What is Multicollinearity? Here's Everything You Need to Know

Home

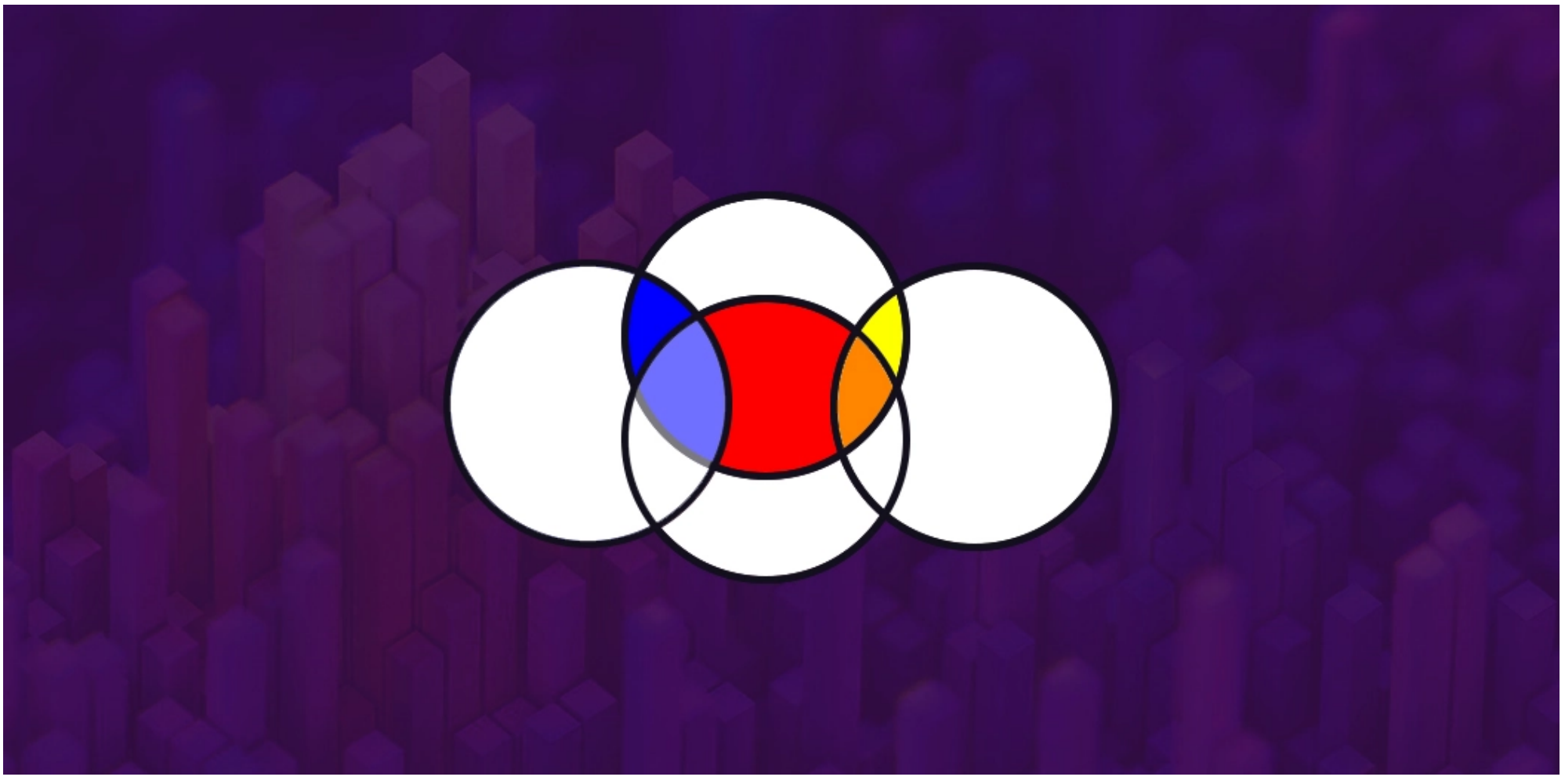
Aniruddha Bhandari — March 20, 2020

Beginner Python Regression Statistics Structured Data Technique

Introduction

Multicollinearity might be a handful to pronounce but it's a topic you should be aware of in the machine learning field. I am familiar with it because of my statistics background but I've seen a lot of professionals unaware that multicollinearity exists.

This is especially prevalent in those machine learning folks who come from a non-mathematical background. And while yes, multicollinearity might not be the most crucial topic to grasp in your journey, it's still important enough to learn. Especially if you're sitting for data scientist interviews!



So in this article, we will understand what multicollinearity is, why it's a problem, what causes multicollinearity, and then understand how to detect and fix multicollinearity.

Before diving further, it is imperative to have a basic understanding of regression and some statistical terms. For this, I highly recommend going through the below resources:

- *Fundamentals of Regression Analysis (Free Course!)*
- *Beginner's Guide to Linear Regression*

Table of Contents

What is Multicollinearity? Here's Everything You Need to Know

Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model.

This means that an independent variable can be predicted from another independent variable in a regression model. For example, height and weight, household income and water consumption, mileage and price of a car, study time and leisure time, etc.

Let me take a simple example from our everyday life to explain this. Colin loves watching television while munching on chips. The more television he watches, the more chips he eats and the happier he gets!

Now, if we could quantify happiness and measure Colin's happiness while he's busy doing his favorite activity, which do you think would have a greater impact on his happiness? Having chips or watching television? That's difficult to determine because the moment we try to measure Colin's happiness from eating chips, he starts watching television. And the moment we try to measure his happiness from watching television, he starts eating chips.

Eating chips and watching television are highly correlated in the case of Colin and we cannot individually determine the impact of the individual activities on his happiness. This is the multicollinearity problem!

So why should you worry about multicollinearity in the machine learning context? Let's answer that question next.

The Problem with having Multicollinearity

Multicollinearity can be a problem in a regression model because we would not be able to distinguish between the individual effects of the independent variables on the dependent variable. For example, let's assume that in the following linear equation:

$$Y = W0+W1*X1+W2*X2$$

Coefficient W1 is the increase in Y for a unit increase in X1 while keeping X2 constant. But since X1 and X2 are highly correlated, changes in X1 would also cause changes in X2 and we would not be able to see their individual effect on Y.

> " This makes the effects of X1 on Y difficult to distinguish from the effects of X2 on Y. "

Multicollinearity may not affect the accuracy of the model as much. But we might lose reliability in determining the effects of individual features in your model – and that can be a problem when it comes to interpretability.

What causes Multicollinearity?

Multicollinearity could occur due to the following problems:

- Multicollinearity could exist because of the problems in the dataset at the time of creation. These problems could be because of poorly designed experiments, highly observational data, or the inability to manipulate the data:
  - For example, determining the electricity consumption of a household from the household income and the number of electrical appliances. Here, we know that the number of electrical appliances in a household will increase with household income. However, this cannot be removed from the dataset
- Multicollinearity could also occur when new variables are created which are dependent on other variables:
  - For example, creating a variable for BMI from the height and weight variables would include redundant information in the model
- Including identical variables in the dataset:
  - For example, including variables for temperature in Fahrenheit and temperature in Celsius
- Inaccurate use of dummy variables can also cause a multicollinearity problem. This is called the **Dummy variable trap**:

What is Multicollinearity? Here's Everything You Need to Know

Creating dummy variables for both of them would include redundant information. We can make do with only one variable containing 0/1 for 'married'/'single' status.

- Insufficient data in some cases can also cause multicollinearity problems

Detecting Multicollinearity using VIF

Let's try detecting multicollinearity in a dataset to give you a flavor of what can go wrong.

I have created a dataset determining the salary of a person in a company based on the following features:

- Gender (0 – female, 1- male)
- Age
- Years of service (Years spent working in the company)
- Education level (0 – no formal education, 1 – under-graduation, 2 – post-graduation)

```
df=pd.read_csv(r'C:/Users/Dell/Desktop/salary.csv')
df.head()
```

Multicollinearity_import.py hosted with ❤ by GitHub view raw

	Gender	Age	Years of service	Education level	Salary
0	0.0	27.0	1.7	0.0	39343.0
1	1.0	26.0	1.1	1.0	43205.0
2	1.0	26.0	1.2	0.0	47731.0
3	0.0	27.0	1.6	1.0	46525.0
4	0.0	26.0	1.5	1.0	40891.0

Multicollinearity can be detected via various methods. In this article, we will focus on the most common one – **VIF (Variable Inflation Factors)**.

> " VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable. "
>
> or
>
> VIF score of an independent variable represents how well the variable is explained by other independent variables.

R^2 value is determined to find out how well an independent variable is described by the other independent variables. A high value of **R^2** means that the variable is highly correlated with the other variables. This is captured by the **VIF** which is denoted below:

$$VIF = \frac{1}{1-R^2}$$

So, the closer the **R^2** value to 1, the higher the value of VIF and the higher the multicollinearity with the particular independent variable.

```
# Import library for VIF
from statsmodels.stats.outliers_influence import variance_inflation_factor

def calc_vif(X):

    # Calculating VIF
    vif = pd.DataFrame()
    vif["variables"] = X.columns
    vif["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]

    return(vif)
```

What is Multicollinearity? Here's Everything You Need to Know

- VIF starts at 1 and has no upper limit
- VIF = 1, no correlation between the independent variable and the other variables
- VIF exceeding 5 or 10 indicates high multicollinearity between this independent variable and the others

```
X = df.iloc[:,:-1]
calc_vif(X)
```

Multicollinearity_VIF_All.py hosted with ❤ by GitHub view raw

	variables	VIF
0	Gender	2.207155
1	Age	13.706320
2	Years of service	10.299486
3	Education level	2.409263

We can see here that the 'Age' and 'Years of service' have a high VIF value, meaning they can be predicted by other independent variables in the dataset.

Although correlation matrix and scatter plots can also be used to find multicollinearity, their findings only show the bivariate relationship between the independent variables. VIF is preferred as it can show the correlation of a variable with a group of other variables.

Fixing Multicollinearity

Dropping one of the correlated features will help in bringing down the multicollinearity between correlated features:

```
X = df.drop(['Age','Salary'],axis=1)
calc_vif(X)
```

Multicollinearity_VIF_Drop.py hosted with ❤ by GitHub view raw

	variables	VIF
0	Gender	2.207155
1	Age	13.706320
2	Years of service	10.299486
3	Education level	2.409263

	variables	VIF
0	Gender	1.863482
1	Years of service	2.478640
2	Education level	2.196539

The image on the left contains the original VIF value for variables and the one on the right is after dropping the 'Age' variable.

We were able to drop the variable 'Age' from the dataset because its information was being captured by the 'Years of service' variable. This has reduced the redundancy in our dataset.

Dropping variables should be an iterative process starting with the variable having the largest VIF value because its trend is highly captured by other variables. If you do this, you will notice that VIF values for other variables would have reduced too, although to a varying extent.

In our example, after dropping the 'Age' variable, VIF values for all the variables have decreased to a varying extent.

Next, combine the correlated variables into one and drop the others. This will reduce the multicollinearity:

```
df2 = df.copy()
df2['Age_at_joining'] = df.apply(lambda x: x['Age'] - x['Years of service'],axis=1)
```

What is Multicollinearity? Here's Everything You Need to Know

Multicollinearity_VIF_Join.py hosted with ❤ by GitHub view raw

	variables	VIF
0	Gender	2.207155
1	Age	13.706320
2	Years of service	10.299486
3	Education level	2.409263

	variables	VIF
0	Gender	2.168068
1	Education level	2.407695
2	Age_at_joining	3.326991

The image on the left contains the original VIF value for variables and the one on the right is after combining the 'Age' and 'Years of service' variable. Combining 'Age' and 'Years of experience' into a single variable 'Age_at_joining' allows us to capture the information in both the variables.

However, multicollinearity may not be a problem every time. The need to fix multicollinearity depends primarily on the below reasons:

1. When you care more about how much each individual feature rather than a group of features affects the target variable, then removing multicollinearity may be a good option
2. If multicollinearity is not present in the features you are interested in, then multicollinearity may not be a problem.

End Notes

Knowledge about multicollinearity can be quite helpful when you're building interpretable machine learning models.

I hope you have found this article useful in understanding the problem of multicollinearity and how to deal with it. If you want to understand other regression models or want to understand model interpretation, I highly recommend going through the following wonderfully written articles:

- Regression Modeling
- Machine Learning Model Interpretability

You should also check out the Fundamentals of Regression (free) course as a next step.

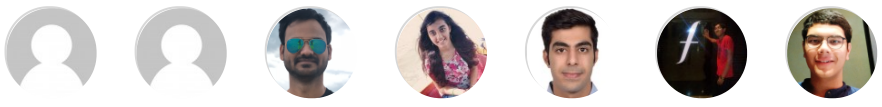
linear regression multicollinearity multicollinearity machine learning multicollinearity statistics python statistics VIF

About the Author

Aniruddha Bhandari

I am on a journey to becoming a data scientist. I love to unravel trends in data, visualize it and predict the future with ML algorithms! But the most satisfying part of this journey is sharing my learnings, from the challenges that I face, with the community to make the world a better place!

Our Top Authors



What is Multicollinearity? Here's Everything You Need to Know

Previous Post

TensorFlow 2.0 Tutorial for Deep Learning

Next Post

Free GPUs for Everyone! Get Started with Google Colab for Machine Learning and Deep Learning

15 thoughts on "What is Multicollinearity? Here's Everything You Need to Know"

Naveen kumar Mamidala says:
March 20, 2020 at 2:34 pm

How does non linear algo handle multi colinearity

Reply

Aniruddha Bhandari says:
March 21, 2020 at 7:37 pm

For tree-based algorithms, multicollinearity wouldn't matter much as they split on the feature that gives higher information gain. However, for other algorithms like polynomial regression and SVM, regularization can be used.

Reply

Christophe Bunn says:
March 22, 2020 at 7:46 pm

Hi Aniruddha, when you wrote "Coefficient W1 is the increase in Y for a unit increase in W1 while keeping X2 constant." didn't you mean "Coefficient W1 is the increase in Y for a unit increase in X1 while keeping X2 constant."? Cheers, Chris.

Reply

Aniruddha Bhandari says:
March 24, 2020 at 9:29 am

Hey Chris, thanks for pointing out the mistake.

