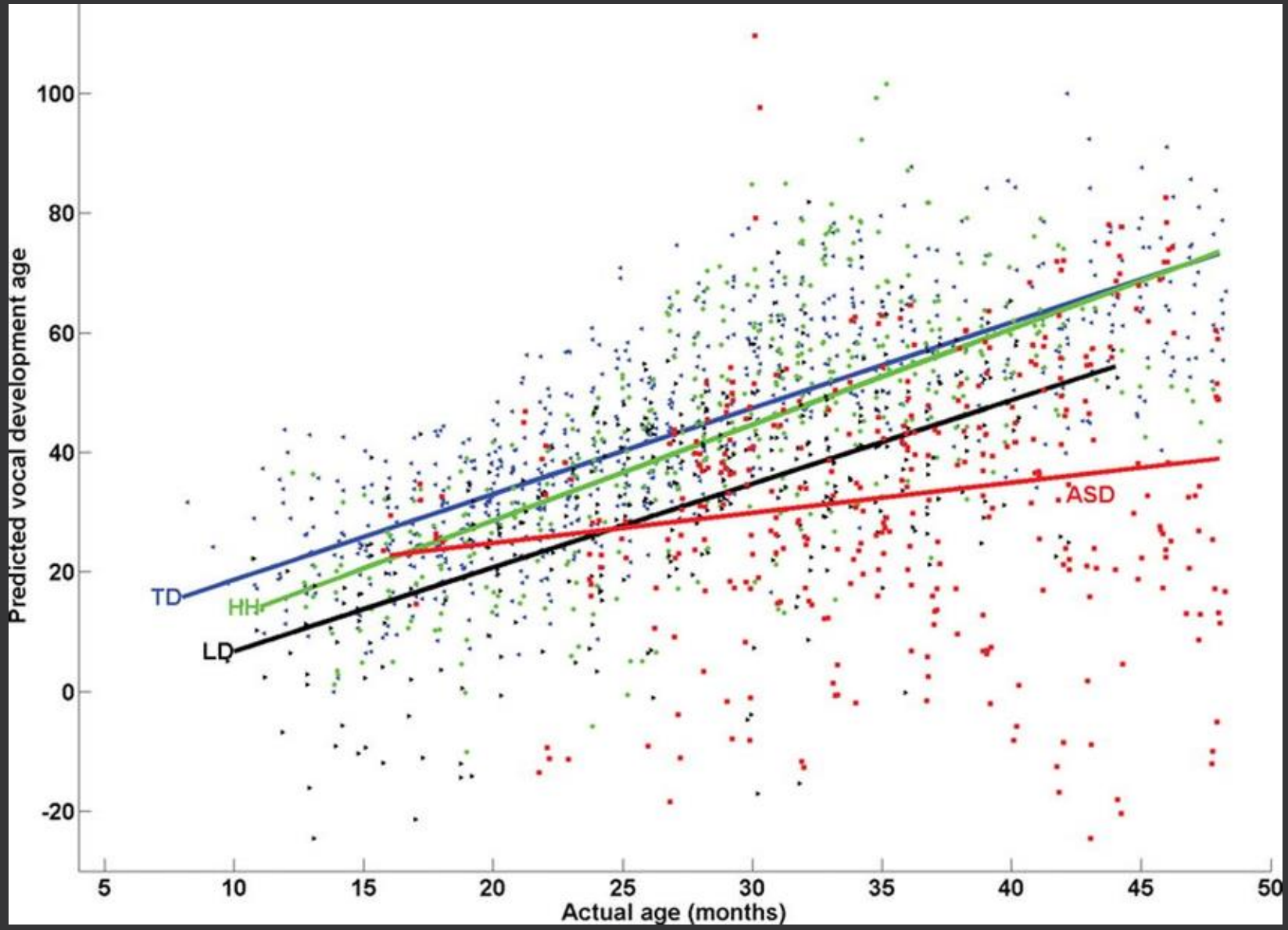
# Regresión lineal múltiple




Óscar Centeno Mora

# Preámbulo

- En temas anteriores repasamos los principios de la regresión lineal bivariada (RLB).
- En el presente, veremos la regresión lineal múltiple (RLM), y algunos aspectos que no se han cubierto sobre la regresión.
- ¿Por qué debemos recurrir a la RML y no solo aplicamos k regresiones bivariadas?
- Una mejora considerable en la explicación de una variable predictora es cuando introducimos dos o más variables explicativa, por lo que conviene analizar este escenario.

# Preámbulo

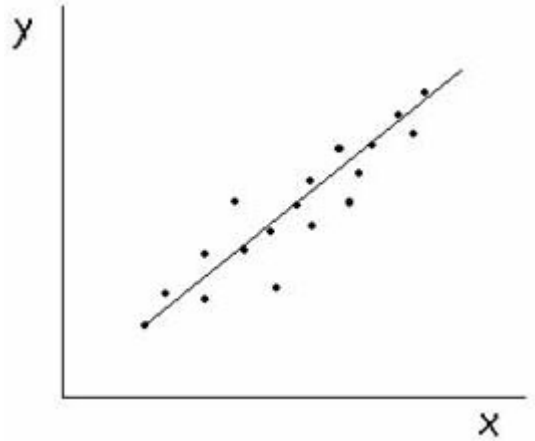
## Regresión bivariada y regresión multivariada

