Modelos Estadísticos I **(Fecha: 24/02/2021)**

Tarea #2

Estudiante: Roberto Vásquez Martínez *NUA:* 424662

Problema 1

(a) Pruebe que las estimaciones de mínimos cuadrados $a_1, \ldots, a_m, b$ de los parámetros $\alpha_1, \ldots, \alpha_n, \beta$ de la familia de rectas

$$\mathbb{E}[Y_i] = \alpha_i + \beta X_i,\ i = 1, 2, \ldots, m,$$

Están dados por

$$b = \frac{\sum_{i=1}^{m}\sum_{u=1}^{n_i}(X_{iu} - \overline{X_i})(Y_{iu} - \overline{Y_i})}{\sum_{i=1}^{m}\sum_{u=1}^{n_i}(X_{iu} - \overline{X_i})^2},$$

y

$$a_i = \overline{Y_i} - b\overline{X_i},$$

donde $(X_{i1}, Y_{i1}), \ldots, (X_{in_i}, Y_{in_i})$ denotan los valores observados de (X_i, Y_i) relacionados con la i-ésima recta, $i = 1, 2, \ldots, m$.

(Solución)
Para cada $i = 1, \ldots, m$ denotamos las medias muestrales como

$$\overline{X_i} = \frac{1}{n_i}\sum_{u=1}^{n_i} X_{iu} \quad \text{y} \quad \overline{Y_i} = \frac{1}{n_i}\sum_{u=1}^{n_i} Y_{iu}.$$

Como buscamos los estimadores de mínimos cuadrados basta minimizar la suma de los cuadrados de los errores sobre todas las rectas, esto es

$$(b, a_1, \ldots, a_m) = \underset{(\beta,\alpha_1,\ldots,\alpha_n)\subset\mathbb{R}^{m+1}\setminus\{0\}}{\operatorname{argmin}} \{\mathcal{L}(\beta, \alpha_1, \ldots, \alpha_m)\}$$

donde

$$\mathcal{L}(\beta, \alpha_1, \ldots, \alpha_m) = \sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu} - a_i - bX_{iu})^2.$$

Derivando respecto a cada variable e igualando a 0 obtenemos el siguiente sistema de ecuaciones

$$\sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu} - a_i - bX_{iu})X_{ui} = 0 \tag{1.1}$$

y

$$\sum_{u=i}^{n_i}(a_i + bX_{iu} - Y_{iu}) = 0 \qquad i = 1, 2, \ldots, m. \tag{1.2}$$

De (1.2) obtenemos que para cada $i = 1, 2, \ldots, m$

$$a_i = \overline{Y_i} - b\overline{X_i}. \tag{1.3}$$

Sustituyendo en (1.1) obtenemos que esta ecuación es equivalente a

$$\sum_{i=1}^{m}\sum_{u=1}^{n_i}(\overline{Y_i} - Y_{iu})X_{iu} + b(X_{iu} - \overline{X_i})X_{iu} = 0.$$

Despejando b se tiene que

$$b = \frac{\sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu} - \overline{Y_i})X_{iu}}{\sum_{i=1}^{m}\sum_{u=1}^{n_i}(X_{iu} - \overline{X_i})X_{iu}}. \tag{1.4}$$

Observamos que

$$\begin{aligned}\sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{ui} - \overline{Y_i})\overline{X_i} &= \sum_{i=1}^{m}\overline{X_i}\sum_{u=1}^{n_i}Y_{iu} - \sum_{i=1}^{m}\sum_{u=1}^{n_i}\overline{X_i}\cdot\overline{Y_i} \\ &= \sum_{i=1}^{m} n_i\overline{X_i}\cdot\overline{Y_i} - \sum_{i=1}^{m} n_i\overline{X_i}\cdot\overline{Y_i} \\ &= 0.\end{aligned}$$

De manera análoga

$$\sum_{i=1}^{m}\sum_{u=1}^{n_i}(X_{ui} - \overline{X_i})\overline{X_i} = 0.$$

Restando 0's de esta forma en el numerador y denominador de (1.4) tenemos que

$$b = \frac{\sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu} - \overline{Y_i})(X_{iu} - \overline{X_i})}{\sum_{i=1}^{m}\sum_{u=1}^{n_i}(X_{iu} - \overline{X_i})^2},$$

este y los estimadores en (1.3) son los estimadores de mínimos cuadrados.

(b) Pruebe además que la suma de cuadrados residual esta dada por

$$s^2 = \sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu} - \overline{Y_i})^2 - b^2\sum_{i=1}^{m}\sum_{u=1}^{n_i}(X_{iu} - \overline{X_i})^2.$$

(Solución)
Por definición tenemos que

$$s^2 = \sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu} - \widehat{Y_{iu}})^2,$$

donde $\widehat{Y_{iu}} = a_i + bX_{iu}$.

