METHODS OF QUANTITATIVE DATA ANALYIS

MSR Course, 2011-2012

Harry B.G. Ganzeboom

Lecture 5: Regression Assumptions and Diagnostics

May 11-14 2012

Regression assumptions

- For an OLS regression equation to be a good model of the data, we have to fulfill assumptions with respect to:
  - Complete specification of X
  - Perfect measurement of X
  - Linearity / additivity
  - Collinearity
  - Residuals
- Regression diagnostics allow us look at the degree to which the assumption are met.
- Note that high explained variance is NOT an assumption!
- Also note that violations of assumptions is not always (very) harmful, in particular (A) because they only have consequences for the SE, rather than for the B’s, (B) much sensitivity to assumptions arises only in low N studies.

Specification

- All the predictors of X and Y have to be included in the model. Violation of this rule is often referred to as “unobserved heterogeneity” or “omitted variables bias”.
  - 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 and the harm is only to the SE.
- 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.

Specification error - repairs

- Specification error cannot be repaired in the analysis, it is a matter of research design:
  - Measure ***all*** the relevant confounders
  - Randomize all the confounders (=experiments)
  - Control all the time-constant confounders by repeated measurement (panel design):
    - Lagged variable (first difference) analysis
    - Fixed effect analysis

Stepwise modeling

- In practice, analysts look quite a bit at badly specified models: in stepwise modeling, we compare models with different specifications (=subsets of X-variables).
- Forward and backward (or even automatic) – not advised.
- If theoretically guided (causal order assumption), this can all be very instructive.
- In publications we often find ‘blockwise’ modeling. It is important to respect causal logic, but it can be done in two ways:
  - First exogenous, than intervening variables
  - First intervening, then confounding variables.
- This depends on how you want to tell the story.

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 should occur 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: it lowers R2 and beta’s (and increases SE), but B’s stay the same.
- Random measurement error biases (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 general statement about the effects of systematic error (and proxy variables) can be made. However, you can often say much about in a specific context.

Measurement error

- Measurement error is an underestimated problem in social science analysis – and even more so in discrete variables analysis and qualitative studies.
- Only calculating cronbach’s alpha is inadequate!!
- Measurement error can be solved if you know the amount of error. This requires designs with repeated measurement.
- But even if you do not have repeated measurement in your design, you can think about and correct consequences of random measurement error on your results.
- I find this an extremely important issue and like linear models (in a SEM context) so much because they offer a solution here.

Collinearity

- The X’s in MR can be uncorrelated (no collinearity), but usually – in observational (=non-experimental) studies – they are correlated. They are ‘collinear’.
- 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.
- If the X’s are uncorrelated (experiments), Multiple Regression is a bit pointless.
- If collinearity is extreme, this implies that we cannot easily separate the individual influences of variables. The SE can become very high, and nothing is significant.
- However, with high collinearity we often see that the effects of X1 and X2 have reversed signs (and both are ‘significant’).
- Methodological texts 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.
- Perfect collinearity arises when one X is totally determined by the other X-vars (e.g. in the case of dummy variables or the APC problem). This is quite different from strong or almost perfect collinearity.

Collinearity: what to do?

- Perfect collinearity needs no repairs, just correct interpretation.
- High collinearity can be countered by more data: this does not repair the collinearity itself, but makes the SE smaller in another way.
- You cannot ‘repair’ collinearity by leaving out one of the variables – this just changes the research question answered.
- Sometimes it helps to center the data (e.g. with polynomial and interaction terms). This trick makes the results a bit more “stable”, although the only thing it does is move the numbers to a more stable area in the regression space.
- You cannot repair collinearity by making small modifications to the data. This is cheating, in particular on yourself.

No <u>perfect</u> multicollinearity

- Three important instances:
  - Dummy representation of X-variables: you have to choose (omit) a reference category.
  - APC or AC: effects of age, cohort and period at the same time: leave out one and refrase your explanation.
  - 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 just need a lot of data to distinguish their partial effects.
- IMPORTANT: this is something that canNOT be avoided by simply omitting one of the collinear variables! This would change the research question.
- However, sometimes changing the question is a good idea: e.g. model the average (joint) effect of father / mothers.

Residuals

- Residuals are the distance between the actual Y from the predicted Y.
- Correct estimation assumes:
  - Residuals are evenly spread around the regression plane (homoskedasticity).
  - Residuals are normally distributed.
  - Residuals are not correlated (=random).
- Violation may affect the estimated SE. There is no harm for the estimated coefficients.