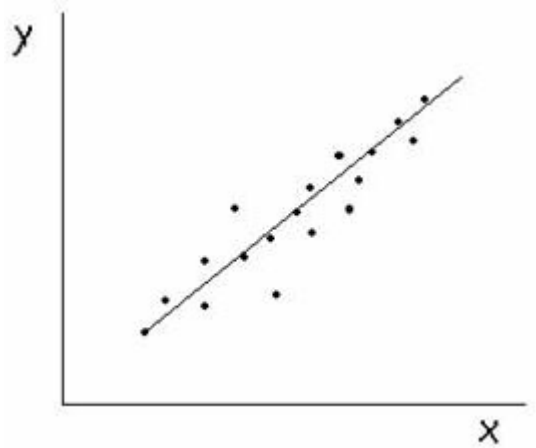
**Regresión bivariada**

Una variable dependiente (Y)
Una variable independiente (X)

$$Y \sim X$$

$$\begin{bmatrix} Y1 \\ Y2 \\ \cdots \\ Y_n \end{bmatrix} \sim \begin{bmatrix} X1 \\ X2 \\ \cdots \\ X_n \end{bmatrix}$$



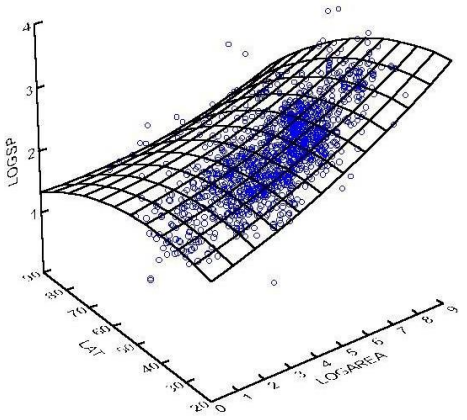




**Regresión multivariada**

Una variable dependiente
Dos o más variables independientes

$$Y \sim X$$

$$\begin{bmatrix} Y1 \\ Y2 \\ \cdots \\ Y_n \end{bmatrix} \sim \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$



# Preámbulo

- Para el caso de la regresión lineal bivariada, habíamos explicado que un buen análisis de regresión, se componía de los siguientes elementos:

1. Relación entre las variables
2. Estimación de la recta de mejor ajuste (modelo de regresión)
3. Diagnóstico del modelo de regresión
4. Medidas remediales
5. Estadísticas de bondad y ajuste
6. Inferencia y prueba de hipótesis de los coeficientes del modelo

- También, y normalmente, se debe de seguir este orden.

- Antes de iniciar un correcto análisis de una RLM, veremos la base de esta técnica.

# Índice


1 Introducción

2 La RLM y sus condiciones o supuestos

3 Estimación de la RLM

4 Valores predichos y residuales

5 Inferencia de la RLM y sus coeficientes

6 Evaluación de la RLM

# Índice

1 Introducción

# Introducción

- Un análisis en la explicación de cierta variable o fenómeno requiere una interacción mayor de sus características:

1. Análisis de la estructura del gasto.
2. Análisis de los factores que influyen en el ingreso medio.
3. Los factores asociados a la pobreza de un país o region.
4. Medir la eficiencia y eficación de una cadena de valor.
5. Los factores asociados a la perdida de grasa.
6. Estudio de la aglomeración de tránsito vehicular en una region.

- En la explicación de lo anterior se podrían emplear técnicas de regression que busquen establecer relaciones asimétricas entre las variables con tal de brindar una aproximación empírica al tema de estudio.

# Introducción

- Antes de iniciar cualquier que sea el análisis, es esencial tener claro, el objetivo que se persigue, y a partir de lo anterior, los insumos con los que se cuenta.
- En la aplicación de métodos de regresión, se debería formular siempre el siguiente cuadro para tener una mayor claridad:

| Características del estudio | Valores asociados |
|---|---|
| Objetivo del estudio: | ------ |
| Que se espera: | ------ |
| Unidad de estudio | ------ |
| Variables** | ------ |
| Variable respuesta (dependiente) | ------ |
| Variable predictoras (independiente) | ------ |

# Introducción

- Como lo comentamos anteriormente, una regresión bivariada es insuficiente en la explicación de una característica. Se deben contemplar otras alternativas. Un único predictor resulta es un modelo o explicación imprecisa de lo que realmente podría llegar a ser.

- Si no se consideran con conjunto de variables independientes o predictores, se podrá potenciar el uso de la regresión.

- En la construcción de una ecuación o modelo lineal de regresión multiple, se deben especificar, en un inicio, 3 aspectos:

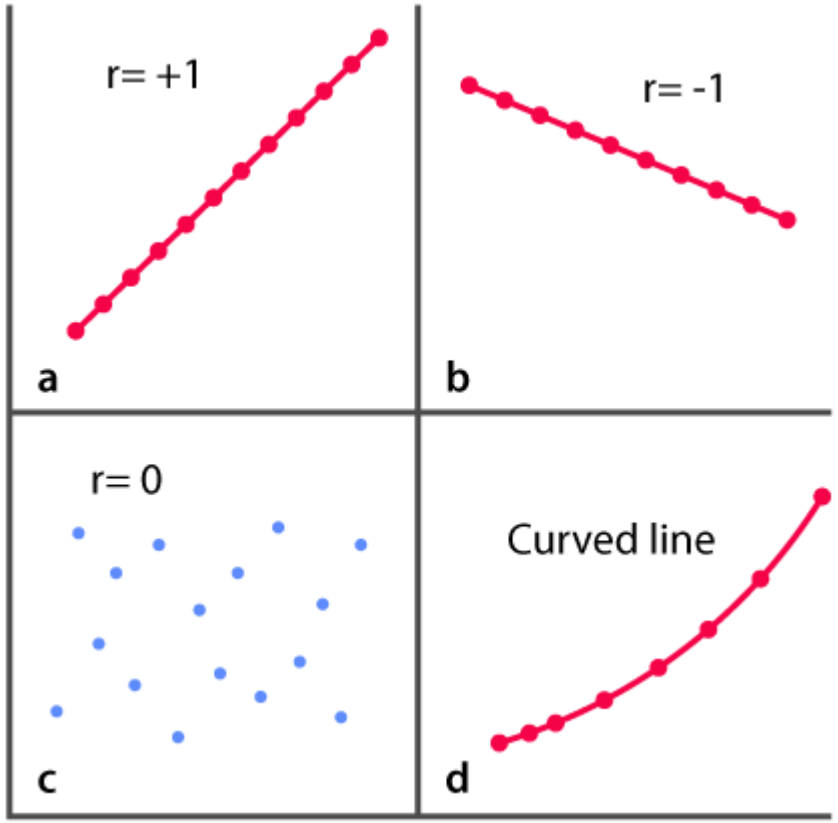
1. Las variables independientes o predictores.
2. La forma funcional o relacional de las variables en la regresión.
3. Y el para qué de la aplicación de la regresión (predecir valores, nivel explicativo de las variables, estandarización de parámetros de análisis, etc.).

