Regresión Avanzada
Práctica 2: Regresión Lineal Múltiple

Ejercicios teóricos

1. (a) Mostrar que $X'(Y - X\widehat{\beta}) = 0$
   (b) Mostrar que $(Y - X\beta)'(Y - X\beta) = (Y - X\widehat{\beta})'(Y - X\widehat{\beta}) + (\widehat{\beta} - \beta)'X'X(\widehat{\beta} - \beta)$.
   (c) Explicar porque la ecuación del item anterior se minimiza en $\beta = \widehat{\beta}$.

2. Mostrar que
   (a) si el modelo incluye una constante $\sum_{i=1}^{n}(y_i - \widehat{y}_i) = 0$.
   (b) si $rg(X) = p$ entonces $\sum_{i=1}^{n} \widehat{y}_i(y_i - \widehat{y}_i) = 0$.

3. Supongamos que cierta variable respuesta y depende linealmente de dos variables regresoras x_1 y x_2, de manera que se verifica el modelo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

   para $i = 1, \ldots, n$ donde los errores, ϵ_i, verifican las hipótesis habituales. Se ajusta por mínimos cuadrados el modelo $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1}$, sin tener en cuenta la segunda variable regresora. Demostrar que el estimador β_1 es, en general, sesgado y determina bajo qué condiciones se anula el sesgo.

4. Dado el modelo $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip} + \epsilon_i$ para $i = 1, \ldots, n$.
   Mostrar que el estadístico F del test $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$ puede escribirse como

$$F = \frac{R^2}{1 - R^2} \frac{n - p}{p - 1}.$$

Ejercicios Prácticos

1. Consideramos el conjunto de datos *trees* ya incluido en R que contiene datos relevados sobre 31 cerezos. Para cada uno de ellos se midió el volumen de madera que se obtuvo de ellos, la altura y el diámetro (medido desde cierta altura del suelo).

Ejercicio 1

a) Muestra que $x'(y - x\hat{\beta}) = 0$

$\hat{\beta} = (x'x)^{-1}x'y \Rightarrow x'(y - x(x'x)^{-1}x'y) = x'y - \underbrace{x'x(x'x)^{-1}}_{Id}x'y = x'y - x'y = 0$

b) Muestra que $(y - x\beta)'(y - x\beta) = (y - x\hat{\beta})'(y - x\hat{\beta}) + (\hat{\beta} - \beta)(x'x)(\hat{\beta} - \beta)$.

En la primer expresión sumo y resto $x\hat{\beta}$ entre 2 términos.

$(\underbrace{y - x\hat{\beta}}_{} + \underbrace{x\hat{\beta} - x\beta}_{})'(\underbrace{y - x\hat{\beta}}_{} + \underbrace{x\hat{\beta} - x\beta}_{}) =$

↑ distribuyo

$= (y - x\hat{\beta})'(y - x\hat{\beta}) + \underbrace{(y - x\hat{\beta})'(x\hat{\beta} - x\beta)}_{(y - x\hat{\beta})'x(\hat{\beta} - \beta)} + \underbrace{(x\hat{\beta} - x\beta)'(y - x\hat{\beta})}_{(\hat{\beta} - \beta)'x'(y - x\hat{\beta})} + \underbrace{(x\hat{\beta} - x\beta)'(x\hat{\beta} - x\beta)}_{(\hat{\beta} - \beta)'x'x(\hat{\beta} - \beta)}$

$= 0$ por el ítem (a) $\quad = 0$ por el ítem (a)

$= (y - x\hat{\beta})'(y - x\hat{\beta}) + (\hat{\beta} - \beta)'x'x(\hat{\beta} - \beta)$

(c) Como $x'x$ es de rango completo y definido positivo entonces $\forall h \neq 0$ $x'x\,h > 0$. Entonces $\forall \hat{\beta} \neq \beta$

$(\hat{\beta} - \beta)x'x(\hat{\beta} - \beta) > 0$.

Luego $(y - x\hat{\beta})'(y - x\hat{\beta}) \leq (y - x\hat{\beta})'(y - x\hat{\beta}) + h'(x'x)h \quad \forall h \neq 0$

Luego la expresión se minimiza en $\underline{\beta = \hat{\beta}}$.

$h = \hat{\beta} - \beta = 0$.

