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.

Multicollinearity

- Multicollinearity occurs when two or more predictors in the model are correlated and provide redundant information about the response.
- Examples of pairs of multicollinear predictors are years of education and income, height and weight of a person, and assessed value and square footage of a house.

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 pvalues for independent variables.

0.1 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.

0.2 Types of multicollinearity

There are two types of multicollinearity:

1. Structural multicollinearity
2. 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!

0.3 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 estimates standard deviation) of an estimated regression coefficient is increased because of collinearity.

0.4 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 vari- ables 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.

0.5 Determing the Variance Inflation Factor (VIF) with R

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

0.6 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 V IF 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.

Tolerance

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

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

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.

0.7 Consequences of high multicollinearity:

Multicollinearity leads to decreased reliability and predictive power of statistical models, and hence, very often, confusing and misleading results.

Consequences of high multicollinearity

1. Increased standard error of estimates of the regression coefficients (i.e. decreased reliability of fitted model).
2. Often confusing and misleading results.

- Multicollinearity will be dealt with in a future component of this course: Variable Selection Procedures.
- This issue is not a serious one with respect to the usefulness of the overall model, but it does affect any attempt to interpret the meaning of the partial regression coefficients in the model.
- 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.

0.8 Variance Inflation Factor (VIF)

- The Variance In ation Factor (VIF) measures the impact of multicollinearity among the variables in a regression model.
- 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.
- 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.
- In many statistics programs, the results are shown both as an individual R2 value (distinct from the overall R2 of the model) and a Variance Inflation Factor (VIF).
- When those R2 and VIF values are high for any of the variables in your model, multicollinearity is probably an issue.
- **Multi-collinearity**: Multicollinearity occurs when two or more predictors in the model are correlated and provide redundant information about the response. Examples of pairs of multicollinear 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**: Multicollinearity leads to decreased reliability and predictive power of statistical models, and hence, very often, confusing and misleading results.
- Multicollinearity will be dealt with in a future component of this course: Variable Selection Procedures.
- This issue is not a serious one with respect to the usefulness of the overall model, but it does affect any attempt to interpret the meaning of the partial regression coefficients in the model.

Variance Infiation Factor

- The variance inflation factor (VIF) is used to detect whether one predictor has a strong linear association with the remaining predictors (the presence of multicollinearity among the predictors).
- VIF measures how much the variance of an estimated regression coefficient increases if your predictors are correlated (multicollinear). VIF = 1 indicates no relation; VIF ¿ 1, otherwise.
- 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.

- You should consider the options to break up the multicollinearity: collecting additional data, deleting predictors, using different predictors, or an alternative to least square regression.

Multicollinearity

- 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. How- ever, correlations amongst the predictor variables are not unusual.
- The term multi-collinearity is used to describe the situation when a high correlation is detected between two or more pre- dictor 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.

Variance Inflation Factor (VIF)

- In many statistics programs, the results are shown both as an individual R^2 value (distinct from the overall R^2 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, multi-collinearity is probably an issue.

Variance Inflation Factor

- The variance inflation factor (VIF) is used to detect whether one predictor has a strong linear association with the remaining predictors (the presence of multicollinearity among the predictors).
- VIF measures how much the variance of an estimated regression coefficient increases if your predictors are correlated (multicollinear). VIF = 1 indicates no relation; VIF ¿ 1, otherwise.
- The variance in ation factor (VIF) quanti

  es 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.

1 Interpreting VIF Multicollinearity

- A common rule of thumb is that if the VIF is greater than 5 then multicollinearity is high. Also a VIF level of 10 has been proposed as a cut off value.

- The largest VIF among all predictors is often used as an indicator of severe multicollinearity.
- Montgomery and Peck [21] suggest that when VIF is greater than 5-10, then the regression coefficients are poorly estimated.
- A common rule of thumb is that if the VIF is greater than 5 then multicollinearity is high. Also a VIF level of 10 has been proposed as a cut off value.
- The largest VIF among all predictors is often used as an indicator of severe multicollinearity.
- Montgomery and Peck [21] suggest that when VIF is greater than 5-10, then the regression coefficients are poorly estimated.
- You should consider the options to break up the multicollinearity: collecting additional data, deleting predictors, using different predictors, or an alternative to least square regression.

2 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.

2.1 How to Identify Multicollinearity

You can assess multicollinearity by examining **tolerance** and the **Variance Inflation Factor** (VIF) are two collinearity diagnostic factors that can help you identify multicollinearity. Tolerance is a measure of collinearity reported by most statistical programs such as SPSS; the variables tolerance is 1-R2. 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. All variables involved in the linear relationship will have a small tolerance. 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.

2.2 The Variance Inflation Factor (VIF)

In many statistics programs, the results are shown both as an individual R2 value (distinct from the overall R2 of the model) and a Variance Inflation Factor (VIF). When those R2 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.

2.3 Variance Inflation Factor

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.

A common rule of thumb is that if the VIF is greater than 5 then multicollinearity is high. Also a VIF level of 10 has been proposed as a cut off value.

2.4 How to Identify Multicollinearity

You can assess multicollinearity by examining tolerance and the Variance Inflation Factor (VIF) are two collinearity diagnostic factors that can help you identify multicollinearity. Tolerance is a measure of collinearity reported by most statistical programs such as SPSS; the variables tolerance is 1-R2. 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. All variables involved in the linear relationship will have a small tolerance. 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.

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.

2.5 Multi-collinearity

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 multi-collinearity 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.

2.6 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!

2.7 Variance Inflation Factor

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.

2.8 Tolerance

A tolerance close to 1 means there is little multicollinearity, whereas a value close to 0 suggests that multicollinearity may be a threat. The reciprocal of the tolerance is known as the Variance Inflation Factor (VIF). 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.

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

2.9 Variance Inflation Factor (VIF)

The Variance Inflation Factor (VIF) measures the impact of multi-collinearity among the variables in a regression model.

There is no formal VIF valu-e for determining presence of multi-collinearity. 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 R^2 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, multi-collinearity is probably an issue.

2.10 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.