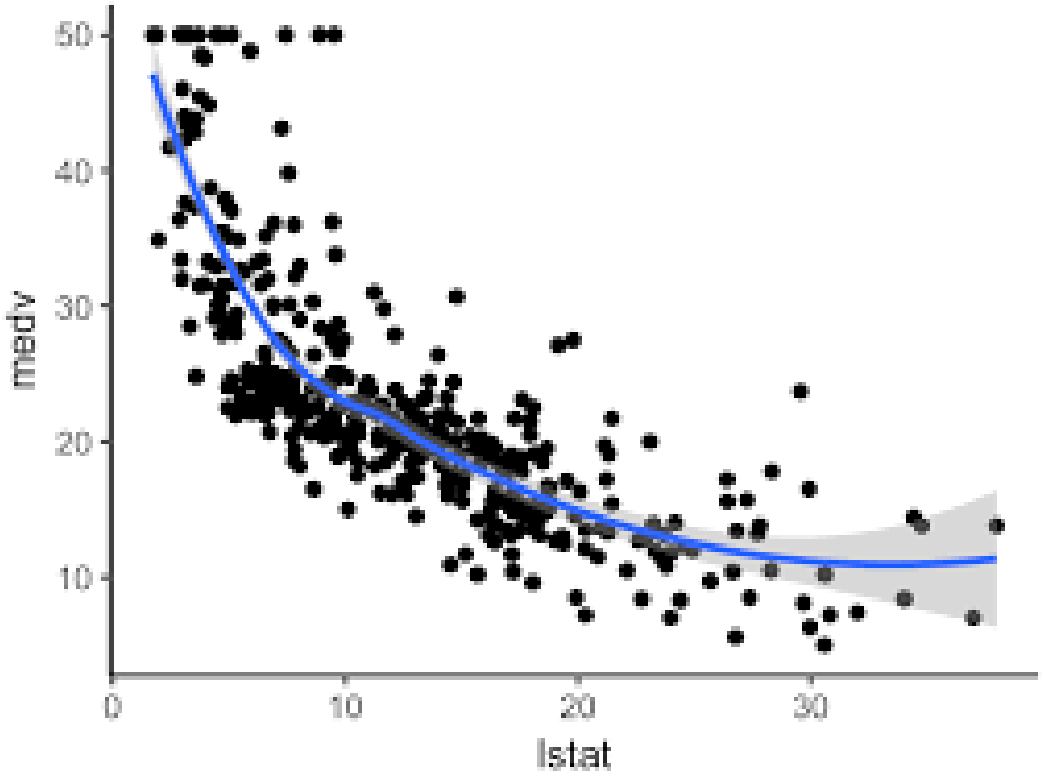
# Introducción

1. **<u>La selección de las variables:</u>** en esta etapa es seleccionar el conjunto de variables que me ayudan a determinar la ecuación o el modelo de regresión. Poseemos una variable dependiente, y *k* variables independientes. Más adelante veremos formas que nos ayudan a determinar un conjunto *k* variables independientes.

