UNIVERSITAT DE BARCELONA

Universitat Oberta de Catalunya

Métodos alternativos de regresión

Francesc Carmona

23 de abril de 2019

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

Ejercicios del libro de Faraway

1. (Ejercicio 7 cap. 7 pág. 110)

   Use the `happy` dataset with `happy` as the response and the other variables as predictors. Discuss possible rescalings of the variables with the aim of helping the interpretation of the model fit.

2. (Ejercicio 1 cap. 8 pág. 130)

   Researchers at National Institutes of Standards and Technology (NIST) collected `pipeline` data on ultrasonic measurements of the depth of defects in the Alaska pipeline in the field. The depth of the defects were then remeasured in the laboratory. These measurements were performed in six different batches. It turns out that this batch effect is not significant and so can be ignored in the analysis that follows. The laboratory measurements are more accurate than the in-field measurements, but more time consuming and expensive. We want to develop a regression equation for correcting the in-field measurements.

   (a) Fit a regression model `Lab ~ Field`. Check for non-constant variance.

   (b) We wish to use weights to account for the non-constant variance. Here we split the range of `Field` into 12 groups of size nine (except for the last group which has only eight values). Within each group, we compute the variance of `Lab` as `varlab` and the mean of `Field` as `meanfield`. Supposing `pipeline` is the name of your data frame, the following **R** code will make the needed computations:

   ```
   > i <- order(pipeline$Field)
   > npipe <- pipeline[i,]
   > ff <- gl(12,9)[-108]
   > meanfield <- unlist(lapply(split(npipe$Field,ff),mean))
   > varlab <- unlist(lapply(split(npipe$Lab,ff),var))
   ```

   Suppose we guess that the the variance in the response is linked to the predictor in the following way:

   $$\text{var}(Lab) = a_0 Field^{a_1}$$

   Regress `log(varlab)` on `log(meanfield)` to estimate a_0 and a_1. (You might choose to remove the last point.) Use this to determine appropriate weights in a WLS fit of `Lab` on `Field`. Show the regression summary.

   (c) An alternative to weighting is transformation. Find transformations on `Lab` and/or `Field` so that in the transformed scale the relationship is approximately linear with constant variance. You may restrict your choice of transformation to square root, log and inverse.

3. (Ejercicio 2 cap. 8 pág. 131)

   Using the `divusa` data, fit a regression model with `divorce` as the response and `unemployed`, `femlab`, `marriage`, `birth` and `military` as predictors.

(a) Make two graphical checks for correlated errors. What do you conclude?

(b) Allow for serial correlation with an AR(1) model for the errors. (Hint: Use maximum likelihood to estimate the parameters in the GLS fit by `gls(...,method="ML", ...)`). What is the estimated correlation and is it significant? Does the GLS model change which variables are found to be significant?

(c) Speculate why there might be correlation in the errors.

4. (Ejercicio 3 cap. 8 pág. 131)

   For the `salmonella` dataset, fit a linear model with `colonies` as the response and `log(dose+1)` as the predictor. Check for lack of fit.

5. (∗) (Ejercicio 4 cap. 8 pág. 131)

   For the `cars` dataset, fit a linear model with `distance` as the response and `speed` as the predictor. Check for lack of fit.

6. (Ejercicio 5 cap. 8 pág. 131)

   Using the `stackloss` data, fit a model with `stack.loss` as the response and the other three variables as predictors using the following methods:

   (a) Least squares

   (b) Least absolute deviations

   (c) Huber method

   (d) Least trimmed squares

   Compare the results. Now use diagnostic methods to detect any outliers or influential points. Remove these points and then use least squares. Compare the results.

7. (∗) (Ejercicio 6 cap. 8 pág. 131)

   Using the `cheddar` data, fit a linear model with `taste` as the response and the other three variables as predictors.

   (a) Suppose that the observations were taken in time order. Create a time variable. Plot the residuals of the model against time and comment on what can be seen.

   (b) Fit a GLS model with same form as above but now allow for an AR(1) correlation among the errors. Is there evidence of such a correlation?

   (c) Fit a LS model but with time now as an additional predictor. Investigate the significance of time in the model.

   (d) The last two models have both allowed for an effect of time. Explain how they do this differently.

   (e) Suppose you were told, contrary to prior information, that the observations are not in time order. How would this change your interpretation of the model from (c)?

8. (Ejercicio 7 cap. 8 pág. 132)

   The `crawl` dataset contains data on a study looking at the age when babies learn to crawl as a function of ambient temperatures. There is additional information about the number of babies studied each month and the variation in the response. Make an appropriate choice of weights to investigate the relationship between crawling age and temperature.

9. (∗) (Ejercicio 8 cap. 8 pág. 132)

   The `gammaray` dataset shows the x-ray decay light curve of a gamma ray burst. Build a model to predict the flux as a function time that uses appropriate weights.

10. (Ejercicio 9 cap. 8 pág. 132)

    Use the `fat` data, fitting the model described in Section 4.2.

    (a) Fit the same model but now using Huber's robust method. Comment on any substantial differences between this model and the least squares fit.

    (b) Identify which two cases have the lowest weights in the Huber fit. What is unusual about these two points?

    (c) Plot weight (of the man) against height. Identify the two outlying cases. Are these the same as those identified in the previous question? Discuss.

11. (Ejercicio 1 cap. 9 pág. 145)

    The `aatemp` data come from the U.S. Historical Climatology Network. They are the annual mean temperatures (in degrees F) in Ann Arbor, Michigan going back about 150 years.