Reply

Parvesh says:
April 10, 2020 at 5:53 pm

Hi Aniruddha I found this article very useful, could you share dataset so that readers may implement code at their end to get maximum out of this article

Reply

Aniruddha Bhandari says:
April 15, 2020 at 11:26 am

Hi Parvesh Glad you liked the article. I created a dummy dataset for this article. You can access it at this link. Thanks

Reply

Tudor Cristian Bogdan says:
July 08, 2020 at 3:06 pm

Thank you for explaining this in such fashion, it helped me understand what is and how to deal with VIF.

Reply

Chris says:
July 29, 2020 at 3:44 pm

Thanks for the article. When you talked categorical data being hot encoded, do we still need to perform the VIF on the encoded variable to see if it is highly correlated to other variables?

What is Multicollinearity? Here's Everything You Need to Know

Jairam Desik says:
August 20, 2020 at 1:18 am

My every doubt regarding Reduction of Multivariate correlation is cleared by this article. Thank You very much.

Reply

Franco Arda says:
August 20, 2020 at 2:36 pm

Hi ANIRUDDHA, excellent post straight to the point. I might add the level of when VIF might become problematic. In general, if the VIF value exceeds 5 - 10 indicates a problem (source: Introduction to Statistical Learning, page 101). Cheerio, Franco

Reply

Aniruddha Bhandari says:
August 24, 2020 at 11:34 pm

Yes, you can check the multicollinearity.

Reply

Aniruddha Bhandari says:
August 24, 2020 at 11:35 pm

Glad you found it useful!

Reply

Aniruddha Bhandari says:
August 24, 2020 at 11:36 pm

Thanks, Franco!

Reply

someone says:
August 29, 2020 at 8:55 pm

great article. thanks.

Reply

Vartika says:
September 17, 2020 at 12:40 pm

very informative .Thanks for such a clear explanation

Reply

Leave a Reply

Your email address will not be published. Required fields are marked *

Comment

Name*

Email*

Website

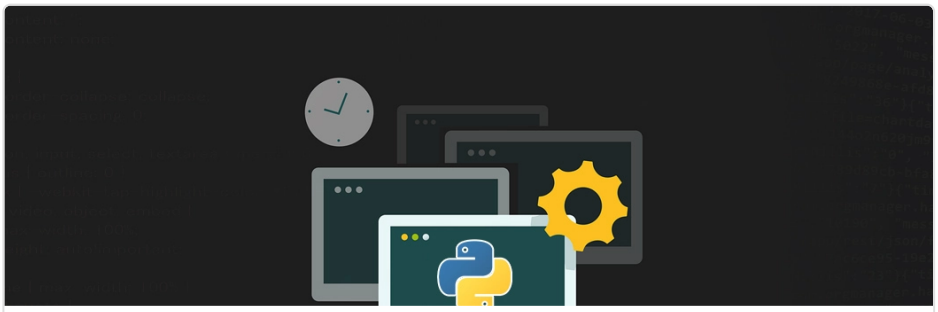
☐ Notify me of follow-up comments by email.

☐ Notify me of new posts by email.

Submit

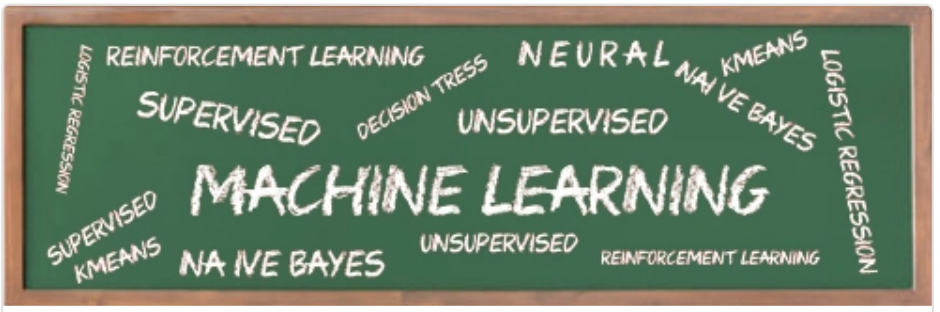
What is Multicollinearity? Here's Everything You Need to Know

Top Resources



Python Tutorial: Working with CSV file for Data Science

Harika Bonthu - AUG 21, 2021

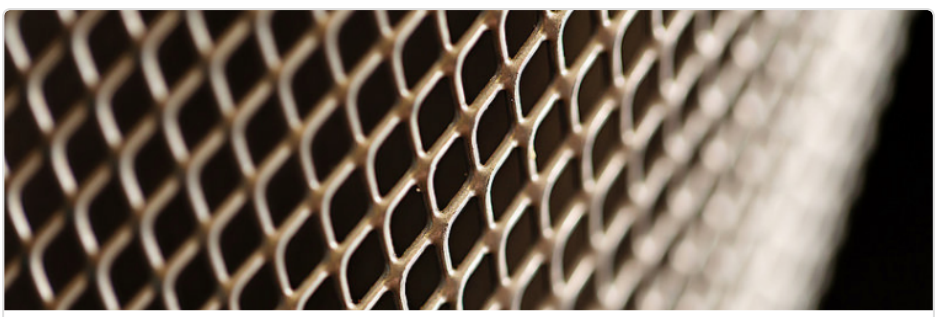


Commonly used Machine Learning Algorithms (with Python and R Codes)

Sunil Ray - SEP 09, 2017

A Comprehensive Guide to PySpark RDD Operations

Rahul Shah - OCT 09, 2021



40 Questions to test a Data Scientist on Clustering Techniques..

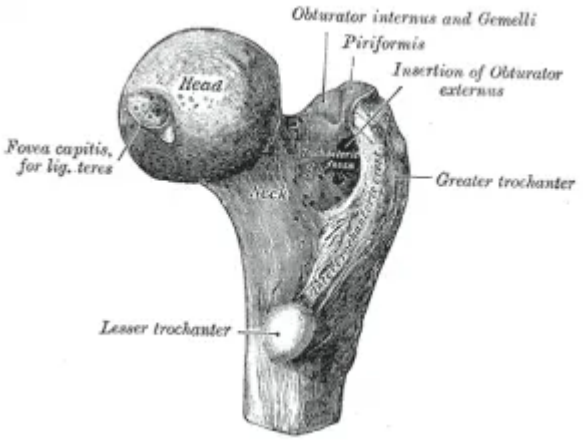
Saurav Kaushik - FEB 05, 2017

Multicollinearity in Regression Analysis: Problems, Detection, and Solutions

By Jim Frost — 171 Comments

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!



0

Multicollinearity in Regression

from **Jim Frost**

mineral
ɔ, pardon
ɔf
andyke

07:17

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.

Related post: What are Independent and Dependent Variables?

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 estimated 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.

If you're performing a designed experiment, it is likely orthogonal, meaning it has zero multicollinearity. Learn more about orthogonality.

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 coefficient 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.

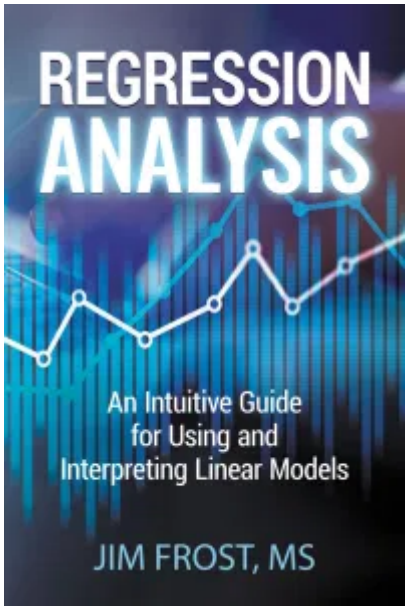
The potential solutions include the following:

- 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.
- LASSO and Ridge regression are advanced forms of regression analysis that can handle multicollinearity. If you know how to perform linear least squares regression, you'll be able to handle these analyses with just a little additional study.

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.

In this post, I use VIFs to check for multicollinearity. For a more in-depth look at this measure, read my post about Calculating and Assessing Variance Inflation Factors (VIFs).

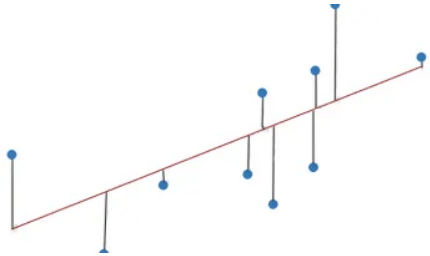
If you're learning regression and like the approach I use in my blog, check out my eBook!



Related

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

Variance Inflation Factors (VIFs)
In "Regression"



Choosing the Correct Type of Regression Analysis
In "Regression"

Filed Under: Regression

Tagged With: analysis example, conceptual, interpreting results

Comments

RABIA NOUSHEEN says
June 28, 2021 at 3:36 am

Hi Jim

I want to ask if main effects (Individual predictors) show multicollinearity greater than the interaction terms in regression analysis then is it sensible to remove the individual terms from the model?

Thank you for your help

Loading...

Reply

0

Taiwo says
June 15, 2021 at 8:15 am

What an awesome simply work. Keep it up

Loading...

Reply

Jim Frost says
June 19, 2021 at 4:18 pm

Thanks, Taiwo!

Loading...

Reply

0

Joblin Omari says
May 12, 2021 at 1:00 am

This is really helpful, for the first time i have seen statistics simply explained

Loading...

Reply

Bayarbaatar says
May 6, 2021 at 4:06 am

Dear Jim,
Thank you for your great explanation about multicollinearity. It was really understandable and made me more confident.

So I would like to ask you about what software (R or SPSS) is best and easy to use to detect VIF on windows operation system

Loading...

Reply

0

Amanda says
April 30, 2021 at 10:45 am

"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 really helped me understand why multicollinearity can be a problem. Thank you for this well-written post!

Loading...

Reply

Jim Frost says
April 30, 2021 at 3:03 pm

Hi Amanda, you're very welcome! So glad to hear that it's been helpful.

Loading...

Reply

muhammad ali raza abid says
April 24, 2021 at 2:35 am

0

Dear Jim I am pretty thankful that you are sharing your experience. I did a big mistake in centering the variables! Felt shameful for that! What I did is that I took the mean of the dataframe (including all IV's) and then subtracted the original ones from the mean which increased VIF.
However Good news for me! I want to tell you that actually VIF's are lowered after subtracting the mean form the variable 1st and then creating the polynomial terms from that variable. However the lowered VIF's are still ranging from 100-600. Before they were in the order of 1000's. This is well in agreement with the fact you mentioned before that the question is not that VIF's are lowered or not after centering but the question is how much VIF's are reduced. So in my case they are not reduced enough probably! I have also noted your comment that 5 degree polynomial might be doing overfitting. Actually my purpose for creating 5 degree polynomials is nothing but to improve the prediction accuracy which I am getting from the model but unfortunately at the expense of high VIF's which are though now reduced a bit after centering but still multicollinearity is there. Any suggestions regarding this case study will be highly appreciated please. Thanks!

Loading...

Reply

muhammad ali raza abid says

April 22, 2021 at 2:52 am

0

Dear Jim, I have rechecked again and found that VIF increases by centering the variables i.e. taking the mean and subtracting them , again testing for VIF. I am also surprised to see my results! Any suggestions to further investigate this issue will be highly appreciated! Actually in my case I have 19 IV's. I found the most important IV and created 5 degree polynomials for that variable. This process significantly increased by Adjusted R Square (about 10%). But the issue of multicollinearity remain there! Thanks for considering my question.

Loading...

Reply

Jim Frost says
April 23, 2021 at 11:27 pm

Hi Muhammad,

I'm still at a loss. However, if your fitting a 5th degree polynomial is extremely unusual. That's almost definitely overfitting your data unless there is a very particular reason that your data requires it. It might also explain why centering isn't helping the VIFs! In practice, I almost never see even a 3rd degree polynomial. Usually a quadratic term is the highest in practice.

Loading...

Reply

0

Batru Wolde says
April 21, 2021 at 11:58 am

Hi Jim Thank you for your contributions. Really the material is help full I used it for my research work. Keep it up

Loading...

Reply

muhammad ali raza abid says
April 21, 2021 at 4:46 am

Hello Jim

Thanks for this wonderful article. In my case I intentionally created polynomial features and observed that my linear regression model performance is significantly increased. However at the same time as you mentioned I observed that the polynomial features have high correlation with the feature from which they are created and also high VIF values of the order of 1000. Then I tried to do the same as mentioned here i.e.

calculated the mean and then subtracted mean from the features to create a new data frame. Then I used that data frame to calculate VIF and correlations. But I observed that VIF values actually increases a bit instead of decreasing, the correlation values on the other hand remains the same as before. Could you please suggest something here!
Does v.high values of VIF i.e. the order of 1000 means something abnormal etc?
Although I did check the p-values and they are well below 0.05.
Looking forward to your kind advice on this. Thanks

Loading...

Reply

Jim Frost says
April 21, 2021 at 2:00 pm

Hi Muhammad,

I have to be honest, I've never heard of a case where centering the variables actually increased the VIFs! Are you sure the centering process was implemented correctly? It should reduce the VIFs and the main question is whether it reduces them by a sufficient amount. Again, I've never seen a case where it increase them and that really shouldn't happen. After you center all of the variables, the mean of each variable should be virtually zero.

I will need to think about this one to see if there are cases where VIFs could increase.

Loading...

Reply

0

David says
March 14, 2021 at 4:47 pm

Dear Jim

this post is early amazing and easy to understand. but I have collinearity problem with my data. my data is a univariate time series of the load consumption(with hourly resolution), I use the sliding technique to make the input vector of NN that includes the lag variables. I want to solve this with two-step correlation analysis, the firs one that is the easiest part and calculate the relevancy (correlation with the target, in which candidate variables with higher correlation value than threshold are retain) but the second one is problematic, and the idea(according to one of journal's paper) is cross correlation in which ' if the correlation index between
any two candidate variables is smaller than a pre-specified value cor2, then both variables are retained; else, only the variable with the largest correlation with respect to the output is retained, while the other is not considered any further'. but according to the correlation matrix it is not possible because all nearly located lags are highly correlated it means that according the assumption of the paper, variables should retain because their correlation index is lower than the second threshold and should be discard because the high correlation according to the recent lags.
for example: t-1, t-2, …,t-24,.., t-48 —> t-2 and t-48 should be retain because they are

not correlated but t-2 should be discard because it is highly correlated with t-1.
what is your opinion ? isn't it a contradict? how can I solve it in general and how can I solve it by correlation analysis?

Bests
David

Loading...

