ESTADISTICA 2

MUESTREO

- Pasos:
  1) Definir problema
  2) Relacion problema - Estimacion [M, T, P, A]
  3) Definir: Elemento, Poblacion, Unidad de muestreo y marco.
  4) Definir Metodologia de Muestreo [MAS, MAE, MS, MC, MVE]
  5) Calcular tamaño de muestra con la muestra piloto. NC
  6) Seleccionar Muestra
  7) Hacer Inferencia

- Definicion: Parte de la estadistica que se encarga del estudio de metodos para la obtencion de una muestra de una poblacion objeto de estudio

Muestra = Parte de la poblacion a partir de la cual hacemos inferencia (sin remplazo)

→ Sacar conclusiones sobre toda la poblacion a partir de la Muestra

→ Coleccion de unidades de muestreo obtenidas a partir de un marco o marcos. (representativa)

Estimamos

M = Media

T = Total$_M$ Suma

P = Proporcion

A = Total$_P$ atributo

Poblacion: Conjunto de elementos acerca de los cuales se realizara la inferencia. Estudiantes de primer Semestre Fac Minas, tiendas visibles en medellin

Unidad de Muestreo: Colecciones no traslapadas de elementos de la poblacion que cubren la poblacion completa. Asignaturas introductorias, Manzanas, Parcelas, intervalos de tiempo

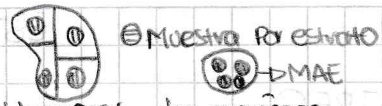
Elemento: Objeto sobre el que se toma medicion o Conteo. Cada estudiante de primer semestre Fac. Minas, tiendas de barrio

Marco Muestral: Lista de unidades de muestreo. Listado asignaturas introductorias. Listado de manzanas

METODOLOGIAS DE MUESTREO

1) Muestreo Aleatorio Simple (MAS) Independiente
   * Poblacion homogenea con respecto a la caracteristica de interes
   * Principio de equiprobabilidad

2) Muestreo Aleatorio Estratificado (MAE):



* Poblacion heterogenea
* La poblacion es divisible en subpoblaciones homogeneas (Estratos) no traslapados
* De cada estrato se selecciona una MAS
* Aumenta la precision.

3) Muestreo Sistematico (MS)

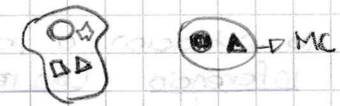
* Para poblaciones en movimiento
* Estimaciones similares a las del MAS

Aleatorio (5)
↳ equidistantes
↳ 10, 15, 20, 25

Control de calidad industrial
- La no respuesta daña el muestreo

4) Muestreo por conglomerados (MC)

* La poblacion esta conformada por subpoblaciones (conglomerados) homogeneas internamente, distintas de unidades de muestreo
* Se selecciona una MAS de conglomerados y se muestrean todas las unidades de los conglomerados seleccionados



5) Muestreo en varias etapas (MVE)

* Disminuye error de MC
* Aumenta varianza

Estratificado → Conglomerado → Simple

PLANIFICACION ENCUESTA

1) Declaracion de objetivos: concisos, simples
2) Poblacion objetivo: poblacion factible seleccionada cuidadosamente
3) Marco: La lista y la poblacion objetivo coincidan lo maximo, La eficiencia puede aumentar con marcos multiples
4) Diseño de muestreo: # de elementos de la muestra que proporcione suficiente ~~[illegible]~~ informacion para los objetivos (representativos)
5) Metodo de Medicion: Entrevistas, cuestionarios personales, telefonicos, correos u observacion
6) Instrumento de medicion: Minimizar la no respuesta y el sesgo por respuesta incorrecta
7) Seleccion y formacion de personal de campo.
8) Pre-test: Muestreo piloto, calificacion entrevistadores, operatividad
9) Organizacion: establecer lineas de autoridad
10) Administracion esquematica de los datos
11) Analisis de datos.

MUESTREO PILOTO

* Validar preguntas
* Estimar varianza
* Determinar tamaño de muestra
* Da ide para el limite de error de estimacion definitivo

TAMAÑO DE MUESTRA

* Depende del tipo de muestreo seleccionado y del muestreo piloto
* Es el numero minimo de elementos que estaran en la muestra
* Es el numero de elementos que permite hacer inferencia

NIVEL DE CONFIANZA Y LIMITE DE ERROR DE ESTIMACION
(NC) (B)

* Son fijados por el investigador (Los dos)
* B, dos desviaciones estandar del estimador

$$\text{Error de estimacion} = |\theta - \hat{\theta}| < B$$

$$P[\text{Error de estimacion}] = 1 - \alpha$$

$$NC = 1 - \alpha$$

OBSERVACIONES GENERALES

- Cada elemento del muestreo provee informacion de los parametros de interes
- Poca informacion impide realizar buenas estimaciones
- Mucha informacion despilfarro de recursos
- La encuesta controla la variacion de los datos
- En la encuesta se debe recolectar informacion adicional que ayude a la pregunta de interes.
- El mejor diseño da la precision necesaria en terminos de B con un coste minimo

MUESTREO ALEATORIO SIMPLE (MAS)

Procedimiento estadistico a partir del cual se selecciona una muestra de tamaño n de una poblacion de N unidades garantizando que cada muestra posible de tamaño n tenga la misma probabilidad de ser seleccionada (Principio de equiprobabilidad)

- En la practica una MAS es seleccionada unidad por unidad
- El muestreo se realiza sin remplazo. (NO se repiten las unidades)

El MAS se utiliza cuando la poblacion es homogenea con respecto a la caracteristica de interes y cuando las estimaciones se refieren a toda la poblacion y no a subgrupos

Estadisticos e intervalos

- Media Muestral (M)

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

→ y_i = Observaciones de la variable de interes
n = tamaño de la muestra

$$Var[\bar{y}] = \frac{\sigma^2}{n} \left(\frac{N-n}{N-1}\right)$$

→ Factor de correccion para poblaciones finitas

$$\hat{\sigma}^2 = \frac{N-1}{N} \cdot S^2$$

N = tamaño de la poblacion

- Varianza Estimada (M)

$$\hat{Var}[\bar{y}] = \frac{\hat{\sigma}^2}{n} \cdot \frac{N-n}{N-1} = \frac{N-1}{N} \cdot S^2 \, \frac{N-n}{N-1}$$

$$\hat{Var}[\bar{y}] = \frac{S^2}{n} \left(\frac{N-n}{N}\right)$$

→ Factor de correccion para poblaciones finitas

- Intervalos de confianza (M)

↳ Limite en el error de estimacion B

↳ $\bar{y} \pm B_M$

Donde B depende de n:

Si $n \geq 30$

$$\boxed{B = Z_{\frac{\alpha}{2}} \sqrt{\frac{S^2}{n}\left(\frac{N-n}{N}\right)}}$$

Entonces

$$\bar{y} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{S^2}{n}\left(\frac{N-n}{N}\right)}$$

Si $n \leq 30$

$$\boxed{B = t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{S^2}{n}\left(\frac{N-n}{N}\right)}}$$

Entonces

$$\bar{y} \pm t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{S^2}{n}\left(\frac{N-n}{N}\right)}$$

- Total de la poblacion (τ)

$$\tau = N\mu$$

$$\hat{\tau} = N\hat{\mu} = N\bar{y}$$

- Varianza Estimada (τ)

$$\hat{Var}[\hat{\tau}] = N^2 \frac{s^2}{n}\left(\frac{N-n}{N}\right)$$

- Intervalo de confianza (τ)

$$\hat{\tau} \pm B_\tau \qquad N(\bar{y} \pm B_\mu)$$

Donde B_τ depende de n

Si $n > 30$

$$\boxed{B_\tau = z_{\frac{\alpha}{2}}\sqrt{N^2 \frac{s^2}{n}\left(\frac{N-n}{N}\right)}}$$

Entonces

$$\hat{\tau} \pm z_{\frac{\alpha}{2}}\sqrt{N^2 \cdot \frac{s^2}{n}\left(\frac{N-n}{N}\right)}$$

Si $n \leq 30$

$$\boxed{B_\tau = t_{\frac{\alpha}{2}, n-1}\sqrt{N^2 \frac{s^2}{n}\left(\frac{N-n}{N}\right)}}$$

Entonces

$$\hat{\tau} \pm t_{\frac{\alpha}{2}, n-1}\sqrt{N^2 \cdot \frac{s^2}{n}\left(\frac{N-n}{N}\right)}$$

- Proporcion Poblacional (p)

$$y_i = \begin{cases} 1, & \text{tiene el atributo} \\ 0, & \text{no tiene el atributo} \end{cases}$$

$$\hat{p} = \frac{\sum y_i}{n} = \frac{\text{\# Unidades con el atributo}}{\text{\# unidades en la muestra}}$$

- Varianza Estimada (p)

$$\hat{Var}[\hat{p}] = \frac{p(1-p)}{n-1}\left(\frac{N-n}{N}\right)$$

- Intervalo de confianza

$$\hat{p} \pm B_p$$

Donde B_p depende de n

Si $n \geq 100$

$$\boxed{B_p = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1} \left(\frac{N-n}{N}\right)}}$$

Entonces

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1} \left(\frac{N-n}{N}\right)}$$

Si $n < 100$

$$\boxed{B_p = t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1} \left(\frac{N-n}{N}\right)}}$$

Entonces

$$\hat{p} \pm t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1} \left(\frac{N-n}{N}\right)}$$

- Total de la Poblacion (A)

$$\hat{A} = N\hat{p}$$

- Varianza estimada (A)

$$\hat{V}ar[\hat{A}] = N^2 \cdot \frac{\hat{p}(1-\hat{p})}{n-1} \left(\frac{N-n}{N}\right)$$

- Intervalo de confianza (A)

$$\hat{A} \pm B_A \qquad N(\hat{p} \pm B_p)$$

Donde B_A depende de n

Si $n \geq 100$

$$\boxed{B_A = z_{\frac{\alpha}{2}} \sqrt{N^2 \frac{(1-\hat{p})\hat{p}}{n-1} \left(\frac{N-n}{N}\right)}}$$

Entonces

$$\hat{A} \pm z_{\frac{\alpha}{2}} \sqrt{N^2 \frac{\hat{p}(1-\hat{p})}{n-1} \left(\frac{N-n}{N}\right)}$$

Si $n < 100$

$$\boxed{B_A = t_{\frac{\alpha}{2}, n-1} \sqrt{N^2 \frac{(1-\hat{p})\hat{p}}{n-1} \left(\frac{N-n}{N}\right)}}$$

Entonces

$$\hat{A} \pm t_{\frac{\alpha}{2}, n-1} \sqrt{N^2 \frac{\hat{p}(1-\hat{p})}{n-1} \left(\frac{N-n}{N}\right)}$$

Conocidos B y el nivel de confianza $1-\alpha$ fijados por el investigador para estimar M, se busca hallar cual debe ser el tamaño de muestra n. Para ello partimos del limite de error de estimacion B

$$B = z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}}$$

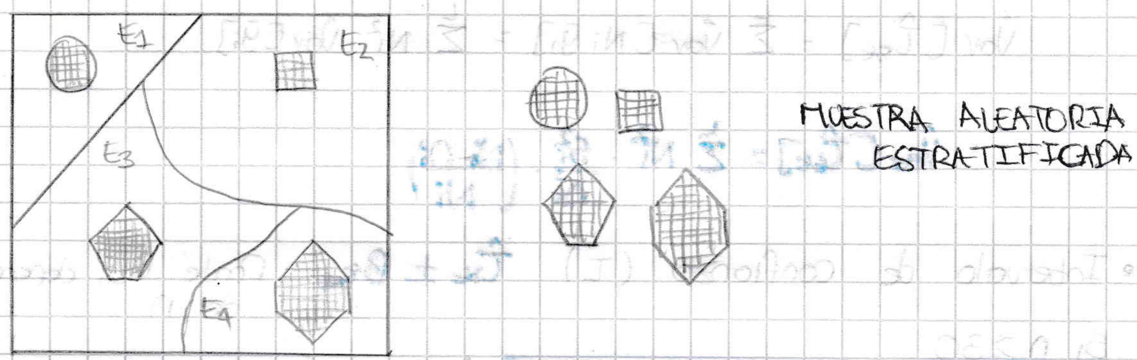
$$n = \frac{\sigma^2 N}{\frac{B^2}{Z^2_{\frac{\alpha}{2}}}(N-1) + \sigma^2} \quad ; \quad n = \frac{1}{\frac{1}{N} + \frac{N-1}{N} \cdot \frac{1}{n_0}} \quad , \quad n_0 = \frac{Z^2_{\frac{\alpha}{2}} \sigma^2}{B^2}$$

Tamaño de muestra cuando la población es infinita

Para una muestra para la media (M)	Para una muestra para la proporción (P)
$\sigma^2 = \hat{S}^2$	$\sigma^2 = \hat{P}(1-\hat{P})$

MUESTREO ALEATORIO ESTRATIFICADO (MAE)

Técnica de muestreo mediante la cual se obtiene una muestra a partir de la separación de los elementos de una población en grupos no traslapados, llamados estratos y la selección posterior de una muestra aleatoria simple, independiente, en cada estrato



- Maximiza la información obtenida
- Minimiza el límite del error de estimación (B)
- La variabilidad de la población es heterogénea
- La variabilidad dentro de los estratos es homogénea
- Disminuye costos
- Permite estimaciones para los subgrupos (estratos)

• Tamaño de la población (N)

$$N = \sum_{i=1}^{L} N_i$$

N = Tamaño de la población
N_i = Unidades muestrales en el i-ésimo estrato
L = # de estratos

- Tamaño de la muestra (n)

$$n = \sum_{i=1}^{L} n_i$$

n_i = # Unidades muestrales seleccionadas del estrato i

- Medidas de resumen por estrato

$\bar{y}_i$ = Media muestral de la MAS seleccionada del estrato i
S_i^2 = Varianza muestral de la MAS seleccionada del estrato i
$P_i = a_i / n_i$ = # Unidades muestrales en la muestra que tienen el atributo

- Total de toda la población (τ)

$$\hat{\tau}_{st} = \sum^{L} \hat{\tau}_i = \sum^{L} N_i \bar{y}_i$$

$$\hat{\tau}_{st} = \sum^{L} N_i \bar{y}_i$$

- Varianza Estimada de τ

$$\hat{Var}[\hat{\tau}_{st}] = \sum^{L} \hat{Var}[N_i \bar{y}_i] = \sum^{L} N_i^2 \cdot \hat{Var}[\bar{y}_i]$$

$$\hat{Var}[\hat{\tau}_{st}] = \sum^{L} N_i^2 \cdot \frac{S_i^2}{n_i} \cdot \left(\frac{N_i - n_i}{N_i}\right)$$

- Intervalo de confianza (τ) $\hat{\tau}_{st} \pm B_{\tau_{st}}$, Donde $B_{\tau_{st}}$ depende de n

Si $n > 30$

$$\boxed{B_{\tau_{st}} = Z_{\alpha/2} \sqrt{\sum^{L} N_i^2 \cdot \frac{S_i^2}{n_i} \left(\frac{N_i - n_i}{N_i}\right)}}$$

Entonces

$$\hat{\tau}_{st} \pm Z_{\frac{\alpha}{2}} \sqrt{\sum^{L} N_i^2 \cdot \frac{S_i^2}{n_i} \left(\frac{N_i - n_i}{N_i}\right)}$$

