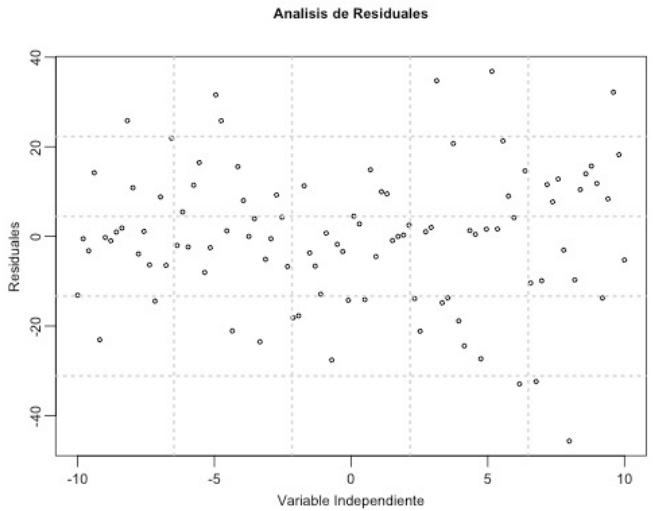
Una vez calculado el modelo de regresión lineal es muy importante llevar a cabo la verificación de las hipótesis que impusimos al principicio, esto es, tenemos que verificar la linealidad de los datos, la normalidad de los errores así como su homocedasticidad y su independencia. Es muy importante verificar todos estos supuestos previo a cualquier inferencia que se quiera hacer con el modelo calculado.

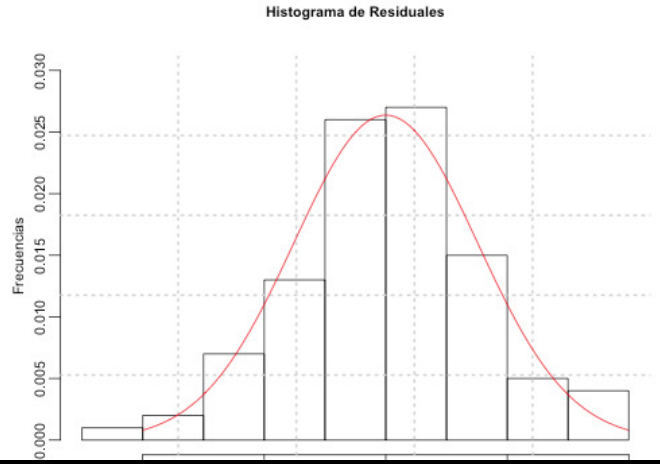
- Una primera validación que debemos de hacer es verificar estadísticamente mediante una prueba de hipótesis que $\beta_1 \neq 0$ ya que si hay evidencia de que $\beta_1 = 0$ entonces se interpretara como **falta de relación lineal** entre las variables y, por lo tanto, supone la inutilidad del modelo de regresión, sin embargo el hecho de rechazar en una prueba que $\beta_1 = 0$ no garantiza por sí sola el ajuste del modelo.
- Para verificar la hipótesis de normalidad homecedasticidad y la independencia de los errores se utilizan los gráficos de los residuos, donde cada residuo, denotado por $e_i$, esta definido como $e_i = y_i - \hat{y}_i$. Generalmente se gráfica $(x_i, e_i)$.

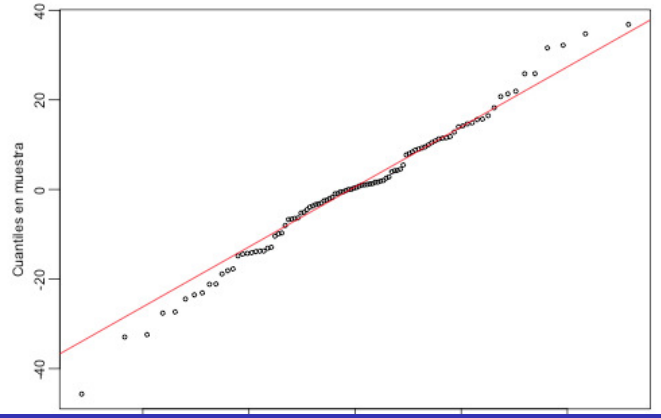
# Validación del Modelo : Normalidad

Parte fundamental de la validación del modelo es verificar la normalidad de los residuales, para ello contamos con pruebas de bondad de ajuste, sin embargo una primera prueba se hace llevando a cabo un histograma de los residuales y verificar que siguen una distribución parecida a la normal.



# Validación del Modelo : Normalidad

Existen muchas pruebas para verificar la normalidad, las mas usadas son:

- Shapiro-Wilk Normality Test (shapiro.test(x))
- Anderson-Darling test for normality (ad.test(x))
- Kolmogorov-Smirnov Tests - Lilliefors (lillie.test(x))

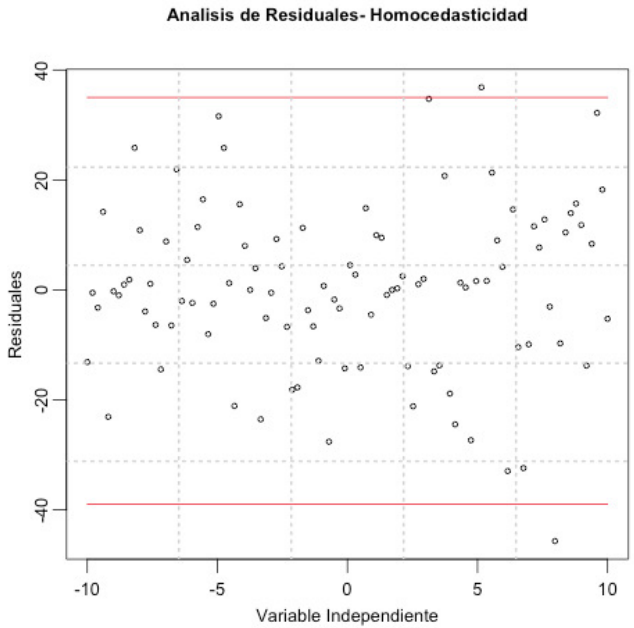
Ejemplo:

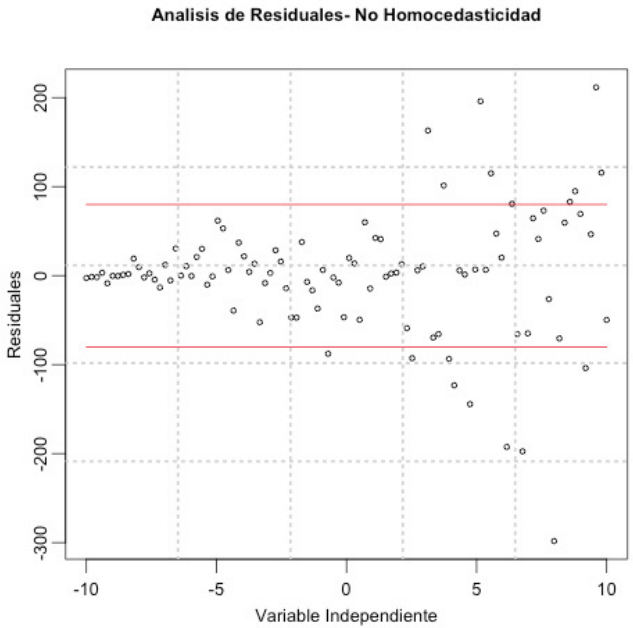
- `Shapiro-Wilk normality test`
  `data: res`
  `W = 0.9887, p-value = 0.558`
- `Anderson-Darling normality test`
  `data: res`
  `A = 0.466, p-value = 0.2475`

Sin embargo, en el contexto que estamos trabajando **no** podemos aplicar teóricamente la prueba pues los residuales no son independientes. Sim embargo puede utilizarse como una medida de que tan normales son.

# Validación del Modelo : Homocedasticidad

Generalemente para verificar que tenemos varianza constante se llevan a cabo pruebas visuales verificando el comportamiento de los residuales conforme variamos la variable independiente

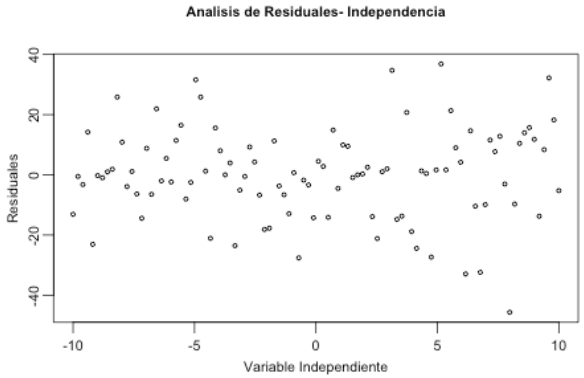


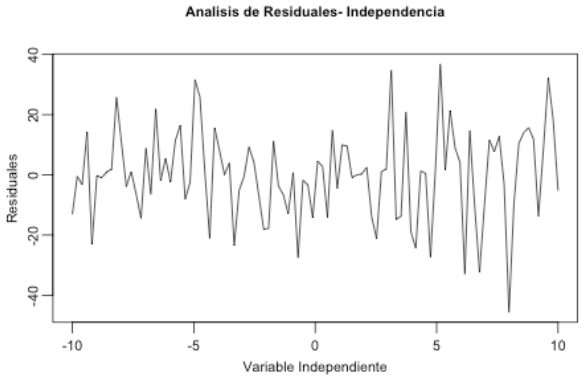



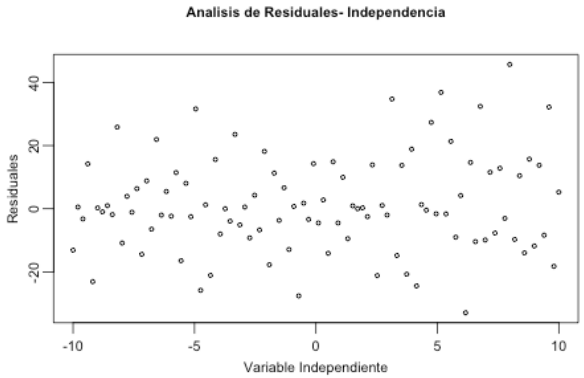

Existen pruebas formales para llevar a cabo el contraste (Ej. Barttlet, Levenne), pero requieren tener definido grupos en la variable independiente $X$, por lo general esta prueba es utilizada en diseño de experimentos pero podemos adaptarla al caso de regresión definiendo grupos arbitrarios, el problema que tiene esto es que deja que el analista determine el número de grupos así como el método para formarlos, lo cual puede no ser objetivo y llegar a cometer errores. (Investigar prueba: BreuschPagan)

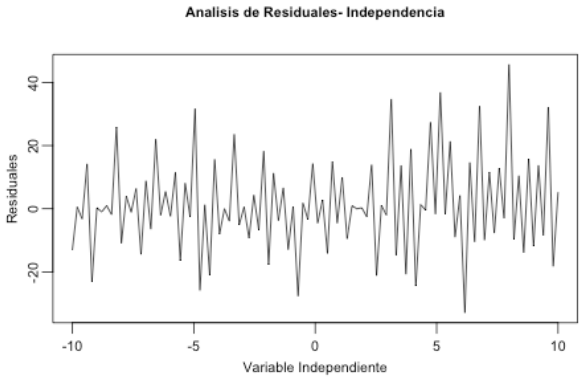
# Validación del Modelo : Independencia

Generalemente para verificar que tenemos independencia de los errores se suele hacer una gráfica de los residuales contra el orden de obtención de los datos que casi siempre es contra la variable $X_1$. Luego en el gráfico generado se busca algún tipo de patron que evidencíe la dependencia de los errores conforme se van realizando las mediciones.
Sin embargo a veces no es fácil detectar un patron por lo que se pueden aplicar pruebas estadísticas para detectar patrones en los datos, las pruebas mas usuales son las que útilizan la función de autocorrelación **(ACF)** y la prueba no parámetrica de **Rachas** que va midiendo el cambio de los signos de los errores.



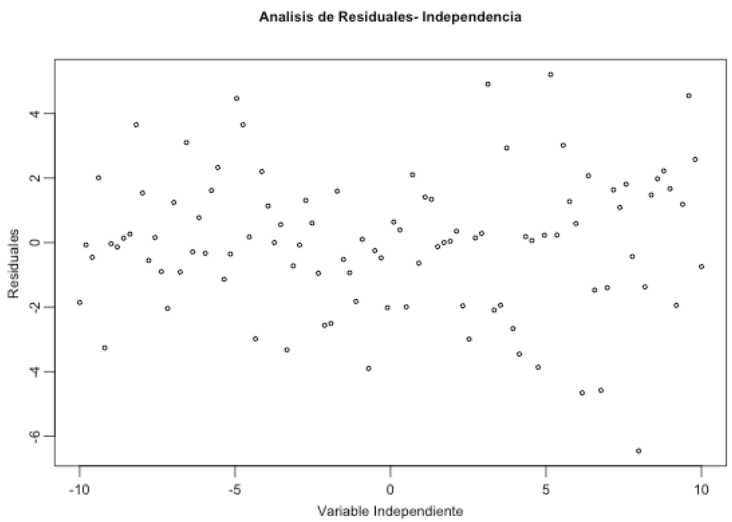
# Validación del Modelo : Independencia

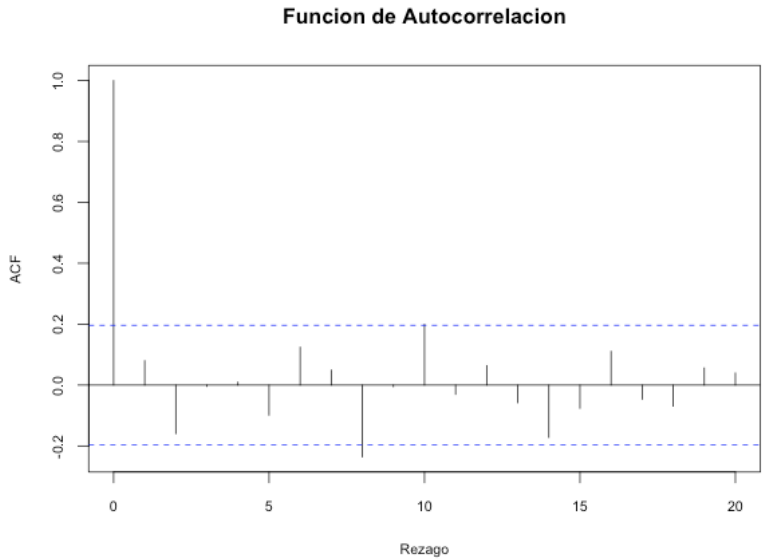
- Función de Autocorrelación (ACF). En estadística, la autocorrelación de una serie de datos discreta de un proceso $X_t$ no es más que el coeficiente de correlación de dicho proceso con una versión desplazada de la propia serie. La forma en como se calcula es la siguiente, suponiendo que tenemos a la serie de datos $(e_1, e_2, \ldots, e_n)$ entonces la función de autocorrelación con rezago k se obtiene como:

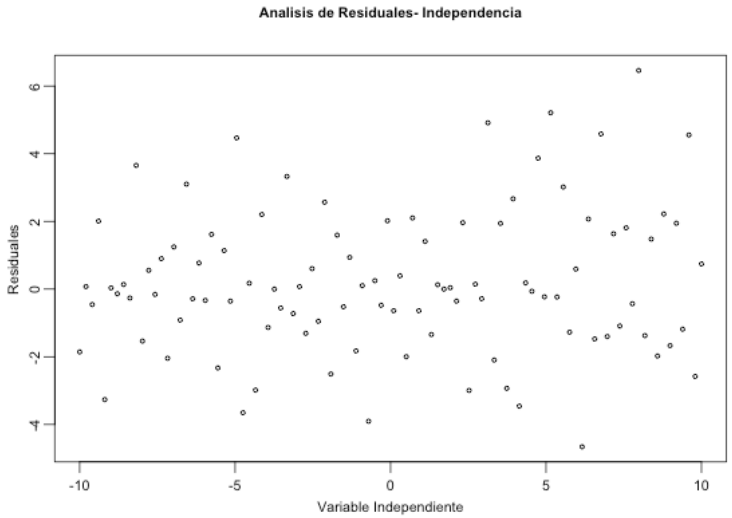
$$\rho_k = \frac{\sum_{i=1}^{n-k} (e_i - \bar{e})(e_{i-k} - \bar{e})}{\sum_{i=1}^{n} (e_i - \bar{e})^2}$$

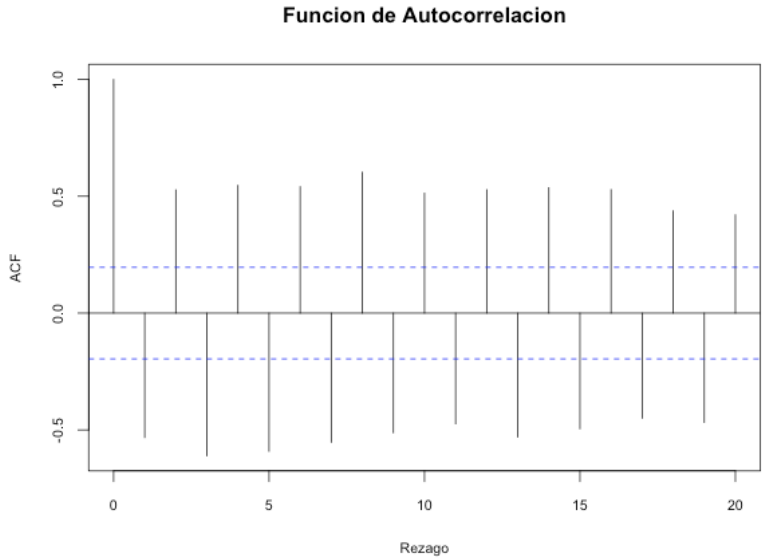
Si $\rho_k$ se aleja del valor 0 hay evidencia de que cada $k$ observaciones hay un patron de los residuales pues dichas observaciones están muy correlacionadas lo que implicaría dependencia de los residuales. La idea entonces es contrastar la hipótesis $\rho_k = 0$ vs $\rho_k \neq 0$, afortunadamente R tiene programada la prueba en la función *acf*

# Validación del Modelo : Independencia :: ACF



# Validación del Modelo : Independencia :: Rachas

Otra forma que tenemos de verificar que los residuales son independientes es probar la aleatoreidad con la que van cambiando de signo los errores, si los residuales son independientes se esperaría que el cambio de signo del residual conforme se va obteniendo la muestra es aleatorio.

**Definition (Racha)**

Se considera una Racha de tamaño $k$ a la secuencia de $k$ de valores consecutivos de un mismo signo siempre y cuando estos sean precedidos y seguidos por valores con signo opuesto a la de la Racha.

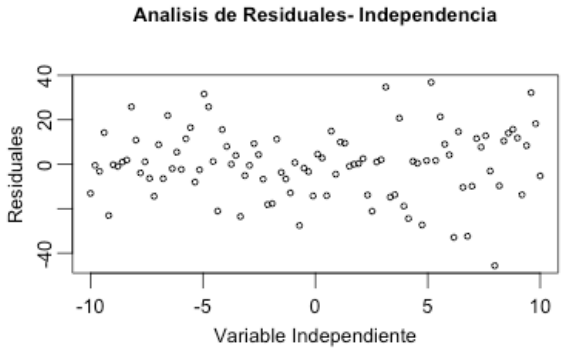
$$\underbrace{+,+}, \underbrace{-}, \underbrace{+}, \underbrace{-}$$

La idea de la prueba es contar el número de rachas en la muestra, luego un número reducido o grande de rachas es indicio de que las observaciones no se han obtenido de forma aleatoria .
Si la muestra es grande y la hipótesis de aleatoriedad es cierta la distribución muestral del número de rachas R, puede aproximarse mediante una distribución Normal de parámetros:

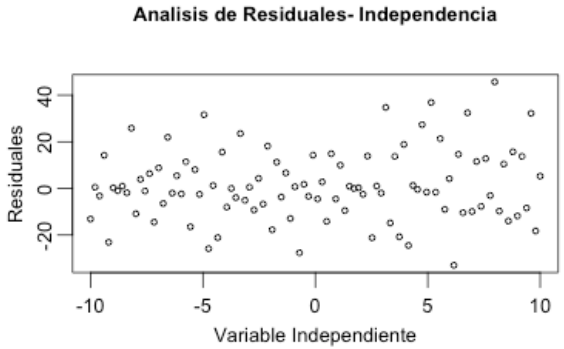
$$\mu_R = \frac{2n_1n_2}{n} \quad \sigma_R^2 = \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}$$

# Validación del Modelo : Independencia :: Rachas

La prueba de Rachas se encuentra programada de R y se encuentra dentro de la librería *tseries* bajo el nombre *runs.test*, esta función recibe un vector con valores binarios indicando el signo del residual.
Ejemplo:




```
library(tseries)
runs.test(as.factor(res>0))
Runs Test
data: as.factor(res > 0)
Standard Normal = 0.004, p-value = 0.9968
alternative hypothesis: two.sided
```




```
library(tseries)
runs.test(as.factor(res2>0))
Runs Test
data: as.factor(res2 > 0)
Standard Normal = 9.8499, p-value < 2.2e-16
alternative hypothesis: two.sided
```

# Transformaciones

En caso en el que el análisis de los residuos no permita validar el modelo, bien por:

- Falta de linealidad en la relación entre las variables $X_1$ e $Y$
- Falta de homocedasticidad
- Falta de Normalidad

En ocasiones se puede obtener un modelo lineal que si verifique las hipótesis a través de transformaciones en $X$ y $Y$ o en ambas. Algunos Modelos Linealizables son:

| Modelo Real (Desconocido) | Transformación | Modelo Lineal |
|---|---|---|
| $Y = \beta_0 + \beta_1 X^k$ | $Z = X^k$ | $Y = \beta_0 + \beta_1 Z$ |
| $Y = \beta_0 + \beta_1 ln(X)$ | $Z = ln(X)$ | $Y = \beta_0 + \beta_1 Z$ |
| $Y = \beta_0 e^{\beta_1 X}$ | $V = ln(Y)$ | $V = ln(\beta_0) + \beta_1 X$ |
| $Y = \beta_0 x^{\beta_1}$ | $V = ln(Y); Z = ln(X)$ | $V = ln(\beta_0) + \beta_1 Z$ |

# Breve introducción al algebra matricial

**Definición (Matriz)**

*Una **matriz** es un arreglo (bidimensional) de números o elementos algebraicos.*

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix}_{n \times m}$$

Se dice que la matriz es cuadrada cuando $n = m$.
El espacio de las matrices con entradas de números reales se le denota como $\mathbb{R}^{(n \times m)}$, por lo que aveces se denota $\mathbf{A} \in \mathbb{R}^{(n \times m)}$, o bien $(a_{ij}) \in \mathbb{R}^{(n \times m)}$
Ejemplos:

- Matriz Identidad:

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}_{n \times n}$$

# Breve introducción al algebra matricial: Matrices

Operaciones Básicas

- **Suma de Matrices.** Sea $\mathbf{A}$ y $\mathbf{B} \in \mathbb{R}^{(n\times m)}$ se define la suma de matrices $\mathbf{A}+\mathbf{B}=\mathbf{C}$ como:

$$(c_{ij}) = (a_{ij}) + (b_{ij}) \quad i \in \{1,\ldots,n\} \quad j \in \{1,\ldots,m\}$$

- **Multiplicación de Matrices.** Sea $\mathbf{A} \in \mathbb{R}^{(n\times m)}$ y $\mathbf{B} \in \mathbb{R}^{(m\times k)}$ se define la multiplicación de matrices $\mathbf{AB}=\mathbf{C} \in \mathbb{R}^{(n\times k)}$ como:

$$(c_{ij}) = \sum_{q=1}^{m} a_{iq}b_{qj} \quad i \in \{1,\ldots,n\} \quad j \in \{1,\ldots,k\}$$

  Ejercicio: (*Expresar el SCE como producto matricial, *Expresar una combinación lineal como producto matricial )

**Matriz inversa:**
Sea $\mathbf{A} \in \mathbb{R}^{n\times n}$ una matriz cuadrada. La inversa de $\mathbf{A}$, denotada por $\mathbf{A}^{-1}$, es otra matriz $n \times n$ tal que:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

Si la inversa existe, es única.

- **Propiedad Distributiva:** Sea $\mathbf{A}$; $\mathbf{B} \in \mathbb{R}^{(n \times m)}$ y $\mathbf{C} \in \mathbb{R}^{(k \times n)}$ entonces:

$$\mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB}$$

Por ejemplo, supongamos $\mathbf{A}$; $\mathbf{B}$; $\mathbf{C} \in \mathbb{R}^{(n \times n)}$ entonces:

$$\left(\mathbf{A}^2 + \mathbf{ABA}\right) = (\mathbf{A} + \mathbf{AB})\,\mathbf{A} = \mathbf{A}(\mathbf{I} + \mathbf{B})\,\mathbf{A}$$

- **Transpuesta de una Matriz**. Sea $\mathbf{A} \in \mathbb{R}^{(n\times m)}$. La matriz transpuesta de $\mathbf{A}$ denotada por $\mathbf{A}^T$, es una matriz en $\mathbb{R}^{(m\times n)}$ tal que sus columnas son los renglones de $\mathbf{A}$ es decr:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix} \Rightarrow \mathbf{A}^T = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \dots & a_{nm} \end{pmatrix}$$

  Observación: Si $\mathbf{A}$ es una matriz $(n \times m)$ y $\mathbf{B}$ una matriz $(m \times k)$ entonces:

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$

  Observación: Si $\mathbf{A}$ es una matriz $(n \times m)$ y $\mathbf{B}$ un matriz $(n \times m)$ entonces:

$$(\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$$

- **Matriz Simétrica**. Sea $\mathbf{A} \in \mathbb{R}^{(n\times n)}$. Se dice que $\mathbf{A}$ se llama simétrica si $\mathbf{A} = \mathbf{A}^T$
- **Matriz Idempotente**. Sea $\mathbf{A} \in \mathbb{R}^{(n\times n)}$. Se dice que $\mathbf{A}$ es idempotente si $\mathbf{A} = \mathbf{AA} = \mathbf{A}^2$
  Observación: Si $\mathbf{A}$ es simétrica e idempotente, entonces $\mathbf{I} - \mathbf{A}$ es tambien simétrica e idempotente.

