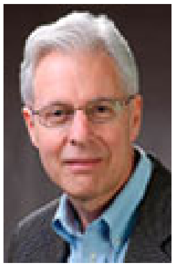
When Can You Safely Ignore Multicollinearity?

statisticalhorizons.com/multicollinearity

September 10, 2012 By Paul Allison



Multicollinearity is a common problem when estimating linear or generalized linear models, including logistic regression and Cox regression. It occurs when there are high correlations among predictor variables, leading to unreliable and unstable estimates of regression coefficients. Most data analysts know that multicollinearity is not a good thing. But many do not realize that there are several situations in which multicollinearity can be safely ignored.

Before examining those situations, let's first consider the most widely-used diagnostic for multicollinearity, the variance inflation factor (VIF). The VIF may be calculated for each predictor by doing a linear regression of that predictor on all the other predictors, and then obtaining the R^2 from that regression. The VIF is just 1/(1-R^2).

Learn more in a seminar with Paul Allison

It's called the variance inflation factor because it estimates how much the variance of a coefficient is "inflated" because of linear dependence with other predictors. Thus, a VIF of 1.8 tells us that the variance (the square of the standard error) of a particular coefficient is 80% larger than it would be if that predictor was completely uncorrelated with all the other predictors.

The VIF has a lower bound of 1 but no upper bound. Authorities differ on how high the VIF has to be to constitute a problem. Personally, I tend to get concerned when a VIF is greater than 2.50, which corresponds to an R^2 of .60 with the other variables.

Regardless of your criterion for what constitutes a high VIF, there are at least three situations in which a high VIF is not a problem and can be safely ignored:

1. **The variables with high VIFs are control variables, and the variables of interest do not have high VIFs.** Here's the thing about multicollinearity: it's only a problem for the variables that are collinear. It increases the standard errors of their coefficients, and it may make those coefficients unstable in several ways. But so long as the collinear variables are only used as control variables, and they are not collinear with your variables of interest, there's no problem. The coefficients of the variables of interest are not affected, and the performance of the control variables as controls is not impaired.

Here's an example from some of my own work: the sample consists of U.S. colleges, the dependent variable is graduation rate, and the variable of interest is an indicator (dummy) for public vs. private. Two control variables are average SAT scores and average ACT scores for entering freshmen. These two variables have a correlation above .9, which corresponds to VIFs of at least 5.26 for each of them. But the VIF for the public/private indicator is only 1.04. So there's no problem to be concerned about, and no need to delete one or the other of the two controls.

2. The high VIFs are caused by the inclusion of powers or products of other variables. If you specify a regression model with both *x* and x^2, there's a good chance that those two variables will be highly correlated. Similarly, if your model has *x*, *z*, and *xz*, both *x* and *z* are likely to be highly correlated with their product. This is not something to be concerned about, however, because the *p*-value for *xz* is not affected by the multicollinearity. This is easily demonstrated: you can greatly reduce the correlations by "centering" the variables (i.e., subtracting their means) before creating the powers or the products. But the *p*-value for x^2 or for *xz* will be exactly the same, regardless of whether or not you center. And all the results for the other variables (including the R^2 but not including the lower-order terms) will be the same in either case. So the multicollinearity has no adverse consequences.

3. The variables with high VIFs are indicator (dummy) variables that represent a categorical variable with three or more categories. If the proportion of cases in the reference category is small, the indicator variables will necessarily have high VIFs, even if the categorical variable is not associated with other variables in the regression model.

Suppose, for example, that a marital status variable has three categories: currently married, never married, and formerly married. You choose formerly married as the reference category, with indicator variables for the other two. What happens is that the correlation between those two indicators gets more negative as the fraction of people in the reference category gets smaller. For example, if 45 percent of people are never married, 45 percent are married, and 10 percent are formerly married, the VIFs for the married and never-married indicators will be at least 3.0.

Is this a problem? Well, it does mean that *p*-values for the indicator variables may be high. But the overall test that *all* indicators have coefficients of zero is unaffected by the high VIFs. And nothing else in the regression is affected. If you really want to avoid the high VIFs, just choose a reference category with a larger fraction of the cases. That may be desirable in order to avoid situations where none of the individual indicators is statistically significant even though the overall set of indicators is significant.

Comments (416)

416 Responses

1. *Arne Mastekaasa* says:

   September 12, 2012 at 11:53 am
   Thanks for interesting and useful comments on this issue. I am not a specialist on this, but I miss one consideration. The VIF is a measure of relative increase in the variance of the estimate. But why should one care about relative increases (a VIF of, say, 3) if the absolute value of the variance (and the standard error) is minute, i.e., if the sample size is sufficiently large? Is it not primarily a matter of statistical power?

   Reply

   - *Paul Allison* says:

     September 17, 2012 at 1:19 pm
     If the model is correctly specified (with the right covariates and the right functional form) then the issue is, in fact, "primarily a matter of statistical power." So with very large samples, multicollinearity should not be a problem. However, multicollinearity also makes the estimates very sensitive to minor changes in specification. Suppose, for example, that two variables, x and z, are highly collinear. Suppose, further, that the effect of x on y is strictly linear but the effect of z on y is slightly non-linear. If we estimate a strictly linear model, the effect of x on y could be greatly exaggerated while the effect of z on y could be biased toward 0. This is sometimes known as the "tipping effect", and it can happen even with very large samples.

     Reply

     ARIYO OLADELE A says:

     February 5, 2015 at 3:23 pm
     Happy day Sir. I am using response surface methodology to find optimality and also for prediction but the result of my first and second order analysis shows a pure error of zero, which means my lack of fit will be zero, can I still make use of the model since my p-values for most of the terms in the model are signficant.

     (2.) I also want to confirm whether the AIC of the reduced model in second order design is always higher compare with the AIC of the initial model.
     Thank you Sir

     Reply

     Paul Allison says:

     February 9, 2015 at 1:20 pm
     Sorry, but I don't have answers to these questions.

     Reply

- *Timon* says:

  September 23, 2012 at 6:42 am
  Well, centering does rdecue multicollinearity, and thus is it not the same in the two models. It is possible to take all the covariance out of the matrix of predictors, but only by taking out a corresponding amount of variance. Thus, no new information is added and the uncertainty remains unchanged. Also, centering can be help computation by improving the convergence of MCMC chains. See, for example, Gelman and Hill.

  Reply

2. *Pat Rubio Goldsmith* says:

   October 1, 2012 at 1:53 pm
   I seem to recall from an old Hanushek book that multicollinearity does not bias coefficients; it inflates their standard errors. These large standard errors make p-values too large. That is a problem when the p-values go above a threshold like .05, but otherwise, the inflated standard errors don't change the interpretation of the results. So, in addition to the comments you made above, multicollinearity does not usually alter the interpretation of the coefficients of interest unless they lose statistical significance.

   Reply

3. *joe scarborough* says:

November 16, 2012 at 11:49 am
in the instance of vif associated with x and x^2— while centering x^2 may have no effect on the parameter or p value for x^2— it may affect the parameter /p value for x— so if one is interested in both— then centering may be preferred

Reply

Paul Allison says:

November 16, 2012 at 12:02 pm
Yes, but x and x^2 should not be seen as separate effects but as one quadratic effect. And that effect (as shown in a graph) will be the same whether you center or not.

Reply

Alex says:

January 17, 2014 at 4:26 pm
In my case I need to test if a relation is quadratic or not. Without centering the x and the x^2 I get significance p values but high VIFs, around 17. If I center the variables only x^2 is still significant while the x term not anymore. Which is the correct result? What does it mean if only the quadratic term is significant? thanks

Reply

Paul Allison says:

January 22, 2014 at 8:55 am
Your result is not surprising. When fitting models with x and x^2, the coefficient (and significance) of x^2 is invariant to centering. But the coefficient (and significance) of x is not. The answer is that the two models are really equivalent and there's no strong reason to prefer one over the other. I would keep x in the model, however.

Reply

4. *Jean-Bernard Chatelain* says:

   December 9, 2012 at 4:11 am
   Thanks for these very interesting comments. To Pat Rubio Goldsmith: near-multicollinearity does not bias coefficients, but it is their interpretation as "ceteris paribus" effects (which is not a theorem) which turns to be an "act of faith", because a shock of one unit of one of the highly correlated regressor is very likely to imply that the other highly correlated regressor will move as well by nearly one unit and will not remain unchanged.
   We have a concern when there are highly correlated regressors which are weakly correlated with the dependent variable (in particular when using interaction terms or quadratic terms). Statistical significance is easily obtained because of the particular shape of the critical region of the t-test in this case. However, the effects may be spurious and/or outliers driven (overfitting) due to an issue initially unrelated to power and large samples. What do you think?

   Can statistics do without artefacts
   http://mpra.ub.uni-muenchen.de/42867/1/MPRA_paper_42867.pdf
   Technical paper: Spurious regressions with near-multicollinearity
   http://mpra.ub.uni-muenchen.de/42533/1/MPRA_paper_42533.pdf

   Reply
5. *KH Lee* says:

   January 14, 2013 at 9:43 am
   Thanks for your comments. It's been very helpful for correcting what I have misunderstood so far.
   I'm now working on writing parts of results of empirical analysis in paper. Here, I'd like to refer to this comments. Can you recommend books or papers that contain above suggestions?
   I'll appreciate you send me an e-mail for this.
   Thanks again.

   Reply

   Paul Allison says:

   January 23, 2014 at 10:12 am
   Wooldridge has a good discussion of multicollinearity in Chapter 3 of his book Introductory Econometrics.

   Reply

6. *Del* says:

   January 17, 2013 at 12:38 pm
   I am doing SEM and my variables are likert with 5 categories. We usually treat these variables as normal but they are at least ordinal. Can we get the VIF for ordinal or categorical data? Do we have to specific the data type in SPSS?

   Reply

   Paul Allison says:

   January 17, 2013 at 2:15 pm
   Well, if you're treating these likert scales as quantitive variables in the regression model, then the VIFs would be calculated just like any other variable. If you're treating them as categorical in the regression, there would be a VIF for each indicator variable.

   Reply

7. *Olatunji, I.A.* says:

   February 11, 2013 at 12:44 am
   Thanks. I found the lecture note, the comments and replies very helpful to interpret my regression results. I have 5 regressors from Factor Analysis of residential choice Optimality determinants, two with VIFs of 4.6 and 5.3 respctvly. Commuting cost has distance, as part of its function and so are highly correlated. Variables of interest are Rental value, income and activity pattern. Can we ignore the high VIFs?.

   Reply

   - *Paul Allison* says:

     February 11, 2013 at 3:09 pm
     I don't think these VIF's should be ignored. Are these really distinct factors?

     Reply

   - *Kabir Opeyemi* says:

     February 18, 2013 at 3:44 pm
     Collinearity is often a data problem. It is purely as a result of ill-conditioning in the data. It could be remedied, sometimes, by redefining variables to reduce or eliminate inter-relations. If x1 and x2 are highly correlated, for example, redefine to Z1 = X1 – X2 or Z1 = X1 + X2. You may transform X1- Xp to principal components PC1 – PCp. PC1 – PCp will be uncorrelated linear combinations of X1- Xp. Then regress y on PC1 – PCp. I recommend R package 'pls'.

     Reply

8. *Chris* says:

February 20, 2013 at 12:38 pm
Hey,

I am working on my thesis and looking for a paper to cite the third point, as the dummy variables in my regression have a high point biserial correlation with the continuous variables and a high VIF. The VIFs of my continuous variables are all below 2, but the VIFs of the dummies are ranged between 4 and 5.

Thanks in advance

Reply

Paul Allison says:

February 20, 2013 at 1:03 pm
Do the dummy variables represent a single categorical variable, or more than one categorical variable?

Reply

Chris says:

February 20, 2013 at 5:33 pm
3 dummies represent one categorical variable.

1 additional dummy variable is independent of these 3 (it's VIF equals 4.57), which is quite high, and will be tested via a robustness test. This dummy variable equals 1 only for a fraction of the data set (5000 out of 100000 observations). Does this reveal a possible reason?

Reply

Paul Allison says:

February 28, 2013 at 3:13 pm
I don't think that's a likely reason. I would do a regression of this variable on all the others to see which variables are highly related to it.

Reply

9. *Denis* says:

March 6, 2013 at 7:41 am
Hi,
ich have a huge problem and I cannot find an answer. I am doing a research on the moderating effect of gender on entrepreneurial intention. The thing is, that the VIF get over 2.5 then I enter the interactionterms. What can I do to reduce the VIF or are there any theories that show, that a high VIF with binary variables as a moderator is not problem. Hope someone can help me.

Reply

Paul Allison says:

March 6, 2013 at 8:57 am
As I said in the post, high VIFs for an interaction and its two components is not usually a problem. But you should be able to reduce the VIF by subtracting the mean of entrepreneurial intention before creating the product variable.

Reply

Denis says:

March 8, 2013 at 4:28 am
Thank you Mr Allison, the thing is I already substracted the mean but it does not really reduce the VIF.

Reply

Paul Allison says:

March 8, 2013 at 10:13 am
Try subtracting the mean for the dichotomous variable as well.

Reply

10. *Fk* says:

    April 2, 2013 at 8:08 pm
    Dear Paul,
    Thanks for your numerous blogs and excellent books.
    How can one assess for collinearity between two categorical predictors in a logistic regression model?

    Reply

    Paul Allison says:

    April 3, 2013 at 7:18 am
    Just like linear regression. Use an ordinary linear regression program and request the variance inflation factor or tolerance. For details and rationale, see pp. 60-63 in my book “Logistic Regression Using SAS”, 2nd edition.

    Reply

11. *Thuan Chu* says:

    April 7, 2013 at 11:24 pm
    Hi Paul,
    Thanks for your great book “Logistic Regression Using SAS”, I bought it last week. I am having problems with variables selection in logistic model. All variables in minimal model are significant terms (using AIC and p-value), but there are some of them are very high collinearity (VIF>300). How can I explain my result to keep those variables in the final model since they are important terms to explain the variation of response variable? (I am applying LG in remote sensing to map burned and unburned areas with spectral bands, sample size = 10 000).

    Reply

    Paul Allison says:

    April 8, 2013 at 10:00 am
    I would carefully examine the bivariate correlations among these variables. A VIF>300 is very high.

    Reply

12. *Thuan Chu* says:

April 8, 2013 at 12:25 pm
There are some high bivariate correlation (0.8<R^2<0.9). What should I do? Thank you.

Reply

Paul Allison says:

April 9, 2013 at 4:13 pm
Hard to say without knowing more about what your objectives are.

Reply

Thuan Chu says:

April 10, 2013 at 1:32 pm
My objective is to use logistic regression to discriminate burned pixel (by value 1), and unburned pixel (by value 0) from a set of explanatory variables which are spectral bands (7 bands) of satellite image. Spectral bands range from visible wavelength to mid-infrared wave length. Thanks.

Reply

Paul Allison says:

April 10, 2013 at 3:29 pm
Well, if your main goal is a model that will predict or discriminate, multicollinearity is less of a concern. But check your criterion for predictive accuracy (e.g., area under the curve, R squares, etc.) to see if you get a noticeable improvement when both highly correlated variables are in the model (as opposed to just one of them).

Reply

13. *Jim Pence* says:

April 9, 2013 at 12:45 pm
Dr. Allison – thanks for your postings on this topic. I have a case where I have 15 levels for my categorical variable. I created 14 dummies for both intercept and slope, so that each level can have (potentially) its own unique slope and intercept. I'm seeing very high VIFs between the intercept term and the slope term within a given level, which makes sense given the high degree of correlation between them. I'm assuming that since this is an "intra-variable" issue, and not an "inter-variable" one, then I don't need to be concerned with the high VIFs?

Reply

Paul Allison says:

April 9, 2013 at 4:12 pm

Not sure I understand the data structure. And is this a mixed model?

Reply

Jim Pence says:

April 18, 2013 at 9:52 am
Here’s an example (with 3 levels) of what I’m doing.

```
data x;
/* Create intercept dummies */
port1 = (port = 1);
port2 = (port = 2);
port3 = (port = 3);

/* Create slope dummies */
slope1_interaction = port1 * HPI;
slope2_interaction = port2 * HPI;
slope3_interaction = port3 * HPI;
run;

proc reg;
/* use port3 as “reference” */
model y = port1 port2 HPI slope1_interaction slope2_interaction;
run;
```

```
OUTPUT
——
Variable VIF
Intercept 0
port1 165
portt2 220
HPI 4
slope1_interaction 166
slope2_interaction 220
```

The high VIFs occur between the intercept dummy and slope dummy for each level of my independent variable “port”.

Reply

> *Paul Allison* says:
>
> April 30, 2013 at 10:17 am
> Try centering the HPI variable before creating the interactions.
>
> Reply

14. *Scott* says:

    April 19, 2013 at 2:32 pm
    Dr. Allison,
    I have 5 dummy variables for my education measure and my reference category is the smallest, but provides the best explanation for the data (all p values for the indicator variables are .001 so significance is not a problem). Similar to Chris above, I am writing my dissertation and hoping to find a citation I can use to justify the multicollinearity. Are you aware of any? Thank you.

    Reply

    Paul Allison says:

    April 30, 2013 at 10:15 am
    Sorry, but I can't think of a citation. Does anyone else have a suggestion?

    Reply

15. *Ivana* says:

    April 29, 2013 at 5:54 pm
    Hi there,
    I was wondering if there's a consensus in the literature on how to approach collinearity in survival models. I'm running a loglogistic AFT model but some of my independent variables seem to be highly correlated with each other (.7-.8), and so I've tried running them in OLS to get the VIF. A few models have a handful of predictors with VIF larger than 10, so should I be orthogonalizing them using the Gram-Schmidt procedure?
    Thanks very much!

    Reply

    Paul Allison says:

    April 30, 2013 at 10:14 am
    No consensus, but the problem is pretty much the same as in OLS regression. You VIFs are quite high. I'm not convinced that Gram-Schmidt really alleviates the problem.

    Reply

16. *Alex* says:

May 1, 2013 at 5:26 am
Hi Paul,

Thank you for making this resource available.

I have 3 dummy variables to describe physical activity. The reference category is ‘inactive’ vs. ‘more active’, ‘active’ and ‘most active’. The VIFs (and tolerance) for these latter three are 12.4 (.080), 12.7 (.079) and 9.1 (.110) respectively.

As they are dummy variables, I am tempted to “safely ignore” this multicollinearity. But are these VIFs really TOO high; should I be concerned?

Thank you!

Alex

Reply

Paul Allison says:

May 1, 2013 at 7:57 am
The question is, what proportion of the cases are in the reference category? If it is very small, that is probably the cause of the multi-collinearity. Maybe you should choose a different reference category. Are any of the cofficients statistically significant? That’s an indication that the collinearity doesn’t matter. Have done a global test for all three dummies? That test is invariant to the choice of the reference category.

Reply

Alex says:

May 2, 2013 at 6:24 am
I’ve redone my groups (as you suggested) and my model makes so much more sense now!