Si $n \leq 30$

$$\boxed{B_{\tau_{st}} = t_{\frac{\alpha}{2}; n-L} \sqrt{\sum^{L} N_i^2 \cdot \frac{S_i^2}{n_i} \left(\frac{N_i - n_i}{N_i}\right)}}$$

Entonces

$$\hat{\tau}_{st} \pm t_{\frac{\alpha}{2}, n-L} \sqrt{\sum^{L} N_i^2 \cdot \frac{S_i^2}{n_i} \left(\frac{N_i - n_i}{N_i}\right)}$$

- Media Muestral (μ)

$$\tau = N \cdot \mu \Rightarrow \mu = \frac{\tau}{N} \; ; \; \mu = \bar{y}_{st} \Rightarrow \bar{y}_{st} = \frac{\tau_{st}}{N}$$

$$\bar{y}_{st} = \frac{1}{N} \cdot \sum^{L} N_i \bar{y}_i$$

$\frac{N_i}{N}$ = Proporción de Unidades Muestrales en el estrato i

• Varianza estimada (M)

$$\mathrm{Var}[\bar{y}_{st}] = \mathrm{Var}\left[\frac{1}{N} \cdot \sum^{L} N_i \bar{y}_i\right] = \frac{1}{N^2} \cdot \hat{\mathrm{Var}}[\hat{\tau}_{st}]$$

$$\hat{\mathrm{Var}}[\bar{y}_{st}] = \sum^{L} \frac{N_i^2}{N^2} \cdot \frac{S_i^2}{n_i} \left(\frac{N_i - n_i}{N_i}\right)$$

• Intervalo de confianza (M)

$\hat{\mu} \pm B_\mu$; Donde B_μ depende de n $\quad \frac{1}{N}(\hat{\tau}_{st} \pm B_{\hat{\tau}_{st}})$

$n > 30$	$n \leq 30$
$\boxed{B_\mu = Z_{\frac{\alpha}{2}} \sqrt{\sum^{L} \frac{N_i^2}{N^2} \frac{S_i^2}{n_i} \left(\frac{N_i - n_i}{N_i}\right)}}$	$\boxed{B_\mu = t_{\frac{\alpha}{2}, n-L} \sqrt{\sum^{L} \frac{N_i^2}{N^2} \cdot \frac{S_i^2}{n_i} \left(\frac{N_i - n_i}{N_i}\right)}}$
$\bar{y}_{st} \pm Z_{\frac{\alpha}{2}} \sqrt{\sum^{L} \frac{N_i^2}{N^2} \frac{S_i^2}{n_i} \left(\frac{N_i - n_i}{N_i}\right)}$	$\bar{y}_{st} \pm t_{\frac{\alpha}{2}, n-L} \sqrt{\sum^{L} \frac{N_i^2}{N^2} \frac{S_i^2}{n_i} \left(\frac{N_i - n_i}{N_i}\right)}$

• Total de la poblacion (A)

$$\hat{A} = \sum^{L} \hat{A}_i = \sum^{L} N_i \hat{P}_i$$

$$\hat{A}_{st} = \sum^{L} N_i \hat{p}_i$$

• Varianza estimada (A)

$$\hat{\mathrm{Var}}[\hat{A}_{st}] = \hat{\mathrm{Var}}\left[\sum^{L} N_i \hat{P}_i\right] = \sum^{L} N_i^2 \cdot \hat{\mathrm{Var}}[\hat{P}_i]$$

$$\hat{\mathrm{Var}}[\hat{A}_{st}] = \sum^{L} N_i^2 \cdot \frac{\hat{P}_i(1-\hat{P}_i)}{n_i - 1} \left(\frac{N_i - n_i}{N_i}\right)$$

• Intervalo de confianza (A)

$\hat{A}_{st} \pm B_{\hat{A}_{st}}$, Donde $B_{\hat{A}_{st}}$ depende de n

Si $n > 30$	Si $n \leq 30$
$\boxed{B_{\hat{A}_{st}} = Z_{\frac{\alpha}{2}} \sqrt{\sum^{L} N_i^2 \cdot \frac{\hat{P}_i(1-\hat{P}_i)}{n_i - 1} \left(\frac{N_i - n_i}{N_i}\right)}}$	$\boxed{B_{\hat{A}_{st}} = t_{\frac{\alpha}{2}, n-L} \sqrt{\sum^{L} N_i^2 \cdot \frac{\hat{P}_i(1-\hat{P}_i)}{n_i - 1} \left(\frac{N_i - n_i}{N_i}\right)}}$
Entonces	Entonces

$$\hat{A}_{st} \pm Z_{\frac{\alpha}{2}} \sqrt{\sum^{L} N_i^2 \frac{\hat{p}_i(1-\hat{p}_i)}{n_i - 1} \left(\frac{N_i - n_i}{N_i}\right)} \quad // \quad \hat{A}_{st} \pm t_{\frac{\alpha}{2}, n-L} \sqrt{\sum^{L} N_i^2 \frac{\hat{p}_i(1-\hat{p}_i)}{n_i - 1} \left(\frac{N_i - n_i}{N_i}\right)}$$

- Proporcion poblacional (P)

$$\hat{A}_{st} = N\hat{p}_{st} \Rightarrow \hat{p}_{st} = \frac{\hat{A}_{st}}{N} \Rightarrow \hat{p}_{st} = \frac{\sum^{L} A_i}{N} = \frac{\sum^{L} N_i \hat{p}_i}{N}$$

$$\hat{p}_{st} = \frac{1}{N} \sum^{L} N_i \cdot \hat{p}_i$$

- Varianza estimada (P)

$$\hat{Var}[\hat{p}_{st}] = \hat{Var}\left[\frac{1}{N} \sum^{L} N_i \hat{p}_i\right] = \sum^{L} \frac{N_i^2}{N^2} \cdot \hat{Var}[\hat{p}_i]$$

$$= \frac{1}{N^2} \cdot \hat{Var}[\hat{A}_{st}].$$

$$\hat{Var}[\hat{p}_{st}] = \sum^{L} \frac{N_i^2}{N^2} \cdot \frac{\hat{p}_i(1-\hat{p}_i)}{n_i - 1} \cdot \left(\frac{N_i - n_i}{N_i}\right)$$

- Intervalo de confianza

$\hat{p} \pm B_{\hat{p}_{st}}$, Donde $B_{\hat{p}_{st}}$ depende de n $\quad \frac{1}{N}(\hat{A}_{st} \pm B_{\hat{A}_{st}})$

Si $n > 30$

$$\boxed{B_{\hat{p}_{st}} = Z_{\frac{\alpha}{2}} \sqrt{\sum \frac{N_i^2}{N^2} \cdot \frac{\hat{p}_i(1-\hat{p}_i)}{n_i - 1} \left(\frac{N_i - n_i}{N_i}\right)}}$$

Entonces

$$\hat{p}_{st} \pm Z_{\frac{\alpha}{2}} \sqrt{\sum \frac{N_i^2}{N^2} \cdot \frac{\hat{p}_i(1-\hat{p}_i)}{n_i - 1} \left(\frac{N_i - n_i}{N_i}\right)}$$

//

Si $n \leq 30$

$$\boxed{B_{\hat{p}_{st}} = t_{\frac{\alpha}{2}, n-L} \sqrt{\sum^{L} \frac{N_i^2}{N^2} \frac{\hat{p}_i(1-\hat{p}_i)}{n_i - 1} \left(\frac{N_i - n_i}{N_i}\right)}}$$

Entonces

$$\hat{p}_{st} \pm t_{\frac{\alpha}{2}, n-L} \sqrt{\sum^{L} \frac{N_i^2}{N^2} \frac{\hat{p}_i(1-\hat{p}_i)}{n_i - 1} \left(\frac{N_i - n_i}{N_i}\right)}$$

Conocidos B y el nivel de confianza $1-\alpha$ fijados por el investigador para μ, se busca hallar cual debe ser el tamaño de muestra n, para ello partimos de la formula del limite del error de estimacion B

$$B = z_{\frac{\alpha}{2}} \sqrt{\hat{V}ar[\bar{y}_{st}]} \quad ; \text{ Entonces } \quad V[\bar{y}_{st}] = \frac{B^2}{z^2_{\frac{\alpha}{2}}}$$

La varianza disminuye conforme aumenta n

$$\frac{B^2}{z^2_{\frac{\alpha}{2}}} = \sum^{L} \left(\frac{N_i}{N}\right)^2 \frac{\sigma_i^2}{n_i} \cdot \frac{N_i - n_i}{N_i} \quad ; \quad \boxed{n_i = w_i \cdot n} \quad \ldots$$

$$n = \frac{\sum^{L} N_i^2 \, \sigma_i^2 / w_i}{D N^2 + \sum^{L} N_i \cdot \sigma_i^2} \quad , \quad D = \frac{B^2}{z^2_{\frac{\alpha}{2}}}$$

, Donde w_i depende de la afijacion

Para μ

$\sigma^2 = S^2$

o

$$\sigma_i = \frac{R_i}{6} = \frac{Y_{i\,max} - Y_{i\,min}}{6}$$

Para P

$\sigma^2 = \hat{P}(1-\hat{P})$

Para totales (τ, A), lo unico que cambia con respecto a (μ, P) es D

$$D = \frac{B^2}{z^2_{\frac{\alpha}{2}} \cdot N^2}$$

- Asignacion Optima con costos Variables

$$\boxed{C = C_0 + \sum^{L} C_i \, n_i}$$

C_i = diferentes,
σ_i = diferentes
$w_i \propto \sigma_i$
$w_i \propto 1/\sqrt{c_i}$
$w_i \propto N_i$

$$w_i = \frac{N_i \cdot \sigma_i / \sqrt{c_i}}{\sum N_k \cdot \sigma_k / \sqrt{c_k}}$$ > Estandarizar

- Asignacion de Neyman

$C_1 = C_2 = C_3 = \ldots = C_n =$

σ_i = diferentes

$$w_i = \frac{N_i \cdot \sigma_i}{\sum N_k \sigma_k}$$

- Asignacion Proporcional

$C_1 = C_2 = C_3 = \ldots = C_n$

σ_i = Iguales

$$w_i = \frac{N_i}{N}$$

REGRESION LINEAL

↳ Tecnica para investigar y modelar la relacion existente entre variables.

Y = Variable respuesta → Dependiente

X = Variable regresora → Independiente

Origen

↳ Sir Francis Galton (1889)

Relacion estatura de niños con estatura de padres

↳ • A padres Altos, los hijos generalmente lo son tambien.

• A padres bajos, los hijos son de menor estatura

• A padres muy altos o muy bajos se percibe una regresion hacia la estatura media de la poblacion

✱ El analisis de regresion permite estudiar las relaciones de asociacion entre Y y x

- ESTA ASOCIACIÓN NO ES RELACION CAUSA EFECTO
  ↳ NO IMPLICA CAUSALIDAD

- La relacion Causa - Efecto, solo aplica para diseño de experimentos

Y → x_1, x_2, x_3, x_k

x_1, x_2, x_3, x_k → No son aleatorios

→ Variables que pueden ser controladas por el investigador

↳ Regresoras, covariables

OTROS MODELOS DE REGRESION:

- Lineales generalizados
- Logistica (Sigmodeas)
- Prueba de Signos
- NO parametricas

• MODELO DE REGRESION LINEAL SIMPLE

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , i = 1, 2, \cdots n$$

Donde

n_i = # de datos en el que se supone que existe una relacion de tipo lineal

Y_i = Observacion

β_0 = Intercepto de la recta de regresion
Se interpreta si $0 \in [X_{min}, X_{max}]$

β_1: Pendiente de la recta de regresion (Cambio en la respuesta media por un cambio unitario en x)

ε_i = Error aleatorio del modelo

β_0, β_1 = Parametros

ε_i = Variable aleatoria

Para estimar los parametros (β_1, β_0) del modelo utilizamos dos metodos: (1) Maxima Verosimilitud (Maximum likelihood) y (2) Minimos Cuadrados

1) Maxima Verosimilitud = Maximiza la probabilidad de ocurrencia

↳ Requiere la verificacion sobre los residuos del modelo (Supuestos): → Resid (Modelo)

→ + Q-Q plot, P-P plot, histograma

- Distribuyen Normal (Shapiro-Wilk)
- Son independientes (Durbin-Watson)
- Tienen varianza Constante (Prueba F)

$E[\varepsilon_i] = 0$
$Var[\varepsilon_i] = \sigma^2$

$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ → Los residuos distribuyen normal (Independientes e Identicamente distribuidos)

$\varepsilon_i \sim NID(0, \sigma^2)$

↳ Consecuencias de los Supuestos

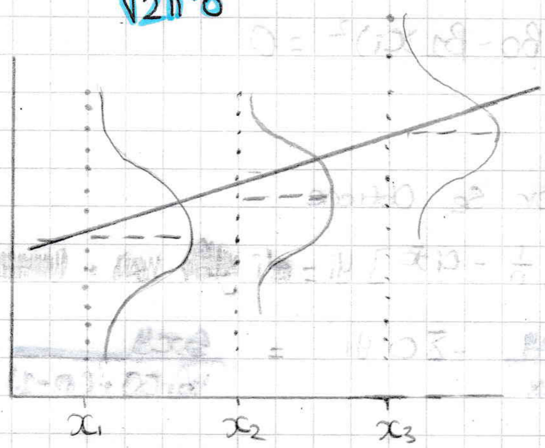
$y_i \sim NI(\beta_0 + \beta_1 x_i, \sigma^2)$ → - Distribucion Normal (No identicos)
- Independientes

↳ $E[y_i] = E[y_i | x_i] = \beta_0 + \beta_1 x_i$

$Var[y_i] = \sigma_i^2$

Fdp

↳ $F(y_i) = \dfrac{e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2}}{\sqrt{2\pi}\,\sigma}$



--- : $E[y_i|x_i]$

OTRAS PRUEBAS DE NORMALIDAD

- Kolmogorov-Smirnov
- Cramer-Von Mises
- Anderson-Darling

- FUNCION DE VEROSIMILITUD

$L(\beta_0, \beta_1, \sigma^2)$ = Verosimilitud

$$L = \prod_{i=1}^{n} F(y_i) = \prod_{i=1}^{n} \frac{e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2}}{\sqrt{2\pi}\,\sigma}$$

$\ell(\beta_0, \beta_1, \sigma^2) = $ Log Verosimilitud

$$\ell = Ln\, L(\beta_0, \beta_1, \sigma^2) = Ln\left[(2\pi\sigma^2)^{-n/2} * e^{-\frac{1}{2\sigma^2}\sum(y_i - \beta_0 - \beta_1 x_i)^2}\right]$$

$$\ell = -\frac{n}{2} Ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum(y_i - \beta_0 - \beta_1 x_i)^2$$

$$\ell = -\frac{n}{2} Ln(2\pi) - \frac{n}{2} Ln(\sigma^2) - \frac{1}{2\sigma^2}\sum(y_i - \beta_0 - \beta_1 x_i)^2$$

Para hallar los valores que maximizan la funcion derivamos e igualamos a cero.

$$\frac{\partial \ell}{\partial \beta_0} = \frac{1}{2\sigma^2}\sum 2(y_i - \beta_0 - \beta_1 x_i)(-1) = 0$$

$$\frac{\partial \ell}{\partial \beta_1} = \frac{1}{2\sigma^2}\sum 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{(\sigma^2)^{-2}\sum(y_i - \beta_0 - \beta_1 x_i)^2}{2} = 0$$

