ST 517: Data Analytics I

Multicollinearity

Outline

Multicollinearity definition

Variance Inflation Factor

Recommendations

Multicollinearity

Multicollinearity between predictor variables is a situation that arises frequently, and can affect the *precision* of coefficient estimates in multiple linear regression.

Multicollinearity means that at least one of the predictor variables is very close to a linear function of some of the other predictor variables, or in other words there is high correlation between that variable and a linear combination of some other variables.

Intuitively, what this means is that if you know the values of a certain subset of your predictor variables, you could make a pretty good guess at the value(s) of some of the other predictor variables.

Multicollinearity Example

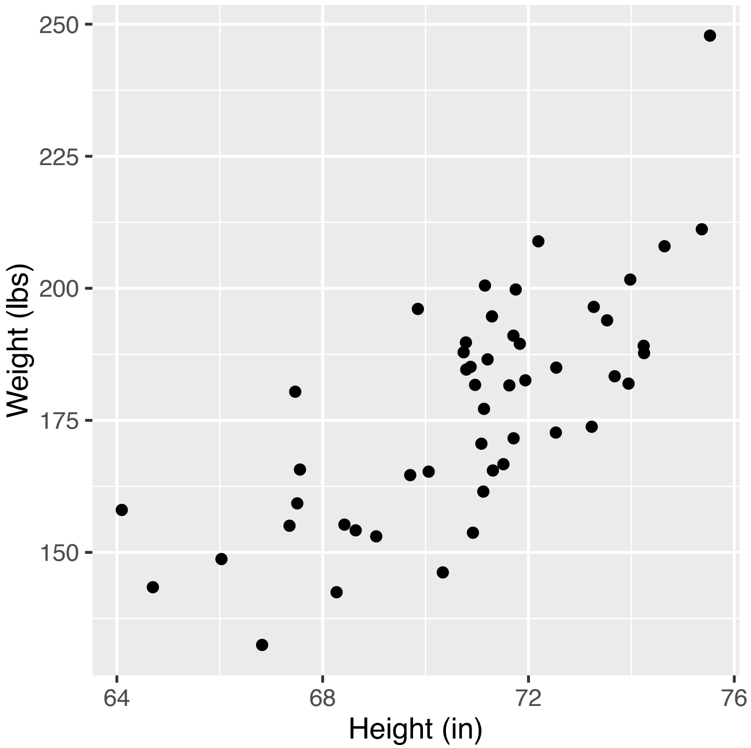
Suppose, as a made-up example, that we are interested in constructing a linear model for professional soccer players' resting heart rate (Y) as a function of predictor variables: player's height (X_1), and player's weight (X_2).

Observe that our predictor variables all tend to be highly correlated with each other:

- Taller people tend to have higher weight

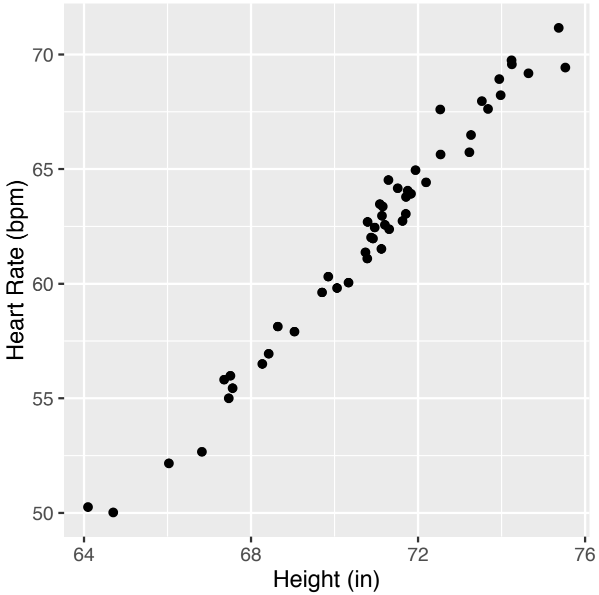
MLS players heights and weights

Here is a scatter plot of height vs. weight for 50 MLS players. Note the high correlation: if I tell you a player's height, you could make a reasonable guess as to their weight, and vice versa.



