UNIVERSITAT DE BARCELONA

Universitat Oberta de Catalunya

Estimación del modelo lineal

Francesc Carmona

27 de febrero de 2019

Ejercicios del libro de Faraway

1. (Ejercicio 1 cap. 2 pág. 30)

   The dataset `teengamb` concerns a study of teenage gambling in Britain. Fit a regression model with the expenditure on gambling as the response and the sex, status, income and verbal score as predictors. Present the output.

   (a) What percentage of variation in the response is explained by these predictors?

   (b) Which observation has the largest (positive) residual? Give the case number.

   (c) Compute the mean and median of the residuals.

   (d) Compute the correlation of the residuals with the fitted values.

   (e) Compute the correlation of the residuals with the income.

   (f) For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?

2. (Ejercicio 2 cap. 2 pág. 30)

   The dataset `uswages` is drawn as a sample from the Current Population Survey in 1988. Fit a model with weekly wages as the response and years of education and experience as predictors. Report and give a simple interpretation to the regression coefficient for years of education. Now fit the same model but with logged weekly wages. Give an interpretation to the regression coefficient for years of education. Which interpretation is more natural?

3. ($*$) (Ejercicio 3 cap. 2 pág. 30)

   In this question, we investigate the relative merits of methods for computing the coefficients. Generate some artificial data by:

   ```
   > x <- 1:20
   > y <- x+rnorm(20)
   ```

   Fit a polynomial in `x` for predicting `y`. Compute $\hat{\beta}$ in two ways — by `lm()` and by using the direct calculation described in the chapter. At what degree of polynomial does the direct calculation method fail? (Note the need for the `I()` function in fitting the polynomial, that is, `lm(y ~ x + I(x^2))`.

4. (Ejercicio 4 cap. 2 pág. 30)

   The dataset `prostate` comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Fit a model with `lpsa` as the response and `lcavol` as the predictor. Record the residual standard error and the R^2. Now add `lweight`, `svi`, `lbph`, `age`, `lcp`, `pgg45` and `gleason` to the model one at a time. For each model record the residual standard error and the R^2. Plot the trends in these two statistics.

5. (Ejercicio 5 cap. 2 pág. 30)

   Using the `prostate` data, plot `lpsa` against `lcavol`. Fit the regressions of `lpsa` on `lcavol` and `lcavol` on `lpsa`. Display both regression lines on the plot. At what point do the two lines intersect?

6. (Ejercicio 6 cap. 2 pág. 30)

   Thirty samples of cheddar cheese were analyzed for their content of acetic acid, hydrogen sulfide and lactic acid. Each sample was tasted and scored by a panel of judges and the average taste score produced. Use the `cheddar` data to answer the following:

   (a) Fit a regression model with taste as the response and the three chemical contents as predictors. Report the values of the regression coefficients.

   (b) Compute the correlation between the fitted values and the response. Square it. Identify where this value appears in the regression output.

   (c) Fit the same regression model but without an intercept term. What is the value of R^2 reported in the output? Compute a more reasonable measure of the good- ness of fit for this example.

   (d) Compute the regression coefficients from the original fit using the QR decomposition showing your **R** code.

7. (Ejercicio 7 cap. 2 pág. 31)

   An experiment was conducted to determine the effect of four factors on the resistivity of a semiconductor wafer. The data is found in `wafer` where each of the four factors is coded as $-$ or $+$ depending on whether the low or the high setting for that factor was used. Fit the linear model `resist ~ x1 + x2 + x3 + x4`.

   (a) Extract the X matrix using the `model.matrix` function. Examine this to determine how the low and high levels have been coded in the model.

   (b) Compute the correlation in the X matrix. Why are there some missing values in the matrix?

   (c) What difference in resistance is expected when moving from the low to the high level of `x1`?

   (d) Refit the model without `x4` and examine the regression coefficients and standard errors? What stayed the the same as the original fit and what changed?

   (e) Explain how the change in the regression coefficients is related to the correlation matrix of X.

8. ($*$) (Ejercicio 8 cap. 2 pág. 31)

   An experiment was conducted to examine factors that might affect the height of leaf springs in the suspension of trucks. The data may be found in `truck`. The five factors in the experiment are set to $-$ and $+$ but it will be more convenient for us to use -1 and $+1$. This can be achieved for the first factor by:

   ```
   truck$B <- sapply(truck$B, function(x) ifelse(x == "-",-1,1))
   ```

   Repeat for the other four factors.

   (a) Fit a linear model for the height in terms of the five factors. Report on the value of the regression coefficients.

   (b) Fit a linear model using just factors `B`, `C`, `D` and `E` and report the coefficients. How do these compare to the previous question? Show how we could have anticipated this result by examining the X matrix.

   (c) Construct a new predictor called `A` which is set to `B+C+D+E`. Fit a linear model with the predictors `A`, `B`, `C`, `D`, `E` and `O`. Do coefficients for all six predictors appear in the regression summary? Explain.

   (d) Extract the model matrix X from the previous model.
   Attempt to compute $\hat{\beta}$ from $(X^T X)^{-1} X^T y$. What went wrong and why?

(e) Use the QR decomposition method as seen in Section 2.7 to compute $\hat{\beta}$. Are the results satisfactory?

