1 Overfitting

Overfitting occurs when a statistical model does not adequately describe of the underlying relationship between variables in a regression model. Overfitting generally occurs when the model is excessively complex, such as having too many parameters (i.e. predictor variables) relative to the number of observations. A model which has been overfit will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data.

- **Multi-collinearity**: Multi-collinearity occurs when two or more predictors in the model are correlated and provide redundant information about the response. Examples of pairs of multi-collinear predictors are years of education and income, height and weight of a person, and assessed value and square footage of a house.
- **Consequences of high multicollinearity**: Multi-collinearity leads to decreased reliability and predictive power of statistical models, and hence, very often, confusing and misleading results.

2 Consequences of Multicollinearity

In statistics, the occurrence of several independent variables in a multiple regression model are closely correlated to one another. Multicollinearity can cause strange results when attempting to study how well individual independent variables contribute to an understanding of the dependent variable. In general, multicollinearity can cause wide confidence intervals and strange $p-$values for independent variables.

3 Multicollinearity

- In multiple regression, two or more predictor variables are colinear if they show strong linear relationships. This makes estimation of regression coefficients impossible. It can also produce unexpectedly large estimated standard errors for the coefficients of the X variables involved.
- This is why an exploratory analysis of the data should be first done to see if any collinearity among explanatory variables exists.
- Multicolinearity is suggested by non-significant results in individual tests on the regression coefficients for important explanatory (predictor) variables.
- Multicolinearity may make the determination of the main predictor variable having an effect on the outcome difficult.
- When choosing a predictor variable you should select one that might be correlated with the criterion variable, but that is not strongly correlated with the other predictor variables.
- However, correlations amongst the predictor variables are not unusual. The term multicollinearity is used to describe the situation when a high correlation is detected between two or more predictor variables.
- Such high correlations cause problems when trying to draw inferences about the relative contribution of each predictor variable to the success of the model.

3.1 Types of multicollinearity

There are two types of multicollinearity:

- Structural multicollinearity
- Data-based multicollinearity

Structural multicollinearity is a mathematical artifact caused by creating new predictors from other predictors such as, creating the predictor x2 from the predictor x. Data-based multicollinearity, on the other hand, is a result of a poorly designed experiment, reliance on purely observational data, or the inability to manipulate the system on which the data are collected. In the case of structural multicollinearity, the multicollinearity is induced by what you have done. Data-based multicollinearity is the more troublesome of the two types of multicollinearity. Unfortunately it is the type we encounter most often!

4 How to Identify Multicollinearity

- You can assess multicollinearity by examining two collinearity diagnostic measures: **tolerance** and the **Variance Inflation Factor** (VIF) .
- Tolerance is a measure of collinearity reported by most statistical programs such as SPSS; the variables tolerance is $1 - R^2$.
- All variables involved in the linear relationship will have a small tolerance.
- **Interpretation:** A small tolerance value indicates that the variable under consideration is almost a perfect linear combination of the independent variables already in the equation and that it should not be added to the regression equation.
- **Interpretation:** Some suggest that a tolerance value less than 0.1 should be investigated further. If a low tolerance value is accompanied by large standard errors and nonsignificance, multicollinearity may be an issue.
- The variance inflation factor (VIF) quantifies the severity of multicollinearity in a regression analysis.
- The VIF provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.

4.1 The Variance Inflation Factor (VIF)

- The Variance Inflation Factor (VIF) measures the impact of collinearity among the variables in a regression model.
- The Variance Inflation Factor (VIF) is 1/Tolerance, it is always greater than or equal to 1.

- There is no formal VIF value for determining presence of multicollinearity. Values of VIF that exceed 10 are often regarded as indicating multicollinearity, but in weaker models values above 2.5 may be a cause for concern.
- In many statistics programs, the results are shown both as an individual R^2 value (distinct from the overall R2 of the model) and a Variance Inflation Factor (VIF).
- When those R^2 and VIF values are high for any of the variables in your model, multicollinearity is probably an issue.
- When VIF is high there is high multicollinearity and instability of the b and beta coefficients. It is often difficult to sort this out.

You can also assess multicollinearity in regression in the following ways:

(1) Examine the correlations and associations (nominal variables) between independent variables to detect a high level of association. High bivariate correlations are easy to spot by running correlations among your variables. If high bivariate correlations are present, you can delete one of the two variables. However, this may not always be sufficient.

(2) Regression coefficients will change dramatically according to whether other variables are included or excluded from the model. Play around with this by adding and then removing variables from your regression model.

(3) The standard errors of the regression coefficients will be large if multicollinearity is an issue.

(4) Predictor variables with known, strong relationships to the outcome variable will not achieve statistical significance. In this case, neither may contribute significantly to the model after the other one is included. But together they contribute a lot. If you remove both variables from the model, the fit would be much worse. So the overall model fits the data well, but neither X variable makes a significant contribution when it is added to your model last. When this happens, multicollinearity may be present.

- Variance inflation factor and tolerance tolerance. One is the reciprocal of the other.

4.2 Determing the Variance Inflation Factor (VIF) with R

```
library(car)
# Evaluate Collinearity
vif(fit) # variance inflation factors
sqrt(vif(fit)) > 2 # problem?
```

4.3 Interpreting Variance Inflation Factors

- We learned previously that the standard errors, and hence the variances, of the estimated coefficients are inflated when multicollinearity exists.

- So, the variance inflation factor for the estimated coefficient b_k, denoted VIF_k, is just the factor by which the variance is inflated.
- Variance inflation factors k greater than 4 suggest that the multicollinearity should be investigated.
- Variance inflation factors greater than 10 are taken as an indication that the multicollinearity may be unduly influencing the least squares estimates.

4.4 Tolerance

- Tolerance is simply the reciprocal of VIF, and is computed as

$$\text{Tolerance} = \frac{1}{VIF}$$

- Whereas large values of VIF were unwanted and undesirable, since tolerance is the reciprocal of VIF, larger than not values of tolerance are indicative of a lesser problem with collinearity. In other words, we want large tolerances.
- A tolerance close to 1 means there is little multicollinearity, whereas a value close to 0 suggests that multicollinearity may be a threat.
- The VIF shows us how much the variance of the coefficient estimate is being inflated by multicollinearity. For example, if the VIF for a variable were 9, its standard error would be three times as large as it would be if its VIF was 1. In such a case, the coefficient would have to be 3 times as large to be statistically significant.