Statistics By Jim

Making statistics intuitive

When Do You Need to Standardize the Variables in a Regression Model?

By Jim Frost — 11 Comments

Standardization is the process of putting different variables on the same scale. In regression analysis, there are some scenarios where it is crucial to standardize your independent variables or risk obtaining misleading results.

In this blog post, I show when and why you need to standardize your variables in regression analysis. Don't worry, this process is simple and helps ensure that you can trust your results. In fact, standardizing your variables can reveal essential findings th you would otherwise miss!

Why Standardize the Variables

In regression analysis, you need to standardize the independent variables when your model contains polynomial terms to model curvature or interaction terms. These terms provide crucial information about the relationships between

the independent variables and the dependent variable, but they also generate high amounts of multicollinearity.

Multicollinearity refers to independent variables that are correlated. This problem can obscure the statistical significance of model terms, produce imprecise coefficients, an make it more difficult to choose the correct model.

When you include polynomial and interaction terms, your model almost certainly has excessive amounts of multicollinearity. These higher-order terms multiply independen variables that are in the model. Consequently, it's easy to see how these terms are correlated with other independent variables in the model.

When your model includes these types of terms, you are at risk of producing misleadi results and missing statistically significant terms.

Fortunately, we're in luck because standardizing the independent variables is a simple method to reduce multicollinearity that is produced by higher-order terms. Although, it important to note that it won't work for other causes of multicollinearity.

Standardizing your independent variables can also help you determine which variable the most important. Read how in my post: Identifying the Most Important Independent Variables in Regression Models.

How to Standardize the Variables

Standardizing variables is a simple process. Most statistical software can do this for y automatically. Usually, standardization refers to the process of subtracting the mean a dividing by the standard deviation. However, to remove multicollinearity caused by higher-order terms, I recommend only subtracting the mean and **not** dividing by the standard deviation. Subtracting the means is also known as centering the variables.

Centering the variables and standardizing them will both reduce the multicollinearity. However, standardizing changes the interpretation of the coefficients. So, for this post center the variables.

Interpreting the Results for Standardized Variables

When you center the independent variables, it's very convenient because you can interpret the regression coefficients in the usual way. Consequently, this approach is e to use and produces results that are easy to interpret.

Let's go through an example that illustrates the problems of higher-order terms and h centering the variables resolves them. You can try this example yourself using the CS data file: TestSlopes.

Regression Model with Unstandardized Independent Variables

First, we'll fit the model without centering the variables. Output is the dependent varia And, we'll include Input, Condition, and the interaction term Input*Condition in the regression model. The results are below.

```
Coefficients

Term                Coef  SE Coef  T-Value  P-Value   VIF
Constant           9.099    0.980     9.29    0.000
Input             1.5359   0.0823    18.67    0.000  2.00
Condition
  B                -2.36     1.39    -1.70    0.093  4.48
Input*Condition
  B                0.469    0.116     4.03    0.000  5.48


Regression Equation

Condition
A          Output = 9.099 + 1.5359 Input

B          Output = 6.740 + 2.0050 Input
```

Using a significance level of 0.05, Input and Input*Condition are statistically significan while Condition is not. However, notice the VIF values. VIFs greater than 5 indicate th you have problematic levels of multicollinearity. Condition and the interaction term bot have VIFs near 5.

Related post: Understanding Interaction Effects

Regression Model with Standardized Variables

Now, let's fit the model again, but we'll standardize the independent variables using th centering method.

Coded Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	25.226	0.463	54.50	0.000	
Input	1.5359	0.0823	18.67	0.000	2.00
Condition					
B	2.567	0.655	3.92	0.000	1.00
Input*Condition					
B	0.469	0.116	4.03	0.000	2.00

Regression Equation in Uncoded Units

Condition
A Output = 9.099 + 1.5359 Input

B Output = 6.740 + 2.0050 Input

Standardizing the variables has reduced the multicollinearity. All VIFs are less than 5. Furthermore, Condition is statistically significant in the model. Previously, multicollinea was hiding the significance of that variable.

The coded coefficients table shows the coded (standardized) coefficients. My softwar converts the coded values back to the natural units in the Regression Equation in Uncoded Units. Interpret these values in the usual manner.

Standardizing the independent variables produces vital benefits when your regressior model includes interaction terms and polynomial terms. Always standardize your variables when the model has these terms. Keep in mind that it is enough to center th variables for a more straightforward interpretation. It's an easy thing to do, and you ca have more confidence in the results.