- **Matriz Ortogonal.** Sea $\mathbf{A} \in \mathbb{R}^{(n\times n)}$. Se dice que $\mathbf{A}$ es ortogonal si $\mathbf{A}^T\mathbf{A} = \mathrm{I}$, en consecuencia, si $\mathbf{A}$ es ortogonal entonces $\mathbf{A}^{-1} = \mathbf{A}^T$
- **Forma Cuadrática** Sea $\underline{y}$ un vector $(n \times 1)$ y sea $\mathbf{A} \in \mathbb{R}^{(n\times n)}$ entonces la función $f : \mathbb{R}^n \to \mathbb{R}$ con regla de correspondencia:

$$f(\underline{y}) = \underline{y}^T\mathbf{A}\underline{y} = \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}y_iy_j$$

  es llamada una forma cuadrática. $\mathbf{A}$ es llamada la matriz de la forma cuadrática. (Ejercicio: Expresar a SCE como una forma cuadrática en función de $\underline{Y}$)

- **Matriz Definida Positiva y Semidefinida Positiva** Sea $\mathbf{A} \in \mathbb{R}^{(n\times n)}$. Se dice que $\mathbf{A}$ es positiva definida si las siguientes condiciones se cumplen:
  - $\mathbf{A} = \mathbf{A}^T$ (Es simétrica)
  - $\underline{y}^T\mathbf{A}\underline{y} > 0 \quad \forall \underline{y} \in \mathbb{R}^n \quad \underline{y} \neq 0$ (Positiva Definida)
  - $\underline{y}^T\mathbf{A}\underline{y} \geq 0 \quad \forall \underline{y} \in \mathbb{R}^n \quad \underline{y} \neq 0$ (Positiva Semidefinida)

- **Traza de una Matriz**. Sea $\mathbf{A} \in \mathbb{R}^{(n \times n)}$. La traza de $\mathbf{A}$, denotada por $tr(\mathbf{A})$, es la suma de los elementos en la diagonal de $\mathbf{A}$

$$tr(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$$

Algunas propiedades de la traza son:

- Si $\mathbf{A} \in \mathbb{R}^{(m \times n)}$ y $\mathbf{B} \in \mathbb{R}^{(n \times m)}$ entonces:

$$tr(\mathbf{AB}) = tr(\mathbf{BA})$$

- Si $\mathbf{A} \in \mathbb{R}^{(m \times k)}$ ; $\mathbf{B} \in \mathbb{R}^{(k \times n)}$ y $\mathbf{C} \in \mathbb{R}^{(n \times m)}$ entonces:

$$tr(\mathbf{ABC}) = tr(\mathbf{CAB})$$

- Si $\mathbf{A} \in \mathbb{R}^{(n \times n)}$ y $\mathbf{B} \in \mathbb{R}^{(n \times n)}$ y $a, b$ dos escalares, entonces:

$$tr(a\mathbf{A} + b\mathbf{B}) = a(tr(\mathbf{A})) + b(tr(\mathbf{B}))$$

- **Rango de Matriz** Sea $\mathbf{A} \in \mathbb{R}^{(m \times n)}$, se define el rango de $\mathbf{A}$ como el número de columnas linealmente independientes, equivalentemente es el número de renglones linealmente independientes. Ejemplo:

$$Rank\left(\mathbf{I}_{n \times n}\right) = n$$

- **Rango de Matriz Idempotente**. Sea $\mathbf{A} \in \mathbb{R}^{(n \times n)}$ una matriz idempotente, entonces:

$$Rank\left(\mathbf{A}\right) = tr\left(\mathbf{A}\right)$$

# Cálculo diferencial matricial

Una función real de varias variables es aquella que asocia a un vector $\underline{x} \in \mathbb{R}^n$ a un único valor real $x \in \mathbb{R}$, es decir:

$$f : \mathbb{R}^n \to \mathbb{R} \quad f(\underline{x}) = x$$

La derivada de una función de variables variables se puede calcular derivando parcialmente respecto a cada componente formando a si un vector columna al que se le denomina gradiente de $f$ es decir:

$$\bigtriangledown f = \frac{\partial}{\partial \underline{x}} f = \left( \frac{\partial}{\partial x_1} f(\underline{x}), \frac{\partial}{\partial x_2} f(\underline{x}), \ldots, \frac{\partial}{\partial x_n} f(\underline{x}) \right)^T$$

Supongamos ahora que tenemos $\mathbf{A} \in \mathbb{R}^{(n \times n)}$ , $\underline{a}$ un vector columna, es decir $\underline{a} \in \mathbb{R}^{(n \times 1)}$ y $\underline{x}$ un vector columna de variables. Entonces:

- Si $f(\underline{x}) = \underline{a}^T \underline{x}$ o si $f(\underline{x}) = \underline{x}^T \underline{a}$ entonces:

$$\bigtriangledown f(\underline{x}) = \frac{\partial}{\partial \underline{x}} f = \underline{a}$$

# Cálculo diferencial matricial

- (Regla de la suma). Supongamos que tenemos dos funciones de varias variables $f : \mathbb{R}^n \to \mathbb{R}$ y $g : \mathbb{R}^n \to \mathbb{R}$. Definamos, la función $h : \mathbb{R}^n \to \mathbb{R}$ como $h(\underline{x}) = f(\underline{x}) + g(\underline{x})$ entonces:

$$\bigtriangledown h(\underline{x}) = \bigtriangledown f(\underline{x}) + \bigtriangledown g(\underline{x})$$

- Si $f(\underline{x}) = \underline{x}^T \underline{x}$ entonces:

$$\bigtriangledown f(\underline{x}) = \frac{\partial}{\partial \underline{x}} f = 2\underline{x}$$

- Si $f(\underline{x}) = \underline{a}^T \mathbf{A} \underline{x}$ entonces:

$$\bigtriangledown f(\underline{x}) = \frac{\partial}{\partial \underline{x}} f = \mathbf{A}^T \underline{a}$$

- (Regla del producto). Supongamos que tenemos dos funciones de varias variables $f : \mathbb{R}^n \to \mathbb{R}$ y $g : \mathbb{R}^n \to \mathbb{R}$. Definamos, la función $h : \mathbb{R}^n \to \mathbb{R}$ como $h(\underline{x}) = f(\underline{x})g(\underline{x})$ entonces:

$$\bigtriangledown h(\underline{x}) = f(\underline{x}) \bigtriangledown g(\underline{x}) + g(\underline{x}) \bigtriangledown f(\underline{x})$$

# Cálculo diferencial matricial

- Si $f(\underline{x}) = \underline{x}^T \mathbf{A} \underline{x}$ entonces:

$$\bigtriangledown f(\underline{x}) = \frac{\partial}{\partial \underline{x}} f = \mathbf{A}\underline{x} + \mathbf{A}^T \underline{x}$$

- Si $f(\underline{x}) = \underline{x}^T \mathbf{A} \underline{x}$ y $\mathbf{A}$ es simétrica entonces:

$$\bigtriangledown f(\underline{x}) = \frac{\partial}{\partial \underline{x}} f = 2\mathbf{A}\underline{x}$$

# Cálculo diferencial matricial

Puntos Críticos de una función de varias variables:

**Teorema (Condiciones necesarias de extremo relativo)**

*Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función real definida en conjunto un abierto de $\mathbb{R}^n$ y $\underline{a} \in \mathbb{R}^n$. Si $f$ admite derivadas parciales en $\underline{a}$ y alcanza un máximo o mínimo relativo en dicho punto, entonces se verifica:*

$$\bigtriangledown f(\underline{a}) = \underline{0}$$

Luego entonces si queremos encontrar puntos críticos de una función debemos resolver el sistema de ecuaciones

$$\bigtriangledown f(\underline{x}) = \underline{0}$$

# Vectores Aleatorios y sus Propiedades

A lo largo de este curso se ha venido trabajando con variables aleatorias reales, es decir, variables que modelan una características numérica de algún fenómeno aleatorio. Por ejemplo $X =$ Número que se obtendrá en el lanzamiento de un dado. Debe observarse entonces que $X$ es una variable que toma valores reales a saber $\{1, 2, 3, 4, 5, 6\}$.
Suponga ahora que queremos modelar varias características numéricas a un fenómeno aleatorio, por ejemplo, se lanza un dardo en un plano cartesiano, sea $X_1 =$ La posición en el eje de las abscisas donde cae el dardo y $X_2 =$ La posición en el eje de las ordenadas donde cae el dardo. Este fenómeno aleatorio se modela entonces por dos variables simultáneamente $(X_1, X_2)$.
Cuando tenemos arreglos de variables aleatorias decimos que tenemos un **vector aleatorio.**

**Definición (Vector Aleatorio)**

*Diremos que* $\underline{X} := (X_1, X_2, \ldots, X_n)^T$ *es un vector aleatorio en* $\mathbb{R}^n$ *si cada componente de este vector* $X_i$ *es una variable aleatoria real*

# Vectores Aleatorios y sus Propiedades

## Definición (Distribución de un Vector Aleatorio)

*Sea* $\underline{X} := (X_1, X_2, \ldots, X_n)^T$ *un vector aleatorio en* $\mathbb{R}^n$*, definimos la función de distribución de* $\underline{X}$ *como:*

$$F_{\underline{X}}(x_1, \ldots . x_n) := \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

## Definición (Densidad de un Vector Aleatorio caso Discreto)

*Sea* $\underline{X} := (X_1, X_2, \ldots, X_n)^T$ *un vector aleatorio en* $\mathbb{R}^n$*, tal que cada* $X_i$ *es una v.a. Discreta. Definimos la densidad de* $\underline{X}$ *como :*

$$f_{\underline{X}}(x_1, \ldots . x_n) := \mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

## Definición (Densidad de un Vector Aleatorio caso absolutamente continuo)

*Sea* $\underline{X} := (X_1, X_2, \ldots, X_n)^T$ *un vector aleatorio en* $\mathbb{R}^n$*, tal que cada* $X_i$ *es una v.a. absolutamente continua. Definimos la densidad de* $\underline{X}$ *como aquella función* $f_{\underline{X}} : \mathbb{R}^n \to \mathbb{R}$ *tal que toda región* $D \in \mathbb{R}^n$ *se tiene que :*

$$\mathbb{P}(\underline{X} \in D) = \int_D f_{\underline{X}}(x_1, x_2, \ldots, x_n)$$

# Vectores Aleatorios y sus Propiedades

## Definición (Esperanza de un Vector Aleatorio)

*Sea* $\underline{X} := (X_1, X_2, \ldots, X_n)^T$ *un vector aleatorio en* $\mathbb{R}^n$. *Se define la esperanza del vector* $\underline{X}$ *como:*

$$\mathbb{E}(\underline{X}) := (\mathbb{E}(X_1), \mathbb{E}(X_2), \ldots . \mathbb{E}(X_n))^T$$

## Definición (Varianza de un Vector Aleatorio)

*Sea* $\underline{X} := (X_1, X_2, \ldots, X_n)$ *un vector aleatorio en* $\mathbb{R}^n$. *Se define la varianza del vector* $\underline{X}$ *como la siguiente matriz:*

$$Var(\underline{X}) := \Sigma_{\underline{X}} := \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & Cov(X_1, X_3) & \ldots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Var(X_2) & Cov(X_2, X_3) & \ldots & Cov(X_2, X_n) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & Var(X_3) & \ldots & Cov(X_3, X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & Cov(X_n, X_3) & \ldots & Var(X_n) \end{pmatrix}$$

$$Var(\underline{X}) := \mathbb{E}\Big((\underline{X} - \mathbb{E}(\underline{X}))(\underline{X} - \mathbb{E}(\underline{X}))^T\Big)$$

Obs: Se puede probar que $\Sigma_{\underline{X}}$ es una matriz simétrica y definida positiva.

# Vectores Aleatorios y sus Propiedades

Propiedades importantes de los vectores aleatorios:
Sea $\mathbf{A} \in \mathbb{R}^{(n\times n)}$ un matriz cuadrada de constantes, $\underline{a}$ un vector columna de constantes $(k \times 1)$ y $\underline{X}$ un vector aleatorio en $\mathbb{R}^n$ tal que $\mathbb{E}(\underline{X}) = \underline{\mu}$ y $\text{Var}(\underline{X}) = \Sigma_{\underline{X}}$

- $\mathbb{E}\left(\underline{a}^T\underline{X}\right) = \underline{a}^T\underline{\mu}$
- $\mathbb{E}(\mathbf{A}\underline{X}) = \mathbf{A}\underline{\mu}$
- $\text{Var}\left(\underline{a}^T\underline{X}\right) = \underline{a}^T\Sigma_{\underline{X}}\underline{a}$
- $\text{Var}(\mathbf{A}\underline{X}) = \mathbf{A}\Sigma_{\underline{X}}\mathbf{A}^T$
- Si $\Sigma_{\underline{X}} = \sigma^2\mathbf{I}$ entonces $\text{Var}(\mathbf{A}\underline{X}) = \sigma^2\mathbf{A}\mathbf{A}^T$
- $\mathbb{E}\left(\underline{X}^T\mathbf{A}\underline{X}\right) = tr\left(\mathbf{A}\Sigma_{\underline{X}}\right) + \underline{\mu}^T\mathbf{A}\underline{\mu}$
- Si $\Sigma_{\underline{X}} = \sigma^2\mathbf{I}$ entonces $\mathbb{E}\left(\underline{X}^T\mathbf{A}\underline{X}\right) = \sigma^2 tr\left(\mathbf{A}\right) + \underline{\mu}^T\mathbf{A}\mu$

# Normal Multivariada y sus propiedades

**Normal Univariante:**
Sabemos que si $X \sim N_1(\mu, \sigma^2)$ entonces $\mathbb{E}(X) = \mu$ y $\text{Var}(X) = \sigma^2 > 0$, y tiene por densidad:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad x \in \mathbb{R}$$

**Normal Multivariante:**
La generalización de la normal univariante al caso multivariado es la siguiente:
Decimos que el vector aleatorio $\underline{X}$ sigue una distribución Normal Multivariada de orden $p$ o $p$-variada denotada por $\underline{X} \sim N_p\left(\underline{\mu}, \Sigma_{p\times p}\right)$, donde $\underline{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)^T$ y $\Sigma$ es una matriz simétrica definida positiva, si la función de densidad de $\underline{X}$ está dada por:

$$f_{\underline{X}}(\underline{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} exp\left(-\frac{1}{2}\left(\underline{x}-\underline{\mu}\right)^T \Sigma^{-1} \left(\underline{x}-\underline{\mu}\right)\right) \quad \underline{x} \in \mathbb{R}^p$$

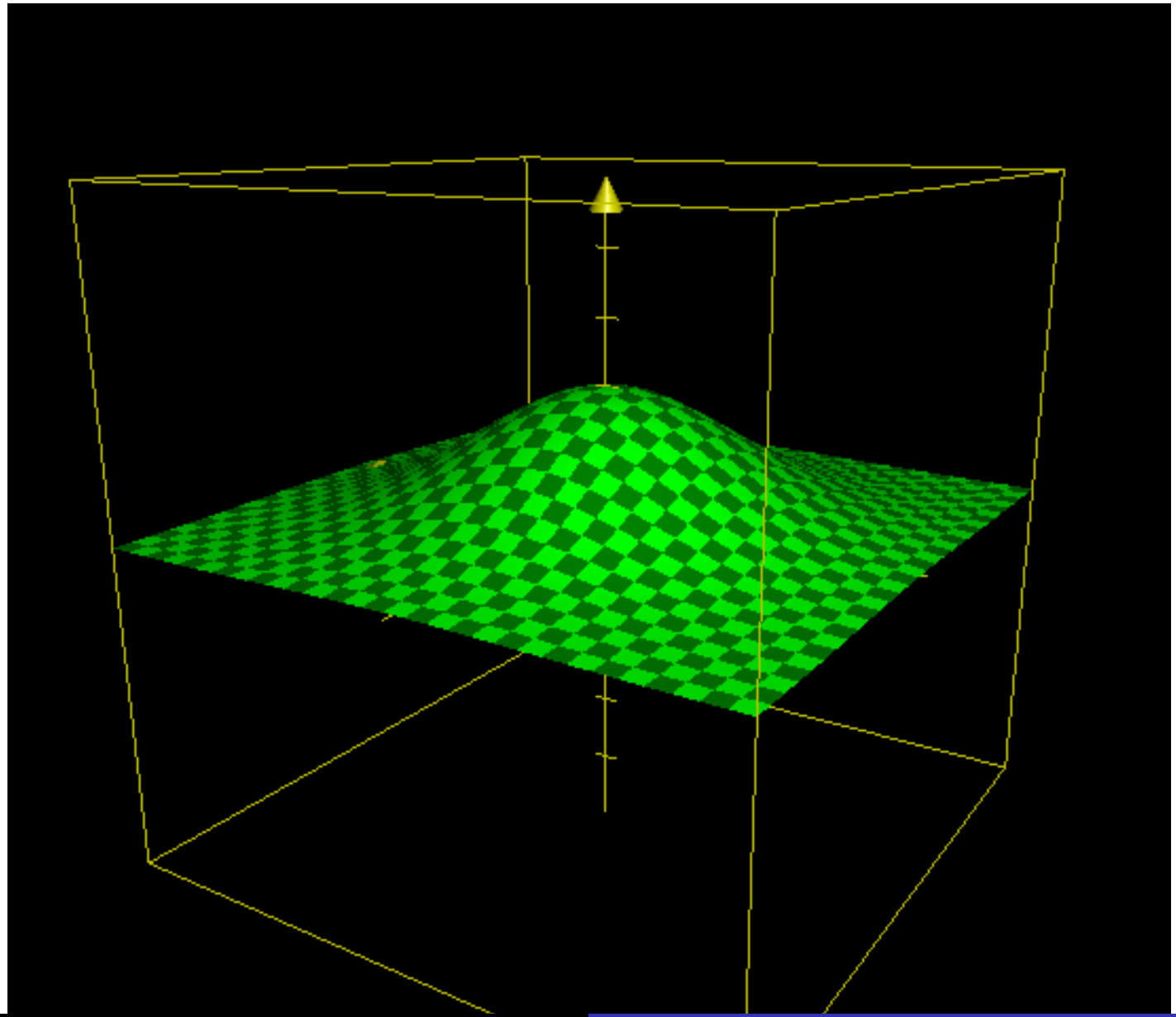
# Normal Multivariada y sus propiedades

**Ejemplo:**
Supongamos que $p = 2$ (Normal Bi-variada) y que $\mu = (0,0)^T$, $\Sigma = \sigma^2 \mathbf{I}_{2\times 2}$ entonces:

$$\begin{aligned} f_{\underline{X}}(x_1,x_2) &= \frac{1}{(2\pi)\sigma^2} exp\left(-\frac{1}{2}(x_1,x_2)^T\left(\sigma^2\mathbf{I}_{2\times 2}\right)^{-1}(x_1,x_2)\right); \quad (x_1,x_2)\in\mathbb{R}^2 \\ &= \frac{1}{(2\pi)\sigma^2} exp\left(-\frac{1}{2\sigma^2}\left(x_1^2+x_2^2\right)\right); \quad (x_1,x_2)\in\mathbb{R}^2 \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{x_1^2}{2\sigma^2}\right)\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{x_2^2}{2\sigma^2}\right); \quad (x_1,x_2)\in\mathbb{R}^2 \\ &= f_{X_1}(x_1) * f_{X_2}(x_2) \end{aligned}$$

Es decir, recuperamos la densidad conjunta como el producto de las marginales. Esta es una propiedad que ya sabamos, pues cuando $x_1, x_2 \sim N_1(0,\sigma^2)$ independientes, entonces la conjunta se obtiene multiplicando las marginales.
Obs: En este caso se verifica el hecho de que si $Cov(x_1,x_2) = 0$ entonces $x_1$ es independiente de $x_2$

# Normal Multivariada y sus propiedades



Suponga que $\underline{X} \sim N_p\left(\underline{\mu}, \Sigma\right)$ y que $\underline{a} = (a_1, a_2, \ldots, a_p)^T$. Entonces:

- La componente i-ésima del vector $\underline{X}$ sigue una distribución normal con parámetros $\left(\mu_i, \sigma_{ii}^2\right)$ es decir $X_i \sim N\left(\mu_i, \sigma_{ii}^2\right)$
- $\underline{a}^T\underline{X} \sim N_1\left(\underline{a}^T\underline{\mu}, \underline{a}^T\Sigma\underline{a}\right)$
- Si para toda $\underline{a} \in \mathbb{R}^p$ (visto como vector columna) se tiene que $\underline{a}^T\underline{Y}$ es normal entonces $\underline{Y}$ es normal $p$-variante
- Supongamos que participamos al vector $\underline{X}$ de la siguiente manera $\left(\underline{X_1}, \underline{X_2}\right)$ donde $\underline{X_1} \in \mathbb{R}^q$ y $\underline{X_2} \in \mathbb{R}^{p-q}$ con $1 \leq q \leq p-1$. Entonces, $\underline{X_1}$ sigue una distribución Normal $q$-variada y $\underline{X_2}$ sigue una distribución Normal $p-q$-variada con los siguientes parámetros.

$$\underline{X_1} \sim N_q\left(\underline{\mu_1}, \Sigma_{11}\right) \qquad \underline{X_2} \sim N_{p-q}\left(\underline{\mu_2}, \Sigma_{22}\right)$$

Donde:

$$\underline{\mu} = \begin{pmatrix} \underline{\mu_1} \\ \underline{\mu_2} \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Suponga que $\underline{X} \sim N_p\left(\underline{\mu}, \Sigma\right)$ y que $\mathbf{A} \in \mathbb{R}^{m \times p}$ (De rango completo). Entonces:

$$\mathbf{A}\underline{X} \sim N_m\left(\mathbf{A}\underline{\mu}, \mathbf{A}\Sigma\mathbf{A}^T\right)$$

Cualquier transformación lineal de un vector aleatorio Normal $p$-variado se distribuye Normal. Ejemplo:

- Si $\mathbf{A} \in \mathbb{R}^{1 \times p}$ (un vector columna en $\mathbb{R}^p$) entonces replicamos la segunda propiedad del slide anterior pues:

$$\mathbf{A}\underline{X} \sim N_1\left(\mathbf{A}\underline{\mu}, \mathbf{A}\Sigma\mathbf{A}^T\right)$$

- Si $\mathbf{A} \in \mathbb{R}^{q \times p}$ con $1 \leq q \leq p-1$ tal que:

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}_{q \times q} & 0_{q \times p-q} \end{pmatrix}$$

Entonces

$$\underline{X}_1 = \mathbf{A}\underline{X} \sim N_q\left(\underline{\mu}_1, \Sigma_{11}\right)$$

Es decir replicamos la cuarta propiedad del slide anterior.

# Planteamiento del Problema

Ahora nos enfrentamos al problema de tener $k$ funciones de las covariables y una variable respuesta. Entonces se plantea ajustar un modelo lineal de la forma:

$$Y_i = \beta_0 + \beta_1 s_1(\underline{z}_i) + \beta_2 s_2(\underline{z}_i) + \cdots + \beta_k s_k(\underline{z}_i) + \varepsilon_i$$

Donde definiendo:

$$x_{ij} = s_j\left(\underline{z}_i\right) \quad i \in \{1, \ldots, n\}; \quad j \in \{1, \ldots, k\}$$

se transforma en el modelo:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

Donde $\varepsilon_i$ una variable aleatoria que representa el error en la relación. $\beta_i$ son parámetros desconocidos en el modelo y tales que

$$\varepsilon_i \sim F\left(0, \sigma^2\right) \quad \mathbb{E}(\varepsilon_i) = 0 \quad var\left(\varepsilon_i\right) = \sigma^2;$$

Obs: Note que las $k$ funciones de las covariables $\underline{z}$ dan origen a $k$ variables que denotamos por $x$, de ahi que digamos que se ajusta un modelo con $k$ variables.