Resolviendo el sistema anterior se obtiene

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x} = \sum^{n}\left[\frac{1}{n} - c_i\bar{x}\right] y_i = d_i y_i$$

$$\tilde{\beta}_1 = \frac{\left(\sum^{n} x_i y_i\right) - (n\bar{x}\bar{y})}{\sum^{n} x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}} = \sum c_i y_i = \frac{S_{xy}}{\hat{Var}[x]\cdot(n-1)}$$

$$\tilde{\sigma}^2 = \frac{\sum^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2}{n}$$

Donde $c_i = \frac{x_i - \bar{x}}{S_{xx}}$; $d_i = \frac{1}{n} - c_i\bar{x}$

Los estimadores $\tilde{\beta}_0$ y $\tilde{\beta}_1$ son insesgados, esto se verifica conforme a los siguientes pasos:

a) $\sum c_i = \sum^n \frac{(x_i-\bar{x})}{Sxx} = \frac{1}{Sxx} \cdot \cancelto{0}{\sum^n (x_i-\bar{x})}$

b) $\sum c_i x_i = \sum^n \frac{x_i-\bar{x}}{Sxx}(x_i) = \frac{1}{Sxx} \cdot \cancelto{Sxx}{\sum^n (x_i-\bar{x})x_i}$

$$\boxed{\begin{array}{l} \sum c_i = 0 \\ \sum c_i x_i = 1 \end{array}}$$

1. Para $\tilde{\beta}_1$

$$E[\hat{\beta}_1] = E\left[\sum^n c_i y_i\right] = \sum^n E[c_i y_i] = \sum^n c_i E[y_i]$$

$$= \sum^n c_i (\beta_0 + \beta_1 x_i) = \beta_0 * \cancelto{0}{\sum^n c_i} + \beta_1 * \cancelto{1}{\sum^n c_i y_i}$$

$$= \beta_1 \checkmark$$

y_i independientes

$$\mathrm{Var}(\tilde{\beta}_1) = V\left[\sum^n c_i y_i\right] = \sum^n c_i^2 * \mathrm{Var}[y_i] = \sum^n \left(\frac{x_i-\bar{x}}{Sxx}\right)^2 * \sigma^2$$

$$= \frac{\cancelto{Sxx}{\sum^n (x_i-\bar{x})^2}}{(Sxx)^2} \cdot \sigma^2 = \frac{\sigma^2}{Sxx}$$

2. Para $\tilde{\beta}_0$

$$E[\tilde{\beta}_0] = E[\bar{y} - \tilde{\beta}_1 \bar{x}] = E\left[\frac{1}{n}\sum^n y_i - \hat{\beta}_1 \bar{x}\right] = \frac{1}{n}\sum^n E[y_i] - \bar{x}E[\hat{\beta}_1]$$

$$= \frac{1}{n}\sum^n (\beta_0 + \beta_1 x_i) - \bar{x}\beta_1 = \beta_0 + \cancel{\beta_1 \bar{x}} - \cancel{\bar{x}\beta_1} = \beta_0 \checkmark$$

$$\mathrm{Var}[\hat{\beta}_0] = V\left[\sum^n d_i y_i\right] = \sum^n d_i^2 * V[y_i] = \sum^n d_i^2 \sigma^2 = \sigma^2 \sum^n d_i^2$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{Sxx}\right)$$

El estimador $\tilde{\sigma}^2$ es sesgado para σ^2

$$E[\tilde{\sigma}^2] = \left(\frac{n-2}{n}\right)\sigma^2$$ Pero es posible construir un estimador insesgado de σ^2

$$\hat{\sigma}^2 = \left(\frac{n}{n-2}\right)\tilde{\sigma}^2 = \frac{\sum^n (y_i - \hat{\beta}_0 - \beta_1 x_i)^2}{n-2} = \frac{\sum^n e_i^2}{n-2} = \frac{\sum^n (y_i - \hat{y}_i)^2}{n-2}$$

e_i = Residuo

$\hat{\sigma}^2$ = MSE = Error Cuadratico Medio

$$MSE = \frac{1}{n-2}\left(S_{yy} - \hat{\beta}_1^2 * S_{xx}\right)$$

Dados los estimadores concluimos ademas

$$\hat{y}_i = E[y_i|x_i] = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

↳ que $\hat{y}_i$ es un estimador insesgado para y_i, esto debido al principio de invarianza

$$E[\hat{y}_i] = E[\hat{\beta}_0 + \hat{\beta}_1 x_i]$$
$$= E[\hat{\beta}_0] + E[\beta_1]x_i$$
$$= \beta_0 + \beta_1 x_i$$

En consecuencia

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx})$$
$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\right)$$

2) Minimos cuadrados = Minimizan la suma de los cuadrados de los errores

↳ No requiere supuestos

Funcion MC: $S(\beta_0, \beta_1) = \sum e_i^2 = \sum^n (y_i - \hat{y}_i)^2$
$$= \sum^n [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i]^2$$

Los estimadores de minimos cuadrados los obteremos de la siguiente manera

$$\frac{\partial s}{\partial \hat{\beta}_1} = \sum^n [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i]^2$$

$$= \sum^n 2[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i](-x_i)$$

$$= \sum^n 2[x_i y_i - \hat{\beta}_0 x_i - \beta_1 x_i^2]$$

Igualamos a cero

$$0 = \sum^n x_i y_i - \beta_0 \sum^n x_i - \hat{\beta}_1 x_i^2$$

$$\sum^n x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 x_i^2 \quad (1)$$

$$\bar{y} = \frac{\sum y_i}{n}$$

$$\bar{y} n = \sum y_i$$

$$\frac{\partial s}{\partial \hat{\beta}_0} = \sum^n 2[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i](-1)$$

Igualamos a cero

$$\sum y_i - \sum \hat{\beta}_0 - \sum \hat{\beta}_1 x_i = 0$$

$$\cancel{n}\bar{y} - \cancel{n}\hat{\beta}_0 - \cancel{n}\bar{x}\hat{\beta}_1 = 0$$

$$\bar{y} - \hat{\beta}_0 - \bar{x}\beta_1 = 0$$

$$\beta_0 = \bar{y} - \beta_1 \quad (2)$$

Remplazamos (2) en (1)

$$\sum^n x_i y_i = \bar{y} n \bar{x} - n\beta_1 \bar{x} + \beta_1 \sum x_i^2$$

$$\sum^n x_i y_i - n\bar{x}\bar{y} = \beta_1 [-n\bar{x} + \sum x_i^2]$$

$$\beta_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}}$$

Tras realizar el procedimiento notamos que los estimadores de Maxima Verosimilitud y de minimos cuadrados son iguales, sin embargo los EMV facilitad las pruebas de significancia y el calculo de intervalos de confianza y de prediccion

SIGNIFICANCIA DE LOS ESTIMADORES (β_0, β_1)

- Para β_1 (Pendiente): Por un aumento de 1 unidad en - - se estima que hay un incremento promedio de (β_1) en - - -

$$\beta_1 \sim N(\beta_1, \sigma^2/S_{xx})$$

β_{10} = Valor fijo

↳ Generalmente cero

↳ Sig. Regresion

Prueba de hipotesis

$H_0: \beta_1 = \beta_{10}$

$H_a: \beta_1 \neq \beta_{10}$

Estadistico de Prueba

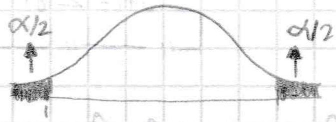
$$t_{01} = \frac{\beta_1 - \beta_{10}}{ee(\hat{\beta}_1)} \sim t_{n-2}$$

$$Var[\hat{\beta}_1] = \frac{MSE}{S_{xx}}$$

Donde $ee(\hat{\beta}_1) = \sqrt{\hat{Var}(\hat{\beta}_1)}$

Criterio de decision = Significancia de la regresion $[(t_{01})^2 = F_0]$

- $RR_1 = \{ t_{0,1} \mid |t_{01}| > t_{\frac{\alpha}{2}, n-2} \}$ // Rechazamos si t_{01} esta en la RR
- $P\text{-value} = P[t_{n-2} > |t_{01}|]$ // P-Value $< \alpha$ $\rightarrow$ Rechazamos Ho



Si Rechazo Ho la regresion es significativa de lo contrario las variables son independientes $(\beta_{10} = 0)$

- Para β_0 (Intercepto) = Es interpretable si $0 \in [X_{min}, X_{max}]$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\right)$$

β_{00} = Valor fijo → generalmente cero

Prueba de hipotesis

Ho: $\beta_0 = \beta_{00}$
H1: $\beta_0 \neq \beta_{00}$

Estadistico de prueba

$$t_{00} = \frac{\hat{\beta}_0 - \beta_{00}}{ee(\hat{\beta}_0)} \sim t_{n-2}$$

$$Var[\hat{\beta}_0] = MSE\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$$

Donde $ee(\hat{\beta}_0) = \sqrt{\hat{var}[\hat{\beta}_0]}$

Criterio de decision

- $RR_{\beta_0} = \{ t_{00} \mid |t_{00}| > t_{\frac{\alpha}{2}, n-2} \}$ // Rechazamos si t_{00} esta en la RR
- $P\text{-value} = P[t_{n-2} > |t_{00}|]$ // P-Value $< \alpha$ $\rightarrow$ Rechazamos Ho

INTERVALOS DE CONFIANZA DE LOS ESTIMADORES (β_0, β_1)

- Para β_1

$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}, n-2} * ee(\hat{\beta}_1)$$

- Si el IC no contiene al cero se dice que hay una relacion lineal en la cual la variable x es adecuada para predecir el comportamiento de y

- Para β_0

$$\hat{\beta}_0 \pm t_{\frac{\alpha}{2}} \, ee(\hat{\beta}_0)$$

- Si $x=0$ tiene sentido y el IC no contiene al cero entonces $\hat{E}[y \mid x=0] = \hat{\beta}_0$

- Si contiene al cero y $x=0$ tiene sentido entonces se recomienda adoptar un modelo de la forma

$$y = \beta_1 x + \varepsilon_i$$

ANALISIS DE VARIANZA (ANOVA)

Recordemos que evaluamos la Significancia de la regresion con una prueba de hipotesis para $\beta_1 = 0$, retomaremos este concepto para el analisis de Varianza. y evaluaremos de forma similar la Significancia de la regresion.

- Se analiza la variabilidad de (y), analizando las desviaciones de cada observacion y_i respecto de su media $\bar{y}$

$$(y_i - \bar{y})$$

La medida de la variacion total en y es la suma de los cuadrados de las desviaciones de los y_i's con respecto a su media muestral

$$S_{yy} = SST = \sum^{n} (y_i - \bar{y})^2 = \hat{V}ar[y] * (n-1) = \sum^{n} y_i^2 - n\bar{y}^2$$

$$\underbrace{\sum^{n} (y_i - \bar{y})^2}_{\substack{SST \\ (n-1)}} = \underbrace{\sum^{n} (\hat{y}_i - \bar{y})^2}_{\substack{SSR \\ (1)}} + \underbrace{\sum^{n} (y_i - \hat{y}_i)^2}_{\substack{SSE \\ (n-2)}}$$ → Identidad Suma de cuadrados

gl

Donde

SST = Suma de cuadrados total
SSR = Suma de cuadrados de la regresion
SSE = Suma de cuadrados del error

$$SSR = \hat{\beta}_1 S_{xy} = \hat{\beta}_1^2 S_{xx}$$ → Cantidad de variacion que alcanzo a explicar con el modelo RLS

$$SSE = SST - SSR = S_{yy} - \hat{\beta}_1^2 S_{xx}$$

Cada una de las sumas tiene asociados grados de libertad y con ellos construimos estimaciones independientes de σ^2, de la siguiente manera:

$$MSR = \frac{SSR}{1} \quad ; \quad MSE = \frac{SSE}{n-2}$$

Para realizar la Prueba de significancia de la regresion partiendo de ela anova utilizamos Ph para β_1; utilizando F_0 como estadistico de Prueb.

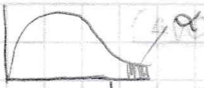
$H_0: \beta_1 = 0$
$H_a: \beta_1 \neq 0$

$$F_0 = \frac{MSR}{MSE} \sim F_{1, n-2}$$

$RR = \{ F_{cal} > F_{1-\alpha, 1, n-2} \}$

$P\text{-Value} = P[F_{1,n-2} > F_{cal}]$

- Tabla ANOVA



$F_0 = (t_{01})^2$

$\hookrightarrow qf(\,) \rightarrow qf(1-\alpha, 1, n-2)$

Fuente de Variacion	SS	gl	MS	F
Regresion o Modelo	SSR	1	$MSR = SSR/1$	$F_{cal} = \frac{MSR}{MSE}$
Error o Residual	SSE	n-2	$MSE = SSE/n-2$	$\sim F_{1,n-2}$
Total	SST	n-1		

Una Medida de bondad de ajuste es el coeficiente de determinacion denominado R^2, interpretado como la proporcion de variacion en la respuesta que alcanza a ser explicada por el modelo de regresion lineal

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = r^2$$

Donde

r: Coeficiente de correlacion de pearson

$\hookrightarrow \frac{S_{xy}}{S_x \cdot S_y}$

Consideraciones

- Pendiente cercana a cero
- Si R^2 tiende a cero la relacion entre x y y es muy pobre
- Si R^2 tiende a uno la recta ajustada se aproxima bien a los puntos
- Pendiente significativo

- R^2 grande no implica pendiente (β_1) grande
- R^2 grande no garantiza que el modelo de RLS ajustado sea el adecuado para los datos

RESPUESTA MEDIA

$\hookrightarrow$ Se permite solo interpolar es decir buscar un $x_0 \in [x_{min}, x_{max}]$ y esta dado por la expresion $\hat{y}_0 = \hat{E}[y \mid x = x_0] = \hat{\beta}_0 + \hat{\beta}_1 x_0$ para una estimacion puntual

Para la estimacion por intervalo tenemos $\hookrightarrow \hat{\beta}_0$ y $\hat{\beta}_1$ son LD

$$\hat{y}_0 \pm t_{\frac{\alpha}{2}, n-2}\, ee(\hat{y}_0)$$

Donde

$$ee(\hat{y}_0) = \sqrt{\hat{Var}(\hat{y}_0)}$$

$$\hat{Var}[\hat{y}_0] = MSE\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$$

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

$\hat{Var}[\hat{y}_0] = \hat{Var}[\sum d_i y_i + \sum c_i y_i x_0]$

$\hat{Var}[\hat{y}_0] = \hat{Var}[\sum (d_i + c_i x_0)\, y_i]$

$= \sum (d_i + c_i x_0)^2 * \hat{Var}[y_i]$

$= \sum \left[\left(\frac{1}{n} - \bar{x} c_i\right) + \left(\frac{x_i - \bar{x}}{S_{xx}}\right) x_0\right]^2 \hat{\sigma}^2$

$= MSE\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$

El intervalo de Longitud Minima se da cuando $x_0 = \bar{x}$

PREDICCIÓN

↳ Interesa una nueva observación y_0 que corresponda a un nivel específico de x, denominado x_0.
El estimador puntual está dado al igual que el anterior por la expresión $\hat{y}_0 = E[y \mid x = x_0] = \hat{\beta}_0 + \hat{\beta}_1 x_0$
↳ Solo se permite interpolar

