Inferencia

Francesc Carmona

7 de marzo de 2019

Los ejercicios con (∗) son opcionales, con (∗∗) además son difíciles.

Ejercicios del libro de Faraway

1. (Ejercicio 1 cap. 3 pág. 48)

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

   (a) Compute 90 and 95% CIs for the parameter associated with `age`. Using just these intervals, what could we have deduced about the p-value for `age` in the regression summary?

   (b) Compute and display a 95% joint confidence region for the parameters associated with `age` and `lbph`. Plot the origin on this display. The location of the origin on the display tells us the outcome of a certain hypothesis test. State that test and its outcome.

   (c) In the text, we made a permutation test corresponding to the F-test for the significance of all the predictors. Execute the permutation test corresponding to the t-test for `age` in this model. (Hint: `summary(g)$coef[4,3]` gets you the t-statistic you need if the model is called `g`.)

   (d) Remove all the predictors that are not significant at the 5% level. Test this model against the original model. Which model is preferred?

2. (Ejercicio 2 cap. 3 pág. 49)

   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. Identify the predictors that are statistically significant at the 5% level.

   (b) `Acetic` and `H2S` are measured on a log scale. Fit a linear model where all three predictors are measured on their original scale. Identify the predictors that are statistically significant at the 5% level for this model.

   (c) Can we use an F-test to compare these two models? Explain. Which model provides a better fit to the data? Explain your reasoning.

   (d) If `H2S` is increased 0.01 for the model used in (a), what change in the `taste` would be expected?

   (e) What is the percentage change in `H2S` on the original scale corresponding to an additive increase of 0.01 on the (natural) log scale?

3. (Ejercicio 3 cap. 3 pág. 49)

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

   (a) Which variables are statistically significant at the 5% level?

   (b) What interpretation should be given to the coefficient for `sex`?

(c) Fit a model with just `income` as a predictor and use an F-test to compare it to the full model.

4. (Ejercicio 4 cap. 3 pág. 49)

   Using the `sat` data:

   (a) Fit a model with `total` sat score as the response and `expend`, `ratio` and `salary` as predictors. Test the hypothesis that $\beta_{salary} = 0$. Test the hypothesis that $\beta_{salary} = \beta_{ratio} = \beta_{expend} = 0$. Do any of these predictors have an effect on the response?

   (b) Now add `takers` to the model. Test the hypothesis that $\beta_{takers} = 0$. Compare this model to the previous one using an F-test. Demonstrate that the F-test and t-test here are equivalent.

5. (∗) (Ejercicio 5 cap. 3 pág. 50)

   Find a formula relating R^2 and the F-test for the regression.

6. (∗) (Ejercicio 6 cap. 3 pág. 50)

   Thirty-nine MBA students were asked about happiness and how this related to their income and social life. The data are found in `happy`. Fit a regression model with `happy` as the response and the other four variables as predictors.

   (a) Which predictors were statistically significant at the 1% level?

   (b) Use the `table()` function to produce a numerical summary of the response. What assumption used to perform the t-tests seems questionable in light of this summary?

   (c) Use the permutation procedure described in Section 3.3 to test the significance of the `money` predictor.

   (d) Plot a histgram of the permutation t-statistics. Make sure you use the the probability rather than frequency version of the histogram.

   (e) Overlay an appropriate t-density over the histogram.
   Hint: Use `grid <- seq(-3, 3, length = 300)` to create a grid of values, then use the `dt()` function to compute the t-density on this grid and the `lines()` function to superimpose the result.

   (f) Use the bootstrap procedure from Section 3.6 to compute 90% and 95% confidence intervals for β_{money}. Does zero fall within these confidence intervals? Are these results consistent with previous tests?

7. (∗) (Ejercicio 7 cap. 3 pág. 50)

   In the `punting` data, we find the average distance punted and hang times of 10 punts of an American football as related to various measures of leg strength for 13 volunteers.

   (a) Fit a regression model with `Distance` as the response and the right and left leg strengths and flexibilities as predictors. Which predictors are significant at the 5% level?

   (b) Use an F-test to determine whether collectively these four predictors have a relationship to the response.

   (c) Relative to the model in (a), test whether the right and left leg strengths have the same effect.

   (d) Construct a 95% confidence region for $(\beta_{Str}, \beta_{LStr})$. Explain how the test in (c) relates to this region.

   (e) Fit a model to test the hypothesis that it is total leg strength defined by adding the right and left leg strengths that is sufficient to predict the response in comparison to using individual left and right leg strengths.

   (f) Relative to the model in (a), test whether the right and left leg flexibilities have the same effect.

   (g) Test for left-right symmetry by performing the tests in (c) and (f) simultaneously.

   (h) Fit a model with `Hang` as the response and the same four predictors. Can we make a test to compare this model to that used in (a)? Explain.

Ejercicios del libro de Carmona

1. (∗∗) (Ejercicio 5.1 del Capítulo 5 página 86)

   Sean $X \sim N(\mu_1, \sigma)$, $Y \sim N(\mu_2, \sigma)$ variables independientes. En muestras de extensión n_1 de X, n_2 de Y, plantear la hipótesis nula

   $$H_0 : \mu_1 = \mu_2$$

   mediante el concepto de hipótesis lineal contrastable y deducir el test t de Student de comparación de medias como una consecuencia del test F.

