ME104 Linear Regression Analysis: Problem Set 4

Multiple linear regression II

Today we use again the dataset parking.dta. 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.
`dipl`	number of diplomats at the countrys UN mission.
`gdp`	GDP per capita in 1998 (in year-2000 US dollars).

1. Revision
   (a) Create a scatterplot of violations against gdp. What do you observe?
   (b) Create a new variable called logviol, which is the logarithm of 1+violations. In Stata, the formula for this is ln(1+violations). Create a scatterplot of logviol against loggdp. What do you observe?
2. Fit a linear model for logviol given corruption dipl and loggdp. Briefly interpret.
3. Check homoscedasticity of residuals using the command `rvfplot, yline(0)`. What are the pattern of the data points?
4. Use a Cook-Weisberg test using the command `estat hettest` to test the residuals for heteroskedasticity. An insignificant result indicates lack of heteroskedasticity. Interpret the test result.
5. Calculate and plot studentised residuals against standardised fitted values (include country codes as labels for the points) for the model in c). Examine the residual plot. Can you identify any outliers?
6. Calculate leverage points to identify observations that will have potential great influence on regression coefficient estimates using the command `predict lev, leverage`. Generally, a point with leverage greater than (2k+2)/n should be carefully examined (where k is the number of predictors and n is the number of observations). Can you identify any leverage points?
7. Use the `lvr2plot` command to check potential influential observations and outliers at the same time. Briefly explain.

8. Calculate Cook's distance against the observation numbers using the command `predict newvarname if obsyear==1998, cooksd .` List the countries which show a cook value beyond the cut-off point 4/n.

9. Perform a model specification test using the command `ovtest`. Interpret the result of the test.

Additional work

1. Drop the problematic cases and re-run the model in c). What do you conclude?

2. Use the commands `kdensity`, `qnorm` and `pnorm` to check the normality of the residuals. What do you conclude?

3. Use a Smirnov-Kolmogorov test using the command `sktest resid` to test the residuals for normality. Interpret.