1. **Datos de Próstata.** En el archivo `prostate.txt` se encuentran los datos correspondientes a 97 pacientes de de cáncer de próstata antes de tener una prostatectomía radical, y se miden las siguientes variables:

   - **lcavol**: log del volumen del tumor
   - **lweight**: log del peso de la próstata
   - **age**: edad
   - **lbph**: log de la cantidad de hiperplasia prostática benigna
   - **svi**: invasión seminal (si o no)
   - **lcp**: logaritmo de la penetración capsular
   - **gleason**: score de Gleason
   - **pgg45**: porcentaje de scores de Gleason 4 or 5.
   - **lpsa**: log del PSA (antígeno prostático)

   El objetivo es predecir el logaritmo del PSA. Para ello, proponemos ajustar un modelo lineal para `lpsa` usando todas las variables explicativas registradas. Asumiendo normalidad de los errores, complete los siguientes items.

   a) Escriba i) el modelo propuesto y ii) el modelo ajustado.

   b) ¿Cuáles son los coeficientes estimados para el intercept y β_{age}, el coeficiente correspondiente a la variable `age`, usando el método de mínimos cuadrados?

   c) Supongamos que los errores tienen varianza σ^2. ¿Cómo se estima σ^2? ¿Cuánto vale la estimación de σ^2 en este ejemplo?

   d) ¿Hay evidencia suficiente a nivel 0.01 para decir que $\beta_{age} \neq 0$? ¿Cuál es el estadístico del test correspondiente y cuál es su distribución bajo H_0? ¿Cuánto vale este estadístico para estos datos? Hallar el $p-$valor e interpretar la conclusión de este test.

   e) Hallar un intervalo de confianza para el intercept β_0 de nivel 0.95.

   f) ¿Cuánto vale el R^2 en este ejemplo y cómo interpretaría esto?

   g) Se quiere testear la significacón de la regresión. ¿Cómo lo haría? ¿Qué estadístico usaría? ¿Cuánto vale en este caso y qué decisión tomaría? Justificar.

h) Sea β_{lcavol} el coeficiente correspondiente a la variable `lcavol`. Mirando la salida, hallar el $p-$valor del test con hipótesis $H_0 : \beta_{lcavol} = 0$ vs. $H_1 : \beta_{lcavol} \neq 0$. ¿Cuál sería el $p - valor$ del test con hipótesis $H_0 : \beta_{lcavol} = 0$ vs. $H_1 : \beta_{lcavol} < 0$.

i) Hallar un intervalo de confianza de nivel 0.95 para $\beta_{lweight}$ (el coeficiente correspondiente a la variable `lweight`).

j) ¿Cuáles parecen ser las variables más relevantes en este modelo?

2. **Leave-one-out Cross-Validation.** En el puerto de la Ciudad de Grand Lakes, en Canadá, se quiere ver cómo influye el peso de un cargamento en el tiempo necesario para descargarlo. Para esto, se registró el peso y el tiempo de descarga para 30 cargamentos Los datos se encuentran en el archivo `glakes.csv`. Para este conjunto de datos, se proponen los siguientes modelos:

   I. $\text{Tiempo} = \beta * \text{Peso} + \varepsilon$

   II. $\text{Tiempo} = \alpha + \beta * \text{Peso} + \varepsilon$

   III. $\log(\text{Tiempo}) = \alpha + \beta * \text{Peso}^{0.25} + \varepsilon$

   IV. $\log(\text{Tiempo}) = \alpha + \beta * \text{Peso}^{0.5} + \gamma * \text{Peso}^{0.25} + \varepsilon$

   Consideremos el siguiente procedimiento como modo de *evaluar* cada modelo:

   a) Sea n el tamaño muestral. Fijado un $i \in \{1, \ldots, n\}$, obtener los coeficientes estimados $\hat{\beta}^{(-i)}$ que correspondan usando todas las observaciones salvo la $i-$ésima.

   b) Con las estimaciones obtenidas $\hat{\beta}^{(-i)}$, predecir el valor de la $i-$ésima observación, llamemos $\hat{Y}_i^{(-i)}$ a dicha predicción. Luego, obtener los residuos $e_i = Y_i - \hat{Y}_i^{(-i)}$.

   c) Hacer variar el i de 1 a n ($n =$ cantidad de observaciones) y para cada i, seguir los pasos anteriores, obteniendo de esta forma los residuos $e_1, \ldots, e_n$.

   d) Computar $W = \sum_{i=1}^{n} e_i^2$.

   Si se tienen varios modelos posibles, se elige aquel que tiene menor W.

   Aplicando este procedimiento a cada uno de los cuatro modelos propuestos, decidir cuál elegiría.

   Nota: Para los modelos (III) y (IV), las predicciones se obtienen haciendo $\hat{y}_i = \exp(\hat{\alpha} + \hat{\beta} * \text{Peso}^{0.25})$ e $\hat{y}_i = \exp(\hat{\alpha} + \hat{\beta} * \text{Peso}^{0.25} + \hat{\gamma} * \text{Peso}^{0.25})$, respectivamente.

