Multicollinearity

Multicollinearity occurs when two or more independent in the model are highly correlated and, as a consequence, provide redundant information about the response when placed together in a model. (Everyday examples of multicollinear independent variables are height and weight of a person, years of education and income, and assessed value and square footage of a home.) 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

- 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.
- 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. Variance In ation 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.
- 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, multi-collinearity is probably an issue.

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.
- 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.

0.1 Tolerance

Tolerance is simply the reciprocal of VIF, and is computed as Tolerance = 1 V IF 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. 10

0.2 Multicollinearity

- **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.

1.10 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. 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)

- The Variance Inflation 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.
- 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 es- timate'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.
- 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.

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.