Steps for analysing data with regression

Data Analysis 2

Fall, 2020

This short step-by-step description tries to give you an overview on how to carry out data analysis for different subjects. These are only guidelines, you have to understand your data quality issues, your research question and modify some of this steps in order to carry out a proper analysis.

A good data analysis does not depend on the impact of its results, but rather on how it was carried out. No-connection/insignificant results are also as important as significant results with large magnitude OR large uncertainty in your prediction might be as valuable as a quite certain prediction. Be honest and report the findings that you can get out of the data. Good luck and try to enjoy as much as it is possible!

Steps

1. Define what is your (broad) research question! What you want to learn from the data or what you want to use it?
   (a) What is the main aim: *prediction* or *causality*?
2. Data collection - getting your data
   (a) What are the possible problems with the data - *data quality*
      i. Representative sample?
      ii. Measurement errors in the variables?
      iii. What they truly capture and what you want to use them for? (e.g. management quality v.s. scale from 1-5 for a specific question)
   (b) Choose your y and your x variables from the potential set of variables.
3. Data descriptives
   (a) Check the summary statistics and distributions of your variables.
   (b) Consciously retain or drop observations. Maybe impute if really necessary.
      i. Take care of missing values and extreme observations.
   (c) Select your (first) estimation sample and change your research question accordingly.
   (d) You may want to scale your variables for easier handling and interpretation.
      i. Change your interpretation as well, e.g. GDP in thousands, or number of cases for 10,000 people, ect.
4. Check the pattern of association between y and each (key) x variables
   (a) Use scatter plot with lo(w)ess or bin-scatter
      i. Have an idea how the two variable is related
   (b) Decide on possible non-linear transformation (if association is non-linear)
      i. Log transformation(s) - only if not contradicts your research question and non-negative values.
      ii. Quadratic or other polinomial

iii. Piecewise linear splines, and check for potential knot places
iv. Think about taking ratios

(c) Weighted regression: if this estimation method helps your interpretation (e.g. not comparing countries, but people living in countries), you can use it. This is also a good robustness check, that you can put into the appendix along with the others.

5. Compare explanatory variables (x-s) - if you do multiple regression

(a) What is the best measure for a certain intended variable?
i. Check how x-s are related to each other, especially if they measure the same thing.
ii. Check collinearity, if present: a) retain one variable and drop the others (you may play along which one to retain), b) create a score variable, by averaging or z-score or create principle components.

(b) Think about possible interactions:
i. Do you have quantitative variables? If yes, a) Do you expect them to have different levels? $\rightarrow$ use simple dummy variables. b) Do they have different slope values? $\rightarrow$ interact with specific variable. If both, then use both of them.
ii. Is there possible interaction effect between two continuous variables? You may interact them. Check how the results changes!
iii. If your research question is aiming to a hypothesis which can be captured by interactions, use them and test it.

6. Variable selection/model choice

(a) IF your goal is prediction:
i. Keep whatever works (meaningful, and you have data on them for possible prediction)
ii. Focus on functional form of explanatory variables.
iii. Try to keep the model as simple as possible and avoid over-fitting your data!
iv. Analysis of residuals - check your lowest and highest residuals. What can you infer from the characteristics?
v. Visualization of fit - $y - \hat{y}$ plot to show your fit.
vi. Check the goodness of fit measures like BIC and compare models based on that.
vii. Talk about uncertainties in your prediction and plot them: confidence interval of predicted values and prediction interval.

(b) IF causality:
i. Think about potential confounders (or omitted variables). If you do not have data on them, how would you expect your coefficient to change? Argue!
ii. Think about measurement errors in the explanatory variables and if there is a potential threat, explain and argue.
iii. Think about the mechanism of cause and effect. Drop bad controls.
iv. Think about functional forms of important confounders.
v. Compare different setups (models w increasing number of variables) and show parameter stability.
vi. Talk about generalization (population/general pattern and external validity)
vii. Test hypothesis: is your parameter of interest significantly different from the value you are interested? (E.g. wage gap supposed to be 0.)

7. Probability models and time-series analysis

   (a) Probability models:

      i. Use LPM, logit and probit models and compare the results
      ii. If very different (marginal) coefficients, think about the functional form of your explanatory variables.
      iii. Analyse predicted probabilities: lowest and highest percents' characteristics, how the probabilities are distributed unconditionally and conditionally
      iv. Check model fit measures: brier score and pseudo-R2 (maybe R2)
      v. Check model diagnostics: bias of prediction and calibration curve.
      vi. May think about categorization and show a confusion table for your chosen model and threshold.

   (b) Time-series models:

      i. Do the additional cleaning and data wrangling: dummy for imputed missing/extreme values and adjusting frequencies.
      ii. Check stationarity: use differences or dummies for seasonality. Use graphs and unit-root tests.
      iii. Make sure SE are proper (Newey-West or including lags of y)
      iv. Model the serial correlation: think about propagation effect in both x and y. If meaningful for x, estimate the cumulative effect and test it.
      v. If your goal is prediction, make a fan-chart.

8. Robustness checks

   (a) You need to check how robust your results. General advices:

      i. If you have many observations (+500), you may use a training (4/5 of sample) and a test (1/5) sample, where you re-run your regression (causality) or predictions (prediction) and check your results. The samples must be a random pool from your original sample and you should not use your test sample for choosing your model. Only for evaluation.
      ii. External validity: if you have in addition such variable for y - other time/space/group - where you can test your results and conclude whether it is valid as well. If no such variable then make a short argument about what would you expect: how is your results change and why.
      iii. If you have competing models or alternative parametrization (e.g. piecewise linear spline vs. using interactions of certain x), you can show how your model performance is (un)changed. Also you may include other xs, that you would not consider in your main model.