PRÁCTICA 2

1. Completar los detalles acerca de las propiedades de la distribución de $\widehat{\beta}_0$ en la demostración del Teorema 2.1 de la teórica.

   a) Mostrar que el estimador $\widehat{\beta}_0$ puede escribirse de la forma

   $$\sum_{i=1}^{n} d_i Y_i, \tag{1}$$

   donde

   $$d_i = \left(\frac{1}{n} - \frac{(X_i - \overline{X})\,\overline{X}}{(n-1)S_X^2} \right), \quad \text{y} \quad s_X^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{2}$$

   b) Probar que los d_i satisfacen

   i. $\sum_{i=1}^{n} d_i = 1$
   ii. $\sum_{i=1}^{n} d_i X_i = 0$

   c) A partir de la expresión de $\widehat{\beta}_0$ obtenida en el ítem (a) para , también puede escribirse a $\widehat{\beta}_0$, pruebe que $\widehat{\beta}_0$ se puede escribir en términos de los ε_i de la siguiente forma,

   $$\widehat{\beta}_0 = \beta_0 + \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(X_i - \overline{X})\,\overline{X}}{n s_X^2} \right) \varepsilon_i. \tag{3}$$

   d) Verificar que

   $$Var\left(\widehat{\beta}_0 \mid X_1, \ldots, X_n\right) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{n s_X^2} \right)$$
   $$\operatorname{cov}\left(\widehat{\beta}_0, \widehat{\beta}_1 \mid X_1, \ldots, X_n\right) = -\frac{\sigma^2\,\overline{X}}{n\, s_X^2}$$

   e) Probar que una expresión alternativa para la varianza condicional de $\widehat{\beta}_0$ está dada por

   $$Var\left(\widehat{\beta}_0 \mid X_1, \ldots, X_n\right) = \sigma^2 \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{n s_X^2}.$$

2. Sean $r_1, r_2, \ldots, r_n$ los residuos del ajuste lineal a la recta de mínimos cuadrados, es decir

   $$r_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i. \tag{4}$$

   Asumimos que las X_i son fijas (o que las esperanzas y varianzas que siguen se toman condicionales a las X_i).

   a) Mostrar que $\sum_{i=1}^{n} r_i = 0$.

   b) Verificar que

   $$\operatorname{Var}(r_k) = \operatorname{Var}(Y_i) + \operatorname{Var}\left(\widehat{\beta}_0\right) + X_i^2 \operatorname{Var}\left(\widehat{\beta}_1\right) - 2\operatorname{Cov}\left(Y_i, \widehat{\beta}_0\right) - 2X_i \operatorname{Cov}\left(Y_i, \widehat{\beta}_1\right) + 2X_i \operatorname{Cov}\left(\widehat{\beta}_0, \widehat{\beta}_1\right).$$

   c) Usar

   - la bilinealidad de la covarianza
   - la independencia de las distintas observaciones
   - la expresión (1) para $\widehat{\beta}_0$ y la expresión análoga vista en clase para $\widehat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i$
   - el ítem 1e)
   - el resultado de covarianza del Teorema 2.1 de la teórica

para probar que

$$\text{Cov}\left(Y_k, \widehat{\beta}_0\right) = \sigma^2 d_k$$
$$\text{Cov}\left(Y_k, \widehat{\beta}_1\right) = \sigma^2 c_k$$
$$\text{Var}(r_k) = \sigma^2 \left[\frac{n-2}{n} + \frac{1}{n s_X^2}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - X_k^2 + 2\overline{X}\left(X_k - \overline{X}\right)\right)\right]$$

Observar que la varianza del residuo r_k depende de X_k y por lo tanto los residuos no son idénticamente distribuidos.

d) Concluir que

$$\frac{1}{n}\sum_{k=1}^{n} \text{Var}(r_k) = \sigma^2 \frac{(n-2)}{n}$$

y finalmente deducir que entonces

$$E\left(\widehat{\sigma}^2\right) = E\left(\frac{1}{n}\sum_{k=1}^{n} r_k^2\right) = \frac{1}{n}\sum_{k=1}^{n} E\left(r_k^2\right) = \frac{1}{n}\sum_{k=1}^{n} \text{Var}(r_k) = \sigma^2 \frac{(n-2)}{n},$$

por lo que $\widehat{\sigma}^2$ resulta ser un estimador sesgado de σ^2 y por lo tanto $S^2 = \frac{n}{(n-2)}\widehat{\sigma}^2$ resulta ser un estimador insesgado de σ^2.