Por otro lado usando la identidad de diferencia de cuadrados y recordando (1.3) obtenemos

$$\begin{aligned}\sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu}-\overline{Y_i})^2 - b^2\sum_{i=1}^{m}\sum_{u=1}^{n_i}(X_{iu}-\bar{X}_i)^2 &= \sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu}-\overline{Y_i})^2 - b^2(X_{iu}-\overline{X_i})^2\\ &= \sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu}-\overline{Y_i}+bX_{iu}-b\overline{X_i})(Y_{iu}-\overline{Y_i}-bX_{iu}+b\overline{X_i})\\ &= \sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu}-a_i-bX_{iu})(Y_{iu}-a_i-bX_{iu}+2b(X_{iu}-\overline{X_i}))\\ &= \sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu}-\widehat{Y_{iu}})^2 + 2b\sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu}-\widehat{Y_{iu}})(X_{iu}-\overline{X_i}).\end{aligned}$$

Observamos que

$$(Y_{iu}-\widehat{Y_{iu}}) = (Y_{iu}-\overline{Y_i}) - b(X_{iu}-\overline{X_i}).$$

Por lo tanto

$$b\sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu}-\widehat{Y_{iu}})(X_{iu}-\overline{X_i}) = b\sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu}-\overline{Y_i})(X_{iu}-\overline{X_i}) - b^2\sum_{i=1}^{m}\sum_{u=1}^{n_i}(X_{iu}-\overline{X_i})^2.$$

Sustituyendo el valor de b que obtuvimos en el inciso (a) obtenemos

$$\begin{aligned}b\sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu}-\widehat{Y_{iu}})(X_{iu}-\overline{X_i}) &= \frac{\left[\sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu}-\overline{Y_i})(X_{iu}-\overline{X_i})\right]^2}{\sum_{i=1}^{m}\sum_{u=1}^{n_i}(X_{iu}-\overline{X_i})^2} - \frac{\left[\sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu}-\overline{Y_i})(X_{iu}-\overline{X_i})\right]^2}{\sum_{i=1}^{m}\sum_{u=1}^{n_i}(X_{iu}-\overline{X_i})^2}\\ &= 0\end{aligned}$$

Concluimos así que

$$\begin{aligned}s^2 &= \sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu}-\widehat{Y_{iu}})^2\\ &= \sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu}-\overline{Y_i})^2 - b^2\sum_{i=1}^{m}\sum_{u=1}^{n_i}(X_{iu}-\bar{X}_i)^2,\end{aligned}$$

que es lo que queríamos verificar.

□

Problema 2

La mayoría de los resultados se obtienen de un código en Python que se anexa a la tarea.

(a) La gráfica de los datos es la siguiente

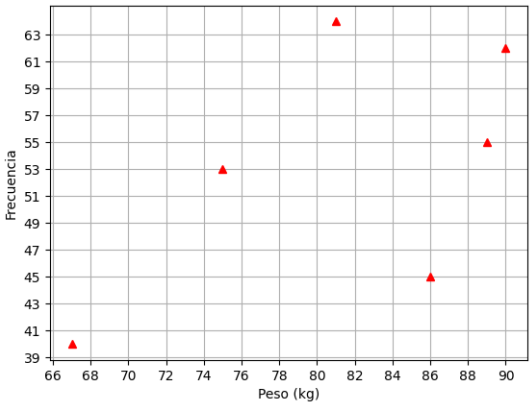

Figura 1: Relación entre peso corporal y frecuencia cardiaca

Parece que puede haber una relación lineal, pues parece que la frecuencia es directamente proporcional al peso corporal.

(b) Los estimadores que definen la recta de regresión son

$$\widehat{\beta}_0 = 4.7990,$$

y

$$\widehat{\beta}_1 = 0.5947.$$

La gráfica que muestra el ajuste de la regresión lineal es

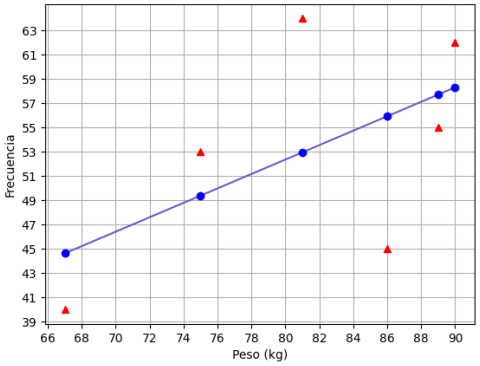

Figura 2: Ajuste lineal a los datos de frecuencia cardiaca

Los resultados muestran que la frecuencia cardiaca aumenta en 0.6 latidos por minuto por kilogramo de peso. Sin embargo el intercepto nos dice que hay pulsaciones a pesar de que no haya peso lo cual no tiene mucho sentido. En este contexto sería más apropiado una regresión por el origen.