Reply

YJ says
February 13, 2021 at 11:16 pm

Hmm, okay, I think I can understand that, it targets structural multicollinearity rather than data multicollinearity.

I'm trying to think through the implications of this technique then of a regression model that exhibits multicollinearity, but not structurally. My intuition is that the difference between y = b0 + b1x1 + b2x2 and y = b0 + b1c(x1) + b2c(x2) is that the coefficients remain the same, but b0 increases by mean(x1)+mean(x2). But honestly, neither here nor there, since that's a data multicollinearity problem! I should just throw PCA at it.

Thanks for responding, by the way, I do appreciate the effort and work you put into these posts; it's hard to describe statistics as straightforwardly as you do!

Loading...

Reply

0

Jim Frost says
February 14, 2021 at 12:06 am

Hi YJ,

Yes, centering the variables changes the constant. But, it does reduce multicollinearity, which helps you accurately determine the statistical significance of your predictors. Multicollinearity reduces the ability to detect significant effects.

And, thanks! I do strive to explain statistics in a straightforward yet completely accurate manner, which can be difficult to both sometimes!

Loading...

Reply

YJ says
February 9, 2021 at 8:13 pm

Jim,

I'm having a really hard time intuiting how standardising the vars would reduce structural multicollinearity without it being a hack in the maths.

For instance, if we have two vars, x1 and x2, and we know them to be collinear, we would expect to see that relationship in the plot of x1 against x2. In effect we'd expect to see a high R2 between the two, instead of the 0 that denotes orthogonality. But mean(x1) and mean(x2) are both constants, so subtracting their means from the observed values would only translate the datapoints left and down: I don't see how that would reduce the R2, or the corr coefficient, s(x1) and s(x2) are still correlated to the same extent.

In thinking about it that only thing I can think in how it addresses that collinearity issue is that it percolates through to the actual regression, and "reduces" the effect this collinearity has on the dependent var by altering x1 and x2's coefficients while decreasing x1 and x2's numerical values, by essentially adding two more constants to the regression, so instead of y = b0 + b1x1 + b2x2 + e, regressing instead on y=b0 + b1x1 + b2x2 – b1.mean(x1) – b2.mean(x2) + e. So we kind of leave the collinearity there, but add more "secondary" constants to account for it. Is this correct?

Loading...

Reply

Jim Frost says
February 11, 2021 at 4:49 pm

Hi YJ,

It reduces the extremeness of the products. Say you're talking about an interaction X1*X2. By centering the values, they're all close to zero. That reduces the impact of the multiplication on larger values. It's an accepted and valid solution for reducing structural multicollinearity.

Note that you don't need to standardize (center and divide by the standard deviation). You only need the centering portion. However it works using either method. The advantage of centering only is that it doesn't change the interpretation of the coefficients.

Loading...

Reply

RABIA NOUSHEEN says
February 6, 2021 at 3:04 am

Hi Jim

It is a very useful article. Thanks for that.

I am recently facing multicollinearity in my data. Can you please share R codes to standardize the categorical variables and their regression. My variables are categorical like Toxicant type(4 levels), Exposure time(4 levels) and concentration(4 levels).

Loading...

Reply

0

Jim Frost says
February 6, 2021 at 11:46 pm

Hi Rabia, unfortunately R isn't my strong suit.

Loading...

Reply

Louise says
December 10, 2020 at 9:40 am

Thank you for a really clear and useful explanation. I have what may be a stupid question (apologies if it is!). I'm doing some analysis regarding a predictive test which generates complex spectroscopic data. Previous analysts have used binomial LR to combine related variables (spectroscopic readings of the same tissue at different frequencies, ie same subject and measurement obtained at more or less same time) to produce a summary measure which is then used to predict a dichotomous clinical outcome. The combined variables are obviously not independent and consequently have severe multicollinearity. But the output probabilities are only being used for prediction. I'm therefore unsure if this is a completely invalid approach or not! Much of

the literature in this field is engineering/computing based and quickly becomes incomprehensible (as someone not in those fields!).
Thanks again for a useful article!

Loading...

Reply

Jim Frost says
December 10, 2020 at 10:15 pm

Hi Louise,

This approach sounds like a good one to me. In my post towards the end, I have three bullet points for potential solutions. The second bullet states, Linearly combine the independent variables, such as adding them together. That's essentially what the researchers are doing. Combining highly correlated variables together. This removes the multicollinearity because you can have correlation between predictors when there is only one predictor! Well, you can have other predictors in the model, but this process combines the correlated predictors into a combined measure.

The downside of this approach is that you don't know the role of each predictor. However, if the primary goal is to predict the outcome rather than understanding the role of each predictor, that's not really a problem here.

So, I don't see anything inherently wrong with the approach for this case.

Thanks for writing with the interesting scenario!

Loading...

Reply

0

Fernanda Hornos says
December 1, 2020 at 10:57 am

Hi Jim! We have two IV that are highly correlated (-0.86) and our VIF is 3.90. Do you think that we continue to trust the data or do we have to make a PCA?

Loading...

Reply

Jim Frost says
December 1, 2020 at 11:34 pm

Hi Fernanda,

I don't think that's a problem. It does sound like a high correlation but the VIF is below 5. You should should be good. With just two IVs, you'd need an even higher correlation to be a problem. I did some calculations and found that the VIFs won't

reach 5 until the correlation reaches an absolute correlation of 0.894 when you're dealing with only two IVs. So, you're getting close but you should be OK. You can stick with regression! Thanks for writing!

Loading...

Reply

Danny Data says
November 29, 2020 at 9:53 pm

Jim,

I was working on finding the VIF in your dataset using Python to follow along with your post. I ran the following and my initial VIF for the Independent Variables was much higher:

import pandas as pd
from statsmodels.stats.outliers_influence import variance_inflation_factor

df = read_csv("MulticollinearityExample.csv")

df['%Fat*Weight kg'] = df['%Fat']*df['Weight kg']

x = df[['Activity','%Fat','Weight kg','%Fat*Weight kg']]

vif_data = pd.DataFrame()

vif_data['feature'] = x.columns

vif_data['VIF'] = [variance_inflation_factor(x.values, i) for i in range(x.shape[1])]

print(vif_data.round(2))

feature VIF
0 Activity 7.48
1 %Fat 53.21
2 Weight kg 66.93
3 %Fat*Weight kg 26.73

After applying initial standardization as noted, my VIF matched your initial VIF from your example:

for iv in ['Activity','%Fat','Weight kg','%Fat*Weight kg']:
df[iv] = [x – df[iv].mean() for x in df[iv]]

feature VIF
0 Activity 1.05
1 %Fat 14.93
2 Weight kg 33.95
3 %Fat*Weight kg 73.06

Applying it one more time I got your same end results:

feature VIF
0 Activity 1.05
1 %Fat 3.32

2 Weight kg 4.75
3 %Fat*Weight kg 1.99

Any idea why this occurred? Thanks for your help!

Loading...

Reply

Jim Frost says
December 2, 2020 at 12:11 am

Hi Danny,

I'm really puzzled as to what is happening. Unfortunately, I don't have sufficient Python knowledge to help there. However, I have double-checked the my VIF values and they're correct. You can try calculating the VIF yourself to see where the problem lies. Take one of the IVs and use that as the DV. Then take the remaining IVs and regress that on the new DV (the former IV). Don't include the original DV at all. From that regression take the R-squared and calculate the VIF: 1 / (1 – R Square). Perhaps that will point you in the right direction.

Loading...

Reply

Heather says
October 30, 2020 at 1:15 pm

0

Hi Jim,

Thanks so much for your response – this is really helpful and makes much more sense now! I checked the residuals following your instructions on the relevant blog page and all look fine – thank you for recommending this.

Best wishes,
Heather

Loading...

Reply

Jim Frost says
November 1, 2020 at 10:12 pm

Hi Heather,

You're very welcome! It's definitely good news that the residuals look good. That means there's no obvious problems!

Loading...

Reply

0

Prince says
October 15, 2020 at 5:31 pm

When doing multiple regression when does multicollinearity matter (i.e. worth examining) and when does it not?

Loading...

Reply

Jim Frost says
October 16, 2020 at 3:46 pm

Hi,

You're looking at the right article to answer your question. You'll find your answer in the section titled, "Do I Have to Fix Multicollinearity?" It doesn't make sense for me to retype it here in the comment but all the details are there! If anything is still not clear, please don't hesitate to ask!

Loading...

Reply

Heather says
October 9, 2020 at 7:49 am

0

Hi Jim,

Thanks so much for the post, it's very helpful. Do you have any advice on how to interpret a model which is significant overall, but non of the predictors reach significance? The highest VIF i have is 3, for two of the seven predictors which, after reading your blog, doesn't seem like something to worry about. I'm generally interested more in the predictors (i.e., which might predict the outcome most strongly). A significant overall model is great, however i'm not sure how to interpret the fact that non of the predictors are significant, or whether this indicates another problem in the model since most advice I am finding online is that this would normally indicate a problem with multicollinearity – but it seems this isn't the problem because VIFs aren't very big?

Thank you!

Loading...

Reply

Jim Frost says
October 13, 2020 at 2:42 pm

Hi Heather,

This situation usually happens when your model is borderline significant. I describe your situation in more detail in my post about the overall f-test of significance.

I agree that multicollinearity is not likely a problem given your VIFs. But, you're correct, when multicollinearity is present, that can happen.

What it probably means is that there is just enough evidence to conclude that your model as a whole predicts the DV better than just using the mean. However, there's insufficient evidence to link it to any individual IV. Technically, the significant overall f-test indicates that all seven of your coefficients are unlikely to equal zero simultaneously. However, there is not enough evidence to identify a specific coefficient that is different from zero. Being different from zero is equivalent to being statistically significant for these tests.

I explain this in more detail in the post that I link to above.

Basically, your model is borderline significant and there's just not enough information to pin down the significance more specifically to specific variables. That's how I'd read it. Another alternative is that your model isn't accounting for curvature or interaction terms. Check those residual plots to look for obvious patterns that you can correct.

I hope this helps!

Loading...

Reply

0

John Grenci says
October 4, 2020 at 6:24 pm

0

Hi Jim, as always much appreciate the articles, and this one in particular is quite interesting. Count me among the many who take issue with the third item listed above. I know that some things in mathematics are not perhaps as intuitive as others… but what if you have two IV's that are correlated. Assume that both have a low p value and for illustration, lets assume the coefficient of variable A is 3 and variable B is 5. Thus, we predict the DV by 3a+5b. However, is it not true that if we cannot trust the p values, then we cannot trust the coefficients, and thus cannot trust the prediction? thanks John

Loading...

Reply

Hans says
October 1, 2020 at 7:29 am

Hi Jim,
I'm running a GLM model. I have used VIF and remove the covariate that was >10.

I checked for correlation and there is still some covariates that is correlated. I remove one of the correlated covariate and check the BIC value, if it decrease I continue with it. At the same time I'm aware on statistical significancy. But two of my covariates is

correlated(0.62), but when I remove one of them from the GLM model BIC increase. All of my VIF values is below 2.5. What can I do?

Loading...

Reply

Jim Frost says
October 1, 2020 at 4:44 pm

Hi Hans,

In some cases, handling multicollinearity can be very difficult to remove. There are times when the treatment is worse than the problem.

A correlation of 0.62 isn't necessarily problematic. Give the VIFs priority over the pairwise correlations. If the VIFs are fine (less than 5), I wouldn't worry about the pairwise correlations.

If removing a correlated predictor improves the goodness-of-fit, that's also a net gain. A simpler model and an improved fit. That sounds promising!

If you really need to include a correlated predictor that you otherwise would need to exclude for the reasons above (may be you need to estimate its effect and test its significance specifically), I'd try Ridge or LASSO regression. These forms of regression analysis are better able to handle multicollinearity. I write a little bit about both of them in my post about Choosing the Correct Type of Regression Analysis. Look for Ridge and LASSO regression in that article to see what it can do

for you. I don't have an article that shows them in action, at least not yet. But, it would be an avenue for you to explore. Using these forms of regression, you'd leave the correlated predictors in the model.

I hope that helps!

Loading...

Reply

purva garg says
September 14, 2020 at 3:17 am

Hi! thanks for such informative article. In sources of multicollinearity does high leverage values also come? I can't understand the intuition behind it. it will be helpful if you could explain how high leverage values are a source of multicollinearity?

Loading...

Reply

Pavithra says
September 13, 2020 at 12:27 pm

0

Hi sir, I have one doubt in my research totally i taken 16 variables in that 11 variables are collinearity . i did structural model equation with 11 variables. is it is good or not sir. kindly give valuable suggestion.

Loading...

Reply

Jim Frost says
September 13, 2020 at 10:13 pm

Hi Pavithra,

You need to check the VIFs as I show in this post. See what those values are to determine whether you have problematic levels of multicollinearity. If you do, you might need to use something like Ridge or LASSO regression.

Loading...

Reply

Angela Kitali says
August 21, 2020 at 11:46 am

0

Hello Elif. You definitely can. The approach is just a little bit different from the usual way we compute VIF. Instead of using VIF, a generalized variance inflation factor (GVIF), proposed by Fox and Monette (1992), is used to check for correlation when all variables are categorical. Refer to this document for more information {Fox, J., & Monette, G. (1992). “Generalized collinearity diagnostics”. Journal of the American Statistical Association, 87(417), 178-183.}.

Loading...

Reply

Sauddi Syamier says
June 17, 2020 at 4:52 am

Hi Jim,

Thank you for a very insightful post!

I am regressing between the relationship of net profit(dependent variable) with revenue, cost of sales and expenditure (independent variables). My analysis shows R-square =1, all the coefficients are 1 and p-values are all 0 which is very weird.

I found out that multicollinearity exists between all of the independent variables as regressing them with each other gives me a high R-square figure (>0.95) and significant p-values so I think this is a case of multicollinearity in the data set.

I noticed that there is not a lot of regression out there on accounting data/equation so im not sure if this type of regression is possible or not. It would be best if you could share with me your professional opinion regarding this matter.

Thanks,
Sauddi

Loading...

Reply

Elif says
June 16, 2020 at 10:24 am

Hello, firstly, thank you very much for this very helpful post. I have categorical independent variables, can i use vif to determine multicolinearity?

Loading...

Reply

Jim Frost says

June 16, 2020 at 3:35 pm

0

Hi Elif,

Unless your software uses a specialized VIF that I'm unaware of, VIFs are only for continuous variables. You might need to use something like Cramer's V to assess the strength of the relationship between categorical variables.

