Multicollinearity in Regression Analysis: Problems, Detection, and Solutions

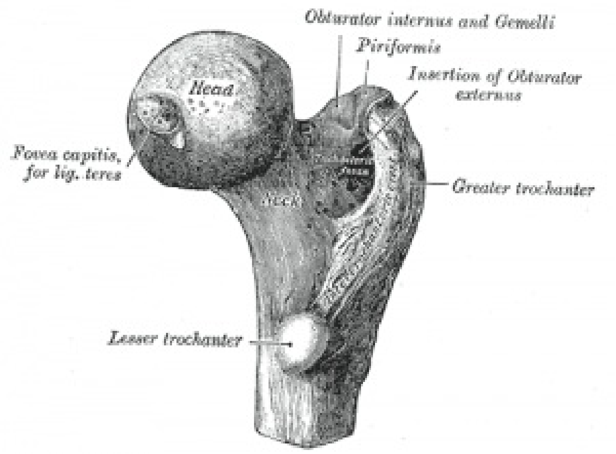
α statisticsbyjim.com/regression/multicollinearity-in-regression-analysis

Jim Frost

April 2, 2017

Multicollinearity occurs when independent variables in a regression model are correlated. This correlation is a problem because independent variables should be *independent*. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.

In this blog post, I'll highlight the problems that multicollinearity can cause, show you how to test your model for it, and highlight some ways to resolve it. In some cases, multicollinearity isn't necessarily a problem, and I'll show you how to make this determination. I'll work through an example dataset which contains multicollinearity to bring it all to life!


Obturator internus and Gemelli
Piriformis
Insertion of Obturator externus
Head
Fovea capitis, for lig. teres
Greater trochanter
Neck
Lesser trochanter

I use regression to model the bone mineral density of the femoral neck in order to, pardon the pun, flesh out the effects of multicollinearity. Image By Henry Vandyke Carter – Henry Gray (1918)

Why is Multicollinearity a Potential Problem?

A key goal of regression analysis is to isolate the relationship between each independent variable and the dependent variable. The interpretation of a regression coefficient is that it represents the mean change in the dependent variable for each 1 unit change in an independent variable when you *hold all of the other independent variables constant*. That last portion is crucial for our discussion about multicollinearity.

The idea is that you can change the value of one independent variable and not the others. However, when independent variables are correlated, it indicates that changes in one variable are associated with shifts in another variable. The stronger the correlation, the more difficult it is to change one variable without changing another. It becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable *independently* because the independent variables tend to change in unison.

There are two basic kinds of multicollinearity:

- **Structural multicollinearity**: This type occurs when we create a model term using other terms. In other words, it's a byproduct of the model that we specify rather than being present in the data itself. For example, if you square term X to model curvature, clearly there is a correlation between X and X^2.
- **Data multicollinearity**: This type of multicollinearity is present in the data itself rather than being an artifact of our model. Observational experiments are more likely to exhibit this kind of multicollinearity.

What Problems Do Multicollinearity Cause?

Multicollinearity causes the following two basic types of problems:

- The coefficient estimates can swing wildly based on which other independent variables are in the model. The coefficients become very sensitive to small changes in the model.
- Multicollinearity reduces the precision of the estimate coefficients, which weakens the statistical power of your regression model. You might not be able to trust the p-values to identify independent variables that are statistically significant.

Imagine you fit a regression model and the coefficient values, and even the signs, change dramatically depending on the specific variables that you include in the model. It's a disconcerting feeling when slightly different models lead to very different conclusions. You don't feel like you know the actual effect of each variable!

Now, throw in the fact that you can't necessarily trust the p-values to select the independent variables to include in the model. This problem makes it difficult both to specify the correct model and to justify the model if many of your p-values are not statistically significant.

As the severity of the multicollinearity increases so do these problematic effects. However, these issues affect only those independent variables that are correlated. You can have a model with severe multicollinearity and yet some variables in the model can be completely unaffected.

The regression example with multicollinearity that I work through later on illustrates these problems in action.

