Regresión Avanzada
Práctica 3: Problemas de la Regresión

Ejercicios con Datos

1. Consideramos el conjunto de datos *mammals* de la librería de R *openintro*, que describe el comportamiento de 39 especies de mamíferos se estudiaron diversas características de comportamiento y físicas. El conjunto original de datos tiene datos faltantes, sacarlos con el comando na.omit.

   (a) Hacer un scatter plot del peso del cerebro de un mamífero **brainWt** en función de su peso corporal **bodyWt**. Describir el gráfico. Ajustas un modelo de regresión lineal. Analizar sus residuos.

   (b) Transformar la variables utilizadas en el ejercicio anterior (para transformar la variable de respuesta se sugiere hallar la transformación de Box y Cox), aplicar la misma transformación a la variable de respuesta. Volver a ajustar el modelo. Repetir el análisis.

   (c) Considerar ahora también las variables *Gestation, LifeSpan* y *TotalSleep*. Analizar si es conveniente sumar alguna de estas variables al análisis.

   (d) A partir del último ajuste detectar la observaciones de alto leverage y analizar posibles outliers. Si no hay outliers analizar los pesos de la regresión robusta.

2. Consideramos nuevamente los datos relativos al fracaso escolar en Madrid (`fracaso.txt`).

   - Hacer un análisis de diagnóstico de los residuos, si fuera necesario o conveniente proponer una mejora del modelo.
   - Construir y graficar intervalos de confianza y predicción a nivel 95% para la variable de respuesta

3. Continuamos con el primer ejercicio (parte práctica) de la ejercitación 2, basado en el conjunto de datos *trees*. Considerando el modelo

$$\log(Volume) = \beta_0 + \beta_1 \log(Girth) + \beta_2 \log(Height) + \epsilon$$

(a) Qué se puede decir sobre si se cumplen o no las hipótesis habituales del modelo de regresión?

(b) Elimina la observación cuya distancia de Cook es máxima y ajusta de nuevo el modelo. Hay mucha diferencia con los resultados obtenidos anteriormente?

(c) Cómo interpretás la existencia de puntos para los que el leverage toma valores altos y, simultáneamente, la distancia de Cook es pequeña?

4. La solubilidad de los alcoholes en agua es importante para comprender el transporte de alcohol en organismos vivos. Este conjunto de datos de (Romanelli et al., 2001) contiene características fisicoquímicas de 44 alcoholes alifáticos. El objetivo del experimento fue la predicción de la solubilidad sobre la base de descriptores moleculares. El conjunto de datos *alcohol* se encuentra en la libreria de R *robustbase*.

   (a) Realizar un análisis exploratorio de los datos. Qué se observa?

   (b) Ajustar un modelo lineal para predecir la solubilidad. Analizar el ajuste. Hacer un análisis de los residuos. Comentar.

   (c) Volver a ajustar el modelo, pero ahora hacerlo en forma robusta. Repetir el análisis del item anterior.

   (d) Analizar los pesos de la regresión robusta. Explicar los resultados obtenidos.

5. Considerar el conjunto de datos *cps09mar* que se encuentran en el libro de Hansen (`https://www.ssc.wisc.edu/~bhansen/econometrics/`). Consideramos un subconjunto de datos de la base *cps09mar* dada por el conjunto de datos *cps09marsubcjto.txt*. La descripción del conjunto se encuentra en `https://www.ssc.wisc.edu/~bhansen/econometrics/cps09mar_description.pdf`. Considerar los datos dados por las variables *aerning, age, education, hours* y *week* a partir de las mismas generar la siguientes variable

   - *wage=earnings/(hours*week)*

   (a) En base a la variable *education* hacer la regresión para predecir *wage*. Hacer un análisis exploratorio. Ajustar el modelo. Analizar los residuos.

(b) Transformar la variable *wage* y evaluar incorporar otras variables las variables regresoras. Encontrar un ajuste que resulte mejor.

6. El conjunto de datos *usch.txt* (que forma parte del conjunto de datos *uschange* de la librería *fpp2*) presenta datos correspondientes al porcentaje de cambios en el gasto trimestral de consumo personal e ingreso personal disponible para los EE. UU. entre 1960 y 2016.

   (a) Hacer el scatter plot del consumo versus el ingreso. Y además, en un mismo gráfico, superponer las curvas de consumo e ingreso en función del tiempo.

   (b) Ajustar el modelo lineal para predecir el consumo a partir del ingreso.

   (c) Analizar los residuos estandarizados. Recordar hacer el gráfico de los residuos en función del tiempo.

   (d) Testear si el coeficiente de correlación entre estas variables es positivo.

   (e) Proponer la siguiente transformación de los datos y repetir el análisis.

   - $consumo_t^* = consumo_t - \widehat{\rho} consumo_{t-1}$
   - $ingreso_t^* = ingreso_t - \widehat{\rho} ingreso_{t-1}$,

   donde $\widehat{\rho}$ es el estimador de la correlación entre el consumo y el ingreso.

7. Consideramos los datos que es `Meuse.all` que se encuentran en la librería `gstat`. Este conjunto de datos proporciona ubicaciones y concentraciones medidas en la zona del suelo de metales pesados(ppm), junto con una serie de variables del suelo y paisaje, recogidas en la margen del río Meuse, cerca de la ciudad de Stein.

   Se busca un modelo para explicar la variable `copper`, que representa la concentración de cobre en la capa superior del suelo (ppm) en función de las variables

   `dist` distancia a la margen del río Meus.

   `elev` elevación relativa.

`om` materia orgánica, como porcentaje.

(a) Ajustar un modelo de regresión múltiple con las variables sin transformar. Describir lo observado.

(b) Subsanar los problemas encontrados transformando las variables que consideres necesarios.

Simulaciones

1. (a) Generar datos que sigan el modelo

$$y_i = 5 + 3x_{1i} - 2x_{2i} + \epsilon_i,$$

con $n = 10$ donde las $x_j \sim U[-4, 4]$ con $j = 1, 2$ independientes y $\epsilon_i \sim Exp(0.5) - 2$ (restamos 2 para que la media sea cero).
Generar además una variable $x_3 \sim U[-4, 4]$ independiente de las anteriores.

(b) Ajustar el modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i,$$

(c) i. Guardar los parámetros estimados.

ii. Construir los intervalos de confianza de nivel 0.90 para los parámetros β_0 y β_3. Para cada uno de estos parámetros poner en un vector un 1 si el verdadero valor pertenece al intervalo y un 0 sino.

(d) Repetir los pasos anteriores $B = 1000$ veces.

(e) i. Analizar la distribución de los parámetros (histogramas, qq-plot, boxplot).

ii. Qué porcentaje de las veces el verdadero valor del parámetro está en el intervalo?

(f) Repetir la simulación ahora para $n = 30$ y $n = 100$.

(g) Repetir las simulaciones anteriores modificando la distribución de los errores $\epsilon_i \sim U[-2, 2]$.