I. Regresion Lineal Simple (Teoria)

1. Suponga que se plantea el modelo lineal de dos covariables de la forma:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Demuestre que las ecuaciones normales para encontrar los estimadores por mínimos cuadrados están dadas por:

$$\left.\begin{array}{l} \sum_{i=1}^n y_i = n\beta_0 + \beta_1 \sum_{i=1}^n x_{1i} + \beta_2 \sum_{i=1}^n x_{2i} \\ \sum_{i=1}^n y_i x_{1i} = \beta_0 \sum_{i=1}^n x_{1i} + \beta_1 \sum_{i=1}^n x_{1i}^2 + \beta_2 \sum_{i=1}^n x_{1i}x_{2i} \\ \sum_{i=1}^n y_i x_{2i} = \beta_0 \sum_{i=1}^n x_{2i} + \beta_1 \sum_{i=1}^n x_{1i}x_{2i} + \beta_2 \sum_{i=1}^n x_{2i}^2 \end{array}\right\}$$

2. Sea $\hat{\beta}_1$ y $\hat{\beta}_0$ los estimadores Máximo Verosimilies encontrados en el modelo de regresión lineal simple, demuestre que:

$$\mathrm{Var}\left(\hat{\beta}_0\right) = \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\sigma^2$$

3. Se tiene el modelo de regresión lineal simple $y = \beta_0 + \beta_1 x + \varepsilon$, con $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$ y tal que $cov(\varepsilon_i, \varepsilon_j) = 0$ con $i \neq j$ Demostrar que:

$$\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$$

$$\sum_{i=1}^n x_i e_i = 0 = \sum_{i=1}^n \hat{y}_i e_i$$

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\overline{x}\sigma^2}{S_{xx}}$$

$$Cov(\overline{y}, \hat{\beta}_1) = 0$$

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

$$\mathbb{E}\left(\sum_{i=1}^n (\hat{y}_i - \overline{y})^2\right) = \sigma^2 + \beta_1^2 S_{xx}$$

Donde $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ y e_i es el residual i, es decir $e_i = y_i - \hat{y}_i$

4. Demuestre la siguiente descomposición de suma de cuadrados:

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

Donde denotamos:

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=SCT=\text{Suma de cuadrados totales}$$

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=SCE=\text{Suma de cuadrados del error}$$

$$\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2=SCR=\text{Suma de cuadrados de la Regresion}$$

5. Se tiene el modelo de regresión lineal simple:

$$y=\beta_0+\beta_1x+\varepsilon\quad\varepsilon\sim N(0,\sigma^2)$$

Demuestre que la distribución del estimador máximo verosímil $\sigma^2_{M.V.}$ es Gamma y encuentre sus parámetros.

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

Hint: De por hecho que:

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

6. Se tiene el modelo de regresión lineal simple:

$$y=\beta_0+\beta_1x+\varepsilon\quad\varepsilon\sim N(0,\sigma^2)$$

Suponga que β_0 es un parámetro ya conocido del modelo por lo que el único parámetro a estimar es β_1

- Determinar el estimador de β_1 por máxima verosimilitud y por mínimos cuadrados.
- ¿El estimador por máxima verosimilitud es combinación lineal de las observaciones y_i?
- Encuentre la esperanza del estimador $\hat{\beta}_1$. ¿El estimador es insesgado?
- Encuentre la varianza del estimador $\hat{\beta}_1$. ¿Este estimador tiene menor varianza que la del estimador de β_1 cuando β_0 es también desconocido?
- Como se distribuye el estimador $\hat{\beta}_1$
- Determinar el estimador de σ^2

7. En el modelo de regresión lineal se define el coeficiente de determinación como $R^2 = \frac{SCR}{SCT}$, por otro lado se define el **Coeficiente de Correlación de Pearson** como :

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

   Demuestre entonces que:

$$r^2 = R^2$$

8. **Intervalo de Confianza para** β_0 . Considere el estimador máximo verosimil para β_0.

   - Demuestre que: $\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2\right)$
   - Asumiendo el hecho que $\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sigma^2} = \frac{(n-2)}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-2}$, demuestre que:

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim \tau_{(n-2)}$$

   - Observe que lo anterior es una cantidad pivotal, muestre entonces que un intervalo al $100(1-\alpha)\%$ de confianza para β_0 es

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

9. Para el modelo lineal simple mostramos que existen dos estimadores para σ^2:

$$\hat{\sigma}^2_{M.V} = \frac{\sum i = 1^n (y_i - \hat{y})^2}{n} \qquad \hat{\sigma}^2 = \frac{\sum i = 1^n (y_i - \hat{y})^2}{n-2}$$

   En términos del error cuadrático medio. ¿Cuál estimado es mejor?