# Planteamiento

Supongamos entonces que observaremos $n$ observaciones de las $k$ funciones de las variables explicativas con su respectiva variable respuesta, es decir:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_k x_{1k} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_k x_{2k} + \varepsilon_2 \\ \vdots &= \vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_k x_{nk} + \varepsilon_n \end{aligned}$$

En nuestro modelo general nuevamente impondremos la hipótesis de no correlación entre los errores, es decir $Cov\left(\varepsilon_i, \varepsilon_j\right) = 0$ con $i \neq j$. Se plantea el problema entonces de encontrar los $k+2$ parámetros asociados al modelo lineal:

$$\left(\beta_0, \beta_1, \ldots, \beta_k, \sigma^2\right)$$

# Planteamiento

Para facilitar notación y dar una solución mas elegante al problema se establece la siguiente notación matricial. $(p = k + 1)$

$$\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$$

Donde:

$$\underline{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}_{n\times p} \qquad \underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1}$$

El vector de parámetros:

$$\underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}_{p\times 1}$$

Ejercicio: (Expresar el modelo lineal simple en forma matricial)

# Estimación por mínimos cuadrados

El objetivo de esta técnica es minimizar la diferencia al cuadrado entre los valores reales y los ajustados por el modelo, para ello definamos el vector de valores ajustados como:

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

Luego entonces el objetivo es minimizar respecto a $\underline{\hat{\beta}}$ a la siguiente expresión:

$$\begin{aligned} f\left(\underline{\hat{\beta}}\right) &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \left(\underline{Y} - \underline{\hat{Y}}\right)^t \left(\underline{Y} - \underline{\hat{Y}}\right) = \left(\underline{Y} - \mathbf{X}\underline{\hat{\beta}}\right)^t \left(\underline{Y} - \mathbf{X}\underline{\hat{\beta}}\right) \\ &= \underline{Y}^t\underline{Y} - 2\underline{\hat{\beta}}^t\mathbf{X}^t\underline{Y} + \underline{\hat{\beta}}^t\mathbf{X}^t\mathbf{X}\underline{\hat{\beta}} \end{aligned}$$

Derivando e igualando a 0 para encontrar el punto crítico:

$$\nabla f\left(\underline{\hat{\beta}}\right) = \frac{\partial}{\partial\underline{\hat{\beta}}} f\left(\underline{\hat{\beta}}\right) = -2\mathbf{X}^t\underline{Y} + 2\mathbf{X}^t\mathbf{X}\underline{\hat{\beta}}$$

$$\nabla f\left(\underline{\hat{\beta}}\right) = \frac{\partial}{\partial\underline{\hat{\beta}}} f\left(\underline{\hat{\beta}}\right) = 0 \Rightarrow -2\mathbf{X}^t\underline{Y} + 2\mathbf{X}^t\mathbf{X}\underline{\hat{\beta}} = \underline{0}$$

# Estimación por mínimos cuadrados

Finalmente obtenemos que el estimador por mínimo cuadrados se obtiene al solucionar la denominada ecuación normal:

$$\mathbf{X}^t\mathbf{X}\underline{\hat{\beta}} = \mathbf{X}^t\underline{Y} \tag{11}$$

$$\underline{\hat{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\underline{Y} \tag{12}$$

Debe notarse que los estimadores $\hat{\beta}_i$ son combinaciones lineales de las variables respuestas y que se tiene solución única a la ecuación normal cuando la matriz $\mathbf{X}\mathbf{X}^t$ tiene inversa lo cual ocurre si y sólo si $\mathbf{X}$ tiene rango completo, lo que equivale a pedir que las variables $(X_1, X_2, \ldots, X_k)$ son linealmente independientes. Esto impone una nueva hipótesis al modelo lineal, la cual nos indica que una variable no puede ser combinación lineal de otras, en caso de que esto ocurra, se deberá analizar el modelo y reducir el número de variables.
(Ejercicio: Encuentre los estimadores por mínimos cuadrados del modelo lineal simple con esta nueva formula y verifique que se obtiene los mismos estimadores)

# El problema de la Multicolinealidad

Ejemplo:
Supongamos que tenemos el siguiente modelo de 3 variables:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Sin embargo suponga que $x_2 = \alpha_0 + \alpha_1 x_1 + \alpha_3 x_3$, entonces, sustituyendo en la ecuación anterior

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 \left(\alpha_0 + \alpha_1 x_1 + \alpha_3 x_3\right) + \beta_3 x_3 + \varepsilon \\ Y &= \underbrace{\left(\beta_0 + \beta_2 \alpha_0\right)}_{\beta_0'} + \underbrace{\left(\beta_1 + \alpha_1 \beta_2\right)}_{\beta_1'} x_1 + \underbrace{\left(\beta_3 + \alpha_3 \beta_2\right)}_{\beta_3'} x_3 + \varepsilon \\ Y &= \beta_0' + \beta_1' x_1 + \beta_3' x_3 + \varepsilon \end{aligned}$$

Lo que indica que en realidad el modelo lineal es de 2 variables explicativas y no de 3. La colinealidad de las variables se puede detectar analizando la matriz de correlación, una alta correlación entre las variables puede ocasionar problemas con el cálculo de la inversa de $\left(\mathbf{X}^t\mathbf{X}\right)$ (Problema Inestable)

# Estimación por mínimos cuadrados

Hemos probado que el estimador por mínimos cuadrados está dado por $\underline{\hat{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\underline{Y}$, luego como $\underline{\hat{Y}} = \mathbf{X}\underline{\hat{\beta}}$ entonces: $\underline{\hat{Y}} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\underline{Y}$ Si definimos:

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t$$

Entonces se tiene que $\underline{\hat{Y}} = \mathbf{H}\underline{Y}$. En la literatura a la matriz $\mathbf{H}$ se le conoce como la matriz sombrero (hat) y tiene la propiedad de mapear al vector de valores observados en el vector de valores ajustados. Observaciones:

- $\mathbf{H}$ es simétrica
- $\mathbf{H}$ es idempotente ($\mathbf{HH} = \mathbf{H}^2 = \mathbf{H}$)
- El vector de residuales $\underline{e}$ pueden expresarse como $\underline{e} = (\mathbf{I} - \mathbf{H})\underline{Y}$

# Propiedades de los estimadores

- $\mathbb{E}(\underline{Y}) = \mathbb{E}\left(\mathbf{X}\underline{\beta} + \underline{\varepsilon}\right) = \mathbf{X}\underline{\beta}$
- $\text{Var}(\underline{Y}) = \text{Var}\left(\mathbf{X}\underline{\beta} + \underline{\varepsilon}\right) = \text{Var}(\underline{\varepsilon}) = \mathbf{I}\sigma^2$
- (Insesgamiento de $\underline{\hat{\beta}}$):

$$\begin{aligned}\mathbb{E}\left(\underline{\hat{\beta}}\right) &= \mathbb{E}\left(\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\underline{Y}\right) = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbb{E}(Y) \\ &= \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{X}\underline{\beta} = \underline{\beta}\end{aligned}$$

- (Matriz de Varianzas- Covarianzas para $\underline{\hat{\beta}}$):

$$\begin{aligned}\text{Var}\left(\underline{\hat{\beta}}\right) &= \text{Var}\left(\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\underline{Y}\right) \\ &= \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\text{Var}(\underline{Y})\left(\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\right)^t \\ &= \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{I}\sigma^2\mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1} \\ &= \sigma^2\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\end{aligned}$$

Observe que si $\mathbf{C} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}$ entonces $Cov\left(\hat{\beta}_i, \hat{\beta}_j\right) = \sigma^2 c_{(i+1),(j+1)}$ donde $c_{(i+1),(j+1)}$ es el elemento $(i+1),(j+1)$ de la matriz $C$, con $i, j \in \{0, 1, \ldots, k\}$

# Teorema de Gauss-Markov

En el modelo lineal $\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$ bajo las siguientes hipótesis:

- $\varepsilon_i$ es variable aleatoria con distribución $F$ para toda $i$
- $\mathbb{E}(\varepsilon_i) = 0$
- La matriz $\mathbf{X}$ es de rango completo
- $\text{Var}(\varepsilon_i) = \sigma^2$ (homocedasticidad)
- $Cov\left(\varepsilon_i, \varepsilon_j\right) = 0 \quad i \neq j$

Entonces, los estimadores de mínimos cuadrados cumplen con lo siguiente:

- Son combinaciones lineales de las observaciones $y_i$
- Son insesgados
- **Son los mejores estimadores lineales (BLUES) - (MELI's)**. Es decir, si damos cualquier otro estimador lineal $\underline{\tilde{\beta}}$ entonces
  $\text{Var}\left(\hat{\beta}_i\right) \leq \text{Var}\left(\tilde{\beta}_i\right)$ con $i \in \{0, 1, \ldots, k\}$

# Teorema de Gauss-Markov :: Demostración

Sea $\underline{\hat{\beta}}$ nuestro estimador por mínimos cuadrados y $\underline{\tilde{\beta}}$ cualquier otro estimador lineal insesgado para $\underline{\beta}$

$$\underline{\hat{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\underline{Y} \qquad \underline{\tilde{\beta}} = \mathbf{K}\underline{Y}$$

Suponemos que $\underline{\hat{\beta}} \neq \underline{\tilde{\beta}}$ por lo tanto $\mathbf{K} \neq \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t$ de donde sabemos existe una matriz $\mathbf{D} \neq \mathbf{0}$ tal que $\mathbf{K} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t + \mathbf{D}$ por lo tanto

$$\underline{\tilde{\beta}} = \left(\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t + \mathbf{D}\right)\underline{Y} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\underline{Y} + \mathbf{D}\underline{Y} = \underline{\hat{\beta}} + D\underline{Y}$$

Pero por hipótesis $\mathbb{E}\left(\underline{\hat{\beta}}\right) = \underline{\beta} = \mathbb{E}\left(\underline{\tilde{\beta}}\right)$, por lo tanto tomando esperanza en la última igualdad obtenemos que:

$$\mathbb{E}\left(\underline{\tilde{\beta}}\right) = \mathbb{E}\left(\underline{\hat{\beta}}\right) + \mathbf{D}\mathbb{E}(\underline{Y}) = \underline{\beta} + \mathbf{DX}\underline{\beta} = \underline{\beta} \Rightarrow \mathbf{DX}\beta = \underline{0}$$

Como esta igualdad es válida para cualquier valor del vector $\underline{\beta}$, entonces $\mathbf{DX} = \mathbf{0}$.

# Teorema de Gauss-Markov :: Demostración

Ahora calculemos $\text{Var}\left(\underline{\tilde{\beta}}\right)$

$$\begin{aligned}
\text{Var}\left(\underline{\tilde{\beta}}\right) &= \text{Var}\left(\left(\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t + \mathbf{D}\right)\underline{Y}\right) \\
&= \left(\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t + \mathbf{D}\right)\sigma^2\mathbf{I}\left(\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t + \mathbf{D}\right)^t \\
&= \sigma^2\left(\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t + \mathbf{D}\right)\left(\mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1} + \mathbf{D}^t\right) \\
&= \sigma^2\left(\left(\mathbf{X}^t\mathbf{X}\right)^{-1} + \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{D}^t + \mathbf{D}\mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1} + \mathbf{D}\mathbf{D}^t\right)
\end{aligned}$$

Pero hemos probado que $\mathbf{DX} = \mathbf{0}$ y por tanto $\mathbf{X}^t\mathbf{D}^t = \mathbf{0}$, entonces

$$\text{Var}\left(\underline{\tilde{\beta}}\right) = \sigma^2\left(\left(\mathbf{X}^t\mathbf{X}\right)^{-1} + \mathbf{D}\mathbf{D}^t\right) = \text{Var}\left(\underline{\hat{\beta}}\right) + \sigma^2\mathbf{D}\mathbf{D}^t$$

Finalmente definamos a la matriz $\tilde{\mathbf{D}} = \mathbf{D}\mathbf{D}^t$ y como los elementos de la diagonal de la matriz $\mathbf{D}\mathbf{D}^t$ son siempre positivos, entonces de la última igualdad de matrices tendrámos que:

$$\text{Var}\left(\tilde{\beta}_i\right) = \text{Var}\left(\hat{\beta}_i\right) + \sigma^2\tilde{d}_{ii}$$

Donde $\tilde{d}_{ii}$ es el elemento i-ésimo de la diagonal de la matriz $\tilde{\mathbf{D}}$ el cual sabemos es positivo. Por lo tanto $\text{Var}\left(\tilde{\beta}_i\right) \geq \text{Var}\left(\hat{\beta}_i\right)$ con $i \in \{0, 1, \ldots, k\}$

# Estimación por máxima verosimilitud

Hasta ahora las hipótesis que hemos utilizado en el modelo lineal son:

- La matriz $\mathbf{X}$ es de rango completo $Rank(\mathbf{X}) = p$
- $\varepsilon_i$ es variable aleatoria con distribución $F$ para toda $i$
- $\mathbb{E}(\varepsilon_i) = 0$
- $\text{Var}(\varepsilon_i) = \sigma^2$ (homocedasticidad)
- $Cov\left(\varepsilon_i, \varepsilon_j\right) = 0 \quad i \neq j$

Ahora impondremos una condición fundamental para el desarrollo de toda la teoría inferencial.

- $\underline{\varepsilon}$ sigue una distribución Normal Multivariante.

Consecuencias:

- $\varepsilon_i \sim N\left(0, \sigma^2\right)$
- $\varepsilon_i$ es independiente de $\varepsilon_j$
- $\underline{\varepsilon} \sim N_n\left(\underline{0}, \sigma^2 \mathbf{I}_n\right)$
- Como $\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$ entonces $\underline{Y} \sim N_n\left(\mathbf{X}\underline{\beta}, \sigma^2 \mathbf{I}_n\right)$

# Estimación por máxima verosimilitud

Dado que ahora conocemos la forma de distribución de $\underline{Y}$, podemos encontrar la función de verosimilitud

$$\underline{Y} \sim N_n\left(\mathbf{X}\underline{\beta}, \sigma^2\mathbf{I}_n\right) \Rightarrow f\left(\underline{Y}; \underline{\beta}, \sigma^2\right) = \frac{1}{(2\pi)^{\frac{n}{2}}\left|\sigma^2\mathbf{I}_n\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\underline{Y}-\mathbf{X}\underline{\beta}\right)^t\left(\sigma^2\mathbf{I}_n\right)^{-1}\left(\underline{Y}-\mathbf{X}\underline{\beta}\right)}$$

Como $\left|\sigma^2\mathbf{I}_n\right| = \sigma^{2n}$ y $\left(\sigma^2\mathbf{I}_n\right)^{-1} = \frac{1}{\sigma^2}\mathbf{I}_n$ entonces la verosimilitud es:

$$\mathscr{L}\left(\underline{\beta}, \sigma^2; \underline{Y}\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\left(\underline{Y}-\mathbf{X}\underline{\beta}\right)^t\left(\underline{Y}-\mathbf{X}\underline{\beta}\right)}$$

Sacamos logaritmo de la verosimilitud:

$$\log\mathscr{L}\left(\underline{\beta}, \sigma^2; \underline{Y}\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left(\underline{Y} - \mathbf{X}\underline{\beta}\right)^t\left(\underline{Y} - \mathbf{X}\underline{\beta}\right)$$

Maximizamos con respecto a $\left(\underline{\beta}, \sigma^2\right)$ para ello derivamos e igualamos a cero

$$\frac{\partial}{\partial\underline{\beta}}\log\mathscr{L}\left(\underline{\beta}, \sigma^2; \underline{Y}\right) = -\frac{1}{2\sigma^2}\left(-2\mathbf{X}^t\underline{Y} + 2\mathbf{X}^t\mathbf{X}\underline{\beta}\right) \quad (13)$$

$$\frac{\partial}{\partial\sigma^2}\log\mathscr{L}\left(\underline{\beta}, \sigma^2; \underline{Y}\right) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\underline{Y} - \mathbf{X}\underline{\beta}\right)^t\left(\underline{Y} - \mathbf{X}\underline{\beta}\right) \quad (14)$$

# Estimación por máxima verosimilitud

De la ecuación (13) obtenemos las ecuaciones normales, es decir:

$$\mathbf{X}^t\mathbf{X}\underline{\beta} = \mathbf{X}^t\underline{Y} \Rightarrow \underline{\beta} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\underline{Y} \tag{15}$$

Por lo tanto el estimador máximo verosímil para $\underline{\beta}$ es $\underline{\hat{\beta}}_{MV} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\underline{Y}$, el mismo que habíamos obtenido por mínimos cuadrados. La ventaja de la estimación por máxima verosimilitud es que nos permite encontrar un estimador para el parámetro desconocido $\sigma^2$, de la ecuación (14) obtenemos:

$$-n + \frac{1}{\sigma^2}\left(\underline{Y} - \mathbf{X}\underline{\beta}\right)^t\left(\underline{Y} - \mathbf{X}\underline{\beta}\right) = 0 \Rightarrow \sigma^2 = \frac{\left(\underline{Y} - \mathbf{X}\underline{\beta}\right)^t\left(\underline{Y} - \mathbf{X}\underline{\beta}\right)}{n}$$

Por lo tanto al sustituir en la ultima igualdad lo que obtuvimos en la ecuación (15) obtenemos que estimador máximo verosímil para $\sigma^2$ es:

$$\hat{\sigma}^2_{MV} = \frac{\left(\underline{Y} - \mathbf{X}\underline{\hat{\beta}}_{MV}\right)^t\left(\underline{Y} - \mathbf{X}\underline{\hat{\beta}}_{MV}\right)}{n}$$

Definiendo $\underline{\hat{Y}} := \mathbf{X}\underline{\hat{\beta}}_{MV}$ entonces:

$$\hat{\sigma}^2_{MV} = \frac{\left(\underline{Y} - \underline{\hat{Y}}\right)^t\left(\underline{Y} - \underline{\hat{Y}}\right)}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

# Propiedades de los estimadores máximo verosimiles

Dado que el estimador máximo verosímil para $\underline{\beta}$ coincide con el estimador por mínimos cuadrados entonces:

$$\mathbb{E}\left(\underline{\hat{\beta}}_{MV}\right) = \underline{\beta} \qquad \text{Var}\left(\underline{\hat{\beta}}_{MV}\right) = \sigma^2 \left(\mathbf{X}^t\mathbf{X}\right)^{-1}$$

Luego como $\underline{Y} \sim N_n\left(\mathbf{X}\underline{\beta}, \sigma^2\mathbf{I}_n\right)$ y como $\underline{\hat{\beta}}_{MV} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\underline{Y}$ se sigue que:

$$\underline{\hat{\beta}}_{MV} \sim N_p\left(\underline{\beta}, \sigma^2\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\right)$$

# Propiedades de los estimadores máximo verosimiles

Por otro lado para determinar comportamiento distribucional del estimador $\hat{\sigma}^2_{MV}$ requerimos el siguiente resultado:

$$\frac{\left(\underline{Y}-\underline{\hat{Y}}\right)^t\left(\underline{Y}-\underline{\hat{Y}}\right)}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2\sim\chi^2_{(n-p)}$$

Para demostrar este resultado requeriremos un teorema clásico de modelos lineales:

**Teorema (Distribución de Formas Cuadráticas)**

*Sea* $\underline{W}\sim N_n\left(\underline{\mu},\boldsymbol{I}\right)$, *sean* $\boldsymbol{A}$, $\boldsymbol{B}\in\mathbb{R}^{n\times n}$ *matrices idempotentes y* $\boldsymbol{D}\in\mathbb{R}^{k\times n}$ *suponga que* $\boldsymbol{A}$ *tiene rango* $\nu$ *y* $\boldsymbol{D}$ *de rango completo, entonces:*

- $\underline{W}^T A\underline{W}\sim\chi^{'2}_{(\lambda,\nu)}$; $\qquad\lambda=\mu^T\boldsymbol{A}\mu$
- $\underline{W}^T\boldsymbol{A}\underline{W}$ *es independiente de* $\underline{W}^T\boldsymbol{B}\underline{W}$ *si y solo si* $\boldsymbol{AB}=\boldsymbol{0}$
- $\underline{W}^T\boldsymbol{A}\underline{W}$ *es independiente de* $\boldsymbol{D}\underline{W}$ *si y solo si* $\boldsymbol{DA}=\boldsymbol{0}$

*Donde* $\chi^{'2}_{(\lambda,\nu)}$ *se le conoce como la distribución* $\chi^2$ *- no central, con parámetro de no centralidad* $\lambda$ *y grados de libertad* $\nu$. *Cuando* $\lambda=0$ *se obtiene la* $\chi^2$ *tradicional.*

# Propiedades de los estimadores máximo verosimiles

Apliquemos el teorema anterior a nuestro modelo lineal.

$$\underline{Y} \sim N_n\left(\mathbf{X}\underline{\beta}, \sigma^2\mathbf{I}_n\right) \Rightarrow \underline{W} = \frac{1}{\sigma}\underline{Y} \sim N_n\left(\frac{1}{\sigma}\mathbf{X}\underline{\beta}, \mathbf{I}_n\right)$$

Por otro lado:

$$\frac{\left(\underline{Y} - \underline{\hat{Y}}\right)^t\left(\underline{Y} - \underline{\hat{Y}}\right)}{\sigma^2} = \frac{\underline{Y}^T\left(\mathbf{I} - \mathbf{H}\right)\underline{Y}}{\sigma^2} = \left(\frac{\underline{Y}^T}{\sigma}\right)\left(\mathbf{I} - \mathbf{H}\right)\left(\frac{\underline{Y}}{\sigma}\right)$$

Por lo tanto, usando el teorema anterior:

$$\left(\frac{\underline{Y}^T}{\sigma}\right)\left(\mathbf{I} - \mathbf{H}\right)\left(\frac{\underline{Y}}{\sigma}\right) \sim \chi'^2_{(\lambda,\nu)}$$

Donde $\lambda = \frac{1}{2\sigma^2}\left(\mathbf{X}\underline{\beta}\right)^T\left(\mathbf{I} - \mathbf{H}\right)\left(\mathbf{X}\underline{\beta}\right) = 0$ y $\nu = Rank\left(\mathbf{I} - \mathbf{H}\right) = n - p$. Por lo tanto concluimos que:

$$\frac{\left(\underline{Y} - \underline{\hat{Y}}\right)^t\left(\underline{Y} - \underline{\hat{Y}}\right)}{\sigma^2} \sim \chi^2_{(n-p)}$$

# Propiedades de los estimadores máximo verosimiles

Por este resultado tenemos que:

$$\begin{aligned}\mathbb{E}(\hat{\sigma}^2_{MV}) &= \mathbb{E}\left(\frac{\left(\underline{Y}-\hat{\underline{Y}}\right)^t\left(\underline{Y}-\hat{\underline{Y}}\right)}{n}\right) = \frac{\sigma^2}{n}\mathbb{E}\left(\frac{\left(\underline{Y}-\hat{\underline{Y}}\right)^t\left(\underline{Y}-\hat{\underline{Y}}\right)}{\sigma^2}\right)\\ &= \sigma^2\frac{n-p}{n}\end{aligned}$$

Luego entonces hemos probado que el estimador máximo verosímil es sesgado. Sin embargo podemos construir un estimador insesgado para ello definimos:

$$\hat{\sigma}^2 = \frac{n}{n-p}\hat{\sigma}^2_{MV} = \frac{\left(\underline{Y}-\hat{\underline{Y}}\right)^t\left(\underline{Y}-\hat{\underline{Y}}\right)}{n-p} = \frac{\sum_{i=1}^n (y_i-\hat{y}_i)^2}{n-p}$$

# Análisis inferencial para $\sigma^2$

Tomando en cuenta nuevamente que:

$$\frac{\left(\underline{Y}-\underline{\hat{Y}}\right)^t\left(\underline{Y}-\underline{\hat{Y}}\right)}{\sigma^2}=\frac{\sum_{i=1}^n\left(y_i-\hat{y}_i\right)^2}{\sigma^2}\sim\chi^2_{n-p}$$

Se prueba que:

| Prueba | Región de Rechazo |
|---|---|
| $H_0:\sigma^2\leq s$ vs $H_1:\sigma^2>s$ | $\frac{\sum_{i=1}^n(y_i-\hat{y}_i)^2}{s}>\chi^{2\ (1-\alpha)}_{n-p}$ |
| $H_0:\sigma^2\geq s$ vs $H_1:\sigma^2<s$ | $\frac{\sum_{i=1}^n(y_i-\hat{y}_i)^2}{s}<\chi^{2\ (\alpha)}_{n-p}$ |
| $H_0:\sigma^2=s$ vs $H_1:\sigma^2\neq s$ | $\frac{\sum_{i=1}^n(y_i-\hat{y}_i)^2}{s}>\chi^{2\ (1-\alpha/2)}_{n-p}$ o $\frac{\sum_{i=1}^n(y_i-\hat{y}_i)^2}{s}<\chi^{2\ (\alpha/2)}_{n-p}$ |

Intervalo de confianza para $\sigma^2$ es:

$$\left(\frac{1}{\chi^{2\ (1-\alpha/2)}_{n-p}}\sum_{i=1}^n\left(y_i-\hat{y}_i\right)^2,\frac{1}{\chi^{2\ (\alpha/2)}_{n-p}}\sum_{i=1}^n\left(y_i-\hat{y}_i\right)^2\right)$$

# Análisis inferencial para $\underline{\beta}$::Pruebas para un coeficiente

El objetivo es contrastar hipótesis de la forma:

$$H_0 : \beta_i = b_i \quad vs \quad H_1 : \beta_i \neq b_i; \quad H_0 : \beta_i \leq b_i \quad vs \quad H_1 : \beta_i > b_i$$

$$H_0 : \beta_i \geq b_i \quad vs \quad H_1 : \beta_i < b_i \quad i \in \{0, 1, \ldots, k\}$$

Como $\underline{\hat{\beta}}_{MV} \sim N_p\left(\underline{\beta}, \sigma^2 \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\right)$, entonces haciendo $\mathbf{C} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}$ y definiendo $C_{ij}$ al elemento $(i, j)$ de la matriz $\mathbf{C} \in \mathbb{R}^{p\times p}$ se tiene que:

$$\hat{\beta}_i \sim N\left(\beta_i, \sigma^2 C_{(i+1)(i+1)}\right) \Rightarrow \frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2 C_{(i+1)(i+1)}}} \sim N(0, 1) \quad i \in \{0, \ldots, k\}$$

Recordando que:

**Theorem (Distribución $\tau$)**

*Sea* $X \sim N(0, 1); \quad Y \sim \chi^2_n$; $X$ *independiente de* $Y$, *entonces*
$T = \frac{X}{\sqrt{\frac{Y}{n}}} \sim \tau_{(n)}$

# Análisis inferencial para $\beta$::Pruebas para un coeficiente

Como $\frac{(n-p)}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-p}$ y $\frac{\hat{\beta}_i-\beta_i}{\sqrt{\sigma^2 C_{(i+1)(i+1)}}} \sim N(0,1)$ entonces:

$$t = \frac{\frac{\hat{\beta}_i-\beta_i}{\sqrt{\sigma^2 C_{(i+1)(i+1)}}}}{\sqrt{\frac{(n-p)}{(n-p)\sigma^2}\hat{\sigma}^2}} \sim \tau_{(n-p)}$$

$$t = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 C_{(i+1)(i+1)}}} \sim \tau_{n-p} \quad i \in \{0, \ldots, k\}$$

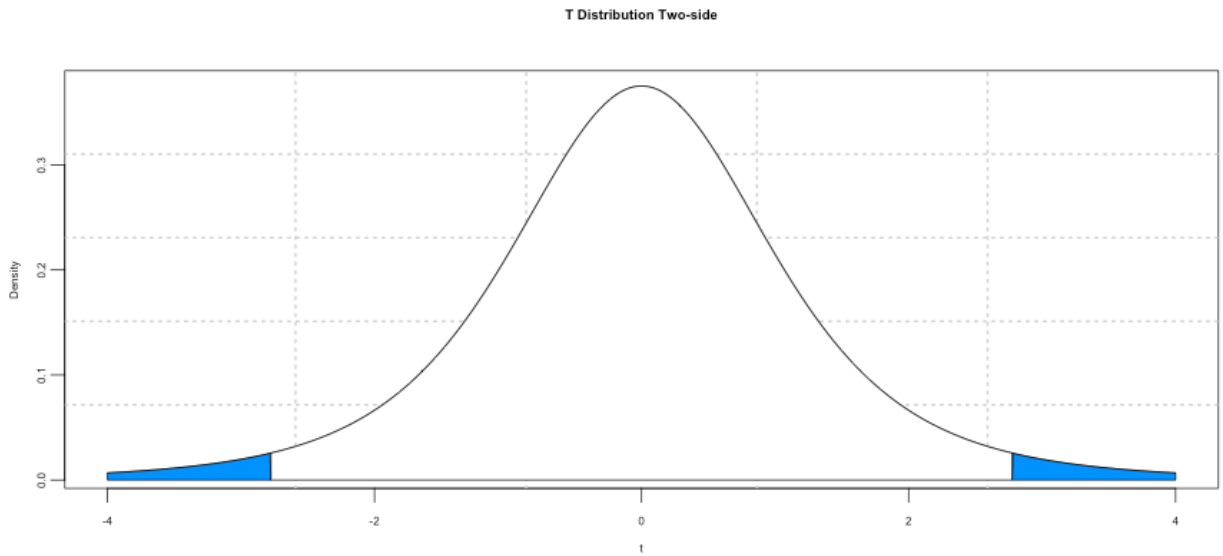
Ojo: Para que esto sea válido hay que probar la independencia entre $\hat{\beta}_i$ y $\hat{\sigma}^2$ (TAREA)

Obs: $t$ así definida *no* es una estadística (Pues depende de $\beta_i$ que es desconocida) sin embargo al fijar $\beta_i$ en una prueba de hipótesis ya puede ser utilizada para construir la región de rechazo.