3. Sean $(X_1, Y_1), \ldots, (X_n, Y_n)$ observaciones que siguen el modelo lineal

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ 1 \leq i \leq n$$

Consideremos la recta ajustada por mínimos cuadrados y los residuos de las observaciones definidos por (4).

a) Probar que la suma de los residuos vale cero, a partir de las ecuaciones normales. (En el ejercicio anterior ya lo probamos usando la definición del estimador $\widehat{\beta}_0$). Deducir que entonces el promedio muestral de los residuos también vale 0.

b) Probar que

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \widehat{Y}_i,$$

siendo $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_i X_i$ el valor predicho i–ésimo.

c) Deducir del ítem anterior que el promedio de las respuestas observadas Y_i, es igual al promedio de los valores predichos, $\widehat{Y}_i$, es decir, $\overline{Y} = \overline{\widehat{Y}}$.

d) A partir de las ecuaciones normales obtener que

$$\sum_{i=1}^{n} \widehat{Y}_i r_i = 0$$

Deducir que entonces

$$\sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{\widehat{Y}}\right) r_i = 0$$

y que por lo tanto la correlación muestral de Pearson entre los predichos y los residuos, $\widehat{\rho}\left(\widehat{Y}_i, r_i\right)$ también vale 0. ¿Cómo se interpreta este resultado?

e) Probar que la recta ajustada siempre pasa por el punto $\left(\overline{X}, \overline{Y}\right)$.

4. El archivo `vapor.txt` contiene datos registrados durante 25 meses en una planta de vapor. Se observaron:

$Y =$ cantidad de vapor consumido (en libras)

$X =$ promedio mensual de temperatura atmosférica (en Farenheit).

a) Realizar el diagrama de dispersión para Y vs. X. ¿ Qué se observa?

b) Hallar la media y el desvío standard de cada una de las variables.

c) Si se plantea un modelo $E(Y_i) = \beta_0 + \beta_1 X_i, \quad i = 1, 2, \ldots, 25$, hallar los estimadores de mínimos cuadrados de β_0 y β_1.

d) ¿Cuánto vale el estimador de σ^2?

e) Calcular la matriz de covarianza de los estimadores obtenidos.

f) Verificar que $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i) = 0$.

g) Centrar las observaciones X_i's y recalcular los estimadores de los parámetros. ¿Con quién coincide $\widehat{\beta}_0$? ¿Cambia el estimador de σ^2? Recalcular la matriz de covarianza de los estimadores y compararla con la obtenida en e).

5. El conjunto de datos `bdims` del paquete `openintro` que se habilita en el workspace del R con los comandos

```
library(openintro)
```

```
data(bdims, package = ''openintro'')
```

Ya trabajamos con estos datos en la Práctica 1.

(a) Realizar un diagrama de dispersión que muestre la relación entre el peso medido en kilogramos (`wgt`) y la circunferencia de la cadera medida en centímetros (`hip.gi`), poner el peso en el eje vertical. Describir la relación entre la circunferencia de la cadera y el peso.

(b) ¿Cómo cambiaría la relación si el peso se midiera en libras mientras que las unidades para la circunferencia de la cadera permanecieran en centímetros?

(c) Ajustar un modelo lineal para explicar el peso por la circunferencia de cadera, con las variables en las unidades originales. Escribir el modelo (con papel y lápiz, con betas y epsilones). Luego, escribir el modelo ajustado (sin epsilones). Interpretar la pendiente estimada en términos del problema. La respuesta debería contener una frase que comience así: "Según el modelo ajustado, si una persona aumenta un cm. de contorno de cadera, en promedio su peso aumentará ... kilogramos".

(d) Superponer la recta ajustada al scatterplot. Observar el gráfico. ¿Diría que la recta describe bien la relación entre ambas variables?

(e) Elegimos una persona adulta físicamente activa entre los estudiantes de primer año de la facultad. Su contorno de cadera mide 100 cm. Predecir su peso en kilogramos.