Many thanks again for being available to answer these queries!

Reply

17. *Andreas* says:

May 1, 2013 at 10:34 am
Dr. Allison,

I have a question regarding multicollinearity, which I did not find an answer for in any of my statisic books.

I'm working with an unbalanced panel dataset (t: 10 years, x: 170 companies) calculating logistic regressions with random effects including "normal" (company age, sales growth, …) as well as dummy independent variables (fiscal year, industry). I checked for multicollinearity and have no VIF above 2.5. Is there anything I need to consider because I'm working with a panel model or is there no difference between the VIFs of a standard logistic model and an logistic model with random effects model?

There is one specific source of multicollinearity I'm curious about: I'm using a company age (in years) variable as well as dummies for the different fiscal years. The VIFs calculated for "company age" do not indicate any problems with multicollinearity and I receive significant results for the "company age" variable in my random effects model. On the other hand as far as I understand the within variance of the "company age" variables should be explained perfectly by the fiscal year dummies. Do I have to change the "company age" variable and what would be the best way to do so? (e.g. variable "company age in 2002" or "company age when entering the sample"; this is a difference, because my sample is unbalanced)

Thank you very much!
Andreas

Reply

Paul Allison says:

May 1, 2013 at 11:54 am
The VIFs are purely descriptive, and I see no serious objection to using them in a logistic random effects model. As for the second question, I'd probably use company age when entering the sample, for the reasons you suggest.

Reply

Andreas says:

May 2, 2013 at 3:48 am
Thank you very much, that helped a lot.

One short additional question: Can I use the same VIFs I would use in a pooled model or do I need to calculate the VIFs using a linear regression with random effects?

Reply

Paul Allison says:

May 2, 2013 at 7:45 am
I'd use the same VIFs.

Reply

18. *Li Li* says:

May 1, 2013 at 2:15 pm
Dr.Allison – thank you so much for this article, very helpful to me indeed. I have a question regarding the high VIF among different levels of a categorical variable. While you said the "…but the overall test that all indicators have coefficients of zero is unaffected by the high VIFs", I wonder if you can explain more what the impact on individual indicator coefficient estimates will be. Will some indicators (individual category levels) will have coefficient estimates of the opposite sign from what it should be. What will be the impact on the absolute value of the estimates? And for my project, I'm more interested in the impact of each independent variables on the dependent, than the predictive power of the model, what will your suggestion be if I have high multicollinearity among my category levels? Many thanks.

Reply

Paul Allison says:

May 1, 2013 at 2:25 pm
My bottom line suggestion is to choose as your reference category the one with the most cases. That will minimize any collinearity problems. In the case of categorical predictors, the collinearity that arises among them will not bias the signs of the coefficients. But it may make it appear that none of the categories is significant when, in fact, there are significant differences among them.

Reply

19. *Alonso Bussalleu* says:

 May 23, 2013 at 3:51 pm
 Dr. Allison Thank you for the great insights on how to deal with collinearity.

 I have a question regarding how to deal with collynearity between two nominal categorical explanatory variables. When I use a contingency analysis, it shows a strong association between them. A pair of binomial variables showed a correlation coefficient of -0.9, which I think should be interpreted as a strong association as well, although I don't think it is the correct way to test for this, since there is no order between categories.
 I haven't coded them as dummy variables, but I think the statistical program I am using (R) estimates parameters for each level treating them in this way, but separately for each variable not solving a a possible erroneous parameter estimation.

 I appreciate any comment on this matter.
 Thanks

 Reply

 Paul Allison says:

 May 28, 2013 at 9:13 am
 If I understand you correctly, this would certainly be an instance of serious multicollinearity. If the variables are binary predictors in a regression, the Pearson correlation between them is an appropriate way to assess their association.

 Reply

20. *Tuan* says:

 June 5, 2013 at 11:27 pm
 Thanks for your good lecture.
 I have a problem here. I use linear regression to find the relation between y and x1 x2 x3 and the result shows that y and x1 is not linear relation and p values of x1 is more than 0.05. Form the graph, I decide to add x1^2 into model. The result is good with all p<0,05. Now, if I center x1 to deal with collinearity then p value of x^2 is not change but p of x is more than 0.05
 Can I ignore collinearity safely?
 Thanks in advance

 Reply

 Paul Allison says:

 June 8, 2013 at 10:40 am
 Yes

 Reply

21. *KOOLE O.K* says:

    June 22, 2013 at 7:58 pm
    I have a question regarding multicollinearity.
    Im doing research based on factor analysis . Im using SPSS to analyse my data.Determinants of my study is 9.627E-017 which I think is 0.000000039855 indicating that multicollinearity is a problem.Field (2000) say if determinant of correlation matrix is below is 0.00001 multicollinearity is a serious case.Im requesting for help.

    Reply

    Paul Allison says:

    June 24, 2013 at 12:34 pm
    Multicollinearity is less of a problem in factor analysis than in regression. Unless they cause total breakdown or "Heywood cases", high correlations are good because they indicate strong dependence on the latent factors.

    Reply

    Carol says:

    June 24, 2015 at 4:34 pm
    Dear Paul, thank you for your excellent blog. The determinant for my correlation matrix is 0 doing PCA on SPSS, but I don't see any of my items correlating more than 0.6 with another. Could you please advise? Thank you.

    Reply

    Paul Allison says:

    June 25, 2015 at 6:51 am
    You can have multicollinearity even if none of the bivariate correlations is high.

    Reply

22. *Jordan* says:

    June 28, 2013 at 4:06 pm
    In regards to the statement that multicollinearity is only a problem for the variables that are colinear: is there a commonly cited reference for this? I am attempting to raise this point in response to a manuscript review and would like to be able to back it up with a published reference, if possible. Thanks.

    Reply

    Paul Allison says:

    June 28, 2013 at 8:22 pm
    Jeffrey Wooldridge (2013) Introductory Econometrics, 5th ed., p. 97.

    Reply

23. *Arul Nadesu* says:

    August 20, 2013 at 12:51 am
    Hi Allison,

    There is no formal cutoff value to use with VIF for determining presence of multi-collinearity. However, I have been using VIF = 2.5 in Logistic Regression modelling to reduce multi-collinearity. Can I completely ignore multi-collinearity when using Classification Trees?
    thanks

    Reply

    - *Paul Allison* says:

      August 20, 2013 at 10:42 am
      Interesting question. As I understand it, classification trees are used primarily for prediction/classification, not for estimating "causal" effects. For that reason, I'm guessing that multicollinearity would not generally be a serious issue for this kind of analysis.

      Reply

24. *Subi* says:

August 21, 2013 at 12:51 pm
Dr. Allison:
I am planning a study where there are three variables of interest in the model: a)allergic rhinitis, b)allergic asthma, and c)allergic rhinitis and allergic asthma (as a composite) variable (plus the other covaraites). There is a possibility of multicollinearity among these variables. How should I take care of this problem during the analysis?

Reply

Paul Allison says:

August 21, 2013 at 1:52 pm
Depends on what your goal is. This sounds like an interaction test in which you have x1, x2, and their product x1*x2. As I wrote in the post, the test for the interaction is invariant to any multicollinearity, so there's nothing really to do. On the other hand, if you want to take a more descriptive approach, you can view this as a four-category variable: neither, asthma only, rhinitis only, and both. Treat neither as the reference category and include indicators for other three. Again, no problem with multicollinearity.

Reply

Subi says:

August 21, 2013 at 3:08 pm
sorry, I was not clear enough. I will use 3 indicator variables as my 4 exposure categories(x1 to x4). It is now clear to me that I will not run into this problem! Thank you so much for verifying this information…. it was really crucial for my dissertation proposal!!

Reply

25. *Arslan* says:

    August 23, 2013 at 1:32 pm
    Dr. Allison,

    I have created an interaction variable between a continuous variable (values ranging from 0 to 17) and a categorical variable (3 ategories 0,1,2).

    Teh question is: Do I need to check VIFs for the three variables together? (e.g. X, Z, and XZ). and if I do that, the VIF is still high (around 7.09 for interaction variable and 5.21 for one of the constituent variable), can I igonre that?

    Also, please tell me how do I cite this article?

    I shall be grateful to you.

    Reply

    - *Paul Allison* says:

      August 23, 2013 at 3:24 pm
      Yes, I would ignore it. As for citing my blog post, here is the American Psychological Association style: Allison, P.D. (2012, September 10). When Can You Safely Ignore Multicollinearity [Web log post]. Retrieved from https://statisticalhorizons.com/blog.

      Reply

26. *Mandy* says:

August 26, 2013 at 4:17 pm
Dear Dr Allison,

Thank you for your blog post and i would like to ask a question relating to point 2.

“The high VIFs are caused by the inclusion of powers or products of other variables” (then it’s ok to ignore)

I still don’t quite understand why it doesn’t matter if an interaction term is product of 2 IVs and thereofre highly correlated. could you explain more please?

I have a dataset with 3 predictors and 2 interaction terms. The interaction term requires multiplying a continuous with a dichotomous variable. I centered all IVs but not the dichotomous variable, so there is still high VIF.

are you suggesting we center ALL variables, including continuous predictors, dichotomous predictors and then create interaction terms using these centered variables?

I read an article that said not to center dichotomous variables… bottom of p.4 – http://www3.nd.edu/~rwilliam/stats2/l53.pdf

I also read an article that suggested that centering doesn’t help… p.71 of the artcile below

https://files.nyu.edu/mrg217/public/pa_final.pdf

I am feeling very confused about all these different sources of information. would you be able to help clarify please?

Many thanks for your help – it’s very much appreciated!!
Mandy

Reply

Paul Allison says:

August 27, 2013 at 9:29 am
Centering does not change the coefficient for the product nor its p-value. In that sense, centering doesn't help. However, the fact that it doesn't change those results (even though it can markedly reduce the collinearity) demonstrates that collinearity is not really an issue. In very extreme cases, the collinearity may be so high that numerical/computational problems can arise, in which case centering may be necessary. Centering does change the interpretation of the main effects, however, often for the better. Without centering, the main effects represent the effect of each variable when the other variable is zero. With centering, the main effects represent the effect of each variable when the other variable is at its mean.

Reply

Mandy says:

August 28, 2013 at 6:00 pm
Thank you for your help. i will do the centering for my variables then. thanks again.

Reply

27. *Arul Nadesu* says:

September 15, 2013 at 6:47 pm
Hi Allison,
I have found the variables (using VIF = 2.5 or less to reduce multi-collinearity)for the Logistic Regression model. The model further improves when I transform four of the intervel variables used in the model. Do I have to calculate VIF values again using transformed values of the variables?
regards
Arul

Reply

Paul Allison says:

January 22, 2014 at 10:25 am
Probably worth doing. Transformations can make a substantial difference.

Reply

28. *Bob* says:

    September 21, 2013 at 5:01 pm

    Paul Allison, My problem using SPSS to center a variable in order to reduce its collinearity with its quadratic term is that by subtracting each case's score from the variable mean, you create a variable with negative as well as positive scores. SPSS informs me it will treat all negative scores as system missing. This greatly reduces the N. How can I get around this problem in SPSS?

    Reply

    Paul Allison says:

    January 22, 2014 at 10:22 am

    I'm no SPSS expert, but I would be extremely surprised if SPSS treated negative values of predictor variables as missing. Negative values occur frequently for many variables.

    Reply

29. *Brian* says:

    September 24, 2013 at 8:29 pm

    Hi Dr. Allison,
    I stumbled upon this post as I was searching for an answer to my dilemma. I am using restricted (natural) cubic splines (Dr. Harrell macro in SAS) for a logistic regression – aimed at prediction only. The model fits well (as determined by a test set) but the VIF of the terms in the cubic spline are HUGE. Is this a non-issue as well (same as you suggest for powers of a variable)? Thanks!

    Reply

    Paul Allison says:

    January 22, 2014 at 10:19 am

    Don't know much about cubic splines. But I'm guessing that high VIF's should not be a concern here.

    Reply

30. *Jessica* says:

October 1, 2013 at 8:48 pm
Dr. Allison

How can I address multicollinearity if my independent variable (depression) and a control variable (anxiety)are highly correlated? What should I do in this case?

Thanks,

Jessica

Reply

Paul Allison says:

January 22, 2014 at 10:03 am
Not a lot that you can do in this situation. Be cautious about your conclusions. Try out different specifications to see how robust your findings are.

Reply

31. *James* says:

    October 5, 2013 at 8:22 am
    Dear Dr Allison,

    Thank you for your article. I've been looking through numerous articles explaining the importance of centring interaction terms due to possible multi-collinearity. However your article makes me feel reassured that I can ignore high VIF's as I am working with interactions. However I still feel uneasy.

    I have four continuous IV's, one categorical DV and three interaction terms based upon the four continuous IVs.

    Some of the IV's are highly correlated to each other, whilst two of them are highly correlated to the DV (e.g above 0.7)
    I cannot throw them out despite them being highly correlated, as they are important variables that represent the model I am testing.

    Despite centring I still have high VIF's mainly with the IV's I have used in the interaction terms. I have tried to centre using the median, but the high VIF still exists. Is it safe to ignore VIF and interpret the standardised coefficients (final step of the stepwise multiple regression model)? or should I be worried as my IV's have high correlations? They are supposed to though I think. As I am working with intention and habit to predict behaviour.

    Many thanks
    James

    Reply

    Paul Allison says:

    January 22, 2014 at 9:49 am
    First of all, if you've got interactions in the model, the standardized coefficients for both the interactions and their related main effects are worthless–ignore them. Second, the p-values for the interactions are what they are. You may have low power to test them, but there's no easy fix.

    Reply

32. *Trần Quang Tuyến* says:

    October 7, 2013 at 3:43 am
    Dear Professor Paul Allsson,

    Could you please show me your book or article telling about the above topic:” When Can You Safely Ignore Multicollinearity?”. I would like to use it as a valid reference for my paper

    Kind Regards

    Tuyen

    Reply
33. *LL* says:

    October 8, 2013 at 12:40 pm
    Dear Dr. Allison,

    I ran a categorical moderated regression my categorial variable has three categories and my reference group has the highest number of participants. I used the same variables to test two different outcomes so I adjusted my alpha to .025 (using the Bonferroni correction). My results indicate that the step that includes the interaction is significant p .025. The VIF of my continuous variable (centered) is 3.17 and the VIF of one of my interaction terms is 2.50. I’m having trouble interpreting this result.

    Thank you!

    Reply

    Paul Allison says:

    January 22, 2014 at 9:45 am
    I wouldn’t be concerned about these VIFs.

    Reply

34. *Ayse* says:

    October 30, 2013 at 10:25 am
    Dear Dr. Allison,
    I'm doing OLS regression and I use the factor notation for interactions.

    The model without the interactions (that doesn't use factor notation) has no problem with MC. But when I include the factor notation the VIF of the primary independent variable skyrockets. The other VIFs are in the acceptable range.
    The STATA code for the models are:

    *reg male_cesd hage hjhs hhs hcol hkurdish Zestrainchg estrainbcz lnhhincome hunempdur2 finsup_male inkindsup_male jobsup_male emosup_male, vce(robust)

    *reg male_cesd c.hunempdur2##(c.hage i.hjhs i.hhs i.hcol i.hkurdish c.Zestrainchg c.estrainbcz), vce(robust)

    it's the c.hunempdur2 variable in the second regression code that is one of the variables of interest and has the skyrocketing VIF. Also, one of the interactions with c.hunempdur2 also has a high VIF.

    Should this be a point of concern?

    Thank you in advance.

    Reply

    Paul Allison says:

    January 22, 2014 at 9:44 am
    I wouldn't be concerned.

    Reply

35. *Chrystel* says:

    November 3, 2013 at 2:42 pm
    Dr. Allison,

    For my master's thesis, I have the exact situation that you describe in point 2. In my model I have an interaction term of the form: x2 * y *z. I am looking for an article to cite to justify a high VIF in this situation and have found none yet. Any suggestions?

    Thank you very much in advance

    Reply

36. *DONIA* says:

    November 7, 2013 at 10:08 am
    Dear Dr.Allison,
    I am using survey data. My outcome variable is continuous (mathematics score) and my predictor variables are ordinal and binary (like possessing a computer, possessing a study desk..parents'highest education level, spend time work on paid jobes…)I have a total of 6 binary variables, 5 ordinal variables , one nominal variable(parents born in country) and one continuous variable(age). I am working on 15 countries individually (I have a cross section data TIMSS 2007), I wonder how to test in my case for multicollinearity.

    Reply

    Paul Allison says:

    November 25, 2013 at 1:34 pm
    I'd recommend the usual statistics, like the variance inflation factor.

    Reply

37. *Sandip* says:

    November 27, 2013 at 2:42 pm
    Dear Dr. Allison ,
    I am working with panel data containing 110 observations (22 cross sections for 5 years) and I need to apply quantile regression (which is robust) to test any non-linear effects.
    Do I need to calculate VIF and conduct Panel Unit root tests ? The reference papers do not report them .

    Reply

    Paul Allison says:

    January 22, 2014 at 9:37 am
    Sorry but I really don't know much about quantile regression. But I'm guessing that multicollinearity could be just as much an issue here as with other regression methods.

    Reply

38. *David* says:

    November 29, 2013 at 2:45 pm
    Dear Dr. Allison,

    Thank you for posting the information on multicollinearity. I am working on learning curve based study and one of the issues that I am running into is the interpretation of the squared term of X. The theory suggests that the relationship between Y and X (experience) is U-shaped. After centering X, we can get two types of extremes, either positive value (e.g., centeredq1=+100) indicating a lot of experience or a extremely negative value (e.g., centeredq2=-100) indicating very minimal experience. While I am fine using the original value and its quadratic term to test learning theory, I think the centered values lead to confusion. If both quadratic terms of q1 and q1 (i.e., both equals to 10000) load with negative coefficients (say .025) at p<.001, are we reaching the conclusions that both high experience and low experience will have the same effects in the long run? I will try to rephrase if my problem is not clear.

    Reply

    Paul Allison says:

    January 22, 2014 at 9:29 am
    In my experience, trying to interpret quadratic results just by looking at the coefficients is asking for trouble. Much better to graph the implied function over the range of x's that are observed.

    Reply

39. *David* says:

    November 29, 2013 at 2:50 pm
    Following above question, I have another question dealing with the practice of multicollinearity. Say, for example, we come to a refined set of Xs in either cross sectional and panel data models. Yes we find out multicollinearity issues after centering Xs, including points 2 & 3 in your discussion. As a results, we have inflated s.e. and our p values are large. How can we proceed with this scenario given that most of academia especially reviewers will definitely pick out the p values in the output. And how would we explain the results if the p value is inflated and therefore not reliable? Thanks!

    Reply

    Paul Allison says:

    January 22, 2014 at 9:34 am
    Hey, the p-values are what they are. You can explain that high p-values could be a result of multicollinearity, but that's not going to provide the kind of evidence that you'd ideally like to have.