(f) Use the function `qr.coef` to correctly compute $\hat{\beta}$.

Ejercicios del libro de Carmona

1. (Ejercicio 2.1 del Capítulo 2 página 41)

   Una variable Y toma los valores y_1, y_2 y y_3 en función de otra variable X con los valores x_1, x_2 y x_3. Determinar cuales de los siguientes modelos son lineales y encontrar, en su caso, la matriz de diseño para $x_1 = 1$, $x_2 = 2$ y $x_3 = 3$.

   a) $y_i = \beta_0 + \beta_1 x_i + \beta_2(x_i^2 - 1) + \epsilon_i$

   b) $y_i = \beta_0 + \beta_1 x_i + \beta_2 e^{x_i} + \epsilon_i$

   c) $y_i = \beta_1 x_i(\beta_2 \text{tang}(x_i)) + \epsilon_i$

2. (Ejercicio 2.4 del Capítulo 2 página 42)

   Cuatro objetos cuyos pesos exactos son β_1, β_2, β_3 y β_4 han sido pesados en una balanza de platillos de acuerdo con el siguiente esquema:

β_1	β_2	β_3	β_4	peso
1	1	1	1	9.2
1	−1	1	1	8.3
1	0	0	1	5.4
1	0	0	−1	−1.6
1	0	1	1	8.7
1	1	−1	1	3.5

   Hallar las estimaciones de cada β_i y de la varianza del error.

3. (∗) (Ejercicio 3.2 del Capítulo 3 página 54)

   En un modelo lineal, la matriz de diseño es

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \end{pmatrix}$$

   Hallar la expresión general de las funciones paramétricas estimables.

4. (∗) (Ejercicio 3.7 del Capítulo 3 página 55)

   Consideremos el modelo lineal

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 + \epsilon_1 \\ y_2 &= \beta_1 + \beta_3 + \epsilon_2 \\ y_3 &= \beta_1 + \beta_2 + \epsilon_3 \end{aligned}$$

   Se pide:

   1) ¿Es la función paramétrica

$$\psi = \beta_1 + \beta_2 + \beta_3$$

   estimable?

   2) Probar que toda función paramétrica

$$\psi = a_1\beta_1 + a_2\beta_2 + a_3\beta_3$$

   es estimable si y sólo si $a_1 = a_2 + a_3$.

5. ($*$) (Ejercicio 3.8 del Capítulo 3 página 56)

   Consideremos el modelo lineal

$$\begin{aligned}
y_1 &= \mu + \alpha_1 + \beta_1 + \epsilon_1 \\
y_2 &= \mu + \alpha_1 + \beta_2 + \epsilon_2 \\
y_3 &= \mu + \alpha_2 + \beta_1 + \epsilon_3 \\
y_4 &= \mu + \alpha_2 + \beta_2 + \epsilon_4 \\
y_5 &= \mu + \alpha_3 + \beta_1 + \epsilon_5 \\
y_6 &= \mu + \alpha_3 + \beta_2 + \epsilon_6
\end{aligned}$$

   (a) ¿Cuando es $\lambda_0\mu + \lambda_1\alpha_1 + \lambda_2\alpha_2 + \lambda_3\alpha_3 + \lambda_4\beta_1 + \lambda_5\beta_2$ estimable?

   (b) ¿Es $\alpha_1 + \alpha_2$ estimable?

   (c) ¿Es $\beta_1 - \beta_2$ estimable?

   (d) ¿Es $\mu + \alpha_1$ estimable?

   (e) ¿Es $6\mu + 2\alpha_1 + 2\alpha_2 + 2\alpha_3 + 3\beta_1 + 3\beta_2$ estimable?

   (f) ¿Es $\alpha_1 - 2\alpha_2 + \alpha_3$ estimable?

   (g) Hallar la covarianza entre los estimadores lineales MC de las funciones paramétricas $\beta_1 - \beta_2$ y $\alpha_1 - \alpha_2$, si éstas son estimables.

   (h) Hallar la dimensión del espacio paramétrico.

   (i) Obtener una expresión del espacio de los errores.

6. ($*$) (Ejercicio 3.10 del Capítulo 3 página 56)

   Un transportista realiza diversos trayectos entre tres poblaciones A, B y C. En cuatro dias consecutivos ha hecho los recorridos que muestra la siguiente tabla:

trayecto	**km**
$A \to B \to A \to C$	533
$C \to A \to C \to B$	583
$B \to C \to A \to C \to A \to B \to A$	1111
$A \to B \to A \to C \to A \to B \to A$	1069

   donde el kilometraje es, por diversas causas, aproximado.

   (a) Proponer un modelo lineal, con la matriz de diseño y las hipótesis necesarias, para estimar las distancias kilométricas entre las tres poblaciones.

   Con los datos proporcionados, ¿es posible estimar las distancias entre las tres poblaciones? ¿Cuales son las distancias o funciones paramétricas estimables (fpe) en este modelo?

   (b) ¿Se puede estimar el kilometraje del trayecto $M_{BC} \to B \to A \to C \to M_{AC}$, donde M_{IJ} es el punto medio entre dos poblaciones? ¿Es una buena estimación? ¿Cual es el *error* de esta estimación?