(c) Repitiendo el ajuste sin el dato $(67, 40)$ obtenemos

$$\widehat{\beta}_0 = 49.1641 \quad \text{y} \quad \widehat{\beta}_1 = 0.0788,$$

que representa una diferencia grande en comparación al ajuste anterior.

(d) La estimación puntual de la respuesta media en $X_0 = 88$ la obtenemos sustituyendo ese valor en la recta de regresión lo que sería

$$Y_0 = 57.1312.$$

Y el intervalo de confianza de nivel $\alpha = 0.05$ en este caso es

$$(Y_0 + t_{\alpha/2,4}s(Y_0), Y_0 + t_{1-\alpha/2,4}s(Y_0)),$$

donde $s(Y_0)$ es la desviación estándar de la estimación, que como vimos en clase es

$$s(Y_0) = \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})}{\sum x_i^2}}s^2,$$

con $n = 6$ y $s^2 = MS(Res)$ el estimador insesgado de σ^2.
Numéricamente el intervalo de confianza es

$$(44.5325, 69.7299).$$

(e) El predictor es el mismo

$$Y_0 = 57.1312,$$

sin embargo, la desviación del predictor viene dada por

$$s(Y_0) = \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})}{\sum x_i^2}},$$

por lo que realizando los cálculos el intervalo de confianza en este caso es

$$(30.0959, 84.1665),$$

que al ser la variabilidad mayor en este caso el intervalo es de mayor longitud que en el inciso anterior.

(f) Como vimos en los últimos dos incisos la variabilidad es menor cuando más cerca se esta de la media. La media de esta muestra es aproximadamente 81, por lo que la menor varianza para Y_0 se tendría con $X_0 = 81$.

□

Problema 3

Use los datos y la ecuación de regresión del Ejercicio 2 y calcule el valor de $\hat{Y}_i$ para cada valor de X. Calcule las siguientes correlaciones

(a) Denotemos por $\hat{\rho}(X, Y)$ la correlación muestral entre esas variables, con los datos del ejercicio anterior y haciendo los cálculos correspondientes en el código llegamos a que

$$\hat{\rho}(X, Y) = 0.5687.$$

(b) Aquí obtenemos

$$\hat{\rho}(Y, \hat{Y}) = 0.5687.$$

(c) Por último, en este caso tenemos

$$\hat{\rho}(X, \hat{Y}) = 1$$

Además se tiene que el coeficiente de determinación es

$$R^2 = 0.3234,$$

y se nota que

$$R^2 = \hat{\rho}(X, Y)^2,$$

que se deduce fácilmente del siguiente desarrollo y la definición del estimador $\hat{\beta}_1$

$$\begin{aligned} R^2 =& \frac{SS(Reg)}{\sum_{i=1}^n (Y_i - \overline{Y})^2} \\ =& \frac{\hat{\beta}_1^2 \sum_{i=1}^n (X_i - \overline{X})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} \\ =& \frac{\left[\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})\right]^2}{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2} \\ =& \hat{\rho}(X, Y)^2. \end{aligned}$$

Se puede ver que la igualdad entre las correlaciones de (a) y (b) se sigue del hecho de que $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ y las siguientes cuentas

$$\begin{aligned} \rho(Y, \hat{Y}) =& \frac{\sum_{i=1}^n (Y_i - \overline{Y})(\hat{\beta}_1 X_i - \hat{\beta}_1 \overline{X})}{\left(\sum_{i=1}^n (Y_i - \overline{Y})^2\right)^{1/2} \left(\sum_{i=1}^n (\hat{\beta}_1 X_i - \hat{\beta}_1 \overline{X})^2\right)^{1/2}} \\ =& \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\left(\sum_{i=1}^n (X_i - \overline{X})^2\right)^{1/2} \left(\sum_{i=1}^n (Y_i - \overline{Y})^2\right)^{1/2}} \\ =& \hat{\rho}(X, Y). \end{aligned}$$

Finalmente se tiene que $\rho(X, \hat{Y}) = 1$ pues $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, entonces claramente las variables están relacionadas linealmente.

□

Problema 4

Pruebe las siguientes relaciones

$$SS(modelo) = n\overline{Y}^2 + \hat{\beta}_1^2 \sum_i (X_i - \overline{X})^2 \tag{1.5}$$

y

$$\sum_i (Y_i - \hat{Y}_i)^2 = \sum_i (Y_i - \overline{Y})^2 - \hat{\beta}_1^2 \sum_i (X_i - \overline{X})^2 \tag{1.6}$$

(Solución)

Primero verifiquemos (1.5). Por definición

$$SS(modelo) = \sum_{i=1}^{n} (\hat{Y}_i)^2,$$

donde $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ y

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

.