Para la estimación del intervalo tenemos el error en la predicción estará dado por $(y_0 - \hat{y}_0)$ por lo que el intervalo quedará:

$$\hat{y}_0 \pm t_{\frac{\alpha}{2}, n-2} * ee(y_0 - \hat{y}_0)$$

Donde

$$ee(y_0 - \hat{y}_0) = \sqrt{\hat{Var}(y_0 - \hat{y}_0)}$$

$$\hat{Var}[y_0 - \hat{y}_0] = MSE \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

Son VA Independi

$$\hat{Var}[y_0 - \hat{y}_0] = \hat{Var}[y_0] + \hat{var}[\hat{y}_0]$$

$$= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

$$= \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

$$= MSE \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

VALIDACIÓN DE LOS SUPUESTOS

↳ Los supuestos son los ya nombrados y los hacemos sobre los residuales → Se hace apenas se ajusta el modelo

↳
- Distribución normal
- Media cero
- Varianza Constante
- Incorrelacionados

→ Ayuda a detectar si
- La función de regresión no es lineal
- Faltan variables predictoras al modelo

- **Normalidad**

1. Shapiro-Wilk : $W = \frac{(\sum a_i e_i)^2}{\sum (e_i - \bar{e})^2}$ Donde se concluye con el criterio del p-value

H_0: Distribuye Normal
H_a: NO Distribuye Normal

2. Q-Q Plot
3. P-P plot
4. Histograma

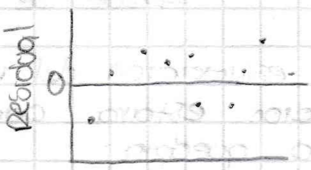
- **Media cero e Incorrelación**
↳ Usualmente no se valida por la forma en que se estiman los parámetros

Si se tiene el orden de registro (t_i) se grafica e_i vs t_i. SI NO HAY TENDENCIA SE CONCLUYE QUE NO HAY CORRELACION ENTRE LOS ERRORES DEL MODELO

$\bar{e} = \frac{1}{n} \sum e_i = 0$, lo cual se cumple si $\sum e_i = 0$

Lo que a su vez implica

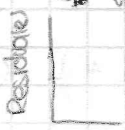
$$\sum^{n} y_i = \sum^{n} \hat{y}_i$$



Orden

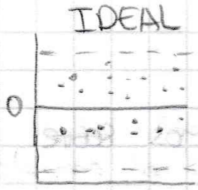
NO TENDENCIA → Incorrelacionados

- Varianza Constante
  - ↳ Verificacion grafica = NO TENDENCIA

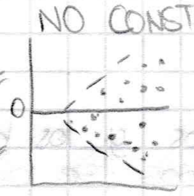


niveles (x)

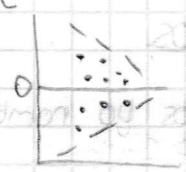
IDEAL



NO CONSTANTE

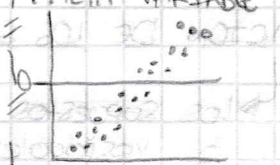


La Varianza va aumentando



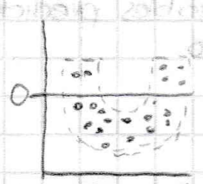
La varianza va disminuyendo

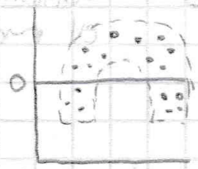
FALTA VARIABLE LINEAL



No Podemos decir si es constante o no

FALTA VARIABLE CUADRATICA





No podemos decir si es constante o no

Transformacion Box-cox cuando no es normal

PRUEBA DE CARENCIA DE AJUSTE

↳ Violacion del supuesto de linealidad

Consideraciones:

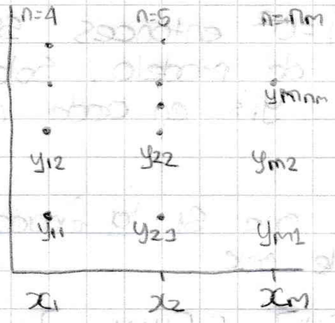
- Se realiza para la validacion del supuesto de linealidad
- Se hace despues de validar Normalidad, Independencia y σ^2 Constante
- Solo se puede realizar cuando se tiene al menos en un nivel de x dos o mas valores de y (Replicas)

$H_0: E[Y|X=x] = \beta_0 + \beta_1 x$
$H_a: E[Y|X=x] \neq \beta_0 + \beta_1 x$

H_0: El modelo de primer orden es apropiado
H_a: El modelo de primer orden no es apropiado

En esta prueba las réplicas son utilizadas para estimar σ^2 independiente del modelo ajustado

x	y				n_i
x_1	y_{11}	y_{12}	$\cdots$	y_{1n_1}	n_1
x_2	y_{21}	y_{22}	$\cdots$	y_{2n_2}	n_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_m	y_{m1}	y_{m2}		y_{mn_m}	n_m



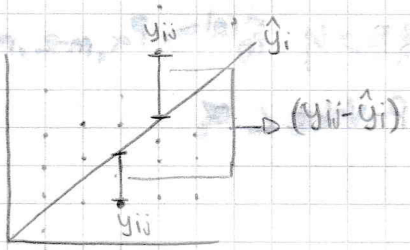
m = Distintos valores de la variable x
n_i = Observaciones de la variable y

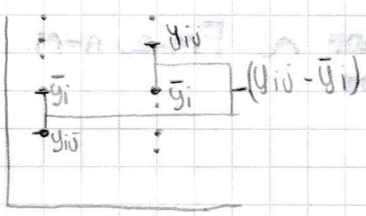
$\sum n_i$ = # total de observaciones
y_{ij} = j-ésima observación de la variable asociada al nivel i-ésimo de la variable regresora x.

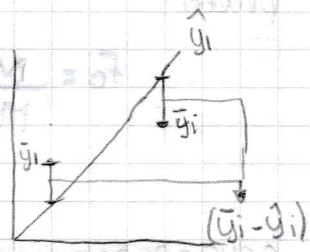
En la prueba de falta de ajuste descomponemos la Suma de Cuadrados del error (SSE)

$$\rightarrow \sum^{m}\sum^{n}(y_{ij}-\hat{y}_i)^2 = \sum^{m}\sum^{n}(y_{ij}-\bar{y}_i)^2 + \sum^{m}\sum^{n}(\bar{y}_i-\hat{y}_i)^2$$

SSE, gl $(n-2)$ | SS_{PE} $(n-m)$ | SS_{LOF} $(m-2)$







SS_{PE} = Suma de cuadrados asociada al error puro (Dentro)

SS_{LOF} = Suma de cuadrados asociado a la falta de ajuste (Ajuste)

MS_{PE} = Cuadrados Medios debido al error Puro
MS_{LOF} = Cuadrados Medios debido a la falta de Ajuste

$$MS_{PE} = \frac{SS_{PE}}{n-m}$$

$$MS_{LOF} = \frac{SS_{LOF}}{m-2}$$

$$E[MS_{PE}] = \sigma^2$$

$$E[MS_{LOF}] = \sigma^2 + \frac{\sum^{m} [E(y_i) - \beta_0 - \beta_1 x_i]^2}{m-2}$$

Si $V(\varepsilon_i) = \sigma^2$ entonces SS_{PE}, es un estimador de σ^2 independiente del modelo, solo se usa la variabilidad de los y_i's en cada nivel de x

Observemos que si la función verdadera es lineal, entonces: se cumple que

$$E[y_i] = E[y \mid X = x_i] = \beta_0 - \beta_1 x_i$$

Y por tanto se obtiene que la esperanza del MS_{LOF} es:

$$E[MS_{LOF}] = \sigma^2 + \sum^{m} [\underbrace{(\beta_0 - \beta_1 x_i) - (\beta_0 + \beta_1 x_i)}_{\to 0}]^2$$

$$E[MS_{LOF}] = \sigma^2$$

Si la función verdadera no es lineal, entonces $E[MS_{LOF}] > \sigma^2$.

Para la Ph. de falta de ajuste tenemos el estadístico de prueba:

$$F_0 = \frac{MS_{LOF}}{MS_{PE}} \sim F_{m-2,\, n-m}$$

$RR = \{ F_0^{cal} > F_{\alpha, m-2, m-n}$

$P\text{-Value} < \alpha$

CONSIDERACIONES

1. El SS_{PE} refleja la variación aleatoria o error experimental puro
2. El SS_{LOF} es una medida de la variación sistemática introducida por términos de distinción lineal o de primer orden
3. Si el modelo de RLS no se ajusta a los datos de manera apropiada, entonces la SSE estará inflada y producirá un estimador sesgado de σ^2 $[MSE = \hat{\sigma}^2]$

TRANSFORMACIONES : $y = \beta_0 + \beta_1 x + \varepsilon$

Usamos transformaciones cuando hay falta de ajuste o cuando no se cumplen los supuestos de normalidad o homogeneidad de varianza:

- Falta de ajuste
- No se cumple el supuesto de normalidad
- No se cumple el supuesto de homocedasticidad

} TRASFORMACION

Notas

- Tenga en cuenta siempre la transformacion a la que se sometieron los datos
- Al hacer inferencia se deben convertir los resultados a la escala original e interpretarlos en el contexto del problema

Las transformaciones las podemos aplicar en x, o en y, o en x y y. Al realizar la transformacion tenemos un modelo de la forma

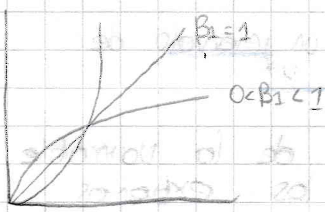
$$y^* = \beta_0^* + \beta_1^* x^* + \varepsilon^*$$

Donde * indica las variables transformadas

1. MODELO POTENCIA

$$Y = \beta_0 x^{\beta_1}$$

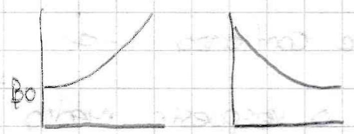
Aplicamos log de tal manera que obtenemos:



$$\text{Log}(y) = \text{Log}(\beta_0) + \beta_1 \text{Log}(x)$$

$$y^* = \text{Log}(\beta_0) + \beta_1 x^*$$

2. MODELO EXPONENCIAL



$$y = \beta_0 e^{\beta_1 x_i}$$

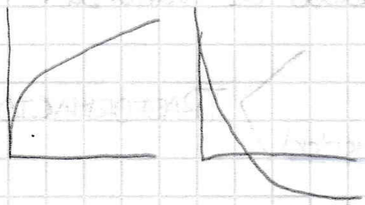
Aplicamos log de tal manera que obtenemos:

$$\text{Log}(y) = \text{Log}\,\beta_0 + \beta_1 x$$

$$y^* = \text{Log}\,\beta_0 + \beta_1 x$$

3. MODELO LOG EN x

$$y = \beta_0 + \beta_1 \log(x)$$

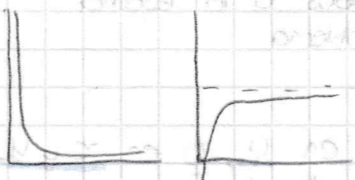


Aplicamos log sobre la variable x:

$$y = \beta_0 + \beta_1 x^*$$

4. MODELO RECIPROCO

$$y = \frac{x}{\beta_0 x - \beta_1}$$



Aplicamos el inverso de tal manera que obtenemos

$$\frac{1}{y} = \beta_0 - \beta_1 \cdot \frac{1}{x}$$

$$y^* = \beta_0 - \beta_1 x^*$$

CONSIDERACIONES DE LAS TRASFORMACIONES

La transformacion de forma general se expresa:

$$y^* = g(y) \quad \text{con } g \text{ invertible}$$

(u, v) es una expresion para un intervalo de la variable transformada y^*

Es decir, que para concluir en terminos de la variable original debemos invertir cada uno de los extremos

- Para intervalos de prediccion el valor de la respuesta futura y en el valor de $x = x_0$ es de la forma:

$$\left(g^{-1}(u) \,,\, g^{-1}(v) \right) \quad \text{con un nivel de confianza } 1 - \alpha$$

- Para intervalos de confianza el valor de la respuesta media $E[y \mid x_0]$ es de la forma

$$\left(K \cdot g^{-1}(u) \,,\, K \cdot g^{-1}(v) \right)$$

Donde K es el factor de correccion que depende de la transformacion aplicada

En el caso de $y^* = \log(y)$ el factor de correccion K es $[e^{MSE/2}]$

• MODELO DE REGRESION LINEAL MULTIPLE

Estos modelos con frecuencia se escriben de forma matricial.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times p} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{p\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

El MODELO DESCRIBE UN HIPERPLANO CON COEFICIENTES DE REGRESION $\beta_0, \beta_1, \beta_p$ y ES LINEAL POR SER UNA FUNCION LINEAL DE β

$$\underbrace{\overset{VA}{\mathbb{y}}}_{n\times 1} = \overbrace{\underbrace{X}_{n\times p}\ \underbrace{\beta}_{p\times 1}}^{FIJO} + \underbrace{\overset{VA}{\varepsilon}}_{n\times 1}$$

Donde

$\mathbb{y}$ = Vector Aleatorio (respuesta)
β = Vector de parametros
ε = Vector error aleatorio

$$\varepsilon \sim N_n(\underline{0}, \sigma^2 I_n)$$

Los errores distribuyen Normal n-variada, con vector de medias cero y varianza constante

$$\to Cov[\varepsilon] = \begin{bmatrix} Var[\varepsilon_1] & Cov[\varepsilon_1,\varepsilon_2] & \cdots & \\ \vdots & Var[\varepsilon_2] & & \\ & & \ddots & \\ Cov[\varepsilon_1,\varepsilon_2] & & & Var[\varepsilon_n] \end{bmatrix}$$

$$\to E[\varepsilon] = \begin{bmatrix} E[\varepsilon_1] \\ \vdots \\ E[\varepsilon_n] \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$Cov[\varepsilon] = \begin{bmatrix} \sigma^2 & & 0 \\ & \sigma^2 & \\ 0 & & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & & 0 \\ & 1 & \\ 0 & & 1 \end{bmatrix}$$

$Cov[X_1, X_2] = 0$
↳ NO implica independencia si incorrelacion
Independencia
↳ implica $Cov[X_1, X_2] = 0$

$$Cov[\varepsilon] = \sigma^2 I_n$$

La esperanza y la varianza del modelo estan dados por:

$$E[\mathbb{y}] = M_{\mathbb{y}} = \begin{bmatrix} E[y_1] \\ \vdots \\ E[y_n] \end{bmatrix} = \text{Vector de Medias}$$

$\boxed{P} + \boxed{P} + \boxed{\frac{P(P-1)}{2}}$

MEDIAS
Varianzas
Covarianzas

$$Var[\mathbb{y}] = \Sigma_{\mathbb{y}} = Cov[\mathbb{y}] = E[(\mathbb{y} - M_{\mathbb{y}})(\mathbb{y} - M_{\mathbb{y}})^T] = \begin{bmatrix} Var[y_1] & Cov[y_1,y_2] & \cdots & Cov[y_1,y_n] \\ Cov[y_2,y_1] & Var[y_2] & \cdots & Cov[y_2,y_n] \\ \vdots & \vdots & \ddots & \\ Cov[y_n,y_1] & & & Var[y_n] \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & & \sigma_{2n} \\ \vdots & & \ddots & \vdots \\ \sigma_{n1} & & & \sigma_n^2 \end{bmatrix}$$

Matriz Varianzas - Covarianzas

Matriz Simetrica y Cuadrada

La Matriz de Varianzas-Covarianzas asociada a un vector aleatorio es una matriz cuadrada y simetrica de orden igual al tamaño del vector aleatorio donde la diagonal principal contiene las varianzas asociadas a cada elemento del vector y por fuera de la diagonal estan las covarianzas entre pares de elementos del vector aleatorio.