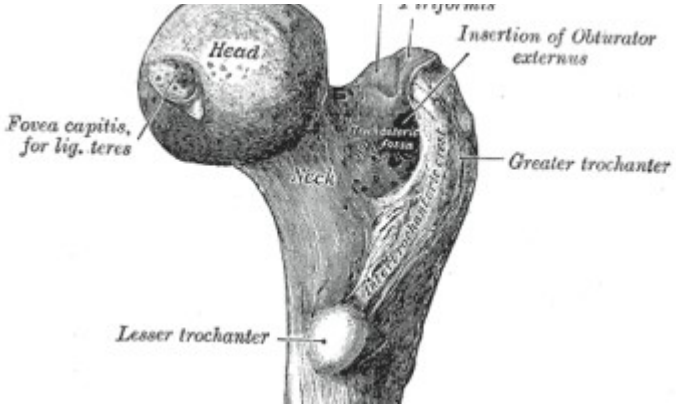
For more information about multicollinearity, plus another example of how standardizi the independent variables can help, read my post: Multicollinearity in Regression Analysis: Problems, Detection, and Solutions. The example in that post shows how multicollinearity can change the sign of a coefficient!

If you're learning regression, check out my Regression Tutorial!

Share this:

- Share 44
- Share
- Tweet
- Save

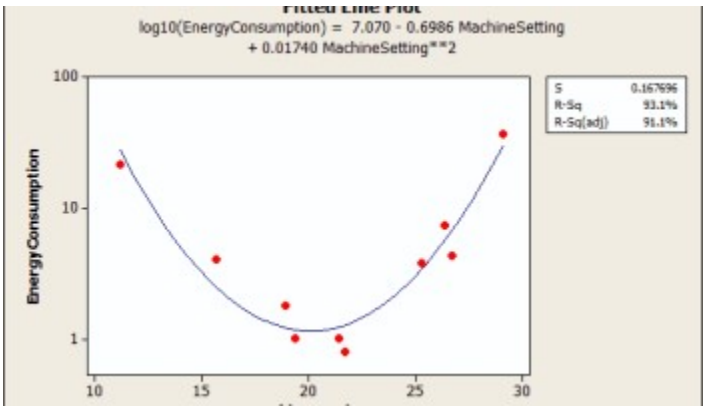
Related



Multicollinearity in Regression Analysis: Problems, Detection, and Solutions

Multicollinearity occurs when independent variables in a regression model are correla This correlation is a problem because independent variables should be independent. the degree of correlation between variables is high enough, it can cause problems wh you fit the model and interpret the results. In this blog post, I'll…

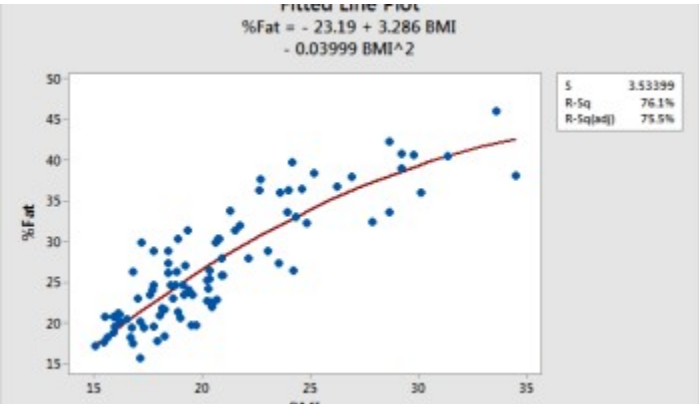
In "Regression"



When Should I Use Regression Analysis?

Use regression analysis to describe the relationships between a set of independent variables and the dependent variable. Regression analysis produces a regression equation where the coefficients represent the relationship between each independent variable and the dependent variable. You can also use the equation to make predictio As a statistician, I…

In "Regression"



Regression Tutorial with Analysis Examples

Regression analysis mathematically describes the relationship between independent variables and the dependent variable. It also allows you to predict the mean value of t dependent variable when you specify values for the independent variables. In this regression tutorial, I gather together a wide range of posts that I've written about…

In "Regression"

Filed Under: Regression Tagged With: analysis example, interpreting results

Comments

1.  Gabriel Samuel says

October 22, 2018 at 12:44 am

Your blog is awesome. I'm grateful I got hooked at this point in my thesis write up. Th
and keep up the good work.

Loading...

Reply

- 

  Jim Frost says

  October 22, 2018 at 1:50 am

  Thank you so much, Gabriel! I'm so happy to hear that my blog has been helpful. Be
  with your thesis!

  Loading...

  Reply

2. 

   Shaarang says

   August 7, 2018 at 10:55 pm

   As an aspiring data scientist, I can not overstate how helpful your setup has been. Th
   ton

   Loading...

   Reply

   - 

     Jim Frost says

     August 8, 2018 at 4:50 pm

     You're very welcome! It makes my day hearing how it has been helpful for you!

     Loading...

     Reply

3. 

   Luke says

   May 7, 2018 at 7:56 pm

   Great article! Thanks for sharing. I do have a question regarding what you said here
   "However, to remove multicollinearity caused by higher-order terms, I recommend or

subtracting the mean and not dividing by the standard deviation. Subtracting the mea also known as centering the variables", would you elaborate how will it cause proble dividing the standard deviation after centering?