    Reply

40. *Sarah* says:

    December 5, 2013 at 11:38 am
    Paul,

    Thank you for your blog post. I’m hoping you can help with a question that I have about multicollinearity. In my study, I create an industry-level measure for industries that have a common feature. This measure is a dummy variable equal to 1 for industries with that specific feature, and 0 otherwise. The model also includes an interaction with a continuous variable as well as several additional control variables. A reviewer asked why we don’t also include industry fixed effects (dummy variables for each specific industry). Including these industry fixed effects makes the VIF values for our industry measure and the related interaction become quite large. My question is: Do we have a good reason to exclude the industry fixed effects since our primary measure is based on an industry trait and these fixed effects create very large VIFs?

    Thank you kindly for your help.

    Reply

    - *Paul Allison* says:

      January 22, 2014 at 9:24 am
      Well, if you have industry fixed effects, you can’t also include any industry-level variables–that creates perfect collinearity. However, you can have industry fixed effects and also include an interaction between the dummy industry level variable and any other variables that are not at the industry level.

      Reply

41. *Kellie* says:

    January 21, 2014 at 12:48 pm
    Dear Dr. Allison –

    Thank you so much for this post. I work in the field of building predictive models. I typically have maybe a hundred thousand observations and a couple thousand predictors. I generally do not worry too much about high VIF as I am optimizing the final result of prediction and not looking to explain. However, I do try to keep from including terms that are too highly correlated in my automated model fitting procedures because I am worried that I may miss an effect (I still use the p-value as a first cut for inclusion in the model) and I am concerned about the length of time the procedure will run. Due to the large number of observations I typically use a cut off of around .001 or .0005 for model inclusion and I try to keep VIF <20 in my group of candidate variables when running automated model fitting procedures. My question is two fold.

    1. I am seeing colleagues run stepwise procedures where they have a variable transformed in two very similar ways (say 1/4 power and ln) producing VIF in the hundreds. I do not want to give bad advice but I am concerned that there will be issues with the procedure being able to find the optimal variables with such high correlation between the variables. What would you consider a good VIF cut off when building a predictive model?

    2. My second concern is computing time. With SAS proc logistic being single threaded it can run for hours even days on one of my large datasets. Do you know how high VIFs impact proc logistics stepwise model fitting speed?

    Thank you!

    Reply

    - *Paul Allison* says:

      January 22, 2014 at 8:52 am
      These are excellent questions, but I'm afraid I don't have very good answers. Regarding (1), I agree that with two or more transformations of the same variable, it can be difficult to reliably determine which one is optimal. However, I don't have any recommendations for a VIF cutoff. Regarding (2) I really don't know how high VIF's impact processing speed. But my guess is that it wouldn't have much effect.

      Reply
    - *Robert Feyerharm* says:

      January 30, 2014 at 11:16 am
      If you're primarily concerned with prediction, and there may be mutlicollinearity issues, then a penalized regression (ridge regression, lasso, or elastic net) may be the way to go). Stepwise regression techniques aren't terribly reliable.

      Reply

42. *Persaud* says:

January 25, 2014 at 4:15 pm
Hello Dr. Allison
I was wondering whether you have a reference for your recommendation of using a VIF of 2.5 as the cutoff value for addressing collinearity issues?
Thanks

Reply

Paul Allison says:

January 27, 2014 at 11:10 am
That’s my personal cutoff, but you can find a reference to it in my book Multiple Regression: A Primer (Sage 1999)

Reply

Christian says:

May 5, 2014 at 4:55 am
I would be deeply interested in citing this as well, but since I’m writing my thesis abroad I’m having limited access to literature. Could you specify which pages that are dealing with this in the above mentioned book.

Many thanks

Reply

Paul Allison says:

May 8, 2014 at 3:00 pm
p. 141

Reply

43. *Nicole* says:

    January 28, 2014 at 10:59 am
    Dr. Allison,

    I am a novice but this post has been very helpful to me. I do have an situation with a very limited dataset I am analyzing, and hopefully you (or someone else reading this post) can help. I have 2 variables – code count and months (software development durations). There is not good relationship between the two, but there was a good (power function) relationship between code and code/month. I think the code/month becomes a proxy for staff size (which I don't have), and productivity i.e. code/hour per person (which I don't have). Are there collinearity issues or some other issues involved by using code count as both my dependent variable and a factor of my independent variable? Data is very limited unfortunately…

    Reply
44. *Robert Feyerharm* says:

    January 30, 2014 at 11:27 am
    Regarding choice of reference categories, I have several variables in my health care dataset with large "Missing" categories. My analysis dataset was built by linking several datasets from multiple State agencies which do not share the same set of variables. For example, my Employment_status variable contains 23 different categories totaling 56,970 obs., then a Missing category with 183,697 obs. What's the wisest choice for a reference category here – choose the largest of the 23 categories for which I have data, or the Missing category?

    Reply
    - *Paul Allison* says:

      January 30, 2014 at 1:57 pm
      I'd probably go with the largest of the 23 categories.

      Reply
45. *Luckmika Perera* says:

    February 5, 2014 at 8:29 pm
    Thanks Paul! This has been very useful!

    Reply

46. *sivasairam* says:

    February 10, 2014 at 10:01 am
    I have only three variables with two of them having Multicollinearity how can I face this situation?

    Reply

    Paul Allison says:

    February 10, 2014 at 10:09 am
    Depends. How bad is the multicollinearity? Are the two variables that are collinear used only as control variables?

    Reply

47. *Sinie* says:

    February 13, 2014 at 12:28 pm
    Dear Dr. Allison,

    For my dissertation paper, the OLS regression model has 6 independent variables and 3 control variables.

    According to the correlation matrix, one of the control variables (Client Size) is highly correlated with one of the independent variables (Board Size), at 0.7213. However, its VIF is 2.9481.

    When I include "Client Size" in the model, all of my variables are insignificant except Client Size. When it is omitted from the model, certain variables are significant and in accordance to prior literature.

    So does it safe to assume the existence of multicollinearity problem and therefore omit "client size" from the model?

    If so, which reference do you recommend to use VIF of 2.5 as a cut-off value in addressing mullticollinearity problem?

    Reply

    Paul Allison says:

    February 13, 2014 at 4:44 pm
    The fact that client size is collinear with other variables does NOT imply that it's safe to omit it. It may be that client size is what's really driving your dependent variable, and the other variables just happen to be highly correlated with it. This is not an easy problem to resolve.

    Reply

48. *David S* says:

    February 20, 2014 at 8:33 am
    Dear Dr Alisson,

    In the second point of your discussion you have stated that:
    “if your model has x, z, and xz, BOTH x and z are likely to be highly correlated with their product. This is not something to be concerned about..”

    But when there is high correlation between x and xz only, is it than also admissible ignore it and to make use of centering or?

    In my case i am dealing with a X variable that is dummy (0 or 1) and Z is a LN(Assets), al my other variables are not correlated with each other.I hope you can help me with my question, advance thanks!

    Reply

    Paul Allison says:

    February 24, 2014 at 10:46 am
    Yes, “when there is high correlation between x and xz only” it is OK to ignore it. If you want to center reduce the VIFs, that’s fine.

    Reply

49. *isomorphismes* says:

    February 26, 2014 at 12:12 pm

    > a high VIF is not a problem and can be safely ignored [if] The variables with high VIFs are control variables, and the variables of interest do not have high VIFs. Here’s the thing about multicollinearity: it’s only a problem for the variables that are collinear.

    Why is this the case?

    Reply

    Paul Allison says:

    February 26, 2014 at 12:45 pm
    If a variable does not have a high VIF, then it’s coefficient estimate is unaffected by collinearity. It’s OK for the other variables to be collinear. Even if their individual coefficients have large standard errors, collectively they still perform the same control function.

    Reply

50. *dorsaf* says:

February 27, 2014 at 12:41 pm
Dear sir,

Thak you very much for these usefull comments. I want to ask you if there is problem if i include an interaction term (x*y) xith x and y being higly correlated.

Thank you

Reply

Paul Allison says:

February 28, 2014 at 12:55 pm
There’s nothing intrinsically wrong with this. But if x and y are already highly correlated, it will be very difficult to get reliable estimates of their interaction.

Reply

51. *waqas* says:

March 14, 2014 at 9:35 am
Sir my question is that whether multicollinearity is present in polynomial model or not?
Sir, Gujrati has written in his book that polynomial model do not violate the assumption of no multicollinearity, but the variables will be highly correlated.
one of the reason behind this statement (told by my colleague) is that the relationship between the explanatory variables in polynomial model is not linear. but Gujrati has also written in sources of multicollinearity that adding polynomial terms to a regression model, especially when the range of x variable is small cause multicollinearity.
so still the query is not solved.
I have stuided from (http://www.public.iastate.edu/~alicia/stat328/Model%20diagnostics.pdf)
that
Multicollinearity is a problem in polynomial regression (with terms of
second and higher order): x and x^2
tend to be highly correlated.
A special solution in polynomial models is to use zi = xi − x¯i
instead of just xi. That is, first subtract each predictor from its mean and then use the deviations in the model.
Sir plz guide me that which opinion is write

Reply

Paul Allison says:

March 16, 2014 at 10:05 am
Multicollinearity is generally not a problem when estimating polynomial functions. That’s because you are not really interested in the effect of, say, x controlling for x-squared, but rather estimating the entire curve. Centering the variables reduces the apparent multicollinearity, but it doesn’t really affect the model that you’re estimating.

Reply

waqas says:

March 16, 2014 at 5:27 pm
whether multicollinearity is present in polynomial model or not?
if yes then why and if no then why?

Reply

Paul Allison says:

March 16, 2014 at 8:10 pm
Present but not a problem.

Reply

- *waqas* says:

  March 18, 2014 at 3:37 pm
  Sir as multicollinearity occurs when there is exact linear relationship between explanatory variables then how can i defend the statement that there is linear relationship between the x-variables in polynomial model, because variables are non linear in it?

- *Paul Allison* says:

  March 19, 2014 at 9:47 am
  Multicollinearity doesn’t have to be exact. Even if variables are non-linearly related, they can have a high linear correlation.

52. *Lisa* says:

    March 14, 2014 at 11:44 am
    I have 3 variables: lnGDP, lnPopulation and lnGDP per capita. They all have a high VIF. But the figures for GDP, population and GDP per capita are actually the product of 2 countries' GDP, population and GDP per capita respectively. Can I ignore multicollinearity then?

    Reply

    Paul Allison says:

    March 16, 2014 at 9:59 am
    I would say no.

    Reply

53. *Jane* says:

    March 18, 2014 at 2:28 am
    If my variable with the high VIF is correlated with another variable, through a division, does situation number 2 apply in this case?

    Reply

    Paul Allison says:

    March 18, 2014 at 10:00 am
    Probably not. If your goal is to get estimates of the effect of each variable controlling for the other, situation 2 does not apply. It's only when you are trying to estimate the JOINT effect of a set of variables that you can ignore multicollinearity.

    Reply

54. *Samar* says:

    March 26, 2014 at 4:36 pm
    This is a very interesting topic. Thank you so much for this great discussion.

    What if the correlation between the independent variable of interest and the controlling variable is greater than 0.6 and once you add the control your main effect turned to not significant? This is the case for adiposity measures such as BMI (control) and volume of visceral adipose tissue (main independent variable)?

    Thank you,

    Reply

    Paul Allison says:

    March 27, 2014 at 4:29 pm
    Is BMI significant?

    Reply

    Samar says:

    January 12, 2016 at 6:47 pm
    Yes BMI is significant. If you dichotomize BMI, associations will stay significant between the main independent variable (visceral adipose tissue)and the outcome, however if you adjust for BMI as a continuous variable the main associations turned to not significant.

    Reply

    Paul Allison says:

    January 25, 2016 at 10:53 am
    Then I would conclude that there is no evidence for the effect of your main independent variable, once you control for BMI.

    Reply

55. *Sean* says:

April 1, 2014 at 4:08 am
Thanks for this post Professor Allison.

I am using pooled cross section data in my paper, and in order to fix auto-correlation I run prais-wintein regression. Do you know if it is necessary to calculate VIF based on the prais-winstein regression?

I use stata and there is no direct way to find VIF after controlling for auto-correlation. I use the command “vif, uncentered” and received several VIFs above 40, the values of which would really invalidate my analysis and seems too high to be correct. I found a post here:

http://stackoverflow.com/questions/20281055/test-for-multicollinearity-in-panel-data-r

that suggests VIF is not needed in structure such as times series or panel. I would sincerely appreciate it if you could share your insight on this.

Thank you,

Reply

Paul Allison says:

April 2, 2014 at 9:18 am
Multicollinearity is a potential problem with ANY kind of regression. But in most software packages, collinearity diagnostics are only available for linear regression. That’s OK for most purposes, however. But if you’re using the vif command in Stata, I would NOT use the VIF option. Things will look much better if you don’t use that.

Reply

56. *Pankaj Vashisht* says:

April 2, 2014 at 2:12 am
Dear Sir,
Thanks a lot for writing this wonderful blog. I have a small question. I have been estimating a model with export and fdi as explanatory variables. I want to use the product of these two variables. As you have mentioned the FDI and it products are highly correlated. Pluse when i am including the product of FDI with export, the efficiency of FDI coefficient increase. Can i go ahead with my model and include the FDI and it product with export in same model, despite they being highly correlated.

Reply

Paul Allison says:

April 2, 2014 at 9:14 am
Yes.

Reply

57. *natasha* says:

April 25, 2014 at 1:54 am
Dear Dr.
I have VIF= -0.0356 and -0.19415 in a linear regression model
shall i ignore the VIF and multicolinearity in this case, since it is linear model??

thanks in advance

Reply

Paul Allison says:

April 25, 2014 at 9:48 am
VIFs are always greater than 1. So I don’t know what diagnostics you are reporting here.

Reply

58. *Eric* says:

April 28, 2014 at 2:41 pm
Dr. Allison,

I am running a hurdle model on zero-inflated species-level abundances. I have VIFs >20 between one of my predictor variables and the intercept of the count model. I have never had this happen before, and was wondering if you had any suggestions.

I've read several of your columns and found them very helpful. Thanks!

Eric

Reply

Paul Allison says:

April 28, 2014 at 2:53 pm
Sorry but this is not something I've encountered before. What software produced these VIFs?

Reply

Eric says:

April 28, 2014 at 3:56 pm
I am using package 'car' in R. For my original analysis I used generalized linear models to predict species richness (specifying poisson distribution) using the same predictor set. VIFs for these models were all <2. I then chose individual taxa to predict with zero-inflated models and I am getting the high VIFs. I can't really find any information on this issue. Normally, VIF tests don't give you results for the intercept, so I'm not sure how much it really matters in this case. The results of this model corroborate those of my original poisson models, so it may not be too much of an issue.

Thank you for your time.

Eric

Reply

Paul Allison says:

April 28, 2014 at 4:02 pm
I'd go with the VIFs for the GLM.

Reply

59. *Gayathri* says:

    April 29, 2014 at 10:29 pm
    Sir,

    I'm doing a research where gender is a moderator variable. I have found that p-value for gender is 0.4517 (greater than 0.05). Can i proceed to test the moderating effects?

    Reply

    Paul Allison says:

    April 30, 2014 at 7:29 am
    Yes. It's quite possible for a moderating variable to have no "main effect" and yet have a significant interaction with some other variable.

    Reply

60. *Annelore Janssens* says:

    May 2, 2014 at 4:22 pm
    Dear Mr. Allison,

    Let's say:
    dependent variable: y
    independent variables: a, b, c, d

    When introducing an interaction variable "ab", do the separate variables "a" and "b" need to be included in the regression? Or can you measure the effect of the interaction variable solely (and then the problem of the high VIFs doesn't appear i suppose)? So the model becomes y = c+d+ab instead of y = a+b+c+d+ab

    Thanks in advance from Belgium!

    Reply

    Paul Allison says:

    May 8, 2014 at 2:55 pm
    No, you really need a and b separately in the model. Otherwise, you force the effect of a to be 0 when b=0, and vice versa.

    Reply

61. *Christine rewolinski* says:

May 15, 2014 at 9:59 am

I conducted a principle components analysis with oblique oblimin rotation. A 5-component solution was the best result, chose oblique because the items are related. I have four items with multicollinearity, non loading values of .43 or less. My purpose was to reduce a data set, not predict. Other than removing these items, what else can I do. I am not sure how the discussion of regression pertains to my issue

Reply

Paul Allison says:

May 23, 2014 at 7:46 am

Collinearity may not be a problem in this kind of application.

Reply

62. *Albert* says:

May 23, 2014 at 9:40 am
Dear Mr. Allison,

When one of the explanatory variables is the lagged dependent variable,

Yt = alpha1 + alpha2*Yt-1 + alpha3*Xt

and correlation between Xt and Yt-1 is very high (=~0.9).

This kind of multicollinearity is such a big deal? How should I deal with this kind of situation?

Reply

Paul Allison says:

May 23, 2014 at 10:01 am
I'd say this is a big cause for concern. But I can't make any recommendations without knowing a lot more about your data and your objectives. Unfortunately, I just don't have the time to review that in this kind of forum.

Reply

Albert says:

May 23, 2014 at 1:21 pm
Actually, I'm not facing this problem in any work. That was just a general doubt that arised when I was reading your post.

I think that this kind of situation may be very common in time series analysis. If the goal is to estimate a causal relation between Yt and Xt, but Yt is a variable with high persistence (big alpha2) and Yt-1 has a high correlation with Xt.

Thank you for the reply Mr. Allison.

Reply

63. *Eva* says:

    May 26, 2014 at 6:48 am
    Hi Paul,
    Thanks for this helpful article.

    I am fitting two glms from the same dataset with different response variables. I am confused because when I calculate the VIFs, I get different results for each model, despite the fact that explanatory variables are exactly the same. The dataset is complete so there are no missing values. If I understand it correctly and VIF is calculated only from regression of the explanatory variables, can you explain why I get different results?

    Cheers,
    Eva

    Reply

    - *Paul Allison* says:

      May 26, 2014 at 10:56 am
      This is indeed puzzling, and I don’t have an immediate answer. Are these linear models? What software are you using?

      Reply

64. *Koray Can Canut* says:

    May 27, 2014 at 11:41 am
    Hi Paul,

    I am using an ordered logit since I have an ordinal variable as DV. I planned to have quadratic variables for 3 IVs. I regressed my variables first without quadratic variables, then with square of demeaned x and lastly with square of just x. The first and second one show very similar results, whereas the coefficient of x1,x2 and x3 diverge a lot in the last regression from the other two. (the quadratic variables have the same coefficients.) In this last regression these 3 coefficients were still significant but the sign was changed. I couldn’t interpret this situation.

    Reply

    - *Paul Allison* says:

      May 27, 2014 at 12:08 pm
      That’s not surprising. When you include x-squared, the coefficient of x alone depends greatly on the zero point of x. I wouldn’t worry about it. The important thing is the implied curve which is invariant to the zero point.