De lo anterior y una sucesión de manipulaciones algebraicas obtenemos

$$\begin{aligned} SS(modelo) =& \sum_{i=1}^{n} (\overline{Y} + \hat{\beta}_1 (X_i - \overline{X}))^2 \\ =& n\overline{Y}^2 + 2\hat{\beta}_1 \overline{Y} \sum_{i=1}^{n} (X_i - \overline{X}) + \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2, \end{aligned}$$

y como

$$\sum_{i=1}^{n} (X_i - \overline{X}) = 0,$$

entonces

$$SS(modelo) = n\overline{Y}^2 + \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2,$$

lo que verifica (1.5).

Ahora veamos que se cumple (1.6).

De la identidad de diferencia de cuadrados tenemos

$$\sum_{i=1}^{n} (Y_i - \overline{Y}) - \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} (Y_i - \overline{Y} + \hat{\beta}_1 \overline{X} - \hat{\beta}_1 X_i)(Y_i - \overline{Y} - \hat{\beta}_1 \overline{X} + \hat{\beta}_1 X_i).$$

Como $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$ y $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ entonces

$$\sum_{i=1}^{n} (Y_i - \overline{Y}) - \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)(Y_i - \overline{Y} - \hat{\beta}_1 \overline{X} + \hat{\beta}_1 X_i).$$

Sumando $\hat{\beta}_1\overline{X}$ y restando $\hat{\beta}_1 X_i$ al segundo factor de cada sumando de la suma anterior, y luego restando y sumando lo correspondiente para no alterar la igualdad y después de agrupar algunos términos y hacer manipulaciones algebraicas tenemos

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\overline{Y})-\hat{\beta}_1^2\sum_{i=1}^{n}(X_i-\overline{X})^2 &= \sum_{i=1}^{n}(Y_i-\hat{Y}_i)(Y_i-\hat{Y}_i-2\hat{\beta}_1\overline{X}+2\hat{\beta}_1 X_i)\\ &= \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2+2\hat{\beta}_1\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(X_i-\overline{X}).\end{aligned} \tag{1.7}$$

Observamos que

$$(Y_i-\hat{Y}_i)=(Y_i-\overline{Y})-\hat{\beta}_1(X_i-\overline{X}).$$

Por lo tanto haciendo algunos cálculos tenemos

$$\hat{\beta}_1\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(X_i-\overline{X}) = \hat{\beta}_1\sum_{i=1}^{n}(Y_i-\overline{Y})(X_i-\overline{X})-\hat{\beta}_1^2\sum_{i=1}^{n}(X_i-\overline{X})^2. \tag{1.8}$$

Como

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(Y_i-\overline{Y})(X_i-\overline{X})}{\sum_{i=1}^{n}(X_i-\overline{X})^2},$$

entonces sustituyendo en (1.8) obtenemos

$$\hat{\beta}_1\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(X_i-\overline{X})=\frac{\left[\sum_{i=1}^{n}(Y_i-\overline{Y})(X_i-\overline{X})\right]^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}-\frac{\left[\sum_{i=1}^{n}(Y_i-\overline{Y})(X_i-\overline{X})\right]^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}=0.$$

De lo anterior y (1.7) concluimos

$$\sum_{i=1}^{n}(Y_i-\overline{Y})-\hat{\beta}_1^2\sum_{i=1}^{n}(X_i-\overline{X})^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2,$$

lo que verifica (1.6) y concluye el problema.

□

Problema 5

Pruebe algebraicamente que cuando la ecuación de regresión lineal simple contiene el término intercepto se tiene que $\sum e_i = 0$. Muestre algebraicamente que esto no es cierto cuando la regresión lineal simple no incluye al intercepto

(Solución)

Primero consideremos la regresión con intercepto. Notamos que

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i).$$

Como $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ y $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$ entonces

$$\begin{aligned}\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n} \left(Y_i - \overline{Y} - \hat{\beta}_1 (X_i - \overline{X}) \right) \\ &= \sum_{i=1}^{n} (Y_i - \overline{Y}) + \hat{\beta}_1 \sum_{i=1}^{n} (X_i - \overline{X}).\end{aligned}$$

Notamos que

$$\sum_{i=1}^{n} (Y_i - \overline{Y}) = 0,$$

y

$$\sum_{i=1}^{n} (X_i - \overline{X}) = 0,$$

de lo que se sigue

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

que es lo que queríamos ver.

Ahora supongamos el modelo de regresión simple sin intercepto, por lo que

$$\hat{Y}_i = \hat{\beta}_1 X_i,$$

con

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.$$

Por lo tanto

$$\begin{aligned}\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n} Y_i - \hat{\beta}_1 X_i \\ &= \frac{\left(\sum_{i=1}^{n} Y_i\right)\left(\sum_{i=1}^{n} X_i^2\right) - \left(\sum_{i=1}^{n} X_i Y_i\right)\left(\sum_{i=1}^{n} X_i\right)}{\sum_{i=1}^{n} X_i^2},\end{aligned}$$

y en general

$$\left(\sum_{i=1}^{n} Y_i\right)\left(\sum_{i=1}^{n} X_i^2\right) \neq \left(\sum_{i=1}^{n} X_i Y_i\right)\left(\sum_{i=1}^{n} X_i\right),$$

pues para que se de la igualdad deberíamos tener

$$\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} X_i Y_i X_j = 0,$$

pero si todos las variables son positivas claramente esto no se cumple.
Por lo tanto se tiene en general que

