Regresión Lineal Múltiple (Teoria)

1. En clase probamos que para el modelo lineal múltiple:

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

El estimador por mínimos cuadrados y el de máxima verosimilitud está dado por:

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

En el caso lineal simple sabemos que:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}_{n\ x\ 2} \qquad \underline{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\ x\ 1} \qquad \underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}_{2\ x\ 1}$$

Demuestre entonces que:

$$\underline{\hat{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \overline{y} - \hat{\beta}_1\overline{x} \\ \frac{S_{xy}}{S_{xx}} \end{pmatrix}$$

2. (Propiedades de la matriz $\mathbf{H}$).

   (a) Pruebe que $\text{Var}\left(\underline{\hat{Y}}\right) = \sigma^2\mathbf{H}$

   (b) Para el caso de la regresión lineal simple $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Muestre que los elementos de la matriz H son:

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{xx}} \qquad h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}}$$

3. Sea $X \sim N(\mu, 1)$. Demuestre que entonces que:

$$X^2 \sim \chi^{2^{'}}_{1,\mu^2}$$

Donde $\chi^{2^{'}}_{1,\mu^2}$ es la distribución Chi-cuadrado no central con 1 grado de libertad y parámetro de no centralidad μ^2

4. Descomposición de la suma de cuadrados en la regresión lineal multiple:

   - En el modelos lineal mulltiple:

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

     Demuestre que: $\mathbf{HX} = \mathbf{X}$

   - Sean $\underline{x}_0, \ldots, \underline{x}_k$ las p columnas de la matriz de diseño $\mathbf{X}$ es decir:

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

     Demuestre que entonces:

$$\mathbf{H}\underline{x}_i = \underline{x}_i$$

- Verifique que en el modelo lineal multiple se cumple que:

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

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

Donde $e_i = y_i - \hat{y}_i$ es el residual i y $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_p x_{pi}$
Hint: Defina al vector columna de 1's como $\mathbf{1} = (1, 1, \ldots, 1)^T$, luego entonces observe que:

$$\sum_{i=1}^{n} e_i = \underline{e}^T \mathbf{1} \qquad \text{Con } \underline{e} = (e_1, e_2, \ldots, e_n)^T$$

$$\sum_{i=1}^{n} e_i \hat{y}_i = \underline{e}^T \underline{\hat{Y}} \qquad \text{Con } \underline{\hat{Y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^T$$

- Utilizando lo anterior verifique que:

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

5. Sea $X_1 \ldots, X_n$ m.a. del modelo $N(\mu, \sigma^2)$ demuestre que:

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} \left(X_i - \bar{X}\right) \sim \chi^2_{n-1}$$

6. Sea $X \sim \chi^{2'}_{k,\lambda}$ encuentre $VAR(X)$

7. El presente ejercicio tiene por objetivo construir el estadístico de prueba F para el contraste de hipótesis (Significancia de la Regresión):

$$H_0 : \beta_1 = \ldots = \beta_k = 0 \qquad vs \qquad H_1 : \beta_j \neq 0 \text{ p.a. } j \in \{1, \ldots, k\}$$

- Exprese a la suma de cuadrados de la regresón (SCR) como una forma cuadrática función del vector columna $\underline{y} = (y_1, \ldots, y_n)$, es decir encuentre la matriz $\mathbf{A}$ tal que :

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 = \underline{y}^T \mathbf{A} \underline{y}$$

- Demuestre que la matriz $\mathbf{A}$ que encontró es idempotente y simetrrica, ademas demuestre que su rango es k donde k es el número de variables explicativas en el modelo lineal multiple.

- Concluya entonces que bajo H_0

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

- En clase probamos que:

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

Demuestre que $\frac{1}{\sigma^2}SCR$ y $\frac{1}{\sigma^2}SCE$ son independientes (Hay que probar independencia de dos formas cuadráticas) y concluya que bajo H_0:

$$F=\frac{\frac{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}{(k)\sigma^2}}{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{(n-p))\sigma^2}}\sim F_{(k,n-p))}$$

8. En el ejercicio anterior se encontró un estadístico de prueba para contrastar la hipótesis:

$$H_0:\beta_1=\ldots=\beta_k=0\qquad vs\qquad H_1:\beta_j\neq 0\text{ p.a. }j\in\{1,\ldots,k\}$$

El presente ejercicio tiene como fin probar que el cociente de verosimilitudes generalizado (Neyman-Pearson) nos lleva a este mismo estadístico de prueba, lo que demuestra que el estadístico de prueba genera la región de rechazo óptima