(f) Esa persona elegida al azar pesa 81kg. Calcular el residuo.

(g) Estimar el peso esperado para la población de adultos cuyo contorno de cadera mide 100 cm.

(h) Estimar la varianza del error (σ^2) con un estimador insesgado.

(i) Hallar un intervalo de confianza para la pendiente de la recta, β_1, asumiendo que los errores del modelo tienen distribución normal.

6. (Del Libro de Weisberg (2005))Durante el período 1893–1898, E. S. Pearson organizó la recolección de las alturas de $n = 1375$ madres en el Reino Unido menores de 65 años y una de sus hijas adultas mayores de 18 años. Pearson y Lee (1903) publicaron los datos para estudiar cómo se traspasaban datos genéticos en la herencia. Los datos (medidos en pulgadas) pueden verse en el archivo de datos `Heights` del paquete `alr4` de `R`. Nos interesa estudiar el traspaso de madre a hija, así que miramos la altura de la madre, llamada `Mheight`, como la variable predictora y la altura de la hija, `Dheight`, como variable de respuesta. ¿Será que las madres más altas tienden a tener hijas más altas? ¿Las madres más bajas tienden a tener hijas más bajas?

(a) Realizar un scatterplot de los datos, con la altura de las madres en el eje horizontal.

i. Como lo que queremos es comparar las alturas de las madres con la de las hijas, necesitamos que en el scatterplot las escalas de ambos ejes sean las mismas (y que por lo tanto el gráfico sea cuadrado).

ii. Si cada madre e hija tuvieran exactamente la misma altura que su hija, ¿cómo luciría este scatterplot? Resumir lo que observa en este gráfico. Superponer la figura que describió como respuesta a la pregunta anterior. ¿Describe esta figura un buen resumen de la relación entre ambas variables?

iii. Los datos originales fueron redondeados a la pulgada más cercana. Si trabajamos directamente con ellos, veremos menos puntos en el scatterplot, ya que varios quedarán superpuestos. Una forma de lidiar con este problema es usar el *jittering*, es decir, sumar un pequeño número uniforme aleatorio se a cada valor. Los datos de la librería alr4 tienen un número aleatorio uniforme en el rango de -0.5 a +0.5 añadidos. Observemos que si se redondearan los valores del archivo `heights` se recuperarían los datos originalmente publicados. En base al scatterplot, ¿parecería ser cierto que las madres más altas suelen tener hijas más altas y viceversa con las más bajas?

(b) Ajustar el modelo lineal a los datos. Indicar el valor de la recta ajustada. Superponerla al scatter plot. ¿Presenta visualmente un mejor ajuste que la recta identidad postulada en el ítem anterior?

(c) Dar los estimadores de los coeficientes de la recta, sus errores estándares, estimar la varianza de los errores.

(d) Halle un intervalo de confianza de nivel 0.95 para la pediente. El comando `confint` aplicado a un objeto `lm` puede ser de utilidad.

(e) Testear la hipótesis $E(\texttt{Dheight} \mid \texttt{Mheight}) = \beta_0$ versus la alternativa que $E(\texttt{Dheight} \mid \texttt{Mheight}) = \beta_0 + \beta_1 \texttt{Mheight}$. Escribir la conclusión al respecto en un par de renglones.

(f) Predecir la altura de una hija cuya madre mide 64 pulgadas. Observar que para que esta predicción sea razonable, hay que pensar que la madre vivía en Inglaterra a fines del siglo XIX.

(g) Una pulgada equivale a 2.54cm. Convertir ambas variables a centímetros (`Dheightcm` y `Mheightcm)` y ajuste un modelo lineal a estas nuevas variables. ¿Deberían cambiar los estimadores de β_0 y β_1? ¿De qué manera? ¿Y los errores estándares? ¿Y los p-valores de los tests para testear si cada uno de ellos es igual o distinto de cero? ¿Y la estimación del desvío estándar de los errores? Comparar ambos resultados, y verificar si las conjeturas realizadas resultaron ciertas. En estadística, que un estimador se adapte al cambio de escala en las variables (covariable y respuesta) se dice: "el estimador es equivariante (afín y por escala)".