      Reply

65. *Hamna* says:

June 4, 2014 at 1:07 pm
Dear Mr. Allison
I want to ask about the Multicollinearity in threshold regression, from your discussion it seems that multi is the problem in linear models so we need not to test Multicollinearity in threshold regression. Am I right?

Reply

Paul Allison says:

June 4, 2014 at 2:44 pm
Multicollinearity can be a problem with any regression method in which one is estimating a linear combination of the coefficients. That includes logistic regression, Cox regression, negative binomial regression, and threshold regression.

Reply

hamna says:

June 5, 2014 at 5:44 am
thanks sir, it means for non-linear relationship we need not to check multi

Reply

Paul Allison says:

June 5, 2014 at 8:12 am
No, just the opposite. In most "nonlinear" regression methods, you are still estimating a linear function of the predictors. So multicollinearity is still a problem.

Reply

66. *Gloria* says:

June 25, 2014 at 1:08 pm
Dear Prof. Allison,

I have 3 independent variables measured during childhood (family conflict at ages 8, 10, and 15) and 1 health-related dependent variable measured in adulthood (all variables are continuous). There is considerable multicollinearity among all three IVs (correlations .65 to .72).

How can I safely analyze this data? All three IVs make some unique contribution to R (which is lost when I average across them).

My research goal is theoretical explanation of potential influences on the health DV and to be able to indicate at which age family conflict has the most potential impact on the DV.

Thanks for this very helpful blog!

Reply

Paul Allison says:

June 25, 2014 at 1:18 pm
Are all three variables statistically significant? If so, then you may be OK. Have you checked the VIFs for this regression? Keep in mind that a large sample size could compensate for multicollinearity.

Reply

Gloria says:

June 25, 2014 at 1:38 pm
The simple correlations are significant but some of the regression coefficients are not. In this particular analysis VIF is about 2.8, but in some similar analyses from this data set VIF can go as high as 6.0. Sample size is about 200

Reply

Paul Allison says:

June 30, 2014 at 1:08 pm
Given this level of collinearity, I think it will be difficult to make strong claims about the effects of your predictors at one age vs. another.

Reply

67. *Lida L. Zhang* says:

June 27, 2014 at 6:34 am
Hi Paul,

Thanks for your interesting and helpful insights about multicollinearity. I have a question about your second comment. You demonstrated the non-influence of multicollinearity on the interpretation of interaction terms with the fact that centering does not change the p-value for xz although it reduces the multicollinearity. However, we know that centering can only remove the nonessential but not the essential multicollinearity. My question is – does essential multicollinearity among x, Z, and xz have any consequence? If not, how to demonstrate it? Thank you!

Reply

Paul Allison says:

July 1, 2014 at 10:10 am
Not sure what you mean by "essential" multicollinearity. It's certainly true that if x and z are highly correlated to begin with, centering them will not remove that multicollinearity. And in that situation, centering on the means will not necessarily bring the VIF for the product xz down to acceptable levels. However, at least in my experience, there exist some numbers that you can center on that will bring the VIF for the product to acceptable levels. Now, I am not advocating doing that in practice, because it really doesn't change the model in any fundamental sense. In particular, the test for the interaction and the predicted y will be the same whether you center or not.

Reply

Lida L. Zhang says:

July 2, 2014 at 5:52 am
Thanks! Cohen, Cohen, West, and Aiken (2003, p. 264) distinguished two types of multicollinearity associated xz. Ones is the amount of correlation produced between x and xz by the nonzero means of x and z (i.e., nonessential multicollinearity), which can be reduced by mean-centering. One is the amount of correlation between x and xz produced by skew in x (i.e., essential multicollinearity), which cannot be reduced by mean-centering. Can centering x on certain numbers (not the means) reduce the amount of correlation between x and xz caused by skew in x?

Reply

Paul Allison says:

July 2, 2014 at 6:26 am
I can't make any general claims about this, but in my experiments with several data sets, there were always centering points at which the VIF for the product term xz would fall well below 2.0.

Reply

68. *Brian* says:

July 2, 2014 at 2:13 am
Hello Professor Allison,

I really appreciate you writing on this subject. I have a question about a series of multiple linear regression analyses I ran on state averages. Multicollinearity is definitely present. In one instance a variable had a VIF of 8.5, even though the standard error was much lower than another instance with a VIF of 4.2 for this variable. So, one question I have is isn't it true that for multicollinearity in a variable to be a serious problem the standard error should go up a lot? The other question I have concerns whether variable interactions can be responsible for a high R^2 and the multicollinearity present in an instance of regression in which no interaction variables are specified. The VIF of 8.5 corresponds to a model with a R^2 of 0.93. I think another omitted variable is causing the multicollinearity, but someone else says the variables are interacting. Any such interactions were not specified, and I think the high R^2 doesn't give much room for them to add to the model much. Specifically, I agree with Seibold and McPhee, 1979, who said, "Interaction effects make no contribution to explained variance in the typical regression equation (not having been specified by the researcher and therefore mathematically excluded), while common effects are nearly always present." Can I exclude interactions as having any role in the multicollinearity and high R^2? I'm also curious about why a VIF of 2.5 stands out for you, rather than 10, as John Neter's Applied Linear Regression Models suggests.

Reply

Paul Allison says:

July 2, 2014 at 6:37 am
1. The VIF tells you how much greater the standard error is compared with what it would be if this variable were uncorrelated with all the others. Just because the VIF is higher in one model than another doesn't mean the standard error should be higher. Changes in the model can dramatically change standard errors.
2. I don't think unspecified interactions are likely to explain high R^2 and multicollinearity.
3. A VIF of 2.5 corresponds to a bivariate correlation of .77 while a VIF of 10 corresponds to a bivariate correlation of .95. In my experience, weird things start to happen when you've got a VIF greater than 2.5. Reasonable people may differ on this, but I think 10 is too high a cut-off.

Reply

69. *Alex* says:

July 2, 2014 at 12:09 pm
Hello Professor Allison

My aim is producing more accurate regression models using Excel in order to forecast calls/units of work (y) better from some version of demand. As it is not a direct output of the data analysis pack, I have ignored VIFs thus far and focussed on finding the strongest drivers, using only 1 or 2 regressors (bank holidays). Given the 2.5 VIF suggestion should I be using all regressors that have an R-squared linked to y above 0.600? Also, given continuous shift changes in my regression line as customers shift to different channels, my population size is small (probably around 30). Does this small sample size affect the use of VIF and the output predictive ability of the regression line?

Reply

Paul Allison says:

July 2, 2014 at 12:21 pm
Hard to answer without more context. As I'll explain in an upcoming post, multicollinearity is often a less serious issue for predictive modeling. On the other hand, the small sample size could make it more important.

Reply

70. *Junaid* says:

July 5, 2014 at 6:57 pm
Hello Professor Allison,
I am conducting a research in which correlation of each independent variable with dependent variable is more than .6 and in some cases it is more than .7 but VIF is less than 3. first I want to ask that is there multicollinearity in my data?
secondly VIF for general linear model is always 1. why this happen? and how can i find multicollinearity for general linear model?

Reply

Paul Allison says:

July 6, 2014 at 9:14 am
Correlations with the dependent variable don't matter. Multicollinearity is all about correlations among the independent variables (although if several variables are highly correlated with the dependent variable, one might expect them to be highly correlated with each other). Do you have multicollinearity? Well, that's always a matter of degree. As I've said repeatedly, I start to get concerned when VIFs are greater than 2.5, but that's simply the point at which I start taking a closer look at what happens when variables are entered and removed from the model. As for your second question, I really don't know what you are talking about. It all depends on the software you are using.

Reply

Junaid says:

July 7, 2014 at 4:14 pm
Sir,
First of all thanks. secondly, Whenever i check VIF for only one one independent and one dependent variable using SPSS result shows VIF=1. how can i check VIF of one independent and one dependent variable.

Reply

Paul Allison says:

July 8, 2014 at 7:18 am
Multicollinearity is about correlations among the independent variables. If you have only one independent variable, there is no issue of collinearity.

Reply

71. *A Das* says:

    August 7, 2014 at 8:13 am
    Sir,
    I am doing a binary logistic regression in the stepwise method. I have selected 2 categorical variables for Block 1 and 2. Only in Block 3 I have selected the covariates of clinical importance. Results show significance for all the covariates selected in Block 3. However the variables selected in Block 1 and 2 show large SE among their categories. Can I ignore them? The categorical variables selected in Block 1 and 2 are used to describe the stratification in the data based on a criteria of clinical importance.
    Thanks

    Reply

    Paul Allison says:

    August 7, 2014 at 8:51 am
    Based on what you have told me, I would say that you can ignore the variables in blocks 1 and 2.

    Reply

72. *Dimitri Putilin* says:

August 10, 2014 at 6:10 pm
Dear Dr. Allison,
I am running a regression where the predictors are:
3 out of 4 indicators representing a nominal variable (leaving the 4th as reference),
four mean-centered and highly correlated (.4-.8) unit factors of a scale,
and their interactions.

I am primarily interested in the simple effects of one of the four unit factors at every one of the four possible reference levels of the other variable.

I get VIFs of 12 to almost 27 for the unit factor of interest when the interaction terms are introduced. Without them, the same VIFs do not exceed 3.

Regressing the unit factor on the other predictors, I see that the interaction terms of the same unit factor with the other indicator variables have the largest standardized coefficients (around .5 to .6), followed by another unit factor (.6).

Thanks for your time and any suggestions!!

Reply

Paul Allison says:

August 11, 2014 at 11:40 am
This is one of those situations where I would not be particularly concerned about the degree of collinearity. It's to be expected that the "main effect" of your unit factor is going to be highly correlated with the interactions. That doesn't invalidate the test for interaction, or the estimated effects of your unit factor at each level of the nominal variable.

Reply

73. *Thomas* says:

August 10, 2014 at 6:49 pm
Hi,

What if you are working in HLM, using composite variables, and two of the variables of interest are highly correlated (e.g., .4-.6.)? HLM doesn’t give VIF statistics.

You could run the regression using regular MLR I suppose, but then you wouldn’t be taking account of how the relationship between each predictor and outcome varies across level-2 units. My inclination is to think that this would have implications for how multicollinearity would affect the data, is that right?

Reply

Paul Allison says:

August 11, 2014 at 2:59 am
To check this, calculate level 2 means of level 1 variables and then do your linear regression at level 2, requesting multicollinearity diagnostics

Reply

74. *Kim* says:

August 22, 2014 at 10:15 am
Hi Dr. Allison,

I would like to assess multicollinearity in a case-control study, where I will be using conditional logistic regression to account for matching between cases and controls. I have one main exposure (3 categories) and many other variables I would like to adjust for.

When I examine multicollinearity, may I do so in the same way as if I were conducting a cohort study? Should I account for the matching when assessing multicollinearity?

Thanks for your help!

Reply

Paul Allison says:

August 25, 2014 at 10:00 am
How many controls for each case?

Reply

Kim says:

August 28, 2014 at 9:46 am
4 controls for each case. Thanks!

Reply

Kim says:

August 28, 2014 at 9:46 am
Or I should say…up to 4 controls for each case. The majority have 4 controls, but some have 1-3 controls per case.

Reply

Paul Allison says:

August 28, 2014 at 9:55 am
OK, here's what I would do. You'll have to decide whether it's worth the effort. For each predictor variable, calculate the cluster-specific mean. Then subtract those means from the original variables to create deviation scores. Estimate a linear regression model (with any dependent variable) and the deviation scores as predictors. Request vif statistics.

75. *Zain* says:

August 22, 2014 at 10:04 pm
can i use vif to check multicollinearity in panel data?

Reply

Paul Allison says:

August 23, 2014 at 4:52 pm

Yes. If you’re doing random effects or gee, just do OLS on the pooled data. If you’re estimating a fixed effects model, it’s a bit trickier. The vifs should be checked for transformed predictors with individual-specific means subtracted.

Reply

Zain says:

August 25, 2014 at 6:11 am
i do not understand how to do it for fixed effexts model. i use eviews

Reply

> *Paul Allison* says:
>
> August 25, 2014 at 9:57 am
> Well, there’s no easy way to do it in eviews or other software. You’ve got to create the mean-deviated variables as described on pp. 17-18 of my book Fixed Effects Regression Models
>
> Reply

- *Zain* says:

  September 1, 2014 at 2:03 am
  Could you explain, why vif can be used directly to model random ang gee?

- *Paul Allison* says:

  September 1, 2014 at 8:07 am
  Because multicollinearity is essentially about correlations among the predictors not about the model being estimated. That said, when doing GEE or random effects, you are actually estimating a model that uses a transformation of the predictors. So ideally, VIF would be done on the transformed predictors. But it’s usually not worth the additional effort of doing that.

- *Zain* says:

  September 1, 2014 at 11:24 am
  why should transform the variables into mean deviated variables for fixed effect model

- *Paul Allison* says:

  September 2, 2014 at 10:16 am
  Because when you estimate a fixed effects model, you are essentially using the predictors as deviations from their cluster means. So vif should be calculated on those variables. It can make a big difference.

76. *Yifan* says:

October 11, 2014 at 4:11 pm
It is of great help for my thesis, many thanks! I use the product to explain causality and not sure if I should identify it as multicollinearity or not. Thank you!

Reply

77. *Valle* says:

October 21, 2014 at 10:37 am
Hi Dr. Allison,

Thank you for your advice on this topic. In the study I am working on, I am examining deviations in the median age at first marriage on several outcomes. To calculate deviation, I subtract the median age at first marriage from the actual marriage age of the respondent. Age at marriage and the deviation measure are highly correlated (0.99). If I include both in the model, the vif = 12.20. Can multicollinearity be ignored in this instance since I would expect the two to be highly correlated (as I used age at marriage to create the deviation measure)? Any advice or suggestions would be greatly appreciated.

Reply

Paul Allison says:

October 22, 2014 at 3:47 pm
If you're just subtracting the same median age for everyone, the new variable should be perfectly correlated with the original one. It wouldn't make sense to put both in the model.

Reply

Valle says:

October 22, 2014 at 10:25 pm
Thank you, Dr. Allison. That was my thought as well, but studies with similar situations included both in the model and I couldn't help to think they would have multicollinearity issues.

Reply

78. *Matt* says:

October 23, 2014 at 8:25 pm
Dr. Allison,

I am currently running a logit model with gender and a five category self reported health variable. When I test for multicollinearity gender gets a VIF of 8.12, no surprise there.

What concerns me is when I include the interactions that produce multicollinearity the predicted ORs for gender dramatically change-no interactions yields OR=1.91; interaction included gives and OR of 3.02. Since these are substantively important changes in a study interested in gender effects this is definitely making me uncomfortable. Is there something I am missing here or a way to determine which estimates to trust?

Reply

Paul Allison says:

October 27, 2014 at 6:50 am
You say no surprise for the VIF of 8.12 for gender, but it's surprising to me. Certainly gender is known to be correlated with self reported health, but not that highly correlated. As for the interaction, hard to comment without more details.

Reply

Matt says:

October 27, 2014 at 12:59 pm
The VIF did not surprise me in this case since it is inflated only after adding the gender x health interaction. The VIF is very low prior to adding the interaction.

As for the interaction- is there any more information I can provide? At the moment is seems that adding interactions with anything other than time to the model produces instability in the coefficients. Right now sometimes the coefficients increase, other times they decrease while standard errors are unsurprisingly inflated. I would be more willing to accept this as variability in gender being accounted for to some extent by the interaction, but the genderXhealth is not often statistically significant and it produces conflicting results depending on how I attempt to model the relationship. So, for instance, because of survey mode changes I've opted to model two waves in one model and two in a second model. Model 1 says womanxpoor health has an OR of 2.0, while the subsequent model says womenxpoor health has an OR of .3. This kind of inconsistency makes me think the results of the Interaction probably shouldn't be trusted because they are introducing some sort of instability to the model.

Reply

Paul Allison says:

November 3, 2014 at 9:41 am
You're results are not surprising, but don't necessarily imply instability. You need to more carefully interpret the interactions and main effects. When there is a product term in the model, the main effect represents the effect of that variable when the other variable has a value of 0. That can be very different than the original main effect. In most cases, when the interaction is not significant, I recommend deleting it in order to avoid difficulties in interpretation.

Reply

79. *George* says:

    November 9, 2014 at 3:45 am
    Hi, Professor Allison,

    Can we include the interaction terms but not the main effects in the models to avoid the multicollinearity problem? Thanks!!

    Reply

    Paul Allison says:

    November 13, 2014 at 3:11 pm
    Definitely not.

    Reply

80. *Madhu G* says:

    November 24, 2014 at 1:23 pm
    I have a few questions in Multicollinearity concept for Logistic Regression. I am dealing with a data where few dummy variables and few numerical variables as independent variables and which leads to the following questions.

    1) To detect multicollinearity in multiple regression we will use VIF. What to do if I am working with Logistic Regression.How to detect muticollinearity among independent variables in Logistic regression?
    2) Also, If I have independent dummy variable then how to find VIF for these type variables. Is there any alter native method?
    4)If I have three categories like 2G, 3G and 4G as a mobile network operator then which category to be considered as reference and why?
    3) Do we need to check muticollinearity first and then subset selection or the other way round?

    Thank you in advance!!!!

    Reply

    Paul Allison says:

    December 1, 2014 at 9:59 am
    1. Run your model as a linear regression and check VIFs.
    2. Nope, same as for other variables.
    4. It doesn’t make a crucial difference, but I’d go with the most common category.
    3. Hard to say. Depends on the situation.

    Reply

81. *Janaka* says:

December 11, 2014 at 4:23 am
Thank you very much for the explanation. Please consider the following issue;

I have included age and age2 in a probit model and the VIF values are very high (50 in both age and age2). However, both age and age2 are highly significant.

Is it OK to continue with the same model? Main concern is about the VIF value (50)…

Reply

Paul Allison says:

December 12, 2014 at 9:52 am
Yes, it's OK to continue with this model. If you'd like to reduce the VIF, try coding age as a deviation from its mean before squaring. It won't change the model in any fundamental way, although it will change the coefficient and p-value for age alone.

Reply

82. *natalia* says:

December 27, 2014 at 7:13 am
Dear prof. Alisson,

I have one quick question regarding the concept of multicollinearity for multinomial logistic regression.
In my data age and tenure with employer (both continuous, though age provided as age bands) correlate at 0.6.
If in stata instead of running mlogit, i run regress and ask for the vif the values for the corresponding coefficients are about 1.6. am in the clear? i find 0.6 to be quite a high value and so am inclined to categorize tenure.

Reply

Paul Allison says:

December 27, 2014 at 7:45 am
1.6 is not bad. I'd feel OK with that. Besides, categorizing tenure is not necessarily going to make things any better.