Loading...

Reply

David de Visser says

June 6, 2020 at 4:09 am

Hi Jim. Thanks a lot in helping me out here! Your articles have been very helpful and I especially appreciate the fast responses. Thanks!

Loading...

Reply

David de Visser says
June 5, 2020 at 3:14 am

0

Hi Jim,

Thanks a lot for the fast response. The pairwise correlation is 0.3629 between X1 and X4, which is significant in my data-set with 136 observations.
The VIFs are as follows:
X1: 1.27
X2: 1.08
X3: 1.82
X4: 2.03

All of them are well below five, so I would think that there is no problem of multicollinearity. Or is there still another way through which the significant correlation between X1 and X4 could bias my results?

The explanation on statistically significant and practically meaningful is also interesting and helpful. Can I see the VIF as 'a sort of' test of whether a significant correlation is practically meaningful and causes problems?

Thanks again! David

Loading...

Reply

Jim Frost says
June 5, 2020 at 7:57 pm

0

Hi David,

Those values are not problematic at all. The VIFs are all good. The correlation is fairly weak. It's statistically significant, so your sample data support the hypothesis that the variables are correlated in the population. However, that doesn't mean it's a strong correlation. So, no need to worry about multicollinearity!

VIFs are actually a more sophisticated way of checking for correlation than pairwise correlations. Imagine that your X1 has a small correlation with X2, X3, and X4. Each of those pairwise correlations are moderate and no big deal on their own. However, now imagine all those correlations result in those three variables (X1, X2, X3) collectively explaining most of the variability in X1. Basically, you can use X1 – X3 to predict X4. VIFs can check for that scenario. VIFs regress each X on the remaining set of Xs. VIFs = 1/1-R^2. So, as R-squared of those regressions increase, so does the VIF.

Now, R-squared measures the strength of that relationship. So, ultimately, VIFs measure the strength of the relationships between each predictor and all the remaining predictors. If you plug in R-squared values into the VIF formula, you'll find that an R-squared of 80% produces a VIF of 5, which is where problems start. Your variable X4 has a VIF of 2.03, which means that if you regress X1, X2, and X3 on X4, that model will have an R-squared of about 50%. You can try that with your variables if you want. If you have more predictors than those listed, you'd need to include them as well.

If you have only two predictors X1 and X2 in your model and they have a Pearson's correlation of 0.9, that equates to an R-squared of 81%, which is again right around where you get VIFs of 5. But, VIFs are better than correlations because they incorporate all the relationships between the predictors, not just one pair at a time.

So, yes, that's a long way of saying that VIFs measure the practically meaningful strength of the relationships between all the predictors.

Loading...

Reply

David de Visser says
June 4, 2020 at 3:50 am

Hi, first of all thank you for your explanation on multicollinearity. At the moment I am struggling with a regression analysis. I am trying to find the effect of X1 on Y and based on literature I should include control variables X2, X3 and X4. However, when looking at the pairwise correlations between all my independent variables I discovered a statistically significant correlation between X1 and X4. However, when looking at the VIF of the regression model I do not see any problem. So, I would say there is no problem of multicollinearity. I am wondering whether the significant pair wise correlation between X1 and X4 still causes a problem in my regression and whether I should exclude X4 based on the correlation?

Loading...

Reply

0

Jim Frost says
June 5, 2020 at 12:33 am

Hi David,

How high is the correlation between X1 and X4? And what are the VIFs for all your variables?

One thing to keep in mind is the difference between statistically significant and practically meaningful. In this context, practically meaningful is a correlation high enough to cause problematic levels of multicollinearity. However, just because a test result (correlation in this case) is statistically significant, it doesn't necessarily mean it is practically significant.

If your VIFs are low enough, then my guess is that the correlation is statistically significant (you can be reasonably sure it exists in the population) but it's not large enough to cause problems with multicollinearity. The VIFs for each IV are the key statistics to assess here.

In a nutshell, it sounds like multicollinearity is NOT a problem when you include all your x variables.

Loading...

Reply

0

Julien Mainguy says
May 29, 2020 at 11:59 am

Hi Jim, thanks for these precisions. I will certainly look into the LASSO and Ridge regression.

In response to your question, I did use previously the approach I've proposed for a published MS of my PhD thesis (see Mainguy et al. 2008 Age- and state-dependent reproductive effort in male mountain goats, Oreamnos americanus. Behavioral Ecology and Sociobiology, 62:935-943) in which the proportion of time spent in reproductive behaviours by males was investigated according to age and age-specific body mass within a same model – with age-specific body mass still explaining a significant part of the variation observed in the dependent variable.

And thanks again for your blog, I've found very interesting and useful information in it.

Regards.

Loading...

Reply

Jim Frost says
May 29, 2020 at 4:11 pm

0

Hi Julien,

Thanks for the extra information. I will have to look into that approach. Honestly, I don't fully understand it. Unless the residuals from the other model are not random, which is usually a bad thing, I'm not quite sure how random residuals would explain variation in a DV. However, I'm glad it helped you out with your thesis!

I can say that using LASSO or Ridge are more common approaches.

Loading...

Reply

Julien Mainguy says
May 27, 2020 at 12:43 pm

Just a quick comment first: congratulations Jim for this very clear and well-written blog about multicollinearity. Well done. Seriously.

And, second, if I may suggest an additional potential solution to deal with multicollinearity: in a model where Y ~ X1 + X2, but X1 and X2 are quite correlated such

that one or both VIF values are > 5 , then one could decide to look at X2 ~ X1, found the predictive equation between these variables and used the residuals of X2 on X1 in the model “Y ~ X1 + (X2 corrected for X1). As an example, if the probability to lay a clutch (or not) for a female bird is examined according to her AGE and BODYMASS, two “independent” variables that are in fact correlated, then one could rather use as independent variables AGE (again) and AGE-CORRECTED BODYMASS in the analysis, as the residuals of BODYMASS against AGE should no longer be correlated in a problematic manner. This sould allow to test both variables in a same model without multicollinearity issues. Would you agree with this potential solution?

Loading...

Reply

Jim Frost says
May 27, 2020 at 5:05 pm

Hi Julien,

I’m not familiar with that technique. I’d have to think about it. Offhand, I’m not sure that I’d refer to it as age corrected body mass using your example. I agree the residuals would not be correlated but they’re just the random error left over from that model. The random error from one model shouldn’t explain anything in another model. If it does, the residuals aren’t random. But, I’d have to think it through. Have you tried this method? If so, how did it work out?

I’d recommend using LASSO or Ridge regression as I describe in my post about choosing the correct type of regression analysis. I hope to wrote a blog post

particular about those techniques at some point!

Loading...

Reply

0

Collin M says
May 22, 2020 at 7:08 am

thanks Jim for ur help.
Now should I take it that multicollinearity can not exist significantly if interaction is absent?

And if so, can I have a decision about the existence of multicollinearity based on whether the interactions are significant in the ANOVA rather than using VIFs.

In other words, can I conclude on existence of multicollinearity based on the P-value of the interaction terms in the ANOVA?

thank you

Loading...

Reply

Jim Frost says

May 22, 2020 at 1:34 pm

0

Hi Collin,

Multicollinearity can definitely exist when there are no interactions. If the IVs are correlated, then multicollinearity exists. No interaction effect is required. However, interaction terms create multicollinearity unless you center the variables. Additionally, if you include an interaction term and the p-value is not significant, the interaction term still creates multicollinearity. It's the fact that the interaction term uses the product of two or more variables that causes that type of multicollinearity rather than the significance of the term.

Just keep in mind that interaction terms are just one possible source of multicollinearity, not the only source.

Don't use p-values to assess multicollinearity. P-values don't assess multicollinearity so it's not the right tool to use. Use VIFs as I show in this post.

Loading...

Reply

Collin M says

May 21, 2020 at 12:34 pm

0

hello Jim, thanks 4 ur great efforts in simplifying statistics and regression analysis.
But am confused about the difference between multicollinearity and interaction of the independent variables, do they mean the same thing? And can they occur at the same time in multiple independent variables.
I will be glad if u help me out on that.
thank you

Loading...

Reply

Jim Frost says
May 21, 2020 at 2:35 pm

Hi Collin,

That's a great question. I've heard similar confusions before about this exact question.

Multicollinearity and interactions are different things.

Multicollinearity involves correlations between independent variables.

Interactions involve relationships between IVs and a DV. Specifically, an interaction effect exists when the relationship between IV1 and the DV changes based on the

value of IV2.

So, each concept refers to a different set of relationships. Within the IVs (multicollinearity) or relationships between IVs and the DV that change based on another IV.

And, yes, they can occur at the same time. In fact, when you include an interaction term, it usually creates multicollinearity. The interaction term itself is the product of at least two IVs. Consequently, that term is correlated with those IVs and thereby creates multicollinearity. However, as I write in this post, centering the variables helps reduce that type of multicollinearity.

I hope that helps!

Loading...

Reply

Prabina says
May 21, 2020 at 8:53 am

Hi Jim,
Is there any condition at which regression coefficient is statistically significant in univariate logistic model, but not correlated with each other. By this what I mean is can I use one of the categorical IV as a dependent variable to check for collinearity. I would

assume that if the regression coefficient is significant , then they are correlated.
Thank you,
Prabina

Loading...

Reply

Jim Frost says
May 21, 2020 at 10:48 am

Hi Prabina,

When a regression coefficient is statistically significant, and assuming there aren't other problems with the model that is biasing the coefficient, that indicates there is a relationship between the IV and DV. So, there is some type of correlation between the two. In your case, it's not the more familiar Pearson's correlation because you're not dealing with two continuous variables. Instead, you've got a categorical and a binary variable. Still, there is a correlation between them. Your model is telling you that if you know the category for the IV, it gives you information about value for the binary variable.

Loading...

Reply

MAHESWAR SETHI says
April 3, 2020 at 11:40 am

0

In a model with 4 independent variables, when there is multicollinearity problem among two independent variables. Is it necessary that such two variables need to be centered to remove multicollinerity problem or centering any one variable can remove the multicolineariry problem?
Further, centering any variables requires all the independent variables to be centered even though they don't sufferer from multicolinearity problem? As it shown in the example that due to multicollinerity Fat%, Weight, and Fat*Weight are to be centered. But in addition to those variables even Activity has been centered which even don't suffers from multicollinerity.

Loading...

Reply

Jim Frost says
April 5, 2020 at 7:07 pm

Hi Maheswar,

Centering the variables only reduces multicollinearity for variables that are in polynomial terms and interaction terms. If your two independent variables are included in an interaction term, you should center both of them. If they're not in an interaction term, centering will not help.

Technically, you only need to center variables with high VIFs. However, typically, if you center one variable, you should center all the continuous IVs because it affects how you interpret the constant. When you don't center IVs, the constant represents the predicted value when all IVs equal zero. However, if you center all the IVs, the constant represents the predicted value when all the IVs are at their means. If you center only some of the IVs but not others, there's no simple interpretation for the constant. However, I always urge caution when interpreting the constant.

Loading...

Reply

gaurav sharma says
March 18, 2020 at 4:14 am

hey I am new to R, how you produced that summary containing vif and all those details. I used this code but my summary is not good and informative as yours. Neither it contains vif. Please tell you to get summary like yours. If you are using any additional package then I would like to know

"`

linear_regressor = lm(femoral_neck ~ fat_perc + weight_kg + activity + (fat_perc * weight_kg), data = df)
summary(linear_regressor)

"`

output –

“`

Call:
lm(formula = femoral_neck ~ fat_perc + weight_kg + activity +
(fat_perc * weight_kg), data = df)

Residuals:
Min 1Q Median 3Q Max
-0.178453 -0.042754 -0.006129 0.033937 0.186795

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.549e-01 1.317e-01 1.176 0.24274
fat_perc 5.571e-03 4.087e-03 1.363 0.17632
weight_kg 1.447e-02 2.852e-03 5.073 2.19e-06 ***
activity 2.238e-05 7.276e-06 3.075 0.00281 **
fat_perc:weight_kg -2.142e-04 7.394e-05 -2.898 0.00476 **
—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07051 on 87 degrees of freedom
Multiple R-squared: 0.5623, Adjusted R-squared: 0.5422
F-statistic: 27.95 on 4 and 87 DF, p-value: 6.242e-15
“`

Loading...

Reply

Akshay Ghalsasi says
February 27, 2020 at 5:52 pm

0

Hi Jim,
Thanks for the excellent and intuitive explanations. I have a question regarding doing PCA to get rid of multicollinearity. Is it always wise to do a PCA before linear regression if you have high multicollinearity? If you keep all your PCA components will you get a better fit?

I recently tried PCA for Naive Bayes classification since Naive Bayes also has no multicollinearity assumption. The Naive Bayes of the PCA transformed features was worse than the non PCA transformed features, counter to my intuition. Do you have anu intuition of why this could be the case?

Thanks,
Akshay

Loading...

Reply

Bianca says
February 15, 2020 at 10:03 am

Hi Jim,

Thank you so much for this article! It explained things a lot better than I've seen it written before so I will definitely be purchasing your book!

I am currently trying to create a model which predicts sales using past weather data. My variables include the daily rainfall, max temperature and min temperature and then I have calculated the sum of rainfall for the month and week prior to the sale, as well as the average max and min temperature for the month and week prior to the sale. I also have interaction factors between the max temperature and the rainfall.

Centering the variables reduced my sum of rainfall variables to an acceptable level – however the average max and min temperature for the week and month prior to the sale still have very high VIF. I guess this is because they are based on the mean anyway. Do you know if there is something I can do in this case? Am I missing something?

Thank you!

Loading...

Reply

Cat Alves says
February 6, 2020 at 5:16 am

Hey, Jim, I need some help with my master's dissertation.

I am using a simple regression, often used in literature that establishes a relationship between Price=B1+B2 Equity+ Net Result.

However, in reviewing the results I have the following problem:
-> If we analyse separately the equity variable(X) in relation to price(Y) it is significant for 5%, however in the model with net result it is no longer significant.
-> In the following year separately or together the equity variable is significant.

I am afraid that in the defence of the dissertation the question will be raised because in a model (year 2017) equity is not significant and in the following year it is.
Can you help me?

p.s: the correlation between variables and own capital is very high but close for the two years 0.8

Loading...

Reply

Jim Frost says
February 8, 2020 at 3:42 pm

Hi Cat,

First, just a minor point of terminology. Simple regression refers to regression models with a single independent variable. Because your model has more than one IV, we call it multiple regression. It's the same underlying analysis but you just have more than one IV!

