Multicollniarity

Venkat Reddy

Note

- This presentation is just the lecture notes form the corporate training on Regression Analysis
- The best way to treat this is as a high-level summary; the actual session went more in depth and contained other information.
- Most of this material was written as informal notes, not intended for publication
- Please send your questions/comments/corrections to venkat@trenwiseanalytics.com or 21.venkat@gmail.com
- Please check my website for latest version of this document

-Venkat Reddy

Contents

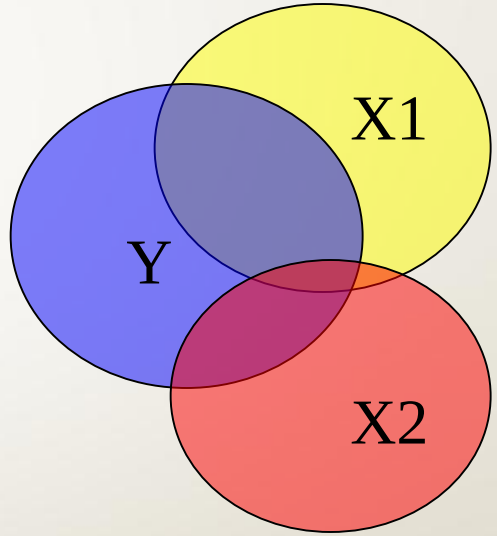
- What is “Multicollinearity”?
- Causes
- Detection
- Effects
- Redemption

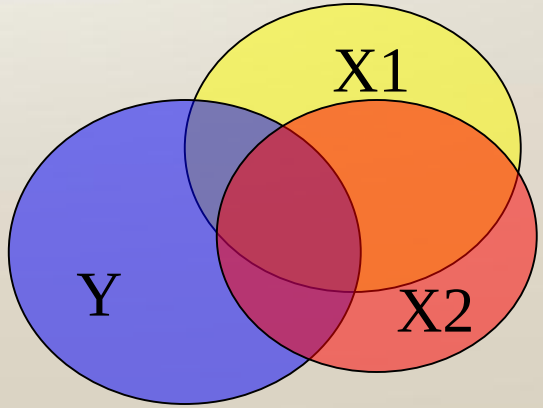
What is “Multicollinearity”?

- Multicollinearity (or inter correlation) exists when at least some of the predictor variables are correlated among themselves
- A “linear” relation between the predictors. Predictors are usually related to some extent, it is a matter of degree.

Multicollinearity-Illustration

- When correlation among X's is low, OLS has lots of information to estimate b. This gives us confidence in our estimates of b
- What is the definition of regression coefficient by the way?
- When correlation among X's is high, OLS has very little information to estimate b. This makes us relatively uncertain about our estimate of b





Perfect Multicollinearity

- Recall to estimate b, the matrix (X'X)-1 had to exist
- What is OLS estimate of b or beta ?
- This meant that the matrix X had to be of full rank
- That is, none of the X's could be a perfect linear function of any combination of the other X's
- If so, then b is undefined- But this is very rare

Causes of Multicollinearity

- Statistical model specification: adding polynomial terms or trend indicators.
- Too many variables in the model – X's measure the same conceptual variable.
- Data collection methods employed.

How to detect Multicollinearity

- A high F statistic or R2 leads us to reject the joint hypothesis that all of the coefficients are zero, but the individual t-statistics are low. (why?)
- **VIF=1/(1-R_k^2)**
- One can compute the condition number. That is, the ratio of the largest to the smallest root of the matrix x'x.
- This may not always be useful as the standard errors of the estimates depend on the ratios of elements of the characteristic vectors to the roots.
- High sample correlation coefficients are sufficient but not necessary for multicollinearity.

Effects of Multicollinearity

- Even in the presence of multicollinearity, OLS is **BLUE** and consistent.
- **Standard errors of the estimates tend to be large.**
- Large standard errors mean large confidence intervals. Large standard errors mean small observed test statistics. The researcher will accept too many null hypotheses. The probability of a type II error is large.
- Estimates of standard errors and parameters tend to be sensitive to changes in the data and the specification of the model.

Multicollniarity Redemption

- **Principal components estimator**: This involves using a weighted average of the regressors, rather than all of the regressors.
- **Ridge regression technique**: This involves putting extra weight on the main diagonal of x'x so that it produces more precise estimates. This is a biased estimator.
- **Drop the troublesome RHS variables**. (This begs the question of specification error)
- **Use additional data sources**. This does not mean more of the same. It means pooling cross section and time series.
- **Transform the data**. For example, inversion or differencing.

Use prior information or restrictions on the coefficients.