- Bajo H_0 el modelo reducido es de la forma $y_i=\beta_0+\varepsilon_i$. Demuestre entonces que los estimadores que maximizan la verosimilitud bajo H_0 son:

$$\hat{\beta}_{0_{MV|H_0}}=\overline{y}\qquad\hat{\sigma}^2_{MV|H_0}=\frac{1}{n}\sum_{i}^{n}\left(y_i-\overline{y}\right)^2=\frac{1}{n}SCT$$

- Usando lo anterior demuestre entonces que:

$$\sup\mathscr{L}\left(\Theta_{H_0}\right)=\left(\frac{1}{\sqrt{2\pi\hat{\sigma}^2_{MV|H_0}}}\right)^n exp\left(-\frac{n}{2}\right)$$

- Bajo H_1 el modelo completo es de la forma $y_i=\beta_0+\beta_1x_{1i}+\ldots+\beta_kx_{ki}+\varepsilon_i$. En clase mostramos que los estimadores que maximizan la verosimilitud son:

$$\underline{\hat{\beta}}_{MV|H_1}=\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\underline{Y}\qquad\hat{\sigma}^2_{MV|H_1}=\frac{1}{n}\sum_{i}^{n}\left(y_i-\hat{y}_i\right)^2=\frac{1}{n}SCE$$

Pruebe entonces que:

$$\sup\mathscr{L}\left(\Theta_{H_1}\right)=\left(\frac{1}{\sqrt{2\pi\hat{\sigma}^2_{MV|H_1}}}\right)^n exp\left(-\frac{n}{2}\right)$$

- Con los puntos anteriores demuestre que el cociente de verosimilitudes generalizados toma la siguiente forma:

$$\Lambda = \frac{\sup \mathscr{L}\left(\Theta_{H_0}\right)}{\sup \mathscr{L}\left(\Theta_{H_1}\right)} \leq K \Leftrightarrow \frac{\hat{\sigma}^2_{MV|H_1}}{\hat{\sigma}^2_{MV|H_0}} \leq K^* \Leftrightarrow \frac{SCE}{SCT} \leq K^{**}$$

- Concluya con ayuda del punto anterior que la region de rechazo obtenida por el cociente de verosimilitudes generalizado es:

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

  Donde $(K, K^*, K^{**}, K^{***})$ son constantes que no dependen de las observaciones $(y_1, \ldots, y_n)$

9. (Regresión Ponderada). Suponga que tiene el siguiente modelo lineal:

$$\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon} \qquad \text{Con } \mathrm{Var}(\underline{\varepsilon}) = \sigma^2\mathbf{V}$$

Donde $\mathbf{V}$ es una matriz simétrica definida positiva tal que puede ser factorizada como $\mathbf{V} = \mathbf{K}\mathbf{K}$, con $\mathbf{K}$ una matriz simétrica no singular, es decir que existe $\mathbf{K}^{-1}$.
Ahora defina lo siguiente:

- $\underline{Z} = \mathbf{K}^{-1}\underline{Y}$
- $\mathbf{A} = \mathbf{K}^{-1}\mathbf{X}$
- $\underline{\delta} = \mathbf{K}^{-1}\underline{\varepsilon}$

(a) Observe que cuando transformamos el modelo original $\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$ multiplicando por la izquierda por la matriz $\mathbf{K}^{-1}$ obtenemos el modelo:

$$\underline{Z} = \mathbf{A}\underline{\beta} + \underline{\delta}$$

Prueba que este nuevo modelo transformado cumple con que $\mathbb{E}(\underline{\delta}) = \underline{0}$ y $\mathrm{Var}(\underline{\delta}) = \sigma^2\mathbb{I}$

(b) Obtenga $\hat{\underline{\beta}}$, el vector de estimadores de $\underline{\beta}$, en términos de las matrices $\underline{Z}$ y $\mathbf{A}$, utilizando el método de mínimos cuadrados, es decir, encuentre $\underline{\beta}$ tal que haga mínima la siguiente expresión :

$$\left\{\left(\underline{Z} - \mathbf{A}\underline{\beta}\right)^T\left(\underline{Z} - \mathbf{A}\underline{\beta}\right)\right\}$$

(c) Haga las transformaciones correspondientes para obtener a $\hat{\underline{\beta}}$ en términos de las matrices $\underline{Y}$ y $\mathbf{X}$

(d) Suponga ahora que en el modelo lineal:

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

Se tienen los siguientes supuestos: $\mathrm{Var}(\varepsilon_i) = \frac{\sigma^2}{w_i}$ (No hay homocedasticidad) y que $Cov(\varepsilon_i, \varepsilon_j) = 0$ con $i \neq j$.