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

□

Problema 6

Use las ecuaciones normales del modelo de regresión simple para probar que

(a) $\sum_{i=1}^{n} X_i Y_i = \sum_{i=1}^{n} X_i \hat{Y}_i$

Las ecuaciones normales del modelo de regresión simple considerando el intercepto son

$$n\hat{\beta}_0 + \left(\sum_{i=1}^{n} X_i\right)\hat{\beta}_1 = \sum_{i=1}^{n} Y_i \tag{1.9}$$

$$\left(\sum_{i=1}^{n} X_i\right)\hat{\beta}_0 + \left(\sum_{i=1}^{n} X_i^2\right)\hat{\beta}_1 = \sum_{i=1}^{n} X_i Y_i \tag{1.10}$$

Como $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, de (1.10) se sigue que

$$\begin{aligned}\sum_{i=1}^{n} X_i \hat{Y}_i &= \sum_{i=1}^{n} X_i(\hat{\beta}_0 + \hat{\beta}_1 X_i) \\ &= \left(\sum_{i=1}^{n} X_i\right)\hat{\beta}_0 + \left(\sum_{i=1}^{n} X_i^2\right)\hat{\beta}_1 \\ &= \sum_{i=1}^{n} X_i Y_i.\end{aligned}$$

Si tenemos el modelo sin intercepto entonces

$$\hat{Y}_i = \hat{\beta}_1 X_i \quad \forall\ i = 1, 2, \ldots, n,$$

con

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2},$$

por lo que en este caso

$$\begin{aligned}\sum_{i=1}^{n} X_i \hat{Y}_i &= \hat{\beta}_1 \sum_{i=1} X_i^2 \\ &= \sum_{i=1}^{n} X_i Y_i,\end{aligned}$$

por lo que en este caso también se cumple el resultado.

(b) $\sum_{i=1}^{n} X_i e_i = 0.$

Del hecho $e_i = Y_i - \hat{Y}_i$ y el inciso (a) se sigue que

$$\begin{aligned}\sum_{i=1}^{n} X_i e_i &= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \hat{Y}_i \\ &= 0,\end{aligned}$$

que se cumple en el modelo con intercepto y en la regresión por el origen.

(c) $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$

Si consideramos el modelo con intercepto entonces

$$\sum_{i=1}^{n} \hat{Y}_i e_i = \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i e_i,$$

del inciso (b) de este problema y por el problema 5 tenemos $\sum_{i=1}^{n} X_i e_i = 0$ y $\sum_{i=1}^{n} e_i = 0$, por lo que

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

Si consideramos la regresión por el origen entonces

$$\sum_{i=1}^{n} \hat{Y}_i e_i = \hat{\beta}_1 \sum_{i=1}^{n} X_i e_i = 0,$$

por el inciso (b) de este problema, y esto es lo que queríamos verificar.

□

Problema 7

Obtenga las ecuaciones normales y las estimaciones mínimo cuadráticas para el modelo

$$Y_i = \mu + \beta_1 x_i + \epsilon_i,$$

donde $x_i = (X_i - \overline{X})$. Compare estos resultados con los que se obtienen en el modelo de regresión lineal usual. El modelo aquí propuesto se conoce como el modelo "centrado", porque los valores de la variable independiente se desplazan para tener media cero.

(Solución)

Notamos que si $\overline{x}$ es la media muestral de las x_i entonces

$$\begin{aligned}\overline{x} =& \frac{1}{n}\sum_{i=1}^{n} x_i \\ =& \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}) \\ =& \overline{X} - \overline{X} \\ =& 0,\end{aligned}$$

de lo que también se puede ver que $\sum_{i=1}^{n} x_i = 0$.

Como en el modelo de regresión usual las ecuaciones normales que deben satisfacer los estimadores de mínimos cuadrados $\hat{\mu}$ y $\hat{\beta}_1$ son

$$\begin{aligned} n\hat{\mu} + \left(\sum_{i=1}^{n} x_i\right)\hat{\beta}_1 &= \sum_{i=1}^{n} Y_i \\ \left(\sum_{i=1}^{n} x_i\right)\hat{\mu} + \left(\sum_{i=1}^{n} x_i^2\right)\hat{\beta}_1 &= \sum_{i=1}^{n} x_i Y_i. \end{aligned}$$

Simplificando estas ecuaciones considerando $\sum_{i=1}^{n} x_i = 0$ obtenemos

$$\begin{aligned} n\hat{\mu} &= \sum_{i=1}^{n} Y_i \\ \left(\sum_{i=1}^{n} x_i^2\right)\hat{\beta}_1 &= \sum_{i=1}^{n} x_i Y_i. \end{aligned}$$