# Análisis inferencial para $\beta$::Pruebas para un coeficiente

$$H_0 : \beta_i = b_i \quad vs \quad H_1 : \beta_i \neq b_i$$

La regla es rechazar $H_0$ cuando $|t| \geq \tau_{n-p}^{1-\alpha/2}$, donde $t = \frac{\hat{\beta}_i - b_i}{\sqrt{\hat{\sigma}^2 C_{(i+1)(i+1)}}}$ ;



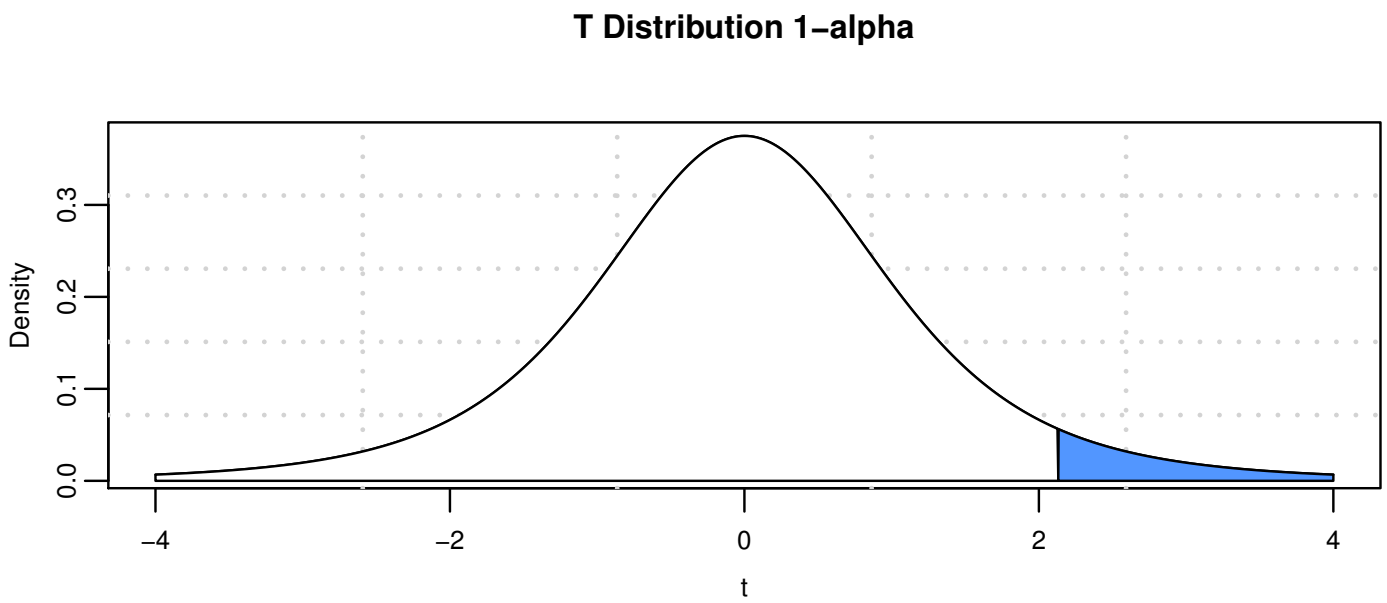

Importante: Esta es la prueba de hipótesis mas importante dentro del análisis de regresión lineal pues cuando hacemos $b_i = 0$ nos ayuda a determinar si la variable asociada a ese coeficiente es *estadísticamente* diferente de cero lo que se traduce en verificar si hay un efecto de la variable $x_i$ en la variable $Y$.

# Análisis inferencial para $\beta$::Pruebas para un coeficiente

$$H_0 : \beta_i \leq b_i \quad vs \quad H_1 : \beta_i > b_i$$

La regla es rechazar $H_0$ cuando $t \geq \tau_{n-p}^{1-\alpha}$, donde

$$t = \frac{\hat{\beta}_i - b_i}{\sqrt{\hat{\sigma}^2 C_{(i+1)(i+1)}}} \quad ;$$

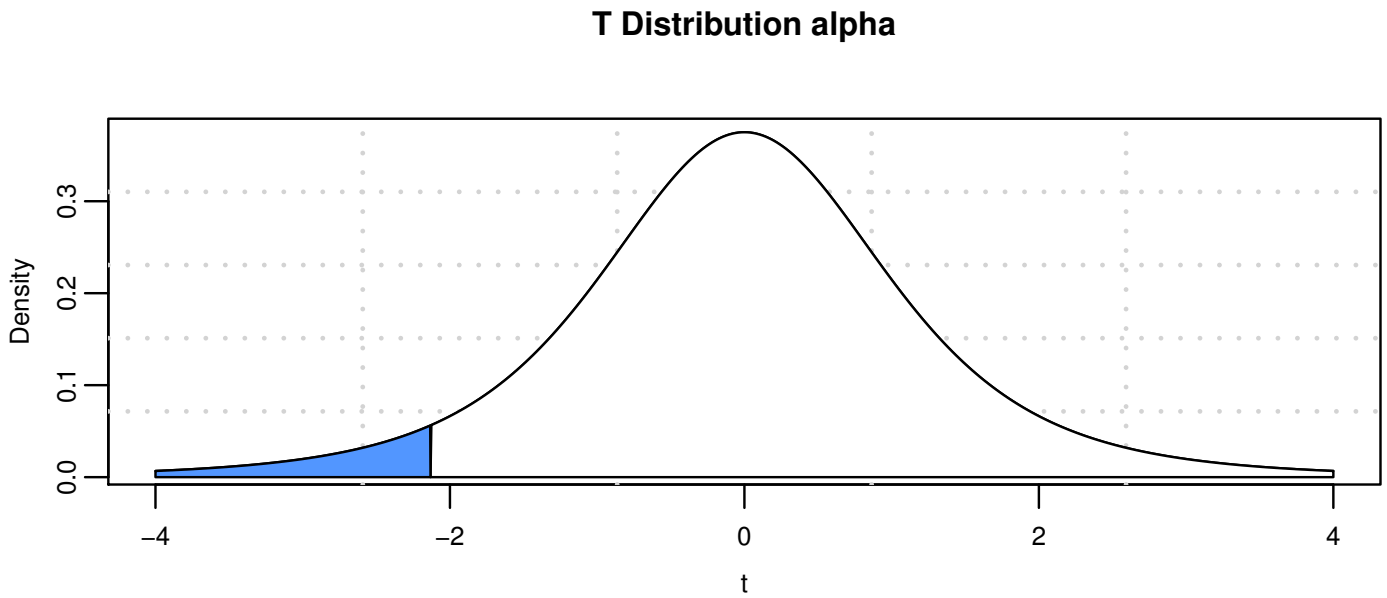


# Análisis inferencial para $\underline{\beta}$::Pruebas para un coeficiente

$$H_0 : \beta_i \geq b_i \quad vs \quad H_1 : \beta_i < b_i$$

La regla es rechazar $H_0$ cuando $t \leq \tau^{\alpha}_{n-p}$, donde

$$t = \frac{\hat{\beta}_i - b_i}{\sqrt{\hat{\sigma}^2 C_{(i+1)(i+1)}}} \quad ;$$



# Análisis inferencial para $\beta$::Intervalo para un coeficiente

La estadística $t = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 C_{(i+1)(i+1)}}}$ es una cantidad pivotal y por tanto puede ser utilizada para construir intervalos de confianza: $i \in \{0, 1, \ldots, k\}$

$$\mathbb{P}\left(-\tau_{n-p}^{1-\alpha/2} \leq \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 C_{(i+1)(i+1)}}} \leq \tau_{n-p}^{1-\alpha/2}\right) = 1 - \alpha$$

$$\mathbb{P}\left(\hat{\beta}_i - \sqrt{\hat{\sigma}^2 C_{(i+1)(i+1)}}\tau_{n-p}^{1-\alpha/2} \leq \beta_i \leq \hat{\beta}_i + \sqrt{\hat{\sigma}^2 C_{(i+1)(i+1)}}\tau_{n-p}^{1-\alpha/2}\right) = 1 - \alpha$$

Por lo tanto un intervalo de confianza al $(1-\alpha)$ % es:

$$\left(\hat{\beta}_i - \sqrt{\hat{\sigma}^2 C_{(i+1)(i+1)}}\tau_{n-p}^{1-\alpha/2}, \hat{\beta}_i + \sqrt{\hat{\sigma}^2 C_{(i+1)(i+1)}}\tau_{n-p}^{1-\alpha/2}\right)$$

Donde $\sqrt{\hat{\sigma}^2 C_{(i+1)(i+1)}}$ se le conoce como error estándar del estimador.

# Análisis inferencial para $\underline{\beta}$

Un problema que tienen la pruebas a cada coeficiente es que da por hecho que los restantes coeficiente están en el modelo, surge entonces la necesidad de desarrollar una teoría para pruebas de hipótesis de un subconjunto de coeficientes. Consideremos el modelo lineal general:

$$\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$$

Queremos determinar si un subconjunto de tamaño $r < p+1$ contribuye en la regresión, para ello particionemos al vector de coeficientes y la matriz $\mathbf{X}$ como sigue:

$$\underline{\beta} = \begin{pmatrix} \underline{\beta_1} \\ \underline{\beta_2} \end{pmatrix}_{p\times 1} \qquad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{pmatrix}_{(n\times K)}$$

Donde $\underline{\beta_1}$ es un vector con $p-r$ entradas y $\underline{\beta_2}$ es un vector con $r$ entradas mientras que la partición de la matriz $\mathbf{X}$ hace que $\mathbf{X}_1$ sea una sub-matriz de $n \times (p-r)$ mientras que $\mathbf{X}_2$ es una sub-matriz de $n \times r$. Con esto el modelo lineal general puede expresarse como:

$$\underline{Y} = \mathbf{X}_1\underline{\beta_1} + \mathbf{X}_2\underline{\beta_2} + \underline{\varepsilon}$$

# Análisis inferencial para $\underline{\beta}$

Nos interesa una prueba de hipótesis de la forma:

$$H_0 : \underline{\beta_2} = \underline{0} \quad vs \quad H_1 : \underline{\beta_2} \neq \underline{0}$$

En términos de modelos estamos haciendo una prueba como sigue:

$$H_0 : \underline{Y} = \mathbf{X}_1\underline{\beta_1} + \underline{\varepsilon} \quad vs \quad H_1 : \underline{Y} = \mathbf{X}_1\underline{\beta_1} + \mathbf{X}_2\underline{\beta_2} + \underline{\varepsilon}$$

Para facilitar el problema plantaremos un caso mas simple pero a la vez muy importante para el modelo lineal. Supongamos que particionamos al vector de parámetros $\underline{\beta}$ como sigue:

$$\underline{\beta_1} = (\beta_0) \qquad \underline{\beta_2} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}_{p\times 1}$$

La prueba de hipótesis asociada bajo esta partición se le conoce como prueba de significancia de la regresión y es calculada por casi todos los paquetes estadísticos con el fin de determinar si existe una relación entre las variables explicativas y la variable respuesta:

$$H_0 : \beta_1 = \beta_2 = \ldots \beta_k = 0 \quad H_1 : \beta_j \neq 0 \text{ para algún } j \in \{1, \ldots, k\}$$

# Análisis de Varianza

Nuestro objetivo se centra en encontrar una estadística de prueba para la significancia de la regresiíon, es decir:

$$H_0 : \beta_1 = \beta_2 = \ldots \beta_k = 0 \quad H_1 : \beta_j \neq 0 \text{ para algún } j \in \{1, \ldots, k\}$$

Recordemos la partición de la suma de cuadrados totales:

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

(Esta igualdad hay que probarla para el modelo lineal multiple: TAREA)
Donde:

- SCT: Suma de cuadrados totales
- SCR: Suma de cuadrados de la Regresión
- SCE: Suma de cuadrados del error

Explicaremos a continuación esta descomposición y construiremos a partir de ésta una estadística para nuestra prueba de hipótesis.

# Análisis de Varianza:: SCT

La suma de cuadrados totales se define como:

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

Tiene la característica de medir la variabilidad total de las observaciones. Para darle una mejor interpretación consideremos lo siguiente, primero recordemos que estamos probando la hipótesis:

$$H_0 : \beta_1 = \beta_2 = \ldots \beta_k = 0 \quad H_1 : \beta_j \neq 0 \text{ para algún } j \in \{1, \ldots, k\}$$

Si $H_0$ fuera cierta entonces el modelo de regresión reducido que estaríamos ajustando es de la forma: $y_i = \beta_0 + \varepsilon_1$,se puede probar que bajo este modelo el mejor estimador para $\beta_0$ es $\bar{y}$, es decir $\hat{y} = \hat{\beta}_0 = \bar{y}$, luego entonces bajo este modelo reducido:

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

Es decir SCT no es mas que la suma de cuadrados del error si hubiéramos ajustado un modelo reducido de la forma $y_i = \beta_0 + \varepsilon_i$, en otras palabras es una medida del error que cometeríamos al ajustar el modelo reducido.

# Análisis de Varianza:: SCT

En muchas ocasiones a la SCT también se le conoce como la suma de cuadrados del modelo reducido bajo $H_0$.
Una característica distribucional importante (**bajo** $H_0$) de esta suma es lo siguiente:

$$\underbrace{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sigma^2}}_{\frac{SCT}{\sigma^2}} \sim \chi^2_{(n-1)}$$

Su demostración se basa en colocar a dicha suma como una forma cuadrática de la forma:

$$SCT := \frac{1}{\sigma^2}\underline{Y}^T(\mathbf{I}-\mathbf{A})\underline{Y}$$

# Análisis de Varianza:: SCE

La suma de cuadrados del error se define como:

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

Recordemos que en este caso estamos bajo el modelo completo es decir:

$$\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}$$

Luego entonces SCE, tiene la característica de medir el error por ajustar el modelo a los datos, es decir, sería el error que no podemos explicar por el modelo de regresión completo y que asumimos se debe enteramente a la variabilidad que se tiene por estimar los parámetros reales mas la variabilidad del error $\varepsilon_i$. Una característica distribucional importante de esta suma es lo siguiente:

$$\underbrace{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sigma^2}}_{\frac{SCE}{\sigma^2}} \sim \chi^2_{(n-p)}$$

Su demostración se basa también en expresar a esta suma como una forma cuadrática.

# Análisis de Varianza:: SCR

La suma de cuadrados de la regresón juega un papel importante en la teoría de los modelos lineales pues su valor es utilizado para validar el modelo, recordemos que

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

Como ya vimos $SCT = SCR + SCE$, luego entonces $SCR = SCT - SCE$. Ya vimos que SCT se interpreta por el error que cometemos por ajustar el modelo reducido bajo $H_0$ mientras que SCE se interpreta por el error que cometemos por ajustar el modelo completo, luego entonces si restamos estas dos cantidades el resultado nos indica cuanto error logramos reducir por irnos de un modelo reducido a aun modelo completo, es decir SCR mide el error que pudimos explicar debido a que incluimos al modelo los coeficiente $\beta_1, \beta_2, \ldots, \beta_p$. Queda claro entonces que entre mas grande sea esta cantidad quiere decir que la regresión bajo el modelo completo está explicando mucha variabilidad y por tanto nos indicara que $H_0$ debería de ser rechazada. Una característica distribucional importante de esta suma es lo siguiente:

$$\underbrace{\frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sigma^2}}_{\frac{SCR}{\sigma^2}} \sim \chi^2_{p-1} = \chi^2_k$$

# Análisis de Varianza

Debe de quedar claro entonces que si el valor de la $SCR$ es grande entonces hay evidencia de que se debería rechazar $H_0$ y por tanto considerar que en efecto existe $\beta_i \neq 0$ para alguna $i \in \{1, 2, \ldots, k\}$. En un primer intento uno quisiera utilizar a la siguiente expresión como estadística de prueba:

$$\underbrace{\frac{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2}{\sigma^2}}_{\frac{SCR}{\sigma^2}} \sim \chi^2_{p-1} = \chi^2_k$$

Sin embargo la anterior expresión en si no es una estadística pues depende del parmetro desconocido $\sigma^2$, para dar solución al problema se lleva a cabo el cociente de $\chi^2$ para dar origen al estadístico $F$. Para ello recordando que $SCE/\sigma^2$ también tiene una distirbución $\chi^2$ entonces construimos:

$$F = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sigma^2(p-1)}}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sigma^2(n-p)}} = \frac{(n-p)\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{(p-1)\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \sim \mathscr{F}_{(p-1,n-p)}$$

Luego entonces rechazamos $H_0$ si el valor de nuestra estadística $F$ es mas grande que el cuantil $\mathscr{F}_{(p-1,n-p)}$

# Análisis de Varianza

La región de rechazo que hemos construido fue de forma intuitiva y sacando provecho de algunas características distribucionales sin embargo se puede probar que esta misma región de rechazo puede ser obtenida mediante el cociente de verosimilitudes generalizado.

$$\frac{\sup \mathscr{L}\left(\Theta_{H_0}\right)}{\sup \mathscr{L}\left(\Theta\right)} < \lambda; \quad \lambda \leq 1$$

Donde:

$$\Theta_{H_0} := \left\{\left(\beta_0, \ldots, \beta_k, \sigma^2\right) : \beta_0 \in \mathbb{R}, \beta_i = 0 \text{ con } i \in (1, 2, \ldots, k), 0 < \sigma^2 < \infty\right\}$$

$$\Theta := \left\{\left(\beta_0, \ldots, \beta_k, \sigma^2\right) : (\beta_0, \beta_1, \ldots, \beta_k) \in \mathbb{R}^p, 0 < \sigma^2 < \infty\right\}$$

Se prueba entonces que se rechaza $H_0$:

$$\frac{\sup \mathscr{L}\left(\Theta_{H_0}\right)}{\sup \mathscr{L}\left(\Theta\right)} < \lambda \Leftrightarrow \frac{SCE}{SCT} \leq K \Leftrightarrow \frac{SCE}{SCE + SCR} \leq K$$

$$\frac{1}{1 + \frac{SCR}{SCE}} \leq K \Leftrightarrow \frac{SCR}{SCE} \geq K^* \Leftrightarrow \frac{(n-p)\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{(p-1)\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \geq K^{**}$$

# Análisis de Varianza

La prueba $F$ que mostramos es sin duda muy importante en el análisis de regresión pues nos indica si realmente existe algún tipo de efecto entre las variables independientes y la variable respuesta, si no rechazamos esta prueba y nos inclinamos por decir que $H_0$ es cierta entonces el modelo de regresión que estamos planteando no es válido. Debido a su importancia, esta prueba esta programada en prácticamente en todos los paquetes estadísticos y se encuentra resumida en la tabla ANOVA.

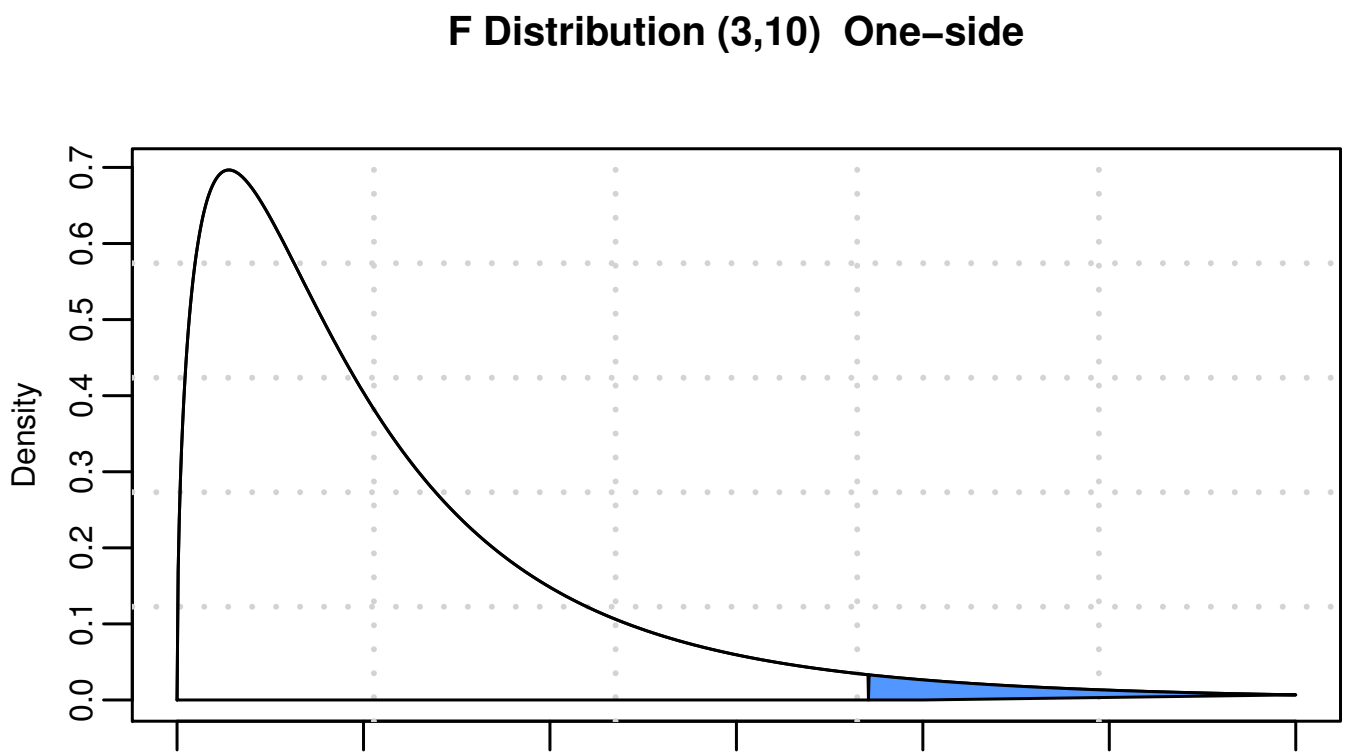
| Var. | S. C. | G. Lib | S.C.M. | F |
|---|---|---|---|---|
| (SCR) | $\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2$ | $p-1$ | $\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{p-1}$ | $\frac{(n-p)\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{(p-1)\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$ |
| (SCE) | $\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$ | $n-p$ | $\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-p}$ | |
| (SCT) | $\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$ | $n-1$ | $\frac{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}{n-1}$ | |

# Análisis de Varianza

En resumen, la tabla ANOVA nos ayuda a contrastar la hipótesis mas importante de la regresión.

$$H_0 : \beta_1 = \beta_2 = \ldots \beta_k = 0 \quad H_1 : \beta_j \neq 0 \text{ para algún } j \in \{1, \ldots, k\}$$

Y se rechaza $H_0$ al nivel de significancia $\alpha$ si el valor del estadístico $F$ toma valores mas grandes que el cuantíl $\mathscr{F}^{1-\alpha}_{(p-1,n-p)}$.



