Linear Models, Non-assessed Practical

Week 3, MT 2020

In addition to the exercises below, there is plenty of other material to practise on – all of the R examples from lectures.

The data set we are considering today describes a cloud seeding experiment aimed at increasing rainfalls, taken from Cook and Weisberg's "Residuals and Inference in Regression" book. It used silver iodide as a catalyst to induce rain, and targeted an area of 3000 square miles north-east of Coral Gable, California for 24 days in the summer of 1975. The following variables were recorded:

- Action (`A`): a classification indicating *seeding* (coded `1`) or *no seeding* (coded `0`).
- Time (`T`): days after the beginning of the experiment.
- Suitability (`SNe`): if `SNe` $\geqslant 1.5$ the day was judged suitable for seeding based on natural conditions.
- Echo coverage (`C`): per cent cloud cover in the area, measured using radar.
- Pre-wetness (`P`): total rainfall in the target area.
- Echo motion (`E`): a classification indicating a moving radar echo (coded `1`) or a stationary radar echo (coded `2`).
- Response (`Y`): amount of rain (in $10^7 m^3$) that fell in the area for a 6-hours period on each suitable day.

Several `R` functions will be suggested and used for the analysis, please use the help (`?fun` or `help("fun")`) to make yourself familiar with them as needed.

Exercise 1: Importing and Exploring the Data

1. Load the data from the file `cloud.seeding.txt`.

   The file is on Canvas, as well as at `http://www.stats.ox.ac.uk/~laws/SB1/data/cloud.seeding.txt`

2. Print the first few lines of the data and explore variable types.

3. Which variables appear to be related to the response variable, and thus may be good choices for an explanatory variable in a linear model? [Use `cor()`.]

4. Perform a graphical inspection of the relationship between the response `Y` and the other variables. Does any variable show a definite trend?

5. Transform `A` and `E` into factors with `as.factor()`. Is `Y` distributed differently for the level of each of these variables?

Exercise 2: Model Estimation

1. Fit a simple linear regression using `Y` as the response variable and `T`; save the model in an object called `mT`; and extract regression coefficients, residuals and fitted values.

2. Describe the main quantities present in the output of `summary(mT)`.

3. Is there any evidence that the rainfalls are increasing with time? Use the regression coefficient for `T` to assess whether there is any significant relationship between `Y` and `T`.

4. Now perform a simple linear regression using first `C`, and then `P`, and save them respectively as `mC` and `mP`. Are the respective regression coefficients significant?

5. Try a few transformations of `C`, such as log(`C`) and `C`^{2}, and then do the same for of `P`; does the model fit the data any better? Does it make sense to compare models after transforming the explanatory variable? [Consider R^2 values.]

6. Now transform `Y` into log(`Y`) and fit a simple linear regression using `C` as the explanatory variable. Does it make sense to compare (using R^2, or the residual standard error) how this model fits compared to previous models?

7. Fit a multiple linear regression with `Y` as the response and `T`, `C` and `P` as explanatory variables, and save it into an object called `mCPT`. Are the regression coefficients the same as in the simple linear regressions fitted above? Why?

8. Include the `A` variable into the previous model, coded as a factor. Describe how it is coded as a contrast. Does it appear to be significant?

9. Fit a model which also includes interaction terms between `A` and the other variables, and describe the resulting set of regression coefficients. [Use `summary()`.]

Exercise 3: Model Validation

1. Consider again the model in the `mCPT` object, and call `par()` and `plot()` to plot all the diagnostic plots generated by `plot(mCPT)` in a single figure.

2. Look at the first and second plots: is there any reason to think that the `cloud` data violate the assumptions of the model?

3. Describe the concepts of leverage and influence. Now look at the last plots, locate observations that look problematic and comment on them.

4. Observations 1, 2 and 15 are labelled as possible outliers. Decide which of these to omit and fit the `mCPT` model again (i.e. without some/all of 1, 2, 15) and call it `mCPT2`; does this new model fit the remaining data better than before?