PROPIEDADES DEL VALOR ESPERADO Y LA VARIANZA DE UN VECTOR ALEATORIO

i. $E[A\,y] = A\mu_y$

ii. $E[A\,y + b] = A\mu_y + b$

iii. $V[A\,y] = A\Sigma_y A'$

iv. $V[A\,y + b] = A\Sigma_y A'$

μ_y = vector de medias

Σ_y = Matriz var-cov

A = Matriz $m \times n$

b = Vector $m \times 1$ (constantes)

DEFINICIONES BASICAS DE TEORIA MATRICIAL

i. $(BA)^T = A^T B^T$

ii. A es simetrica si $A^T = A$
A es idempotente si $AA = A$

iii. Si A es simetrica e idempotente entonces $(I - A)$ tambien lo es

iv. **FORMA CUADRATICA**: La funcion $x^T A x = \sum \sum a_{ij} x_i x_j$ es la forma cuadratica de x con a_{ij} la ij-esima componente de A

v. Si $x^T A x > 0, \forall x$, A es definida positiva
Si $x^T A x \geq 0, \forall x$, A es definida semipositiva

A = Matriz constante $(n \times n)$

B = Matriz constante $(m \times n)$

x = vector de variables $(n \times 1)$

a = vector de constantes $(n \times 1)$

I = Matriz identidad (n)

MODELOS MULTIRESPUESTA

DERIVADAS VECTORIALES O MATRICIALES

a) $\dfrac{\partial (a^T x)}{\partial x} = a$

b) $\dfrac{\partial (x^T x)}{\partial x} = 2x$

c) $\dfrac{\partial (x^T A x)}{\partial x} = Ax + A^T x$. Si A es simetrica el resultado es: $2Ax$

a = Vector de constantes $(n \times 1)$

A = Matriz de constantes $(n \times n)$

x = Vector de variables $(n \times 1)$

RESULTADOS DISTRIBUCIONALES PARA VECTORES ALEATORIOS

Sea $\mathbb{Y}$ un vector aleatorio normal n-variado con media μ_y y matriz no singular de Var-Cov Σ_y, es decir

$$\mathbb{Y} \sim N_n(\mu_y, \Sigma_y)$$

Sea A una matriz nxn de constantes y sea U una forma cuadrática de $\mathbb{Y}$ definida por

$$U = \mathbb{Y}^T A \mathbb{Y}$$

Entonces:

i. Si $A\Sigma_y$ o $\Sigma_y A$ es una matriz idempotente de rango p, entonces

$U \sim \chi^2_{p,\lambda}$ donde $\lambda = \mu^T A \mu$ es el parametro de no centralidad de la distribucion chi-cuadrado

ii. Sea $\Sigma_y = \sigma^2 I$, lo cual es una suposicion tipica, si A es idempotente y de rango p entonces

$$\frac{U}{\sigma^2} \sim \chi^2_{p,\lambda} \quad \text{donde } \lambda = \mu^T A \mu / \sigma^2$$

iii Sea B una matriz mxn y W la forma lineal definida por: $W = B\mathbb{Y}$, entonces la forma cuadrática $U = \mathbb{Y}^T A \mathbb{Y}$ son independientes entre si

$$B\Sigma_y A = 0 \qquad 0 = \text{Matriz nula } (m \times n)$$

iv. Sea B una matriz nxn y sea $V = \mathbb{Y}^T B \mathbb{Y}$, entonces las dos formas cuadráticas U y V son independientes entre si

$$A \Sigma_y B = 0$$

NOTA: Si $\Sigma_y = \sigma^2 I$, entonces U y V son independientes si $AB = 0$

CONSIDERACIONES MRLM

1. Se dice lineal ya que la parte deterministica es una funcion lineal de los parametros desconocidos ($\beta_0, \beta_1 \ldots \beta_p$) denominados coeficientes de regresion
2. Describe un hiperplano en el espacio de K-dimensiones de las variables regresoras x's

3. B_j: Representa el cambio esperado en la variable respuesta y, por un cambio unitario en x_j, cuando las demás variables regresoras x_k's (con $k \neq j$) se mantienen constantes

---

ENFOQUE MATRICIAL DEL MODELO RLS

$$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_3$$
$$\vdots$$
$$y_n = \beta_0 + \beta_1 x_n + \varepsilon_n$$

$$\mathbb{y} = \underset{n\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}} \qquad \mathbb{B} = \underset{2\times 1}{\begin{bmatrix} B_0 \\ B_1 \end{bmatrix}} \qquad \mathcal{E} = \underset{n\times 1}{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}$$

$$\mathcal{E} \sim N_n(\underset{n\times 1}{\underline{0}}, \underset{n\times n}{\sigma^2 I_n})$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} B_0 + B_1 x_1 \\ B_0 + B_1 x_2 \\ \vdots \\ B_0 + B_1 x_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \Leftrightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_2 \end{bmatrix} \begin{bmatrix} B_0 \\ B_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\underset{n\times 1}{\mathbb{y}} = \underset{n\times 2}{\mathbb{X}} \underset{2\times 1}{\mathbb{B}} + \underset{n\times 1}{\mathcal{E}}$$

AHORA

$$\hat{\mathbb{B}} = \begin{bmatrix} \hat{B}_0 \\ \hat{B}_1 \end{bmatrix} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{y}$$

$$\bar{x} = \frac{\sum x_i}{n}$$

$$S_{xx} = \sum x_i^2 - n\bar{x}^2$$

$$\mathbb{X}^T \mathbb{X} = \underset{2\times n}{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_2 \end{bmatrix}} \underset{n\times 2}{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

$$(\mathbb{X}^T \mathbb{X})^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

$$\mathbb{X}^T \cdot \mathbb{y} = \underset{2\times n}{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}} \cdot \underset{n\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

Norma

$$\hat{\beta} = \begin{bmatrix} \hat{\beta_0} \\ \hat{\beta_1} \end{bmatrix} = \frac{1}{n\sum x_i^2 - n^2\bar{x}^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta_0} \\ \hat{\beta_1} \end{bmatrix} = \frac{1}{n[\sum x_i^2 - n\bar{x}^2]} \begin{bmatrix} (\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i) \\ (-\sum x_i)(\sum y_i) + (\sum x_i y_i)(n) \end{bmatrix}$$

$$\hat{\beta}_1 = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n[\sum x_i^2 - n\bar{x}^2]} = \frac{\cancel{n}\sum x_i y_i - \cancel{n}\bar{x}\, n\bar{y}}{\cancel{n}[\sum x_i^2 - n\bar{x}^2]}$$

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

$$\hat{\beta}_0 = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{n[\sum x_i^2 - n\bar{x}^2]} = \frac{\cancel{n}\bar{y}\cdot\sum x_i^2}{\cancel{n}\, S_{xx}} - \frac{\cancel{n}\bar{x}\ \sum x_i y_i}{\cancel{n}\, S_{xx}} \ \ldots$$

$$\hat{\beta}_0 = \bar{y} - \beta_1\bar{x}$$

Como mencionamos anteriormente el MRLM se considera lineal siempre y cuando este en funcion lineal de los parametros, por lo que es valido encontrar modelos de la forma

1. Polinomica de tercer orden

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

POLINOMICA NO RECOMENDADA NO ES PARCIMONIOSA

Si hacemos $x_1 = x$, $x_2 = x^2$ y $x_3 = x^3$ tenemos que

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \rightarrow$ EL CUAL ES UN MRLM

2. MODELO CON INTERACCION

EL AUMENTO DE VARIABLES DISMINUYE LA PRECISION EN LA ESTIMACION DEL MSE

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2 + \varepsilon$$

Si hacemos $x_1 = x_1$, $x_2 = x_2$, $x_3 = x_1 x_2$ tenemos que

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \rightarrow$ EL CUAL ES UN MRLM

Para estimar los parametros β del MRLM utilizamos dos metodos: (1) Maxima Verosimilitud y (2) Minimos Cuadrados los cuales resultan en los mismos estimadores.

1) Minimos Cuadrados: Minimiza la suma de cuadrados de los errores

$$S(\beta) = S(\beta_0, \beta_1, \dots \beta_p) = \sum^{n} \varepsilon_i^2 = \varepsilon^T \varepsilon$$
$$= (y - X\beta)^T (y - X\beta)$$
$$= y^T y - \beta^T X^T y - y^T X \beta + \beta^T X^T X \beta$$

Para obtener una estimacion de β derivamos vectorialmente respecto a β y obtenemos:

$$\left.\frac{\partial S(\beta)}{\partial \beta}\right|_{\hat{\beta}} = -2X^T y + 2X^T X \hat{\beta}$$

Igualando a cero obtenemos

$$X^T X \hat{\beta} = X^T y$$ → ECUACIONES NORMALES

Si $X^T X$ es invertible o definida positiva entonces podemos hallar β

Y por tanto EL VECTOR estimado es

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

NOTA: Otra forma de hallar los estimadores β_p's es utilizando la funcion $S(\beta_0, \beta_1, \dots \beta_p)$ y derivar con respecto a cada parametro β_p

$$S(\beta_0, \beta_1, \dots \beta_p) = \sum^{n} \varepsilon_i^2$$
$$= \sum^{n} (y_i - E(y_i))^2$$
$$= \sum^{n} \left(y_i - \beta_0 - \sum^{p} \beta_e x_{ij}\right)^2$$

Cuando hay problemas de multicolinealidad REGRESION RIDGE

$(X^T X)^{-1}$ + K veces identidad

CONSIDERACIONES EMC → Estimadores de mínimos cuadrados

1. $(X^T X)^{-1}$ = Existe, siempre que los regresores X_j's sean LI

2. Los elementos de la diagonal de $X^T X$ son las sumas de cuadrados de los elementos de las columnas de X

3. Los elementos fuera de la diagonal de $X^T X$ son las sumas de los productos cruzados de los elementos de las columnas de X

4. Los elementos de $X^T y$ son las sumas de los productos cruzados de los elementos de las columnas de X con los elementos de y

INTERPRETACION $\hat{\beta}_j$'s

El valor de $\hat{\beta}_j$ con $j=1,2,\dots,K$ se interpreta como el cambio esperado (o efecto parcial) en la respuesta promedio debido a un incremento unitario en x_j, cuando las demás predictoras se mantienen fijas.

El valor de $\hat{\beta}_0$ solo tiene sentido práctico interpretativo si $x_j = 0$ $\in [\min\{x_{ij}\}, \max\{x_{ij}\}]$ $\forall j = 1,2,\dots K$ y en ese caso se interpreta como la respuesta media cuando $x_j=0$, $\forall j=1,2,\dots,K$.

PROPIEDADES de $\hat{\beta}$

i $\hat{\beta} = (X^T X)^{-1} X^T y$ es insesgado para β, es decir $E[\hat{\beta}] = \beta$

$$
\begin{aligned}
E[\hat{\beta}] &= E[(X^T X)^{-1} X^T y] \\
&= (X^T X)^{-1} X^T E[y] \\
&= (X^T X)^{-1} X^T E[X\beta + \varepsilon] \\
&= (X^T X)^{-1} X^T [X\beta + E[\varepsilon]] \quad E[\varepsilon] \to 0 \\
&= (X^T X)^{-1} X^T X \beta \\
&= I\beta = \beta
\end{aligned}
$$

ii $\hat{\beta}$ es el mejor estimador lineal-insesgado de β, en el sentido de que $\hat{\beta} = (X^T X)^{-1} X^T y$ tiene mínima varianza entre todos los estimadores insesgados de β

$$\text{cov}(\hat{\beta}) = \text{cov}\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \mathbb{y}\right)$$

$$= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}\ \text{cov}(\mathbb{y})\left[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\right]^T$$

$$\text{cov}[\mathbb{y}] = \text{cov}(\mathcal{E}) = \sigma^2 I$$

$$= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\sigma^2 I\ \mathbb{X}(\mathbb{X}\mathbb{X})^{-1}$$

$$= \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$$

$$= \sigma^2[c_{ij}]$$

iii La matriz de Varianzas - Covarianzas de $\hat{\beta}$ esta dada por

$$\text{cov}(\hat{\beta}) = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1} = \sigma^2[c_{ij}]$$

En dende los c_{ij} son los elementos de la matriz $(\mathbb{X}^T\mathbb{X})^{-1}$

$$C = (\mathbb{X}^T\mathbb{X})^{-1} = \begin{bmatrix} C_{00} & C_{01} & \cdots & C_{0K} \\ C_{01} & C_{11} & \cdots & C_{1K} \\ \vdots & \vdots & \ddots & \\ C_{0K} & C_{1K} & \cdots & C_{KK} \end{bmatrix}$$

$$\text{cov}(\hat{\beta}) = \sigma^2 C$$ Donde $V(\hat{\beta}_j) = \sigma^2 C_{jj}$ y

$$\text{cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 c_{ij}$$

$$\text{Var}[\hat{\beta}] = \text{Var}\left[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{y}\right]$$

$$= \text{Var}[A\mathbb{y}] \quad \text{con } A = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$$

$$= A\,\text{var}[\mathbb{y}]\,A^T$$

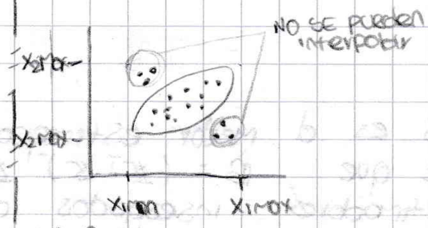
$$= A\,\sigma^2 I\,A^T$$

$$= \sigma^2 AA^T$$

$$= \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}$$

$$= \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$$

PARA INTERPOLAL



NO BASTA CON QUE ESTE EN EL RANGO DE x_1, x_2 SINO QUE ESTE EN LA NUBE

Norma

MODELO DE REGRESION LINEAL MULTIPLE (SEGUNDA PARTE)

En general observamos un modelo del tipo

$$\hookrightarrow y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} + \varepsilon_i$$

Donde

$$\varepsilon_i \overset{iid}{\sim} (0, \sigma^2), \quad i = 1, 2, 3, \ldots, n$$

Matricialmente lo escribimos

$$\hookrightarrow \mathbb{y} = \mathbb{X}\beta + \varepsilon$$

Donde

$$\mathbb{X} = \begin{bmatrix} 1 & 1 & 1 & & 1 \\ \vdots & x_1 & x_2 & \cdots & x_K \\ 1 & 1 & 1 & & 1 \end{bmatrix}_{n\times p}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{p\times 1} ; \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

Ecuaciones Normales

$$\mathbb{X}^T \mathbb{X} \hat{\beta} = \mathbb{X}^T \mathbb{y}$$

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

P = # de parametro = $K+1$

K = # de predictores

- $E[\hat{\beta}] = \beta$
- $\mathrm{cov}[\beta] = \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1}$

