STAT 302 - Chapter 3 : Multiple Regression - Part 4

Harsha Perera

Key Topics

- Correlated Predictors
- Multicollinearity
- Detecting Multicollinearity
- Variance Inflation Factor (VIF)
- Dealing with Multicollinearity

Correlated Predictors

- In a regression model the response variable is related to predictor variables.
- What if predictor variables itself related to each other.
- This can lead to difficulty in fitting and interpreting the model.
- **Example 3.15 - House Prices in NY** illustrates some behaviour when correlated predictors are in a multiple regression model.

Multicollinearity

- We say that a set of predictors exhibits **Multicollinearity** when one or more of the predictors is strongly correlated with some combination of the other predictors in the set.

- If one predictor has an exact linear relationship with one or more other predictors in a model, the least squares process to estimate the coefficients in the model does not have a unique solution.

Detecting Multicollinearity

- Examine scatterplots of pairs of variables to look for associations among the predictors or look at pairwise correlations between the predictors. **Example 3.16 - House Prices in NY - Correlations**
- When the predictors are related, we should also take care when interpreting the individual coefficients (Read the explanation on first paragraph on page 125).
- Sometimes dependence between predictors can be more subtle. To investigate that we can consider regression models for each individual variable in the model using all of the other variables as predictors.
- Any measure of these models (e.g : R^2) could indicate which predictors might be strongly related to others in the model.
- One calculation is **Variance Inflation Factor (VIF)** which reflects the association between a predictor and all of the other predictors.

Variance Inflation Factor (VIF)

For any predictor X_i

$$VIF_i = \frac{1}{1 - R_i^2}$$

R_i^2: cofficient of multiple determination for a model to predict X_i using the other predictors in the model

Higher R_i^2: higher VIF
Lower R_i^2: VIF will be closer to 1

In general, **if** $VIF > 5$ **we suspect multicollinearity**.

i.e. $R_i^2 > 80\%$

Example 3.17 - Doctors and hospitals in counties : VIF

Dealing with Multicollinearity

What should we do if we detect multicollinearity in a set of predictors ?

Multicollinearity is not a bad thing. It does not necessarily produce poor models. In some occasions correlated predictors are important to the model (interaction terms and polynomials).

Some options for dealing with correlated predictors :

- Drop some predictors
- Combine some predictors
- Discount the individual coefficients and T-tests