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 of B a single R2 represents all residual variation adequately. In heteroskedastic data this is a simplification.
- 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.

A special case: dichotomous Y

- In a dichotomous variabe, variance = p*(1-p). This number becomes very small when p is close to 0 or 1.
- When the Y is dichotomous, the residuals are typically reduced with extreme X-values. This is a form of hetero-skedasticity.
- Adequate technique here is logistic regression.
- However, the 'linear probability' models often works well in cases that Y has an overall mean between 0.80 and 0.20.
- While the linear probability is for some (journals) a big no-no, I would encourage any used of LR to OLS regression and compare the conclusions.

Normality of the residuals

- Distribution of the residuals is often assumed to be normal. You can check this by plotting the residuals (formal testing is a little hard).
- This assumption is only needed in small samples for correct estimation of SE.
- In small samples (with appropriate sampling design), correct SE's can also be estimated using resampling (jackknife & bootstrap).
- IMPORTANT: one important exception to normality is the presence of outliers.

No autocorrelation (among residuals)

- No autocorrelation: you cannot predict the size of a residuals from another one; residuals are a truly random draw.
- Exceptions may occur in:
  - Time-series data, panel-data
  - Network data
  - Geographical data
  - Multi-level (hierarchical) data
- Fortunately, autocorrelation can be repaired in GLS-estimation (e.g. time series analysis). In fact, taking into account autocorrelation may improve model estimation considerably and more so if autocorrelation is higher. This is more of less the relevance of panel analysis.

Outliers

- An outlier is a point with a high residual that is not fit by the regression model.
- The first point to understand about this is that regression models often do fit extreme data-point – so you do not see outliers.
- Outliers are relatively rare in dependent variables with a limited range – such as attitude scales.
- We should be particularly sensitive to outliers in low N studies with ratio variables (such as comparative studies of countries or organizations).

Leverage

- We speak of leverage (‘hefboomwerking’), when a datapoint is an outlier only in the X-space, i.e. far removed from the other data-point.
- Typically, leverage does not lead to outliers in Y!!
- Leverage can be measured by the distance from the centroid. SPSS measures this distance by:
  - Cook
  - Manahalobi
  - Leverage
- High leverage points are typical candidates for influence / jacknife analysis.

Influence

- Influence denotes that a data-point determines the solution strongly: if we leave it out, the model changes dramatically.
- This may be so because a data point has much leverage OR because of outlie-ing residual.
- SPSS measure influences by Cooks distance and by studentized deleted residualized.

SPSS

- UniAnova offers:
  - Residuals: raw, standardised, studentized, deleted
  - Leverage
  - Cook’s distance
- Regression offers the same plus:
  - Studentized deleted
  - Mahalanobi’s distance
  - Influence statistics: 5x
- All of this works best in small N problems – in which case you can also eyeball.

Is it a problem?

- Outliers, high leverage and influence make the model unstable, but do not make the model incorrect.
- Again, most of the effect is on the SE: these may be incorrectly estimated.
- Again: resampling techniques (bootstrap) may be the solution to obtain correct SE. But in small N studies, it may also be instructive to present models on restricted samples (=do your own jacknife).
- An alternative may be to re-express all variables in P-scores, which removes most extremities. Then re-estimate the model and compare standardized solutions. This is a easy form of ‘robust estimation’.

Resampling techniques

- Resampling techniques are becoming increasingly popular to estimate correct SE.
- The traditional approach is that SE are ‘analytically’ derived (via mathematical formula) assuming simple random sampling (and normality).
- In resampling, you generate an ‘empirical’ sampling distribution by repeating the SRS process a great many times. This also assumes SRS, although more complex sampling design can be accommodated.
- With increased computational power, this is now also do-able to high N studies (but still takes a lot of computer time, try).

Resampling techniques

- Bootstrap. Samples of size N (!!) are drawn from your effective sample with replacement ("met teruglegging") and this is repeated very often.
- Jack-knife: leave out 1, 2, 3 .. observations.
- Both lead to a similar empirical sampling distribution, which may have different SE than the analytical one.
- Jack-knife is reproducible, bootstrap has a random component.
- In SPSS, only bootstrap is available (in UNIANOVA).
- Jack-knifing with 1 observation omitted is actually easy to do and can be quite instructive about influence.
- IMPORTANT: Resampling techniques are NOT a repair for faulty sampling designs. They have the same assumptions about the sampling designs as analytical procedures.

Further reading?

- Berry, William D. (1993) Understanding Regression Assumtions. Sage University Papers #92.
- Allison, Paul D. (1999), Multiple Regression, a Primer. Sage.