Ejercicio 2:

(a) Este resultado se desprende de forma inmediata de los resultados

sumados ya que la primera columna de la matriz de diseño tiene todos unos porque corresponde al intercept. $X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1(p-1)} \\ \vdots & \vdots & & \\ 1 & x_{n1} & \cdots & x_{n(p-1)} \end{pmatrix}$

$$X'X\hat{\beta} = X'y$$

Tenemos que

$$\sum_{i=1}^{n} r_i = \sum y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} - \cdots - \hat{\beta}_{p-1} \sum_{i=1}^{n} x_{(p-1)i} \qquad (1)$$

La primera ecuación del sistema de ecuaciones normales nos dice que

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} + \cdots + \hat{\beta}_{p-1} \sum_{i=1}^{n} x_{(p-1)i} = \sum_{i=1}^{n} y_i \qquad (2)$$

De (1) y (2) surge que

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i \Rightarrow \bar{y} = \bar{\hat{y}}$$ que es lo que queríamos ver.

(b) Si $rg(X) = p \Rightarrow \sum_{i=1}^{n} \hat{y}_i (y_i - \hat{y}_i) = 0$

Es decir que $\sum_{i=1}^{n} \hat{y}_i (y_i - \hat{y}_i) = \hat{y}' r_i = (Hy)'(1-H)y = y' \underbrace{H'}_{H}(1-H)y =$

$= y' \underbrace{(H - HH)}_{H} y = 0.$

$\underbrace{\qquad\qquad}_{=0}$

Ejercicio 3:

Al ajustar β_1 con el modelo restringido, sin la variable x_2 tenemos

que $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(y_i - \bar{y})}{(n-1)\, s_{x_1}^2}$

Calculamos la esperanza para calcular el sesgo. Recordemos que los x's son fijos

$$E(\hat{\beta}_1) = \frac{1}{s_{x_1}^2 (n-1)} \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)\, E(y_i - \bar{y}) \quad (*)$$

Esta esperanza hay q calcularla bajo el modelo completo.

Luego

$$E(y_i - \bar{y}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \underbrace{E(\varepsilon_i)}_{=0} - \frac{1}{n} \sum_{i=1}^{n} E(y_i) =$$

$E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \underbrace{E(\varepsilon_i)}_{=0}$

$$= \cancel{\beta_0} + \beta_1 x_{i1} + \beta_2 x_{i2} \cancel{+ \beta_0} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2 =$$

$$= \beta_1 (x_{i1} - \bar{x}_1) + \beta_2 (x_{i2} - \bar{x}_2)$$

Volvemos a $(*)$

$$E(\hat{\beta}_1) = \frac{1}{s_{x_1}^2 (n-1)} \sum_{i=1}^{n} (x_{i1} - \bar{x}) \left[\beta_1 (x_{i1} - \bar{x}_1) + \beta_2 (x_{i2} - \bar{x}_2) \right] =$$

$$= \beta_1 \frac{1}{s_{x_1}^2 (n-1)} \underbrace{\sum_{i=1}^{n} (x_{i1} - \bar{x})^2}_{(n-1) s_{x_1}^2} + \beta_2 \frac{1}{s_{x_1}^2 (n-1)} \sum (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

$$= \beta_1 + \beta_2 \frac{\hat{cov}(x_1, x_2)\, \cancel{(n-1)}}{s_{x_1}^2 \cancel{(n-1)}}$$

Entonces $B(\hat{\beta}_1) = E(\hat{\beta}_1) - \beta_1 = \cancel{\beta_1} + \beta_2 \dfrac{\hat{cov}(x_1, x_2)}{s_{x_1}^2} \cancel{- \beta_1} =$

$$B(\hat{\beta}_1) = \beta_2 \frac{\widetilde{cov}(x_1, x_2)}{s_{x_1}^2}$$

Para que no insesgado la covariaza de x_1 y x_2 have que ser 0. Si fueran independientes esta condición se satisface.

Ejercicio 4:

Para el modelo propuesto sabemos que

$$TSS = ESS + RSS \quad \Rightarrow \quad RSS = TSS - ESS$$

y $$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Luego $$F = \frac{ESS/p-1}{RSS/n-p} = \frac{ESS}{RSS} \frac{n-p}{p-1} =$$

↓ dividir numerador y denominador por TSS

$$= \frac{ESS/TSS}{RSS/TSS} \frac{n-p}{p-1} = \frac{R^2}{(TSS-ESS)/TSS} \frac{n-p}{p-1} =$$

$=$

Girth: Diámetro en pulgadas.