# Introducción

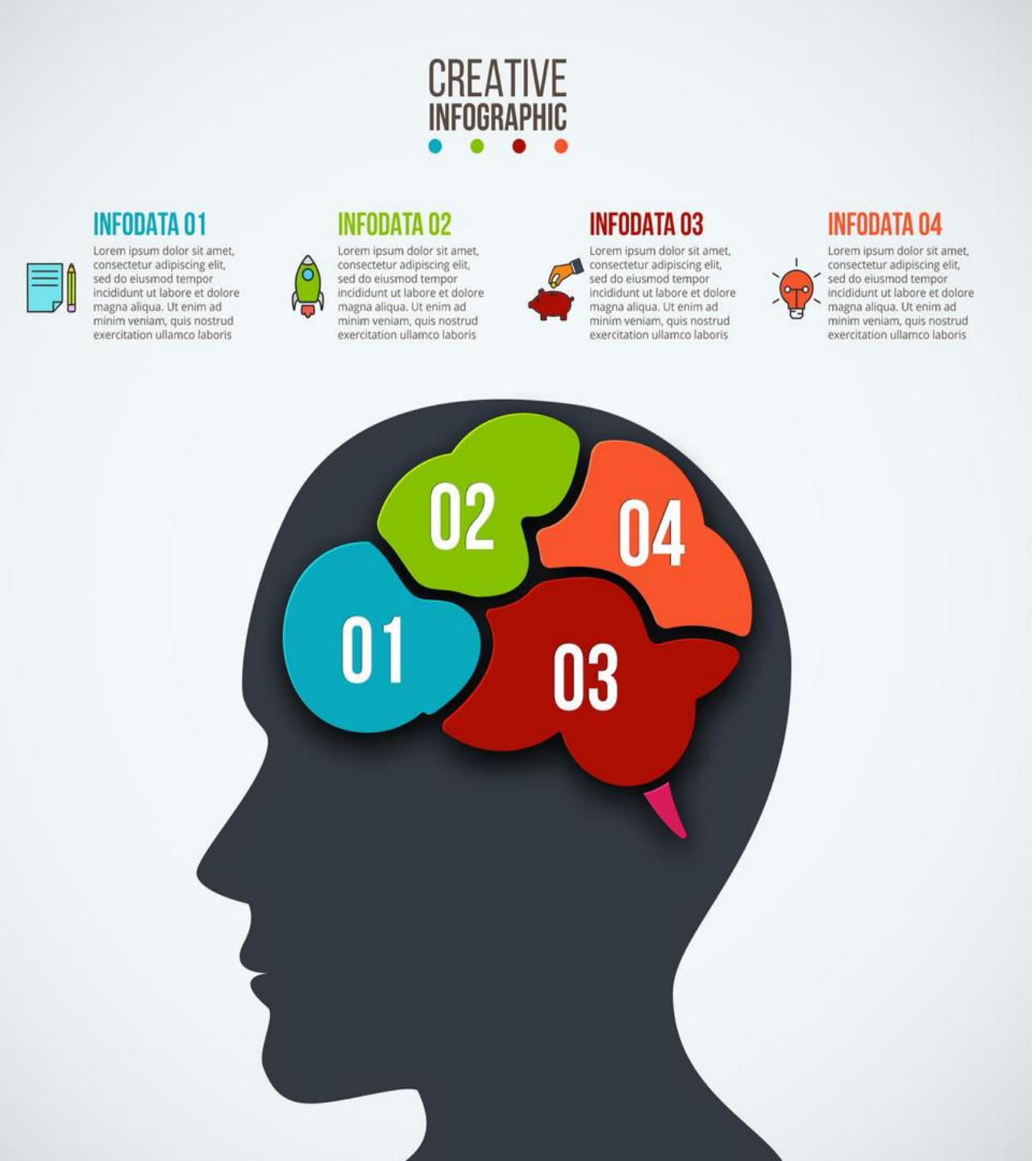
2. **<u>Forma funciona de la relación en la regresión</u>**: en la relación con la variable dependiente *Y*, es posible que las otras variables independientes no posean siempre una forma completamente lineal, sino pueden ser de otra forma. Es importante tenerlo en cuenta para próximas transformaciones, o interpretaciones.



# Introducción

**3. El alcance del modelo de regresión**: llegaran al mundo laboral, y se les dirá “aplique un modelo de regresión”, y ustedes contestarán, “¿para qué?...” Una ecuación de regresión sirve para muchas cosas: predecir valores medios, medir el efecto de una variable predictora, establecer estándares de uso, etc., etc., etc., pero debemos tener claro, el para qué vamos a construir o implementar el modelo de regresión.




# Introducción

RECORDAR: un modelo de regresión, si bien nos aporta una aproximación empírica a la situación que nos interesa analizar, SIEMPRE tendrá un trasfondo conceptual…. no midamos cosas por medirlas…

# Índice

1 Introducción

2 La RLM y sus condiciones o supuestos

# Aplicación de la RLM y sus condiciones

- ¿Hemos sido bastante enfáticos en que es un regresión **lineal** múltiple?

# Aplicación de la RLM y sus condiciones

- Para una regresión lineal bivariada, es sencillo tanto su interpretación en la ecuación, así como su visualización. Para una RLM, es un tanto más laborioso…

- Una regresión lineal múltiple se expresa de la siguiente forma:

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

- De forma simplificada, solemos denotar :

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

# Aplicación de la RLM y sus condiciones

- Solemos partir a partir de los siguientes insumos, o datos:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \dots & x_{2,p-1} \\ \dots & \dots & \dots & & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n,p-1} \end{pmatrix}$$

- La labor de la RLM, es encontrar los parámetros $\beta$, con tal de poder dar una explicación funcional a nuestro problema de fono. A partir de los :

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_{p-1} \end{pmatrix}$$

es que podemos llegar a la ecuación de la RLM:

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

# Aplicación de la RLM y sus condiciones

- La ecuación de regresión, aunque útil, no es del todo flexible, dado que supone unos condiciones que se deberían cumplir para su correcto uso:

1. Relación lineal entre los predictores y la respuesta
2. Variancia constante.
3. Normalidad en sus residuos.
4. No presencia de multicolinealidad.
5. Homoscedasticidad o variancia constante de los errores.
6. Independencia de errores o no autocorrelación de los errores (aplicable para datos temporales).

- Todas esas condiciones, se resumen en la siguiente ecuación de supuestos de la RLM:

$$\varepsilon \sim N(0, \sigma^2 I)$$


# Aplicación de la RLM y sus condiciones

Y, ¿si la RLM no cumple con lo anterior, entonces no podemos utilizar dicho método de estimación?

# Índice


1 Introducción

2 La RLM y sus condiciones o supuestos

3 Estimación de la RLM

# Estimación de la RLM

- Hablar de estimar la RLM, es estimar los coeficientes $\beta$ de la recta. Es similar al caso de la regression lineal bivariada: si se conocen los $\beta$, se obtiene por lo tanto la ecuación de la RLM:

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

- Cuando hablamos de la estimación de la RLM, existen diversos métodos posibles:
  1. Estimación por mínimos cuadrados ordinaries (MCO).
  2. Estimación por máxima verosimilitud.
  3. Estimación por mínimos cuadrados generalizados.
  4. Estimación por estimadores robustos (no lineales en $Y$).
  5. Estimaciones sesgadas.
  6. Etc…

- Las dos primera brindan una solución analítica, las otras son estimaciones por aproximaciones numéricas.