# Coeficiente de Determinación vs $R^2_{Adj}$

El coeficiente de determinación es un número entre 0 y 1 que comúnmente arrojan los paquetes estadísticos y lo denotan como $R^2$ el cual sirve como una medida del ajuste del modelo. Este valor puede ser obtenido de la siguiente manera utilizando la Tabla ANOVA:

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

El coeficiente de determinación tiene el problema de que siempre crece conforme vamos incluyendo mas variables lo cual crea un conflicto al momento de querer seleccionar modelos pues si utilizamos al coeficiente de determinación como un indice para determinar el mejor modelo caeríamos en la regla de ajustar siempre el modelo con mayor número de variables lo cual no pude ser siempre lo mejor pues sobreparametriza el modelo. Para resolver este problema los paquetes también arrojan un coeficiente de determinación ajustado que penaliza el número de variables que tiene el modelo. Se define como:

$$R^2_{Adj} = 1 - \frac{SCE/(n-p)}{SCT/(n-1)} = 1 - \frac{SCME}{SCMT}$$

# ANOVA sin intercepto

Es muy importante notar que la tabla ANOVA que hemos desarrollado supone un modelo con ordenada al origen, es decir $\beta_0$ debe de estar presente en el modelo para que todo funcione correctamente, en algunos casos será necesario llevar a cabo la prueba ANOVA cuando el modelo no tiene la ordenada al origen es decir queremos probar:

$$H_0 : \beta_1 = \beta_2 = \ldots \beta_k = 0 \quad H_1 : \beta_j \neq 0 \text{ para algún } j \in \{1, \ldots, k\}$$

Donde ahora el modelo es de la forma:

$$y_i = \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

En estos casos se muestra (TAREA) que ahora la descomposición de la suma de cuadrados es:

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

# ANOVA sin intercepto

En el modelo:

$$y_i = \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i \qquad \varepsilon_i \sim N\left(0, \sigma^2\right)$$

Cuando se contrasta:

$$H_0 : \beta_1 = \ldots \beta_k = 0 \quad H_1 : \beta_j \neq 0 \text{ para algún } j \in \{0, \ldots, k\}$$

La Tabla ANOVA es la siguiente:

| Var. | S. C. | G. Lib | S.C.M. | F |
|---|---|---|---|---|
| (SCR) | $\sum_{i=1}^n \hat{y}_i^2$ | $k$ | $\frac{\sum_{i=1}^n \hat{y}_i^2}{k}$ | $\frac{(n-k)\sum_{i=1}^n \hat{y}_i^2}{k\sum_{i=1}^n (y_i-\hat{y}_i)^2}$ |
| (SCE) | $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ | $n-k$ | $\frac{\sum_{i=1}^n (y_i-\hat{y}_i)^2}{n-k}$ | |
| (SCT) | $\sum_{i=1}^n y_i^2$ | $n$ | $\frac{\sum_{i=1}^n y_i^2}{n}$ | |

Donde el estadístico $F$ se distribuye bajo $H_0$ como $\mathscr{F}_{(k,n-k)}$ y rechazamos si $F > \mathscr{F}^{1-\alpha}_{(k,n-k)}$

# ANOVA

Imaginemos nuevamente que tenemos el modelo con intercepto de la forma:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i \qquad \varepsilon_i \sim N\left(0, \sigma^2\right)$$

Supongamos que deseamos probar la hipótesis:

$H_0 : \beta_0 = \beta_1 = \ldots = \beta_k = 0 \qquad vs \qquad H_1 : \beta_i \neq 0$ para alguna $j \in \{0, 1, 2, \ldots, k\}$

Nuevamente en este caso la matrix **H** tiene $p$ columnas y por tanto se podrá probar que la SCE del modelo completo cumple con:

$$\frac{1}{\sigma^2} SCE = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \sim \chi^2_{(n-p)}$$

Ahora en este caso la suma de cuadrados del error bajo del modelo reducido ( SCT) dado que no tenemos intercepto hace que nuestro modelo siempre ajuste 0 a cada $y_i$ en ese caso se puede probar que bajo $H_0$ :

$$\frac{1}{\sigma^2} SCT = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - 0)^2 \sim \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i)^2 \sim \chi^2_{(n)}$$

En este caso se prueba (TAREA) que entonces (bajo $H_0$) que:

$$\frac{1}{\sigma^2} SCR = \frac{1}{\sigma^2} (SCT - SCE) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \hat{y}_i^2 \sim \chi^2_{(p)}$$

# ANOVA

En el modelo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i \qquad \varepsilon_i \sim N\left(0, \sigma^2\right)$$

Cuando se contrasta:

$$H_0 : \beta_0 = \beta_1 = \ldots \beta_k = 0 \quad H_1 : \beta_j \neq 0 \text{ para algún } j \in \{0, \ldots, k\}$$

La Tabla ANOVA es la siguiente:

| Var. | S. C. | G. Lib | S.C.M. | F |
|---|---|---|---|---|
| (SCR) | $\sum_{i=1}^{n} \hat{y}_i^2$ | $p$ | $\frac{\sum_{i=1}^{n} \hat{y}_i^2}{p}$ | $\frac{(n-p)\sum_{i=1}^{n} \hat{y}_i^2}{p\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$ |
| (SCE) | $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ | $n-p$ | $\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-p}$ | |
| (SCT) | $\sum_{i=1}^{n} y_i^2$ | $n$ | $\frac{\sum_{i=1}^{n} y_i^2}{n}$ | |

# Análisis inferencial para $\underline{\beta}$

Retomemos nuestro problema inicial que consiste en probar lo siguiente:

$$H_0 : \underline{\beta_2} = \underline{0} \quad vs \quad H_1 : \underline{\beta_2} \neq \underline{0}$$

Queremos entonces una estadística de prueba que nos ayude a determinar si los coeficientes dentro del vector $\underline{\beta_2}$ son significativos. Para dar solución al problema nuevamente recurriremos a la descomposición de la suma de cuadrados. Para el modelo completo $\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$ sabemos que $SCE$ es una medida del error del modelo completo, llamemos $SCE_{H_1}$ a dicha suma. Ahora ajustemos el modelo reducido bajo $H_0$ es decir elaboremos la tabla ANOVA del modelo $\underline{Y} = \mathbf{X}_1\underline{\beta_1} + \underline{\varepsilon}$, con dicha tabla encontramos la $SCE$ asociada y la denominamos $SCE_{H_0}$ la cual es una medida del ajuste del modelo reducido (bajo $H_0$). Definamos ahora $SCE_{H_1|H_0}$ como

$$SCE_{H_1|H_0} = SCE_{H_0} - SCE_{H_1}$$

Intuitivamente $SCE_{H_1|H_0}$ mide el error que logramos explicar por haber incluido los coeficientes de $\underline{\beta_2}$ al modelo y por tanto entre mas grande sea este número hay evidencia de que $H_0$ deberá ser rechazada.

# Análisis inferencial para $\underline{\beta}$

Se puede probar usando formas cuadráticas que:

$$\frac{SCE_{H_1|H_0}}{\sigma^2} = \frac{SCE_{H_0} - SCE_{H_1}}{\sigma^2} \sim \chi^2_r$$

Donde recordemos que $r$ son el número de parámetros que tiene el vector $\underline{\beta_2}$, luego entonces siendo consistentes con lo que hemos estado haciendo, la estadística que surge para probar la hipótesis es:

$$F = \frac{\frac{SCE_{H_0} - SCE_{H_1}}{r\sigma^2}}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sigma^2(n-p)}} = \frac{(SCE_{H_0} - SCE_{H_1})}{r\hat{\sigma}^2} \sim \mathscr{F}_{(r,n-p)}$$

Luego entonces el procedimiento para hacer la prueba es el siguiente:

- Ajustar el modelo completo y calcular ANOVA de donde encontramos $SCE_{H_1}$ y $SCME := \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}$
- Ajustar el modelo reducido bajo $H_0$ y calcular el ANOVA de donde encontramos $SCE_{H_0}$
- Calcular el estadístico F y compararlo con el cuantíl $\mathscr{F}^{1-\alpha}_{r,n-p}$

# Análisis inferencial para $\underline{\beta}$

Si nuestro modelo no tiene intercepto y queremos hacer la prueba de hipótesis

$$H_0 : \underline{\beta_2} = \underline{0} \quad vs \quad H_1 : \underline{\beta_2} \neq \underline{0}$$

Lo que resulta lógico es ahora ajustar el ANOVA sin intercepto que conocemos tanto al modelo completo como al reducido y de ahi extraer las SCR o las SCE respectivamente para obtener el estadístico:

$$F = \frac{\frac{SCE_{H_0} - SCE_{H_1}}{r\sigma^2}}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sigma^2(n-k)}} = \frac{(SCE_{H_0} - SCE_{H_1})}{r\hat{\sigma}^2} \sim \mathscr{F}_{(r,n-k)}$$

# Análisis inferencial para $\underline{\beta}$

El caso mas general de pruebas de hipótesis es el siguiente:

$$H_0 : \mathbf{T}\underline{\beta} = \underline{0} \quad vs \quad H_1 : \mathbf{T}\underline{\beta} \neq \underline{0}$$

Donde $\mathbf{T}$ es una matriz de $(r \times p)$. Este tipo de pruebas surgen cuando queremos verificar si alguna combinación lineal de los paramétros es igual cero, por ejemplo, suponga que tenemos que el modelo lineal:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Y queremos plantear la hipotétesis

$$H_0 : \beta_1 + \beta_3 = 0 \quad vs \quad H_1 : \beta_1 + \beta_3 \neq 0$$

Entonces definimos $\mathbf{T}$ como sigue:

$$\mathbf{T} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \end{pmatrix} \quad \Rightarrow T\underline{\beta} = \beta_1 + \beta_3$$

La forma de atacar este problema es similar, es decir ajustar el modelo reducido y completo, extraer una medida de ajuste de cada modelo y comparar su diferencia con un cuantíl de la distribución $\mathscr{F}$. Queda fuera del alcance del curso para mas información revisar el libro *Introduction to Linear Regression Analysis* en la sección 3.3.4.

# Región de confianza simultaneo para $\underline{\beta}$

Como hemos visto, los intervalos de confianza individuales para los coeficientes tienen el problema que se deben de interpretar de forma independiente, surge entonces la necesidad de crear regiones de confianza que nos de una idea de por donde pueden estar los coeficientes de nuestro modelo. Para resolver este problema requerimos una cantidad pivotal para $\underline{\beta}$. Para ello recordemos lo siguiente:

$$\underline{\hat{\beta}} \sim N_p\left(\beta, \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)$$

Entonces por propiedades de la Distribución Normal Multivariada sabemos que:

$$\frac{1}{\sigma}\left(\mathbf{X}^T\mathbf{X}\right)^{\frac{1}{2}}\left(\underline{\hat{\beta}} - \underline{\beta}\right) \sim N_p\left(\underline{0}, \mathbf{I}\right)$$

Por lo tanto:

$$\left(\frac{1}{\sigma}\left(\mathbf{X}^T\mathbf{X}\right)^{\frac{1}{2}}\left(\underline{\hat{\beta}} - \underline{\beta}\right)\right)^T \frac{1}{\sigma}\left(\mathbf{X}^T\mathbf{X}\right)^{\frac{1}{2}}\left(\underline{\hat{\beta}} - \underline{\beta}\right) \sim \chi^2_p$$

# Región de confianza simultaneo para $\underline{\beta}$

Concluimos entonces que:

$$\frac{1}{\sigma^2}\left(\underline{\hat{\beta}}-\underline{\beta}\right)^T\left(\mathbf{X}^T\mathbf{X}\right)\left(\underline{\hat{\beta}}-\underline{\beta}\right)\sim\chi^2_p$$

Desafortunadamente esto no nos sirve como cantidad pivotal pues depende de $\sigma^2$ que es desconocida, por lo tanto procedemos a hacer el conciente entre otra $\chi^2$, en este caso utilizaremos el hecho que:

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

Luego como $\underline{\hat{\beta}}$ y $\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2\sim\chi^2_{n-p}$ son independientes concluimos que:

$$\frac{(n-p)\left(\underline{\hat{\beta}}-\underline{\beta}\right)^T\left(\mathbf{X}^T\mathbf{X}\right)\left(\underline{\hat{\beta}}-\underline{\beta}\right)}{p\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}\sim\mathscr{F}_{(p,n-p)}$$

La cual ya es una cantidad pivotal para $\underline{\beta}$

# Región de confianza simultaneo para $\underline{\beta}$

Entonces sabemos que si $F^{1-\alpha}_{(p,n-p)}$ es el cuantíl $(1-\alpha)$ de una distribución $\mathscr{F}$ con $(p, n-p)$ grados de libertad entonces:

$$\mathbb{P}\left(\frac{(n-p)\left(\underline{\hat{\beta}}-\underline{\beta}\right)^T\left(\mathbf{X}^T\mathbf{X}\right)\left(\underline{\hat{\beta}}-\underline{\beta}\right)}{p\sum_{i=1}^n\left(y_i-\hat{y}_i\right)^2}\leq F^{1-\alpha}_{p,n-p}\right)=1-\alpha$$

Por lo tanto:

$$\frac{(n-p)\left(\underline{\hat{\beta}}-\underline{\beta}\right)^T\left(\mathbf{X}^T\mathbf{X}\right)\left(\underline{\hat{\beta}}-\underline{\beta}\right)}{p\sum_{i=1}^n\left(y_i-\hat{y}_i\right)^2}\leq F^{1-\alpha}_{p,n-p}$$

Es la región de confianza para $\underline{\beta}$. Debe notarse que como función de $\underline{\beta}$ la ecuación:

$$\frac{(n-p)\left(\underline{\hat{\beta}}-\underline{\beta}\right)^T\left(\mathbf{X}^T\mathbf{X}\right)\left(\underline{\hat{\beta}}-\underline{\beta}\right)}{p\sum_{i=1}^n\left(y_i-\hat{y}_i\right)^2}=F^{1-\alpha}_{(p,n-p)}$$

No es más que la expresión de una elipse en el espacio $\mathbb{R}^p$

# Región de confianza simultaneo para $\underline{\beta}$

Supongamos que ahora queremos construir una región de confianza para un subconjunto de parámetros $\underline{\beta_1} \in \mathbb{R}^q$ de $\underline{\beta}$, $(q < p)$ en ese caso simplemente recordemos que:

$$\underline{\hat{\beta}_1} \sim N_q\left(\beta_1, \sigma^2 \mathbf{C}_{11}\right)$$

Donde $\mathbf{C}_{11}$ es la sub-matrix en $\mathbb{R}^{q\times q}$ de la matriz $\left(\mathbf{X}^t\mathbf{X}\right)^{-1}$ es decir:

$$\left(\mathbf{X}^t\mathbf{X}\right)^{-1} := \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}$$

Luego en este caso, la cantidad pivotal que utilizamos es:

$$\frac{(n-p)\left(\underline{\hat{\beta}_1} - \underline{\beta_1}\right)^T \mathbf{C}_{11}^{-1}\left(\underline{\hat{\beta}_1} - \underline{\beta_1}\right)}{q\sum_{i=1}^n \left(y_i - \hat{y}_i\right)^2} \sim \mathscr{F}_{(q,n-p)}$$

# Intervalo de confianza para la respuesta media

Consideremos una observación de las covariables $(X_1, X_2, \ldots, X_k)$ y construyamos al vector $\underline{x}_0$ a partir de dicha observación:

$$\underline{x}_0 = \begin{pmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{pmatrix}$$

Supongamos que ahora queremos hacer inferencia de la variable respuesta $y$ dado que observamos $\underline{x}_0$, el valor ajustado por nuestro modelo para este punto es el siguiente:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \ldots + \hat{\beta}_k x_{0k} = \underline{x}_0^t \underline{\hat{\beta}}$$

Queremos encontrar un intervalo para $\mathbb{E}(y \mid \underline{x}0)$, queda claro que por ser $\underline{\hat{\beta}}$ insesgado que $\hat{y}_0$ también un estimador insesgado para $\mathbb{E}(y \mid \underline{x}_0)$, en efecto pues:

$$\mathbb{E}(\hat{y}_0) = \mathbb{E}\left(\underline{x}_0^t \underline{\hat{\beta}}\right) = \underline{x}_0^t \mathbb{E}\left(\underline{\hat{\beta}}\right) = \underline{x}_0^t \underline{\beta} = \mathbb{E}(y \mid \underline{x}_0)$$

Por otro lado su varianza es:

$$\mathrm{Var}(\hat{y}_0) = \mathrm{Var}\left(\underline{x}_0^t \underline{\hat{\beta}}\right) = \underline{x}_0^t \mathrm{Var}\left(\underline{\hat{\beta}}\right) \underline{x}_0 = \sigma^2 \underline{x}_0^t \left(\mathbf{X}^t \mathbf{X}\right)^{-1} \underline{x}_0$$

# Intervalo de confianza para la respuesta media

Surge la pregunta de saber como se distribuye $\hat{y}_0$, como $\underline{\hat{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\underline{Y}$ entonces:

$$\hat{y}_0 = \underline{x}_0^t\underline{\hat{\beta}} = \underline{x}_0^t\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\underline{Y} = \mathbf{C}\underline{Y}$$

Como $\underline{Y}$ es un vector normal-multivariado se sigue que $\hat{y}_0$ sigue una distribución normal, luego entonces concluimos que:

$$\hat{y}_0 \sim N\left(\mathbb{E}(\,y \mid \underline{x}_0), \sigma^2\underline{x}_0^t\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\underline{x}_0\right)$$

Por lo tanto:

$$\frac{\hat{y}_0 - \mathbb{E}(\,y \mid \underline{x}_0)}{\sqrt{\sigma^2\underline{x}_0^t\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\underline{x}_0}} \sim N(0,1)$$

Asumiendo que $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sigma^2} \sim \chi^2_{(n-p)}$ entonces:

$$\frac{\hat{y}_0 - \mathbb{E}(\,y \mid \underline{x}_0)}{\sqrt{\hat{\sigma}^2\underline{x}_0^t\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\underline{x}_0}} \sim \tau_{(n-p)}$$

La cual ya es una cantidad pivotal de donde construiremos el intervalo de confianza para la respuesta media $\mathbb{E}(\,y \mid \underline{x}_0)$.

# Intervalo de predicción para la Respuesta Media

$$\mathbb{P}\left(-\tau_{(n-p)}^{1-\frac{\alpha}{2}} \leq \frac{\hat{y}_0 - \mathbb{E}(\,y \mid \underline{x}_0)}{\sqrt{\hat{\sigma}^2 \underline{x}_0^t \left(\mathbf{X}^t\mathbf{X}\right)^{-1} \underline{x}_0}} \leq \tau_{(n-p)}^{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

Despejando obtenemos que el Intervalo al $(1-\alpha)\,\%$ de confianza para la respuesta media es:

$$\left(\hat{y}_0 - \tau_{(n-p)}^{(1-\frac{\alpha}{2})} \sqrt{\hat{\sigma}^2 \underline{x}_0^t \left(\mathbf{X}^t\mathbf{X}\right)^{-1} \underline{x}_0}, \hat{y}_0 + \tau_{(n-p)}^{(1-\frac{\alpha}{2})} \sqrt{\hat{\sigma}^2 \underline{x}_0^t \left(\mathbf{X}^t\mathbf{X}\right)^{-1} \underline{x}_0}\right)$$

# Intervalo de Predicción para nuevas observaciones

Ahora supongamos que queremos construir un intervalo de predicción para una nueva observación $y_0$ (cuando las covariables toman el valor de $\underline{x}_0$) Sabemos que una estimación natural para la variable respuesta $y_0$ es

$$\hat{y}_0 = \underline{x}_0^t \hat{\underline{\beta}}$$

Notemos que como $y_0$ es una nueva observación entonces $y_0$ es independiente de las observaciones anteriores $(y_1, \ldots, y_n)$ y como los estimadores $\hat{\beta}_0, \ldots, \hat{\beta}_k$ son combinaciones lineales de las observaciones $(y_1, \ldots, y_n)$ , se sigue entonces que $y_0$ es independiente de $\hat{y}_0$, por lo tanto $(y_0 - \hat{y}_0)$ sigue una distribución normal. Luego como:

$$\begin{aligned}
\mathbb{E}((y_0 - \hat{y}_0)) &= \mathbb{E}\left(\underline{x}_0^t \underline{\beta} + \varepsilon_0 - \underline{x}_0^t \hat{\underline{\beta}}\right) = 0 \\
\text{Var}((y_0 - \hat{y}_0)) &= \text{Var}(y_0) + \text{Var}(\hat{y}_0) = \sigma^2 + \text{Var}\left(\underline{x}_0^t \hat{\underline{\beta}}\right) \\
&= \sigma^2 + \sigma^2 \underline{x}_0^t \left(\mathbf{X}^t \mathbf{X}\right)^{-1} \underline{x}_0 \\
&= \sigma^2 \left(1 + \underline{x}_0^t \left(\mathbf{X}^t \mathbf{X}\right)^{-1} \underline{x}_0\right)
\end{aligned}$$

# Intervalo de Confianza para nuevas observaciones

Finalmente hemos probado que:

$$(y_0 - \hat{y}_0) \sim N\left(0, \sigma^2\left(1 + \underline{x}_0^t\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\underline{x}_0\right)\right)$$

Asumiendo que $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sigma^2} \sim \chi^2_{(n-p)}$ entonces:

$$\frac{(y_0 - \hat{y}_0)}{\sqrt{\hat{\sigma}^2\left(1 + \underline{x}_0^t\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\underline{x}_0\right)}} \sim t_{(n-p)}$$

De donde obtenemos una cantidad pivotal para construir el siguiente intervalo de confianza para nuevas observaciones:

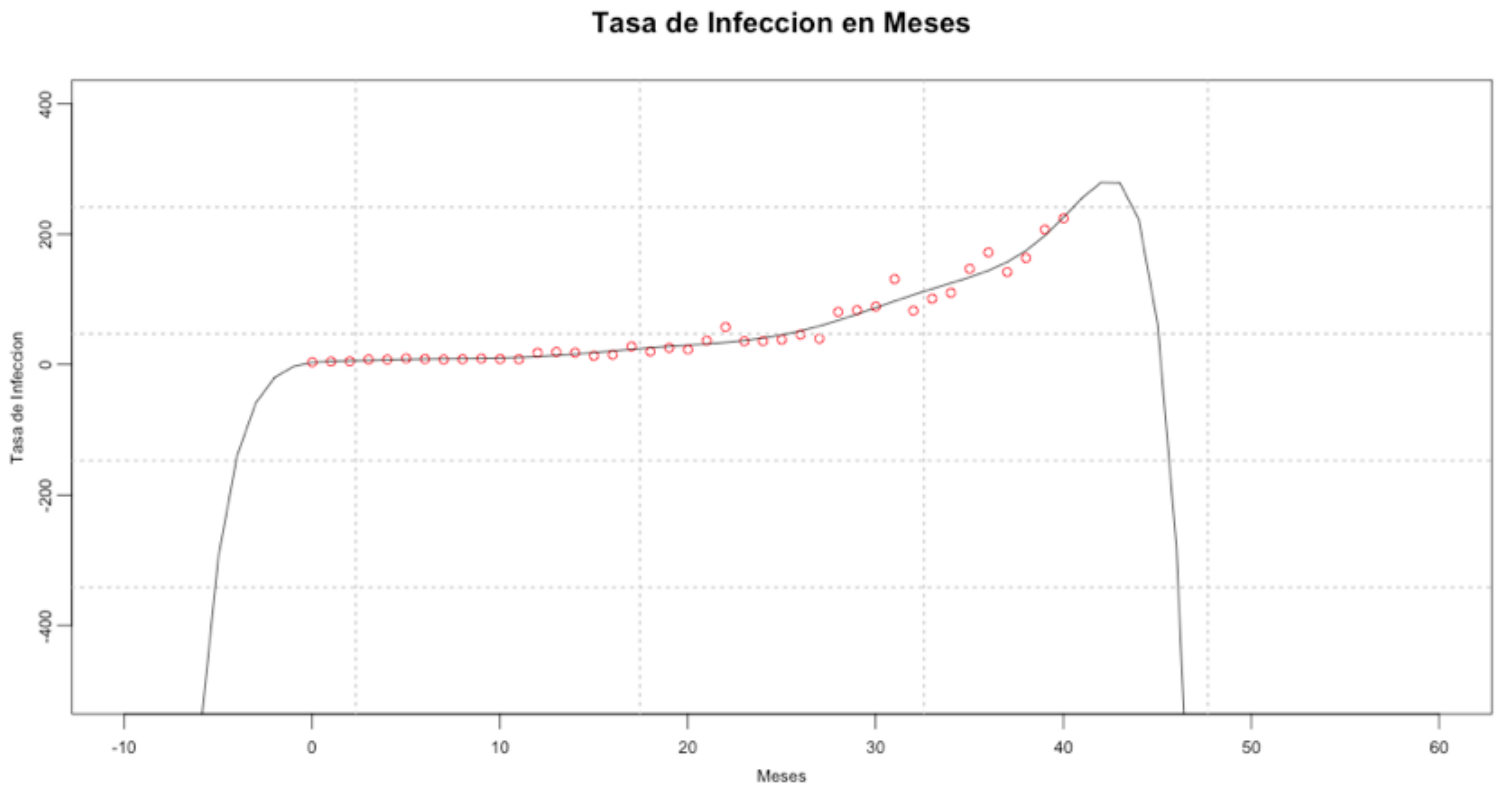
$$\left(\hat{y}_0 - \tau_{(n-p)}^{(1-\frac{\alpha}{2})}\sqrt{\hat{\sigma}^2 + \hat{\sigma}^2\underline{x}_0^t\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\underline{x}_0}, \hat{y}_0 + \tau_{(n-p)}^{(1-\frac{\alpha}{2})}\sqrt{\hat{\sigma}^2 + \hat{\sigma}^2\underline{x}_0^t\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\underline{x}_0}\right)$$

Observaciones de los intervalos de confianza:

- Este intervalo es similar al encontrado anteriormente solo que agrega mas variabilidad debido a toma en cuenta la variabilidad de los estimadores así como de la variabilidad que tiene el modelo en su definición $\varepsilon_0$
- Hay que tener mucho cuidado con la extrapolación.