    (a) Is there a linear trend?

    (b) Observations in successive years may be correlated. Fit a model that estimates this correlation. Does this change your opinion about the trend?

    (c) Fit a polynomial model with degree 10 and use backward elimination to reduce the degree of the model. Plot your fitted model on top of the data. Use this model to predict the temperature in 2020.

    (d) Suppose someone claims that the temperature was constant until 1930 and then began a linear trend. Fit a model corresponding to this claim. What does the fitted model say about this claim?

    (e) Make a cubic spline fit with six basis functions evenly spaced on the range. Plot the fit in comparison to the previous fits. Does this model fit better than the straight-line model?

12. (Ejercicio 6 cap. 9 pág. 145)

    Use the `odor` data for this question.

    (a) Fit a second order response surface for the odor response using the other three variables as predictors. How many parameters does this model use and how many degrees of freedom are left?

    (b) Fit a model for the same response but now excluding any interaction terms but including linear and quadratic terms in all three predictors. Compare this model to the previous one. Is this simplification justified?

    (c) Use the previous model to determine the values of the predictors which result in the minimum predicted odor.

Ejercicios del libro de Carmona

1. (∗) (Ejercicio 7.1 del Capítulo 7 página 131)

    Calcular los parámetros de la recta resistente de los tres grupos con los datos del ejercicio 6.10. Discutir su necesidad frente a la recta mínimocuadrática.

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

2. (∗) (Ejercicio 7.2 del Capítulo 7 página 131)

    Calcular la recta resistente de los tres grupos con los datos (c) de la tabla 6.4. Comparar el resultado con el de la tabla 6.5 y con la recta de regresión mínimocuadrática sin la observación 16.

3. (∗) (Ejercicio 10.1 del Capítulo 10 página 183)

   Calcular la regresión con M-estimadores y diferentes métodos o funciones de ponderación a elegir de la tabla 10.2 sobre los datos del ejercicio 6.10. Discutir su necesidad frente a la recta mínimocuadrática.

   Realizar también una regresión robusta mínimocuadrática recortada.

4. (∗) (Ejercicio 10.2 del Capítulo 10 página 183)

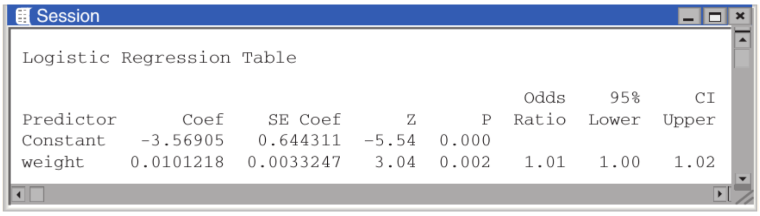
   Calcular la regresión con M-estimadores y diferentes métodos o funciones de ponderación a elegir de la tabla 10.2 sobre los datos (c) de la tabla 6.4. Comparar el resultado con el de la tabla 6.5 y con la recta de regresión mínimocuadrática sin la observación 16.

   Realizar también una regresión robusta mínimocuadrática recortada.

Regresión logística

A study screened a random sample of adult subjects for adultonset, type 2 diabetes with the objective of identifying some of the risk factors associated with the disease. The file `ex28_49.dat` contains data[1] about diabetes status, weight (in pounds), waist circumference (in inches), and cholesterol ratio (ratio of total cholesterol to HDL cholesterol in blood) for all 386 subjects. Fifty-nine of the 386 were identified as having type 2 diabetes (diabetes = 1).

1. Adult-onset, or type 2, diabetes has been associated with obesity. The output in the figure below shows the results of a logistic regression analysis with response variable `diabetes` and explanatory variable `weight`. Give the equation of the logistic regression model. Is the coefficient for the slope significantly different from zero? Does a higher weight appear to increase the odds of having type 2 diabetes?



```
Logistic Regression Table

                                                   Odds    95%     CI
Predictor       Coef    SE Coef      Z      P     Ratio  Lower  Upper
Constant    -3.56905   0.644311  -5.54  0.000
weight     0.0101218  0.0033247   3.04  0.002      1.01   1.00   1.02
```

2. Use software to obtain the logistic regression analysis with response variable `diabetes` and explanatory variable `weight`. Make sure that you obtain the same results as those shown in the previous exercise, up to rounding error.

   a) Research suggests that where body fat is stored is an important factor in predicting diabetes. Use software to run a model with both `weight` and `waist` as explanatory variables. What is the equation of this new logistic regression model? Do both variables contribute significantly to the model? Explain how adding the variable `waist` could change the relative contribution of the variable `weight`.

   b) Use software to fit a logistic regression model to predict `diabetes` from `waist`. Interpret the output in the context of the study. Which of the three models described in this exercise would you favor and why?

3. Among the list of potential risk factors for type 2 diabetes is a person's cholesterol ratio (`cholratio`), the ratio of total cholesterol to good, or HDL, cholesterol. Use software to fit a logistic regression model to predict `diabetes` from both `waist` and `cholratio` as explanatory variables. What is the equation of this new logistic regression model? Are both coefficients significant? Explain in your own words what the two corresponding odds ratios mean in the context of this study.

[1] Estos ejercicios pertenecen al capítulo suplementario sobre regresión múltiple y logística del libro *The Practice of Statistics for Life Sciences* de Baldi y Moore.