ME104 Linear Regression Analysis: Problem Set 2

The Classical Linear Regression Model

For this lab we will use the Gallup et al. (1999) Geography and Economic Development data set (`geodata.dta`). We will investigate the impact of several variables on the countries' GDP per capita. You are strongly encouraged to write your code for this exercise into a `.do` file (e.g. `lab2.do`) using the Stata Do-File editor.

`gdp95`	Adjusted GDP per capita in 1995 ($).
`urbpop95`	% of population living in urban areas, 1995, from World Bank (1998).
`ztropics`	The proportion of the country's land area within the geographical tropics.
`open6590`	The proportion of years that a country is open to trade during 1965-90, by the criteria in Sachs and Warner (1995b).
`airdist`	Distance (*km*) to the closest major port.
`lifex65`	Life expectancy (*years*).
`zwater`	Water availability.
`dens65i`	Inland population density (measured as population per km^2).

Lab work

1. Open the `geodata.dta` dataset and perform the following regression using the `regress yvar xvar` command (or the menu system – `Statistics --> Linear Models and Related --> Linear Regression` if you prefer):

$$gdp95_i = \beta_0 + \beta_1 urbpop95_i + \varepsilon_i$$

   (a) Report OLS estimators of β_0 and β_1 for this example.

   (b) Identify the t-test statistics, its P-value and the 95% confidence interval for β_1 and the coefficient of determination R^2.

   (c) Provide a substantive interpretation of the results, which has to be consistent with the above results, but which should be interpretable for anyone who does not understand regression analysis.

   (d) Create a scatterplot for x and y adding a regression line using the `twoway (scatter yvar xvar) (lfit yvar xvar)`. Discuss the results.

2. Regress `gdp90` on `open6590`, `airdist` and `ztropics`.

   (a) What percentage of variation in the response variable is explained by the three predictors?

   (b) Which coefficients are statistically significant and what does this mean?

   (c) Using the stata command `predict [type]` *newvarname, [option]*; compute the fitted values (no *option* needed) and the residuals of the fitted regression model (*option*=`resid`).

(d) Now try plotting residuals versus fitted using `rvfplot` and compare these results to the previous graph.

(e) Plot the residuals versus fitted values using the command `rvpplot` *varname*, for the variables `open6590` and then again for `ztropics`.

(f) Based on just visual inspection of these plots, evaluate the extent to which the OLS assumptions hold for this model. Briefly discuss these results.

(Optional) Additional work

1. Regress life expectancy on water availability and inland population density. Briefly discuss the results.
2. Create a new dependent variable taking the log of the life expectancy variable. What is the theoretical and empirical range of the new variable compared to the range of the old?
3. Re-run the model from above with the new dependent variable. How does the coefficient of the water availability change? How do the fit statistics from a regression using the new variable compare to the fit statistics from the previous model?