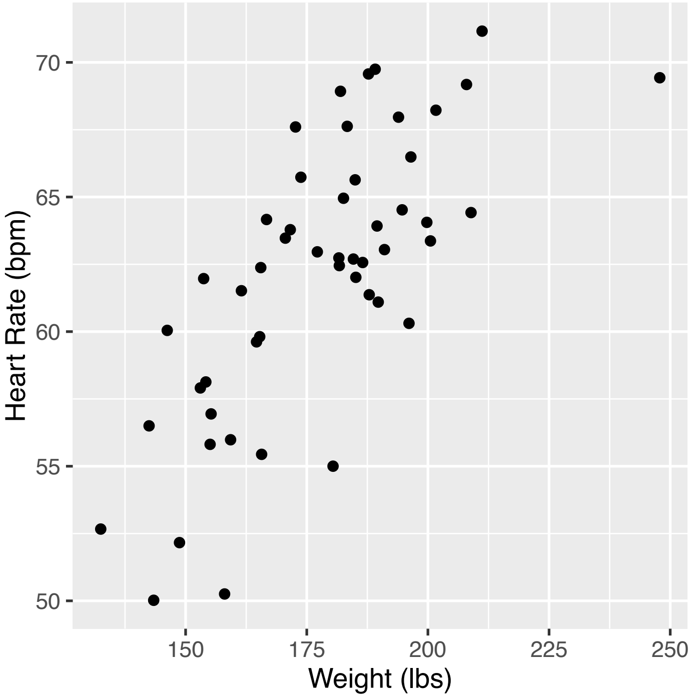
Multicollinearity Example

Now, if we consider the scatter plots of heart rate (HR) vs. height and HR vs. weight, what do you expect to see?



We see a pretty strong linear trend that *taller* players have lower HR. Would you expect that heavier players have lower HR?

Multicollinearity Example



The issue that arises when we have multicollinearity is that it can become difficult to separate the effects of variables that are correlated: Are player’s HRs lower because they’re shorter, or because they weigh less?

Multicollinearity Effects

The effect of multicollinearity in a multiple regression model is to *increase the variance of the coefficient estimates*.

This means that we have less precision in estimating the coefficients when there is multicollinearity, which makes sense: if it is difficult to separate the effects of two (or more) predictor variables, we are less certain about the effect of each individual predictor variable.

We can quantify the effect of collinearity by considering the estimated standard error of our regression coefficients:

$$SE\left(\hat{\beta}_j\right) = \hat{\sigma}\sqrt{\frac{1}{1 - R^2_{(X_j)}}}\sqrt{\frac{1}{\sum_{i=1}^{n}\left(X_{ij} - \bar{X}_j\right)^2}}$$

Here $R^2_{(X_j)}$ denotes the R^2 value you would obtain if you fit a multiple regression model with X_j as the outcome, and the remaining predictors used as predictors for X_j.

Multicollinearity Effects

When X_j can be well-explained as a linear function of the other predictors, $R^2_{(X_j)}$ is ***large***, and this makes the standard error of our coefficient estimate $\hat{\beta}_j$ ***large***.

The quantity

$$VIF = \frac{1}{1 - R^2_{(X_j)}}$$

is called the **variance inflation factor**, and it measures how much the variance of our estimate of β_j is increased because of multicollinearity.

A large $R^2_{(X_j)}$ corresponds to a large VIF, which corresponds to a more variable (less precise) estimate of β_j.

Multicollinearity Effects

It is important to note that multicollinearity in the predictors does not decrease the predictive accuracy of the linear model as a whole: it just makes it more challenging to determine which predictors are most “important”.

Remember the interpretation of β_j: β_j is the additional effect of variable j after accounting for all of the other predictors. When there is multicollinearity involving variable j, it is possible that we might not be able to detect an effect of variable j after we have accounted for all the other variables.

Diagnosing Multicollinearity

- The variance inflation factor can be one useful way to detect multicollinearity issues: some typical rules of thumb are that a VIF > 5 or 10 indicates a multicollinearity problem.
- It is possible for the linear model to have a large R^2 value and a very significant *F*-test result, but none of the *t*-tests for the single coefficient hypotheses $H_0\!: \beta_j = 0$ are significant. This is another good indication that there is some multicollinearity among the predictors.

What to do about Multicollinearity?

If the primary goal of your linear model is prediction (i.e., making a guess as to the value of Y for certain values of X_1, … ,X_p), multicollinearity is not problematic unless it is so severe that the model cannot be fit (i.e. near perfect multicollinearity: one of the predictors is almost perfectly explained by the others).

In this case, you could consider eliminating predictors with very high VIF sequentially (starting from the highest), until the model can be stably fit.

There are several other options you could consider here, including **penalized regression**, **principal component regression**, and **partial least squares regression**.

What to do about Multicollinearity?

If the primary goal of your linear model is inference regarding the effect of specific predictor(s) of interest on the outcome, multicollinearity can be more of a problem (particularly if your predictor(s) of interest have a high VIF).

In this case, you have options including the following:

- Get more data. With a larger sample size, you can get a more precise estimate of β_j even in the presence of multicollinearity.
- Gather data differently: If you can control the values of your predictors, you can force them to not be collinear. This will only work in an experimental setting where at least some of the predictor values can be intentionally set.
- Drop variables that are highly correlated with your predictor of interest...but note that the question you are answering changes if you do this!