What Are the Effects of Multicollinearity and When Can I Ignore Them?

Jim Frost (http://blog.minitab.com/blog/adventures-in-statistics-2) · 2 May, 2013

Multicollinearity is problem that you can run into when you're fitting a regression model, or other linear model. It refers to predictors that are correlated with other predictors in the model. Unfortunately, the effects of multicollinearity can feel murky and intangible, which makes it unclear whether it's important to fix.

My goal in this blog post is to bring the effects of multicollinearity to life with real data! Along the way, I'll show you a simple tool that can remove multicollinearity in some cases.

How Problematic is Multicollinearity?

Moderate multicollinearity may not be problematic. However, severe multicollinearity is a problem because it can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable and difficult to interpret. Multicollinearity saps the statistical power of the analysis, can cause the coefficients to switch signs, and makes it more difficult to specify the correct model.

My goal in this blog post is to bring multicollinearity to life with real data about bone density.

Do I Have to Fix Multicollinearity?

The symptoms sound serious, but the answer is both yes and no—depending on your goals. (Don't worry, the example we'll go through next makes it more concrete.) In short, multicollinearity:

- can make choosing the correct predictors to include more difficult.
- interferes in determining the precise effect of each predictor, but...
- doesn't affect the overall fit of the model or produce bad predictions.

Depending on your goals, multicollinearity isn't always a problem. However, because of the difficulty in choosing the correct model when severe multicollinearity is present, it's always worth exploring.

The Regression Scenario: Predicting Bone Density

I'll use a subset of real data that I collected for an experiment to illustrate the detection, effects, and removal of multicollinearity. You can read about the actual experiment here (http://blog.minitab.com/blog/adventures-in-statistics/the-mysteries-of-variability-and-power) and the worksheet is here (//cdn2.content.compendiumblog.com/uploads/user/458939f4-fe08-4dbc-b271-efca0f5a2682/479b4fbd-f8c0-4011-9409-f4109cc4c745

Might Also Like:

When Should You Fit a Non-Hierarchical Regression Model? (http://blog.minitab.com/blog/adventures-in-statistics-2/when-should-you-fit-a-non-hierarchical-regression-model)
Regression Analysis Tutorial and Examples (http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-tutorial-and-examples)
Multiple Regression Analysis: Use Adjusted R-Squared and Predicted R-Squared to Include the Correct Number of Variables (http://blog.minitab.com/blog/adventures-in-statistics-

2/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables)
How to Identify the Most Important Predictor Variables in Regression Models (http://blog.minitab.com/blog/adventures-in-statistics-2/how-to-identify-the-most-important-predictor-variables-in-regression-models)

/File/a99cb461c18026661da9938dfeaef811/multicollinearity.MTW). (If you're not already using it, please download the free 30-day trial of Minitab (http://www.minitab.com/en-us/products/minitab/free-trial/) and play along!)

We'll use Regression to assess how the predictors of physical activity, percent body fat, weight, and the interaction between body fat and weight are collectively associated with the bone density of the femoral neck.

Given the potential for correlation among the predictors, we'll have Minitab display the variance inflation factors (VIF) (http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regression-and-correlation/model-assumptions/what-is-a-variance-inflation-factor-vif/), which indicate the extent to which multicollinearity is present in a regression analysis. A VIF of 5 or greater indicates a reason to be concerned about multicollinearity.

General Regression Results

Here are the results of the Minitab analysis:

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

Regression Equation

Femoral Neck = 0.154934 + 0.00557106 %Fat + 0.014468 Weight kg + 2.23771e-005 Activity - 0.000214237 %Fat*Weight kg

Coefficients

Term	Coef	SE Coef	T	P	VIF
Constant	0.154934	0.131729	1.17616	0.243	
%Fat	0.005571	0.004087	1.36324	0.176	14.9316
Weight kg	0.014468	0.002852	5.07277	0.000	33.9484
Activity	0.000022	0.000007	3.07546	0.003	1.0530
%Fat*Weight kg	-0.000214	0.000074	-2.89762	0.005	75.0593

Summary of Model

S = 0.0705118 R-Sq = 56.23% R-Sq(adj) = 54.22%
PRESS = 0.489461 R-Sq(pred) = 50.48%

In the results above, Weight, Activity, and the interaction term are significant while %Fat is not significant. However, three of the VIFs are very high because they are well over 5. These values suggest that the coefficients (http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients) are poorly estimated and we should be wary of their p-values (http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients).

Standardize the Continuous Predictors

In this model, the VIFs are high because of the interaction term. Interaction terms and higher-order terms (http://blog.minitab.com/blog/adventures-in-statistics/curve-fitting-with-linear-and-nonlinear-regression) (e.g., squared and cubed predictors) are correlated with main effect terms because they include the main effects terms.