Por lo que en el modelo centrado los estimadores de mínimos cuadrados son

$$\begin{aligned}\hat{\mu} =& \overline{Y} \\ \hat{\beta}_1 =& \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}.\end{aligned}$$

Si quisieramos centrar las Y's en el estimador $\hat{\beta}_1$ basta notar que

$$\sum_{i=1}^{n} x_i \overline{Y} = \overline{Y} \sum_{i=1}^{n} x_i = 0,$$

por lo que tendríamos

$$\hat{\beta}_1 = \sum_{i=1}^{n} \frac{x_i y_i}{\sum_{i=1}^{n} x_i^2},$$

donde $y_i = Y_i - \overline{Y}$, que coincide con el estimador para la pendiente del modelo de regresión simple usual, mientras que el estimador para el intercepto si varía, pues al considerar los datos centrados se hace un desplazamiento proporcional a la media, que en este caso es 0.

□

Problema 8

Recalcule la ecuación de regresión y el análisis de varianza para el ejemplo de datos de ozono contra rendimiento de plantas de soya usando el modelo centrado

$$Y_i = \mu + \beta_1 x_i + \epsilon_i,$$

donde $x_i = (X_i - \overline{X})$. Compare y comente los resultados que se obtienen con ambos modelos.

(Solución)

Los resultados se obtienen a través del código escrito en Python que se adjunta a la tarea

Los resultados que se obtienen con el *modelo usual* son

Fuente	df	Suma Cuadrados	Cuadrado Medio
Total	3	1014.75	
Reg	1	799.1381	799.1381
Residual	2	215.6119	107.8059

Los estimadores y varianzas son

Estimador	Valor	Varianza
β_0	253.4340	115.9422
β_1	-293.5310	11623.2809

Mientras que en el *modelo centrado* se obtiene lo siguiente

Fuente	df	Suma Cuadrados	Cuadrado Medio
Total	3	1014.75	
Reg	1	799.1381	799.1381
Residual	2	215.6119	107.8059

Los estimadores y varianzas son

Estimador	Valor	Varianza
β_0	227.75	26.9515
β_1	-293.5310	11623.2809

Aquí observamos que en el modelo centrado se obtiene un diferente valor del intercepto que habíamos anticipado en el problema anterior, además de eso se obtiene una varianza menor en el modelo centrado esto porque

$$V(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n x_i^2} \right) s^2,$$

y en el caso del modelo centrado $\overline{X} = 0$, luego en el modelo centrado la varianza es

$$V(\hat{\beta}_0) = \frac{s^2}{n},$$

que es menor a la varianza en el modelo usual.

□

Problema 9

Pruebe que

$$t = \frac{\hat{\beta}_1}{s/\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}},$$

es igual a

$$t = \frac{\sqrt{n-2}r}{\sqrt{1-r^2}},$$

donde

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2},$$

y

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]\left[\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right]}}.$$

(Solución)

Sustituyendo

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

y el valor de s en t además de realizando algunas simplificaciones tenemos

$$t = \frac{\sqrt{n-2}\frac{\sum_{i}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}}}{\sqrt{SS(Res)}}.$$

Multiplicando y diviendo por $\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$ y recordando la definición de r tenemos

$$t = \frac{\sqrt{n-2}r}{\sqrt{\frac{SS(Res)}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}}.$$

Del problema 3 se tiene que

$$r^2 = \frac{SS(Reg)}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2},$$

Además sabemos que

$$SS(Res) = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 - SS(Reg).$$

De esta dos últimas identidades y haciendo calculos simples llegamos a que

$$t = \frac{\sqrt{n-2}r}{\sqrt{1-r^2}},$$

que es lo que se queríamos probar.

□

Problema 10

Considere el modelo de regresión lineal

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

donde $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ son variables aleatorias independientes con distribución normal con media cero y varianza común σ^2.

(a) Obtenga los estimadores de máxima verosimilitud de los parámetros del modelo β_0, β_1 y σ^2.

(Solución)

Notamos que $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, si f_i es la correspondiente función de densidad entonces

$$\prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{ \frac{-(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right\}.$$

Si $\mathcal{L}(\beta_0, \beta_2 1, \sigma^2)$ es la verosimilitud entonces

$$L(\beta_0, \beta_1, \sigma^2) = (\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right\},$$

Sea $v = \sigma^2$, luego log-verosimilitud $l(\beta_0, \beta_1, v)$ con este cambio de variable es

$$l(\beta_0, \beta_1, v) = -\frac{n}{2} \log v - \frac{\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2}{2v}. \tag{1.11}$$

Sean $\hat{\beta}_0, \hat{\beta}_1$ los estimadores de mínimos cuadrados. Observamos que

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0, \beta_1) \in \mathbb{R}^2 \setminus \{0\}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \right\},$$

en consecuencia por (1.11) se tiene que

$$l(\beta_0, \beta_1, v) \leq l(\hat{\beta}_0, \hat{\beta}_1, v) \quad \forall v \in \mathbb{R}^+,$$

con igualdad si y sólo si $\beta_0 = \hat{\beta}_0$ y $\beta_1 = \hat{\beta}_1$.

