Sources: <http://blog.minitab.com/blog/understanding-statistics/handling-multicollinearity-in-regression-analysis> (Eston Martz)

<https://onlinecourses.science.psu.edu/stat501/node/347> (Penn State online)

<http://blog.minitab.com/blog/adventures-in-statistics/what-are-the-effects-of-multicollinearity-and-when-can-i-ignore-them> (Jim Frost)

<http://www.personality-project.org/r/book/chapter5.pdf> (Dr. William Revelle)

<http://www.niaoren.info/pdf/Beauty/9.pdf> (Dr. Robyn A. Dawes)

5. 1 In a regression model, what are some of the ways predictors can affect one another?

If no two predictors are correlated, each makes a unique contribution to the response, and the R2 statistic is equal to the sum of the individual r2

In practice, predictors are usually correlated, and, as we include more and more predictors in our model, R2 does not increase as much as it would have, were the predictors uncorrelated.

One type of effect a predictor may have on other predictors is *suppression.* A *suppressor* does not correlate with the response, but because it does correlate with the other predictors, it increases their variance, and results in decreased beta values for the other predictors.

Partial correlation:

Given three variables y, xi , and xj , the relationship between y and xi might change when xj is removed from the model. If we predict xi and y from xj using linear regression, and correlate the residuals, then the correlation of the residuals

xi. = xi – xi\_*hat*

and

y.j = y – yj\_*hat*

is a measure of the strength of the relationship between xi  and y, when the effect of xj  has been removed. This is known as the *partial correlation*, because the effect on xi  and y of all other variables has been *partialled out.*

5.2 How is multicollinearity diagnosed, and when is it a problem?

Eliminating all multicollinearity is not necessary, or productive.

Multicollinearity can make it difficult to determine the precise weight of each predictor. However, it does not affect the overall fit of the model, or the accuracy of predictions made.

Severe multi-collinearity can increase the variance of the model, making predictions susceptible to minor changes in the model.

One of the most common methods for detecting multicollinearity is the Variance Inflation Factor (VIF).

For the following simple linear regression model:

yi=β0+βkxik+ϵi

The variance of *bk*is:

Var(bk)= σ2/∑ni=1(xik−x¯k)2 \* 1/(1−*R2k*)

* where *R2k* is the R2-value obtained by regressing the kth predictor on the remaining predictors.

VIF for the kth predictor is defined as follows:

VIFk = 1 / (1 - R2k)

5.2 What are some techniques to handle multi-collinearity?

*Eliminate it*:

The most common method to handle multi-collinearity is to eliminate it – either manually, or by an algorithmic method.

For instance, if two predictors are known to be highly correlated, one of them can be eliminated from the model manually. Techniques such as stepwise regression are helpful in doing this systematically.

Further, techniques such as PCA (Principal Component Analysis) and PLS (Partial Least Squares regression) are useful in cutting down the number of predictors, and retaining only the most useful (and mutually independent) ones.

*Model it*:

When predictors have multiplicative effects, the regression model can be changed, to include *product terms.*

For instance, if the effect of *x* on *y* depends upon a third variable *z*, such an effect may be examined by using the product terms of *x* and *z*. The preliminary procedure involves *zero-centring* the predictors, so that the product terms are not correlated with the additive terms.

*Linear Regression’s Robustness*:

A long known property of linear regression models is their *robust beauty* (source: Robyn M. Dawes, the University of Oregon, 1979). In effect, this means that even a ‘unit weighted’ linear regression model – i.e. a model in which each of the predictors is, by default, weighted equally – will deliver predictions superior to those obtained by using subjective methods (e.g. clinical intuition) when predicting a numerical criterion from numerical predictors. By extension, even ‘improperly weighted’ regression models – in which the weights of the predictors are chosen by non-optimal methods (e.g. simulation of a clinical judge’s predictions, intuition, or equal weighting) outperform clinical intuition as the basis for decision-making.