Reply

natalia says:

December 27, 2014 at 8:35 am
categorizing drops it to 1.1 . i'm just surprised that the 0.6 correlation did not translate into a higher vif

Reply

83. *donchoa* says:

January 5, 2015 at 10:15 pm
Dear prof. Alisson,

I wish run a multiple linear regression with DV= abnormal returns and continuous IV= ROA, board size, firm size, leverage, market-to-book, % independant directors and dummy IV: cash, cross border

I wish to create a interaction term between ROA and % independant directors

My question is: Should centered all continuous (DV + IVs) or just variables included in interaction term

Thank you in advance

Reply

Paul Allison says:

January 15, 2015 at 10:47 am
No need to center all the variables. In fact, it’s not essential to center the variables in the interaction, although that can improve the interpretability of the results.

Reply

84. *Mike* says:

February 3, 2015 at 2:04 am
Dear Dr. Allison,

I have a quick question regarding (2). While we are able to reduce the VIF of x and x^2 by subtracting a constant from x before squaring, (x-a)^2 is just a linear combination of x, x^2 and the intercept. If the variability of (x-a)^2 that is explained by the other regressors is lower than that of x^2, the VIF should go down. But the variability in (x-a)^2 should be reduced as well, and the variance of the coefficient estimate should be unchanged. In this case, isn't the VIF just hidden in the reduced variability of (x-a)^2? We could do a QR factorization for any full rank set of regressors and find some linear combinations of the variables with VIF 1, but we should still have trouble estimating the coefficients of the original regressors if there is multicollinearity in the original data.

I've spent some time thinking about this, but I don't understand why the ability to reduce VIFs by centering implies that collinearity between polynomial terms is not an issue. If you could help with this, it would be greatly appreciated.

Reply

Paul Allison says:

February 3, 2015 at 7:25 am
Interesting point. My argument was that since the high VIF's often found in this situation are a consequence of coding decisions that are essentially arbitrary, we shouldn't be concerned about them. But there is an inherent difficulty in estimating the coefficient of x^2 when you are already "controlling" fof x. Unfortunately, there is no way around that.

Reply

85. *Claire* says:

February 12, 2015 at 12:28 pm
Dear Dr Allison,

I am performing a longitudinal analysis using Generalized Estimating Equations (n=64 cases). The outcome of interest is a binary variable and the predictor variable we are most interested in is a categorical variable with 6 levels (i.e 5 dummy variables). The categorical variable does not have a significant effect alone (borderline insignificant with an alpha cut-off of 0.05). However, it does when an additional numerical variable is included in the model. VIF values are 5.0 for the numerical variable and 2.8, 1.5, 1.4, 2.0, 4.5 and 1.6 for the 5 dummy variables. The reference category has a small number of cases (n=5), but from a clinical/biological perspective this is the one that makes the most sense to use as a reference.

1. Why would it be that the categorical variable has a significant effect when the numerical variable is included, but not without the numerical variable?

2. I understand that the VIF values of 2.8 are indicative of multicolinearity and mean that the standard errors of the estimates are likely to be inflated, but if my understanding is correct, this doesn't necessarily invalidate the significant effects of the predictors. Is it valid to report the model, including the VIF values for each of the predictors, and include a statement about the effect of multicolinearity in reducing the precision of the estimates, especially for the numerical variable with a VIF of 5.0?

Many thanks in advance for your help. Any advice would be greatly appreciated!

Kind regards,
Claire

Reply

Paul Allison says:

February 18, 2015 at 1:09 pm

1. This can easily happen, especially given the degree of collinearity in your data.
2. Yes, but it's important to do a global test for the null hypothesis that all the coefficients for the categorical variable are zero.

Reply

Claire says:

February 20, 2015 at 7:11 am

Many thanks Dr Allison. Just to clarify for (2), do you mean the Wald statistic and p-value for coefficients for each of the categorical variables?

Reply

Paul Allison says:

February 23, 2015 at 11:13 am

No, I mean a single Wald test for the null hypothesis that all the coefficients for the categorical variable are 0. Some packages will report this automatically. For others, there are special commands for doing such a test. Such tests are important because they are invariant to the choice of the reference category.

Reply

86. *Andrea* says:

February 16, 2015 at 11:22 am

Dear Professor Allison,
first of all, thank you for this extremely helpful and insightful blog on the issue. I have a very quick question on multicollinearity in panel (fixed effects) data. As I understand from your previous reply, the VIFs in case of fixed effects should be calculated on the regressor matrix after applying the within transformation. May you please point me to a citation for this statement?
Thanks for your consideration and kind regards,
Andrea

Reply

Paul Allison says:

February 18, 2015 at 1:06 pm

Sorry, but I don't have a citation for that claim. It just makes sense given the nature of fixed effects estimation.

Reply

87. *Grazia* says:

February 26, 2015 at 3:31 pm
Dear Dr Allison,

I have a very high VIF (>50) which is mainly due to the inclusion of year dummies and their interaction with regional dummies. Can I ignore this very high collinearity?
Thank you very much
Grazia

Reply

Paul Allison says:

March 6, 2015 at 7:18 am
Hard to say. If you're only including these variables for control purposes, it may be OK. Or if you're merely testing the existence of interaction (i.e., null hypothesis is that all interactions are 0), that may be OK too. But if you're actually trying to interpret the interactions, it could be problematic. Try changing your reference categories to the ones with the highest frequency counts.

Reply

88. *Lorena* says:

March 9, 2015 at 5:53 am
Dear Dr. Allison, I am using PROC SURVEYLOGISTIC for a clustered stratified sample and IVs are categorical. What would you recommend as the best way to check for collinearity among them? Thank you

Reply

Paul Allison says:

March 10, 2015 at 9:52 am
I would just use PROC REG, possibly with weights.

Reply

89. *Antonis Athanasiadis* says:

March 10, 2015 at 9:18 am
Dear Professor,

First of all, your analysis is very useful for understanding the multicollinearity. However, I have one question which refers to my model. I use dummies with two categories (not three). Should I ignore the high VIF of these variables or not? In addition, what is your opinion about condition indeces for detecting multicollinearity? Some people believe that it is better measure than VIF.

Thank you

Antonis

Reply

Paul Allison says:

March 10, 2015 at 9:56 am
Is the collinearity only among the dummies? Try changing the reference category to one that has more cases. Condition indices can be useful in complex situations, but most of the time, I think VIFs do the job.

Reply

90. *Paul won* says:

March 11, 2015 at 8:32 pm
Hi Professor, I have a question about multicollinearity in the ordinal logistic regression.

I have a cohort binary variable (i.e. 1 = canada, 0 = U.S.). I then have variable with 3 levels called it Smoking(1 = yes, 2= no, 3 = Unknown) which I define as a class variable in the SAS modelling syntax. But for this variable, there is no Unknown for those who have value of 0 in the cohort variable (i.e. there is no Unknown for those who are U.S.), so I have something like this in the bivariate table:

US Canada

Yes 50 723

No 2342 334

Unknown 0 234

I fit an ordinal logistic regression in SAS using the following:

```
proc logistic data=smoking_data desc;
class cohort(ref=”U.S.”) smoking(ref=’no’) / param=ref;
model Overall_score = cohort smoking cohort*smoking /rl;
run;
```

And I received this error message:

“””””””””””

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

cohortU.S.SmokingUnknown = SmokingUnknown
“”””””””””””””””

I know it has to do with there is no observation in the Unknown category of Smoking in the U.S. cohort . But I am not sure how to fix the problem, should I drop the interaction term? Also I am not sure how to show the interaction variable (i.e. cohort*smoking) to be equal to a linear combination of the variable cohort and smoking.

Also my output is like this:

Analysis of Maximum Likelihood Estimates

Parameter DF Estimate Standard
Error Wald
Chi-Square Pr > ChiSq

```
Intercept 6 1 -3.8714 0.0423 8385.7823 <.0001
Intercept 5 1 -2.2556 0.0211 11465.3319 <.0001
Intercept 4 1 -1.6454 0.0171 9209.7246 <.0001
Intercept 3 1 -0.8562 0.0142 3615.6776 <.0001
Intercept 2 1 0.2594 0.0133 377.8203 <.0001
Intercept 1 1 1.7399 0.0174 9982.0962 <.0001
cohort Canada 1 -0.3224 0.0297 117.7935 <.0001
smoking unknown 1 -0.2553 0.0541 22.2341 <.0001
smoking yes 1 0.6449 0.0768 70.4543 <.0001
cohort*smoking Canada unknown 0 0 . . .
cohort*smoking Canada yes 1 -0.1063 0.0921 1.3320 0.2484
```

I am not sure if I can use the output giving from SAS in this case? It seems that SAS still give an estimate of the interaction when Smoking Yes is comparing to Smoking No status, but Smoking Unknown vs Smoking No is set to Zero or set to Missing by SAS. So I am not sure if I should still use these estimate of coefficients from SAS.

Is it Ok if you could give some suggestions?

thank you for your time

Reply

Paul Allison says:

March 12, 2015 at 11:47 am

Well, I think you can just use the results as they are. SAS has removed the coefficient that's not estimable and what's left is OK. But you would probably do just as well by removing the unknowns before doing the estimation (listwise deletion).

Reply

91. *Matilda Estephane* says:

March 19, 2015 at 11:44 pm
Hi Dr Allison,
In your text you spoke about a latent variable for multicollinearity but I am havign difficulties understanding the concept. I wanted to know if you would please expand on the topic. Thank you so much for what you are doing here. I was amazed when I stumbled on this site. This is extraordinary.

Reply

Paul Allison says:

March 20, 2015 at 8:22 am
The latent variable approach is most useful when you have two or more predictor variables that are highly correlated, and you also believe that they are, in some sense, measuring the same underlying construct. For example, you might have two different scales that measure "depression". You then postulate an unobserved, latent variable depression that causally affects each of the two scales. Using specialized software, like LISREL or Mplus, you estimate a model in which the predictor variable in the regression is the single latent variable, rather than the two separate scales. You also get estimates of the effects of the latent variable on the two observed indicators.

Reply

92. *Scott Fiesler* says:

March 20, 2015 at 9:02 am
Dear Dr. Allison,
This article is very helpful, thanks for posting it.
I have a question regarding multicollinearity when using lagged principle components is a simple linear regression. Specifically, I have performed PCA on a basket of market rates and reduced teh basket to 4 PCs. I then regress use the 4 PCs and a 3 period lag for the first two PCs against a time series of bank CD rates and the results look good, but the lag terms of course have high VIFs. Can I consider this similar to your situation #2 above?

Thanks so much!

Reply

Paul Allison says:

March 20, 2015 at 10:59 am
Yes, it's similar. It sounds like your main goal is to develop a model that will enable you to make good predictions, not to test scientific hypotheses. In that setting, multicollinearity is less of an issue. The question then becomes whether each higher order lag contributes significantly to improving your predictive capability. This can be evaluated both with p-values and with measures of predictive power. The high VIFs should not be treated as somehow invalidating your model.

Reply

93. *Ann* says:

March 23, 2015 at 12:13 pm

I am using three lagged variables of dependent variables and three explicit variables, the equation is like this:

y2014 = x2013 x2012 x2011 z2013 z2013 z2012 z2011

where x variable is lagged variable of y and z is also lagged but explicit.

This is financial data of 300+ companies and violates the normality assumption.
I am developing prediction models. companies are my cross section whereas variables are arranged yearwise for lagged behaviour.

X2008 X2007 x2006 x2005 x2004 x2003 z2008 z2007 z2006 z2005 z2004 z2003 a2008 a2007 a2006 a2005 a2004 a2003

On excel sheet and eviews, i arrange , variables in this way along with 300+ companies as across sections , which keeps the sample size=310

However, i have created lagged regression models for lag one, two and three. In such case, the lag one contains 10 regression models, lag two contains 9 and lag three contains 8 regression models which i then average for R-square value for each lag.

When i add explicit lagged variables to the models to make a new model, it show very high VIF of 190 to 250 for a few explicit variables in different models whereas it shows no significant increase in standard error of regression. Besides lag three model reduces SE significantly. For these models, i am not getting homegenity of residuals. For insatnce , in lag 3 model with 6 degrees of freedom, i am getting 8 regression models, 5 of which pass white test and three fail. When i am predicting dependent variable on lagged and explicit lagged variables, what shpuld i do? Should i eliminate variables? I have tried all transformations, tried to remove outliers but the results increase standard errors, reduce R-square, and incraese VIF more if i remove some outliers.

I also observe that p-value of explicit variables and most variables is not significant which may be due to increase in degrees of freedom as i read Tabanick , and Davidson and Meckinson.

The explicit variable has different unit from x variable. I have tried standardizing of variables to get results but results are same even after standardizing while standardizing reduces R-square. Should i standardize or not?

What should i do? Do you think OLS is good when normality of residuals violates, Durbin Watson test is successful, VIF violates in few lag three models due to explicit lagged variable, p-values do not show variables significant most often.Standard error of regression and sum of squared residuals is not high say 0.0045 or less. If i am extrapolating any variable or any outlier, it is increasing autocorrelation,VIF and Standard errors. Please reply to my queries at earliest.

Reply

Paul Allison says:

March 25, 2015 at 11:10 am

I’m really sorry but this is just way more detail than I have time to address. Others are welcome to make comments or suggestions.

Reply

94. *OSY* says:

April 20, 2015 at 2:45 am

Dear Dr. Allison,
I am doing my thesis with logistic regression. Should I check multicollinearity if all my dependent and independent variables are dichotomous variables? There are no continuous variables. Thanks

Reply

Paul Allison says:

April 20, 2015 at 1:17 pm

You can have serious multicollinearity with dichotomous predictors. It’s worth checking.

Reply

95. *Tarik* says:

April 21, 2015 at 11:16 am
Dear Dr. Allison,

I wish run the following multiple linear regression:

Y = a0 + a1 X1+a2 X2+a3 X3+ a4 X1*Z+ a5 X2*Z+ a6 X3*Z+e

My problem is that interaction terms (Xi *Z) are correlated .

My question is: can I orthogonilize interaction terms by regressing: (X2 * Z)=a+b(X1 *Z) and (X3 * Z)=a+b(X1 *Z)? Is there any other method for my case?

Many thanks in advance for your answer.

Best

Reply

Paul Allison says:

April 21, 2015 at 1:51 pm

Orthogonalizing won’t help. Yes, it will reduce the correlation but it will also reduce the variance of the residual so much that the effect on the standard errors will be the same. I don’t have a good solution for this. I notice that your model doesn’t have Z alone. Why not?

Reply

Tarik says:

April 21, 2015 at 2:23 pm
Hi Professor,

Thank you for your answer. The model does not have Z alone by construction. The idea is to express the effects of Xi's on Y as functions of Z. The model can be re-written as:
Y= a0 + (a1 +a4*Z)X1+(a2 +a5*Z)X2+(a3 +a6*Z)X3+e

Ridge regression reduces significantly the VIFs of my coefficients, but I need standard errors to assess the statistical significance of my coefficients….

Is there any other methods that can resolve the problem of multicollinearity and provide accurate standard errors?

Thanks once again.

Reply

Paul Allison says:

April 21, 2015 at 2:34 pm
For a fair test of the interaction, you really ought to have the “main effect” of Z in the model as well. Otherwise, the apparent effects of the interactions could be do the suppressed main effect of Z. I don’t know of any other good methods in this case.

Reply

- *Tarik* says:

  April 21, 2015 at 4:27 pm
  Thank you Professor.

  Just a last question. What if I split my model into three sub-models:

  (1) Y = a0 + a1 X1+a2 X2+a3 X3 + a4 X1*Z+e

  (2) Y = …………………….a5 X2*Z+e

  (3) Y = …………………….a6 X3*Z+e

  And I assess the effect of each interaction separately??

- *Paul Allison* says:

  April 22, 2015 at 9:12 am
  Could be helpful. But I’d still like to see the main effect of Z in these models. And by the way, for the model with all three interactions, it would be useful to test the null hypothesis that all three are 0. This can be accomplished in a variety of ways, depending on your software. This test would be robust to any multicollinearity.

96. *Jae Lee* says:

May 16, 2015 at 4:14 pm
Have a follow-up question for Dr. Allison.

I have five independent variables (A, B, C, D, E), and three moderators (C, D, E) moderating A and B. When I use all these variables after mean-centering A and B, VIF for A and B are 17.5 and 15.5. So I mean-centered C, D, E as well, and ran the same model. VIF for A and B reduced to 2.5 and 3. Is it acceptable to mean center this way?

Reply

Paul Allison says:

May 18, 2015 at 8:54 am
Yes.

Reply

97. *Jhon Snoew* says:

May 19, 2015 at 8:59 am
Hi Dr. Allison,

I plan to try several regression models on my data. I have a large number of explanetory variables (40) and most of them are oridnal. Obviously 40 is alot for regression and I'm looking to reduce the number based on correlations/VIF's. So my question is, should I base the chosen variables of of VIF's at all, and on correlations(spearman)? The model will be used for predictive purposes and not so much for understanding the effects on my response variable.

Reply

Paul Allison says:

May 19, 2015 at 9:33 am
I would probably use some form of stepwise regression to develop the model, rather than relying on bivariate correlations. I wouldn't worry about VIFs until I had some candidate models to choose among. You didn't say anything about sample size. If your sample is large (e.g., 10,000), you could easily manage a model with close to 40 predictors. But if it's small (e.g., 100), you will need to do a lot of pruning.

Reply

Jhon Snoew says:

May 19, 2015 at 10:48 am
Thank you,
My sample size is very large 200.000-500.000 depending on what I decide to do with missing values. The point is I am using R and computationally 40 predicters seems too much.(running a GLM with probit link at the moment). I guess the most reasonable approach is to go over the predictors and find an ecological appropriate subset of variables. This would however correspond to similar effects as going of of correlation, because it is reasonable they correlate and might prove redundant. Wouldn't a stepwise approach already suffer from collinearity starting at step 1, hence converge to an inapproapriate model?

Reply

Paul Allison says:

May 20, 2015 at 9:35 am
If you do forward stepwise (sounds like you can't do backward), on step 1 it will pull out the strongest predictor. Then on step 2, it will select the the predictor that has the smallest p-value when added to the model, thereby taking into account any correlation with the first. And so on for the remaining steps. Especially, given your sample size, I'd want to use a very low p-value as your criterion. If you want your overall Type I error rate to be .05, that leads to a Bonferroni criterion of .05/40=.00125.

Reply

98. *Tino* says:

May 21, 2015 at 5:17 am
i am not really a specialist in this area but i have one question. how can the issue of multicollinearity be addressed when dealing with independent variables between different biological metrices

Reply

99. *Camilo Ortiz* says:

May 25, 2015 at 9:27 pm

In your very clear article, you say that the variance is the square of the standard error. Don’t you mean that it is the square of the standard deviation?

Also, you are clear on what the VIF is telling you, but can you say why having an inflated variance of a predictor is a problem? Thank you so much!

Reply

Paul Allison says:

May 26, 2015 at 6:15 am

Well, the standard error is the standard deviation of the sampling distribution of the coefficient. The variance of the coefficient is, in fact, the square of its standard error. An inflated variance is a problem because it leads to high p-values and wide confidence intervals. Also, it makes the model more sensitive to mis-specification.

