Roadmap to Data Analysis

For a structured start to your empirical project and a good general orientation, see (among other) the ungraduated textbooks:

Wooldridge: *Introductory Econometrics*
Greene: *Econometric Analysis*

Important topics are found

- Writing an Empirical Paper
  - Wooldridge, Chapter 19
- Panel Data:
  - Wooldridge, Chapters 13 and 14; Greene, Chapter 13
- Instrumental Variable estimations:
  - Wooldridge, Chapter 15
- Discrete Choice:
  - Wooldridge, Chapter 15; Greene, Chapter 21

The following list of points provides a rough practical roadmap to empirical project and addresses recurring concerns, questions and mistakes.

1. Preparatory Data Work

a. Acquisition

   i. Download date from an online source or other.

   ii. Document where you downloaded the data and when.

b. Handling

   i. Store data in the form and format as downloaded in a separate folder (e.g., called *Raw Data* and possible sub-folders).
   Data in this folder will *never* be changed, manipulated or overwritten.

   ii. Insert the data in a statistical software (Stata or other), possible saved in a convenient format.

   Stata command: import, insheet, save

   iii. Use the software to combine data from different files and sources.
   Never combine data by hand.

   Stata commands: merge, joinby, append

   iv. All output from the econometric analysis (see below) goes in a separate folder.
   Produce tables and figures in an automated way, *never* by hand.

c. Review and description

   i. Produce a Table summarizing
   Exact variable definition, Source, Units, Coverage

   ii. Produce a table of the summary statistics including information on
   Units (again), Mean, Median, Variance, Min, Max

   Stata command: outreg2 using…

   iii. Produce a table with unconditional correlation of your central variables.

   Stata command: corr, pwcorr, estout

iv. Plot the unconditional correlation of your central variables (e.g, unemployment and growth rate).
    Do so for all central pairs of variables you will use in the analysis (e.g., dependent and independent variable; instrumented variable and instrument)
    Stata command: scatter, twoway,…
v. Describe observed correlation (strength, positive or negative, outliers).

2. Data Analysis

a. Description of econometric model.
   Address the following question
   i. What do variables stand for? What type of variables are they (e.g., dummy, categorical, continuous)? What are the measure units (e.g. Euros, Millions of Euros, metric tons, percentage points)?
   ii. What are their units of observation (e.g., year, quarter, month, country…)?
   iii. What do indices stand for (time, countries, firms…)?
   iv. What is the assumption on the error term?
b. Motivation of model choice (beyond reference to previous literature).
   Address the following questions:
   i. Why is this model suitable to address the research question?
   ii. Why estimating in levels or changes?
   iii. What are possible omitted variables?
   iv. What economic mechanisms may generate reverse causality?
   v. Why estimating with random effects, fixed effects, time effects (in the case of panel data)?
   vi. What determines and motivates the length of time periods?
   vii. What are your assumptions on the instrument (in the case of IV estimations)?
   viii. What are your assumptions a diff-in-diff approach (exogeneity of treatment, SUTVA)?
   …if applicable:
   i. What is the purpose of interaction effects, structural breaks (if applicable)?
   ii. What type of exogeneous variation in the independent variable may allow for a causal interpretation of the coefficient (if applicable)?
c. Regression
   i. OLS
      Stata command: reg
   ii. Panel data (with random effects, fixed effects)
      Stata command: xtreg (,fe)
   iii. Instrument Variable
      Stata command: ivregress, ivregress gmm
d. Regression results.
   Generate tables reporting the regression results, containing (at least) the following information:
   Stata command: outreg2, outreg, estout
   i. Dependent variable (with units)
   ii. Specification (different Columns, e.g., sample and sub-sample, split of periods)
   iii. Independent variable, Number of Obs., adj. R-squared (Rows)
   iv. F-Test score if applicable. (Mandatory in case of IV estimations.)
e. Possible robustness checks

i. Drop, censor outliers.
ii. Chose sub-samples and sub-periods.
iii. Alternative definition of key variables.
iv. Modifying model (e.g., estimation in changes instead of levels, fixed effects instead of random effects).

3. Interpretation

a. Narrow econometric interpretation.
   Address the following questions:
   i. What is the change of the dependent variable associated with a change in the independent variable?
   ii. Is the effect statistically significant? In all, some specifications?
   iii. What may be the reasons if coefficients or significance change from one specification to the other?

b. Economic interpretation.
   Address the following questions:
   i. How strong are your estimated results? Are they realistic?
   ii. What are potential limitation of our approach (data limitations)?
   iii. Are they likely to generalize to other setups, i.e., countries, industries etc. (external validity)?
   iv. How does your result compare to existing empirical literature?
   v. To what extent are your result in line with existing theory (theories)?

Note: Some commands suggested above require the installation in Stata the first time they are used. For this, it is necessary to install the command using *ssc install*

e.g., ssc install outreg2