ECUACIONES NORMALES

$$\mathbb{X}^T \mathbb{X} \hat{\beta} = \mathbb{X}^T \mathbb{y}$$

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

CURVA DE REGRESION AJUSTADA

$$\begin{aligned} \hat{y}_i &= \hat{E}[y \mid x_{i1}, x_{i2}, \ldots x_{iK}] \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \ldots + \hat{\beta}_K x_{iK} \\ &= [1 \;\; x_{i1} \ldots x_{iK}]\, \hat{\beta} \\ &= x_{0i}^T \hat{\beta} \end{aligned}$$

$$\begin{aligned} \hat{\mathbb{y}} &= \hat{E}[\mathbb{y} \mid \mathbb{X}] \\ &= \mathbb{X}\beta \\ &= \mathbb{X}(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{y} = \mathbb{H}\mathbb{y} \end{aligned}$$

Propiedades de $\mathbb{H}$
(Matriz hat o sombrero)

- Es simetrica
- Es idempotente $\mathbb{H}\mathbb{H} = \mathbb{H}$
- Es cuadrada $(n\times n)$

$$\hat{\mathbb{y}} = \mathbb{H}\mathbb{y}$$

Residuos

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

$$\mathbb{e} = \mathbb{y} - \hat{\mathbb{y}}$$

$$\mathbb{e} = \mathbb{y} - \mathbb{H}\mathbb{y}$$

$$\mathbb{e} = \underbrace{(\mathbb{I} - \mathbb{H})_{n\times n}}_{} \mathbb{y}_{n\times 1}$$

↳ simetrica
idempotente

DISTRIBUCION DE $\mathbb{e}$

$$E[\mathbb{e}] = E[(\mathbb{I}-\mathbb{H})\mathbb{y}]$$
$$= (\mathbb{I}-\mathbb{H})E[\mathbb{y}]$$
$$= (\mathbb{I}-\mathbb{H})\mathbb{X}\beta$$
$$= \mathbb{X}\beta - \mathbb{H}\mathbb{X}\beta \quad \rightarrow \mathbb{H}\mathbb{X} = \mathbb{X}$$
$$= \mathbb{X}\beta - \mathbb{X}\beta$$
$$= \mathbb{0} \rightarrow (E[\hat{y}_i] = E[y_i])$$

$$cov[\mathbb{e}] = Cov[(\mathbb{I}-\mathbb{H})\mathbb{y}]$$
$$= (\mathbb{I}-\mathbb{H})\,cov[\mathbb{y}]\,(\mathbb{I}-\mathbb{H})$$
$$= (\mathbb{I}-\mathbb{H})\,\sigma^2\mathbb{I}\,(\mathbb{I}-\mathbb{H})$$
$$= \sigma^2(\mathbb{I}-\mathbb{H})(\mathbb{I}-\mathbb{H})$$
$$= \sigma^2(\mathbb{I}-\mathbb{H})$$

Los residuos son correlacionados
→ hay que estandarizarlos para ver tendencia

$$\mathbb{e} \sim N_n(\mathbb{0}, \sigma^2(\mathbb{I}-\mathbb{H}))$$ → Los residuos distribuyen normal n variado, con vector de medias cero y varianza constante

$$Var[e_i] = \sigma^2(1-h_{ii})$$

h_{ii}: Diagonal Matriz Sombrero

$$Cov[e_i, e_j] = \boxed{\sigma^2(1-h_{ij})} \rightarrow \neq 0$$ → Los e_i son correlacionados

$i \neq j$

$$Cov(\mathbb{e}) = \sigma^2(\mathbb{I}-\mathbb{H})$$

EN R se calcula la inversa generalizada vía Métodos Numéricos

VERIFICACION DE SUPUESTOS

i) Distribucion Normal de los Errores $[e_i = y_i - \hat{y}_i]$

$H_0: \varepsilon_i \sim N$
$H_a: \varepsilon_i \nsim N$

> Utilizamos la prueba de Shapiro Wilk y concluimos con p-valor

↓

Estandarizamos los residuos

↓

$$r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

La prueba F no es significativa si no se cumple la Normalidad de los residuos

No olvidemos que los $\{e_i\}$ son correlacionados

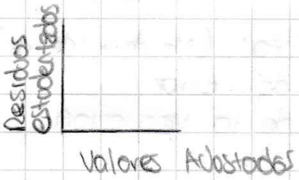
Donde h_{ii} es la diagonal de $\mathbb{H}$

ii) Varianza Constante $Var[\varepsilon_i] = \sigma^2$

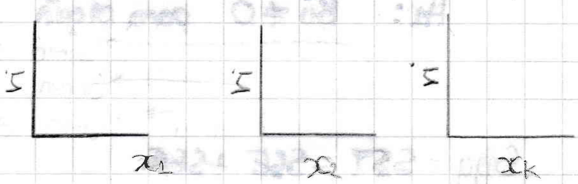
(Homogenidad de varianzas)

"Cuando el modelo no tiene falta de ajuste se llama adecuado"

* Contrastamos e_i vs $\hat{y}_i$

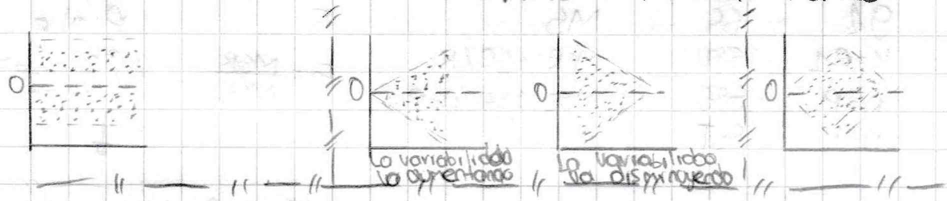


* Contrastamos e_i vs x_j, $j=1,2,3 \dots K$

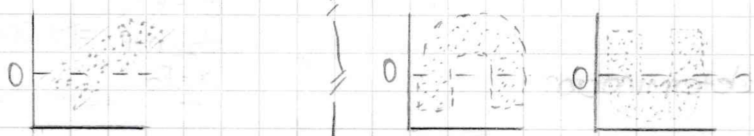


Estos graficos tambien sirven para detectar "Outliers" y nos dicen si necesitamos transformaciones para la respuesta ($\hat{y}_i$), sobre una variable regresora (x_j) o sobre varios variables regresoras.

TENDENCIA IDEAL | NO CONSTANTE (PARLANTE) | Rombo



FALTA VARIABLE LINEAL | FALTA VARIABLE CUADRATICA

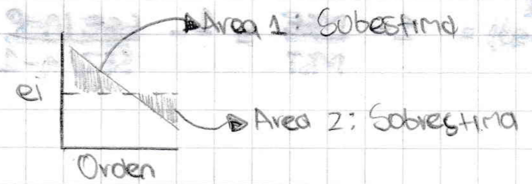


Con estos graficos tambien podemos evidenciar lo linealidad del modelo

iii) Incorrelacion de los e_i

Solo se realiza cuando se tiene el orden en el que se obtuvo cada dato (observacion)

IDEAL



↳ Podemos utilizar (row.names)

Hay dos opciones en este caso ⇒ ① Incluir correlacion en el modelo ② Repetir la toma de datos

PRUEBA DE SIGNIFICANCIA DE LA REGRESION

$H_0: \beta_1 = \beta_2 = \dots = \beta_K = 0$ → Rechazo H0: Al menos uno de los predictores es significativo para explicar la variabilidad.

$H_1: \beta_j \neq 0$ para algún $j = 1,2,3,\dots,K$

$$S_{yy} = SST = SSE + SSR$$

- SST → Suma de cuadrados Total (No depende de la regresion)
- SSE → Suma de cuadrados del Error
- SSR → Suma de cuadrados de la regresion

Grados de libertad: SST → $n-1$; SSE → $n-K-1 = n-p$; SSR → $K = p-1$

Un estimador insesgado de σ^2 es

$$\sigma^2 = MSE = \frac{SSE}{n-p}$$

- Tabla anova (Funcion miAnova) → Funcion creada para la clase

	gl	SS	MS	F	P-value
Regresion	$K=p-1$	SSR	$MSR = SSR/K$	$F_0 = \frac{MSR}{MSE}$	$P(F_{K,n-p} > F_0)$
Error	$n-p$	SSE	$MSE = SSE/n-p$		
Total	$n-1$	SST			

P-value → App Probability Distributions:
$d_1 = K$
$d_2 = n-p$ ⇒ $P(X > x) = $ P-value
$x = F_0$

$$F_0 = \frac{MSR}{MSE} \sim F_{K,n-p}$$

- Coeficiente de determinacion

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$ → Proporcion de varianza explicada por el modelo, en el caso de RLM el R^2 aumenta a medida que aumenta el numero de variables regresoras

$$R^2_{adj} = 1 - \frac{MSE}{MST} = 1 - \frac{SSE/n-p}{SST/n-1}$$ → No es proporcion de variacion explicada, es un criterio de seleccion de modelo

PRUEBA DE SIGNIFICANCIA DE CADA PARAMETRO (T_{0j})

$H_0: B_j = 0$
$B_j \neq 0$

$$t_{0j} = \frac{\hat{B}_j - B_j}{ee(\hat{B}_j)}$$ (B_j → 0)

IC: $B_j \pm t_{n-p, \frac{\alpha}{2}} \cdot ee(\hat{B}_j)$

→ $\boxed{\text{Summary (modelo)}}$

P-value $= 2P(|t_{n-p}| > |t_{0}|)$

$Var[B] = \begin{bmatrix} C_{11} & & \\ & C_{22} & \\ & & C_{jj} \end{bmatrix}$

C_{jj} → Componente (j,j) de la matriz C

Donde

$$ee(\hat{B}_j) = \sqrt{Var[\hat{B}_j]} = \sqrt{\sigma^2 \cdot C_{jj}} = \sqrt{MSE \cdot C_{jj}}$$

$\beta_0 =$ ↗ NO significativo, NO interpretable | ↗ significativo, interpretable | ↗ significativo, NO interpretable ! $0 \in [X_{j\,min}, X_{j\,max}] \ \forall j = 1, \ldots, k$ si uno solo no cumple ya no es interpretable

La interpretacion de la prueba la expresamos de la siguiente manera

Por cada unidad de incremento de la variable regresora (X_k), se espera que la cantidad de la variable explicada, en promedio, se ____*____ de manera significativa, en el (valor del parametro), cuando las demas variables regresoras se mantienen constantes.

En general un parametro es interpretable solo si es significativo

* Si el signo del parametro es positivo = INCREMENTE
Si el signo del parametro es negativo = DISMINUYA

"Por un aumento unitario en X_j se estima hay un cambio promedio de β_j en la respuesta cuando los demas pred estan constes"

INFERENCIA DE UN SUBCONJUNTO DE PARAMETROS

Lo utilizamos para probar la significancia de un subconjunto de parametros o para ver si el efecto de una variable es igual al de otra variable.

Para la construccion del estadistico de prueba se utiliza la metodologia de sumas de cuadrados extra, la cual cuantifica el incremento o disminucion en la SSR o SSE al incluir una o varias variables al modelo de regresion.

EN GENERAL

"$\boxed{\begin{array}{l} H_0 : \beta_2 = \emptyset \\ H_a : \beta_2 \neq \emptyset \end{array}}$"

Donde β_2 contiene los componentes de β a los que se desea analizar la no significancia en el modelo de regresion particionando β

Por tanto, en esta situacion comparamos el modelo completo (FM) contra un modelo reducido (RM)

β_2 = Contiene las condiciones de interes

β_1 = contiene β_0 y las demas componentes de β que no estan en β_2

$$\begin{aligned}
SS_{extra} &= SSR(\beta_2 \mid \beta_1) \\
&= SSR(\beta_1, \beta_2) - SSR(\beta_1) \\
&= SSR(FM) - SSR(RM) \\
&= [SST - SSE(FM)] - [SST - SSE(RM)] \\
&= SSE(RM) - SSE(FM) \\
&= MSE(RM) * (n - P_{RM}) - MSE * (n - P_{FM})
\end{aligned}$$

Estadistico de prueba

$$F_{parcial} = \frac{SSR(\beta_2 | \beta_1)/M}{MSE(\beta_1, \beta_2)} \sim F_{M,\ n-P_{FM}}$$

Donde:

$$M = \dim(\beta_2)$$

La suma de cuadrados de la regresion se descompone como sumas de cuadrados extra

$$SSR(\beta) = \underbrace{SSR_E(\beta_1 | \beta_0, \beta_2 \ldots \beta_K)}_{1\ gl} + \underbrace{SSR_E(\beta_2 | \beta_0, \beta_1, \beta_3, \ldots \beta_K)}_{1\ gl} + \ldots + \underbrace{SSR_E(\beta_K | \beta_0, \beta_2, \beta_2, \ldots \beta_{K-1})}_{1\ gl}$$

Veremos dos tipos de sumas de cuadrados

*) Suma de cuadrados tipo I (SS_1) No muy util

Es una suma de cuadrados secuencial de cada variable, dado que en el modelo estan presentes las variables anteriores

En R anova(modelo)

*) Suma de cuadrados tipo II (SS_2)

Es una suma de cuadrados parcialmente secuencial de cada variable, dado que en el modelo estan presentes las demas variables

En R Anova(modelo)

Podemos probar la significancia de un parametro con SSextra, de forma homologa como lo haciamos con la prueba T. y se llega a la misma conclusion.

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

El modelo reducido correspondiente seria

$$y_i = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots \beta_K x_{iK}$$

Estadistico de prueba

$$F_0 = \frac{[SSE(RM) - SSE(FM)]/1}{MSE(FM)} \sim F_{1,\ n-P_{FM}}$$

Los grados de libertad del numerador en este caso son el # de parametros evaluados (β_1)

- Significancia de un subconjunto de parametros

Buscamos probar la no significancia de un subconjunto de parametros, en general la hipotesis que subyace es la siguiente

Partiendo de un modelo de 6 variables regresoras podemos probar la significancia de digamos 3 de ellos al mismo tiempo.

$H_0: \beta_4 = \beta_5 = \beta_6 = 0$

$H_a:$ Al menos un $\beta_j \neq 0$

Como se ve la prueba es similar a la Ph de significancia de la regresion y concluimos de la misma manera

↓

Si Rechaza Ho

↳ Al menos uno de los β_j en estudio es significativo para explicar la Variabilidad

Estadistico de prueba

$$F_0 = \frac{[SSE(RM) - SSE(FM)]/r}{MSE(FM)} \sim F_{r, n-P_{FM}}$$

Donde r es el numero de parametros que estamos evaluando

↳ Para este caso $r=3$

$RR \{ F_0 / F_0 > F_{\alpha, r, n-P_{FM}} \}$

$V_p = P(F_{r, n-P_{FM}} > F_0)$

El modelo reducido (RM) se ve de la siguiente forma

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon$

$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

- Prueba Lineal General

Con esta prueba verificamos si el efecto de una variable es igual al de otra.

