Analysis of Continuous Data

Homework 4

24/11/2015

This homework consists of a data analysis exercise in SAS, and a small theoretical exercise. The main purpose of the data analysis exercise is to help you understand first how linear regression can bring insight in a practical question. At a more technical level it is designed to have you reflect on the different interpretations of the parameters in a multiple regression set up and on the impact of violations of model assumptions on the model output. Finally, you will be expected to demonstrate the use of partial residual plots in regression diagnostics and the Forward stepwise selection method in the model building process.

For the data analysis exercise, it is again important that you write your results in the requested report format with your report limited to 4 pages. Note that this homework should be made individually, even if some consultation along the road with staff and students is allowed, your answers should be personal and written in your own words. This homework is due on the minerva dropbox on **01 December before Midnight**. It should be sent through that route to Els Goetghebeur and Emmanuel Abatih. The marks that you will receive for this homework contribute to your final score for Analysis of Continuous Data.

1 A data analysis exercise in SAS

The dataset caschool.xlsx contains a random sample of California elementary school districts. The data consists of test scores (Y: testscr), class sizes (

X_1: stratio) and the percentage of non-native English speakers among the students in each district (X_2:nonep).

- Y: (testscr) districtwide average score of reading and math scores on the Stanford achievement test
- X_1: (stratio) the total number of students in the district divided by the number of teachers.
- X_2: (nonep) the percentage of non native English speakers among the students in each district

The test score is a districtwide average of reading and math scores on the Stanford achievement test, a test utilized by school districts in the USA. The student-teacher ratio, i.e. the total number of students in the district divided by the number of teachers, is used as a measure of the (overall) class size in the district. Policy makers are interested in the question of whether reducing class size, for instance by hiring more teachers, improves the student's education. Skeptics worry that reducing class size will increase costs without producing substantial benefits.[1]

1. Explore the variables in the model to get a better understanding of their distribution and of the bivariate relationships. Briefly describe your findings.
2. Fit a multiple linear regression model with stratio and nonep as explanatory variables (no interaction). Discuss briefly your results (e.g. interpretation of parameters, quality of the model fit, outliers, assess model assumptions (Linearity, constancy of error variance, normality of residuals etc)...). Are any remedial measures necessary? if so what will be the impact of the remedial measure on the overall fit of the model.
3. Fit a simple linear regression model with only stratio as explanatory variable. Compare the estimates of the regression coefficients and quality of the model fit between this model and the model in the previous

[1] You can find more information on the research conducted on this subject at http://www.ed.gov/pubs/ReducingClass/index.html

step (2). What is the difference in interpretation of the regression coefficient of stratio ? Explain the issue of confounding in this particular context.

4. Fit again a multiple linear regression model with stratio and nonep, but now also include their interaction effect.
   - What is the interpretation of the effect of X_1 and X_2 on Y?
   - Graphically present the results of the interaction effect in a way that key features can be easily revealed to non-statisticians
5. Use the F-test to perform a Forward stepwise selection procedure to decide if nonep (X_2) should be added to a model that already has (X_1). Are the results similar to those obtained when using the Akaike Information Criterion(AIC) for model selection?
6. Suppose a school district has a student-teacher ratio of 20 and 50% of non-native English speakers. What is the predicted score for such a schooldistrict. Report the outcome with an appropriate prediction interval (with a coverage of 95%).
7. How might your analysis inform any decision on whether or not to reduce the student -teacher ratio ? Can you think of a different study that would be better for this purpose? Explain briefly.
8. **Bonus Question**:
   - How would you use Cross-Validation for model selection?

2 Theoretical Part: Some Small Questions

Are following statements right or wrong ? Explain.

1. Based on a regression analysis with one regressor, the null hypothesis $H_0 : \beta_1 = 0$ is rejected in favor of the alternative H1 : $\beta_1 > 0$ with a very small p-value (p < 0.0001). The tted model will give thus very good predictions.

2. In a study the subject’s age is known to be a confounder. The statistician includes age as a covariate in the regression model, but it turns out that the regression coefficient of age is not significant at the 5% level of significance (say, p = 0.69). He/she can therefore remove age from the model.

Here is yet another theoretical exercise.

1. In a 2008 paper by Cameriere et al., the authors wrote that ’This model had the lowest Akaike Information Criterion (AIC) value among the considered multiple regression models. Indeed Eq. (1) explained 93% of the total variance ($R^2 = 0.93$).’ In which way are the AIC and the R^2 connected? Does a high AIC imply a high R^2 or vice versa? Explain.