# Estimación de la RLM

- De forma equivalente para la RLB, en la RLM la recta de mejor ajuste, se puede, para el método de MCO, obtener mediante la minimización de los errores.
- Se aproximan los $\beta_i$ de tal forma que estos sean la recta de major ajuste de la RLM que minimiza la distancia entre los valores estimados ($\hat{Y}$), y las observaciones ($Y$).
- Sea $Q$, el componente que queremos minimizer :

$$e'e = \begin{pmatrix} e_1 & e_2 & \dots & e_n \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{pmatrix} = \sum_{i=1}^{n} e_i^2 = Q \longrightarrow Q = \sum \left[ Y_i - \hat{Y}_i \right]^2 = \sum e_i^2$$

$$Q = \sum \left[ Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1}) \right]^2$$

$$Q = (Y - X\beta)'(Y - X\beta)$$

$$Q = Y'Y - \beta' X'Y - Y' X\beta + \beta' X' X\beta$$

$$Q = Y'Y - 2\beta' X'Y + \beta' X' X\beta$$

# Estimación de la RLM

¿Cómo minimizamos $Q$?

# Estimación de la RLM

- En la búsqueda del vector de los $\beta_i$, o los estimadores de los parámetros de $\beta$ que minimice la suma de cuadrados residual (SCR), debemos deriver con respect a $\beta$ e igual a 0 la expression anterior. Sea la siguiente formulación:

$$e'e = Y'Y - 2\beta' X'Y + \beta' X' X\beta$$

$$\frac{\partial(e'e)}{\partial\beta} = -2X'Y + 2X' X\beta = 0$$

Al despejar el $\beta$, la solución analítica para esta ecuación son las estimaciones de mínimos cuadrados de los coeficientes de la RLM.

$$\hat{\beta} = \left(X' X\right)^{-1} X'Y \longrightarrow \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_{p-1} \end{pmatrix}$$

# Estimación de la RLM

- Otra forma de obtener los coeficientes $\beta_i$ en la recta de la RLM, es mediante la estimación por máxima verosimilitud (*likelihood*). Para encontrar los valores en **β** por el método de máxima de verosimilitud se debe escribir la función de **densidad conjunta** o **función de verosimilitud**:

$$\varepsilon_i \sim N(0,\sigma^2) \longrightarrow f_{\varepsilon_i}(\varepsilon_i = e_i) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left[-\frac{1}{2\sigma^2} e_i^2\right]$$

$$L(\beta,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left[-\frac{1}{2\sigma^2} e_i^2\right] = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum e_i^2\right]$$

- La maximización de la función de verosimilitud es equivalente a la maximización de su logaritmo:

$$l(\beta,\sigma^2) = \ln\left[\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\right] - \frac{1}{2\sigma^2} \sum e_i^2$$

- Para encontrar los estimadores de los β's, se considera el término que contiene la sumatoria de los errores cuadráticos pues éstos están en función de los β's. Entonces el primer término se puede obviar por ser una constante (no depende de los β's ) y se debe maximizar:

$$-\frac{1}{2\sigma^2} \sum e_i^2 \longrightarrow$$ **Esto es equivalente a minimizar cuyo resultado corresponde al obtenido por mínimos cuadrados ordinarios.**

# Estimación de la RLM

- Ejemplo:

  "Una compañía tiene estudios fotográficos en 21 ciudades de tamaño medio. La compañía está considerando expandirse a otras ciudades de tamaño medio y desea investigar si las ventas (Y) en una comunidad pueden predecirse a partir del número de personas de edad 16 o menores en la comunidad ($X_1$) y el ingreso per capita disponible en la comunidad ($X_2$). "

# Estimación de la RLM

- Ejemplo de una estimación por MCO en la RLM.

$$X = \begin{bmatrix} 1 & 68{,}5 & 16.7 \\ 1 & 45.2 & 16.8 \\ \dots & \dots & \dots \\ 1 & 52.3 & 16.0 \end{bmatrix} \qquad Y = \begin{bmatrix} 174.4 \\ 164.4 \\ \dots \\ 166.5 \end{bmatrix}$$

---

$$X'X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 68.5 & 45.2 & \dots & 52.3 \\ 16.7 & 16.8 & \dots & 16.0 \end{bmatrix} \begin{bmatrix} 1 & 68{,}5 & 16.7 \\ 1 & 45.2 & 16.8 \\ \dots & \dots & \dots \\ 1 & 52.3 & 16.0 \end{bmatrix} = \begin{bmatrix} 21.0 & 1302.4 & 360.0 \\ 1302.4 & 87707.9 & 22609.2 \\ 360.0 & 22609.2 & 6190.3 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 29.7289 & .0722 & -1.9926 \\ .0722 & .00037 & -.0056 \\ -1.9926 & -.0056 & .1363 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 68.5 & 45.2 & \dots & 52.3 \\ 16.7 & 16.8 & \dots & 16.0 \end{bmatrix} \begin{bmatrix} 174.4 \\ 164.4 \\ \dots \\ 166.5 \end{bmatrix} = \begin{bmatrix} 3820 \\ 249643 \\ 66073 \end{bmatrix}$$