Volviendo al modelo con 6 regresoras analicemos

$H_0: \beta_1 = \beta_2 \wedge \beta_5 = \beta_6$

↳ Esta escritura dificulta el juego de hipotesis, entonces se escribe de forma matricial

$\beta_1 - \beta_2 = 0$

$\beta_5 - \beta_6 = 0$

↓

$$\underbrace{\begin{bmatrix} 0 & 1 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix}}_{\beta} = \underbrace{\begin{bmatrix} 0 \\ 0 \end{bmatrix}}_{\emptyset}$$

De forma Matricial las hipotesis quedan de la forma

$H_0: \quad \Pi \beta = \emptyset$

$H_a \quad \Pi \beta \neq \emptyset$

Forma general

El Modelo reducido (RM) se ve de la Siguiente forma

$y_i = \beta_0 + \beta_1 (x_{i1} + x_{i2}) + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 (x_{i5} + x_{i6})$

$$F_0 = \frac{[SSE(RM) - SSE(FM)]/m}{MSE(FM)} \sim F_{m,\, n-P_{FM}}$$

Donde :

M= Es el # de filas LI de Π
$\dim(\Pi)$

Regresion por minimos Cuadrados ponderados

Recordemos que para la inferencia de un subconjunto de parametros sea cual sea siempre tenemos un β_2 y que los grados de libertad r o m siempre son $\dim(\beta_2)$, solo que en el caso de la significancia de un subconjunto de parametros $\dim(\beta_2)$ es igual al # de parametros que analizamos

OBSERVACIONES O VALORES ATIPICOS ("outliers")

Estas observaciones o valores atipicos afectan considerablemente los supuestos del Modelo ya que tienen efecto sobre:

- Estimaciones de los β_j's
- Aumento considerable de $ee(\hat{\beta}_j's)$
- Supuestos del modelo

Antes de abordar en la identificacion de los outliers miremos los tipos de residuales

- $e_i = y_i - \hat{y}_i$ → R. Crudos

- $r_i = \dfrac{d_i}{\sqrt{1-h_{ii}}} = \dfrac{e_i}{\sqrt{MSE(1-h_{ii})}} \sim t_{n-p}$ → R. Estudentizados

- $d_i = \dfrac{e_i}{\sqrt{MSE}}$ → R. Estandarizados

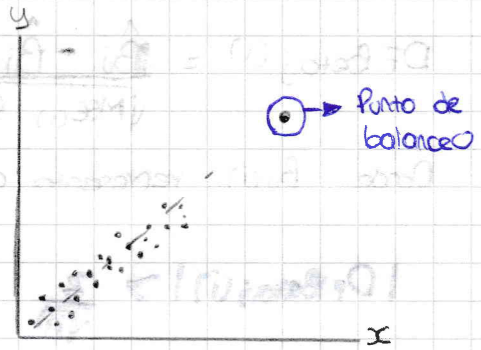
Con h_{ii}: Es el elemento de la diagonal de $\mathbb{H} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$

Representa la distancia al centro de los datos

Las observaciones atípicas se clasifican en puntos de balanceo y puntos influyentes

- Punto de balanceo (PB)

Punto alejado del resto de los datos. Posiblemente no afecta los $\hat{\beta}_j$'s pero si el R^2 y los $ee(\hat{\beta}_j's)$



Tiene un valor moderadamente inusual de las variables predictoras y en la variable respuesta.

CRITERIO

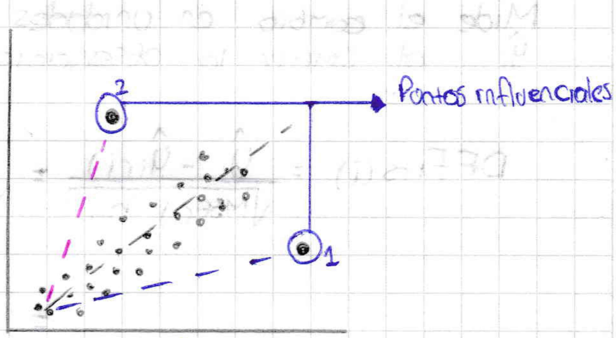
$$h_{ii} > 2\left(\frac{K+1}{n}\right) \equiv h_{ii} > \frac{2p}{n}$$

Valor inusual en x y en y

- Punto influyente (PI)

Tiene un impacto importante en los β_i's ya que el punto influyente "**hala**" el modelo en su dirección.

Afectan los supuestos del modelo



CRITERIO

Para ser candidato a PI se debe cumplir que

- h_{ii} grande ($h_{ii} \approx 1$)
- r_i grande ($|r_i| > 3$)

⟩ Candidato a PI (99%)

h_{ii} grande, $|r_i| > 2$ ⟩ Candidato a PI (95%)

Medidas para detectar observaciones influyentes

1) Distancia de cook

Utilizamos (D_i), si su valor es grande la observación es influyente

$$D_i = \frac{(\hat{y} - \hat{y}_{(i)})^T(\hat{y} - y_{(i)})}{(K+1)\,MSE} = \frac{(\beta - \beta_{(i)})^T X^T X(\hat{\beta} - \beta_{(i)})}{(K+1)\,MSE} = \frac{r_i^2}{K+1} \cdot \frac{h_{ii}}{1-h_{ii}}$$

Donde el numerador es una distancia euclidiana y $\hat{y}_{(i)}$ es la estimación quitando la iésimo observación

D_i es alto si: $D_i > F_{0,50,\,K+1,\,n-p} \approx 1 \equiv D_i > 1$

↳ Entonces es INFLUYENTE

2) DF Betas:

Mide el cambio del $\hat{B}j$ en unidades estandar al omitir la observacion i

$$DF\,Beta_j(i) = \frac{\hat{B}j - \hat{B}j(i)}{\sqrt{MSE(i)\; C_{j+1,j+1}}} = \frac{r_{j,i}}{\sqrt{r_j r_j}} \cdot \frac{e_i}{\sqrt{MSE \cdot (1-h_{ii})}}$$

Donde $Bj(i)$ representa al modelo cuando le quite la iesimo observacion

$$|DF\,Beta_j(i)| > \frac{2}{\sqrt{n}}$$; Si es influyente en el parametro es influyente en el modelo.

3) DF Fits

Mide el cambio en unidades estandar del valor ajustado $\hat{y}_i$ al omitir la observacion i

$$DF\,Fits(i) = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{MSE(i)\; h_{ii}}} = \sqrt{\frac{h_{ii}}{1-h_{ii}}} \cdot \frac{e_i}{\sqrt{MSE(i)\,(1-h_{ii})}}$$

$$= \sqrt{\frac{h_{ii}}{1-h_{ii}}} \cdot r_i$$

$$|DF\,Fits_{(i)}| > 2\sqrt{\frac{K+1}{n}}$$

Si la observacion se clasifica influyente con alguno de los 3 criterios la observacion sera influyente.

IC

$$\hat{y}_0 \pm t_{\frac{\alpha}{2},n-p} * ee(\hat{y}_0)$$

Donde $\hat{y}_0 = X_0 \cdot \hat{\beta} = $ Yohat

Se.yohat $= ee(\hat{y}_0) = \sqrt{MSE\,(X_0\,(X^T X)^{-1} X_0^T)}$

Con una confianza del $(1-\alpha)\%$ la respuesta media (Promedio) se encuentra entre (U,L) Para las condiciones dadas

IP

$$\hat{y}_0 \pm t_{\frac{\alpha}{2},n-p} * ee(\hat{y}_0 - y_0)$$

Donde

$$ee(\hat{y}_0 - y_0) = \sqrt{MSE + ee^2(\hat{y}_0)}$$

Con una confianza del $(1-\alpha)\%$ se Predice que la respuesta se encuentra entre (U,L) Para los condiciones dadas

No olvidemos que los IC y los IP solo se pueden calcular si las condiciones dadas son interpolables, lo cual verificamos de la siguiente manera

$$\text{hoo. value: } \mathbb{X}_0^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}_0 \leq \text{Max } h_{ii}$$

Donde $\mathbb{X}_0$ contiene las condiciones de interes

MULTICOLINEALIDAD.

Esta presente cuando en el modelo de RLM hay dependencias lineales entre las variables regresoras. Si hay multicolinealidad en el modelo se pueden presentar inestabilidad numérica en el calculo de la inversa asociada a $\mathbb{H}$

$\hookrightarrow (\mathbb{X}^T\mathbb{X})$

En R para evitar los problemas de multicolinealidad en la matriz se calcula la inversa generalizada

---

- Fuentes de multicolinealidad

* Metodo de recoleccion de los datos
* Restricciones en el modelo o poblacion → Consumo Electrico = # habitantes Tamaño casa Regresoras
* Modelo sobredefinido $(K > n)$
* Eleccion del modelo (Adicion de terminos polinomiales) ↳ Potencias

- Deteccion de multicolinealidad

* Varianza de los estimadores inflada (muy grande o muy pequeño con respecto a la estimacion de uno de los parametros)
  - Parametros resultan ser significativos cuando no lo son
  - Parametros resultan no significativos cuando realmente lo son
* Cuando al realizar la prueba F resulta significativo y los parametros individuales (usando t_0) resultan no significativos.
* Signos de los parametros estimados son contrarios a lo esperado.

- Diagnostico de Multicolinealidad

1) Analisis de correlacion de las variables regresoras

↳ Funcion cor(Datos)

Si $r_{ij} \approx 1$, entonces hay indicios de multicolinealidad

$R = [r_{ij}] = cor(x_i, x_j)$

$|r_{ij}| > 0{,}5$ ⟶ Normalmente se espera que entre los variables regresoras sea pequeña la correlacion

La correlacion de las variables regresoras con la variable explicada debe ser alta

2) VIF (Factor de inflacion de varianza) ⟶ Funcion miscoeficientes()
↳ (modelo, Datos)

Se ajusta un modelo tipo combinacion lineal

$$x_j \text{ vs } x_1, x_2, \ldots, x_{j-1}, x_{j+1}, x_K$$

Para $j = 1$

$$x_1 = \alpha_2 x_2 + \alpha_3 x_3 \ldots + \alpha_K x_K + C_0 + e$$

Con el modelo se calcula ⟶ R_j^2 = Coeficiente de determinacion multiple del modelo j

el cual se utiliza para determinar el VIF.

$$VIF = \frac{1}{1 - R_j^2}$$

$VIF_j < 5$ No hay multicolinealidad

$5 \leq VIF_j < 10$ Hay multicolinealidad moderada

$VIF_j \geq 10$ Hay multicolinealidad grave.

3) Analisis de los valores propios de $(X^T X)$

↳ Funcion misDiagnosticolin(modelo)

Este analisis se puede realizar sin centrar o centrado.

$\frac{x_{ij} - \bar{x}_j}{S_{x_j}}$

↳ Se hace centrado cuando el intercepto no tiene interpretacion y sucede cuando $0 \notin [x_j min, x_j max]$ para algun j
↳ (Centrar = T)

i) Numero de condicion

$$K = \frac{\lambda \, max}{\lambda \, min}$$

Criterio

$K < 100$ no hay multicolinealidad

$100 \leq K < 1000$ Multicolinealidad moderada

$K \geq 1000$ Multicolinealidad grave

La funcion arroja $\sqrt{K}$

ii) Indicadores o indices de condicion

cond. index: $K = \frac{\lambda_{max}}{\lambda_j}$, $j = 1, 2, 3 \ldots, K$

$K_j < 100$ $\forall j = 1, 2, 3, \ldots K$ no hay multicolinealidad

$100 \leq K_j < 1000$ $\forall j = 1, 2, 3, \ldots K$ Hay multicolinealidad

$K_j \geq 1000$ $\forall j = 1, 2, 3, \ldots K$ Multicolinealidad grave o severa.

iii) Proporcion de descomposicion de varianza (π_{ij})

↳ Cuando se ve multicolinealidad grave o severa

Representa la proporcion de varianza de cada $\hat{B}_j$ (o de cada VIF) debida al iesimo valor propio

Pi. xj

MULTICOLINEALIDAD

- Indicios
  - Matriz de correlacion
    - Cor(Datos)
- ¿En que grado?
  - Factores de inflacion de varianza VIF
    - miscoeficientes(modelo, Datos)
    - VIF
      - ≤ 5 no hay multicolin.
      - $(5, 10)$ hay multicolin moderada
      - ≥ 10 hay multicolin grave o severa
  - Indices de condicion
    - misDiagnostcolin(modelo)
    - $\sqrt{K_j}$
      - ≤ 10 no hay muticolin.
      - $(10, 31.6)$ hay muticolin moderada
      - ≥ 31.6 hay multicolin grave o severa
- ¿Quieres?
  - Proporcion de descomposicion de varianza π_{ij}
    - misDiagnostcolin(modelo)
    - $\pi_{ij} > 0.5$

SELECCION DE MODELO

Suponga que se tiene un modelo con K variables regresoras y desea hallar un modelo reducido que utilice menos de K variables regresoras.

$$\binom{K}{1}+\binom{K}{2}+\binom{K}{3}+\dots+\binom{K}{K} = \sum^{K} \binom{K}{j} 1^{K} 1^{K-j}$$
$$= \boxed{2^K-1}$$

Para la seleccion de modelos Dos metodos:

- METODO DE TODAS LAS REGRESIONES POSIBLES
  ↳ allregtable (Modelo, Respuesta)
  * R^2
  * R^2_{adj}

  Estimaciones o hacer inferencia

  * Cp de Mallows ⟶ Hacer predicciones

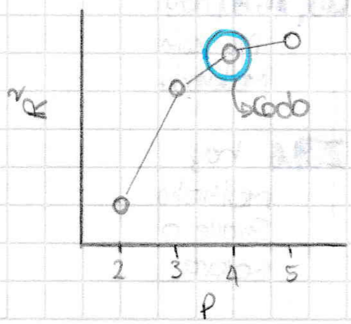
- METODOS SECUENCIALES
  * Forward (Hacia adelante)
  * Backward (Hacia atras)
  * Stepwise (Adelante, atras) ⟶ Elimina Multicolinealidad

- TODAS LAS REGRESIONES POSIBLES

Se compara cada modelo y se analiza R^2, R^2_{adj}, MSE, CP.

* R^2 = se compara cada modelo con el modelo completo y si hay un submodelo donde

$|R^2_p - R^2|$ es pequeño entonces se prefiere el modelo mas pequeño

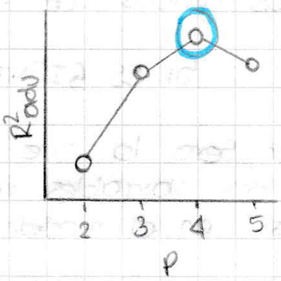


Buscamos el codo

NOTA: Analizar el R^2 es equivalente a analizar el MSE

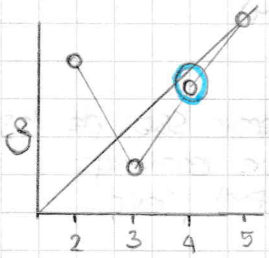
* R^2_{adj} = Se busca el que tenga un mayor valor, ya que el R^2_{adj} penaliza en funcion del numero de variables

→ $1 - \frac{MSE}{MST}$



Buscamos el modelo que tenga el mayor R^2_{adj}

* Cp de Mallows: El mejor modelo es el de cp mas pequeño comparado con el p correspondiente



$|Cp - p| \approx 0$

El modelo mas cercano a cero es insesgado

p = # de parametros del modelo analizado

Buscamos el modelo con Cp mas cercano a p dentro de todos los modelos.

El cp del modelo completo siempre es igual a p

El cp es una medida del sesgo del modelo de regresion, es decir

$(E[\hat{y}_i] - E[y_i])$

$$CP = \frac{SSEP}{MSE} - (n - 2P)$$

└ Modelo completo

Se puede demostrar que $E[Cp \mid sesgo = 0] = P$ por ello se eligen modelos.