- Encuentre la matriz $\mathbf{V}$ tal que $\text{Var}(\underline{\varepsilon}) = \sigma^2\mathbf{V}$
- Encuentre la matriz $\mathbf{K}$ tal que $\mathbf{V} = \mathbf{KK}$
- Utilizando los incisos anteriores encuentre el estimador por mínimos cuadrados para $\underline{\beta}$ en términos de $\underline{Y}$ y de $\mathbf{X}$

Regresión Lineal Multiple (Pactica)

10. Utilizando la teora de modelos lineales, encuentre la ecuación de la parabola $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ que pasa por los puntos:

$$(0, 1), (1, 6), (2, 17)$$

(Ver figura 1)

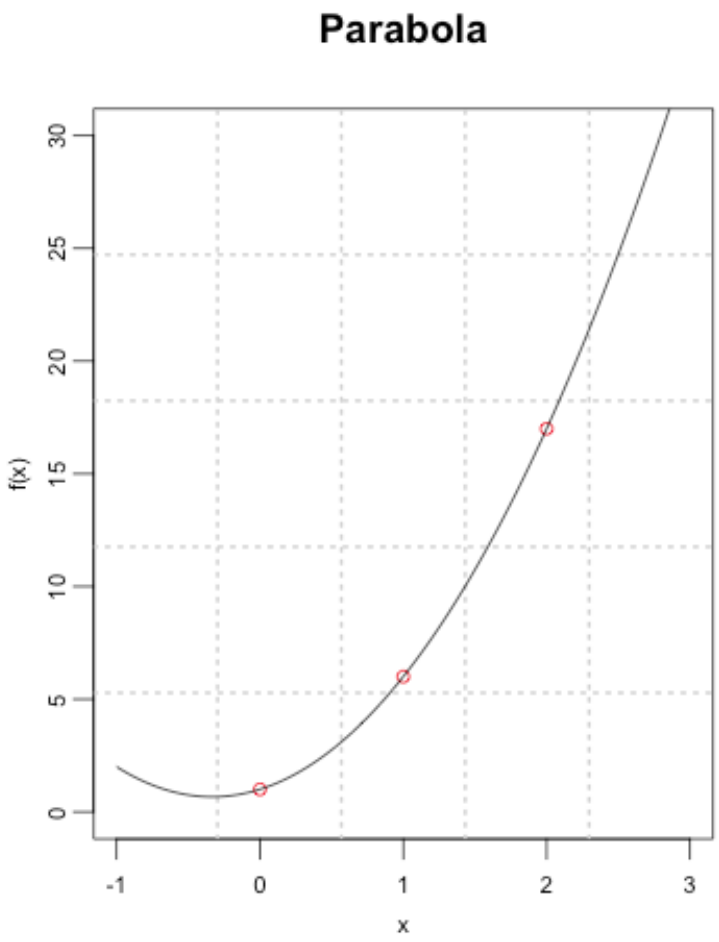

Figure 1: Ajuste de Parabola

11. Complete la siguiente tabla ANOVA:
`anova(lm(y x))`

```
Analysis of Variance Table

Response:  y
```

Fnte. de Var.	Df.	Sum Sq	Mean Sq	F value	Pr(>F)
x	3		1600.81		< 2.2e-16 ***
Residuals	36	146.9			

Responda lo siguiente

- ¿Con cuantas observaciones se hizo el ajuste?
- ¿Con cuantas variables se hizo el ajuste?
- Tomando $\alpha = 0.01$. ¿Rechazaría la hipótesis H_0?
- De un estimador insesgado para σ^2
- De el estimador Máximo Verosimil para σ^2
- Construya un intervalo de confianza al 95% para σ^2
- ¿Cuanto vale $S_{yy} := \sum_{i=1}^{n} (y_i - \overline{y})^2$?
- ¿Qué porcentaje de la variabilidad es explicada por el modelo?

12. 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.

- Ajuste un modelo lineal multiple que relaciona el número de juegos ganados con
  - Yardas por aire del equipo (x_2)
  - El porcentaje de Yardas por Tierra (x_7)
  - Las Yardas por tierra del contrario (x_8)

$$y_i = \beta_0 + \beta_1 x_2 + \beta_2 x_7 + \beta_3 x_8 + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$$

- Construya la tabla ANOVA y haga la prueba de significancia de la regresión. ¿ Si $\alpha = 0.05$, rechazaría H_0?
- Calcule R^2 y R^2 ajustado
- Contraste la prueba de hipótesis:

$$H_0 : \beta_1 = \beta_3 = 0 \qquad H_1 : \beta_1 \neq 0 \text{ o } \beta_3 \neq 0$$