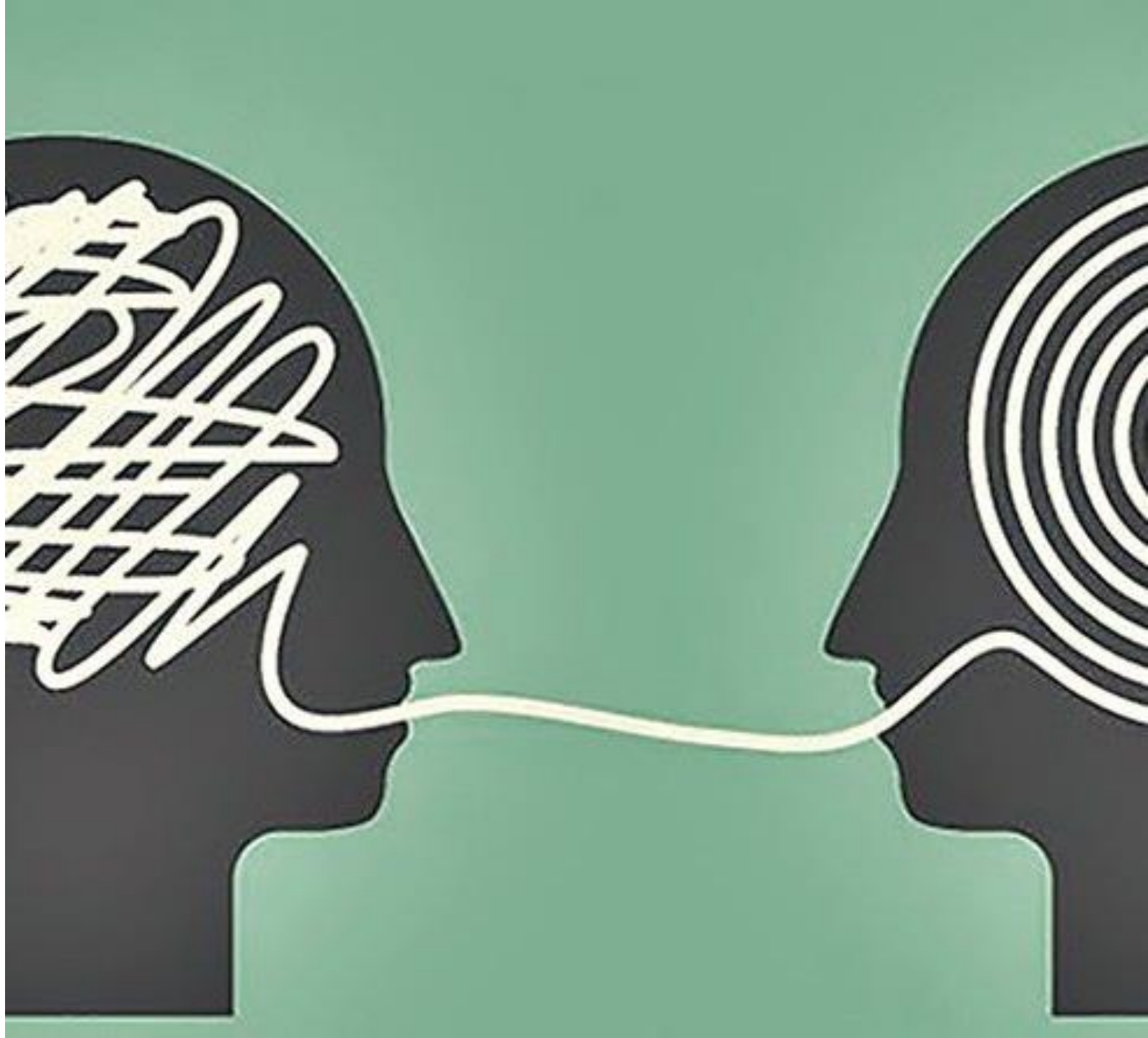
---

$$\hat{\beta} = (X'X)^{-1}(X'Y) = \begin{bmatrix} 29.7289 & .0722 & -1.9926 \\ .0722 & .00037 & -.0056 \\ -1.9926 & -.0056 & .1363 \end{bmatrix} \begin{bmatrix} 3820 \\ 249643 \\ 66073 \end{bmatrix} = \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix} \longrightarrow \hat{Y}_i = -68.857 + 1.455X_{i1} + 9.366X_{i2}$$

# Estimación de la RLM

- ¿Cómo interpretamos los resultados?

- Se espera que las ventas promedio aumenten $1,455,000 cuando la población meta aumenta mil personas de 16 años o menos, manteniendo constante el ingreso per cápita disponible.
- Similarmente se espera que las ventas promedio aumenten $9,366,000 cuando el ingreso per cápita disponible aumente mil dólares, manteniendo constante la población meta.



# Índice


1 Introducción

2 La RLM y sus condiciones o supuestos

3 Estimación de la RLM

4 Valores predichos y residuales

# Valores predichos y residuales

- Estimado la ecuación para la RLM, el análisis, tanto de los valores predichos como de sus residuales, sería la siguiente etapa.
- Los valores predichos son de utilidad para así poder predecir ciertas observaciones en $Y$. La ecuación para obtener el valore predicho estaría dada la ecuación por:

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

- El estudio de los residuos es fundamental para saber si nuestra RLM se está adecuando o no las condiciones del modelo. La ecuación para los valores residuales estaría dada por:

$$\varepsilon_i = Y - \hat{Y}$$

- Para el caso especial de la RLM, una forma bastante conveniente de estudiar los valores predichos y los residuals, es mediante la matriz $\boldsymbol{H}$.