Do I Have to Fix Multicollinearity?

Multicollinearity makes it hard to interpret your coefficients, and it reduces the power of your model to identify independent variables that are statistically significant. These are definitely serious problems. However, the good news is that you don’t always have to find a way to fix multicollinearity.

The need to reduce multicollinearity depends on its severity and your primary goal for your regression model. Keep the following three points in mind:

1. The severity of the problems increases with the degree of the multicollinearity. Therefore, if you have only moderate multicollinearity, you may not need to resolve it.
2. Multicollinearity affects only the specific independent variables that are correlated. Therefore, if multicollinearity is not present for the independent variables that you are particularly interested in, you may not need to resolve it. Suppose your model contains the experimental variables of interest and some control variables. If high multicollinearity exists for the control variables but not the experimental variables, then you can interpret the experimental variables without problems.
3. Multicollinearity affects the coefficients and p-values, but it does not influence the predictions, precision of the predictions, and the goodness-of-fit statistics. If your primary goal is to make predictions, and you don’t need to understand the role of each independent variable, you don’t need to reduce severe multicollinearity.

Over the years, I’ve found that many people are incredulous over the third point, so here’s a reference!

> The fact that some or all predictor variables are correlated among themselves does not, in general, inhibit our ability to obtain a good fit nor does it tend to affect inferences about mean responses or predictions of new observations. —Applied Linear Statistical Models, p289, 4th Edition.

Testing for Multicollinearity with Variance Inflation Factors (VIF)

If you can identify which variables are affected by multicollinearity and the strength of the correlation, you’re well on your way to determining whether you need to fix it. Fortunately, there is a very simple test to assess multicollinearity in your regression model. The variance inflation factor (VIF) identifies correlation between independent variables and the strength of that correlation.

Statistical software calculates a VIF for each independent variable. VIFs start at 1 and have no upper limit. A value of 1 indicates that there is no correlation between this independent variable and any others. VIFs between 1 and 5 suggest that there is a moderate correlation, but it is not severe enough to warrant corrective measures. VIFs greater than 5 represent critical levels of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable.

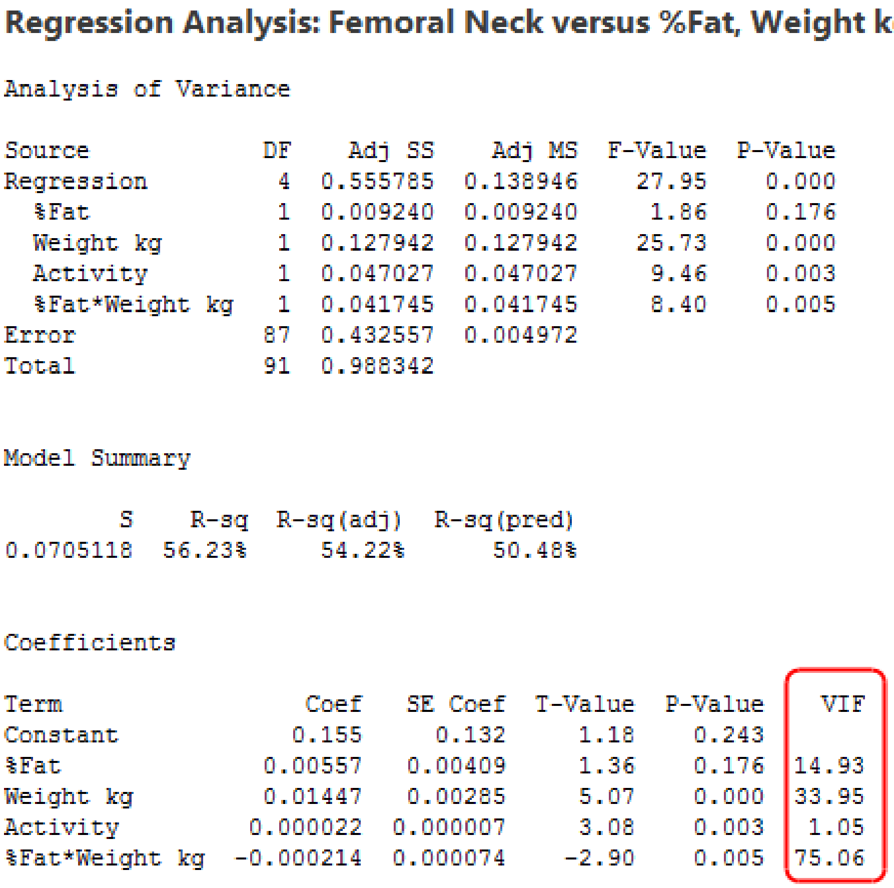
Use VIFs to identify correlations between variables and determine the strength of the relationships. Most statistical software can display VIFs for you. Assessing VIFs is particularly important for observational studies because these studies are more prone to having multicollinearity.