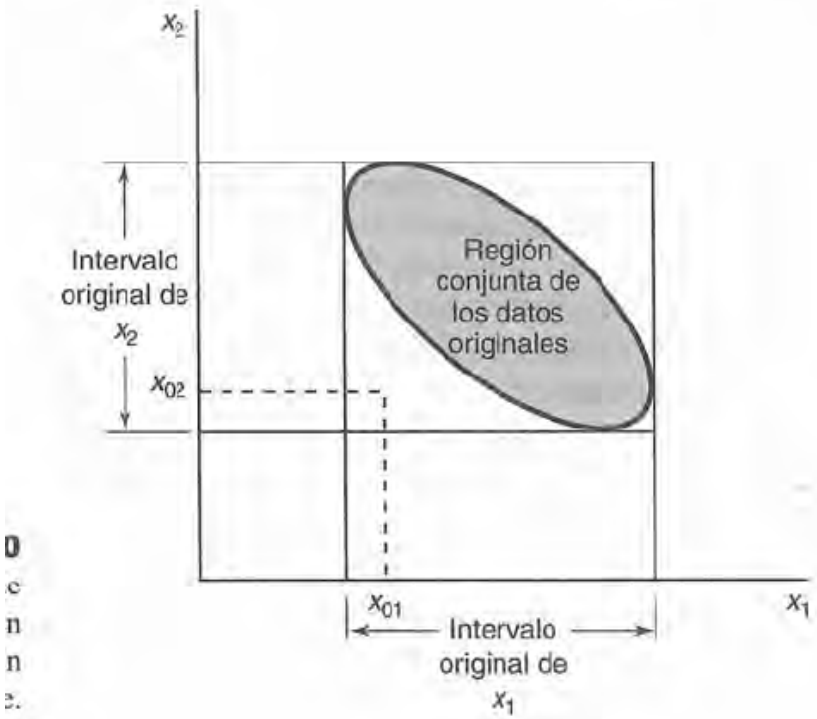
# Extrapolación

# Extrapolación



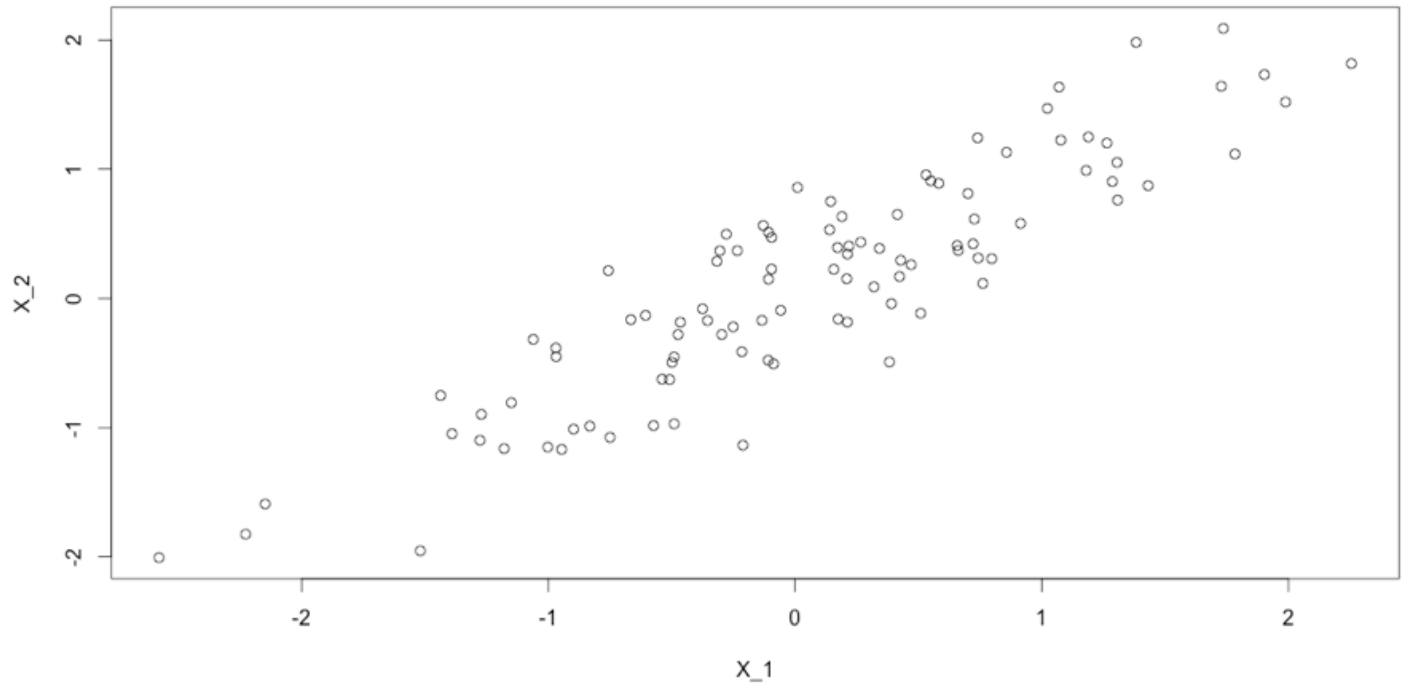
# Extrapolación:: Regresión Multiple

Al hacer inferencia sobre nuevas respuestas en un punto dado $\underline{x}_0^t$, se debe de tener cuidado de la extrapolación pues es posible que los datos ajusten bien pero solo en la regón en donde se tomó la muestra y no fuera de ella. Surge el problema de cómo determinar la región donde el modelo es válido.
Ejemplo, supongamos que tenemos solo dos variables explicativas:



# Extrapolación:: Regresión Multiple

Necesitamos encontrar la region mínima en el espacio de las variables explicativas que contenga a todas las observaciones en muestra.




Encontrar dicha region es díficil por lo que se estila encontrar a la *elipse* de menor área que cubra a todos los puntos

# Extrapolación:: Regresión Multiple

Cook [1979] y Weisberg [1985], consideran esta elipse

$$\underline{x}^t \left(\mathbf{X}^t\mathbf{X}\right)^{-1} \underline{x} \leq h_{max}$$

Donde $h_{max} := \text{máx}\{h_{11}, \ldots h_{nn}\}$ y $h_{ii}$ son los elementos de la diagonal de la matriz Sombrero $\mathbf{H}$.
Ellos demuestran que esta elipse encierra a todos los puntos de la muestra pero no garantizan que sea la miníma, sin embargo es una buena aproximación para solucionar el problema.
Luego entonces, suponga que queremos inferir a $y$ en el punto $\underline{x}_0$, para verificar si nuestra predicción es válida, tenemos que calcular:

$$h_{00} = \underline{x}_0^t \left(\mathbf{X}^t\mathbf{X}\right)^{-1} \underline{x}_0$$

Luego si:

- $h_{00} > h_{max}$ entonces, $x_0$ está fuera de la región de muestra
- $h_{00} \leq h_{max}$ entonces, $x_0$ está cercano a la región de muestra por lo que la estimación de $y$ se considera buena

# Supuestos del Modelo (Segunda Parte)

OJO: En General, no se pueden detectar desviaciones respecto a las premisas básicas examinando los estadísticos estándar de resumen como por ejemplo los estadísticos $t$, $F$ o $R^2$

- La relación entre la respuesta $y$ y los regresores es lineal, (al menos en forma aproximada)
- Independencia (ACF, DurbinWatson, Rachas)
- Varianza Constante (Homocedasticidad, Barttlet, Levene)
- Normalidad (QQPlot, Historgramas )
- Multicolinealidad
- *Outliers* y observaciones influyentes

# Residuales vs $\varepsilon_i$

Recordemos que nuestro modelo general es de la forma:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

Luego entonces:

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_{i1} - \ldots + \beta_k x_{ik}$$

Sin embargo $(\beta_0, \ldots, \beta_k)$ son desconocidos luego entonces $\varepsilon_i$ no son observables al momento de hacer los analísis. Tenemos el problema de que algunos de los supuestos del modelo recaen en suponer que hay Normalidad y Homocedasticidad en las $\varepsilon_i$. Para solucionar esto se suele utilizar a los residuales como una aproximación a la realización de las variables $\varepsilon_i$ y por tanto el análisis recaé sobre el vector de residuales definido como:

$$\underline{e} = \underline{Y} - \underline{\hat{Y}} = \underline{Y} - \mathbf{X}\underline{\hat{\beta}} = (\mathbf{I} - \mathbf{H})\,\underline{Y}$$

Ahora bien, recordando que $\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$ entonces:

$$\underline{e} = (\mathbf{I} - \mathbf{H})\,\underline{Y} = (\mathbf{I} - \mathbf{H})\left(\mathbf{X}\underline{\beta} + \underline{\varepsilon}\right) = (\mathbf{I} - \mathbf{H})\,\underline{\varepsilon}$$

Por lo tanto

$$\mathrm{Var}(\underline{e}) = \mathrm{Var}((\mathbf{I} - \mathbf{H})\,\underline{\varepsilon}) = \sigma^2\,(\mathbf{I} - \mathbf{H})$$

# Residuales vs $\varepsilon_i$

Obs:

$$\underline{e} \not\sim N_n\left(\underline{0}, \sigma^2\left(\mathbf{I}-\mathbf{H}\right)\right)$$

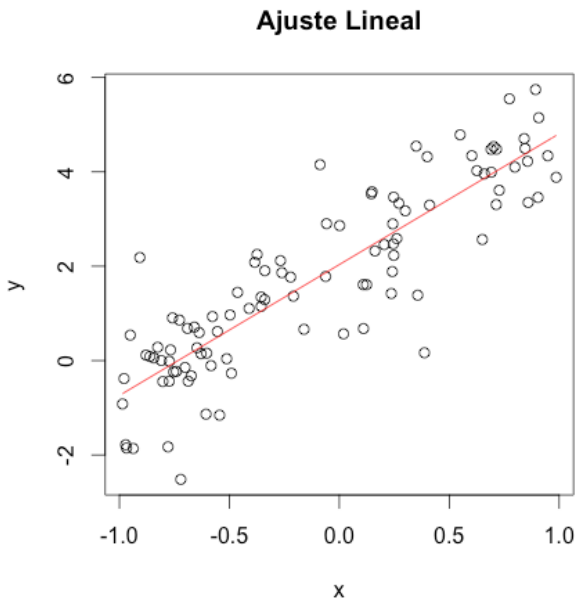
En general la matrix $(\mathbf{I}-\mathbf{H})$ no es diagonal, luego entonces, teóricamente los residuales no tienen correlación 0 y ademas tienen distinta varianza, (pues esta en función de la matriz $\mathbf{H}$). Concluimos entonces que los residuales no son independientes, sin embargo esta **no** independencia de los residuales tiene poco efecto en su aplicación para comprobar la adecuación del modelo, siempre y cuando $n$ no sea pequeña en relación con la cantidad de parámetros $k$

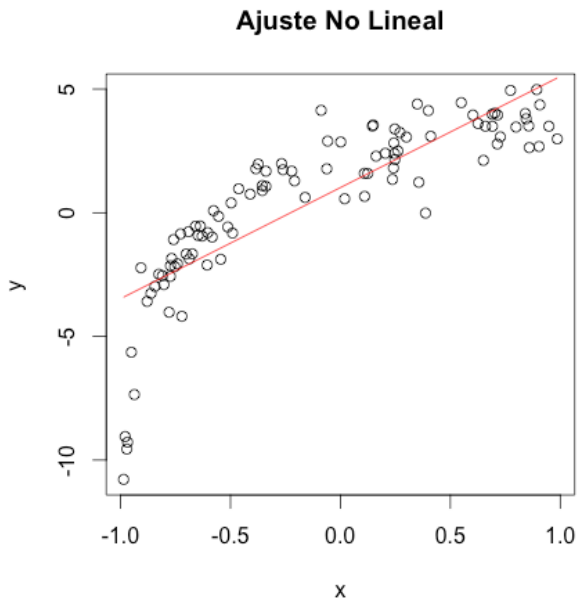
# Verificación del Supuesto de Linealidad

Recordemos que el modelo lineal que ajustamos es de la forma:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

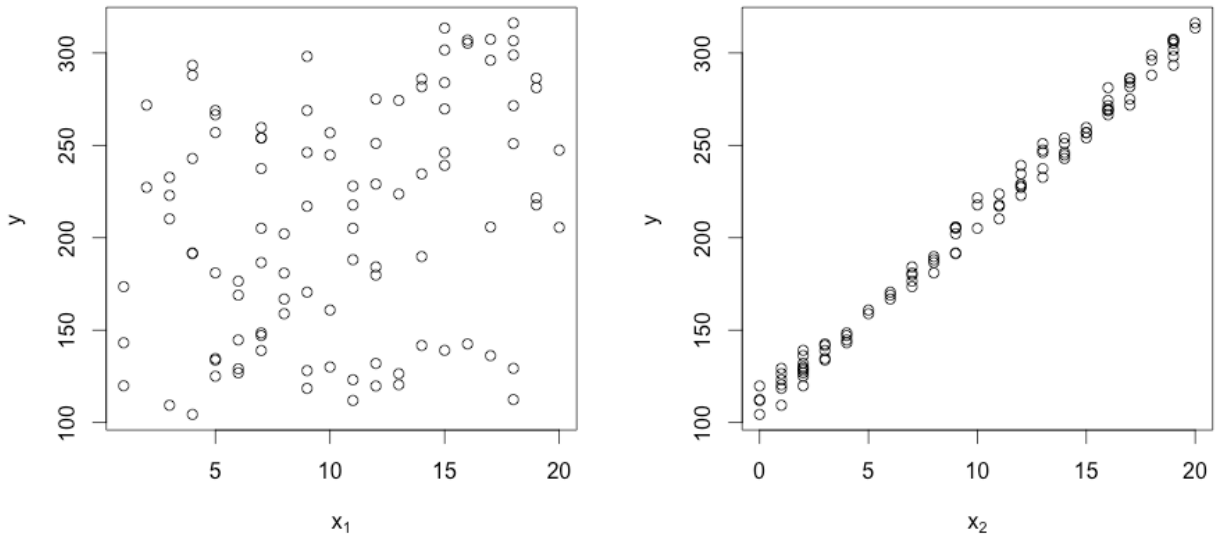
Si $k = 1$ (Modelo lineal simple) es relativamente fácil verificar linealidad, basta con realizar una gráfica de los puntos $(y_i, x_{i1})$ para verificar si en realidad existe una recta asociada a la relación.



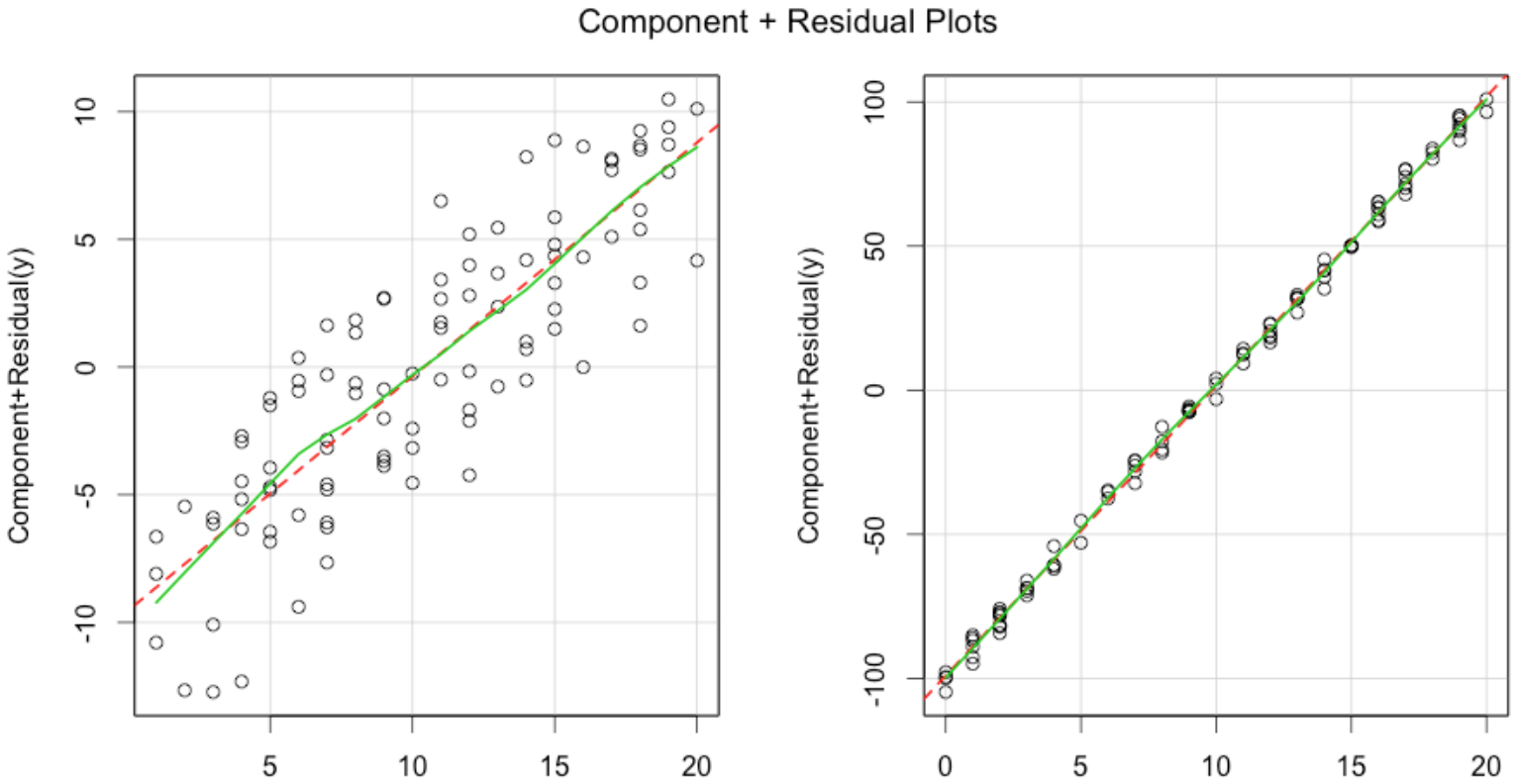
# Verificación del Supuesto de Linealidad

Cuando tenemos 2 o mas variables explicativas la verificación de linealidad se complica debido a que se requiere graficar en dimensiones mayores. Una solución inicial al problema es graficar cada una de las variables explicativas contra la variable respuesta y verificar ahi la linealidad, sin embargo dichos graficos pueden llevarnos a conclusiones erróneas cuando los coeficientes tienen magnitudes distintas. En la siguiente figura se muestran las gráficas de las covariables vs la variable respuesta de un modelo **lineal**




Para solucionar esto recurriremos al análisis de residuales parciales.

Supongamos que queremos ver si existe una relación lineal entre la variables $X_q$ y $Y$, entonces lo que hacemos es ajustar el modelo **eliminando** a la variable $X_q$ del modelo, luego encontramos los residuales de dicho modelo (denotamos por $e_i^{(q)}$) y luego graficamos esos residuales contra los valores de la variables $X_q$, es decir visualizamos los puntos $\left(x_{iq}, e_i^{(q)}\right)$. Si la relación es lineal entre $X_q$ y $Y$ se espera que esta gráfica de residuales parciales presente una relación lineal, en caso contrario, es una indicación de que esa variable posiblemente necesite ser transformada para ayudar al ajuste.



# Verificación del Supuesto de Independencia

En el modelo de regresión lineal simple se plantearon las pruebas para identificar independencia utilizando el ACF (Función de autocorrelación ) así como la prueba no parámetrica de Rachas.
Otra prueba muy utilizada para la identificación de independencia (no correlación) es utilizar la prueba de *Durbin Watson*. Esta prueba pretende ver si los valores presentan algún tipo de dependencia en cuanto al orden de obtención. La prueba se plantea como sigue: Supongamos que los errores $\varepsilon_i$ siguen un modelo AR(1) entonces:

$$\varepsilon_i = \rho\varepsilon_{i-1} + \delta_i \qquad \delta_i \sim N(0, \sigma^2)$$

Luego entonces planteamos:

$$H_0 : \rho = 0 \qquad vs \qquad H_1 : \rho \neq 0$$

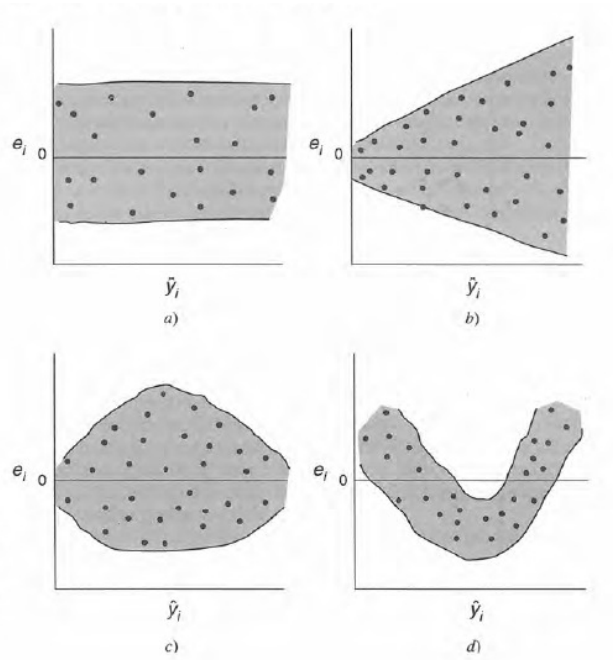
Estadístico de prueba:

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

El comando utilizado en R es *durbinWatsonTest(model)* y esperamos ver un p-value alto para no rechazar. (Ojo: Estas prueban suponen Normalidad)

# Verificación del Supuesto de Homocedasticidad

Generalmente para verificar que tenemos varianza constante se llevan a cabo pruebas visuales verificando el comportamiento de los residuales conforme se van observando, o bien graficar contra $\hat{y}_i$.
Dado que ahora tenemos mas variables explicativas, también se suele graficar los residuales contra cada una de las variables explicativas y verificar le homocedasticidad ahi.




Hay que recordar que existen pruebas formales para verificar la homocedasticidad (Ej. Barttlet, Levene) pero requieren tener definido grupos lo cual puede quitarle credibilidad a la prueba. Ver prueba ncvTest.

# Verificación del Supuesto de Normalidad

- Realizar histrogramas de los residuales esperando ver un comportamiento normal con media en 0
- Realizar la grafica QQ-Plot
- Pruebas de Bondad de Ajuste (Solo como medida del ajuste)

# Multicolinearidad

Dentro de los supuesto del modelo lineal se requirió que la matriz de diseño $\mathbf{X}$ tuviera rango completo, esto con el fin de garantizar la invertibilidad de la matriz $\mathbf{X}^t\mathbf{X}$ y obtener una solución única a las ecuaciones normales $\left(\underline{\hat{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\underline{Y}\right)$. El hecho de pedir rango completo a la matriz $\mathbf{X}$ nos obliga a verificar que no existe columnas en $\mathbf{X}$ que sean combinación lineal de otras. En la practica esto casi no ocurre sin embargo lo que si suele ocurrir es que existan columnas que *casi* son combinaciones lineales de otras columnas. Teóricamente al no haber un igualdad exacta se garantiza la existencia de la matriz inversa. Sin embargo numéricamente esto ocasiona muchos problemas.

$$\mathbf{X} = (\underline{X}_0, \underline{X}_1, \ldots, \underline{X}_k)$$

$$\underline{X}_q \approx \alpha_0\underline{X}_0 + \ldots + \alpha_{q-1}\underline{X}_{q-1} + \alpha_{q+1}\underline{X}_{q+1} + \ldots + \alpha_k\underline{X}_k$$

Verificar la multicolineridad no es fácil a simple vista sin embargo existen formas de detectarla.

