Regression Analysis

Harry Ganzeboom

Research Skills #4

November 12-20 2008

Intro

- This lecture is a comment on materials provided by Berry & Feldman.
- I like these materials because they are relatively non-technical / non-mathematical.
- But the materials are not superficial!
- Occasionally, I have IMPORTANT comments.

Simple and multiple regression

- A regression equation describes the functional relationship between a Y-variable (dependent) and one or more X-variables using an linear, additive equation.
- An equation with one X is a simple regression; if there is two of more X's, it is a multiple (not multivariate!) regression.
- Multivariate regression: when there is more than 1 Y-variable. We do not discuss this case, but will later look at a related model, factor analysis.

Simple regression - DS

- Descriptive statistics you need to be familiar with:
  - Intercept
  - Slope
  - Residual
  - Sums of squares: total = model + residual
  - R-squared and R
  - Standard Error of the Estimate = Root Mean Squared Residual
  - Unstandardized and standardized coefficients.

Simple regression - IS

- Inferential statistics you need to be familiar with:
  - Standard Error = Sampling Error
  - Sampling distribution ("steekproe<u>ven</u>verdeling")
  - T-value and associated probability
  - Significance testing
  - Mean squares, F-value and associated probability.

Multiple regression - basics

- Intercept: expected Y when all X's are 0.
- Always try to interpret the intercept, even when it is useless (out of the range of the data).
- Partial slopes: effect of X on Y when all the other X's are held constant ("controlled").
- Two ways to see what this means:
  - Conditional means: averaged simple regressions
  - Regression between residuals.

Collinearity

- The X's in MR can be uncorrelated (no collinearity), but usually – in observational (non-experimental) studies – they are not.
- If the X's are uncorrelated, MR is a bit pointless.
- The materials may leave the impression that collinearity is a bad thing and needs to be avoided: this is a wrong impression.
- Even if collinearity is strong, it is something you will have to deal with, in stead of avoid.
- Multi-collinearity: the degree to which X's depend upon one another in a case of three or more X's. You cannot directly judge this from the correlation matrix.

Regression and causality

- There is a close relationship between the regression model and causal (cause-effect) analysis; however, it is not identical.
- Regression models are only about partial associations.
- In order to give it a causal interpretation, one needs theory, causal order assumptions and a research design that fits these assumptions.
- I will say more about this in the future – the materials in Berry & Feldman are not very informative on this.

OLS – estimation of the coefficients

- OLS is the standard way to find the best fitting equation.
- Extended OLS: WLS and GLS.
- Major alternative: maximum likelihood estimation. Gives the same results in most cases.
- Take note about what LS methods do:
  - Give relatively much weight to large residuals (outliers)
- OLS produces the SE formulas in Berry & Feldman, p. 13. These formulas are very useful to understand the effects of (multi)collinearity.

SE's

- Note the ingredients of SE's and their role:
  - N: as N becomes larger, SE is smaller, with function 1/sqrt(N).
  - VAR(X): if VAR(X) becomes smaller, SE gets larger.
  - VAR(res): if VAR(res) becomes smaller, SE gets smaller.
  - Multicollinearity: if R2(XX) goes up, SE goes up.

Qualities of estimators

- Unbiasedness: on average at the population value.
- Efficient: estimation procedure has minimum variation (smallest possible SE's).
- Theoretically, there is a possibility that biased estimators are more efficient (better!) than unbiased estimators.
- BLUE: Best Linear Unbiased Estimator. OLS is BLUE in the regression case.

Goodness of fit

- $R2 = SS(reg) / SS(total)$
- Varies between 0 and 1.
- There are no ‘good’ or ‘satisfactory’ R2’s: it is just what it is.
- R2 is sensitive to measurement error in Y.
- Adj R2: $= (SS(reg) – k*MS(res)) / SS(total)$
  - Adjust for the number of variables used.
  - Can go down with more predictors and can be negative!
  - Are useful to judge improvement between models at first glance.

