ME104 Linear Regression Analysis: Problem Set 9

Models for ordinal data and models for count data.

1. Models for ordinal data.
   Cricket is a bat-and-ball game between two teams of 11 players. Test cricket is a form of the game which involves matches lasting up to five days, played between national teams. A test series is a set of 1-6 consecutive test matches played in one country between the same two teams. Each match, and thus also each series, ends in a win for one of the two teams, or a draw when neither team wins. Each series is played in the country of one of the two teams, so we can talk about the home team and the visiting team of the series. The data file is `cricket.dta` The data set shows information on every test series of at least 2 matches played between 9 of the test-playing nations in the period between 1959 and 2011. The dataset includes the following variables:

`series`	ID number of the series.
`year`	the year when the series was played.
`home and visitor`	the home team and the visiting team.
`matches`	number of matches in the series.
`winner`	winner of the series (Draw if the series was drawn).
`result`	result of the series (1=Win for the visiting team; 2=Draw; 3=Win for the home team). This will be treated as the response variable Y in the analyses.
`hrating and vrating`	the ratings of the home and visiting teams before the series began, based on the results of each team in the preceding 3-4 years. Higher ratings correspond to more successful teams. These ratings will be treated as continuous, interval-level variables.
`drating`	the difference hrating-vrating.
`period`	period in which the series was played, approximately a decade. Dummy variables for different periods are also included.

   Suppose we want to consider two simple research questions: (i) how well do the team ratings predict results of test series, and (ii) what is the extent of home advantage in test match cricket, and whether it varies over time. We will also examine whether the result of a test series can reasonably be treated as an ordinal variable in such analyses.

   (a) Fit an ordinal regression model for `result` given `drating` as the only explanatory variable using the `ologit` command. What is the interpretation of the estimated coefficient of drating?

   (b) Create two plots of fitted values:

      i. The two cumulative probabilities of the visitor winning (i.e. $Y \leq 1$), and of a draw or visitor winning ($Y \leq 2$).

      ii. The probabilities of the individual categories of Y, i.e. the three values of `result`.

In broad terms, how would you summarise the information in the graphs?

(c) There is home advantage in test cricket if the probabilities of the results are more favourable for a team if it plays at home than if it plays away against the same opponent. We can examine this with the plot produced in b(ii). If there is no home advantage, the plot should be symmetric around 0. What do you conclude from the plot, is there evidence of home advantage?

(d) We next examine whether the extent of home advantage has varied over different periods. This would be the case if period had a significant association with the result even after controlling for the strengths of the teams.

i. Fit a model which includes both period (as a categorical explanatory variable) and drating. Use this model to calculate fitted probabilities of the different results in each of the periods for a series where `drating`=0. When does the home advantage appear to have been the strongest?

ii. Carry out a likelihood ratio test to compare this model to the model in a, which only includes drating. Report the test statistic and the P-value. Is the partial effect of period statistically significant?

(e) Even though the response variable is ordinal, we could ignore its ordering and model it using a multinomial logistic model. We can then compare this to the corresponding ordinal logistic model, to examine whether the simpler ordinal model appears to be adequate. Are the probabilities similar or different? (Hint: Compare the two models creating a plot which includes the fitted probabilities from both.)

(f) Whether an ordinal model is appropriate depends not only on the response variable but also on the explanatory variables. To illustrate this, consider a model with both drating and matches as explanatory variables. As in question e, fit both an ordinal and a multinomial model, and create a plot of the fitted probabilities where matches varies from 2 to 6, and drating is fixed at 0. What do you conclude from the comparison this time? Where are the largest differences in the probabilities?

2. Models for count data.
We use now the `dollhill3` (the dataset is available for download within Stata using the command `use http://www.stata-press.com/data/r11/dollhill3, clear`. The dataset contains data associating coronary heart disease deaths and smoking originally reported in Doll and Hill (1966).
The age categories are defined as follows:

Agecat	age/years
1	35-44
2	45-54
3	55-64
4	65-74
5	75-84

(a) Fit a Poisson regression for `deaths` given `smokes`, `i.agecat` using `pyears` as exposure. You can obtain incidence-rate ratios running again the same model followed by the `irr` option.

(b) What is your interpretation of the estimated regression coefficients?

(c) To understand the model better, you can use the `margins` command. Use the `margins` command to calculate the predicted counts at each level of `agecat`, holding all other variables in the model at their means. Briefly interpret.

(d) Find the predicted number of events with the commands `predict` and then plot them. Briefly interpret.

(e) To help assess the fit of the model, the `estat gof` command can be used to obtain the goodness-of-fit chi-squared test.

(f) You can use the `vce(robust)` option to obtain robust standard errors for the parameter estimates to control for mild violation of the distribution assumption that the variance equals the mean.

(g) The assumption that the conditional variance is equal to the conditional mean can be checked using several tests including the likelihood ratio test of over-dispersion parameter alpha by running the same regression model using negative binomial distribution `nbreg`. Would you conclude that the negative binomial model is more appropriate than the Poisson model?