Reply

100. *Stefan Kruse* says:

June 1, 2015 at 8:03 am
Dear Mr. Allison,
thank you very much for this enlightening article and the multiplicity of comments! Against this backdrop I'd like to ask another question, which refers to the use of interaction terms and the issue of multicollinearity.

I estimate the following model using OLS with a lagged dependent variable:
logY = B1logX + B2Z + B3logX*Z… + control variables

The coefficients indicate a compensatory effect: B1 & B2 are negative, B3 however positive.
Plotting the marginal effect indicates a negative and significant marginal effect of dy/dx for values of Z below the median of Z (i.e. for almost 60% of the distribution of Z), which decreases in strength (because B3>0). However, for high values of Z (above median) the marginal effect turns positive but insignificant. Focusing only on the range of significant marginal effects, the negative marginal effect seems theoretically plausible.
But since the vif goes literally through the roof for X and XZ>70 I checked mean centering.

After mean centering logX, Z and logXZ (i.e. I first take the logarithm of X and then subtract the mean or take the logarithm of X and generate the cross-product with Z and then subtract the mean), the model is
logY = logXc + Zc + logX*Zc… + control variables
The signs of B1 & B2 are similar, B3 is identical of course and the vif now is <10. However, the marginal effect plot has moved up, so that now already at very low values of the (continuous variable Z) dy/dx is positiv, however, almost never significantly different from zero.

What would you suggest in such a situation? Which of these results would you trust more? As far as I understand a high vif leads to higher standard errors and increases the size of confidence intervals (making it more unlikely to show significant results). So if the SE in the uncentered model are actually overestimated but still lead to significant results, how can the results be even less significant in the centered model?
Is it possible that the log-transformation followed by centering might be a source of bias?

Thank you very much in advance!

Reply

Paul Allison says:

June 2, 2015 at 8:15 am
Mean centering should not change the interpretation of the model, or the significance of an effect at particular values of your variables. It's hard to follow your description of what is happening, but I'm wondering if you just need to take appropriate account of the translation from the original metric to the new metric.

Reply

101. *Philipp* says:

June 9, 2015 at 5:21 pm
Dear Mr. Allison,

in my master thesis I would like to qoute your point two:
“The high VIFs are caused by the inclusion of powers or products of other variables”

Is this statement also mentioned inside your book: “Multiple regression”?

Thanks so much!

Reply

Paul Allison says:

June 10, 2015 at 9:55 am
No, it is not mentioned in the book.

Reply

102. *Jenny* says:

June 17, 2015 at 2:28 pm
Good evening,
I am on the first year of an Open University degree in statistics and calculus. But we've not got very far yet! I'm also doing a six-sigma black belt; and this is where my question stems from.

Because there is a perception in six-sigma that we don't need to understand the mathematics behind the tool – after all "isn't that what MiniTab's for?" and my degree hasn't yet covered it; I get that high VIF could mean I'm measuring the same thing in two different ways and looking for a corrolation (i.e. motor output rpm and conveyor speed etc.) but I'm struggling with the road map I've been given for multiple linear regression; so, in a nutshell:

1. Run linear regression in MiniTab
2. Turn on VIF
3. VIF > 5 is bad because you can't trust the p-value (type 1 and type 2 errors abound!)
4. Use the 4 in 1 plot to graph the individuals
5. Look for unusual observations
a. X = skewing values
b. R = high +/- StDev from the line
6. If all VIF 5 start eliminating one at a time
10. Check R^2 adj, VIF and p-value
11. Add back x factor if R^2 adj goes down
12. Continue until all VIFs are < 5
13. You can now trust the p-value

So, after that very convoluted summary to set the scene. What is the connection between VIF and p-value?

Thanks and regards,
Jenny

Reply

Paul Allison says:

June 25, 2015 at 7:04 am
The standard errors for variables with high VIFs tend to be higher than they would otherwise be. Consequently, for those variables, the p-values tend to be high. Keep in mind, however, that this is only a problem for the variables with high VIFs. It doesn't impugn the rest of the regression.

Reply

103. *Julio* says:

June 22, 2015 at 7:56 pm
Dear Paul,

Thank you so much for taking the time to respond so many questions over the years.
I am running a simple cross sectional estimation using OLS were the variable of interest is an interaction between two variables, one is a country specific characteristic and the other is an industry specific characteristic. The depended variables is US imports by country and industry. I would like to run this estimation including country fixed effects and industry fixed effects. However, this approach produces a high multicollinearity in the interaction term.

Is there any way to reduce the multicollinearity in this case?

Thank you very much for your help in advance.

Reply

Paul Allison says:

June 23, 2015 at 5:02 pm
Interesting problem. You might try centering the variables before multiplying, although this probably wouldn't change the coefficient and its standard error.

Reply

104. *thomas* says:

July 2, 2015 at 7:05 pm
Dear Dr. Allison,

Thank you for a great article. Much has been said by many about how collinearity affects estimation of regression coefficients. My question is how collinearity may impact on prediction of responses which seems less touched. Furthermore, does elimination of collinearity, if successfully done, help with the prediction? Is there a good reference on this topic?

Reply

Paul Allison says:

July 4, 2015 at 3:33 pm
Good question but I'm not familiar with any literature on this topic. Maybe other readers can suggest something.

Reply

105. *kashif* says:

July 8, 2015 at 8:01 pm
what about high p value maximum variability explain and low vif?
confusing scenario

Reply

Paul Allison says:

July 9, 2015 at 6:34 am
Sorry, I don't understand this. You'll have to elaborate.

Reply

106. *Sara* says:

July 9, 2015 at 4:46 am
Dear Dr. Allison,

the info reported here is extremely useful for me. Please, could you suggest me in which book can I find it for report it cited?

Thank you.

Reply

Paul Allison says:

July 9, 2015 at 6:38 am
Paul Allison, Multiple Regression: A Primer (Sage 1999), p. 141.
Jeffrey Wooldridge (2013) Introductory Econometrics, 5th ed., p. 97.

Reply

107. *Shabir hussain* says:

July 26, 2015 at 6:40 am
respected Dr. Allison
will you please tell me what is the acceptable limit of multicolleniarity between two independent variables.

Reply

Paul Allison says:

July 27, 2015 at 2:24 pm
Opinions vary widely on this. I start to get concerned when a VIF is greater than 2.5.

Reply

108. *Steve* says:

August 31, 2015 at 10:48 am
Professor Allison,

Thanks very much for this helpful information.

Regarding your point that "multicollinearity has no adverse consequences" when it is caused by the inclusion of powers of other variables:

This surprised me, as almost everywhere else I've looked on-line seems to recommend that I should centre my variables in a polynomial regression, as otherwise collinearity will cause problems.

Can I ask why this advice is so common, whereas you would suggest that this collinearity is not an issue?* Just trying to reconcile the two sets of views.

Many thanks for you time,

Steve

* Please note that I am not challenging your conclusions. I've run a simple simulation in Excel, comparing a polynomial regression to one in which the variables are centred, and satisfied myself that this does not affect the other variables' coefficients (or their p-values) or the R2, as you say.

Reply

Paul Allison says:

September 3, 2015 at 4:18 pm
Well, centering can be useful in evaluating the impact of lower-order terms. Maybe that's why it's so commonly recommended.

Reply

109. *Lawal Mohammed* says:

September 6, 2015 at 3:45 am
Good day Dr. Allison,

I have followed your site and your posts are remarkably helpful and insightful to us, your followers.

I am currently working on data analysis where 2 two-way and 1 three-way interaction terms were used. It is also not a very large observation (42) but on panel data structure of 6 cross sections and 7 longitudinal units.

The VIF for the interaction terms turned out quiet high even after centering the data. Should they be ignored?

A paper by Donald F. Burrill, suggested some methods for addressing the observed multicollinearity, does one really need to bother to correct it?

Thank you so much for the good work. keep ’em coming!

Reply

Paul Allison says:

September 25, 2015 at 12:35 pm
I’m going to have to pass on this one. The data structure is just too unusual for me to make any confident suggestion.

Reply

110. *John* says:

September 17, 2015 at 10:49 pm
Dear Dr. Allison,
I am running a regression that has high order interactions (2-, 3-ways). I also need to impute missing values before running this analysis. I am using multiple imputation. It is said that when you do multiple imputation, your model must include all your analytic variables, which means that my imputation model must include all the interaction terms (2-, 3-ways). As you can imagine, the VIFs of my imputation model are out of roof for the terms involved in interactions (no amount of transformations can do anything about this). My concern at this time is the collinearity in the multiple imputation process. If I see these high VIFs (in the 100~500 range), should I still include the interaction terms as a part of the imputation model, even though technically they ought to be? Could I ever make a justification for not including them by saying that my coefficient estimates for interaction terms are downward biased due to exclusion of the interaction terms in MI? (When I imputed without including the interactions and ran the analysis, I still obtained quite a bit of statistical significant coeff. est. for the interaction terms, despite the downward bias). Thank you in advance for any kind of suggestion.

Reply

Paul Allison says:

September 25, 2015 at 12:19 pm
Collinearity for an imputation model is generally much less serious a problem than for the analysis model. What happens if try to do the multiple imputation with the interactions?

Reply

111. *Alex Shenkar* says:

September 30, 2015 at 9:47 am
My team is validating the credit risk default model. We found a few highly collinear variables with VIF= 32000. This means that the standard error for the coefficient of that predictor variable is 178 times as large as it would be if that predictor variable were uncorrelated with the other predictor variables. The variables represent the age of loan and its transformations to account for maturation.

The developers cited your publications saying that it is Ok to ignore collinearity.

Any feedback will be very helpful. Thank you in advance.

Reply

Paul Allison says:

October 2, 2015 at 11:09 am
It’s hard to answer this without a more detailed assessment. What are the transformations?

Reply

Alex Shenkar says:

October 2, 2015 at 3:58 pm
Thank you for your response.

There are total 3 variables involved.

S_age1 – month of books or age of the loan (VIF=97.5).

For 2nd and 3rd variables we are using a standard cubic age splines in order to best approximate each of the non-linear segments defined by the selected knots. The VIF=35679 and VIF=32441 respectively.

The age related variables within the model are defined as follows:

For a given set of N spline knots(k_1,…,k_N), N-1 variables will be created.

The first spline is a linear function of age, 〚Sage〛_1=age.

For 〚Sage〛_i, where i=2,…,N-1, we have:

〚Sage〛_i={g(age-k_(i-1))-([g(age-k_(N-1))(k_N-k_(i-1))/((k_N-k_(N-1)))]-[g(age-k_N)(k_(N-1)-k_(i-1))/((k_N-k_(N-1)))])}/(k_N-k_1)^2

where the function g(.) is defined as:
g(z)=z^3,if z>0
g(z)=0,otherwise

Restricted cubic spline function in terms of independent age spline variables can be written as:

f(age)= β_0+β_1 〚Sage〛_1+β_2 〚Sage〛_2+⋯+β_(k-1) 〚Sage〛_(k-1)

Thanks again for your time!

Reply

Paul Allison says:

October 4, 2015 at 3:43 pm
OK, my question is: do you need something this complicated to represent the nonlinearity. If you have good evidence for that need, then I wouldn't be concerned about the collinearity.

Reply

- *Alex Shenkar* says:

  October 4, 2015 at 4:03 pm
  I did not built that model, but from review perspective complexity was not justified. If that won't change, I assume that your answer is "yes" – collinearity is a concern. Am I reading your response correctly?

  Eternally grateful!

- *Paul Allison* says:

  October 5, 2015 at 12:54 pm
  Given that you're going with this model, I would not worry about the collinearity. Collinearity is primarily a concern when you are trying to separate out the effects of two variables. Here you are trying to model a single variable with multiple terms. My concern is whether you need a model of this complexity to adequately represent the nonlinear effect of the variable.

112. *rebecca* says:

October 8, 2015 at 8:20 pm
Hi Professor Allison,

I am a PhD student working on a sociolinguistic variation research. I used the glm and vif function in R to check if there's multicollinearity issue in my dataset. I used to have binary dependent variable for analysis and worked fine. Now, I have a multinominal dependent variable (5 categorical variables). Initially, I thought glm can only use for binary data, so I created dummy variables to make 5 binary dep var. The results report different VIF values and some showed big multicollinearity problem, but not all. Later, I learnt that glm can deal with multinominal dependent variable. So, I calculated the VIF again and the problem was not as big as some of those in the earlier calculation.

Then, my question is 'can I report the VIF of the later calculation to show the issue of multicollinearity? (instead of reporting 5 different values for the same dependent variable?)' As I'll be using Rbrul to do multiple logistic regression later which require the dependent variable to be binary, so I'll have to use the 5 dummy variables I've created to do the analysis later. Would this be something I need to consider when choosing which VIF to report?

I hope I have described my questions clear enough and I hope it makes sense to you.

THANK YOU VERY MUCH FOR YOUR TIME AND HELP IN ADVANCE.

Reply

Paul Allison says:

October 9, 2015 at 11:34 am
Most logistic regression programs don't even have options to examine multicollinearity. I usually just do it within a linear regression framework. Multicollinearity is almost entirely about correlations among the predictor variables, so characteristics of the dependent variable shouldn't matter much. I'd just go with what you got from the multinomial model.

Reply

113. *Naeem* says:

October 14, 2015 at 2:59 am
Would you be kind to advise whether muticollinearity (VIF > 10) between one of the main effects and its products terms with other main effects is a cause of concern?
In my (three way interaction) model there exists multicollinearity (even after centering and standardization) between one of the main effects (v1) and its two way (v1*v2)and three way (v1*v2*v3) product terms with other main effects. Is it a cause of concern? If so, how can one handle it?

Reply

Paul Allison says:

November 19, 2015 at 9:00 pm
I don’t think there’s any inherent problem here. But it may mean that you have low power to test the three-way interaction.

Reply

114. *Asad Ali* says:

November 24, 2015 at 6:00 pm
Hi everyone. its really nice discussion. I am finalizing my results for my paper. i seem to have multicolinearity problem with my control variables (one of the control variable have VIF=10). can i ignore this or should it be a concern. also advise what can i do if this is a potential problem?

Reply

Paul Allison says:

December 10, 2015 at 4:27 pm
If the principal variables of interest do not have high VIF’s, I wouldn’t be concerned about collinearity among the control variables.

Reply

115. *Joel Castellon* says:

December 7, 2015 at 5:34 am
No upper bound? Look at:

VIF_j \leq k(X^TX) where k(A) is the condition number.

To prove it use: the matrix norm e.g
||A|| = sup||Ax|| (where ||x|| = 1) = \sigma _1
which is the largest singular value of A

Reply

Paul Allison says:

December 10, 2015 at 4:38 pm
I stand corrected.

Reply

116. *Joe* says:

December 16, 2015 at 3:46 pm
Hi Paul,

Thanks for the great article and discussion!
I apologize if this was covered somewhere in the 200+ comments…I did not read them all.

How about this for a scenario when it is ok to ignore high-ish VIF: VIF is in the 3-4 range, for a predictor you want to interpret, but the p value is still "small" (<0.05). Could you state that your coefficient is inflated and thus a conservative estimate of the effect of x on y? Could you still make a statement about the directionality of the effect, even though it is inflated, or is the estimate simply not to be trusted / interpreted?

Thanks!
Joe

Reply

Paul Allison says:

December 16, 2015 at 5:11 pm
Well, it's not the coefficient that's "inflated" but the standard error, which makes it harder to get small p-values. So, in that sense, the result is conservative. However, the thing to be cautious about is that collinearity makes your results more sensitive to specification errors, such as non-linearities or interactions that are not properly specified. So you still need to be more tentative about interpreting results when your predictors of interest have high VIFs. It would be desirable to explore alternative specifications.

Reply

117. *Vikrant Singh* says:

December 21, 2015 at 1:06 pm
Hi Paul,
Thank you for looking at my query. I have gone through one of your research papers “Testing for Interaction in Multiple Regression”. I am running the following regression

1)–Y = a +b*x1+c*x2+d*x1*x2.
When I run the above regression I get all the estimates to be significant.

2)–When I run them independently “Y= a + b*x1, Y= a+c*x2, Y = a+d*X1*X2 ” , none of the X1,X2 or X1*X2 come out to be significant.

3)–After reading your blog and thinking that it might be caused because of multicollinearity I run the regression after centering the variables in interaction term.

Y = a +b*x1+c*x2+d*(x1-xbar)*(x2-xbar).
The interaction variable and X1 are still significant but x2 is not.

4) –I also check by running the regression :
Y = a+ b*(x1-xbar)*(x2-xbar). The centered interaction term is significant.

After running all the steps mentioned above, I feel the regression in step1 might be appropriate one. Can you please guide me whether regression in step is completely wrong or it is valid but should be used with some caution.

Thank you for your time and guidance.

Reply

Paul Allison says:

December 22, 2015 at 9:45 am
I would go with the regression in Step 1. This looks like a “suppressor effect.” Each variable has a positive effect for some ranges of the other variable, and a negative effect under other ranges. The two effects cancel out in the bivariate models.

Reply

118. *Christina Mack* says:

December 23, 2015 at 8:08 am
Dear Dr. Allison and everybody who gives advice in this blog,

i am estimating the trade flows for panel data (10 years):
– the dependent variable is the volume of trade
– the independent variables are GDP, as well as other variables like distance, common language,… and I also control for exporter and importer fixed effects.
Now I also want to include the research question:
How did the trade change in a certain year due to a great economic sanction which hindered trade to important trading partners: I expect that there is a redistribution towards countries that did not pose the sanction and that this redistribution depends on the political affinity (index variable).
Therefore, I have to include an interaction dummy, which consists of three variables:
Year*country group*political affinity
However, I also have to include (if I understand correctly) all variations between these three variables:
– Year
– Political affinity
– Country group
– Year*political affinity
– Year* Country group
– Political affinity *Country group
As you might expect, that there is high correlation between the variables. However, as you said, this generally poses no problem.
Nevertheless, I am especially concerned with the "Year*country group" and the final interaction term "Year*country group*Political affinity".
The problem is that both variable estimated separately support my hypothesis (sig. and positive) (I still include the other "control" variables for the interaction term except the problematic "Year*country group"). If I include both however, there is a very high correlation between these two variable and both loose their significance. The variable "Year*country group" also gets an unexpected sign (negative).
Despite knowing that you generally should not do this, I would like to exclude "Year*country group" because I argue that
– the first variable "Year*country group" explains that there has been a redistribution towards a certain country group in a certain year
– The second variable "Year*country group* Political affinity" exlplains the same, but just adds the aspect of the pattern of redistribution.
Therefore, I am not sure if I have to include the variable "Year*country group, because they basically describe the same, the latter just adds one aspect. This is why I think that including both unnecessarily "takes the significance" of these variables. Can I exclude the variable?
Thank you very much
Best regards
Christina Mack

Reply

Paul Allison says:

December 28, 2015 at 7:12 am

I would not exclude the 2-way interaction. If you exclude it, then the estimate for the 3-way interaction may be picking up what should have been attributed to the 2-way interaction.