# Valores predichos y residuales

- La matriz H está compuesta por elementos que son una combinación de todos los predictores. Se le llama matriz sombrero y a sus elementos se les llama influencia (leverage) pues se utilizarán para determinar que tanta influencia tiene una observación sobre los resultados de la regresión.
- H tiene la característica de ser simétrica e **idempotente**:

$$H = X\left(X'X\right)^{-1}X' \longrightarrow HH = X\left(X'X\right)^{-1}X'X\left(X'X\right)^{-1}X' = XI\left(X'X\right)^{-1}X' = H$$

- Para encontrar un valor ajustado (estimación particular de Y) basta usar el modelo de regresión con valores específicos de los predictores en X:

$$\hat{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1}$$

# Valores predichos y residuales

- Se quiere estimar las ventas para el estudio en la primera ciudad que tiene una población de 68500 jóvenes de 16 años o menos ($X_1$=68.5) y el ingreso per cápita disponible es de $16,700 ($X_1$=16.7). Se sustituyen los valores de $X_1$ y $X_2$ en la ecuación de regresión:

$$\hat{Y}_1 = -68.857 + 1.455X_{11} + 9.366X_{12}$$

$$\hat{Y}_1 = -68.857 + 1.455 \times 68.5 + 9.366 \times 16.7 = 187.2$$

- Aunque las ventas reales de esta sucursal fueron de $174,400, el modelo lo estima en $187,200. Se podría esperar que si hubiese muchas sucursales con estas características de población e ingreso, las ventas promedio serían de $187,200.

- Si se quisiera estimar las ventas que podría percibir un estudio en una nueva ciudad para la cual se conoce su población de jóvenes de 16 años o menos ($X_1$=65.4) y el ingreso per cápita disponible ($X_2$=17.6). También se sustituyen los valores de $X_1$ y $X_2$ en la ecuación de regresión:

$$\hat{Y} = -68.857 + 1.455 \times 65.4 + 9.366 \times 17.6 = 191.1$$

# Valores predichos y residuales

- Para encontrar las estimaciones de Y para todas las observaciones en la matriz X se puede multiplicar esta matriz de predictores X por el vector de coeficientes estimados:

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

- El vector de valores estimados se puede expresar en términos de la matriz H:

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

- Esta forma de expresar $\hat{Y}$ indica que la estimación de un valor particular de $Y$ puede verse como una combinación lineal de todas las respuestas. Así por ejemplo, si se quisiera estimar las ventas para todos los 21 estudios de la compañía:

$$\hat{Y} = X\hat{\beta} = \begin{bmatrix} 1 & 68{,}5 & 16.7 \\ 1 & 45.2 & 16.8 \\ \dots & \dots & \dots \\ 1 & 52.3 & 16.0 \end{bmatrix} \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix} \longrightarrow \hat{Y} = \begin{bmatrix} 187.2 \\ 154.2 \\ \dots \\ 157.1 \end{bmatrix}$$

# Valores predichos y residuales

- Los residuos son estimaciones de los errores y se calculan mediante las diferencias entre los valores observados y los estimados: $\varepsilon_i = Y_i - \widehat{Y}_i$ .
- La matriz de residuos se puede expresar en términos de la matriz H:

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

$$e = Y - X(X'X)^{-1}X'Y = \left[1 - X(X'X)^{-1}X'\right]Y$$

$$e = (1 - H)Y$$

- La matriz 1-H también es simétrica e idempotente. Se quiere calcular los residuos para todos los 21 estudios de la compañía:

$$e = \begin{bmatrix} 174.4 \\ 164.4 \\ \dots \\ 166.5 \end{bmatrix} - \begin{bmatrix} 187.2 \\ 154.2 \\ \dots \\ 157.1 \end{bmatrix} = \begin{bmatrix} -12.8 \\ 10.2 \\ \dots \\ 9.4 \end{bmatrix}$$

# Índice


1 Introducción

2 La RLM y sus condiciones o supuestos

3 Estimación de la RLM

4 Valores predichos y residuales

5 Inferencia de la RLM y sus coeficientes

# Inferencia de la RLM y sus coeficientes

- En la regresión lineal bivariada vimos que podemos hacer dos tipos de inferencias: prueba de hipótesis en los coeficientes, e intervalos de confianza de los coeficientes.
- Se cuál se lo que se quiera determinar, el proceso de realizar inferencias estadísticas parte del hecho de conocer, el valor del estimador (el o los $\beta$), y claro está, su variancia muestral o su error de muestreo muestral.
- Para esto, lo primero es conocer la suma de cuadrados de error (SCE):

$$SCE = e'e = Y'(1-H)'(1-H)Y = Y'(1-H)Y$$

- Luego, solemos estimar la variancia muestral:

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

# Inferencia de la RLM y sus coeficientes

- Para el presente caso, veamos el cálculo de la SCE:

$$SCE = e'\hat{e} = \begin{bmatrix} -12.8 & 10.2 & \dots & 9.4 \end{bmatrix} \begin{bmatrix} -12.8 \\ 10.2 \\ \dots \\ 9.4 \end{bmatrix} = 2180.93$$

- Y por lo tanto, podemos obtener la estimación de la variancia muestral como sigue:

$$\hat{\sigma}^2 = \frac{SCE}{n-p} = \frac{2180.93}{21-3} = 121.1626$$

- A partir de lo anterior, podemos obtener los valores de la matriz de variancia y covariancia de los coeficientes, y el error estándar de un coeficiente en particular.