# Multicolinearidad

- Una forma muy simple para detectar multicolinearidad es inspeccionar la matriz de correlaciones entre las variables explicativas. ( `cor(X)` , `pairs(X)`)
- Hacer regresiones lineales de $X_q$ con el resto de las variables y obtener el coeficiente de determinación de dicha regresión $R_q^2$, un coeficiente de deteminación cercana a 1 nos dice que hay un ajuste muy bueno por lo que se puede decir que:

$$\underline{X}_q \approx \alpha_0 \underline{X}_0 + \ldots + \alpha_{q-1} \underline{X}_{q-1} + \alpha_{q+1} \underline{X}_{q+1} + \ldots + \alpha_k \underline{X}_k$$

Una forma de obtener indices es calcular lo que se denomina el $VIF_q$ (Factor de Inflación de la varianza) definido como:

$$VIF_q = \frac{1}{1 - R_q^2} \quad q \in 0, 1, \ldots, k$$

Luego entonces si $R_q^2 \approx 1$ eso se traduce en un $VIF_q$ grande, en la práctica se suele tomar como punto de corte 5 ó 10, es decir obtener un $VIF_q > 10$ nos habla de que la variable $X_q$ puede ser expresada aproximadamente como combinación lineal de las restantes y por tanto podemos tener problemas numéricos en las estimaciones.

# Multicolinearidad

- Otra forma de detectar la multicolinearidad es con el índice de condición de una matriz $\left(\mathbf{X}^t\mathbf{X}\right)$ el cual se definie como:

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}} \quad kappa\left(t(X)\,\%*\,\%X\right)$$

Generalmente un indice de condición menor a 100 no tiene problemas de multicolinealidad, si está entre 100 y 1000 implica una moderada multicolinearidad y si excede a 1000 entonces hay problemas importantes y podemos tener problemas númericos

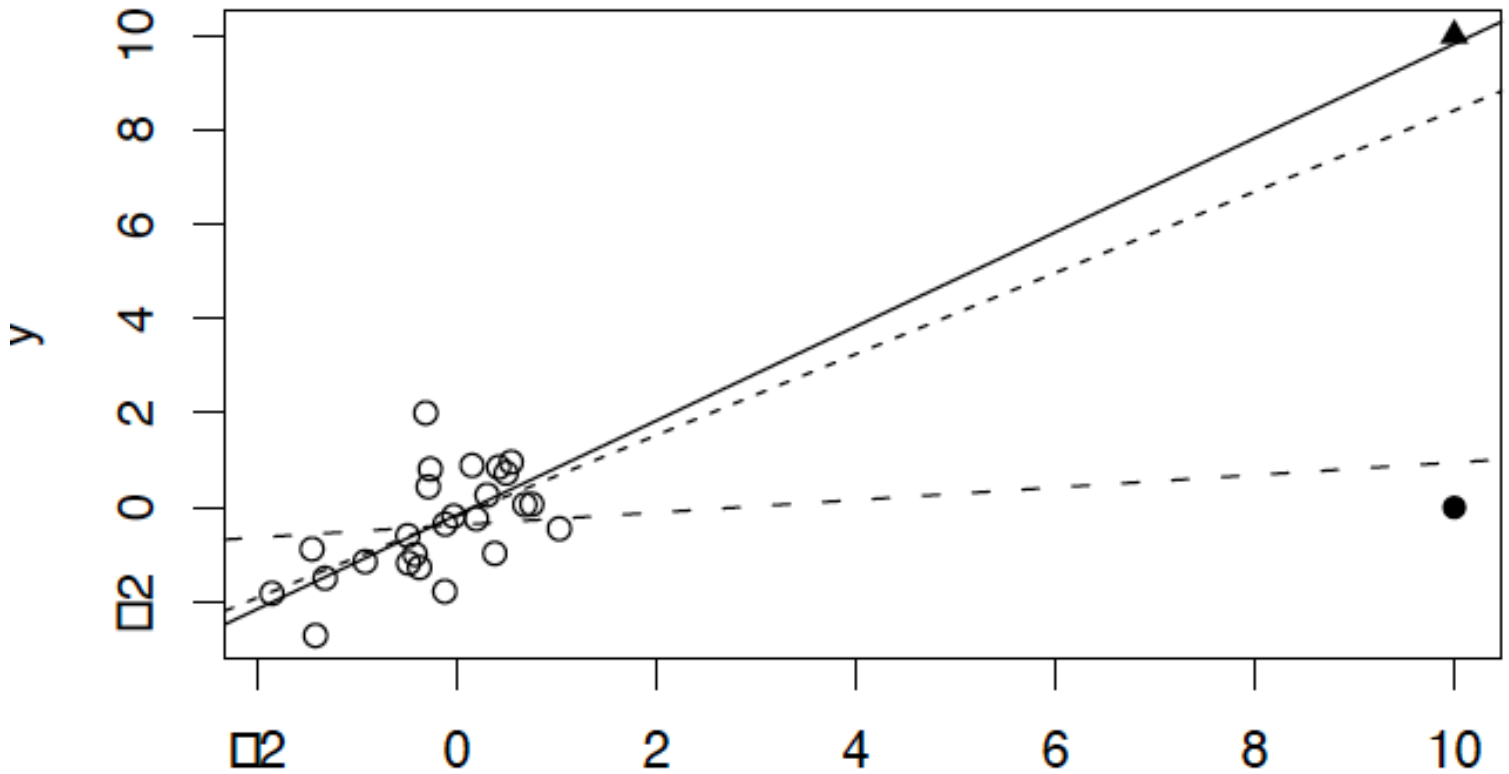
# Outliers y Observaciones influyentes

Hay dos propósitos principales para tratar de identificar las observaciones que no pertenecen al modelo

- Proteger la integridad del modelo de los efectos de los puntos que no pertenecen a éste. (Observaciones que no pertenecen al modelo) Frecuentemente exhiben residuales grandes.
- Identificar las deficiencias del modelo . Posiblemente se necesite incluir nuevas variables o se requiere llevar a cabo transformaciones.

# Outliers y Observaciones influyentes

El primer paso para identificar observaciones atípicas es tener una idea de la región en donde estamos muestreando, la idea es encontrar una métrica que nos ayude a decidir si una observación de las variables independientes está muy alejada de la región de muestreo. En general una observación que se encuentra lejos de donde esta la masa de puntos ocasiona que el modelo cambie de manera significativa (Aunque no necesariamente).



# Outliers y Observaciones influyentes

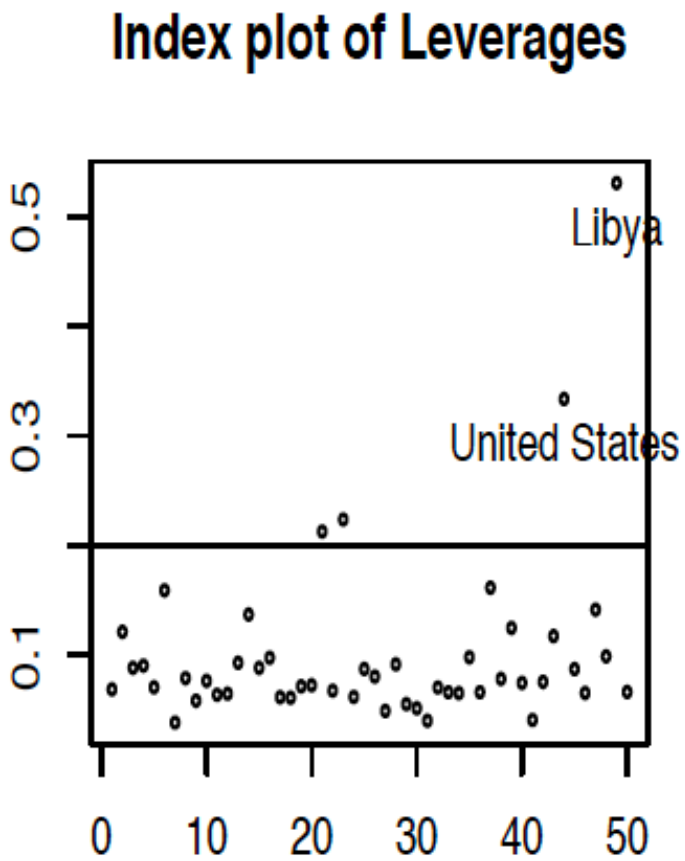
Para detectar los puntos de muestreo que están alejados se utiliza la matriz sombrero, $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t$. A los elementos de la diagonal de $\mathbf{H}$ $(h_{ii})$ se les conoce como el *leverage* o la influencia de cada observación. De este modo un *leverage* grande nos habla de una observación alejada de la masa de los puntos de muestreo.
Como:

$$tr(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = p$$

Por lo tanto, en promedio los *leverages* toman el valor de $\frac{p}{n}$. Existe una regla muy utilizada que dice que aquellos $h_{ii}$ que sean mayores a dos veces al promedio son puntos alejados y por tanto deben de ser analizados. $\left(h_{ii} > 2\frac{p}{n}\right)$

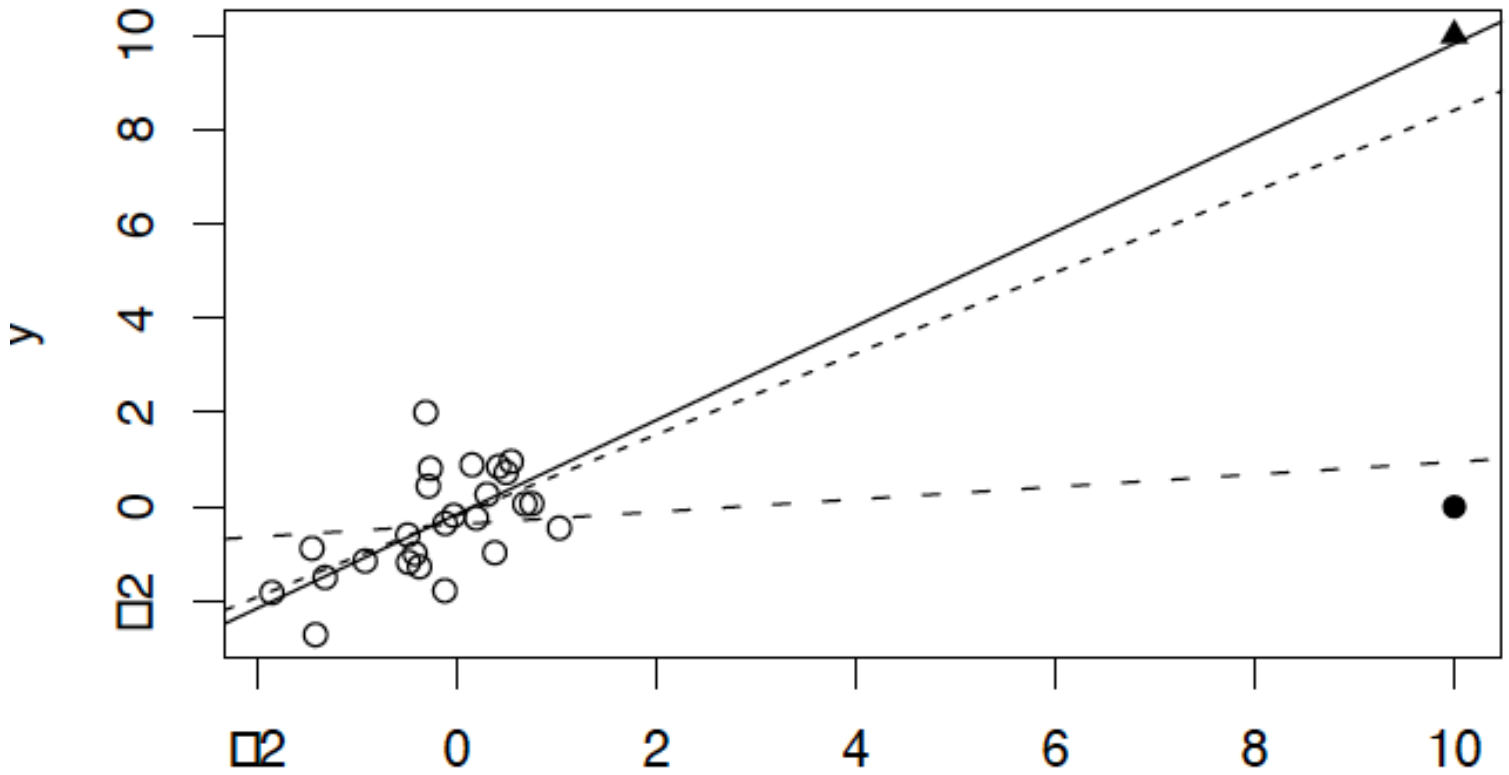
# Outliers y Observaciones influyentes

En la siguiente gráfica se observa como 4 observaciones se sale de la region definida, esas cuatro observaciones deben de ser analizadas y verificar si es correcto que estén en muestra. Si es correcta la observación lo siguiente es verificar si se trata de un outlier para el modelo o bien una observación influyente.



# Outliers y Observaciones influyentes

Outlier: Es un punto que no ajusta en el modelo.



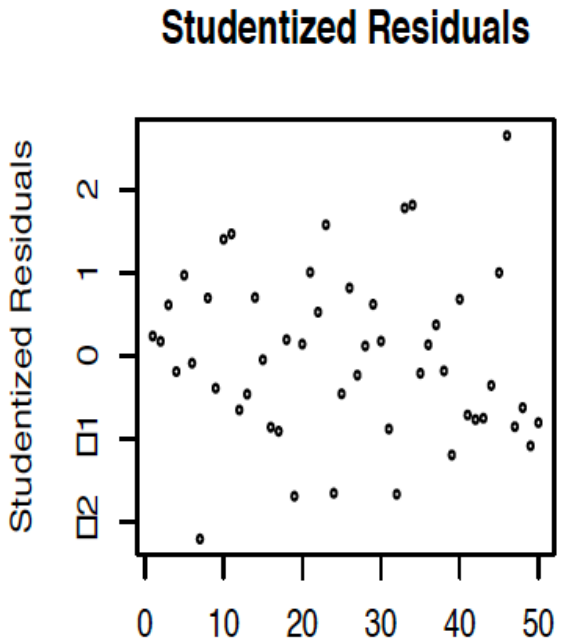

Como detectar outliers?

# Outliers y Observaciones influyentes

Un primer intento para encontrar este tipo de observaciones consiste en *studentizar* a los residuales:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

La idea de esto es homologar la varianza, pues teóricamente los residuales no tiene varianza constante. Luego se espera que estos residuales $r_i$ se encuentren en una banda entre $-3$ y 3, aquellos fuera de estas bandas serán candidatos a ser analizados como outliers del modelo. (Falta de Ajuste del modelo? Errores de captura? )



# Outliers y Observaciones influyentes

El método anterior tiene el problema de que posiblemente el modelo ajustado este siendo influenciado por el dato outlier y por tanto no sea detectado en los residuales, para solucionar esto se estila utilizar lo que se denomina el residual de *jackknife* (Residual de validación cruzada)

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)} \left(1 + \underline{x}_i^T \left(\mathbf{X}_{(i)}^T \mathbf{X}_{(i)}\right)^{-1} \underline{x}_i\right)^{1/2}}$$

Donde $\hat{y}_{(i)} = \underline{x}_i^t \underline{\hat{\beta}_{(i)}}$ es el valor ajustado de la $i-$ésima observación quitando la esa misma observación del ajuste, por lo que $\underline{\hat{\beta}_{(i)}}$ y $\hat{\sigma}_{(i)}$ son los estimadores de $\underline{\hat{\beta}}$ y $\sigma$ quitando la $i$-ésima observación. De la misma forma $\mathbf{X}_{(i)}$ es la matriz de diseño sin el $i-$ésimo renglón. Afortunadamente existe una forma fácil de calcular este residual por medio del residual studentizado.

$$t_i = r_i \left(\frac{n-p-1}{n-p-r_i^2}\right)^{1/2}$$

La gran ventaja de esta última fórmula es que no requiere estar ajustando $n$ regresiones. (Ver función: rstudent)

# Outliers y Observaciones influyentes

Finalmente, se puede probar que bajo el supuesto de que $\underline{\varepsilon} \sim N_n\left(\underline{0}, \sigma^2 \mathbf{I}_n\right)$ se tiene que:

$$t_i \sim t_{n-p-1}$$

Luego entonces, para detectar un outlier se suele utilizar la corrección de Bonferroni y comparar contra el cuantil $\alpha/n$ de una distribución $t$. Por ejemplo si $\alpha = 0.05$ entonces se dice que la i-ésima observación es outlier si

$$|t_i| > t_{n-p-1}^{(1-\alpha^*)} \qquad \text{con } \alpha^* = \frac{0.05}{2n}$$

Notas:

- Dos o mas outliers juntos pueden ocultarse entre ellos.
- Un outlier en un modelo puede dejar de serlo en otro cuando hubieron transformaciones, se recomienda hacer el análisis de outliers cada vez que se hace alguna transformación al modelo
- Revisar funcion *outlierTest*.

# Outliers y Observaciones influyentes

Un punto influyente es aquel que siendo removido del modelo causa un cambio importante en todo el ajuste. Algunas medidas para detectar la influencia de una observación son:

- Cambio en el vector de los coeficientes ajustados: $\underline{\hat{\beta}} - \underline{\hat{\beta}}_{(i)}$
- Cambio en el vector de observaciones $\mathbf{X}^t \left( \underline{\hat{\beta}} - \underline{\hat{\beta}}_{(i)} \right) = \underline{\hat{Y}} - \underline{\hat{Y}}_{(i)}$

Surge el problema de definir una distancia que nos ayude a determinar la diferencia entre vectores. Distancia de Cook:

$$D_i = \frac{\left( \underline{\hat{\beta}} - \underline{\hat{\beta}}_{(i)} \right)^T (\mathbf{X}^T \mathbf{X}) \left( \underline{\hat{\beta}} - \underline{\hat{\beta}}_{(i)} \right)}{p\hat{\sigma}^2}$$

$$D_i = \frac{1}{p} r_i^2 \frac{h_i}{1 - h_i}$$

(R: cooks.distance; lm.influence ) Un regla usada para la detección de observaciones influyentes es considerar $D_i > 1$ ( Cook, R. Dennis; and Weisberg, Sanford (1982)). Pero se ha visto que es muy conservadora, algunos entonces utilizan $D_i > 4/n$ (Bollen, Kenneth A.; and Jackman, Robert W. (1990);)

# Outliers y Observaciones influyentes

Belsey, Kuh y Welch introducen otras dos medias de influencia. La primera es una estadiística que indica cuanto cambia el coeficiente de regresión $\hat{\beta}_j$ cuando la $i$-ésima observación es removida.

$$DFBETA_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{\hat{\sigma}^2_{(i)} C_{jj}}}$$

Donde $C_{jj}$ es el $j$-ésimo elemento de la matriz $\left(\mathbf{X}^t\mathbf{X}\right)^{-1}$. Luego entonces un valor (en magnitud) grande de $DFBETA_{j,i}$ indica que la observación $i$ tiene un influencia sobre el coeficiente $j$ Los autores sugieren un punto de corte igual a $2/\sqrt{n}$

$$|DFBETA_{j,i}| > 2/\sqrt{n}$$

# Outliers y Observaciones influyentes

También podemos medir la influencia de una observación sobre los valores ajustados $\hat{y}_i$. El diagnostico propuesto es:

$$DFFITT_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{\hat{\sigma}^2_{(i)} h_{ii}}}$$

Donde $\hat{y}_{(i)}$ es el valor ajustado para $y_i$, obtenido sin usa la $i$-ésima observación. El denominador no es mas que una estandarización pues se puede probar que $\text{Var}(\hat{y}_i) = \sigma^2 h_{ii}$. (Recuerde que $h_{ii}$ son los elementos de la diagonal de la matriz Sombrero).
La interpretación para el $DFFITT$ es que mide la cantidad de desviaciones estándar que cambia el valor ajustado $\hat{y}_i$ si se elimina la observación $i$. Obviamente un valor grande (magnitud) nos indica una influencia de la observación $i$. Los autores sugieren que merece investigarse toda observación tal que:

$$|DFFITS_i| > 2\sqrt{p/n}$$

# Outliers y Observaciones influyentes

Los diagnósitos que hemos visto permiten conocer el efecto de las observaciones sobre los coeficientes y las estimaciones y no proporcionan información sobre la *precisión* de la estimación. Recordemos que la teoría inferencial nos indica que un estimador es mejor que entro si este último tiene una menor varianza, luego entonces resultaría util saber si quitando una observación la estimación mejora.
Surge entonces la necesidad de tener una medida escalar de la varianza de un vector. Lo que se estila utilizar como una medida escalar es utilizar el determinante de la matriz de varianzas y covarianzas . Definimos entonces la *varianza generalizada* como:

$$GV\left(\underline{\hat{\beta}}\right) = \left|\mathrm{Var}\left(\underline{\hat{\beta}}\right)\right| = \left|\sigma^2\left(\mathrm{X}^t\mathrm{X}\right)^{-1}\right|$$

Finalmente para tener una medida de cuanta precision ganamos o perdemos por quitar una observación definimos lo siguiente:

$$COVRATIO_i = \frac{\left|\left(\mathrm{X}_{(i)}^t\mathrm{X}_{(i)}\right)^{-1}\sigma_{(i)}^2\right|}{\left|\left(\mathrm{X}^t\mathrm{X}\right)^{-1}\sigma^2\right|}$$

# Outliers y Observaciones influyentes

Queda claro entonces que un $COVRATIO_i < 1$ indica que la $i$-ésima observación mejora la precisión de la estimación. No es fácil obtener valores de corte para este índice pero Belsley, Kuh y Welsh (1980) sugieren que si:

$$COVRATIO_i > 1 + 3p/n \quad \text{o} \quad COVRATIO_i < 1 - 3p/n$$

Entonces se debe considerar al punto $i$ como influyente. (Estos valores solo se recomiendan para muestras grandes).
http://www.statmethods.net/stats/rdiagnostics.html

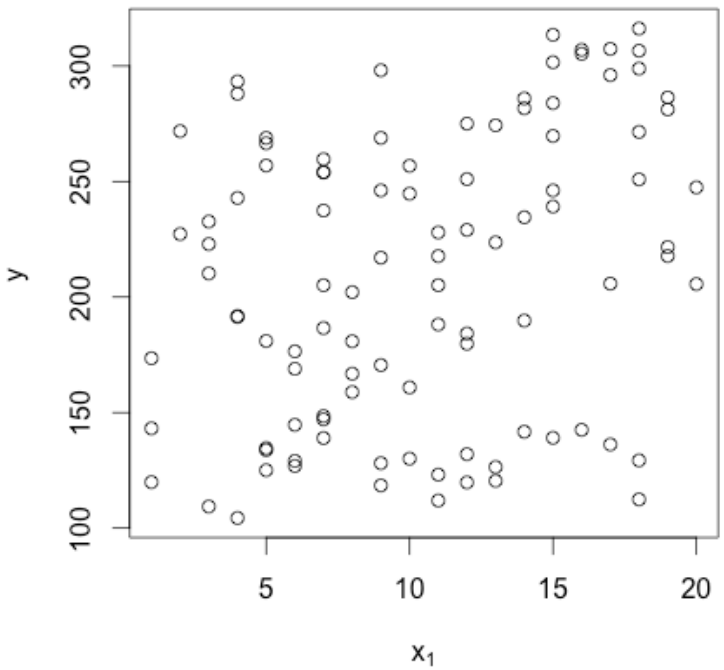
# Técnicas de Transformación de variables

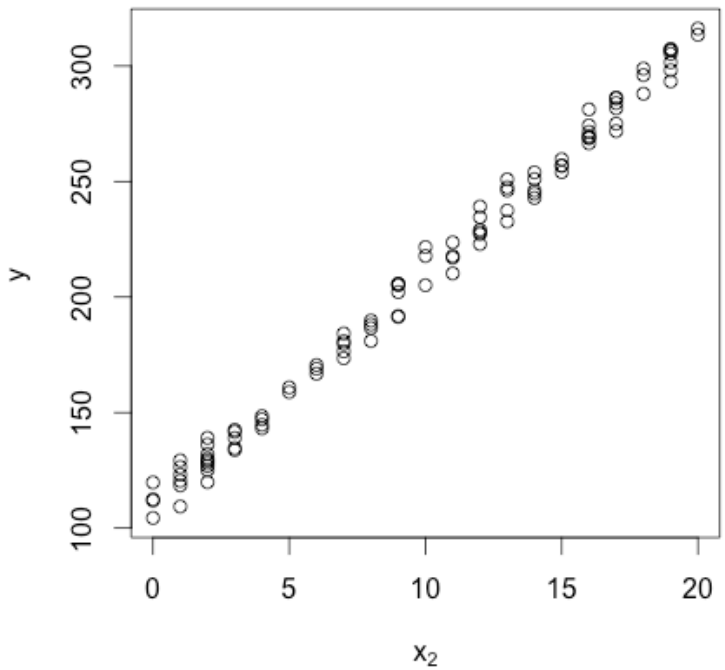
Muchas veces, durante el análisis de residuales, surgirá la necesidad de llevar a cabo alguna transformación de nuestro modelo para poder sustentar las hipótesis del modelo lineal. Desafortunadamente no existe una **receta** que nos indique que tipo de transformación aplicar ante las distintas situaciones a las que nos estemos enfrentando, sin embargo podemos explicar las técnicas principales. En el caso lineal simple es fácil identificar cuando no se cumple la linealidad y visualizar la transformación que se pueda adecuar para tener un mejor modelo.

| Modelo Real (Desconocido) | Transformación | Modelo Lineal |
|---|---|---|
| $Y = \beta_0 + \beta_1 X^k$ | $Z = X^k$ | $Y = \beta_0 + \beta_1 Z$ |
| $Y = \beta_0 + \beta_1 ln(X)$ | $Z = ln(X)$ | $Y = \beta_0 + \beta_1 Z$ |
| $Y = \beta_0 e^{\beta_1 X}$ | $V = ln(Y)$ | $V = ln(\beta_0) + \beta_1 X$ |
| $Y = \beta_0 x^{\beta_1}$ | $V = ln(Y); Z = ln(X)$ | $V = ln(\beta_0) + \beta_1 Z$ |

# Técnicas de Transformación de variables

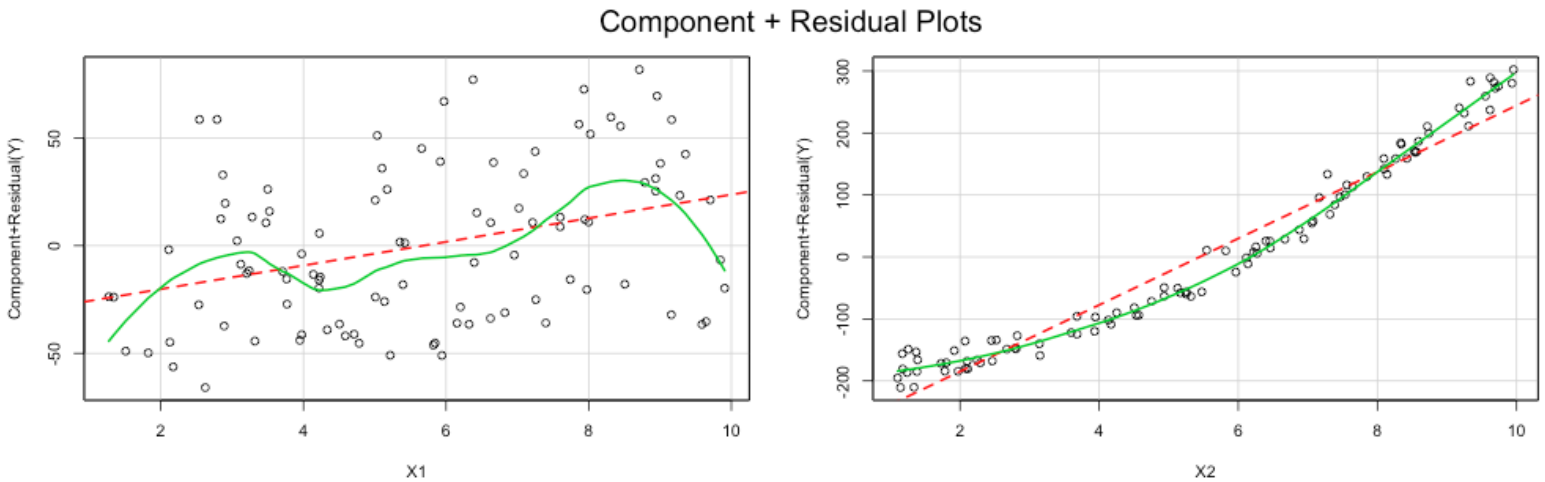
Desafortunadamente cuando tenemos varias variables explicativas es dificil identificar cual variable requiere una transformación para mejorar la linealidad del modelo. Un primer intento puede ser graficar $Y$ contra cada una de las variables explicativas $X_i$ y ver en cual no se está cumpliendo la linealidad, sin embargo debido a la posible correlación entre las variables este gráfico puede no servir.