Because your IVs are correlated, it's not surprising that the significance changes when you exclude one. That's known as omitted variable bias. Given the high

correlation, you should check the VIFs to check for severe multicollinearity. Multicollinearity can reduce the statistical power of the analylsis, and might account for the insignificant result. This post talks about VIFs.

Are you fitting separate models for each year? If so, try including year as an IV and fit the model for all years and see if that helps.

Best of luck with your dissertation!

Loading...

Reply

Ravi Prasad says
February 5, 2020 at 3:01 am

Hi Jim,

Does the assumption ‘multi-collinearity doesn’t impact the predictive power of the model even’ hold even for data sets having severe multi-collinearity among variables.

The way I see it is – Multi-collinearity certainly increases variance of the model. Before using the model built on variables having severe multi-collinearity, we should make sure the model is built on decent amount of data. Otherwise, even small changes in input data may make the model go awry.

Would like to know your views on this.

Thanks.

Loading...

Reply

Jim Frost says
February 5, 2020 at 11:32 am

Hi Ravi,

First, to clarify, that's not an assumption of the analysis, it's just how things work out. Multicollinearity (no dash) increases the variance of the coefficient estimates but it does ***not*** increase the variance of the entire model, which is why it doesn't affect the goodness-of-fit statistics.

However, your other point is valid. It is more difficult fitting the correct model in the presence of multicollinearity because the coefficient estimates and p-values are unstable. Small changes in the model can affect the coefficients and their significance greatly. It's harder to know which model is correct! A larger sample size can help but the problem won't go away completely. I talk about this aspect of multicollinearity in my post about stepwise and best subsets regression because it affects how well those procedures can select the correct model.

Loading...

Reply

0

Shreya Gupta says
November 23, 2019 at 11:24 pm

Hi Jim,

I checked my residual vs fitted plot for the model with all the predictor variables and the plot is a parallel line to y-axis. What does a plot like that represent. Does it represent any issues with model. Could you please share some insights on this.

Thanks

Loading...

Reply

Jim Frost says
November 24, 2019 at 2:54 am

Hi Shreya,

There's no obvious problem from your description. Is the spread of values relatively consistent across the full range of fitted values? Read this post about residual plots for more information and graphs about what they should look like.

Loading...

Reply

0

Shreya Gupta says
November 20, 2019 at 10:49 pm

Thanks a lot Jim !

Loading...

Reply

Shreya Gupta says
November 19, 2019 at 1:27 am

Hi Jim,

Thanks again for an in-depth explanation. Yes, by inverted sign i mean the sign of coefficient is negative (opposite) in both the models but when i plot the graph between X2 and target variable , it shows a positive linear relationship.
My question is if i use the second model with X3 removed, does inverted sign of X2 in the final regression equation to predict target cause any problem? Can i get accurate

value of target with second model?
What should be done in that case?

This is the VIF after i removed X3 from base model.
Variance inflation factor for X1: 34.47
Variance inflation factor for X2: 2.66
Variance inflation factor for x4: 70.09

Loading...

Reply

Jim Frost says
November 19, 2019 at 11:10 am

Hi Shreya,

There's multiple considerations. So, let's go through them.

Negative coefficient: To start, let's refer to positive and negative coefficients, which is based on their sign. "Inverted coefficient" is non-standard terminology and using it will just confuse people. There's nothing inherently wrong with negative coefficients in general. What you need to do is determine whether the signs and magnitudes of the coefficients are consistent with theory. I write about this in my post about choosing the correct model.

Graph vs. Negative coefficient: What you're seeing is a form of omitted variable bias. When you fit the model, the model accounts, or controls for, all other

variables before estimating the coefficient for X2. However, when you graph just X2 by Y, those other variables are uncontrolled. When you exclude important variables and you have a triangle of correlation between the excluded predictors, included predictors, and Y, you get omitted variable bias. Usually you talk about this bias in terms of regression coefficients. But, in this case, you're seeing it in your graph. Given the multicollinearity, it's not surprising you're experiencing it. For more information, read my post about omitted variable bias.

Multicollinearity and coefficient signs and significance: In general, multicollinearity can cause coefficients to have the opposite sign than what they should have and tend to reduce significance. The fact that your three variables are still statistically significant despite the multicollinearity is notable. Even after removing X3, your model has extreme multicollinearity, as indicated by VIFs greater than 5. X2 is not severely affected by multicollinearity. That's interesting because it has the unexpectedly negative coefficient sign. I can't tell you if X2's coefficient sign should be negative or not, but it's not multicollinearity that's causing it.

Multicollinearity and predictions: Yes, to make predictions, the model will use the all the coefficients, including the negative X2 coefficient. As I mention in this post, multicollinearity can flip coefficient signs and obscure statistical significance, but it all works out OK in terms of predictions. However, multicollinearity doesn't severely affect X2, as indicated by its low VIF. So, that coefficient is something you should look into.

Here's your to do list:

Check those residual plots for obvious problems.
Figure out whether theory suggests you'd expect a negative coefficient for X2. Maybe it's not even a problem?

Consider using LASSO or Ridge regression because they can handle multicollinearity.
Because you're heavily using regression analysis, consider getting my ebook about regression analysis. I think it'll be helpful!

Loading...

Reply

Shreya Gupta says
November 18, 2019 at 11:08 pm

Hi Jim,

Thanks for your reply. I removed the predictor with the highest VIF which is 256.46(X3).
Variance inflation factor for X1: 43.01
Variance inflation factor for X2: 2.66
Variance inflation factor for X3: 256.46
Variance inflation factor for X4: 140.84

Initially the adjusted r squared value was 0.901. After removing this predictor(X3) it becomes 0.898. Same is the case with RMSE, the difference is very little. In both the models, coefficient of predictor variable(X2) is inverted. Out of these two models, I don't understand which one is better, the one which suffers from multicollinearity(base model) or the one in which i have removed the predictor with the highest VIF.

Which model should be considered, though both models have negative coefficient of X2 when it should be positive. Does inverted sign of X2 create a problem in calculating Y(target variable). Also , its a very small data set , only 82 rows.

Thanks
Shreya

Loading...

Reply

Jim Frost says
November 18, 2019 at 11:48 pm

Hi Shreya,

Thanks for the extra information. Yes, that is severe multicollinearity. You need to also check the VIFs after you remove X3 to see if that resolved the problem. Removing the predictor with the highest VIF won't necessarily remove all the multicollinearity.

X2 actually isn't affected by multicollinearity. Any unusualness you're seeing with that variable isn't due to multicollinearity. When you say the coefficient is inverted, do you mean that it has the opposite sign that you'd expect theoretically? If that's the case, it's not due to multicollinearity, but there are other potential reasons. The "inverted" coefficient is only a problem if it contradicts theory. If so, you need to figure out what is happening. If it's consistent with theory, no problem! Did you check the residual plots for any obvious problems?

In terms of explanatory power, the model without X3 loses very little predictive power. The decrease in adjust R-squared is essentially zero. It is hard to say definitively which model is better. In fact, it's possible that even the smaller model might still have too much multicollinearity. So, perhaps neither is the best. And, as I write in the post, if you're mainly interested in predicting Y, multicollinearity isn't necessarily a problem.

In general, I'd favor the smaller model because X3 was not significant to begin with and it was causing multicollinearity. Removing it has virtually no impact on the goodness-of-fit too. So, I'd lean in that direction. when you can fit a simpler model that explains the outcome virtually as well, it's often the better model. Simplicity is a good goal. But, I don't know what degree of multicollinearity the smaller model has nor fully understand what you mean by "inverted" coefficient and, therefore, don't know if that is problematic.

Loading...

Reply

Shreya Gupta says
November 18, 2019 at 6:23 am

Hello Jim.

Thanks for a great explanation.
I have built one OLS regression model with four features.It suffers from multicollinearity and one feature's p-value is greater than 0.05. when I remove the feature which is insignificant , i get another model where all the features are significant. But the adjusted r squared value is higher for the first model and RMSE(root mean square error) is also lower for that model.
How do I now decide which model is better, the one with multicollinearity or the one without.

Thanks

Loading...

Reply

Jim Frost says
November 18, 2019 at 3:52 pm

Hi Shreya,

This is a tough issue. Multicollinearity can make things murky! It's entirely possible that the variable with the the insignificant p-value should actually be significant. A classic hallmark of multicollinearity is that it makes variables appear to be insignificant even when they're not. However, the model fit statistics, such as adjusted R-squared and RMSE are not affected by multicollinearity. That *might* explain why removing a variable that appears to be insignificant causes a large reduction in the goodness-of-fit statistics.

There's also a number specifics that matter. How much multicollinearity do you have? Look at the VIFs with all the variables and then after you remove the one. How much do the goodness-of-fit statistics change? If it's a relatively small amount, then it's probably not a meaningful difference. If it's larger, it might mean something, as I discuss above. If you're using the model to make predictions and don't need to understand the specific role of each variable, you probably want to leave them all in.

Unfortunately, in OLS, the variety of solutions for multicollinearity all have some type of drawback and it can be difficult to settle on the best model. Consider using LASSO or Ridge regression. Unfortunately, I don't as yet have a blog post about using them. However, I describe them in my post about choosing the correct type of regression analysis. Those analyses can handle multicollinearity. You can include all your variables and obtain relatively precise estimates with the tradeoff being the introduction of a small amount of bias.

Loading...

Reply

Talha Haroon Khan says
November 15, 2019 at 12:38 am

Dependent variable is a utility function which will be obtained from these coefiicients. From that utility function traffic model will be assessed.

Loading...

Reply

0

Ryan says
November 14, 2019 at 4:36 pm

Multicollinearity does cause inflated variance in the model which makes it unstable. Although it does not affect a model's ability to make predictions, if we introduce new data from a new sample, would it still make accurate predictions?

Loading...

Reply

Jim Frost says
November 14, 2019 at 4:42 pm

Hi Ryan, that's correct. Multicollinearity doesn't affect the model's ability to make predictions. If a model with and without multicollinearity have the same prediction precision, they should make equally good predictions. Use prediction intervals to assess that type of precision.

Loading...

Reply

0

Talha Haroon Khan says
November 12, 2019 at 12:11 am

Thanks a lot Sir!

Loading...

Reply

Talha Haroon Khan says
November 10, 2019 at 5:33 am

Respected Sir,
I am performing Regression Analysis on a Traffic Model.
I have 5 independent variables, i.e. Travel Time (TT), Travel Cost (TC), age, education and income level.
First two are explanatory variables and last three are just used as controllers.
The beta value for TT is opposite to the real scenario (both its sign and value), similarly, its p value is greater than 0.05 and t less than 1.96.
How to adjust this beta for TT so that all these checks are okay?
Thanks in anticipation.

Regards,
Engr. Talha Haroon Khan

Loading...

Reply

0

Jim Frost says
November 10, 2019 at 6:50 pm

Hi Talha,

What's the dependent variable? You didn't say.

There's a variety of reasons for why you can obtain unexpected coefficient signs and p-values. To figure out the specific reason for your data will take some investigation. As I discuss in this post, multicollinearity can cause both unexpected signs and p-values. You might also be misspecifying the model, overfitting, not fitting curvature, and so on.

I'd start by checking the residual plots and the least squares assumptions. They can provide clues.

I have an entire chapter in my ebook about regression analysis dedicated to identifying and solving problems with your model. You should consider getting it.

Best of luck with your analysis!

Loading...

Reply

0

Bharat says
October 4, 2019 at 2:18 am

is it possible that the differences between groups are due to the different sample sizes ?

Loading...

Reply

Jim Frost says
October 5, 2019 at 4:18 pm

Hi Bharat,

As you increase the sample size, you'd expect the range of the data displayed in the graph to increase. However, for this example, this differences in sample size don't account for the observed differences.

Loading...

Reply

0

Brenda Barr says
September 18, 2019 at 10:22 pm

Hi Mr. Frost, I am a PhD candidate who is now working on the assumptions for multiple regression analysis. I have purchased your book and have read this article to hopefully help me with testing multicollinearity in the data. I have a problem that I hope you can at least help me shed light on. I chose to conduct a multiple regression analysis for my study in which I have 6 independent variables and one dependent variable. In testing the assumption of multicollinearity, the following are the numbers for Variance and for VIF. My concern are the VIF statistics for Avoidance, Distraction and Social Diversion Coping which appear to be very high. I'm lost on how to proceed.

Independent variables Tolerance VIF
Gender .935 1.070
Task Coping .434 2.304
Emotion Coping .717 1.394
Avoidance Coping .018 56.095
Distraction Coping .049 20.233
Social Diversion .050 19.904

Thanks for any insight you can offer.

Brenda Barr

Loading...

Reply

0

Leila says
September 3, 2019 at 1:21 am

Hi Jim,

This article is such an easy read for those of us who are statistically- challenged. Could I please ask for some advice?
I have a model with 4 IV's and one DV. I'm looking at the model with two different groups (i.e., I'm comparing the fit of the model for the 2 groups). The second group has much fewer participants (50) than the first (94).
I've found high correlations between the IV's for the second group (.65 to .82) but not for the first group. The VIF's, however, are between 2 and 4.3 for the second group. I'm unsure whether to interpret this as high multicollinearity? If so, would it be reasonable for me to simply make a composite score by adding the four IV's and seeing how this score correlates with the DV?
Also, is it possible that the differences between groups are due to the different sample sizes (94 for group 1 and 50 for group 2)?

Loading...

Reply

Miss Nadine Osei says
August 30, 2019 at 1:25 am

0

Hello Jim, thank you for the clarity. but i have question, it may seem damn but I want clarity on it please because I do not know which regression method to use.
When conducting my Primary research, I asked three different questions (based on my item of measure) in relation to dependent variable (purchase intention). now that i want to run a regression test, do you recommend i it three times since there are three questions for the dependent variable, or there is a way to merge.
Thanks

Loading...

Reply

Jim Frost says
August 30, 2019 at 2:47 pm

Hi, there are several different ways to answer this question.

One option is to do as you mention, create three models, one for each dependent variable.

You can linearly combine the three DVs (e.g., sum or average), if that makes sense for your research question. Then fit one model for that new DV. Going this route depends on using your subject-area knowledge and determining that a summed or average DV is somehow better than separate DVs.

You can also use a multivariate approach. For example, I write about the benefits of using multivariate ANOVA (MANOVA) when you have correlated DVs. You can extend this approach to including covariates if needed.

