UNIVERSITAT DE BARCELONA

Universitat Oberta de Catalunya

Regresión lineal simple y múltiple

Francesc Carmona

22 de marzo de 2019

Los ejercicios con $(*)$ son opcionales, con $(**)$ además son difíciles.

Ejercicios del libro de Carmona

1. (Ejercicio 6.8 del Capítulo 6 página 118)

   Hallar la recta de regresión simple de la variable respuesta *raíz cuadrada de la velocidad* sobre la variable regresora *densidad* con los datos de la tabla 1.1 del capítulo 1.

   Comprobar las propiedades del ejercicio 6.4 sobre las predicciones $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ y los residuos $e_i = y_i - \hat{y}_i$ para estos datos.

   (i) La suma de los residuos es cero: $\sum e_i = 0$.

   (ii) $\sum y_i = \sum \hat{y}_i$

   (iii) La suma de los residuos ponderada por los valores de la variable regresora es cero: $\sum x_i e_i = 0$.

   (iv) La suma de los residuos ponderada por las predicciones de los valores observados es cero: $\sum \hat{y}_i e_i = 0$.

   Calcular la estimación de σ^2 y, a partir de ella, las estimaciones de las desviaciones estándar de los estimadores de los parámetros $\hat{\beta}_0$ y $\hat{\beta}_1$.

   Escribir los intervalos de confianza para los parámetros con un nivel de confianza del 95 %.

   Construir la tabla para la significación de la regresión y realizar dicho contraste.

   Hallar el intervalo de la predicción de la respuesta media cuando la densidad es de 50 vehículos por km. Nivel de confianza: 90 %.

2. $(*)$ (Ejercicio 6.9 del Capítulo 6 página 118)

   Comparar las rectas de regresión de hombres y mujeres con los logaritmos de los datos del ejercicio 1.4.

3. (Ejercicio 6.10 del Capítulo 6 página 118)

   Se admite que una persona es *proporcionada* si su altura en cm es igual a su peso en kg más 100. En términos estadísticos si la recta de regresión de Y (altura) sobre X (peso) es

   $$Y = 100 + X$$

   Contrastar, con un nivel de significación $\alpha = 0.05$, si se puede considerar válida esta hipótesis a partir de los siguientes datos que corresponden a una muestra de mujeres jóvenes:

X :	55	52	65	54	46	60	54	52	56	65	52	53	60
Y :	164	164	173	163	157	168	171	158	169	172	168	160	172

   Razonar la bondad de la regresión y todos los detalles del contraste.

4. (Ejercicio 6.11 del Capítulo 6 página 119)

   El período de oscilación de un péndulo es $2\pi\sqrt{\frac{l}{g}}$, donde l es la longitud y g es la constante de gravitación. En un experimento observamos t_{ij} $(j = 1, \ldots, n_i)$ períodos correspondientes a l_i $(i = 1, \ldots, k)$ longitudes.

   (a) Proponer un modelo, con las hipótesis que se necesiten, para estimar la constante $\frac{2\pi}{\sqrt{g}}$ por el método de los mínimos cuadrados.

   (b) En un experimento se observan los siguientes datos:

longitud	período
18.3	8.58 7.9 8.2 7.8
20	8.4 9.2
21.5	9.7 8.95 9.2
15	7.5 8

   Contrastar la hipótesis $H_0 : \frac{2\pi}{\sqrt{g}} = 2$.

5. (Ejercicio 8.4 del Capítulo 8 página 157)

   Se dispone de los siguientes datos sobre diez empresas fabricantes de productos de limpieza doméstica:

Empresa	V	IP	PU
1	60	100	1.8
2	48	110	2.4
3	42	130	3.6
4	36	100	0.6
5	78	80	1.8
6	36	80	0.6
7	72	90	3.6
8	42	120	1.2
9	54	120	2.4
10	90	90	4.2

   En el cuadro anterior, V son las ventas anuales, expresadas en millones de euros, IP es un índice de precios relativos (Precios de la empresa/Precios de la competencia) y PU son los gastos anuales realizados en publicidad y campañas de promoción y difusión, expresados también en millones de euros.

   Tomando como base la anterior información:

   1) Estimar el vector de coeficientes $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ del modelo

   $$V_i = \beta_0 + \beta_1 IP_i + \beta_2 PU_i + \epsilon_i$$

   2) Estimar la matriz de varianzas-covarianzas del vector $\widehat{\boldsymbol{\beta}}$.

   3) Calcular el coeficiente de determinación.

6. $(*)$ (Ejercicio 8.5 del Capítulo 8 página 157)

   Dado el modelo

   $$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$$

   y los siguientes datos

Y_t	X_{1t}	X_{2t}
10	1	0
25	3	−1
32	4	0
43	5	1
58	7	−1
62	8	0
67	10	−1
71	10	2

obtener:

(a) La estimación MC de $\beta_0, \beta_1, \beta_2$ utilizando los valores originales.

(b) La estimación MC de $\beta_0, \beta_1, \beta_2$ utilizando los datos expresados en desviaciones respecto a la media.