Reply

119. *karin* says:

December 28, 2015 at 1:24 pm

Dr. Allison:

I have found collinearity (VIF=11) for LN(diameter of tres) and for LN(height), however the p is 0.0000 and p=0.0085 for each one. Shoul I be concerned for collinearity eventhough the coefficients are significant?? Thanks in advance for your answer

Reply

Paul Allison says:

December 30, 2015 at 2:35 pm

Well, the fact that both are significant is encouraging. What you should be concerned about, however, is the degree to which this result is sensitive to alternative specifications. What if you used height instead of LN(height)? What about other plausible transformations? When the VIF is that high, the picture can change dramatically under different specifications.

Reply

120. *Aman* says:

January 17, 2016 at 7:05 pm

Hi Paul,

I am carrying out regression which involves interactions of the variables and also the quadratic and cubic terms of multiple variables. I am getting values of VIF of the order of 6000-7000. Can I ignore these as square and cubic terms are highly correlated with the original variable.

Reply

Paul Allison says:

January 25, 2016 at 10:57 am

Well, you may be OK but do you really need a model this complex? My concern is that your power to test the interaction may be low with the inclusion of the quadratic and cubic terms.

Reply

121. *Grace* says:

January 23, 2016 at 7:46 am
Is it valid to assess collinearity in a mixed model (with partially cross-classified random factors) by examining the VIF values? It has been suggested to me that VIF is not an appropriate measure to use in a multi-level model as it assumes the errors are independent identically distributed.

Is this the case?

Reply

Paul Allison says:

January 25, 2016 at 11:10 am
I use VIF simply a rough guide to how much collinearity there is among the predictor variables. So it could still be useful in a multi-level situation. However, in that case, it would not have exactly the interpretation as the multiplier of the sampling variance of a coefficient when predictors are uncorrelated.

Reply

122. *Yvonne* says:

February 4, 2016 at 2:37 pm
Hi Paul,

If the VIF is below 2.5 but the tolerance exceeds 0.9, is this a problem? Can tolerance be ignored so long as the VIF is fine?

Thank you

Reply

Paul Allison says:

February 8, 2016 at 8:23 am
Tolerance is just the reciprocal of VIF, so they cannot give inconsistent results. Thus a VIF of 2.5 corresponds to a tolerance of .40. You want the tolerance to be high, not low.

Reply

123. *Magnus* says:

February 9, 2016 at 10:53 am
Hi Paul,

If you specify a regression model with x, x^2, and z as explanatory variables, there is a good chance that the x and x^2 variables will be highly correlated. If this is the case, x and x^2 will get large VIF's, and you write that this can be safely ignored. But is this necessarily true when z is present in the model too, and z is correlated with x and x^2? If not, how to check whether the large VIF's are partly caused by the presence of z in the model?

Thank you!

Reply

Paul Allison says:

February 9, 2016 at 2:23 pm
Typically, what you would expect to find is a high VIF for both x and x^2 and a much lower VIF for z. Then you're OK. If the VIF for z is high also, then z is strongly predicted by x and x^2. That collinearity should not be ignored because z is conceptually distinct from the x and x^2.

Reply

Magnus says:

February 10, 2016 at 2:30 am
Let me rephrase the question. Assume that I have good reasons for adding z^2 to the previous model with x, x^2, and z as explanatory variables. Now, not only x and x^2 get large VIF's, but also z and z^2.

Is this a model in which high VIF values are not a problem and can be safely ignored?

Thank you!

Reply

Paul Allison says:

February 10, 2016 at 10:43 am
Run the model with just x and z (and possibly other variables). If VIFs are low, then you know the high VIFs that occur when you add the squared terms can be ignored. For further checking, run a model with x, x^2, and z, and another model with x, z and z^2. If the variables without the squared terms have low VIFs then you're in good shape.

Reply

124. *Jerry Bediako* says:

March 31, 2016 at 11:56 am
hi Paul,
I have a little bit of a problem. in all instances i have seen the intercept in a multiple regression does not have a vif. could there ever be a situation where the intercept will have a vif? please help.

Reply

Paul Allison says:

April 1, 2016 at 10:45 am
The VIF is based on the R-squared for predicting each predictor variable from all the other predictor variables. But since the intercept is a constant, the R-squared would be 0 and the VIF would be exactly 1 in all cases.

Reply

125. *Ezza* says:

April 2, 2016 at 2:40 am
Dear Sir,
I refer to your point on the below title;-
“The variables with high VIFs are control variables, and the variables of interest do not have high VIFs”

My VIF result is as follow; which my td is my variable of interest and other variables are my control..

Kindly advice what should I do as my td is also high (more than 10)? Should I ignore the multicollinearity problem?

And if I need to drop the variables, which one should I drop;- either based on the highest VIF result or looking at the lowest t-stat in the regression?

TQVM

Variable VIF 1/VIF

tangible 68.53 0.014591
ndts 36.62 0.027307
td 14.57 0.068643
inflation 2.42 0.414004
eco 2.27 0.441046
profit 1.91 0.524752
z 1.37 0.730983
size 1.35 0.743484
gdp 1.17 0.856032
growth 1.10 0.910267

Mean VIF 13.13

Reply

Paul Allison says:

April 4, 2016 at 10:50 am
You certainly can’t ignore this multicollinearity because it affects your main variable of interest, TD. Hard to say what to do about it. You need to think carefully about how this variable is related to the other two variables with high VIFs, TANGIBLE and NDTS. Are they measuring the same thing, or are they conceptually distinct? If the former, then you might be OK with dropping one or both. If the latter, it’s hard to justify dropping them. But experiment and see what happens to the TD coefficient.

Reply

126. *joleen ng* says:

   May 12, 2016 at 5:18 am

   We have generated a five factor structure for a questionnaire called PANSS. We have calculated the five factor scores to predict other outcome. Howerver there is collinearity among the five factor scores resulting in high value like 0.77 in multiple regression model although the factor scores are significantly correlated with the outcome in linear regression. What is the solution for this? Thank you

   Reply

   - *Paul Allison* says:

     May 18, 2016 at 10:45 am

     What are your VIFs?

     Reply

127. *Jimena* says:

June 22, 2016 at 7:58 am
Dear Sir,

I have a probit model where I want to estimate the probability of cesarean section of women who delivered in public and private hospitals.
I have a high correlation, 0.66, between hospital and age (older women are most likely to have a job, they have health insurance and they deliver in private hospitals).
Probit and OLS give me similar results, but I am a bit concerned about this correlation.
When using the regression:

reg ces i.hospital i.conv i.hospital*i.conv age i.cespref

both, age and hospital are significant at 5% and my VIF are 4,30 for hospital and 4,61 for the interaction term (the other ones are near 1).
But, the uncentered VIF are 8 for hospital, 7 for the interaction, and 43 for age.

Can I still interpret the coefficients if they are significant?

I am particularly interested in the interaction term.

Thank you very much in advance

Notes:
hospital: dummy (0 public, 1 private)
conv: dummy (convenient day or not)
cespref: dummy (preference for CS or not)

Reply

Paul Allison says:

July 4, 2016 at 7:41 pm
Hard to say without more investigation. 0.66 is not a terribly high correlation. And I’m not usually concerned about high VIFs for variables involved in an interaction. But a VIF of 43 is very high for age, and it’s not in the interaction.

Reply

128. *to* says:

July 1, 2016 at 10:59 am
Hi, sir

I have a regression model with panel data (fixed effect model). The dependent variable is y, while the independent variables are x, x^2, v, w, and z. But x has high correlation with v, over 0.90. When I check the VIF, x and x^2 have high VIF, more than 10. But w, v, and z have low one, less than 10. What do you think? Can I ignore the multicollinearity?
And i want to ask, should non-multicollinearity assumption be tested/fulfilled in panel data? Or no need to do?
Thanks.

Reply

Paul Allison says:

July 4, 2016 at 7:49 pm
Well, the VIF for v has to be close to 10, and that's enough to be concerned. I don't think this should be ignored.

With panel data, you don't have to worry about the over-time correlation for the predictors. And high over-time correlations for the dependent variable can lead to problems, but not quite the same as for a simple regression model.

Reply

129. *Yann le Polain* says:

August 5, 2016 at 8:31 pm
Dear Prof. Allison,

I am working with a time and individual fixed-effects model of agricultural area at county level vs. controls and a dummy representing a group of counties (an ecoregion) after 2006, for which I believe there is a "structural shift" after this date. So I'd like to see if this dummy is significant, indicating that there was something qualitatively different for this group after 2006 that is not explained by the fixed effects and control variables. I am using STATA xtreg for this.

I include time trends for different groups of counties, including the group of interest, specified as year*group_dummy. these trends have significant collinearity with the dummy however (high VIF for both the trend and the variable), and I am not sure whether I should take that into account and remove the trend variables, or whether I should consider that this is normal and expected, and calculate the VIF on the other variables but not the trends. It seems almost inevitable that there would be some multicollinearity when including such variables.

Thanks in advance for your help!

Reply

Paul Allison says:

September 21, 2016 at 9:44 am
Yeah, I wouldn't worry about the multicollinearity in this case.

Reply

130. *Michael Kreminski* says:

September 8, 2016 at 3:37 am
Sir,

I have a regression model with 3 IVs; strength of identity (continuous), corptype (binary) and Sense of Connectedness (binary). There is also an interaction term of strength of identity and corptype. When I ran this model the VIFs were:

Strength of Identity: 2.622
Corptype: 6.083
Sense of Connectedness: 1.028
Corptype*Strength of Identity: 7.976

I note earlier you mentioned that when the IVs are binary, if their interactions have a higher VIF score, it is not that significant. Also, when I take out this interaction term, none of the VIFs in my IVs exceed 1.02. Therefor should I be concerned about multicollinearity (and therefor centering my variables)?

Thankyou in advance.

Michael

Reply

Paul Allison says:

September 21, 2016 at 9:35 am
I would not be concerned about this multicollinearity. It is not essential to center the variables.

Reply

131. *Teamchy* says:

October 8, 2016 at 9:10 pm
Dear Prof. Allison,

I am working on a regression with three independent key variables, and let's call them a, b, and c, which are used to predict z, the dependent variable.

In the baseline model, the vif of these three variables is relatively low, ranging from 1 to 4.

However, in the interaction model, when I add the interaction between a and b, and a and c, the vif of the interaction terms increase to over 10.

By the way, the correlation between b and c is 0.7.

Could this be a problem?

Thank you for your attention!

Best,
Teamchy

Reply

Paul Allison says:

November 7, 2016 at 9:41 am
The interaction is probably not a problem.

Reply

132. *Kwaku Adu Agyei* says:

February 13, 2017 at 2:31 am
Paul, thank you for this insight. However I have some few concerns. I predicted an inverse mills ratio after estimating a multinomial logistic regression model which I plugged it in my regression model (using the BFG approach). Unfortunately the four mills ratio had high VIFs (over hundred). I cannot simply remove them because it will further tell me whether to use ESR or PSM to estimate the impact of adoption on income based on its significance level. What can I do to address this problem?

Thanks

Reply

Paul Allison says:

February 14, 2017 at 12:59 pm
The Mills ratios are only included to adjust for bias. If they are highly correlated with each other, that shouldn't be a problem. I would only be concerned if the variables of primary interest had high VIFs.

Reply

133. *George* says:

March 9, 2017 at 3:52 am
Dear Prof. Allison,
I have count panel data and I am going t use xtpoisson or xtnbreg in Stata. Do I still need to check for multicollinearity according to your analysis? If I am not wrong, Poisson and negative binomial models belong to generalised linear models family.

Reply

Paul Allison says:

March 9, 2017 at 8:20 am
Multicollinearity is just as potentially serious in these models as with standard linear models.

Reply

134. *Roger* says:

March 22, 2017 at 12:51 pm
Dear Dr. Allison,
May I ask you a question?
My model is y=ax1+bx2+cx1*x2， I concern about the coefficients a and c, and the result is in line with my expectations, that is, a is insignificant, c is significant, but I found the VIF of the x1 and x1 * x2 are between 5 and 6, I worry about whether there is a collinearity. I know that collinearity does not affect the significance of the interaction, but the coefficient of the lower order term is also my concern.Although after“centering”the VIF are <2. However, the meaning of the coefficient of the lower order term has changed,for example， after “centering”， the meaning of the x1's coefficient becomes when x2=mean x2 the x1's effect to the y.However it makes no sense for my model to discuss the situation when x2=mean x2.It makes sense when x2=0.So I am in a dilemma, no centering, and the coefficient of the lower order term encounters a co-linear.If centering, coefficient of low-term is without practical meaning.
Thanks

Reply

Paul Allison says:

March 22, 2017 at 4:07 pm
I think you’re OK. The one consequence of the high VIF is that the standard error for the main effect will be large, which means a wider confidence interval and a lower p-value.

Reply

Roger says:

March 23, 2017 at 8:34 am
Is it ok to ignore multiple collinearity problems if the lower term and the interaction are all significant?

Reply

Paul Allison says:

March 23, 2017 at 10:48 am
Yes.

Reply

135. *Vikas Rai Bhatnagar* says:

April 19, 2017 at 7:34 am
Dear Paul,
Is multicollinearity a concern while developing a new scale? Should one be cautious that there is no multicollinearity during the EFA stage?
Your insights will be deeply valued.
Regards,
Vikas.

Reply

Paul Allison says:

April 24, 2017 at 4:32 pm
For scale development, you typically want high correlations among the potential items for the scale.

Reply

136. *Carlo* says:

May 18, 2017 at 3:44 pm
Dear Paul,

I have four main explanatory variables (all ratios) which I derive from a categorical variable with four levels. (These variables are computed in GIS software and represent the share (0%-100%) of an ethnic group's common border with other ethnic groups holding certain characteristics divided by the total border length of the observed group).

In the regression model I leave out the "base category" of these four variables (otherwise it's dropped automatically due to collinearity). The VIFs are high up to 8, but when I manually change the base category the VIFs go down to 1.2. I decided though to keep the initial base category because it's more intuitive and the variables are stat. significant (the hight VIFs therefore remain).

When I found the high VIFs I somehow related to your point on categorical variables with more than three levels and thought that the way I did it was alright. However, the more I think about it I doubt whether including all but one of these ratio-variables is "statistically valid." The reason is that the base category which I left out is 0 in 90% of the observations. Practically then the three variables included do most of the explanatory work and in most of the cases the cover every category in the observed data. However Stata does not seem to have problem with that. I am not sure whether I should have a problem with it.

Not sure, if I got my point across, but I would really appreciate your opinion.

Kind regards,
Carlo

Reply

Paul Allison says:

June 5, 2017 at 12:21 pm
This is related to the categorical variable situation that I described in my post. I think what you've done is OK.

Reply

137. *David* says:

July 12, 2017 at 9:51 am
sir, just want to know, if i have census data (population data), do i need to go for assumptions like normality homocedasticity or multicollinearity in multivariate regression analysis?

Reply

Paul Allison says:

July 17, 2017 at 11:14 am
In my opinion, these assumptions are relevant even if you have complete population data.

Reply

138. *Megan* says:

September 13, 2017 at 10:24 pm
In addition to VIFs, I am looking at Condition Indices to identify multicollinearity. If a condition index is high (say, >30) but the only parameters with high variance proportions (say, >0.6) is one variable and the intercept, is this a problem? All VIFs are < 2.
I haven't found the importance/relevance (or otherwise) of the intercept in these condition indices discussed. Any one have any input? Thanks.

Reply

Paul Allison says:

September 19, 2017 at 2:27 pm
Shouldn’t be a problem.

Reply

139. *JEREMIAH OMWOYO* says:

October 28, 2017 at 8:04 am
Thanks so much Allison for this information. I wish to ask if two variables have a strong negative correlation say -0.9 do we say there is multicollinearity?

Reply

Paul Allison says:

November 17, 2017 at 8:08 am
I would say that.

Reply

140. *Skylar* says:

November 21, 2017 at 6:46 pm
Dear Dr.Allison

Thank you for your helpful posts and suggestions on this blog! Hope you can find my question pretty soon

I'm using a discrete-time logistic regression model for my (unbalanced) longitudinal/panel data analysis – having 800 organizations observed for a decade

Excepting for a binary DV, I'm checking VIF for potential multi-collinearity concerns. In my case, should I check the VIF including all the independent variables (no interaction term) and 'year dummies'? OR is it just among the independent variables without considering year dummies?

For example, in stata, using -collin- diagnostic command, should I include IV1, IV2, IV3, … Year dummy 2, 3, 4, …10? OR can I remove the year dummy parts?

I'd really appreciate if you could give any suggestions or let me know any references to search.

Thank you!

Reply

Paul Allison says:

November 24, 2017 at 9:19 am
I'd probably do it with the year dummies. Shouldn't make much difference, however, unless the other variables are strongly correlated with time.

Reply

141. *Lonnie S* says:

November 24, 2017 at 9:06 am

Hi Dr. Allison, Thanks so much for posting this resource! If a set of interval-level variables represents the count of events within each category of a mutually exhaustive set of categories within aggregated units (each variable reflects count of each dummy), can this be simultaneously included into a regression model or do they need to be entered individually in separate equations? Similarly, would the same principles apply for variables representing the spatial density of these same variables (with spatial autocorrelation handled in the equations)? Thanks!

Best,

Lonnie

Reply

Paul Allison says:

November 24, 2017 at 9:17 am

Not sure that I fully understand this. Could you make it more concrete by way of an example.

Reply

142. *Lonnie S.* says:

November 30, 2017 at 4:24 pm

Let’s say you want to examine the effect of schools on crime in census block groups. These schools are distributed around the city; so, to capture that you take a count of schools within 1/2 mile of each block group. From this you produce 3 variables – one for each ‘type/level’ of school representing the count of that type (elementary, middle, high) within the search radius for each census block. Autocorrelation issues aside, can you safely include all 3 count variables as predictors into an equation to determine what the independent effects of each school type are controlling for the effects of the other types? Or, is this multicollinear. It makes sense to me that you could include a set of dummy variables to measure absence/presence of all types of schools into an equation, but taking a count converts those ‘categories’ to ratio level measurement and I wasn’t sure if the consequence would be problematic multicollinearity.

Reply

Paul Allison says:

February 7, 2018 at 8:52 am

I don’t see any a priori reason why this would produce multicollinearity. In this case, it’s an empirical question. Try it and see.

Reply

143. *Zhi Hui* says:

January 4, 2018 at 3:09 pm
Hi Dr. Allison,

Thanks for this insightful post. I'm having trouble understanding this part of the post:

"What happens is that the correlation between those two indicators gets more negative as the fraction of people in the reference category gets smaller."

How does a smaller fraction of people in the reference category cause the correlation of the other two indicators to become more negative? And why does selecting a variable with a larger fraction of people as the reference category fix that?