Multicollinearity Example: Predicting Bone Density in the Femur

This regression example uses a subset of variables that I collected for an experiment. In this example, I'll show you how to detect multicollinearity as well as illustrate its effects. I'll also show you how to remove structural multicollinearity. You can download the CSV data file: MulticollinearityExample.

I'll use regression analysis to model the relationship between the independent variables (physical activity, body fat percentage, weight, and the interaction between weight and body fat) and the dependent variable (bone mineral density of the femoral neck).

Here are the regression results:



Regression Analysis: Femoral Neck versus %Fat, Weight kg, Activity

```
Analysis of Variance

Source            DF    Adj SS    Adj MS  F-Value  P-Value
Regression         4  0.555785  0.138946    27.95    0.000
  %Fat             1  0.009240  0.009240     1.86    0.176
  Weight kg        1  0.127942  0.127942    25.73    0.000
  Activity         1  0.047027  0.047027     9.46    0.003
  %Fat*Weight kg   1  0.041745  0.041745     8.40    0.005
Error             87  0.432557  0.004972
Total             91  0.988342


Model Summary

        S    R-sq  R-sq(adj)  R-sq(pred)
0.0705118  56.23%     54.22%      50.48%


Coefficients

Term                 Coef   SE Coef  T-Value  P-Value    VIF
Constant            0.155     0.132     1.18    0.243
%Fat              0.00557   0.00409     1.36    0.176  14.93
Weight kg         0.01447   0.00285     5.07    0.000  33.95
Activity         0.000022  0.000007     3.08    0.003   1.05
%Fat*Weight kg  -0.000214  0.000074    -2.90    0.005  75.06
```

These results show that Weight, Activity, and the interaction between them are statistically significant. The percent body fat is not statistically significant. However, the VIFs indicate that our model has severe multicollinearity for some of the independent variables.

Notice that Activity has a VIF near 1, which shows that multicollinearity does not affect it and we can trust this coefficient and p-value with no further action. However, the coefficients and p-values for the other terms are suspect!

Additionally, at least some of the multicollinearity in our model is the structural type. We've included the interaction term of body fat * weight. Clearly, there is a correlation between the interaction term and both of the main effect terms. The VIFs reflect these relationships.

I have a neat trick to show you. There's a method to remove this type of structural multicollinearity quickly and easily!

Center the Independent Variables to Reduce Structural Multicollinearity

In our model, the interaction term is at least partially responsible for the high VIFs. Both higher-order terms and interaction terms produce multicollinearity because these terms include the main effects. Centering the variables is a simple way to reduce structural multicollinearity.

Centering the variables is also known as standardizing the variables by subtracting the mean. This process involves calculating the mean for each continuous independent variable and then subtracting the mean from all observed values of that variable. Then, use these centered variables in your model. Most statistical software provides the feature of fitting your model using standardized variables.

