THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 50
Detecting Multicollinearity

Chris A. Mack
Adjunct Associate Professor

http://www.lithoguru.com/scientist/statistics/

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Problems with Multicollinearity

- Adding or deleting predictor variables changes the regression coefficients
- The standard errors of the regression coefficients become large
- The individual regression coefficients may not be significant even if the overall model is significant
- Some regression coefficients may be significantly different than expected (even the wrong sign)

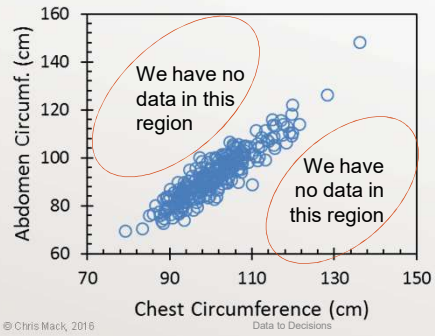
THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Causes of Multicollinearity

- Sampling: we only sampled regions where the predictors are correlated
- The model (or population) demands that certain predictors are correlated
- We are not using the best model
  - Shows up more if the range of predictors is small (e.g., x is correlated with x^2)

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Sampling-induced Multicollinearity



If we simply failed to sample these regions, then the multicollinearity is not inherent in the model or population

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

How to Detect Multicollinearity

- Generate a correlation matrix – simple pairwise coefficients of correlation
  - Won't tell you about more complicated relationships (e.g., if x_1 is highly correlated with x_2 + x_3, but not either variable separately)
- Variance Inflation Factor (VIF)
  - Includes more complicated correlations
- Eigensystem Analysis
  - We'll use this for principle component analysis

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Variance Inflation Factor (VIF)

- How much is the variance of the k[th] model coefficient inflated compared to the case of no correlation?

$$VIF_k = \frac{1}{1-R_k^2} = diagonal\ element\ of\ (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}$$

where R_k^2 is the coefficient of determination when x_k (the k[th] predictor) is regressed against all of the other predictor variables
  - $1/VIF_k$ is sometimes called the tolerance

Variance Inflation Factor (VIF)

- When $R_k^2 = 0$ we have no correlation and $VIF_k = 1$
- When $R_k^2 \to 1$ we have perfect correlation and VIF_k goes to infinity
  - The k[th] regressor adds no new information
- The largest VIF_k is used to indicate the severity of multicollinearity
  - If > 4, we investigate; if > 10, we act
- If the mean value of all the VIF_k is much bigger than 1, we worry as well

Eigensystem Analysis

(see Belsley, Kuh, & Welsch, Regression Diagnostics, Wiley, 1980)

- For a pxp matrix $\boldsymbol{A}$, the eigenvalues are the p-roots of the equation

determinant

$$|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0$$

- Find the eigenvalues of $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$, the correlation matrix
  - If all the eigenvalues are about the same magnitude, we have no multicollinearity
- Calculate the *condition number*, defined as $\kappa = \lambda_{max}/\lambda_{min}$ (some people use $\sqrt{\lambda_{max}/\lambda_{min}}$)
  - If bigger than ~100, we have a problem

Eigensystem Analysis Example

- For a two regressor model,

$$\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$$ Correlation matrix

$$|\boldsymbol{A} - \lambda \boldsymbol{I}| = \begin{vmatrix} 1-\lambda & r_{12} \\ r_{12} & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - r_{12}^2 = 0$$

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}} = \frac{1+|r_{12}|}{1-|r_{12}|}$$

Lecture 50: What have we learned?

- What are the advantages and disadvantages of using the correlation matrix for detecting multicollinearity?
- How do the Variance Inflation Factors address the disadvantage of the correlation matrix?
- How do we use eigenvalues and the condition number to detect multicollinearity?