Loading...

Reply

- 

  Jim Frost says

  May 8, 2018 at 10:09 am

  Hi Luke,

  All I meant by that was that if you just center the variables, the interpretation of the coefficients doesn't change from their normal interpretation that a coefficient indicate mean change in the dependent variable given a one-unit change in the independent However, if you also divide by the standard deviation, the interpretation of the coeffic changes. For that case, the coefficient represents the mean change in the DV for a 1 standard deviation change in the IV.

  I write about how standardizing your continuous IVs can be helpful in a post about H Identify the Most Important Independent Variables in Your Model. You can read more that approach in that post.

  I hope this helps!

  Loading...

  Reply

4. 

   Douglas AMULI says

   February 9, 2018 at 4:15 am

   I found very helpful your post.
   Concerning it i have two questions:

   – Is it a problem if one runs a regression model where some independent variables a standardized and others are not ?

   – Imagine a particular case of a mimic model with standardized causes but not stanc indicators. Are results negatively affected ?

Thanks in advance for your reply.

Loading...

Reply

5. 

   Visar says

   January 22, 2018 at 9:57 am

   This was very useful. Thanks a lot and keep up the good work!

   Loading...

   Reply

   ◦ 

   Jim Frost says

   January 22, 2018 at 10:02 am

   Thank you, Visar!

   Loading...

   Reply

6. 

   Karien says

   December 27, 2017 at 12:52 pm

   Hi Jim,

   Thank you for your posts. I have to do a statistical analysis for a project and I have n
   to delve so deep into statistics before. Your plain English explanations really help a l

   When dealing with interactions, do you first get the interactions between the variable
   center them as well? Or do you center the independent variables and then get the
   interactions? Or, are the interactions from the original independent variables and the
   the independent variables are centered? I am very confused about the order of thing
   Also if you have more than one interaction that is significant, does it become another
   the regression equation?

   Thanks,
   Karien

   Loading...

Reply

Jim Frost says

December 27, 2017 at 1:07 pm

Hi Karien,

I'm so glad to hear that my blog posts are helpful!

To answer your question, some of it depends on your statistical software. If it can do things for you automatically, then you don't have to worry about it.

However, if you need to do them manually, here's to correct order.
1. Create a new column for each continuous independent variable you need to cente
2. Center the continuous variables in the new columns.
3. Create a new column for each interaction term.
4. Create the interaction term by multiplying the appropriate columns. Be sure to use centered variables.

Again, many software packages can do some or all of these steps for you automatica you might not need to worry, but do check the documentation for the software.

I hope this helps!
Jim

Loading...
Reply

Leave a Reply

Enter your comment here...

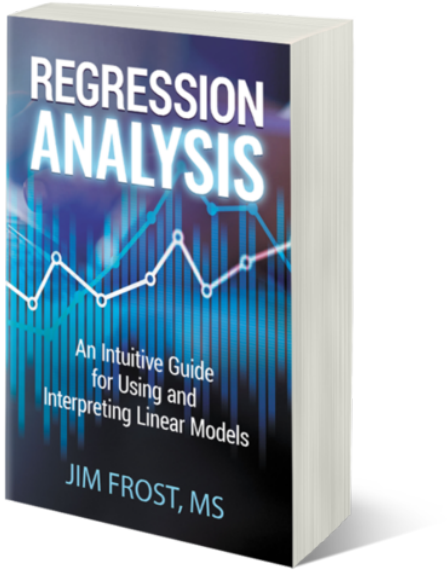
Meet Jim

I'll help you intuitively understand statistics by focusing on concepts and using

plain English so you can concentrate on understanding your results.

Read More…

Buy My eBook!



Regression Analysis: An Intuitive Guide [ebook]

Over the course of this full-length ebook, you'll progress from a beginner to a skilled practitioner. I'll help you intuitively understand regression analysis by focusing on…

$14.00 USD

Buy it now

Search this website

Subscribe via Email!

Enter your email address to receive notifications of new posts by email.

Email Address

Subscribe

Follow Me

Facebook

RSS Feed

Twitter

Popular

Latest

How To Interpret R-squared in Regression Analysis

How to Interpret P-values and Coefficients in Regression Analysis

Understanding Interaction Effects in Statistics

How to Interpret the F-test of Overall Significance in Regression Analysis

Measures of Central Tendency: Mean, Median, and Mode

Multicollinearity in Regression Analysis: Problems, Detection, and Solutions

The Importance of Statistics

Recent Comments

Jim Frost on Heteroscedasticity in Regression Analysis

Luca romen on How To Interpret R-squared in Regression Analysis

phabdallah on Heteroscedasticity in Regression Analysis

Jim Frost on Choosing the Correct Type of Regression Analysis

Syed Abbas on Choosing the Correct Type of Regression Analysis