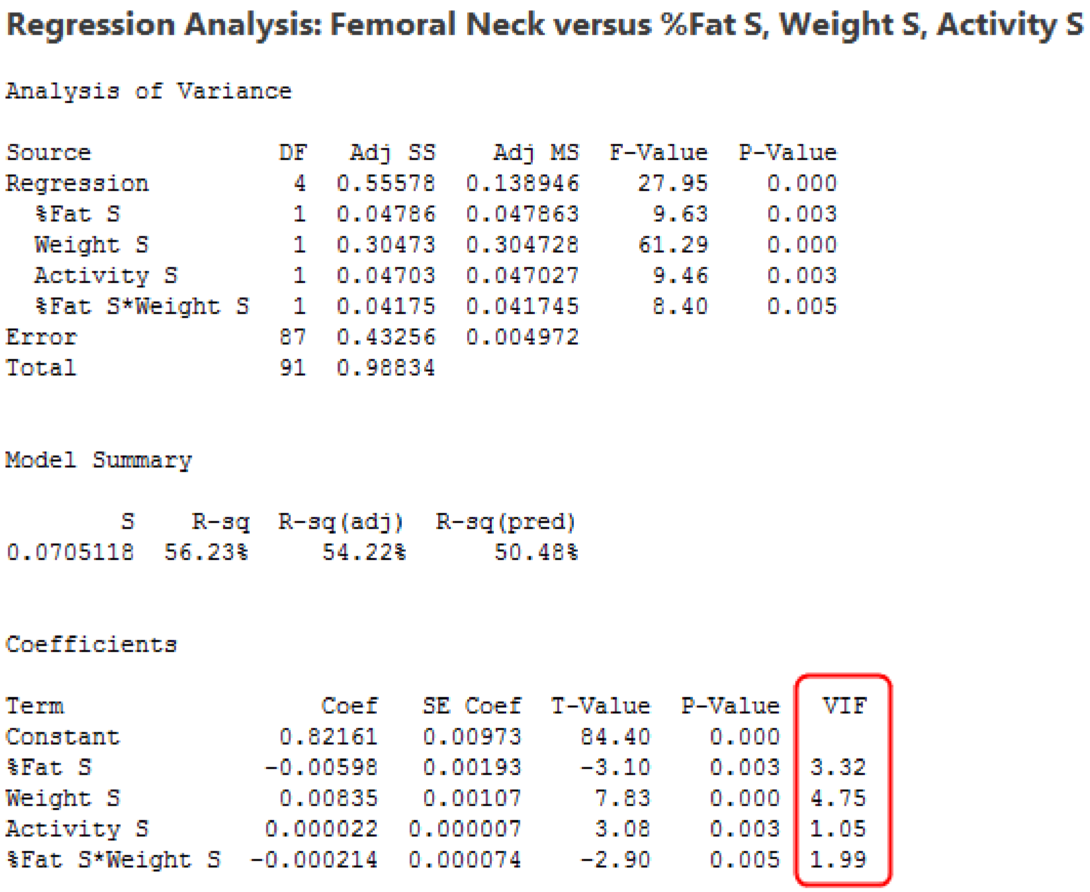
There are other standardization methods, but the advantage of just subtracting the mean is that the interpretation of the coefficients remains the same. The coefficients continue to represent the mean change in the dependent variable given a 1 unit change in the independent variable.

In the worksheet, I've included the centered independent variables in the columns with an S added to the variable names.

For more about this, read my post about standardizing your continuous independent variables.

Regression with Centered Variables

Let’s fit the same model but using the centered independent variables.



Regression Analysis: Femoral Neck versus %Fat S, Weight S, Activity S

```
Analysis of Variance

Source              DF   Adj SS    Adj MS  F-Value  P-Value
Regression           4  0.55578  0.138946    27.95    0.000
  %Fat S             1  0.04786  0.047863     9.63    0.003
  Weight S           1  0.30473  0.304728    61.29    0.000
  Activity S         1  0.04703  0.047027     9.46    0.003
  %Fat S*Weight S    1  0.04175  0.041745     8.40    0.005
Error               87  0.43256  0.004972
Total               91  0.98834


Model Summary

        S    R-sq  R-sq(adj)  R-sq(pred)
0.0705118  56.23%     54.22%      50.48%


Coefficients

Term                  Coef   SE Coef  T-Value  P-Value   VIF
Constant           0.82161   0.00973    84.40    0.000
%Fat S            -0.00598   0.00193    -3.10    0.003  3.32
Weight S           0.00835   0.00107     7.83    0.000  4.75
Activity S        0.000022  0.000007     3.08    0.003  1.05
%Fat S*Weight S  -0.000214  0.000074    -2.90    0.005  1.99
```

