Department of Statistics

COURSE STATS 330/762

Model Answers Assignment 3, 2011

Your task in this assignment was to write me a report of not more than four pages (excluding appendices). The report needed to provide an answer to the question "Is the proportion of new FAIR plan policies in the zip code district related to the racial composition"? To answer the question, you should fit a regression model to the data, taking whatever remedial steps seem necessary.

When you are satisfied with your model, you should have interpret the coefficients in the appropriate way.

Your report should consist of the following parts: (1) an executive summary, (2) a main part (introduction, main section, conclusions) and a technical appendix, containing the details of the statistical analysis, including the R code. Don't make the first two parts too technical. You may find the presentation "Report writing presentation" on the course web site helpful.

The assignment was worth 40 marks, split up as follows:

Report: good structure 10 marks, clear arguments 10 marks

Analysis: Thorough analysis, fixing up any deficiencies in the model: 10 marks. Correct conclusions drawn from the analysis: 10 marks.

What I was looking for in the report:

- An executive summary which clearly summarises your conclusion
- A main part, consisting of (i) an introduction backgrounding the problem, and the methods you are going to use to reach your conclusion, (ii) A section describing the data and the analysis you have done in reasonably non-technical terms, with references to the appendix for technical details.
- A section describing your conclusions.
- An appendix in which you can put the details of your analysis.

You got 10 marks if the report was structures as described above, and 10 marks if your conclusion was clearly expressed in the executive summary and the conclusions section.

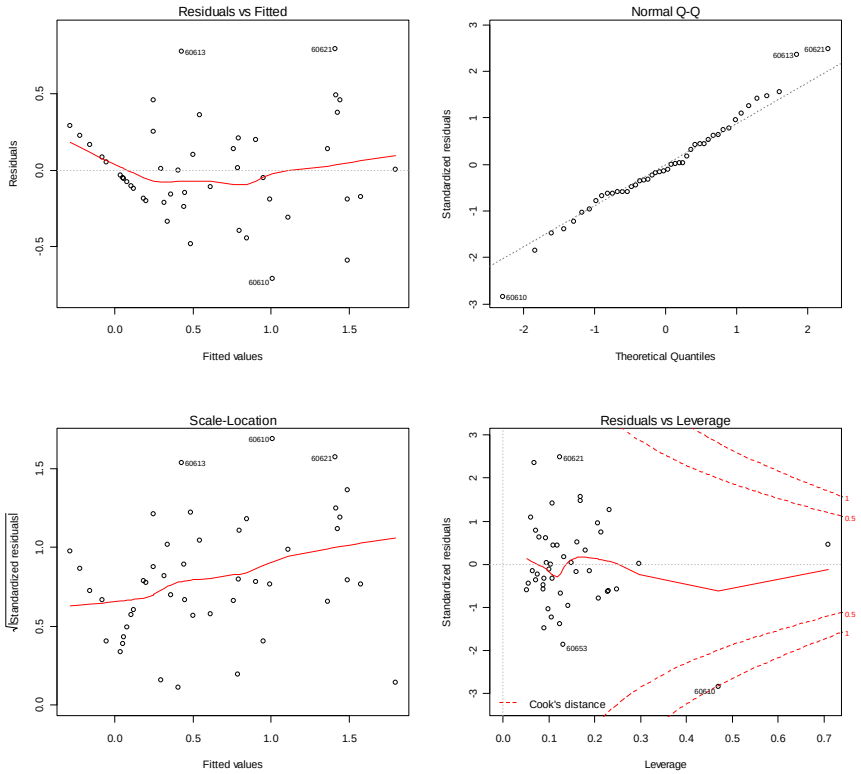
The other 20 marks came from the analysis: 10 marks for finding a reasonable model, and 10 marks for reaching a conclusion supported by the analysis.

Some comments on the analysis are given below.

A pairs plot of the data shows one grossly outlying point, which is the value for “theft” on zip code 60607. I suggest deleting this point before proceeding. The pairs plot also reveals a curved relationship between fire and involact.

The basic question to be answered is “is race and insurance related, after the other variables have been taken into account”. It is legitimate to decline insurance on the basis of risk, bit it is not legitimate to decline on te basis of race, given the risks are equal. This suggests fitting a model with involact as response, race as the explanatory variable of interest, and the risk variables as confounders (variables that offer an alternative expalanation of the responses behavior.) The basic strategy is to interpret the coefficient of race, with the risk held constant (i.e. the usual interpretation of the coefficient).

