Linear Models: Regression Analysis

Exam Project

KU Leuven
Academic year 2014-2015

Stefan Van Aelst, Kris Peremans

The project consists of two parts. First, analyzing multiple datasets by using the tools you have learned throughout the course and second, writing a report with your conclusions. The report should describe your analyses (including results and interpretations) of different datasets available on Toledo. For each dataset some hints or questions are given to guide you through the analysis procedures. Reporting the results is not sufficient, you should clearly describe and interpret the results. The analysis should be performed in **R**. Write the code with structure! Attach your **R** code as an appendix to the report.

The written report with appendix should be handed in on or before **15 January 2015** as one single pdf file. Please upload the written report before the deadline on Toledo. It is not necessary to sent me your report by e-mail or make a hard copy of it. If you have any questions, you can mail me at kris.peremans@wis.kuleuven.be or you can come by at my office (mathematics department 200B, room 02.30). Good luck!

Exercise 1: senic data

The primary objective of the Study on the Efficacy of Nosocomial Infection Control (senic project) was to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection in United States hospitals. This dataset consists of a random sample of 113 hospitals selected from the original 338 hospitals surveyed. We consider 4 variables: average length of stay of the patients (Y), the average age of patients (X_1), the average estimated probability of acquiring infection (X_2) and the number of available facilities and services (X_3).

(a) Explore the senic dataset and model a linear relationship. Give the regression estimates and the standard errors. Interpret the output. Calculate 99% confidence intervals of the regression parameters.

(b) Test with 10% significance level whether there is a linear relation between the response and the predictor variables. At the 99% level, can X_1 and X_3 be dropped from the model together?

(c) Examine the validity of the fitted regression model and corresponding inference. Graphically explore whether interactions would be useful to add to this model.

(d) Build a regression model that can predict well the length of stay of future patients based on these three predictor variables.

(e) Investigate whether there are any isolated outliers that have a large influence on the regression analysis.

(f) Explain (in words) the meaning of the breakdown value. Compute the reweighted least trimmed squares estimator with 25% breakdown value. Identify the four different types of observations based on a diagnostic plot.

Exercise 2: baseball data

Consider the set of Major League Baseball players who played at least one game in both the 1991 and 1992 seasons, excluding pitchers. This dataset contains the salaries for these players, along with performance measures for each player from 1991. Four categorical variables indicate how free each player was to move to other teams. The dataset is split into a training and validation set by random sub-sampling. The last variable indicates whether the observation belongs to the training set or not.

Variable number	Description
1	Salary (in thousands of dollars)
2	Batting average
3	On-base percentage (OBP)
4	Number of runs
5	Number of hits
6	Number of doubles
7	Number of triples
8	Number of home runs
9	Number of runs batted in (RBI)
10	Number of walks
11	Number of strike-outs
12	Number of stolen bases
13	Number of errors
14	Indicator of "free agency eligibility"
15	Indicator of "free agent in 1991/2"
16	Indicator of "arbitration eligibility"
17	Indicator of "arbitration in 1991/2"
18	Player's name (in quotation marks)
19	Training observation (1) or not (0)

(a) Model the relationship between salary and all other variables. If necessary use transformations, but keep all variables in the model. Interpret the results.

(b) Compute the standardized and studentized residuals and detect possible outlying observations. Remove the outliers from the dataset.

(c) Allow for interactions between number of runs and the four dummy variables and build stable models that reflect well the relation between salary and the most relevant predictors in the dataset. Which unique models do you obtain? Validate these models.

(d) Employ the model with the best prediction abilities to predict the salary of a new baseball player (use the entire outlier-free dataset). The performance measures of the new baseball player are available on Toledo. Compute 90% prediction intervals too.

Exercise 3: air pollution data

Consider the air pollution data providing properties of 60 Standard Metropolitan Statistical Areas in the United States. The data include information on the social and economic conditions in these areas, on their climate and some indices of air pollution potentials. The goal is to study the effect of air pollution on mortality.

Variable name	Description
prec	Average annual precipitation in inches
jant	Average January temperature in degrees F
jult	Average July temperature in degrees F
ovr95	Percentage of 1960 SMSA population aged 65 or older
popn	Average household size
educ	Median school years completed by those over 22
hous	Percentage of housing units which are sound and with all facilities
dens	Population per square mile in urbanized areas
nonw	Percentage non-white population in urbanized areas
wwdrk	Percentage employed in white collar occupations
poor	Percentage of families with income < 3000 dollars
hc	Relative hydrocarbon pollution potential
nox	Relative nitric oxides pollution potential
so	Relative sulfur dioxide pollution potential
humid	Annual average percentage relative humidity at 1 pm
mort	Total age-adjusted mortality rate per 100 000

(a) Fit a linear model. Consider transformations of predictors if they improve the model fit. Check the model assumptions and check for multicollinearity problems.

(b) Perform principal component regression. Select an optimal model. Consider using a validation set (take the last 10 observations) and leave-one-out cross-validation. Do different selection techniques lead to the same model? Hint: for cross-validation you can use the option `validation="LOO"` in the **R** command `pcr` from the package `pls`.

(c) Perform ridge regression. Based on the validation set of the previous question, calculate the RMSEP.

(d) Which solution do you prefer, principal component regression or ridge regression, and explain why?

Exercise 4: military lottery data

We consider data with draws performed in 1970 by the U.S. government to determine which men will be called first for military duty in Vietnam. If the procedure (drawings from an urn) would be completely at random, then there should not exist any dependence between day of birth and day of calling. There is suspicion that the drawing was not done correctly. Men that were born in the last months of the year seem to have a higher chance of being called earlier.

(a) Investigate this suspicion by fitting a non-parametric curve to the data.

(b) Is local quadratic or linear regression preferred?

(c) Construct confidence bands for the best non-parametric fit.

(d) What is your conclusion about the randomness of the lottery?