$\boxed{Cp \approx P}$

- METODOS SECUENCIALES

* FORWARD (Hacia adelante)

Se parte de un modelo con intercepto de la forma

$$y_i = \beta_0 + \varepsilon_i$$

Paso 1. El parametro candidato a entrar sera el que tenga mayor SSR ≡ Menor SSE dentro de los parametros con una variable regresora y probamos su significancia en el modelo utilizando SSextra

$$F_0 = \frac{SSR(\beta_{j1} \mid \beta_0)}{MSE(FM)} \sim F_{1, n-p}$$

$H_0: \beta_{j1} = 0$
$H_a: \beta_{j1} \neq 0$

NO olvidemos que el modelo completo en el primer paso tiene β_0, β_{j1} portanto $P=2$, ademas recordemos que para cada paso los grados de libertad del numerador siempre son "1"

SI ES SIGNIFICATIVA ENTRA

Paso 2. La candidata a entrar sera la que tenga menor SSE de los modelos con dos variables regresoras que contenga β_{j1}. Los parametros en el modelo (FM) seran $\beta_0, \beta_{j1}, \beta_{j2}$ por tanto $p=3$

$$H_0: \beta_{j2} = 0 \qquad F_0 = \frac{SSR(\beta_{j2} \mid \beta_0, \beta_{j1})}{MSE(FM)} \sim F_{1,n-3}$$
$$H_a: \beta_{j2} \neq 0$$

SI ES SIGNIFICATIVA ENTRA

Paso 3. La candidata a entrar sera la que tenga menor SSE de los modelos con tres variables regresoras que contenga β_{j1}, β_{j2}. Los parametros en el modelo (FM) seran $\beta_0, \beta_{j1}, \beta_{j2}, \beta_{j3}$ por tanto $P=4$

$$H_0: \beta_{j3} = 0 \qquad F_0 = \frac{SSR(\beta_{j3} \mid \beta_0, \beta_{j1}, \beta_{j2})}{MSE(FM)} \sim F_{1,n-4}$$
$$H_a: \beta_{j3} \neq 0$$

SI ES SIGNIFICATIVA ENTRA

En general se van probando variables una a una y si son significativos entran, esto hasta que ya no hayan mas variables por probar.

PASO 1 = Prueba sobre β_{j1}
FM: β_{j1}, β_0
$F_{1,n-2}$
Significativo: Entra

PASO 2 = Prueba sobre β_{j2}
FM: $\beta_{j1}, \beta_{j2}, \beta_0$
$F_{1,n-3}$
Significativa: Entra

PASO 3 = Prueba sobre β_{j3}
FM: $\beta_{j1}, \beta_{j2}, \beta_{j3}, \beta_0$
$F_{1,n-4}$
Significativa: Entra

PASO 4 = Prueba sobre β_{j4}
FM: $\beta_{j1}, \beta_{j2}, \beta_{j3}, \beta_{j4}, \beta_0$
$F_{1,n-5}$
Significativo: Entra

⋮

Hasta que no hayan mas variables regresoras (K)
o
Falle alguno de los pasos, entonces se detendra

La primera candidata por lo general es la de F parcial mas grande

* Backward (Hacia atrás)

Se parte de un modelo de la forma

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_K x_{iK}$$

↳ El modelo ajustado con todas las variables (Modelo Inicial)

PASO 1. El parametro candidato a salir será el que no este en el modelo con Mayor SSR $\Leftrightarrow$ Menor SSE que tiene K-1 Variables regresoras

$$F_0 = \frac{SSR(\beta_{j1} \mid \beta_0, \beta_1, \dots, \beta_{j1-1}, \beta_{j1+1}, \beta_K)}{MSE\,(FM)} \sim F_{1,n-p}$$

↳ MODELO INICIAL

$H_0: \beta_{j1} = 0$ Si no es Significativo sale

$H_a: \beta_{j1} \neq 0$

No olvidemos que el modelo completo en el primer paso tiene todas las variables regresoras recolectadas por tanto $p = K+1$, además recordemos que para cada paso los grados de libertad del numerador siempre son "1".

PASO 2. El parametro candidato a salir será el que no este en el modelo con menor SSE que tiene K-2 Variables regresoras en el cual no está β_{j1}

$$F_0 = \frac{SSR(\beta_{j2} \mid \beta_0, \beta_1, \beta_{j1-1}, \beta_{j1+1}, \dots, \beta_{j2-1}, \beta_{j2+1}, \dots)}{MSE\,(FM)} \sim F_{1,n-K}$$

↳ modelo sin β_{j1}

$H_0: \beta_{j2} = 0$ Si no es significativa sale

$H_a: \beta_{j2} \neq 0$

PASO 3. El parametro candidato a salir será el que no este en el modelo con menor SSE que tiene K-3 Variables regresoras en el cual no están β_{j1} y β_{j2}

$$F_0 = \frac{SSR(\beta_{j3} \mid \beta_0, \beta_1, \beta_{j1-1}, \beta_{j1+1}, \beta_{j2-1}, \beta_{j2+1}, \dots, \beta_{j3-1}, \beta_{j3+1}, \beta_{K-2})}{MSE\,(FM)} \sim F_{1,n-K-1}$$

↳ modelo sin β_{j1} y β_{j2}

$H_0: \beta_{j3} = 0$

$H_a: \beta_{j3} \neq 0$ Si no es Significativo sale.

En general se van probando variables una a una y si no son significativas salen, esto hasta llegar a un modelo solo con intercepto o hasta que falle alguno de los pasos.

Paso 1 = Prueba sobre β_{j1}
FM: Modelo inicial
$F_{1, n-k+1}$
no significativo: sale

Paso 2 = Prueba sobre β_{j2}
FM: Modelo sin β_{j1}
$F_{1, n-k}$
no significativa: sale

Paso 3 = Prueba sobre β_{j3}
FM: Modelo sin β_{j1} y β_{j2}
$F_{1, n-k-1}$
no significativo: sale

Paso 4 = Prueba sobre β_{j4}
FM: modelo sin $\beta_{j1}, \beta_{j2}, \beta_{j3}$
$F_{1, n-k-2}$
no significativo: sale

⋮

Hasta llegar al modelo solo con β_0
o
falle algún paso, entonces se detiene. → Que no se elimine una variable

* STEPWISE
↳ El mejor = Elimina las que tienen multicolinealidad

En los primeros pasos es igual al método Forward y después analizo significancia de subconjunto de parámetros

PASO 1 = Prueba sobre β_{j1}
FM = β_{j1}, β_0
$F_{1, n-2}$
Significativo = entra

PASO 2 = Prueba sobre β_{j2}
FM = $\beta_{j2}, \beta_{j1}, \beta_0$
$F_{1, n-3}$
Significativo = entra

PASO 3: Se determina si el efecto de β_{j1} es significativo en la presencia de β_{j2}

↓
Eliminación

$H_0: \beta_{j1} = 0 \mid \beta_{j2}$
$H_a: \beta_{j1} \neq 0 \mid \beta_{j2}$

$$F_0 = \frac{SSR(\beta_{j1} \mid \beta_0, \beta_{j2})}{MSE(FM)} \sim F_{1, n-3}$$
↳ $\beta_0, \beta_{j1}, \beta_{j2}$

SI ES SIGNIFICATIVA NO SALE

PASO 4. Igual a paso 3 en forward: Prueba sobre β_{j3}
FM: $\beta_{j3}, \beta_{j2}, \beta_{j1}, \beta_0$
$F_{1, n-4}$
Significativa = Entra

Paso 5. Se analiza la significancia de los parametros β_{J1} y β_{J2} dado que en el modelo esta β_{J3}, se hace

Eliminacion Individual y para los dos al tiempo.

- Prueba 1.

Siempre son modelos con intercepto

$H_0: \beta_{J1} = 0 \mid \beta_{J2}, \beta_{J3}$

$H_a: \beta_{J1} \neq 0 \mid \beta_{J2}, \beta_{J3}$

$$F_0 = \frac{SSR(\beta_{J1} \mid \beta_{J2}, \beta_{J3})}{MSE(FM)} \sim F_{1, n-4}$$

$\beta_0, \beta_{J1}, \beta_{J2}, \beta_{J3}$

SI ES SIGNIFICATIVA
NO SALE

- Prueba 2.

$H_0: \beta_{J2} = 0 \mid \beta_{J1}, \beta_{J3}$

$H_a: \beta_{J2} \neq 0 \mid \beta_{J1}, \beta_{J3}$

$$F_0 = \frac{SSR(\beta_{J2} \mid \beta_{J1}, \beta_{J3})}{MSE(FM)} \sim F_{1, n-4}$$

$\beta_0, \beta_{J1}, \beta_{J2}, \beta_{J3}$

SI ES SIGNIFICATIVA
NO SALE

- Prueba 3.

$H_0: \beta_{J1} = 0 \wedge \beta_{J2} = 0 \mid \beta_{J3}$

$H_a: \beta_{J1} \neq 0 \vee \beta_{J2} \neq 0 \mid \beta_{J3}$

$$F_0 = \frac{SSR(\beta_{J1}, \beta_{J2} \mid \beta_{J3})}{MSE(FM)} \sim F_{2, n-4}$$

$\beta_0, \beta_{J1}, \beta_{J2}, \beta_{J3}$

SI SON SIGNIFICATIVAS
NO SALEN

Si se corrobora la significancia de los dos parametros dado que entro β_{J3} entonces no elimino ninguna de las dos variables y sigo con el procedimiento [Entro, Elimino, Entro...]

REGRESION CON VARIABLES INDICADORAS

Existen problemas que ademas de tener un conjunto de variables $(x_1, x_2, x_3, \dots, x_k)$ tienen diferentes niveles de una variable cualitativa o categorica, por ejemplo:

- Sexo: Hombre
  Mujer
- Turno: A = Mañana
  B = Tarde
  C = Noche
- Estrato: E_1, E_4
  E_2, E_5
  E_3, E_6
- Nacionalidad:
- Tipo de Maquina
- Tipo de llanta

Otras regresiones
- Modelo lineal generalizado
- Logistico

a = Categorias
a-1 = Indicadoras

Cada una de las variables se incorpora usando variables indicadoras

$$I = \begin{cases} 1 & \text{Categoria a} \\ 0 & \text{e.o.p} \end{cases}$$

Para variables cualitativas con dos niveles tenemos

$$I = \begin{cases} 1 & \text{Categoria 1} \\ 0 & \text{Categoria 2} \end{cases} \rightarrow \text{Categoria de referencia}$$

En el cual solo es necesario definir una variable indicadora, tomando una categoria de referencia con la que se va a comparar

El modelo para este caso seria de la forma

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 I + \beta_3 I x_{i1} + \varepsilon$$

Con $I=1$; categoria 1, Modelo:

$$y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{i1} + \varepsilon$$

Donde

β_2 = Mide el cambio en el intercepto con la categoria 1

β_3 = Mide el cambio en la pendiente con la categoria 1

Con $I=0$; Categoria 2, Modelo:

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

Hallando los valores esperados

- categoria 1

$$E[y|x] = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x$$

- categoria 2

$$E[y|x] = \beta_0 + \beta_1 x$$

PREGUNTAS

↳ dependencia de los categoria en la relacion y vs x

- Es necesario la variable categorica para explicar y con x

↳ Igualdad rectas de regresion

$$H_0: \beta_0 = \beta_0 + \beta_2 \wedge \beta_1 = \beta_1 + \beta_3 \equiv \beta_2 = 0 \wedge \beta_3 = 0$$
$$H_a: \beta_0 \neq \beta_0 + \beta_2 \vee \beta_1 \neq \beta_1 + \beta_3 \qquad \beta_2 \neq 0 \vee \beta_3 \neq 0$$

- Los pendientes de ambas rectas son iguales

↳ Efecto de x sobre y para las categorias

$$H_0: \beta_1 = \beta_1 + \beta_3 \equiv \beta_3 = 0$$
$$H_a: \beta_1 \neq \beta_1 + \beta_3 \qquad \beta_3 \neq 0$$

- Los pendientes de ambas rectas son iguales y no significativas

$$H_0: \beta_1 = \beta_1 + \beta_3 = 0 \equiv \beta_1 = 0 \wedge \beta_3 = 0$$
$$H_a: \beta_1 \neq 0 \vee \beta_1 + \beta_3 \neq 0 \vee \beta_1 \neq \beta_1 + \beta_3 \qquad \beta_1 \neq 0 \vee \beta_3 \neq 0$$

Para variables cualitativas con tres niveles tenemos

Es necesario definir 2 variables indicadoras

$$I_1 = \begin{cases} 1 & \text{categoria 1} \\ 0 & \text{e.o.p} \end{cases} \qquad I_2 = \begin{cases} 1 & \text{categoria 2} \\ 0 & \text{e.o.p} \end{cases}$$

La categoria de referencia es la 3 y se obtiene cuando $I_1 = 0$ $\wedge$ $I_2 = 0$, esto es lo visualizamos de la siguiente forma:

I_1	I_2	
1	0	categoria 1
0	1	categoria 2
0	0	categoria 3

Modelo de segundo orden: con una variable regresora

$$y = \beta_0 + \beta_1 x + \beta_2 I_1 + \beta_3 I_2 + \beta_4 I_1 x + \beta_5 I_2 x + \varepsilon$$

MODELOS

- Categoria 1: $I_1 = 1$ $I_2 = 0$

$$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) x + \varepsilon$$

- Categoría 2: $I_1 = 0 \quad I_2 = 1$

$$y = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)x + \varepsilon$$

- Categoría 3: $I_1 = 0 \quad I_2 = 0$

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

PREGUNTAS

- Igualdad de las rectas de regresion : Necesidad de incorporar la variable categorica en el modelo de regresion.

$H_0: \beta_0 = \beta_0 + \beta_3 = \beta_0 + \beta_2 \wedge \beta_1 = \beta_1 + \beta_4 = \beta_1 + \beta_5$

$$\Updownarrow$$

$H_0: \beta_3 = 0 \wedge \beta_2 = 0 \wedge \beta_4 = 0 \wedge \beta_5 = 0$

$H_a: \beta_3 \neq 0 \vee \beta_2 \neq 0 \vee \beta_4 \neq 0 \vee \beta_5 \neq 0$

- Las pendientes de las rectas son iguales
  con esta prueba se puede analizar el efecto de x sobre y de acuerdo a las diferentes combinaciones de las categorias

$H_0: \beta_1 = \beta_1 + \beta_4 = \beta_1 + \beta_5 \quad \equiv \quad \beta_4 = 0 \wedge \beta_5 = 0$

$H_a: \cancel{\beta_1 \neq \beta_1 + \beta_4 \neq \beta_1 + \beta_5} \quad \quad \beta_4 \neq 0 \vee \beta_5 \neq 0$

- Las pendientes de las rectas son iguales y no significativos

$H_0: \beta_1 = \beta_1 + \beta_4 = \beta_1 + \beta_5 = 0$

$H_a: \beta_1 \neq 0 \vee \beta_1 + \beta_4 \neq 0 \vee \beta_1 + \beta_5 \neq 0 \vee \beta_1 \neq \beta_1 + \beta_4 \vee \beta_1 \neq \beta_1 + \beta_5 \vee \dots$

$$\Updownarrow$$

$H_0: \beta_1 = 0 \wedge \beta_4 = 0 \wedge \beta_5 = 0$

$H_a: \beta_1 \neq 0 \vee \beta_4 \neq 0 \vee \beta_5 \neq 0$