Quantitative Analysis and Empirical Methods

Multicollinearity and Heteroscedasticity

Jan Rovny

Sciences Po, Paris, CEE / LIEPP

Overview

- Multicollinearity
- Diagnostics
- Remedies
- Heteroscedasticity
- Diagnostics
- Remedies

What is Multicollinearity

- *Multicollinearity* exists when a predictor is a perfect linear combination of one or more of the remaining predictors. That is to say when a predictor is highly correlated with others.
- High level of correlation between predictors x_1 and x_2 limits our ability to determine the proper relationship between x_1 and y while controlling for x_2 and vice versa, because x_1 does not vary independently of x_2.

What does Multicollinearity do?

- If there is perfect multicollinearity, we have an unidentified model, because there is no non-redundant portion of predictor x_1, with respect to predictors x_2 and x_3. Thus OLS estimator cannot be defined.
- With high correlation there is large standard error which leads to rejection of relationships which may be true.
- It is important to remember that our model as a whole is fine, the problem concerns the standard errors of particular predictors. We may thus see a larger R^2, but have no significant predictors.
- DEMONSTRATION

Collinearity Diagnostics

- The easy thing to do is to look at correlations of our predictors.
    - But be careful, correlations only tell you about the pairwise relationships between predictors, not about the all relationships between predictors
- Subsequently, it is better to check the **Variance Inflation Factor** of each predictor k
    - The logic is that we regress each predictor on all other predictors in the model. This produces R_k^2 which are compared with the model R^2. If $R_k^2 > R^2$ there is evidence of multicollinearity
    - $VIF_k = \frac{1}{1-R_k^2}$ Generally, we need to worry if $VIF_k > 10$
    - Even better is to look at Tolerance $= 1/VIF$
    - R: `library(car) [Return] vif(model)`

Collinearity Remedies

- First reaction to multicollinearity is to drop predictors. This might work, but might also mis-specify your model – which is not just bad, but REALLY BAD.
- To overcome the problem of insignificant t-tests on individual predictors, we can do a joint F-test on the block of problematic predictors. That way we can test whether they – together – explain variance on y or not.
- Predictors which are correlated somehow measure a similar thing. We can thus think of them as forming one common dimension. It might make sense to combine these predictors into one and use it in our regression model. To do this, we perform **Principle Component Analysis**
  DEMONSTRATION

What is Heteroscedasticity

- OLS assumes that the variance of the error is constant $V(\epsilon_i) = \sigma^2$
- Heteroscedasticity means ‘non-constant variance’
- Heteroscedasticity is caused by many things
  - Data pooling – DV across different countries can have substantially different variation
  - Different level of determination – better predictions can be obtained for some units than for other (rich have greater variance on spending on luxury products than poor)
  - Different measurement error – when measurement error is not constant, variance fluctuates (more educated have smaller error variance on questionnaires then less educated respondents)
  - Learning processes – respondents are more erratic on first questions etc.
- Heteroscedasticity biases the standard errors of our estimates and therefore precludes proper hypothesis testing

Diagnosing Heteroscedasticity

- The first useful thing to do is to plot the residuals against the fitted values and against all predictors
  - If you see 'fanning' of the errors at a certain side of the values of a predictor, you have evidence of heteroscedasticity
- The second thing is to run a statistical test for heteroscedasticity – the **Breusch-Pagan Test**
  - `library(car) [Return] ncvTest(model)`
  - Here H_o is homoscedasticity (i.e. constant error variance). Ideally, we wish to fail to reject this H_o. We thus want a high p-value.

Remedies for Heteroscedasticity

- **Robust standard errors** do not remove heteroscedasticity, but correct standard errors to make them consistent (increases significance of truly significant parameters).
- This involves a different estimation of the Variance Covariance Matrix of Errors – we leave this with the econometricians.
- See *R Demonstration*