# Inferencia de la RLM y sus coeficientes

- Si recordamos, un intervalo de confianza siempre tendrá la siguiente forma:

$IC = estimador \ \pm Nivel\ de\ confianza * error\ de\ muestreo$

$IC$ = [ Límite Interior ; Límite Superior]

- Para el caso de los coeficientes de los $\beta$, la estimación punctual, la matriz de variancia y covariancia de los coeficientes y el error estándar están dados por las formulas:

- Estimación punctual: $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$

- Matriz de variancia y covariancia de los coeficientes $\beta$ : $Var(\hat{\beta}) = \sigma^2 \left(X^T X\right)^{-1}$

- Error estándar de un coeficiente particular: $E.E.(\hat{\beta}_i) = \sqrt{\sigma^2 \left(X^T X\right)^{-1}_{ii}}$

# Inferencia de la RLM y sus coeficientes

- Si lo vemos desde nuestro ejemplo, tenemos que la variancia-covariancia de los coeficientes sería la siguiente:

$$Var(\hat{\beta}) = \sigma^2 \left(X'X\right)^{-1} = 121.1626 \begin{bmatrix} 29.7289 & .0722 & -1.9926 \\ .0722 & .00037 & -.0056 \\ -1.9926 & -.0056 & .1363 \end{bmatrix}$$

$$Var(\hat{\beta}) = \begin{bmatrix} 3602.0 & 8.748 & -241.43 \\ 8.748 & .0448 & -.679 \\ -241.43 & -.679 & 16.514 \end{bmatrix}$$

- Y por lo tanto, los errores estándar para los $\beta$ serían

Error estándar para $\hat{\beta}_1$: $\sqrt{0.0448} = 0.212$

Error estándar para $\hat{\beta}_2$: $\sqrt{16.514} = 4.06$

# Inferencia de la RLM y sus coeficientes

- Finalmente, el intervalo de confianza para un determinado coeficiente $\hat{\beta}_i$, se expresa de la forma:

$$\hat{\beta}_i \pm t_{\alpha/2,n-p} \times E.E.(\hat{\beta}_i)$$

- Para nuestro caso:

$$t_{\alpha/2,n-p} = t_{0.025,18} = 2.101$$

$$\hat{\beta}_1 \pm t_{0.025,18} \times E.E.(\hat{\beta}_1) = 1.455 \pm 2.101 \times 0.212 = [1.01, 1.90]$$

$$\hat{\beta}_2 \pm t_{0.025,18} \times E.E.(\hat{\beta}_2) = 9.366 \pm 2.101 \times 4.06 = [0.84, 17.9]$$


# Inferencia de la RLM y sus coeficientes

¿Y la prueba de hipótesis en los coeficientes de la RLM?

# Índice


1 Introducción

2 La RLM y sus condiciones o supuestos

3 Estimación de la RLM

4 Valores predichos y residuales

5 Inferencia de la RLM y sus coeficientes

6 Evaluación de la RLM

# Evaluación de la RLM

- En el proceso de estimar una RLM, nos podríamos preguntar: ¿es mi modelo estimado realmente bueno?
- La pregunta es MUY difícil de responder, pero un primer indicar que nos dice qué porcentaje de variancia explicada se determina por el modelo, es el Coeficientes de Determinación $R^2$.
- Para entender el significado del $R^2$, debemos tal vez antes explicar la partición de las sumas de cuadrados. La desviación total de una observación respecto a su media se puede descomponer en dos componentes:
  - Desviación del valor ajustado respecto a la media
  - Desviación de la observación respecto a la línea de regresión o valor ajustado.

$$Y_i - \overline{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})$$

Observado respecto al ajustado

Ajustado respecto a la media

# Evaluación de la RLM

- La suma de cuadrados total de la respuesta se puede descomponer en dos fuentes de variación:
  - La variación de la línea de regresión alrededor de la media denota la parte de la variabilidad de Y que está asociada con la línea de regresión (variabilidad explicada) y es medida por la SCR (suma de cuadrados de regresión).
  - La variación aleatoria que no logra ser explicada por las variables incluidas en el modelo y es medida por la SCE (suma de cuadrados de error).

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2$$

$$SCT \quad = \quad SCE \quad + \quad SCR$$

- El coeficiente de determinación múltiple es el porcentaje de la variancia total de la respuesta explicada conjuntamente por los predictores incluidos en el modelo de regresión

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

Porción no explicada

# Evaluación de la RLM

El $R^2$ es un indicador que va de 0-1, lo solemos pasar a porcentaje, pero ¿qué es lo que realmente este nos indica?

# Evaluación de la RLM

- El $R^2$, aunque útil, puede llegar a ser problemático, dado que podemos introducir 3413235244341312 3123 variables, pero esto lo que produce es que su valor prácticamente llegue a valer 1… se debe evaluar otra forma de ver la evaluación del modelo de la RLM.
- El $R^2$ se puede ajustar como una medida que toma en cuenta el número de predictores. El $R^2$ ajustado "castiga" a aquéllos modelos que tienen más variables y sólo los "premia" si ese aumento de variables va a compañado por una reducción importante en la SCE. El $R^2$ ajustado podría reducirse cuando se introduce un nuevo predictor en el modelo que contribuye a una disminución de la SCE.

$$R_a^2 = 1 - \frac{\frac{SCE}{n-p}}{\frac{SCT}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SCE}{SCT} = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2)$$

$$R_a^2 = 1 - \frac{\frac{SCE}{n-p}}{\frac{SCT}{n-1}} = 1 - \frac{CME}{S_Y^2}$$

# Conclusión

- En el presente capítulo de introducción a la RLM, se estudio los principios de la RLM:
  - La razón de aplicar la RLM.
  - Condiciones
  - Estimación de la RLM.
  - Valores predichos y residuales
  - Inferencia en la RLM
  - Evaluación de la RLM.

- Un análisis más detallado de una RLM, necesita hacer un análisis más extenso de las variables a seleccionar dentro de la RLM, ver si se cumple las condiciones, posibles medidas remediales, adecuar las medidas de bondad y de ajuste.

- Estos teams, serán estudiados en mayor detalle en las siguientes clases.


The End