There might well be other options but those are ones that come to mind. To figure out the best option, you'll need to use your subject-area knowledge and research goals. If possible, see how other researchers have handled research in your area.

I hope this helps!

Loading...

Reply

suvichar in hindi says
August 21, 2019 at 6:57 am

Thanks a lot for sharing.I discovered a great deal of intriguing data here. A great post, extremely grateful and confident that you will compose a lot more posts like this one.

Loading...

Reply

Jim Frost says

August 21, 2019 at 11:51 am

0

Hi Suvichar,

I'm glad you found my website to be helpful! I'll definitely be writing a lot more! ☺

Loading...

Reply

Jerry says

August 12, 2019 at 9:41 am

Thank you so much. I am grateful sir.

Loading...

Reply

Engr.Waleed Khan says

July 29, 2019 at 11:32 am

Thank you very much. You explained it in very clear and easy words.

Loading...

Reply

0

Jerry says
July 24, 2019 at 8:12 am

Thank you sir for your wonderful contribution to Statistics.
My name is Jerry, I am working on Multicollinearity using Ridge Regression and R package for the analysis.
I am using the Longley Built in economic data already installed in R.
I detected severe presence of Multicollinearity by the help of VIF which were all greater than 10 (Rule of thumb).
Thereafter, I tried to use the ridge regression technique to reduce and to solve the Multicollinearity problem found.
Now, as learnt, after solving the problem of Multicollinearity in a regression model, the ridge regression coefficients are expected to be reduced and less than the OLS coefficients.
So, please I would like to know because, my ridge regression coefficients are only less than my OLS when observing the absolute coefficient values of the OLS and the Ridge Regression. So sir, is it right to make absolute comparison of our coefficients?
Thanks for your educative response.

Loading...

Reply

0

Jim Frost says
July 31, 2019 at 11:35 am

Hi Jerry, sorry for the delay in replying but sometimes life gets busy!

Using Ridge Regression is a great choice for multicollinearity. However, it doesn't allow you to compare the absolute values of your coefficients–at least not in most cases. For more information on what you can do, read my post about identifying the most important variables in a regression mode.

Loading...

Reply

ali says
July 13, 2019 at 6:37 am

I have one questions .. if i have mediating variables am i should included at the test of multicollinearity . or only independent variab;e

Loading...

Reply

0

Jim Frost says
July 15, 2019 at 11:07 am

Hi Ali,

In this post, I show how it's important to assess multicollinearity when you have interaction effects (a.k.a. mediating variables). So, yes, definitely check for multicollinearity for that type of variable. I can guarantee you that your model will almost definitely have high multicollinearity when it contains interaction effects or polynomials.

Fortunately, there's a simple solution–standardized your continuous IVs. I show an example of that in this post and also write about it in a post about standardizing your variables.

Loading...

Reply

Raghad says
May 27, 2019 at 6:09 pm

hi sir
thanks for the important and valuable information.
I have a question please:
I have a time series data with 25 observations. the original model has a binary dependent variable and six independent variables (not binary). There is a high collinearity between variables, so I used the PCA technic to solve the problem.
Now I have a problem at the final step in the model, which is: after I got the PCA matrix and I have to run the new regression using pc1 , pc2 , and pc3 that represent 88% of the variables. shall I run the model just with pc's and deduct the othet three independent variable? because when I regret the model with 3 pc's and 3 original variables I still have multicollinearity problem.

with my respect

Loading...

Reply

philoinme says
April 28, 2019 at 4:20 am

There is this point mentioned in the third bullet under Do I have to fix multicollinearity? – "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."

I am afraid I might be a bit off-topic but I want to know it from a data science perspective.

Will this be a disadvantage when predictions are my main goal and I don't interpret coefficients properly or care about the statistical significance of different independent variables?

Loading...

Reply

Anastasia says
March 24, 2019 at 11:30 pm

Hi Jim!

Thank you for this post, it is definitely very helpful. I see a few people asked about detecting collinearity for categorical variables, and the suggested solution is the chi-square test. My biggest problem is that when I calculated VIFs for my linear mixed-effects model with up to 4-way interactions, some interactions of categorical variables correspond to ginormous VIFs. The interaction term seems to be collinear with one of the main effect terms included into this interaction. Those are essential interactions for my hypothesis, so I'm not sure what to do about it (sort of gave up on this for now and switched to random forests…). Am I right evaluating the collinearity in my model using the VIF? Or can/should I use the chi-square test for that?

Thanks a lot,
Anastasia

Loading...

Reply

Jim Frost says
March 24, 2019 at 11:48 pm

Hi Anastasia,

Because you're referring to VIFs, I'll assume you have continuous interactions. (VIFs are calculated only for continuous variables. Chi-squared is only for categorical variables.) Whenever you include interaction terms, you'll produce high VIFs because the main effects that comprise the interaction terms are, of course, highly correlated with the interaction terms. Fortunately, there is a very simple solution for that, which I illustrate in this blog post. Simply center all of your continuous independent variables and fit the model using the centered variables. That should cut down those VIFs by quite a bit. Often centering will get the VIFs down to manageable levels. So, reread that section of this post!

I hardly ever see models with three-way interactions, much less four-way interactions. Be sure that they're actually improving the model. If the improvement is small, consider removing them. Higher-ordered interaction terms like those are very difficult to interpret. Of course, your model might be much better with those terms in it. But, at least check because it is fairly unusual.

I hope this helps!

Loading...

Reply

0

JP says
March 11, 2019 at 11:54 am

Hi Jim, Thanks for the great article, I am working on multiple regression model where my VIF values are less than 2 for all the variables. From the model one variable changes sign of the coefficient (from the theory) even when the VIF value is less than 2 , I checked that one variable with DV it is showing the right coefficient sign and it is not statistically significant on its own . Should i look into interaction with that variable.

Loading...

Reply

Jim Frost says
March 11, 2019 at 3:16 pm

Hi JP,

With VIFs that small, it's unlikely that multicollinearity is flipping the signs of the coefficients.

Several possibilities come to mind that can flip signs. The two most common are issues that I've written about: Omitted variable bias and incorrectly modeling curvature that is present. For a quick check on the curvature issue, check the residual plots! And, yes, it's possible that not including an interaction term that should be in the model can cause this problem. All of these issues are forms of model specification error (i.e., your model is incorrect and giving you biased results). I'd check those things first.

I'm assuming that the independent variable that has the incorrect sign is statistically significant?

Loading...

Reply

Zara says
March 7, 2019 at 3:45 pm

Hello! Im doing a regression analysis where I have 2 measurements of the same concept. Since the measurements are very correlated I was thinking of creating a composite variable by averaging their z-scores. Is this a good step and would the interpretation of the regression change if I'm using standardized scores?

Loading...

Reply

0

Jim Frost says
March 7, 2019 at 4:15 pm

Hi Zara,

The approach you describe is definitely an acceptable one. First, I'd fit the model with both measures and check the VIFs as I describe in this post. Be sure that's actually a problem. The two variables have to be very highly correlated to cause problems.

If it is a problem, I'd also compare the goodness-of-fit for models with both measurements versus models with just one of the measurements. If they are very highly correlated, you might not lose much explanatory power by simply removing one because they're supply highly redundant information to begin with. If the change in goodness-of-fit is small, you can consider including only one measure. This approach gives you the ability to estimate the effect of one measurement.

However, if you do decide to combine the measurements as you describe, it does change the interpretation. You're standardizing the measurements. For standardized data, the coefficients represent the mean change in the dependent variable given a change of one standard deviation in the independent variable. That's still a fairly intuitive interpretation. However, you'd be averaging the z-scores, so it's not quite that straightforward. Instead, the interpretation will be the mean change in the DV given the average change of one standard deviations across those

two measurements. You'll lose the ability to link the change in a specific IV to the DV. However, it's possible you'll gain more explanatory power. Before settling on this approach, I'd check to be sure that it actually improves the fit of the model.

It's possibly a good approaching depending on whether the VIFs are problematic, the loss of explanatory power by just using one measurement, and the gains in explanatory power by averaging the two.

Best of luck with your analysis!

Loading...

Reply

Rui Fang says
March 6, 2019 at 11:45 pm

Hi Jim,

Can I use VIF to test for multicollinearity between caterogical independent variables?

Loading...

Reply

Jim Frost says
March 7, 2019 at 9:45 am

Hi Rui,

VIFs only work for continuous IVs. I think you'd need to use chi-squared test of independence with categorical variables. It is harder to determine with categorical variables.

Loading...

Reply

Nhat T.Tran says
February 28, 2019 at 7:24 am

Hi Jim,

Thank you for your reply. If it would not take up too much of your time, I would like to ask you more one question.

As you mentioned in the example, using standardized variables does not influence the predictions, so my primary goal is to investigate relationships of independent variables with a dependent variable.

In your example, after using the centered independent variables, the sign for %Fat has changed from positive to negative. Using subject-area knowledge, if %Fat has a negative relationship with the bone mineral density of the femoral neck, we may pick the standardized solution. On the other hand, if %Fat has a positive relationship with the bone mineral density of the femoral neck, we may choose the uncoded model.

I wonder whether the above solutions are appropriate or not. Therefore, I would appreciate any assistance or comments you could give me. Thank you again for your time and consideration.

Best regards,
Nhat Tran.

Loading...

Reply

Jim Frost says
February 28, 2019 at 11:27 am

Hi Nhat,

Unfortunately, because you're interested in understanding the relationships between the variables, this is not a case where you can choose between these two models based on theory. A model with excessive multicollinearity is one that has a problem that specifically obscures the relationships between the variables. Consequently, you don't want to choose the model with multicollinearity because it

happens to agree with your theoretical notions. That's probably just a chance agreement!

What you really need to do is find a method that both resolves the multicollinearity and estimates relationships that match theory (or at least you come up with an explanation for why it does not). If I wasn't able to use centering to reduce the multicollinearity, I'd probably need to use something like Ridge or LASSO regression to accomplish this task.

If the model with acceptable multicollinearity produces estimates that don't match theory, consider the possibility that you're specifying the incorrect model. That would be the next issue I'd look into. But, don't go with the model that has problematic multicollinearity just because it happens to agree with your expectations.

Loading...

Reply

Nhat T.Tran says
February 23, 2019 at 7:05 am

Hi Jim,
Thank you so much for creating a great blog for statistics.

In the ANOVA result of your example, you used the adjusted sums of squares (Adj SS). I also run the same test for your data with Minitab using sequential sums of squares (Seq SS). However, two results are different as follows.

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

Analysis of Variance

Source DF Seq SS Seq MS F-Value P-Value
Regression 4 0.55578 0.138946 27.95 0.000
%Fat 1 0.20514 0.205137 41.26 0.000
Weight kg 1 0.24506 0.245059 49.29 0.000
Activity 1 0.06384 0.063843 12.84 0.001
%Fat*Weight kg 1 0.04175 0.041745 8.40 0.005
Error 87 0.43256 0.004972
Total 91 0.98834

The ANOVA results using Seq SS shows a statistically significant effect of %Fat on the Femoral Neck (P value = 0.000), while The ANOVA results using Adj SS indicates there is not a statistically significant effect of %Fat on the Femoral Neck (P value = 0.176).

Therefore, I wonder that when will we use “sequential sums of squares” and when will we use “adjusted sums of squares”.

If you could answer all two of my questions, I would be most grateful.
Best regards,
Nhat Tran.

Loading...

Reply

Jim Frost says
February 24, 2019 at 6:56 pm

Hi Nhat,

Quick definitions first.

Adjusted sums of squares: Calculates the reduction in the error sum of squares for each variable based on a model that already includes all of the other variables in the model. The procedure adds each IV to a model that already contains all of the other IVs, and determines how much variance it accounts for.
Sequential sums of squares: Calculates this reduction by entering the variables in the specific order that they are listed. The procedure enters the first variable first, then the second variable, third, and so on.

The analysis uses these sums of squares to calculate F-values and t-values, which in turn determines the p-values. So, it’s not surprising that changing the sums of

squares affects the p-values.

The standard choice you almost always want to use is the Adjusted Sums of Squares (Type III). Using this method, the model can determine the unique portion of variance that each variable explains because it calculates the variance reduction for each variable when it is entered last. This type of SS is used for at least 99% of the regression models! Basically, use this type unless you know of very strong reasons to the sequential sums of squares.

I don't have a strong case for using sequential sums of squares. You'd need really strong theoretical reasons for why the variables need to be entered into the model in a specific order. This option is almost never used that I'm aware of.

I hope this helps!

Loading...

Reply

akshay thakar says
February 2, 2019 at 8:11 am

Thanks for such a quick response !! The chi square test for Independence can involve only 2 categorical variables at a time, so should I take multiple pairs one by one to

check for multi collinearity…?? Or is there any way to do the chi square test for multiple variables..??

Loading...

Reply

Jim Frost says
February 4, 2019 at 2:57 pm

Hi again Akshay,

Yes! You can definitely use additional variables. In chi-squared, they're referred to as layer variables. Although, it can get a bit unwieldy when you have many, I'd try that approach. I can help show if you potentially have a problem and where to look for it!

Loading...

Reply

akshay thakar says
February 1, 2019 at 10:36 am

Hi Jim,
Your blog is amazing !!
I am trying to run a regression analysis however I am facing the issue of multi-collinearity between categorical variables , are there any tests to identify multi-collinearity between categorical variables ??

Loading...

Reply

Jim Frost says
February 1, 2019 at 11:53 am

Hi Akshay, thanks for the kind words!

The chi-squared test of independence would be a good way to detect correlations between categorical variables. I cover this method in this blog post: Chi-squared Test of Independence and an Example.

Loading...

Reply

Lola says
January 29, 2019 at 1:47 pm

0

Thank you! Thank you!

Loading...

Reply

Lola says
January 28, 2019 at 5:26 am

Thanks a bunch Jim!

Is that (pairwise correlation) the same as producing a correlation matrix of all the independent variables from the regression model?

I'd also like to know if and how you take classes . I'd really want to hone my statistical 'skills'.

Loading...

Reply

Jim Frost says
January 28, 2019 at 9:30 am

0

Hi Lola, you're very welcome!

You could certainly do a matrix of all the variables as you describe. However, that might provide more correlations than you need.

Suppose X1 is the IV with the high VIF and that the others have low VIFs. You'd really just need the correlations of X1 with X2, X3, X4, and so on.