I. Regresion Lineal Simple (Práctico)

1. La tabla FootballLeague.csv contiene los datos sobre el desempeño de los equipos de la liga nacional de fútbol de E.U.A. durante 1976. Se sospecha que el número de yardas ganadas por los oponentes (Variable x_8) tiene un efecto sobre el número de juegos ganados (Variable y).

   - Ajuste un modelo lineal simple que relacione el número de juegos ganados con el número de yardas ganadas por los oponentes, es decir ajuste el modelo:

$$y_i = \beta_0 + \beta_1 x_8 + \varepsilon$$

   - Construya el intervalo de confianza (α=0.05) para β_1. Verifique si el 0 está en el intervalo construido, dada la información obtenida ¿ Puede asegurar que el número de yardas ganadas tiene un efecto sobre el número de juegos ganados?

- Construya el intervalo de confianza (α=0.05) para β_0.
- Bajo el contexto de este problema, ¿Cuál sería la interpretación que se le daría a β_0?
- Construya un intervalo de confianza (α=0.05) para σ^2. (Punto Extra. Encuentre el intervalo de confianza con menor longitud)
- Construya la tabla ANOVA correspondiente y lleve a cabo la prueba de hipótesis

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

  Obtenga el p-value de la prueba. Tomando ($\alpha = 0.05$) ¿Rechazaría H_0?
- ¿Qué porcentaje de la variabilidad total de y es explicada por el modelo que ajustó?
- Encuentra un intervalo al 95% de confianza para $\mathbb{E}(Y \mid X = 2000)$.
- Encuentra un intervalo al 99% de confianza para el numero de juegos ganados para un equipo que logró que sus oponentes ganasen 2072 yardas

2. La siguiente tabla ANOVA fue obtenida por un paquete estadístico:
   `anova(lm(y x))`
   `Analysis of Variance Table`

   `Response: y`

Fnte. de Var.	Df.	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	20.107	20.1069	4.1248	0.05649
Residuals	19	92.618	4.8746		

   Se sabe a demás que $S_{xx} = 770$. Responda lo siguiente:

   - ¿Con cuantas observaciones se hizo el ajuste?
   - Tomando $\alpha = 0.05$. ¿Rechazaría la hipótesis $H_0 : \beta_1 = 0$?
   - De un estimador insesgado para σ^2
   - Construya un intervalo de confianza al 95% para σ^2
   - ¿Cuanto vale $S_{yy} := \sum_{i=1}^{n} (y_i - \overline{y})^2$?
   - ¿ Cuanto vale $\left|\hat{\beta}_1\right|$?
   - Calcule el error estándar del estimador de β_1 es decir encuentre $\sqrt{\frac{\hat{\sigma^2}}{S_{xx}}}$
   - ¿Qué porcentaje de la variabilidad es explicada por el modelo?

3. (Simulación) Considere el modelo de regresión lineal $y = 5 + 2x + \varepsilon$, donde

$$\varepsilon \sim NID(0, \sigma^2 = 9)$$

   Genere 1 000 muestras de este modelo, donde $x \in \{1, 1.5, 2, 2.5, 3, \ldots, 9.5, 10\}$ (n=19).

- Para cada muestra calcule los estimadores máximo verosímiles $(\hat{\beta}_0, \hat{\beta}_1)$ de tal forma que se obtengan 1 000 estimaciones para β_0 y 1 000 estimaciones para β_1 y elabore un par de histogramas.¿Los histogramas que obtiene son congruentes con la teoría que indica que ambos estimadores siguen una distribución normal ?

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

- Para cada muestra calcule un intervalo al 95% de confianza para β_1. ¿Cuantos de estos intervalos contiene al verdadero valor $\beta_1 = 2$?
- Para cada muestra, encuentre el estimador de $\mathbb{E}(y \mid x = 5)$. Construya un histograma. ¿El histograma que obtiene es congruente con la teoría que indica que el estimador siguen una distribución normal?
- Para cada muestra construya un intervalo de confianza al 95% para $\mathbb{E}(y \mid x = 5)$. ¿Cuántos de estos intervalos contiene al verdadero valor de $\mathbb{E}(y \mid x = 5) = 15$?