Por lo tanto los estimadores de máxima verosimilitud de β_0 y β_1 coinciden con los de mínimos cuadrados.

Para obtener el estimador de máxima verosimilitud de v basta con maximizar $l(\hat{\beta}_0, \hat{\beta}_1, v)$ en la variable v derivando e igualando 0 obtenemos que el estimador de máxima verosimilitud es

$$\hat{v} = \frac{\sum_{i=1}^{b} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n}.$$

Notamos que

$$l(\hat{\beta}_0, \hat{\beta}_1, \hat{v}) - l(\hat{\beta}_0, \hat{\beta}_1, v) = -\frac{n}{2}\left(\log\frac{\hat{v}}{v} + 1 - \frac{\hat{v}}{v}\right).$$

Recordamos que

$$x \leq e^{x-1}, \quad x \in \mathbb{R}^+$$

luego

$$\log x \leq x - 1,$$

de esto se sigue que

$$l(\hat{\beta}_0, \hat{\beta}_1, \hat{v}) - l(\hat{\beta}_0, \hat{\beta}_1, v) \geq 0 \quad \forall v \in \mathbb{R}^+.$$

Por lo tanto

$$\hat{v} = \frac{\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n},$$

es el estimador de máxima verosimilitud de la varianza.

(b) Compare los estimadores de máxima verosimilitud y mínimos cuadrados. Comente esta comparación.

(Solución)
Los estimadores de máxima verosimilitud y de mínimos cuadrados de β_0 y β_1 coinciden, sin embargo, en el caso de la varianza no pasa eso.

El de mínimos cuadrados es

$$s^2 = \frac{\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n-2},$$

que es un estimador insesgado de la varianza, mientras que el de máximo verosimilitud es

$$\hat{v} = \frac{\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n},$$

que no es un estimador insesgado.