If you did a matrix, you get all the correlations with say X2 and X3, X2 and X4, etc. You might not need those. It doesn't hurt to obtain extra correlations, it's just more numbers to sort through!

Just keep in mind that only the terms in the model with high VIFs are actually affected by multicollinearity. You can have some terms with high VIFs and others with low VIFs,, and you don't need to worry about those.

Loading...

Reply

Lola says
January 26, 2019 at 11:08 am

0

Hi Jim, this page is absolutely brilliant. Thanks for this initiative.

However, I have a question on something a bit unclear regarding the VIF interpretation. In the first regression model above, it can be seen that 'Fat', 'Weight' and 'Fat*Weight' have "worrirsome" VIF values., which depict multicollinearity. And as explained above, multicollinearity involves 2 or more independent variables.
the question is – For each, Independent variable with worrisome VIF values, how does one determine which of the other IVs, it is highly correlated with?

Loading...

Reply

Jim Frost says
January 27, 2019 at 11:03 pm

Hi Lola,

That is a great question! The VIF for a specific term in model shows the total multicollinearity across all of the other terms in the model. So, you're right, seeing a high VIF indicates there is a problem but it doesn't tell which variable(s) are the primary culprits. At that point, you should probably calculate the pairwise correlations between the independent variable in question and the other IVs. The variables with the highest correlations would be the primary offenders.

Loading...

Reply

0

Ioakim Boutakidis says
January 13, 2019 at 1:20 pm

Greetings Jim…just wanted to drop a quick note and say that you have laid out some excellent content here. I landed on your cite after a student of mine came across it looking for info on multicollinearity, and so I felt I had to check it out to make sure she was getting legitimate information. I have to say I am very impressed. It looks like you are helping a lot of people do better research, and that's something you should be very proud of.

Loading...

Reply

Jim Frost says
January 13, 2019 at 9:25 pm

Hi Ioakim, thank you so much for you kind words. They mean a lot to me! I'm glad my site has been helpful for your students!

Loading...

Loading...

Reply

0

Michal says

December 31, 2018 at 2:54 am

Very much! Thank you!

Loading...

Reply

Michal says

December 30, 2018 at 9:35 am

Could it be that two highly related independent variables (r=0.834!!) yield a VIF of 3.82? Of course, it makes my life easier that I don't have to deal with the multicollinearity problem, but I don't understand how this can happen….

Loading...

Reply

Jim Frost says
December 31, 2018 at 2:30 am

0

Hi Michal,

Yes, that might be surprising but it is accurate. In fact, for the example in this blog post, the %fat and body weight variables have a correlation of 0.83, yet the VIF for a model with only those two predictor variables is just 3.2. That's very similar to your situation. When you have only a pair of correlated predictors in your model, the correlation between them has to be very high (~0.9) before it starts to cause problems.

However, when you have more than two predictors, the collective predictive power between the predictors adds up more quickly. As you increase the number of predictors, each pair can have a lower correlation, but the overall strength of those relationships accumulates. VIFs work by regressing a set of predictors on another predictor. Consequently, it's easier to get higher VIFs when you have more predictors in the model. No one predictor has to "work very hard" to produce problems.

But, when you have only two predictors, the relationship between them must be very strong to produce problems!

I hope this helps!

Loading...

Reply

0

Ben says
November 9, 2018 at 5:27 pm

What if the sole purpose of the regression is to identify the “rank” of the contributions of the independent variables? If all the varibales (including the dependent variable) are all correlated with each other, does this “drivers analysis” still hold?

Loading...

Reply

Javed Iqbal says
November 5, 2018 at 5:36 am

Q L K
11 12.2 10.1
34.6 30.2 28.2
21.9 23 24
28.2 22.3 21.3
14.7 15.7 14.3
20.2 20.8 18.4
9.7 11.5 10

22.2 25.9 24
17.3 21.5 20.3
19.5 22.4 20.5
13.6 14.4 12.2
34 29.5 29.2
35.1 26.8 25.5
10.6 12.7 10.8
18.6 19.6 19.9
22.9 25 24
27.4 25.7 23.2
16.4 18 16.2
22 18.3 19.4
27 19.7 17.2
27.1 23.7 25
15.6 21.2 20.5
13.2 23 22.1
27.3 26.3 24.3
15.4 22.6 20.8
30.6 30.5 28.9
24.4 28.6 28.1
36.1 26.7 27.9
24.8 21.7 20.7
21 18.5 17.1
10.2 13.5 11.1
20.4 13.4 11.7
14.3 15.7 15.4

This is the data ‘cobb’ from Hill-Grifith-Lim, Principles of Econometrics. The estimate of the Production function results in the following (with R-sq of 0.69 and overlall F of 33.12 with p-valu of 0.000). This is a classical case of multi collinearity. as non of the individual coefficient are significant. The sample correlation b/w log of L and log K is 0.985 and the VIFs are 35.15 for both variables. I will appreciate if you could comment on resolving the multicollinearity issue.

Variable Coefficient Std. Error t-Statistic Prob.

C -0.128673 0.546132 -0.235608 0.8153
LOG(L) 0.558992 0.816438 0.684672 0.4988
LOG(K) 0.487731 0.703872 0.692925 0.4937

Loading...

Reply

derekness says
October 29, 2018 at 4:20 am

Hi Jim,
thanks again for the useful input.
I have been playing with the model and trying to see how it responds when I do put in co-linear data in. If I put in the two strong peaks for A and B twice. It goes mad and falls over. This is good. I think it tells me that the additional data contained in the ratio of the two strong peaks is different than the data contained in the two individual peaks. This is

good as it really appears to help the model work and give me great predictive powers. It gives me confidence that all is working well.

I am also working to improve the performance of the model on how it handles independent data.
We have to make calibration mixes of A and B in the lab. We can do this very accurately and have a great machine for making a super homogenous air free mixes. The model uses these mixes with varying amounts of A and B to then be able to test mixes made in a production environment from a large metering and mixing machine. Unfortunately these do not “look” like the perfect lab made mixes and the analysis gives me a wide variations in the compositions of A and B. I think that this is not real. I therefore am now tuning the model so that it handles the production mixes better. This makes the Rsq. values for the calibration model worse, but I now get much better Rsq. values from the production mixes. This approach has worked well, and the model is now insensitive to whatever is different in the real world production mixes.

This has been a great learning process for me (and also a lot of fun), but I always am cautious as it is a bit of a “black box” and I have no idea what it is up to!

Oh and yes we also use PLS treatments for some analysis’s, and that can be really good on tricky materials. For this one I have to use the ILS method, but it seem pretty good. With the PLS work you have more insight to what is going on withe the factors and PRESS data, the ILS software doesn’t let on how it does it!

regards,

Derek.

Loading...

Reply

0

Jim Frost says
October 29, 2018 at 11:10 am

Hi Derek,

Thanks for the follow-up! I always enjoy reading how different fields use these analyses. While the methods are often the same, the contexts can be so different!

It does sounds like you have a promising model. And, the super accurate machine you have explains how you can obtain a very high R-squared when you assess the lab mixes. As you found, you'd expect the real life mixtures to have a lower R-squared.

That actually reminds me of research done in the education field. When some researchers tested a new teaching method, they did so in a very controlled lab-like setting. It worked in that setting. However, when they tried it in the field, their results were not nearly as good. They learned that because the new method had to work in the field that it had to be robust to variations you'd expect in the field. You always think of reducing the variability for experiments, but there's also the need to reflect the actual conditions. And, that sounds kind of like what you're dealing with.

It's great that it's been a fun experience for you! That's the way I see it too and what I always try to convey through my blog. That statistics can be fun by helping you discover how something works. It's the process of discovery.

Thanks for sharing!

Loading...

Reply

0

Derek Ness says
October 28, 2018 at 8:18 am

Jim, thank for that. We have a good grasp of the chemistry but the maths that we use we don't understand and have no feel for at all. The FTIR spectra is just a collection of peaks. We do one for component A and another for component B. We look for areas where we get a strong peak in one component and nothing in the other. We then use these peaks to put into the Inverse Least Squares (I think it is actually a MLR) treatment, on a sreies of mixes with varying amounts of A and B in. I do this and I get an OK model. If I then add in the ratio of the two peaks as well as the individual peaks, the model gets amazingly good, Rsq. 0.9999 from Rsq.0.9. It also gets very good at analysing my set of independent data. I am mentally struggling with whether this is real or is the treatment cheating, and it will all fall apart when I get a new set of data to analyse.
I will probably work up both approaches and use them on the next application and see if the super one keeps looking great!

Loading...

Reply

Jim Frost says
October 28, 2018 at 4:39 pm

0

Hi Derek,

I don't know enough about the subject to have an opinion on whether it is "cheating" or valid. You should think through the logic of the model and determine whether it makes sense. I wouldn't say there is anything inherently "cheating" about including the ratio. But, does that make sense from a subject matter point of view. What was your rationale for including it? And, do the coefficient estimates fit your theory?

Also, consider your sample size. Just be sure that you're not overfitting the model. It doesn't sound like you have too many predictors, but just something to consider.

Ultimately, yes, I think cross validation with a new dataset is the best way to evaluate a model. I think that's always true. But, even more so when you have questions like the ones you have! Sounds like you've got a great plan to address that! I'd be curious to hear how that goes if you don't mind sharing at a later point?

Loading...

Reply

Derek Ness says
October 27, 2018 at 6:39 am

0

Jim, I am working on FTIR spectra of mixes and we are using a ILS treatment to build an analysis model. We can select peaks to include in the model we can also use ratios of peaks. As I see it if I use two strong peaks from component A and two strong peaks from component B I have a colinearity issue as these peaks are related. Does this cause a problem?
Right now I am using a single peak from A and a single peak from B. This works OK, but if I then use a ratio of these 2 peaks as well the model looks amazing with an Rsq. of 1. This looks suspicious to me. Is colineraity effects causing this and is it thus dangerous to use the ratio and the 2 individual peaks. The cross validation for this also looks great and it appears to analyse an independent set of data really well.

Loading...

Reply

Jim Frost says
October 27, 2018 at 5:03 pm

Hi Derek,

First, let me say that I know so little about the subject area, which will limit what I can say. Statistics should always be a mix of statistical knowledge and subject-area expertise.

If your goal is to make a good prediction, then you don't need to worry about multicollinearity. It's often surprising, but multicollinearity isn't bad when it comes to the goodness-of-fit measures, such as R-squared and S. It does affect the precision of the coefficient estimates and their p-values. But, if the coefficients/p-values aren't your main concern, multicollinearity isn't necessarily a problem.

Again, I don't know the subject area, but for physical phenomenon, it's not impossible to obtain very high R-squared values if there's very little noise/random error in the process. That's something you'll have to determine using your subject-area knowledge. Maybe research what others have done and the results they obtained. But, again, the very high R-squared might not be a problem. I did write a post about reasons why your R-squared value might be too high, which you should read because it covers other potential reasons why it could be too high. But, the fact that your cross validation looks good is a great sign!

Finally, I'm also aware that analysts often use partial least squares (PLS) regression for analyzing spectra because of both a large number of predictors and multicollinearity. This form of regression is a mix between principle components analysis and least squares regression. I'm not sure if it would be helpful for your analysis, but it's an analysis to consider. Unfortunately, I don't have much firsthand experience using PLS so I don't have much insight to offer. But, if you need to consider other forms of analysis (which you might not), it's one I'd look in to.

I hope this helps. Best of luck with your analysis!

Loading...

Reply

0

Jerry Avura says
October 26, 2018 at 4:43 am

Hello sir, how can one calculate for VIF using R? Thanks in anticipation.

Loading...

Reply

Jim Frost says
October 26, 2018 at 10:41 am

Hi Jerry,

Unfortunately, I don't use R and don't know the answer. Sorry.

Loading...

Reply

Jerry Avura says
October 2, 2018 at 3:08 am

0

thank you sir. Looking forward to it

Loading...

Reply

Jerry Avura says
October 1, 2018 at 9:05 am

Thank you sir for the nice job, it was so clear and perfect. I'm Jerry, an MSc Student of Statistics. I'm currently working on Multicollinearity. Please can you throw more light on "Ridge Regression "?
Thanks in anticipation.

Loading...

Reply

Jim Frost says
October 1, 2018 at 9:20 pm

Hi Jerry, at some point I'll try to write a blog post about Ridge Regression, but I have a bit of a backlog right now! I do talk about it a bit in my post about choosing the correct type of regression analysis.

Loading...

Reply

John Komlos says
September 24, 2018 at 11:13 am

than's interesting. thank you very much. I did not know that. do you have a citation for me by any chance? thanks in advance, John

Loading...

Reply

Jim Frost says
September 24, 2018 at 11:39 am

You bet! That's a generally recognized property of multicollinearity so any linear model textbook should discuss this issue. In this post, I include a reference to my

preferred textbook for another issue. That's the one I'd recommend, but any good textbook will talk about this issue. I don't know of any articles offhand.

Loading...

Reply

0

John Komlos says
September 24, 2018 at 10:39 am

Thank you Jim, appreciate the explanation. One more question: would it be possible for the two variables to be significant in spite of multicollinearity?
Thanks.
Best regards,
John

Loading...

Reply

Jim Frost says
September 24, 2018 at 11:00 am

Hi, it's definitely possible. While multicollinearity weakens the statistical power of the analysis, it's still possible to obtain significant results–it's just more difficult. Additionally, the coefficient estimates are erratic and can swing widely depending on which variables are in the model. While you can obtain significant results, this instability makes it more difficult for you to be confident in which specific estimates are correct.

Loading...

Reply

John Komlos says
September 23, 2018 at 10:29 pm

I wonder if two multicollinear variables can be both statistically significant. one is large and negative while the other is large and positive and both significant. i have a feeling that like magnets they repel each other. is that possible?

Loading...

Reply

Jim Frost says
September 24, 2018 at 10:31 am

0

Hi John,

It's certainly possible for multicollinear variables to have opposite signs like you describe. However, there is no propensity for that situation to occur. That is to say, having different signs or the same signs are equally likely and just depends on the nature of the correlations in your data. The real issue is that you can use one independent variable to predict another. It's really the absolute magnitude of the correlation coefficient that is the issue rather than the signs themselves.

It actually gets a bit more involved than that. VIFs aren't just assessing pairs of independent variables. Instead, VIF calculations regress a set of independent variables on each independent variable. It's possible that two or more independent variables collectively explain a large proportion of the variance in another independent variable. In a VIF regression model, it's possible to have a mix of positive and negative signs!

