

Modelos de regresión

Tarea 3

Fecha de entrega: 15 de noviembre

1. Considerando el modelo de regresión simple dado por

$$y = \beta_0 + \beta_1 x + \epsilon,$$

donde $\mathbb{E}(\epsilon) = 0$ y $\text{Var}(\epsilon) = \sigma^2$. Realiza lo siguiente.

- (a) Resuelve de nueva cuenta el problema de mínimos cuadrados para obtener a $\hat{\beta}_0$ y $\hat{\beta}_1$.
- (b) Demuestra que $\hat{\beta}_0$ y $\hat{\beta}_1$ son insesgados y obtén sus varianzas.
- (c) Demuestra que $SS_{\text{Res}} = SS_{\text{T}} - \hat{\beta}_1 S_{xy}$, donde

$$S_{xy} = \sum_{i=1}^n y_i x_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}.$$

2. Encuentra los estimadores máximo verosímiles para el modelo de regresión múltiple dado por

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

donde \mathbf{y} es un vector de tamaño n , \mathbf{X} es la matriz de diseño de tamaño $p \times n$ y $\epsilon \sim N_k(\mathbf{0}, \sigma^2 \mathbf{I})$.

3. Demuestra que la matriz sombrero $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ es simétrica e idempotente. Con este resultado demuestra que $\mathbf{I} - \mathbf{H}$ también lo es.
4. Considera el archivo *rendimientoGasolina.txt*, el cual contiene el rendimiento de la gasolina por kilómetro de 32 automobiles (variable y), así como 11 variables explicativas (denotadas por x_i) y representando las siguientes características.
 - x_1 : cilindrada.
 - x_2 : caballos de fuerza.
 - x_3 : torsión.

- x_4 : relación de compresión.
- x_5 : relación de eje trasero.
- x_6 : carburador.
- x_7 : número de velocidades en la transmisión.
- x_8 : longitud.
- x_9 : ancho.
- x_{10} : peso.
- x_{11} : tipo de transmisión (1 = automático, 0 = manual).

- (a) Ajusta un modelo de regresión lineal simple que relacione el rendimiento de la gasolina y la cilindrada. ¿Es significativa la regresión? ¿Qué puedes decir de los residuales?
- (b) Determina un intervalo de confianza y un intervalo de predicción, ambos al 95%, para el rendimiento de la gasolina si la cilindrada es de 275.
- (c) Ahora se desea ajustar un modelo de regresión múltiple. Para esto, primero realiza un proceso de selección hacia adelante, hacia atrás y por segmentos. Comenta en los resultados obtenidos y selecciona el que creas es el mejor modelo. ¿Es significativa la regresión? ¿Mejora los resultados vistos anteriormente?

5. La CFE está interesada en entender la relación entre la temperatura promedio diaria y el consumo eléctrico de una ciudad con el objetivo de crear un modelo que permita predecir el consumo eléctrico usando el pronóstico del tiempo. Para llevarlo a cabo se le proporciona la base de datos *luz.txt*. Ajuste y valide un modelo de regresión lineal simple para los datos. Si los supuestos de la regresión se cumplen, construya un intervalo de confianza y un intervalo de predicción, ambos al 95%.
6. Considerando la base *regresionPolinomial.txt*, ajusta un modelo de regresión polinomial. Para esto ten en cuenta las siguientes preguntas: ¿Qué orden parece ser el adecuado? ¿Todas las variables son significativas para la regresión? ¿Qué sucede con los residuales? ¿Se requiere alguna transformación?
7. Considera el modelo de regresión lineal simple, donde los errores se asumen normales de media cero y varianza σ^2 constante. Demuestra que la prueba por el método de cociente de verosimilitudes, para probar

$$H_0 : \beta_1 = \beta_{10} \quad \text{vs} \quad H_1 : \beta_1 \neq \beta_{10},$$

depende de una estadística t . (*Hint: Cuando se tenga una versión simplificada del cociente, habrá dos sumas de cuadrados, una en el numerador y otra en el denominador, en esta última sume y reste $\hat{\beta}_1(X_i - \bar{X})$.*)

8. Verifica si para la base de datos *licuefaccion.txt* existe el problema de multicolinealidad. Esta base 27 mediciones y 8 variables que son:

- $y = CO_2$.
- $x_1 =$ tiempo (en min).
- $x_2 =$ temperatura (en grados Celsius).
- $x_3 =$ porcentaje de solvatación.
- $x_4 =$ aceite.
- $x_5 =$ carbón.
- $x_6 =$ solvente.
- $x_7 =$ consumo de hidrógeno.

9. Considerando el archivo *mediciones.txt* realiza lo siguiente.

- (a) Ajusta un modelo de regresión lineal simple. Comenta en los resultados de la regresión.
- (b) ¿Se cumplen todos los supuestos de la regresión? Si no es el caso, encuentra una transformación adecuada y ajusta un nuevo modelo de regresión.
- (c) Construye los intervalos de confianza y predicción al 95% de confianza en la(s) nueva(s) variable(s).
- (d) Con la inversa de la transformación que elegiste, regresa tus resultados a las variables originales.

10. Considera la base de datos *calidadRioja.txt*, la cual contiene 38 observaciones acerca de la calidad, aroma, cuerpo, sabor y la región de algunos vinos Rioja.

- (a) Ajusta un modelo de regresión que explique la calidad en término de alguna (o de algunas) de las variables, aroma, cuerpo y sabor.
- (b) Ahora introduce la variable región a tu modelo. ¿Cómo afecta el introducir la región a la variable o variables que habías seleccionado previamente?