ME104 Linear Regression Analysis: Problem Set 4

Multiple linear regression

1. The first data file used today is FEAROFCRIME. It contains data for 32,824 respondents in the 2001/02 British Crime survey.

   (a) Fit a linear model for fear of crime given age and sex using the command `regress`. How do you interpret the coefficient of female?

   (b) We also want to investigate possible interactions and nonlinear effects of age and sex. To do this, create two new variables: (i) the product of age and female (agexfemale), and (ii) the square of age (age2) using the command `generate`.

   (c) Fit a model which includes the main effects of age and sex and their interaction, by adding agexfemale to the model. Is the interaction term statistically significant?

   (d) Give an interpretation of the estimated regression coefficients of age in the model with the interaction, separately for men and women.

   (e) Summarise the model by plotting the fitted values of fear of crime as a function of age, separately for men and women (Hints: Here we calculate two sets of fitted values using the command `predict`, and plot them in the same graph.).

   (f) In broad terms, how would you describe the effect of sex and age on fear of personal crime?

2. The second data file used today is PARKING. The data concern diplomats from 146 countries stationed at the United Nations in New York City.

`violations`	the number of parking tickets which were issued to diplomatic vehicles from a country and which were not paid (annual average number for the period 11/199711/2002).
`corruption`	a measure of the level of corruption in a country for 1998, with higher levels indicating higher levels of corruption. This reflects both social norms and level of legal enforcement, but only the contribution of the norms may be transferred to a diplomats environment in New York.
`logdipl`	logarithm of the number of diplomats at the countrys UN mission.
`loggdp`	the logarithm of the countrys GDP per capita in 1998 (in year-2000 US dollars).

   (a) Create scatterplots of violations against the three explanatory variables, using the variable countrycode as labels for the points. Also create a histogram of violations. What do you observe?

   (b) Fit a linear model for violations given the three explanatory variables, and create a plot of studentised residuals against standardised fitted values (include country codes as labels for the points). Examine the residual plot.

   (c) Create a new variable called logviol, which is the logarithm of 1+violations.

(d) Fit a model for logviol given the same explanatory variables as in b, again creating also the residual plot. Does the plot now suggest that the problem of heteroscedasticity has been resolved? What do you conclude about the size and significance of the estimated partial effects of the explanatory variables?

3. Additional work

(a) Interpret the estimated regression coefficients of the model fitted in Exercise 2d.

(b) The residual plot obtained in 2d suggested that one country (which one?) might be a mild outlier. To check whether it had a noticeable effect on the results, refit the model without that country. Are there any major changes in the model?