T- and F-test

- T-test: b / SE > 2?
  - One-tailed and two-tailed tests
- F-test: MS(reg)/MS(res) > critical F?
  - F-table is complicated (two degrees of freedom).
  - Magic number: F(1,many) = 3.84 = 1.96*1.96.
- The overall F-test is rather useless, because it only tells you that ‘something’ is going on.
- F-tests (íncremental F-tests) are useful when judging improvement between different models (formula 1.22).

Measurement assumptions

- Interval measurement of X and Y
  - Includes dichotomous measurement of X and dummy variables for nominal X.
  - Whether you can apply the model to ordinal measures is a matter of interpretation.
- No measurement error in the X-variables
- Measurement error in Y has consequences, but these are not as severe as in X.

Measurement error in X / Y

- Measurement error:
  - Random (unreliability)
  - Systematic (invalidity – you are measuring another variable than you intend to).
- Random measurement error in Y is subsumed in the residuals: lowers R2, but B's stay the same.
- Random measurement error weakens effects of the X-variable (downward bias); how this works out in multiple regression is predictable, can be repaired (if you know the size of the random error), but is still complicated.
- No <u>general</u> statement about the effects of systematic error (and proxy variables) can be made. However, you can often say much about in a specific context.

Specification

- All the predictors of Y are included in the model.
  - IMPORTANT: when you leave out predictors of Y that are not correlated with the other X-variables in the model, there is very little harm.
- No irrelevant X-variables are included in the model.
  - In fact, there is usually little harm here. It leads to some inefficiency. You can see (SE's) how much.

Stepwise modeling

- In practice, analysts look quite a bit at badly specified models: in stepwise modeling, we compare models with different specifications (set of X-variables).
- Forward and backward.
- If theoretically guided (causal order assumption), this is all very instructive.

No <u>perfect</u> multicollinearity

- Two important instances:
  - Dummy representation of X-variables: you have to choose (omit) a reference category.
  - You cannot have more predictors than data-points. In fact, it is advisable to have many more (at least 10x) data-points than predictors.
- All of this is a matter of research design and proper interpretation.

High multi-collinearity

- High multicollinearity can be produced by the data and can be repaired by (more) data.
- If two X-variables are highly correlated you need a lot of data to distinguish their partial effects.
- IMPORTANT: this is something that canNOT be avoided by omitting one of the collinear variables!

Residual of Error?

- Note that regression texts waver between the use of ‘residual’ and ‘error’ for the same thing.
- ‘Error’ (“fout”) suggest that we are wrong – this is appropriate for measurement error or sampling error.
- Residual (“rest”) suggest what we do not know about the other determinants.
- However, Berry (8-9) stresses the distinction beween the true model (with residuals) and the estimated model (where residuals also contain error).

The mean residual

- The mean (expectation) of the residuals:
  - 0 for all combination of X-variables
  - 0 over-all
- These assumptions are more a matter of defining residuals than substantive.
- They are important to detect (A) non-linearity, (B) non-additivity.

Variance of the residuals

- VAR(res) is expected to be constant for all combinations of X (homo-skedasticity / hetero-skedasticity).
- Intuitively: in the formula’s for the SE’s a single R2 represents all residual variation adequately.
- WLS en GLS take into account that the expected variance fluctuates by X-combinations. This makes for more complicated estimation procedures and more complicated formulas for SE’s.

Normality of the residuals

- Distribution of the residuals is often assumed to be normal.
- Berry points out that this assumption is only needed in small samples for IS. Even without normality, the estimates are BLUE.
- IMPORTANT: one important exception to normality is the presence of outliers.

Covariances of X and residuals

- Each X is uncorrelated with the residuals.
- If not, it seems more like a specification problem than a statistical problem.
- Note that in OLS the observed residuals are all uncorrelated with the X variables by design.

No autocorrelation (among residuals)

- No autocorrelation: you cannot predict the size of a residuals from another one; residuals are a truly random draw.
- Exceptions:
  - Time-series data, panel-data
  - Network data
  - Geographical data
- Fortunately, autocorrelation can be repaired in GLS-estimation (e.g. time series analysis). In fact, autocorrelation may improve model estimation considerably.