Height: Altura en pies.

Volume: Volumen de madera en pies cúbicos.

El objetivo es explicar el volumen de madera obtenido en función de las otras variables.

(a) Hacer un análisis exploratorio de las variables. Qué se observa?

(b) Pensar (usando matemática del colegio) un modelo para explicar el volumen. A partir de l proponer un modelo lineal.

(c) Ajustá el modelo propuesto

(d) Contrasta la hipótesis de que los coeficientes de la regresión son significativos a nivel 0.01.

(e) Qué valores es de esperar que tomen
estos coeficientes? Se parecen los estimadores obtenidos a estos valores?

(f) Calcula intervalos de confianza de nivel 0.9 para
los coeficientes de la regresión (recorda el comando *confint*).

(g) Contrasta, a nivel 0.9, que el valor de cada uno de los coeficientes coincide con su valor esperado según el modelo teórico que propusiste.

(h) Calcula la suma de cuadrados explicada por la regresión y la correspondiente media de cuadrados. Cuńtos grados de libertad le corresponden?

(i) Calcula la matriz de correlaciones del vector de los coeficientes estimados.

(j) Planteá un test de hipótesis (simultáneo) que tenga sentido de acuerdo al modelo teórico planteado.

2. Considerar los datos del archivo `fuel.txt`, que contiene observaciones correspondientes a las siguientes variables vinculadas al consumo de combustible en el Distrito de Columbia en el año 2001. (este conjunto de datos también se puede bajar de la library(alr3) donde hay más información sobre las variables help(fuel2001).

Drivers: número de licencias de conducir en el estado.

FuelC: combustible vendido para el uso vehicular (en miles de galones).

Income: ingreso anual personal por año (correspondiente al año 2000).

Miles: millas de autopista en el estado.

MCP: millas conducidas estimadas per cápita.

Pop: población mayor a 16 años.

Tax: impuestos estaduales a la gasolina, centavos por galón.

(a) Hacer un análisis exploratorio de los datos (medidas resúmenes, scatter plot, recordar el comando pairs).

(b) Ajustar un modelo de regresión para predecir *FuelC* a partir de *Drivers, Income, Miles, MPC, Tax*. Analizar lo observado.

(c) Repetir el análisis excluyendo la variable *Drivers*, encontrar una explicación para esto.

(d) Determinar si la regresión es significativa a nivel 5%

(e) Determinar si los coeficientes son significativos a a nivel 5%.

(f) Construir intervalos de confianza de nivel 95% para aquellos coeficientes significativos.

(g) Cuál es la matriz de covarianzas estimadas del vector $(\widehat{\beta}_2, \widehat{\beta}_3)$?

(h) Constrastar $H_0 : \beta_2 = \beta_3$,

(i) Construir una elipse de confianza de nivel 95% para el vector (β_2, β_3).

(j) Predecir el valor de FuelC para $Drivers = 2718209$, $Income = 27871$, $Miles = 78914$, $MPC = 10458.4$ y $Tax = 20$. Construir los correspondientes intervalos de confianza y predicción de nivel 90%.

(k) Analizar la tabla de varianzas para esta regresión.

3. Se busca determinar si se puede predecir el suministro de agua del sur de California en años futuros a partir de datos anteriores.

   Un factor que afecta la disponibilidad de agua es la escorrentía de la corriente. Si se pudiera predecir la escorrentía, los ingenieros, planificadores y responsables políticos podrían hacer su trabajo de manera

más eficiente. Se han utilizado modelos de regresión lineal múltiple a este respecto. El conjunto de datos*water.txt* contiene 43 años de mediciones de precipitación tomadas en seis sitios en el Valle de Owens (etiquetados como APMAM, APSAB, APSLAKE, OPBPC, OPRC y OPSLAKE) medido en pulgadas, y un volumen de escorrentía en un sitio cerca de Bishop, California (BSAAM) medido en acree-feet .

(a) Hacer un análisis exploratorio de los datos (medidas resúmenes, scatter plot, recordar el comando pairs).

(b) Considerar la regresión de la variable de respuesta BSAAM en función de OPBPC, OPRC y OPSLAKE.