# Técnicas de Transformación de variables

Para solucionar este problema se estila hacer la gráfica de residuales parciales. (Ver función crPlots() ). La idea de los residuales parciales es simple. Supongamos que queremos ver si existe una relación lineal entre la variables $X_q$ y $Y$, entonces lo que haces es ajustar el modelo **eliminando** a la variable $X_q$ del modelo, luego encontramos los residuales de dicho modelo (denotamos por $e_i^{(q)}$) y luego graficamos esos residuales contra los valores de la variables $X_q$, es decir visualizamos los puntos $\left(x_{iq}, e_i^{(q)}\right)$. Si la relación es lineal entre $X_q$ y $Y$ se espera que esta gráfica de residuales parciales presente una relación lineal, en caso contrario, es una indicación de que esa variable posiblemente necesite ser transformada para ayudar al ajuste.



# Técnicas de Transformación de variables

La técnica de los residuales parciales se basa en lo siguiente, imaginemos que tenemos el siguiente modelo lineal

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + \varepsilon_i$$

Evidentemente no hay una relación lineal entre la variable $X_2$ y $Y$ pero si la hay entre $X_1$ y $Y$. Imaginemos que queremos hacer los residuales parciales repecto a $X_1$, para ello escribamos el modelo original de la siguiente forma y definamos $\varepsilon_i^{(1)} = \varepsilon_i + \beta_1 x_{1i}$ entonces

$$y_i = \beta_0 + \beta_2 x_{2i}^2 + (\varepsilon_i + \beta_1 x_{1i}) \Rightarrow y_i = \beta_0 + \beta_2 x_{2i}^2 + \varepsilon_i^{(1)}$$

La ecuación de la derecha no es mas que el modelo sin la variable $X_1$, si ajustamos este modelo, sabemos que los residuales (denotados por $e_i^{(1)}$) son una aproximación de $\varepsilon_i^{(1)}$, entonces:

$$e_i^{(1)} \approx \varepsilon_i^{(1)} = \varepsilon_i + \beta_1 x_{1i}$$

Luego entonces los residuales del modelo eliminando $X_1$ guardan una relación lineal con $X_1$.

# Técnicas de Transformación de variables

Ahora hagamos lo mismo pero para $X_2$.Imaginemos que queremos hacer los residuales parciales repecto a $X_2$, para ello escribamos el modelo original de la siguiente forma y definamos $\varepsilon_i^{(2)} = \varepsilon_i + \beta_2 x_{2i}^2$ entonces

$$y_i = \beta_0 + \beta_1 x_{1i} + \left(\varepsilon_i + \beta_2 x_{2i}^2\right) \Rightarrow y_i = \beta_0 + \beta_2 x_{2i}^2 + \varepsilon_i^{(2)}$$

La ecuación de la derecha no es mas que el modelo sin la variable $X_2$, si ajustamos este modelo, sabemos que los residuales (denotados por $e_i^{(2)}$) son una aproximación de $\varepsilon_i^{(2)}$, entonces:

$$e_i^{(2)} \approx \varepsilon_i^{(2)} = \varepsilon_i + \beta_2 x_{2i}^2$$

Luego entonces los residuales del modelo eliminando $X_2$ guardan una relación no lineal con $X_2$. (Aquí es donde detectamos la no linealidad de la variables $X_2$!!!!). Una vez detectada, el analista viendo el comportamiento de la gráfica debe de proponer una transformación, a aveces se estila proponer un polinomio de grado no mayor a 4 y luego hacer una selección de variables ahi. (Ver ejemplo: crplots.R)

# Técnicas de Transformación de variables

El método anterior solo nos ayuda a encontrar posibles transformaciones para las covariables. Surge la necesidad de tener un método que nos ayude a encontrar una buena transformación para la variable $Y$ en caso de que sea necesario. En 1964, Box y Cox introdujeron una transformación de la variable respuesta con el objetivo de satisface la suposición de normalidad del modelo de regresión. La transformación está definida como:

$$w = \frac{y^{\lambda} - 1}{\lambda} \quad si \quad \lambda \neq 0 \quad w = ln\,(y) \quad si \quad \lambda = 0$$

En este caso $\lambda$ es otro parámetro en el modelo de regresión es decir:

$$w = \frac{y^{\lambda} - 1}{\lambda} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

Donde ahora los parámetros son: $\left(\lambda, \beta_0, \ldots, \beta_p, \sigma^2\right)$. La forma en como se estima $\lambda$ es por maxima verosimilitud. (Ver función boxCox)

# Metodos de selección de variables

Hasta ahora hemos venido trabajando suponiendo que contamos con todas las variables explicativas que influyen en el modelo. Por lo general esta situación solo ocurre en contadas ocasiones donde la experiencia previa del analista son útiles para seleccionar las variables que deben de estar ajustando en el modelo, sin embargo, en la mayora de los casos , el analísta tiene una variedad de variables candidatas que podrían o no influir en la variable respuesta.
Al problema de encontrar un subconjunto adecuado de variables explicativas se le conoce como el *el problema de selección de variables*

# Metodos de selección de variables

El algoritmo general para llevar a cabo la selección del modelo se puede resumir en lo siguiente:

- Se emplea un criterio de selección de variables en particular
- El modelo resultante se revisa para verificar que las especificaciones funcionales sean correctas (Análisis de Residuales) y que no existan outliers ni observaciones de alta influencia.
- Si el modelo no pasa el punto anterior, se debe de repetir el proceso de selección de variables omitiendo el modelo resultante de las iteraciones anteriores.

Debe de quedar claro que ninguno de los procedimientos garantizarán que se encuentre la mejor ecuación de regresión, además el analísta no debe de confiar demasiado en los resultados de un procedimiento en particular de selección de variables. La experiencia previa del analista, juicios personales siempre deben de entrar en el problema de selección de la mejor ecuación de regresión.

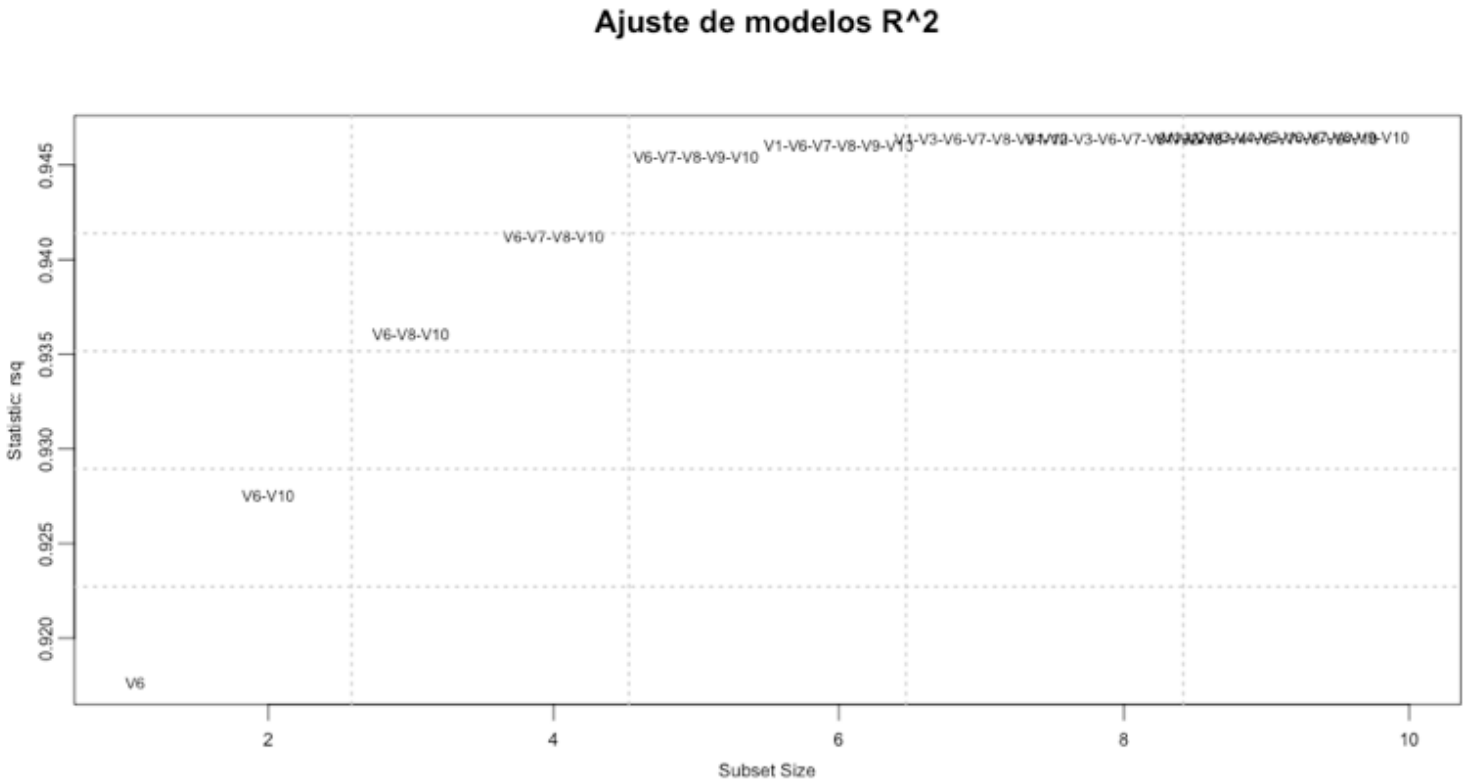
# Cirterios para Evaluar Modelos de Regresión

A continuación se presentan criterios para evaluar y comprar modelos de regresión, debe de quedar claro que todos estos modelos deben de ser submodelos (modelos incompletos) de un modelo general saturado:

- Coeficiente de Determiación Múltiple. $R^2$
- $R^2$ Ajustado (Penaliza la entrada de variables)
- SCE (Suma de cuadrados del error)

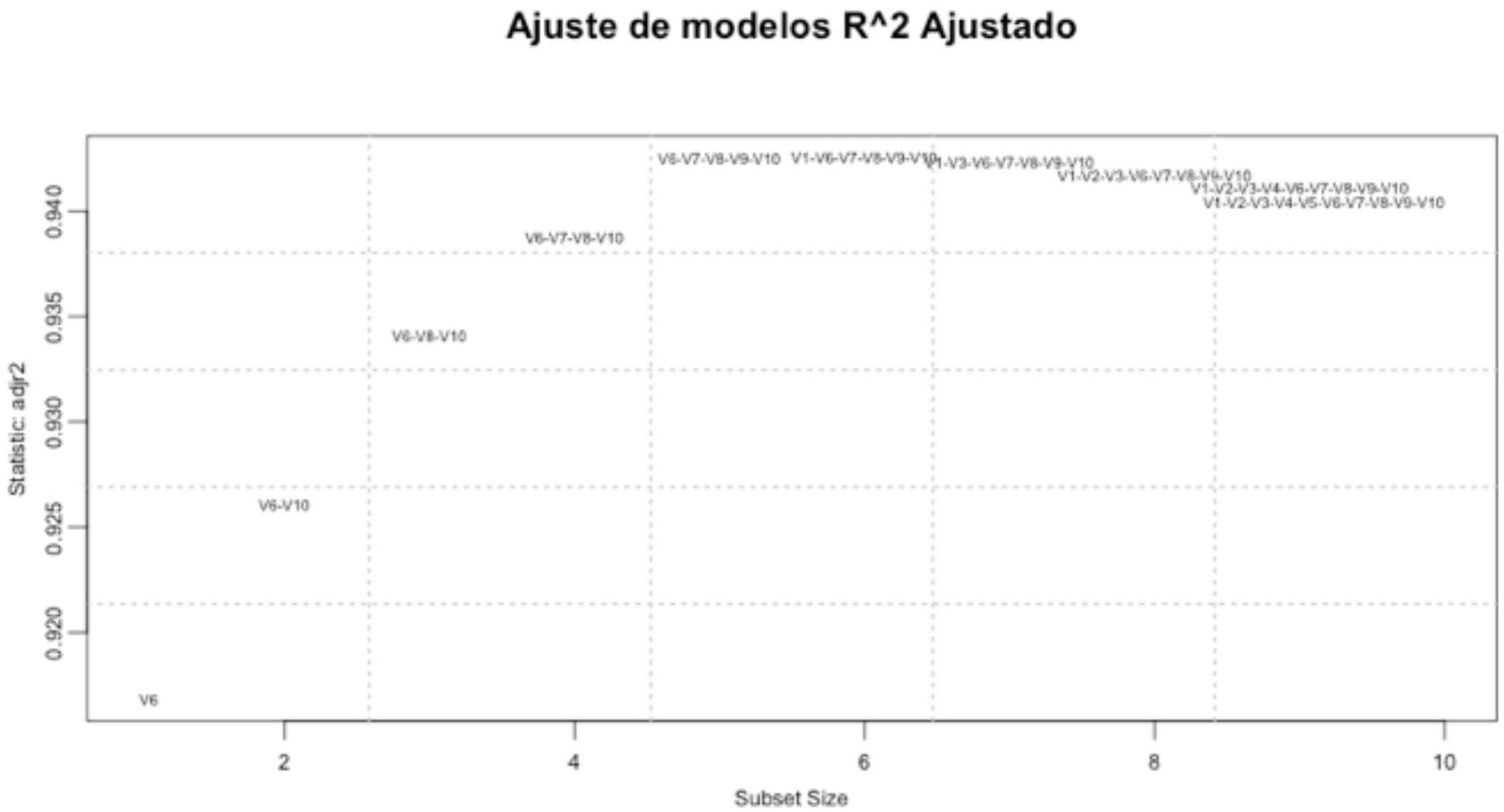
# Cirterios para Evaluar Modelos de Regresión

Cuando el número de variables lo permite $p$ pequeña ($< 30$) lo que se puede hacer es realizar un análisis exhaustivo de todos los modelos posibles: $2^{30} - 1 = 1,073,741,824$ y escoger el mejor bajo algún criterio. Por ejemplo:

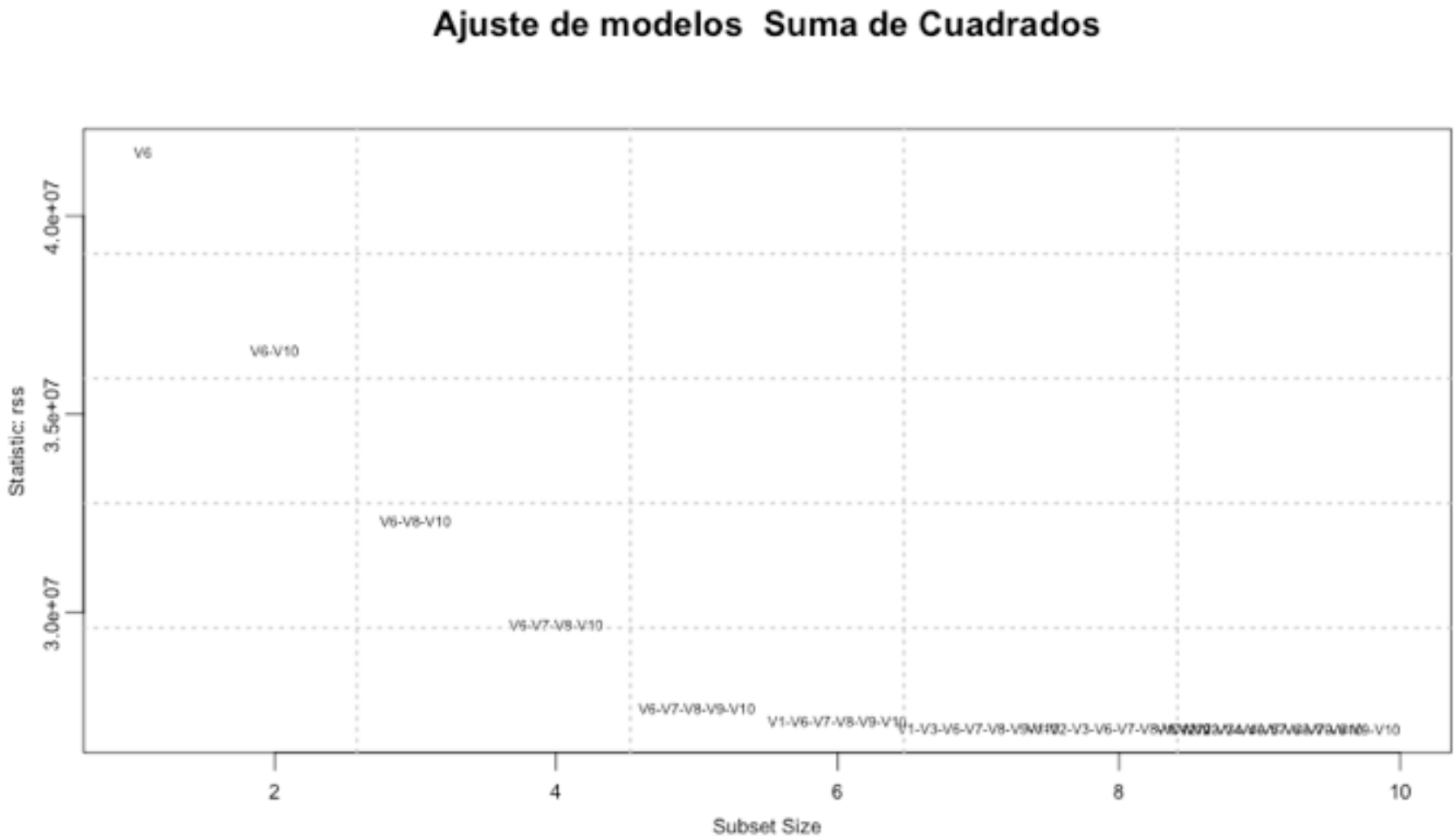



(ver función regsubsets de la libreria leaps)

# Cirterios para Evaluar Modelos de Regresión



# Cirterios para Evaluar Modelos de Regresión



# Cirterios para Evaluar Modelos de Regresión

Los métodos anteriores tiene el problema que no nos dan un punto de corte que decida con cuantas variables ajustemos. Para resolver este problema en la literatura existen algoritmos que determinan el mejor modelo con base en ciertas pruebas estadísticas. Los métodos mas utilizados son:

- Forward
- Backward
- Stepwise

Ojo: Los algoritmos no siempre garantizan encontrar el mejor modelo además de que cada metodología puede arrojar distintas soluciones. El analísta no debe de confiar demasiado en los resultados de un procedimiento en particular de selección de variables. La experiencia previa del analista y juicios personales siempre deben de entrar en el problema de selección de la mejor ecuación de regresión.

# Metodo Forward

El método Forward (o de selección hacia adelante): Inicialmente no hay ninguna variable seleccionada. Se comienza eligiendo la variable que más correlacionada está con la variable respuesta , y si su p-value es pequeño (menor a $\alpha$) entonces se queda en el modelo en caso de que no sea significativa el algoritmo termina. A continuación selecciona la segunda variable que junto con la primera tiene un valor de la estadística $F$ mas grande, lo que se traduce en un incremento grande en la SCR. El algoritmo finaliza cuando ninguna de las variables no seleccionadas agrega un incremento significativo en la SCR. Ejemplo:

$$H_0 : y_i = \beta_0 + \beta_q x_{iq} + \varepsilon_i \quad vs \quad H_1 : y_i = \beta_0 + \beta_q x_{iq} + \beta_j x_{ij} + \varepsilon_i \quad j \in \{1, \ldots, k\} - \{q\}$$

$$F = \frac{\frac{SCR_{H_1} - SCR_{H_0}}{\sigma^2}}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sigma^2(n-(2+1))}} = \frac{(SCR_{H_1} - SCR_{H_0})}{\hat{\sigma}^2} = \frac{(SCE_{H_0} - SCE_{H_1})}{\hat{\sigma}^2} \sim \mathscr{F}_{(1,n-(2+1))}$$

# Metodo Backward

El método Backward (o de eliminación hacia atrás) actúa de forma inversa. Se comienza seleccionando todas las variables. En cada paso se elimina una variable del modelo, en este caso se decide sacar la que menos significancia tiene, es decir un $F$ pequeño. El algoritmo finaliza cuando quitar alguna otra variable del modelo ocasiona un reducción significativa en la SCR. Ejemplo:

$$H_0 : y_i = \beta_0 + \beta_1 x_{i1} \ldots \beta_{k-1} x_{i(q-1)} + \beta_{q+1} x_{i(q+1)} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

$$H_1 : y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

$$F = \frac{\frac{SCR_{H_1} - SCR_{H_0}}{\sigma^2}}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sigma^2(n-p)}} = \frac{(SCR_{H_1} - SCR_{H_0})}{\hat{\sigma}^2} = \frac{(SCE_{H_0} - SCE_{H_1})}{\hat{\sigma}^2} \sim \mathscr{F}_{(1,n-p)}$$

Se elimina la $k$-ésima variable si el estadístico $F$ es mas chico que el cuantil $\mathscr{F}^{1-\alpha}_{(1,n-p)}$ y si además es el valor $F$ mas pequenño que se obtuvo entre todas las posibles $q \in \{1, \ldots, k\}$

# Metodo Stepwise

El método Stepwise (o regresión por pasos), fue originalmente propuesto por Efroymson en 1960 [**?**], ha estado disponible en los paquetes estadísticos desde hace muchos años. Este método utiliza una combinación de los dos algoritmos anteriores: en cada paso se introduce o elimina una variable dependiendo de la significación con la que se está trabajando. Debido a la posible correlación de las variables es posible *arrepentirse* de decisiones tomadas en pasos anteriores, bien sea eliminando del conjunto seleccionado la variable introducida en un paso anterior del algoritmo, o bien sea seleccionando una variable previamente eliminada. El algoritmo puede ser inicializado con todas las variables en cuyo caso el primer paso será la eliminación de una variable o bien también se puede inicializar el algoritmo con ninguna variable, en cuyo caso el primer paso siempre será incluir una variable. Es proceso tiene el problema de que en alguna interación se llegue a un ciclo repetido sacando y metiendo la misma variable, en cuyo caso se decide terminar el algoritmo en ese momento.