(c) La estimación insesgada de σ^2.

(d) El coeficiente de determinación.

(e) El coeficiente de determinación corregido.

(f) El contraste de la hipótesis nula $H_0 : \beta_0 = \beta_1 = \beta_2 = 0$.

(g) El contraste de la hipótesis nula $H_0 : \beta_1 = \beta_2 = 0$ utilizando datos originales.

(h) El contraste de la hipótesis nula $H_0 : \beta_1 = \beta_2 = 0$ utilizando datos en desviaciones respecto a la media.

(i) La representación gráfica de una región de confianza del 95 % para β_1 y β_2.

(j) El contraste individual de los parámetros β_0, β_1 y β_2.

(k) El contraste de la hipótesis nula $H_0 : \beta_1 = 10\beta_2$.

(l) El contraste de la hipótesis nula $H_0 : 2\beta_0 + 2\beta_1 + 7\beta_2 = 50$.

(m) El contraste de la hipótesis nula conjunta $H_0 : \beta_1 = 10\beta_2, 2\beta_0 + 2\beta_1 + 7\beta_2 = 50$.

Ejercicios del libro de Faraway

1. (Ejercicio 1 cap. 4 pág. 56)

   For the `prostate` data, fit a model with `lpsa` as the response and the other variables as predictors:

   (a) Suppose a new patient with the following values arrives:

   ```
    lcavol lweight      age    lbph     svi      lcp
   1.44692 3.62301 65.00000 0.30010 0.00000 -0.79851
   gleason    pgg45
   7.00000 15.00000
   ```

   Predict the `lpsa` for this patient along with an appropriate 95% CI.

   (b) Repeat the last question for a patient with the same values except that he is age 20. Explain why the CI is wider.

   (c) For the model of the previous question, remove all the predictors that are not significant at the 5% level. Now recompute the predictions of the previous question. Are the CIs wider or narrower? Which predictions would you prefer? Explain.

2. (Ejercicio 2 cap. 4 pág. 57)

   Using the `teengamb` data, fit a model with `gamble` as the response and the other variables as predictors.

(a) Predict the amount that a male with average (given these data) status, income and verbal score would gamble along with an appropriate 95% CI.

(b) Repeat the prediction for a male with maximal values (for this data) of status, income and verbal score. Which CI is wider and why is this result expected?

(c) Fit a model with `sqrt(gamble)` as the response but with the same predictors. Now predict the response and give a 95% prediction interval for the individual in (a). Take care to give your answer in the original units of the response.

(d) Repeat the prediction for the model in (c) for a female with `status = 20`, `income = 1`, `verbal = 10`. Comment on the credibility of the result.

3. (Ejercicio 3 cap. 4 pág. 57)

The `snail` dataset contains percentage water content of the tissues of snails grown under three different levels of relative humidity and two different temperatures.

(a) Use the command `xtabs(water ~ temp + humid, snail)/4` to produce a table of mean water content for each combination of temperature and humidity. Can you use this table to predict the water content for a temperature of 25°C and a humidity of 60%? Explain.

(b) Fit a regression model with the water content as the response and the temperature and humidity as predictors. Use this model to predict the water content for a temperature of 25°C and a humidity of 60%?

(c) Use this model to predict water content for a temperature of 30°C and a humidity of 75%? Compare your prediction to the prediction from (a). Discuss the relative merits of these two predictions.

(d) The intercept in your model is 52.6%. Give two values of the predictors for which this represents the predicted response. Is your answer unique? Do you think this represents a reasonable prediction?

(e) For a temperature of 25°C, what value of humidity would give a predicted response of 80% water content.

4. (∗) (Ejercicio 4 cap. 4 pág. 57)

The dataset `mdeaths` reports the number of deaths from lung diseases for men in the UK from 1974 to 1979.

(a) Make an appropriate plot of the data. At what time of year are deaths most likely to occur?

(b) Fit an autoregressive model of the same form used for the airline data. Are all the predictors statistically significant?

(c) Use the model to predict the number of deaths in January 1980 along with a 95% prediction interval.

(d) Use your answer from the previous question to compute a prediction and interval for February 1980.

(e) Compute the fitted values. Plot these against the observed values. Note that you will need to select the appropriate observed values. Do you think the accuracy of predictions will be the same for all months of the year?

5. (∗) (Ejercicio 5 cap. 4 pág. 58)

For the `fat` data used in this chapter, a smaller model using only `age`, `weight`, `height` and `abdom` was proposed on the grounds that these predictors are either known by the individual or easily measured.

(a) Compare this model to the full thirteen-predictor model used earlier in the chapter. Is it justifiable to use the smaller model?

(b) Compute a 95% prediction interval for median predictor values and compare to the results to the interval for the full model. Do the intervals differ by a practically important amount?

(c) For the smaller model, examine all the observations from case numbers 25 to 50. Which two observations seem particularly anomalous?

(d) Recompute the 95% prediction interval for median predictor values after these two anomalous cases have been excluded from the data. Did this make much difference to the outcome?