That's probably more than you want to know! But, to your question, yes, it's possible but it's really the absolute magnitude that is the issue.

Loading...

Reply

Patrik Silva says

September 21, 2018 at 5:13 am

0

Hi Jim, Thank you, I got your point! You are helping me a lot!

Loading...

Reply

Patrik Silva says

September 20, 2018 at 10:23 am

Hi, Jim

I would like to know if, when you mentioned: Fat S, Weight S , Activity S and FatS * WeightS. The Fat S multiplied by Weight S (Fat S * Weight S) is calculated using the Fat S (Standardized) * Weight S (Standardized) or is the standardized of the two variable together by taking the (Fat * Weight) S.

I do not know if you got my point!

Thank you!

Loading...

Reply

Jim Frost says
September 20, 2018 at 3:33 pm

Hey Patrick, it's the first scenario that you list.

Loading...

Reply

Patrik Silva says
September 20, 2018 at 8:59 am

Dear Jim, you are making people love statistic! Every time I come here to read something, I am getting more love to "Stats". You explain statistics so easy, but so easy that I feel like I am reading/hearing a beautiful story.

Thank you Jim!

Loading...

Reply

Sinan says

August 20, 2018 at 4:59 am

0

Hi Jim,
It can be used methods such as backward elemination for property selection in multiple linear regresssion (MLR) . In these methods, features are removed from the model, just like in Multicolliniarity. The question I want to ask is: In an MLR application, which one to do first? Multicolliniarity or model selection?

Loading...

Reply

Jim Frost says
August 23, 2018 at 2:17 am

Hi Sinan,

This is a tricky situation. The problem is that multicollinearity makes it difficult for stepwise regression (which includes the backward elimination method) to fit the best model. I write about this problem in my post that compares stepwise and best subsets regression. You can find it in the second half of the post where I talk about factors that influence how well it works.

However, removing multicollinearity can be difficult. But, I would try to remove the multicollinearity first.

There is another approach that you can try–LASSO regression. This method both addresses the multicollinearity and it can help choose the model. I describe in my post about choosing the right type of regression analysis to use. I don't have hands-on experience with it myself, but it might be something you can look into if it sounds like it can do what you need it to do.

I hope this helps!

Loading...

Reply

Veikko says
August 8, 2018 at 6:44 am

Hi, do you have an author for your reference "Applied Linear Statistical Models"?

Loading...

Reply

Jim Frost says
August 8, 2018 at 10:36 am

Michael H. Kutner et al.

Loading...

Reply

0

John Velez says
April 9, 2018 at 12:31 pm

Hi Jim,

Again, thank you for such a great explanation!

My question is as follows:

Say you have two “severely” correlated IVs (X1 and X2) but you’re interested in examining each one individually. How would controlling (i.e., enter as a covariate) X2 while examining X1 influence your coefficients and p values? I would assume a loss of power but what else may occur? I’m also interested in any potential pitfalls of using this approach.

Thanks for your time!
John

Loading...

Reply

Jim Frost says
April 9, 2018 at 4:05 pm

0

Hi John,

There are several problems with including severely correlated IVs in your model. One, it saps the statistical power. However, it also makes the coefficients unstable. You can change the model by including or excluding variables and the coefficients can swing around wildly. This condition makes it very difficult have confidence in the true value of the coefficient. So, the lower statistical power and unstable coefficients are the major drawbacks. Basically, it's hard to tell which IVs are correlated with the DV and the nature of those relationships.

One thing I don't mention in this post, but I should add, is that you can try Ridge regression and Lasso regression, which are more advanced forms of regression analysis that are better at handling multicollinearity. I don't have much first hand experience using them for that reason but they could be worth looking into. I mention them in my post about choosing the correct type of regression analysis.

I hope this helps!

Loading...

Reply

seyi says
March 26, 2018 at 3:15 pm

0

Hi Jim,

Just to thank you for the clear explanation on your articles.

Loading...

Reply

Jim Frost says
March 28, 2018 at 3:02 pm

Hi Seyi, you're very welcome! Thanks for taking the time to write such a nice comment! It means a lot to me!

Loading...

Reply

Filmon says
March 15, 2018 at 2:08 am

You are just amazing Mr. Jim Thank you for the wonderful and exhaustive note

Loading...

Reply

Jim Frost says
March 15, 2018 at 10:26 am

You're very welcome! Thanks for taking the time to write such a kind comment!

Loading...

Reply

Vijay says
March 8, 2018 at 5:46 am

Hello Jim Frost,

It's Awesome explanation regarding Multicollinearity.
But i have one doubt, is there any other method to detect multicollinearity except VIF?

Thanks in advance.

Regards,
Vijay

Loading...

Reply

0

Jim Frost says
March 8, 2018 at 10:22 am

Hi Vijay, thank you for your kind words! Thanks for writing with the great question!

VIFs really are the best way because they calculation the correlation between each independent variable with ALL of the other independent variables. You get a complete picture of the combined strength of the correlation.

You can also assess the individual simple correlations between pairs of IVs. This approach can tell you if two IVs are highly correlated, but can miss out on correlations with multiple IVs. For instance, suppose that individually, IV1 has a moderate (but not problematic) simple correlation with IV2 and IV3. However, collectively IV1 has problematic levels of correlation with both IV2 and IV3 combined. VIFs will detect the problem while simple correlation will miss it.

I hope this helps!

Loading...

Reply

0

Blake Wareham says
February 1, 2018 at 5:21 am

Hi, Jim! Third year MA/PhD student here, very much appreciating the time and work you've put into communicating these concepts so effectively. I've been reading through a few of your posts for information about a three-way interaction in a regression used for my thesis, and will certainly cite your pages. I'm also wondering if it would also be possible to provide other sources in these posts that reiterate or elaborate on the concepts? I imagine at least some of it includes Dawson, Preacher, and perhaps the Aiken and West paper.

Thanks again for what you're doing here.

Best,
Blake

Loading...

Reply

Jim Frost says
February 1, 2018 at 11:38 am

Hi Blake,

Thanks for the kinds words! They mean a lot!

I cover two-way interactions in a previous post. However, I find the interpretation of three-way interactions to be much more complicated. Suppose the three-way interaction A*B*C is statistically significant. This interaction indicates that the effect of A on the dependent variable (Y) depends on the values of both B and C. You can also come up with similar interpretations for the effects of B and C on Y.

In my post about two-way interactions, I show how graphs are particularly useful for understanding interaction effects. For three-way interactions, you'll likely need even more graphs. If you're using categorical variables in ANOVA, you can also perform post-hoc tests to determine whether the differences between the groups formed by the interaction term are statistically significant.

Down the road, I might well write a more advanced post about interpreting three-way interactions. However, in the meantime, I hope this explanation about interpreting three-way interaction effects is helpful.

Loading...

Reply

Nivedan says

December 20, 2017 at 7:07 am

Hi Jim!

What if a categorical independent variable (converted to dummies) has 2, 3 or more category levels and are severely correlated?

Regards
Nivedan.

Loading...

Reply

Jim Frost says
December 20, 2017 at 1:33 pm

With correlated categorical, independent variables, you'll face the same problems as with any correlated independent variables. You can also use the same potential solutions–partial least squares, Lasso regression, Ridge regression, removing one of them, combining them linearly, etc.

Loading...

Reply

Avinash says

November 28, 2017 at 8:09 am

0

Thanks, Jim!

Loading...

Reply

Avinash says

November 27, 2017 at 8:13 pm

Hi Jim,

If one independent variable is a subset of another independent variable, then does that also cause problems? Say for ex, I'm looking at no. of people who have completed high school and also no. of people with a bachelor's degree. The former is clearly a subset of the latter. Should they be separately considered in 2 different models to study the effects?

And a second question – what about a cluster analysis? Does it cause any problem to have 2 correlated variables?

Loading...

Reply

Jim Frost says
November 28, 2017 at 2:01 am

0

Hi Avinash,

In your education example, yes, those variables would clearly be correlated. Individuals with a bachelor's degree must also have HS degree. And those without a HS degree cannot have a bachelor's degree. Usually, analyses will include a variable for the highest degree obtained (or worked on), which eliminates that problem.

Correlated variables can affect cluster analysis. Highly correlated variables are not sufficiently unique to identify distinct clusters. And, the aspects associated with the correlated variable will be overrepresented in the solution.

I hope this helps!
Jim

Loading...

Reply

prabuddh says
October 20, 2017 at 5:45 pm

i am still trying to figure out the solution and will share here once i get it. thank you for your response though. and great articles to understand concepts.!

Loading...

Reply

Jim Frost says
October 21, 2017 at 11:12 pm

I would love to hear what you do. Please do share! Best of luck!

Loading...

Reply

Prabuddh says
October 15, 2017 at 7:05 am

In case of marketing mix where we want to understand the impact of say radio, TV, internet and other media campaigns individually, we cannot remove or merge these variables. How to go about them? And obviously these variables highly correlated

Loading...

Reply

0

Jim Frost says
October 15, 2017 at 12:57 pm

Hi, first you should make sure that they're correlated to a problematic degree. Some correlation is OK. Check those VIFs! If they're less than 5, then you don't need to worry about multicolliearity. I'm not sure that I'd necessarily expect the correlation between the different types of media campaigns to be so high as to cause problems, but I'm not an expert in that area.

If it is too high, you need to figure out which solution to go with. The correct solution depends on your specific study area and the requirements of what you need to learn. It can be a difficult choice and not one that I can make based on a general information. In other words, you might need to consult with a statistician. You'd have to choose among options such a removing variables, linearly combining the measures, using principle components analysis, and partial least squares regression.

You've already ruled a couple of those out but there are still some options left. So, I'd look into those.

I hope this helps!
Jim

Loading...

Reply

0

Sibashis Chakraborty says
September 20, 2017 at 3:52 pm

say I am regressing Y on two variables x1 and x2 and x1 and x2 has a high correlation between them. If I run a regression of Y on x1 and then Y on X2 and finally Y on x1 and x2 I will get different values for coefficients that will affect my prediction. don't you think?

Loading...

Reply

Jim Frost says
September 22, 2017 at 9:04 pm

Hi, it depends on all of the specifics. However, if X1 and X2 are highly correlated, they provide very similar information in model. Consequently, using X1, X2 or, X1 and X2 might not change the predictions all that much. As the correlation between the two increases, this becomes ever more the case.

I hope this helps!
Jim

Loading...

Reply

0

Comments and Questions

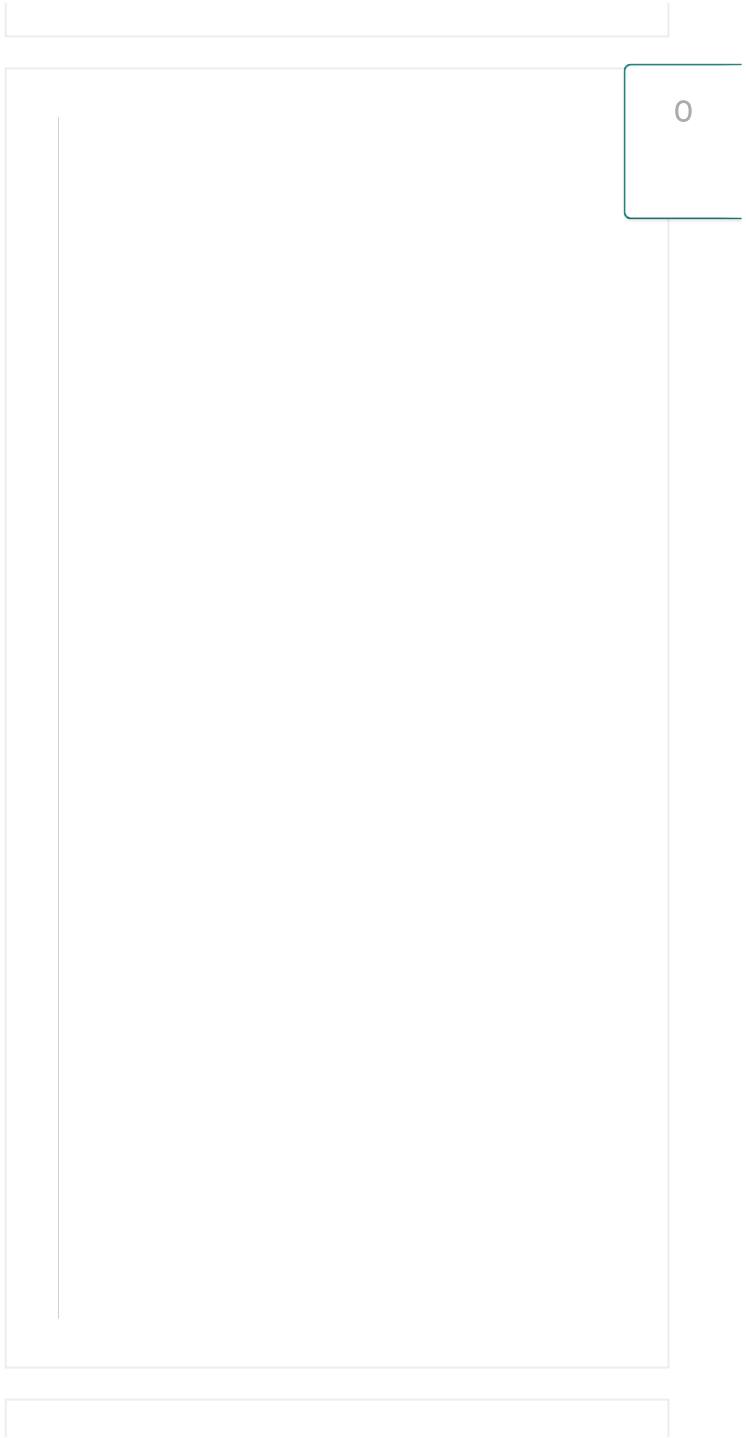
Meet Jim

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...

0

Search this website


0

Understanding Interaction Effects in Statistics

0

Recent Comments

D Clark on 7 Classical Assumptions of Ordinary Least Squares (OLS) Linear Regression

Jim Frost on 7 Classical Assumptions of Ordinary Least Squares (OLS) Linear Regression

D Clark on 7 Classical Assumptions of Ordinary Least Squares (OLS) Linear Regression

Erin on Standard Error of the Mean (SEM)

Jim Frost on Using Confidence Intervals to Compare Means

0

0

Images are still loading. Please cancel your print and try again.