□

```
############
# TAREA 2 #
############

import numpy as np
import math as mt
import matplotlib.pyplot as plt
import matplotlib.ticker as ticker
import statistics as st
from scipy.stats import t

###############
# PROBLEMA 2 #
###############

# Datos Pesos vs Frecuencias
X=np.array([90,86,67,89,81,75])
Y=np.array([62,45,40,55,64,53])

print("PROBLEMA 2:\n")

# Inciso a)
figA= plt.figure()
ax=figA.add_subplot(111)
ax.set_xticks(np.arange(min(X)-1,max(X)+1,2))
ax.set_yticks(np.arange(min(Y)-1,max(Y)+1,2))
ax.plot(X,Y,'^r')
plt.xlabel("Peso (kg)")
plt.ylabel("Frecuencia")
plt.grid(True)
plt.show()
figA.savefig("GraficaDatos.png")

# Inciso b)
b1 =lambda X,Y: sum((X-np.mean(X,dtype=np.float64))*(Y-np.mean(Y,dtype=np.float
    64)))/sum((X-np.mean(X,dtype=np.float64))**2)
b0=lambda X,Y,p:np.mean(Y,dtype=np.float64)-p*np.mean(X,dtype=np.float64)
estY=lambda X,b0,b1:b0+b1*X

print("Los coeficientes son b0=%.4f y b1=%.4f" %(b0(X,Y,b1(X,Y)),b1(X,Y)))

pendiente=b1(X,Y)
intercepto=b0(X,Y,pendiente)

figB=plt.figure()
ax=figB.add_subplot(111)
ax.set_xticks(np.arange(min(X)-1,max(X)+1,2))
ax.set_yticks(np.arange(min(Y)-1,max(Y)+1,2))
ax.plot(X,Y,'^r')
domX=np.arange(min(X),max(X)+1)
ax.plot(domX,estY(domX,intercepto,pendiente),color='slateblue')
hatY=estY(X,intercepto,pendiente)
ax.plot(X,hatY,'ob')
plt.xlabel("Peso (kg)")
```

```
plt.ylabel("Frecuencia")
plt.grid(True)
plt.show()
figB.savefig('lineaRegress.png')


# Inciso c)
Xprima=np.delete(X,2)
Yprima=np.delete(Y,2)
print("\nLos coeficientes son b0=%.4f y b1=%.4f (quitando (67,40))" %(b0(Xprima,Yprima,b1(Xprima,Yprima)),b1(Xprima,Yprima)))


est88=estY(88,intercepto,pendiente) # Respuesta media

def mssRes(X,Y):
    beta1=b1(X,Y)
    beta0=b0(X,Y,beta1)
    hatY=estY(X,beta0,beta1)
    return sum((Y-hatY)**2)/(len(X)-2)

# Varianza de estimacion y prediccion
def varY(x,X,Y,i):
    if i==1: # Estimacion
        aux=(x-np.mean(X,dtype=np.float64))**2/(sum((X-np.mean(X,dtype=np.float64))**2))
        return (1/len(X)+aux)*mssRes(X,Y)
    if i==2: # Prediccion
        aux=(x-np.mean(X,dtype=np.float64))**2/(sum((X-np.mean(X,dtype=np.float64))**2))
        return (1+1/len(X)+aux)*mssRes(X,Y)

# Intervalos de confianza de 95%
def confIntervalY(x,X,Y,alpha,i): # i=1 Estimacion, i=2 Prediccion
    hatY=estY(x,b0(X,Y,b1(X,Y)),b1(X,Y))
    desv=mt.sqrt(varY(x,X,Y,i))
    aux=t.interval(1-alpha,len(X)-2)
    cfInterval=[hatY+aux[0]*desv,hatY+aux[1]*desv]
    return cfInterval

# Respuesta al inciso d)
print("\nLa estimacion en X=88 es: %.4f" %(est88))
print("\nEl intervalo de confianza para la estimacion es: ", confIntervalY(88,X,Y,0.05,1))

# Respuesta al inciso e)
print("\nEl intervalo de confianza para la prediccion es: ", confIntervalY(88,X,Y,0.05,2))

###############
# PROBLEMA 3 #
###############

from scipy.stats.stats import pearsonr
```

```
print("\n\nPROBLEMA 3:\n")

hatY=estY(X,b0(X,Y,b1(X,Y)),b1(X,Y))

#inciso a)
print("La correlacion entre Xi y Yi es: %.4f" %(pearsonr(X,Y)[0]))

#inciso b)
print("\nLa correlacion entre Yi y hatYi es: %.4f" %(pearsonr(Y,hatY)[0]))

#inciso c)
print("\nLa correlacion entre X y hatYi es: %.4f" %(pearsonr(X,hatY)[0]))

# Coeficiente de determinaci n
R2=((pendiente**2)*sum((X-np.mean(X,dtype=np.float64))**2))/sum((Y-np.mean(Y,
    dtype=np.float64))**2)
print("\n\nEl coeficiente de determinacion es: %.4f" %(R2))
print("\nLa correlacion de Xi y Yi al cuadrado es: %.4f" %(pearsonr(X,Y)[0]**2)
    )

###############
# PROBLEMA 8 #
###############

print("\n\nPROBLEMA 8:\n")

Xoz=np.array([0.02,0.07,0.11,0.15])
Yred=np.array([242,237,231,201])

def mssReg(X,Y):
    value=(b1(X,Y)**2)*(sum((X-np.mean(X,dtype=np.float64))**2))
    return value

def varb0(X,Y):
    sigma2=mssRes(X,Y)
    aux=((np.mean(X,dtype=np.float64))**2)/sum((X-np.mean(X,dtype=np.float64))
    **2)
    value=(1/len(X)+aux)*sigma2
    return value

def varb1(X,Y):
    sigma2=mssRes(X,Y)
    value=sigma2/sum((X-np.mean(X,dtype=np.float64))**2)
    return value
# Regresion sin centrar
print("\nModelo sin centrar:")

print("\nIntercepto: %.4f, Pendiente: %.4f" %(b0(Xoz,Yred,b1(Xoz,Yred)),b1(Xoz,
    Yred)))

print("\nSS(Res): %.4f, SS(Reg): %.4f" %((len(Xoz)-2)*mssRes(Xoz,Yred),mssReg(
    Xoz,Yred)))

print("\nMSS(Res): %.4f, MSS(Reg): %.4f" %(mssRes(Xoz,Yred),mssReg(Xoz,Yred)))
```

```
152
153 print("\nVarianza intercepto: %.4f, Varianza pendiente: %.4f" %(varb0(Xoz,Yred)
    ,varb1(Xoz,Yred)))
154
155 # Regresion centrada
156 XozC=Xoz-np.mean(Xoz,dtype=np.float64)
157
158 print("\n\nModelo centrado:")
159
160 print("\nIntercepto: %.4f, Pendiente: %.4f" %(b0(XozC,Yred,b1(XozC,Yred)),b1(
    XozC,Yred)))
161
162 print("\nSS(Res): %.4f, SS(Reg): %.4f" %((len(XozC)-2)*mssRes(XozC,Yred),mssReg
    (XozC,Yred)))
163
164 print("\nMSS(Res): %.4f, MSS(Reg): %.4f" %(mssRes(XozC,Yred),mssReg(XozC,Yred))
    )
165
166 print("\nVarianza intercepto: %.4f, Varianza pendiente: %.4f" %(varb0(XozC,Yred
    ),varb1(XozC,Yred)))
```

Listing 1: Código Tarea 2