- Calcule el vector de valores ajustados por el modelo $\underline{\hat{Y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$
- Calcule el coeficiente de correlación lineal de Pearson entre y_i, $\hat{y}_i$.
- Verifique el cuadrado del coeficiente de correlación lineal de Pearson y R^2 (Coeficiente de Determinación) coinciden
- Calcule el vector de residuales $\underline{e} = (\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n)$, donde $e_i = y_i - \hat{y}_i$
- Verifique por medio de pruebas estadísticas los siguiente:

  - Normalidad de los residuales (Prueba ShapiroWilk, Anderson Darling, Prueba Lilliefors)
  - Homocedasticidad de los residuales (Prueba Levene, Prueba Barttlet)
  - Independencia de los Residuales (ACF, Prueba de Rachas)
- Encuentre los intervalos al 98% de confianza para cada uno de los parámetros $\beta_0, \beta_1, \beta_2, \beta_3$
- Encuentre el intervalo al 92% de confianza para la respuesta media de numero juegos ganados cuando $x_2 = 2300$, $x_7 = 56$ y $x_8 = 2100$
- Encuentre el intervalo al 93% de confianza para el numero de juegos ganados (nueva observación) cuando $x_2 = 2100$, $x_7 = 60$ y $x_8 = 2000$

13. Considere el siguiente modelo lineal múltiple:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i \qquad i \in \{1, 2, \ldots, 40\}$$

, donde $\underline{\varepsilon} \sim N_n(\underline{0}, \sigma^2 \mathbb{I})$.
Se ajustó el modelo anterior utilizando un paquete estadístico y arrojó la siguiente TABLA ANOVA:

Fnte. de Var.	Df.	Sum Sq	Mean Sq	F value	Pr(>F)
x		18053.2			< 2.2e-16 ***
Residuals					
Total		18876.6			

(2 Puntos) Complete la Tabla ANOVA y responda lo siguiente:

- ¿Cuál es la hipótesis nula que se contrasta en esta tabla ANOVA?
- Tomando $\alpha = 0.01$. ¿Rechazaría la hipótesis nula anterior? (Justifique su respuesta)
- Proporcione el valor de la estimación de σ^2 por Máxima Verosimilitud
- ¿Cuánto vale $\sum_{i=1}^{n} (y_i - \overline{y})^2$ y $\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$?
- ¿Qué porcentaje de la variabilidad es explicada por el modelo completo?
- ¿Cuánto vale R^2 ajustada del modelo completo?

El paquete estadístico también arrojó la siguiente información:

Coefficients	Estimate	Std Error	T value	Pr(>|t|)
(Intercept)	4.18132	2.71979	1.537	0.1332
X_1	1.03215	0.06223	16.586	<2e-16
X_2	-0.11146	0.06241	-1.786	0.0828
X_3	-0.08882	0.05222	-1.701	0.0979
X_4	1.04013	0.05106	20.371	<2e-16

(1 Punto) Tomando $\alpha = 0.04$, responda lo siguiente (Justifique su respuesta):

- ¿Rechazaría la hipótesis $H_0 : \beta_0 = 0$ vs $H_0 : \beta_0 \neq 0$?

- ¿Rechazaría la hipótesis $H_0 : \beta_2 = 0$ vs $H_0 : \beta_2 \neq 0$?
- ¿Rechazaría la hipótesis $H_0 : \beta_4 = 0$ vs $H_0 : \beta_4 \neq 0$?

Con los mismos datos, se procedió ahora ajustar un modelo reducido eliminando las variables X_2 y X_3

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_4 x_{4i} + \varepsilon_i \qquad i \in \{1, 2, \ldots, 40\}$$

El paquete estadístico arrojó la siguiente TABLA ANOVA para este modelo reducido:

Fnte. de Var.	Df.	Sum Sq	Mean Sq	F value	Pr(>F)
x		17912.1			< 2.2e-16 ***
Residuals					
Total		18876.6			

(2 Puntos) Complete la Tabla ANOVA y responda lo siguiente:

- ¿Cuál es la hipótesis nula que se contrasta en esta tabla ANOVA?
- Tomando $\alpha = 0.01$. ¿Rechazaría la hipótesis nula? (Justifique su respuesta)
- De el estimador insesgado para σ^2 de este modelo reducido
- Calcule el estadístico de prueba para contrastar la hipótesis:

$$H_0 : \beta_2 = \beta_3 = 0 \qquad H_1 : \beta_2 \neq 0 \quad o \quad \beta_3 \neq 0$$

- ¿Tomando $\alpha = 0.05$ rechazaría la hipótesis anterior?
- ¿Cuál modelo elegiría: el modelo completo o el modelo reducido? (Justifique)