2. (∗) (Ejercicio 5.2 del Capítulo 5 página 86)

   Una variable Y depende de otra x (variable control no aleatoria) que toma los valores $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, $x_4 = 4$ de acuerdo con el modelo lineal normal

   $$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

   Encontrar la expresión del estadístico F para la hipótesis

   $$H_0 : \beta_2 = 0$$

3. (∗) (Ejercicio 5.5 del Capítulo 5 página 87)

   Dado el siguiente modelo lineal normal

   $$\begin{aligned} \beta_1 + \beta_2 &= 6.6 \\ 2\beta_1 + \beta_2 &= 7.8 \\ -\beta_1 + \beta_2 &= 2.1 \\ 2\beta_1 - \beta_2 &= 0.4 \end{aligned}$$

   estudiar si se puede aceptar la hipótesis $H_0 : \beta_2 = 2\beta_1$.

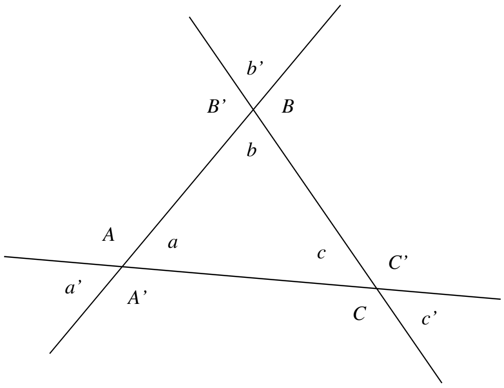
4. (∗) (Ejercicio 5.6 del Capítulo 5 página 87)

   Continuación del ejercicio 3.10 de Estimación:

   El transportista discute con un amigo que afirma que el doble de la distancia entre A y B es equivalente a la distancia del trayecto $A \to C \to B$. ¿Podemos aclarar en términos estadísticos su discusión?

5. (∗∗) (Ejercicio 5.10 del Capítulo 5 página 88)

   Supongamos que cada uno de los valores $x_1, x_2, \ldots, x_{12}$ son las observaciones de los ángulos a,a',A, A',b,b',B,B',c,c',C,C' del triángulo del gráfico adjunto. Los errores de las observaciones $\epsilon_1, \ldots, \epsilon_{12}$ se asume que son independientes y con distribución $N(0, \sigma)$.



Antes de escribir el modelo asociado a estos datos observemos que, aunque aparentemente hay 12 parámetros $a, a', \ldots,$ éstos están ligados por las conocidas propiedades de un triángulo, es decir

$$a = a' \qquad A = A' \qquad a + A = 180 \qquad a + b + c = 180$$

y de forma similar para b, b', B, B' y c, c', C, C'. El conjunto de estas relaciones nos conduce a que, realmente, sólo hay dos parámetros independientes, les llamaremos α y β. Si trasladamos a la izquierda las cantidades 180 y con estos parámetros, el modelo es

$$\begin{array}{llll}
y_1 = \alpha + \epsilon_1 & y_2 = \alpha + \epsilon_2 & y_3 = -\alpha + \epsilon_3 & y_4 = -\alpha + \epsilon_4 \\
y_5 = \beta + \epsilon_5 & y_6 = \beta + \epsilon_6 & y_7 = -\beta + \epsilon_7 & y_8 = -\beta + \epsilon_8 \\
y_9 = -\alpha - \beta + \epsilon_9 & y_{10} = -\alpha - \beta + \epsilon_{10} & y_{11} = \alpha + \beta + \epsilon_{11} & y_{12} = \alpha + \beta + \epsilon_{12}
\end{array}$$

donde

$$\begin{array}{llll}
y_1 = x_1 & y_2 = x_2 & y_3 = x_3 - 180 & y_4 = x_4 - 180 \\
y_5 = x_5 & y_6 = x_6 & y_7 = x_7 - 180 & y_8 = x_8 - 180 \\
y_9 = x_9 - 180 & y_{10} = x_{10} - 180 & y_{11} = x_{11} & y_{12} = x_{12}
\end{array}$$

Deseamos contrastar la hipótesis de que el triángulo es equilátero, es decir, que $a = b = c = 60$. Pero si $a = 60, b = 60$, c es automáticamente 60, luego la hipótesis es

$$H_0 : \alpha = \beta = 60$$

con 2 grados de libertad, no 3. Resolver el contraste.

Otros ejercicios

1. En los ejemplos 5.3.2 y 5.6.3 del libro de Carmona y con los datos del diseño cross-over simplificado considerar el modelo en el que el efecto de la interacción es distinto cuando primero se administra el tratamiento **a** y a continuación el tratamiento **b**, que cuando se hace al revés. Es decir, hay dos parámetros distintos γ_{ab} y γ_{ba}.

   Contrastar en ese modelo la hipótesis $H_0 : \gamma_{ab} = \gamma_{ba}$. Comprobar primero que es una hipótesis contrastable.