3. **Efectos especiales.**

Errores Heteroscedásticos. Con este ejercicio nos proponemos evaluar el efecto de suponer homocedasticidad cuando en realidad las varianzas de los errores no son constantes. Para ello haremos una pequeña simulación a fin de comprobar numéricamente lo que ocurre.

a) i) Generar una muestra de tamaño $n = 50$ de pares (y_i, x_i) que correspondan al modelo

$$y_i = 1 + 50x_i + \epsilon_i$$

donde $x_i \sim N(0, 1)$ son independientes de los errores $\epsilon_i \sim N(0, 9 \cdot i^2)$.
Ajustar a las respuestas y_i un modelo lineal simple basado en las covariables x_i usando el método de mínimos cuadrados ordinario y el método de mínimos cuadrados pesados, utilizando para el i- ésimo par el peso $w_i = \frac{1}{i}$.

ii) Guardar las estimaciones obtenidas para los parámetros en cada modelo.

ii) Repetir $Nrep = 1000$ y realizar un boxplot de las estimaciones de la intercept y el coeficiente que acompaña a x bajo cada método. Comparar con el valor verdadero de cada uno de estos parámetros.

iv) ¿Qué ocurre al no tener en cuenta el supuesto de homocedasticidad?

Sesgo por Variables Omitidas. Con este ejercicio nos proponemos evaluar el efecto de omitir variables relevantes en el modelo y para ello haremos una pequeña simulación a fin de comprobar numéricamente el siguiente hecho.

Supongamos que

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon\,,$$

si calculamos el estimador de β_1 usando solamente a X_1 tendríamos

$$\widetilde{\beta_1} = (X_1' X_1)^{-1} X_1' Y\,.$$

En este caso, se puede comprobar que el sesgo por variable omitida que se introduce es:

$$\delta\,\beta_2 = (X_1' X_1)^{-1} X_1' X_2 \beta_2\,,$$

con lo que el sesgo depende de la correlación de X_2 con X_1 y por otra parte, de la relevancia de X_2 en el modelo.

a) i) Generar una muestra de tamaño $n = 50$ de ternas (y_i, x_{i1}, x_{i2}) que correspondan al modelo

$$y_i = 1 + x_{i1} + 10x_{i2} + \epsilon_i$$

donde $x_{i1} \sim N(0,1)$, $x_{i2} \sim N(0,1)$ independientes entre sí e independientes de los errores $\epsilon_i \sim N(0, 0.25^2)$. Por el método de mínimos cuadrados ajustar los siguientes dos modelos:

$$\begin{aligned} \mathcal{M}_1 &: y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i \\ \mathcal{M}_2 &: y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \end{aligned}$$

ii) Guardar las estimaciones obtenidas para los parámetros en cada modelo.

iii) Repetir $Nrep = 1000$ y realizar un boxplot de las estimaciones de la intercept y el coeficiente que acompaña a x_1 bajo cada modelo. Comparar con el valor verdadero de cada uno de estos parámetros.

iv) ¿Qué ocurre al omitir la variable x_2?

b) Repetir a) considerando el modelo

$$y_i = 1 + x_{i1} + 0.1 x_{i2} + \epsilon_i$$

c) Repetir a) y b) considerando generando covariables correlacionadas usando el siguiente modelo

$$x_{i2} = 0.5 x_{i1} + u_i \ , u_i \sim N(0, 0.5^2)$$

d) ¿Qué conluye?

4. **Seleccionando modelos.** Consideremos el conjunto de datos del archivo `prostate.txt` e implementemos el siguiente procedimiento de comparación de modelos:

Paso 1: Dividir las observaciones en dos grupos: I) de *Entrenamiento* y II) de *Testeo*, de manera tal que el grupo *Entrenamiento* conste aproximadamente de 2/3 del total de las observaciones (elegidas aleatoriamente), mientras que restantes se asignan al grupo *Testeo*. Llamar k al tamaño del grupo de *Testeo* y $n - k$ al de *Entrenamiento*.

Paso 2: Dado un modelo, obtener los coeficientes estimados usando nicamente las observaciones del grupo *Entrenamiento* y con dichas estimaciones, predecir cada una de las observaciones del grupo *Testeo*. Finalmente, con los valores predichos $\hat{y_1}, \ldots, \hat{y_k}$, sobre las observaciones del grupo *Testeo* calcular $W = \sum_{i=1}^{k} (y_i - \hat{y_i})^2$.

Consideremos todos los modelos lineales con intercept cuyas variables explicativas son un subconjunto de $\{$`lcavol`, `lweight`, `lbph`$\}$ (son 8 modelos en total) . Calcular el W de cada uno de ellos y decidir cuál de estos 8 modelos preferiría (Esto es: entre los modelos considerados, elegir el que tiene menor W).