(c) Examinar los diagramas de dispersión de las variables, sus matrices de correlación y explicar los tests de los coeficientes individuales.

(d) Realizar el test F, determinar si la regresión es significativa.

(e) Usando la salida del último problema, testear la hipótesis de que los coeficientes para OPRC y OPBPC son cero vs la alternativa de que alguno es distinto de cero.

(f) Concluir con que modelo te quedarías y en ese caso dar los intervalos de confianza para los parámetros de nivel 95%. Además construir los intervalos de confianza y predicción de la variable de respuesta de nivel 95% para los valores medios de las variables regresoras.

1 2

4. Consideramos el conjunto de datos `esperanza2015.txt` extraídos de `https://www.gapminder.org/data/` que presenta las variables que se describen debajo para 187 países.

life: esperanza de vida al nacer (en años).

child: tasa de mortalidad de 0 a 5 años, por cada mil niños nacidos vivos en el año.

income: producto bruto interno per cápita (en USD).

[1]Escorrentía= Corriente de agua que se vierte al rebasar su depósito o cauce.
[2]Acre.feet= medida de volumen que equivale a 1.233,4818375475 metros cúbicos.

dtp3: porcentaje de niños de un año inmunizados con tres dosis de vacuna contra la difteria, tétanos y pertussi (triple bacteriana) (DTP3) medida en 2010.

school: años de escolaridad promedio en hombres de 25-34 años. estándar.

status: grado de desarrollo del país (developed - not developed).

El objetivo de este ejercicio es modelizar la variables `life`

(a) Hacer un análisis exploratorio de los datos y dar las principales características. Como la variables `status` es dicotómica, se puede omitir su uso a esta altura del curso y retomarlo más adelante.

(b) Considerar los modelos

$$life = \beta_0 + \beta_1 child + \epsilon$$

y

$$life = \beta_0 + \beta_1 child + \beta_2 child^2 + \epsilon$$

Decidir cual parece más adecuado. Graficar ambos. (HINT= para el segundo modelo se puede usar el comando $I(x^2)$).

(c) Considerar el modelo

$$life = \beta_0 + \beta_1 dtp3 + \epsilon$$

Analizarlo.

(d) Hacer un modelo que incluya en forma conjunta a las variables `dtp3` y `child`. Explicar lo observado.

(e) Analizar si incorporando otras variables o quitando algunas se tiene un modelo mejor. Para ello se puede analizar el R_{ajust} y hacer las comparaciones mediante y mediante las tablas de ANOVA, que consideres pertinentes.

(f) Retomar el modelo

$$life = \beta_0 + \beta_1 child + \beta_2 child^2 + \epsilon$$

para una grilla adecuada de valores de `child` graficar en un mismo esquema los datos, el modelo ajustado y los intervalos de confianza y predicción de nivel 95% para la variable child.

5. Consideramos un subconjunto de los datos recopilados por Galton en 1886 `http://www.medicine.mcgill.ca/epidemiology/hanley/galton/notebook/index.html`, se consideraron las alturas de 963 hijos, correspondientes a 205 familias. El conjunto de datos `AlturaDalton.txt` presenta para cada familia:

father: altura del padre.

mother: altura del madre.

midparentHeight: altura promedio de la madre y el padre, calculada como (father + 1.08*mother)/2.

gender: sexo del hijo.

childHeight: altura del hijo.

donde para cada familia unicamente si informa la altura para el hijo y/o hija mayor. Si bien el objetivo inicial de Galton era estudiar la relación entre las alturas de los padres, los datos se han hecho conocidos para predecir la altura de los hijos. Considerar por separado las datos de los hijos y de las hijas.

(a) Hacer un análisis exploratorio de los datos.

(b) Hacer un modelo de regresión lineal para predecir la altura de los hijos, en función de las alturas de la madre, del padre y de la altura promedio de ambos. Observar lo que ocurre. Quitar alguna de las variables y comparar los modelos.

(c) Estudiando la descomposición de la varianza elegir el mejor modelo.

(d) En un mismo gráfico, hacer un diagrama de dispersión de los datos `childHeight` versus `midparentHeight` en diferentes colores de varones y mujeres. Superponiendo las respectivas rectas de regresión y las bandas de predicción y confianza de nivel 0.95% para la altura de hijos e hijas.