To reduce high VIFs produced by interaction and higher-order terms, you can standardize the continuous predictor variables. In Minitab, it's easy to standardize the continuous predictors by clicking the **Coding** button in Regression dialog box and choosing the standardization method.

For our purposes, we'll choose the **Subtract the mean** method, which is also known as centering the variables. This method removes the multicollinearity produced by interaction and higher-order terms as effectively as the other standardization methods, but it has the added benefit of not changing the interpretation of the coefficients. If you subtract the mean, each coefficient continues to estimate the change in the mean response per unit increase in

X when all other predictors are held constant.

I've already added the standardized predictors in the worksheet we're using; they're in the columns that have an S added to the name of each standardized predictor.

Regression with Standardized Predictors

We'll fit the same model as before, but this time using the standardized predictors.

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

Regression Equation

```
Femoral Neck  =  0.821609 - 0.00598238 %Fat S + 0.00834826 Weight S +
                 2.23771e-005 Activity S - 0.000214237 %Fat S*Weight S
```

Coefficients

Term	Coef	SE Coef	T	P	VIF
Constant	0.821609	0.0097344	84.4028	0.000	
%Fat S	-0.005982	0.0019281	-3.1027	0.003	3.32387
Weight S	0.008348	0.0010664	7.8288	0.000	4.74565
Activity S	0.000022	0.0000073	3.0755	0.003	1.05300
%Fat S*Weight S	-0.000214	0.0000739	-2.8976	0.005	1.99106

Summary of Model

```
S = 0.0705118     R-Sq = 56.23%       R-Sq(adj) = 54.22%
PRESS = 0.489461  R-Sq(pred) = 50.48%
```

In the model with the standardized predictors, the VIFs are down to an acceptable range.

Comparing Regression Models to See the Effects of Multicollinearity

Because standardizing the predictors effectively removed the multicollinearity, we could run the same model twice, once with severe multicollinearity and once with moderate multicollinearity. This provides a great head-to-head comparison and it reveals the classic effects of multicollinearity.

The standard error of the coefficient (SE Coef) indicates the precision of the coefficient estimates. Smaller values represent more reliable estimates. In the second model, you can see that the SE Coef is smaller for both %Fat and Weight. Also, %Fat is significant this time, while it was insignificant in the model with severe multicollinearity. Also, its sign has switched from + 0.005 to - 0.005! The %Fat estimate in both models is about the same absolute distance from zero, but it is only significant in the second model because the estimate is more precise.

Compare the Summary of Model statistics between the two models and you'll notice that S (http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-to-interpret-s-the-standard-error-of-the-regression), R-squared (http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit), adjusted R-squared (http://blog.minitab.com/blog/adventures-in-statistics/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables), and the others are all identical. Multicollinearity doesn't affect how well the model fits. In fact, if you want to use the model to make predictions (http://blog.minitab.com/blog/adventures-in-statistics/how-to-predict-with-minitab-using-bmi-to-predict-the-body-fat-percentage-part-1), both models produce identical results for fitted values and prediction intervals (http://blog.minitab.com/blog/adventures-in-statistics/when-should-i-use-confidence-intervals-prediction-intervals-and-tolerance-intervals)!

Multicollinear Thoughts

Multicollinearity can cause a number of problems. We saw how it sapped the significance of one of our predictors and changed its sign. Imagine trying to specify a model with many more potential predictors. If you saw signs that kept changing and incorrect p-values, it could be hard to specify the correct model! Stepwise regression does not work as well with multicollinearity. (http://blog.minitab.com/blog/adventures-in-statistics/which-is-better%2C-stepwise-regression-or-best-subsets-regression)

However, we also saw that multicollinearity doesn't affect how well the model fits. If the model satisfies the residual assumptions and has a satisfactory predicted R-squared, even a model with severe multicollinearity can produce great predictions.

You also don't have to worry about every single pair of predictors that has a high correlation. When putting together the model for this post, I thought for sure that the high correlation between %Fat and Weight (0.827) would produce severe multicollinearity all by itself. However, that correlation only produced VIFs around 3.2. So don't be afraid to try correlated predictors—just be sure to check those VIFs!

For our model, the severe multicollinearity was primarily caused by the interaction term. Consequently, we were able to remove the problem simply by standardizing the predictors. However, when standardizing your predictors doesn't work, you can try other solutions such as:

- removing highly correlated predictors
- linearly combining predictors, such as adding them together
- running entirely different analyses, such as partial least squares regression or principal components analysis

When considering a solution, keep in mind that all remedies have potential drawbacks. If you can live with less precise coefficient estimates, or a model that has a high R-squared but few significant predictors, doing nothing can be the correct decision because it won't impact the fit.

If you're learning about regression, read my regression tutorial (http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-tutorial-and-examples)!