Fitting the model using all the variables but without point 60607 without this point gives the following diagnostic plots



From these we can conclude that

- There is a slight tendency for the errors to increase with the mean;
- One point (Zip 60610) has a high Cooks distance, due to its large residual;
- The rather strange band of points in the residuals/fitted value plot is due to there being several zero responses.

At this point we might do an "all possible regressions" to eliminate some of the risk variables. This indicates that using fire alone may be reasonable. Gam plots indicate (as did the pairs plot) that a quadratic in fire might be required. A Box-Cox plot suggests a 1/3 power or square root of the response, so a reasonable final model is

sqrt(involact) ~ race + poly(fire,2)

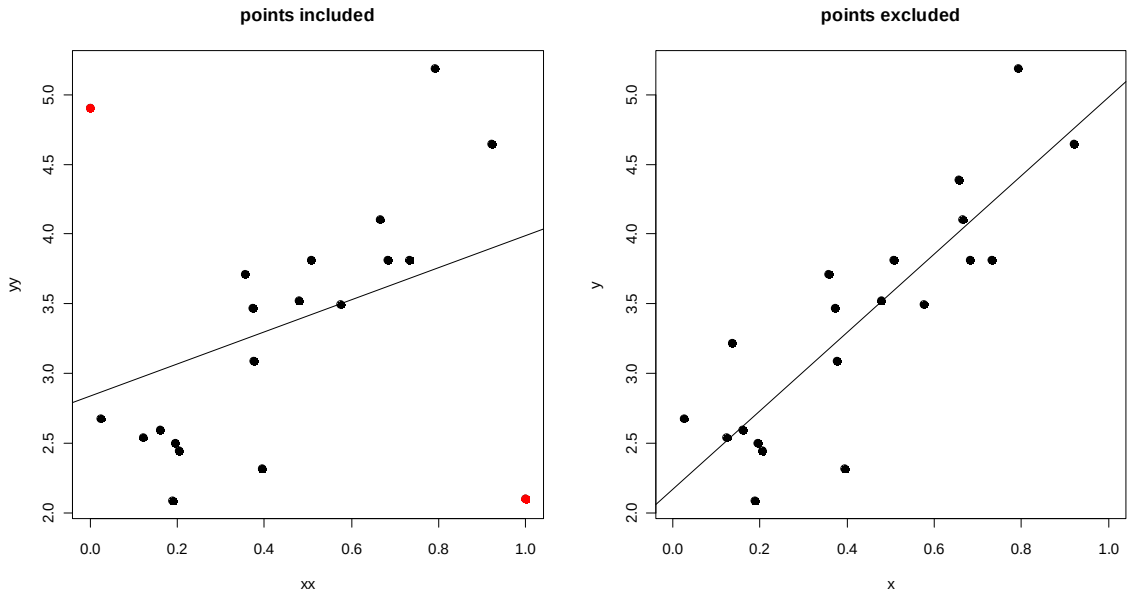
(the square root gives a better R2 than the 1/3 power or a log) Fitting this indicates that there may be problems with outliers/influential points, with points 6, 7, 23 (zips 60610, 60611, 60612) being problematic.

When we start deleting points, we see that for some combinations the p-values for race and the R^2 vary considerably.

I experimented with deleting various combinations of 6, 7 and 23 (in addition to 24), with the following results

Points removed	P-value for Race	R^2
24	0.0120	0.7985
6, 24	0.0113	0.8281
6,7,24	0.0456	0.8576
6, 23,24	0.0196	0.8195
7,24	0.0474	0.8279
7, 23, 24	0.1730	0.8403
23,24	0.0398	0.8021
6,7,23,24	0.1001	0.8528

From this we see that removing 7 and 23 together makes the coefficient of race non-significant. The combined effect of these two points is to shift the regression surface by a considerable amount. The picture overleaf illustrates this in a simplified form: the combined effect of removing two is greater than just moving one at a time, as the diagnostics do.



Thus, the significant p-values seem to be driven by one or two points and therefore shouldn't be trusted.

There is another issue here: the interpretation of p-values after model selection has been done. However, broadly the same results hold true if the other risk variables are included in the model.

Conclusion:
In view of the ambiguities in the analysis, the conclusion should be that insufficient evidence exists to be sure that the insurance companies are guilty of racially motivated decisions.

This is a difficult analysis with some puzzling features. If you fitted a reasonable model, and interpreted the p-value for race correctly, then you got most of the 20 marks for this part of the assignment. I didn't penalize students for not fully teasing out the "remove two points" issue.

There is a discussion of this data set in the book by Julian Faraway (Faraway, J.J. (2005). *Linear Models with R.* Chapman and Hall, Boca Raton.) I have put a scan of the relevant chapter on the web site.