The most apparent difference is that the VIFs are all down to satisfactory values; they’re all less than 5. By removing the structural multicollinearity, we can see that there is some multicollinearity in our data, but it is not severe enough to warrant further corrective measures.

Removing the structural multicollinearity produced other notable differences in the output that we’ll investigate.

Comparing Regression Models to Reveal Multicollinearity Effects

We can compare two versions of the same model, one with high multicollinearity and one without it. This comparison highlights its effects.

The first independent variable we’ll look at is Activity. This variable was the only one to have almost no multicollinearity in the first model. Compare the Activity coefficients and p-values between the two models and you’ll see that they are the same (coefficient = 0.000022, p-value = 0.003). This illustrates how only the variables that are highly correlated are affected by its problems.

Let’s look at the variables that had high VIFs in the first model. The standard error of the coefficient measures the precision of the estimates. Lower values indicate more precise estimates. The standard errors in the second model are lower for both %Fat and Weight. Additionally, %Fat is significant in the second model even though it wasn’t in the first model. Not only that, but the sign for %Fat has changed from positive to negative!

The lower precision, switched signs, and a lack of statistical significance are typical problems associated with multicollinearity.

Now, take a look at the Summary of Model tables for both models. You’ll notice that the standard error of the regression (S), R-squared, adjusted R-squared, and predicted R-squared are all identical. As I mentioned earlier, multicollinearity doesn’t affect the predictions or goodness-of-fit. If you just want to make predictions, the model with severe multicollinearity is just as good!

How to Deal with Multicollinearity

I showed how there are a variety of situations where you don’t need to deal with it. The multicollinearity might not be severe, it might not affect the variables you’re most interested in, or maybe you just need to make predictions. Or, perhaps it’s just structural multicollinearity that you can get rid of by centering the variables.

But, what if you have severe multicollinearity in your data and you find that you must deal with it? What do you do then? Unfortunately, this situation can be difficult to resolve. There are a variety of methods that you can try, but each one has some drawbacks. You’ll need to use your subject-area knowledge and factor in the goals of your study to pick the solution that provides the best mix of advantages and disadvantages.

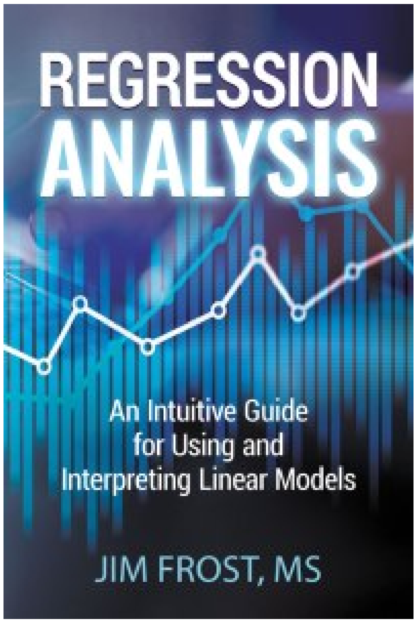
The potential solutions include the following:

- Remove some of the highly correlated independent variables.
- Linearly combine the independent variables, such as adding them together.
- Perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression.

As you consider a solution, remember that all of these have downsides. If you can accept less precise coefficients, or a regression model with a high R-squared but hardly any statistically significant variables, then not doing anything about the multicollinearity might be the best solution.

Do you have experience dealing with multicollinearity?

If you’re learning regression and like the approach I use in my blog, check out my eBook!


REGRESSION
ANALYSIS
An Intuitive Guide
for Using and
Interpreting Linear Models
JIM FROST, MS