Thank you in advance,
ZH

Reply

Paul Allison says:

February 7, 2018 at 8:58 am
Consider a three-category variable, with dummies for categories D1, D2, and D3. Let category 3 be the reference category, so we're just using D1 and D2. In the extreme case where category 3 has zero cases, D1 and D2 will be perfectly correlated–if you're in category 1, you can't be in category 2. If category 3 has just 1 case, the correlation between D1 and D2 will be very high.

Reply

144. *Claire* says:

February 7, 2018 at 5:02 am

Hello Dr. Allison,
Firstly thanks a lot for your posting on the multicollinearity issues. My study deals with a panel dataset, and VIF value for firm size, which is one of the control variables, increases a lot (from 1 to 10) if I add firm dummies in the regression model. In this case, can I ignore the high VIF value of the variable? Other VIF values look fine with the firm dummies. I really wonder the reason behind the increase in VIF after including firm dummies. Is there anything I can do?
Thank you again!

Reply

Paul Allison says:

February 7, 2018 at 9:32 am

If firm size were time-invariant, you'd have PERFECT collinearity if you included the dummies. So, I'm guessing that there is some over-time variability in firm size, but not a lot. I'm also guessing that the standard error for the firm size coefficient goes up a lot when you include the dummies. This is simply a consequence of the fact that most of the variation in size is between firms rather than within firms. Nothing you can do about that.

Reply

145. *Trang* says:

March 1, 2018 at 12:28 am
Hi Dr.Allison,
Thanks a lot for this informative post. I've been working on a survival analysis with mortality as outcome and nutrients as predictors (Cox model). It's common that the correlations among nutrient data are fairly high. I found this statement " Because multicollinearity is all about linear relations among the covariates, it is not necessary to evaluate it within the context of a survival analysis" in a book of you ("Survival Analysis", page 417, 1st paragraph, last sentence).
Would you please explain a bit more on this statement? Are there any statistical simulation examples on this issue? Couldn't we ignore it in survival analysis?
Thank you in advance.

Link of the book is as below https://pdfs.semanticscholar.org/7ca8/af7fe2f4f8aa219dc7a929de9ef7806e99aa.pdf

Reply

Paul Allison says:

April 2, 2018 at 2:44 pm
No, you can't just ignore the multicollinearity. My suggestion was to run a linear regression (with any dependent variable) and check the multicollinearity diagnostics in that setting.

Reply

146. *Jay C* says:

March 14, 2018 at 2:16 am
Dear Dr. Allison,

Thank you for your post. I learned a lot from it.

I am writing this to ask a question about high correlations between an IV1(continuous) and a IV2(binary:developed country1 and developing country0).

When these two have a high correlation, can't I use IV2 as a moderator? I wanted to see if the effect of IV1 on DV has different slopes depending on whether the country is a developed or developing country.

Or should I separate the sample into two groups and run two different regression models (one for the developed and the other for the developing)?

Thank you a lot! I wish you a great day.

Reply

Paul Allison says:

April 2, 2018 at 2:40 pm
You can do it either way.

Reply

147. *Martin* says:

April 3, 2018 at 5:07 pm
Dear Dr Allison
Very interesting post. Could one use VIF factor for logistic regression? I do not see such a command on SAS. ..
Many thanks

Reply

Paul Allison says:

May 25, 2018 at 12:56 pm
No, you would have to use PROC REG. That's OK because VIF is about the correlations among the predictors, not about how they relate to the dependent variable.

Reply

148. *Shay B* says:

May 1, 2018 at 6:04 pm
Thank you so much for this information. I am doing a study with the state level data in the US with the time span of 4 years. I am adding a dummy variable for each state and each year to account for state-level and year-level fixed effects. The dummy variables cause very high VIF values (larger than 10) for my continuous independent variables. Without the dummy variables the VIF values are lower (between 5 and 10). This means that my independent variables are highly correlated. What can I do at this point? I tried Ridge Regression but the problem with Ridge Regression is that it does not give me any p-vale to test for the significance of my variables. I am looking for a suggestion which helps me to overcome the multicollinearity issue and gives me a way for testing the significance of each variable. Thank you.

Reply

Paul Allison says:

May 25, 2018 at 12:31 pm
Sorry but I don't think there are any easy fixes here.

Reply

149. *Robin* says:

June 4, 2018 at 9:19 am
Dear Dr. Allison,

Is it true that multicollinearity affects standard errors of individual variables, but does not affect the model's SSE (and hence not the model significance)?
I have heard that some scientists claim that multicollinearity between IVs can explain variance in the DV, which implies that the R2 of the model can increase.

Many thanks

Reply

Paul Allison says:

June 4, 2018 at 12:22 pm
Adding a variable that is highly collinear with a variable that is already in the model can, indeed, increase the R2 substantially. This can happen when the two variables have a positive correlation but coefficients that are opposite in sign. Or when the two variables have a negative correlation but coefficients of the same sign.

Reply

150. *Sennett* says:

July 18, 2018 at 11:32 am
Dear Dr. Allison,

I am currently doing a hierarchical regression model, where control variables are entered in the first block, key independent predictors and moderator variables in the second block, and one interaction term in the last block.

I found out that one of the moderator variables which is inputted in the second step of the model has VIF of 2.536. Is it something I need to be concerned about? If so, what do I need to do? Thank you so much!

Reply

Paul Allison says:

October 2, 2018 at 4:15 pm
I presume you mean mediator variables because moderators are necessarily in interactions. I'd probably be OK with a VIF at that level.

Reply

151. *Mary* says:

November 15, 2018 at 4:08 am
Dear Dr Allison,

I'm running multiple linear regression- Step 1: personality variables; step 2 instructional variables (all scale variables). I have positive correlations, low VIF and tolerance stats but still get negative beta values on one of the instructional variables at step 2. The beta values are insignifcant and in fact adding the variables to the model makes no change in terms of the variance explained. Can I ignore this or is it still valid to report the results, given that the lack of variance is in itself a useful finding? Thank you!

Reply

Paul Allison says:

January 25, 2019 at 12:59 pm
Either way is OK. Your choice.

Reply

152. *Index* says:

November 27, 2018 at 6:50 pm
We are creating an index of neighborhood change in a metro area. We use 7 variables, 3 of which are likely to covary: education, occupation, income. The index is calculated for all the census geographies of the metro area.

Is collinearity a problem in this case?

I realize we're double counting to some extent but we're doing so for every tract. But what if the co-variance is very different in different tracts? What if in some tracts people are educated but blue collar, and in others uneducated but rich? Etc.

Reply

Paul Allison says:

January 25, 2019 at 12:55 pm
High correlations among components of an index are usually not considered to be a problem.

Reply

153. *Marisha Johnson* says:

December 10, 2018 at 2:06 pm
Hello Dr. Allison,

I'm conducting multilevel models in HLM, and I am trying to assess the degree of multicollinearity between a level 1 predictor and it's level 2 counterpart. In other words, I have a variable included at level 1 and a composite of that variable at level 2. The point of the analysis is to see how the estimated influence of that variable at each level simultaneously. Is there a way to check for multicollinearity here? I saw your response to a former comment about calculating the level 2 means of level 1 variables and then doing the linear regression at level 2, requesting multicollinearity diagnostics, but that doesn't work if I'm trying to assess collinearity between the level 1 variable and it's level 2 mean. Thanks!!

Reply

Paul Allison says:

December 28, 2018 at 11:38 am
Well, you're estimating the HLM with level 1 outcomes and level 2 means, right? I would just run that regression with OLS and request the VIFs.

Reply

154. *Srikant* says:

December 28, 2018 at 4:30 pm
Hi Paul,

Is multicollinearity an issue that I need to address (via PCA for example) if I am only interested in building a good predictive model?

I was told recently that unstable parameter estimates in the presence of multicollinearity would result in unstable predictions as well but I have also read elsewhere (blog posts etc) that multicollinearity is not an issue if I am only interested in making predictions.

Is there a standard reference (preferably open access) that discusses these issues?

Thanks
Srikant

Reply

Paul Allison says:

January 25, 2019 at 12:53 pm
I don’t know any good references. But simulations that I have done persuade me that high multicollinearity can produce some increase the standard errors of predictions. However, the effect is not dramatic.

Reply

155. *Jason Aizkalns* says:

January 24, 2019 at 11:13 am
This is a great little summary. I have a question regarding point #3 that I am hoping you can clarify. There is also a question/discussion posted on https://stats.stackexchange.com located here:

https://stats.stackexchange.com/q/388821/31007

Specifically, since the VIF for each level of a factor variable depends on which level is set to the reference, what is the point of looking at by-level VIFs for factor variables? Do we ever eliminate just a particular level of factor and/or re-categorize the levels? I believe looking at the factor's VIF as-a-whole will remain unchanged.

Reply

Paul Allison says:

January 25, 2019 at 12:50 pm
You're correct that the VIF for individual levels of a factor variable depend on the choice of the reference category. And as I note in the post, if the reference category has a small number of cases, that can produce a high VIF. What's really needed is a VIF for the factor as a whole, but most software packages won't produce that.

Reply

156. *Ivy* says:

March 21, 2019 at 9:57 am
I did VIF tests both in construct and item level. The VIFs of every construct are below 3.0, but some VIFs of item are high than 10. Is it serious? How can I deal with this problem? Thanks.

Reply

Paul Allison says:

March 21, 2019 at 11:02 am
I would expect high VIF's at the item level because the items should be highly correlated with those in the same construct. But if you're using the constructs as the predictors, the item-level VIF's are not a problem.

Reply

157. *Rachael* says:

March 25, 2019 at 11:41 am
Hello, I am running a generalized linear mixed effects model with the negative binomial as the family. When I run a vif on the model, I get a very very high VIF value for predictor variable of interest (87). The predictor is a factor variable with 4 levels. I'm wondering if that is why the VIF is so high?

Reply

Paul Allison says:

March 25, 2019 at 12:43 pm
Hard to say without more information. Is the high VIF for just one out of the 3 coefficients? How many cases in the reference category? Are there any other variables with high VIFs?

Reply

158. *Madeleine* says:

April 5, 2019 at 5:04 am
Hi Paul,
I am working on e thesis where I am examining the effect of GDPR (represented by a dummy variable) on a companies revenue. The multiple regression model used to examine this is only based on categorical time variables such as weekday, month, year, holiday, etc (the data used is daily revenue). The result shows a highly significant impact of GDPR on revenue. However, it can be seen that when examining the covariance matrix GDPR and the indicator variable for 2018, which was the year GDPR was introduced, has a simple correlation of -0.81. The VIF between GDPR and year overall is 4.73. Does this mean I can´t trust my results or is there a way to come around this problem?

Thank you in advance!

Reply

Paul Allison says:

April 11, 2019 at 6:19 pm
Sounds to me like what you have, in effect, is a regression discontinuity design. This can be tricky, and there's a whole literature on it. Check it out in Wikipedia, where you'll also find good suggestions for further reading. BTW, Statistical Horizons will be offering a course on this topic in the fall, taught by one of the leading authorities on the topic.

Reply

159. *Sofia* says:

May 10, 2019 at 6:41 am
Dear Professor Paul Alisson,

Thank you very much for your interesting and clear posts. I was struggling because I am testing a moderator effect Y ~ X*M, i.e., Y ~ X + M + X:M, where Y is dichotomous, and X and M are continuous variables. The sample consists of 171 participants and for the product term (X:M), I obtained a p-value of .0532. VIF values were 12.6, 6.8 and 21.9 for X, M, X:M, respectively. So, my first thought was that p-values were being inflated due to multi-collinearity issues. Then, I was struggling because I was not being able to find a way to prove the significance of this moderator.

After reading your post, I centered the variables and, as expected, I obtained the same p-values and no collinearity issues: VIF values decreased to 1.08, 1.05, 1.03! So, thank you so much for sending the "multicollinearity" Ghost away.

However, now I have another problem because I will have to leave with the "borderline p-value" Ghost. I would like to know your wise opinion about borderline p-values. In my opinion, this effect is relevant and deserves to be highlighted. However, I am afraid of future referee recommendations because some weeks ago, a referee told me not to report marginally significant results (p<.1) because "a result is significant or not". Do you agree that the criterion p=.05 must be taken so rigorously in order to decide whether an effect is significant/important/deserves attention or not?

I already read some papers where borderline significance is further assessed with Bayesian tools. In this case, I would obtain a significant interaction with Posterior Probability = .98. Can you please shed some light in the right way to follow?

Thanks in advance!

Reply

Paul Allison says:

May 10, 2019 at 4:25 pm
Well, there's nothing magical about .05. But personally I like to see p-values that are much lower than that. I would say that the evidence for this interaction (moderation) is very weak.

Reply

160. *Cara* says:

May 20, 2019 at 11:26 am
Hi! Does VIF apply to Bayesian models? Thank you for your input.

Reply

Paul Allison says:

May 22, 2019 at 7:58 am
Just like non-Bayesian models. However, there are Bayesian approaches to dealing with multicollinearity that may be somewhat different. See, e.g.,
https://www.tandfonline.com/doi/abs/10.1080/15598608.2011.10483741

Reply

161. *Gergo* says:

August 20, 2019 at 2:14 pm
Dear Paul,

I am working on a meta-analysis and I would like to account somehow for multicollinearity present in the included studies. Especially, as there are some studies with small sample size (e.g. 400-500) but controlling for a lot (7-8) of closely related indicators, which may distort the pooled effect size.

Is there any rule of thumb indicating how may potentially correlated factors at a certain level of aggregation (e.g. neighbourhood) can be balanced with a particular sample size (e.g. 2 factors by n=1000, 3 by n=2000)?

Thank you very much for your help!

Reply

Paul Allison says:

September 5, 2019 at 2:42 pm
Sorry, but I don’t know of any rule of thumb for this.

Reply

162. *L. J. Coombs* says:

September 9, 2019 at 11:26 am
My question concerns point 3.

What if instead of dummy coded predictors you have aggregate data like %married and %never married as your predictors. I usually still exclude a reference group, is this correct? And is collinearity acceptable is this situation?

Reply

Paul Allison says:

October 31, 2019 at 12:44 pm
Yes, it's basically the same situation.

Reply

L. J. Coombs says:

November 15, 2019 at 2:23 am
But in this case, most software does not give you an overall test. Is there a way to get an overall test?

Reply

Paul Allison says:

November 18, 2019 at 7:44 am
Not sure what kind of test you are looking for. You can certainly get the usual collinearity diagnostics. And most software can produce (upon request) a test that all the coefficients for the percentage variables are zero.

Reply

163. *Camille WILLIAMS* says:

December 20, 2019 at 1:17 pm
Thank you for your informative post. You seem to explain that the coefficients of other variables are not influenced by this centering. I have two questions.

To “deal” with multicolinearity, I did Age_Adj = age – mean(age) and Age2_Adj = (age – mean(age))^2 .

Q1: Why do the estimates I obtain for Age & Adj_Age differ between model 1 and 2?

Model 1 – DV~ Age + Age2
Model 2 – DV~ Adj_Age + Adj_Age2

Q2: When I run the following models, the Sex and Age2 estimates across models are identical and the Age estimates are very close.
Model 3 – DV~ Age + Age2 + Sex
Model 4 – DV~ Adj_Age + Adj_Age2 + Sex

However, when I include interactions, only the Age, Sex, and Age * Sex Estimates differ across models. Why is this so?
Model 5 – DV~ Age * Sex + Age2 * Sex
Model 6 – DV~ Adj_Age * Sex + Adj_Age2 * Sex

Please feel free to provide any ressources on the subject. Thank you very much for your time.

Reply

Paul Allison says:

December 20, 2019 at 1:25 pm
Q1 – When you estimate a model with x and x^2, the coefficient of x represents the .5*slope of the curve when x=0. But when you center the variable, you change the zero point.

Q2 – Models 5 and 6 should also include the main effects of each of the variables in the interaction. Make that change and ask me again. .

Reply

Camille WILLIAMS says:

December 22, 2019 at 5:29 am
Hi,

I apologize for the lack of clarity. I wrote my models like you input them in R, but they do include main effects. Here is the complete equation.

Model 5 – DV~ Age + Sex + Age2 + Age * Sex + Age2 * Sex
Model 6 – DV~ Adj_Age + Sex + Adj_Age2 + Adj_Age * Sex + Adj_Age2 * Sex

I find that:
1. The Age estimate in Model 5 differs from the Adj_Age estimate in Model 6.
2. The Sex estimate in Model 5 differs from the Sex estimate in Model 6.
3. The Age * Sex estimate in Model 5 differs from the Adj_Age * Sex estimate in Model 6.
4. All other estimates between Models are identical.

Thank you for your response.

Reply

Paul Allison says:

December 23, 2019 at 9:00 am
All these results are to be expected. The general principle is this: When you have interactions and polynomials as predictors, the highest order terms are invariant to the 0 point of the variables. But all lower order terms will depend on the 0 point of each variable in the higher order terms.

Reply

164. *Rado P Kotorov* says:

January 2, 2020 at 10:36 am
What about lagged variables? For example, you lag the trading price of an asset to predict whether the market will move up or down.

Reply

Paul Allison says:

January 2, 2020 at 2:52 pm
If you have multiple lags, you can easily run into multicollinearity problems, and this is not a problem that can be ignored.

Reply

165. *Felix Bures* says:

January 11, 2020 at 7:09 am
Dear Professor Allison,

Thank you very much for this insightful article. This is by far the most helpful resource I found on the topic of multicollinearity.

I am wondering if case 3 applies in my situation.

My indepdendent variables are three dummy variables presenting the membership in four groups (the control condition is the reference category and is not included as a variable). The N is almost identical for all four groups with 606, 585, 604, and 603. The dependent variable is a count variable (I am using a negative binomial regression model) The aim of my project is to determine the impact of membership in one of the treatment groups on the number of a specific action taken by users in a mobile application.

While I have low VIF values of 1.5, 1.49 and 1.5 for the three dummy variables, I am still concerned about collinearity as the coefficients in my regression model change quite drastically (in size and significance levels but not direction) if I run models with different combinations of the independent variables.

For example: Treatment 1 does not seem to have an effect on the dependent variable, whereas treaments 2 and 3 have a strong positive effect according to my full model (and looking at the despriptive statistics)

If I now run a model including only the Treatment 1 variable, it becomes more negative and highly significant.

While my notion would be to simply base my interpretation on the full model, and report the low VIF factors as proof that there is no collinearity issue, I am not sure if this would be statistically sound (this project is part of a Master Thesis)

Thank you very much in advance for your assistance!

Reply

Paul Allison says:

January 11, 2020 at 10:57 am
This is not a collinearity issue. It’s all about changing the reference category. You should first do an overall test of the three dummies (the null hypothesis is that all three coefficients are zero). If that’s statistically significant, you can then proceed to pairwise tests of the four groups. This can be done with test command in stata or the test statement in SAS.

Reply

166. *charithkrish* says:

March 29, 2020 at 11:54 am
Dear sir
Is it necessary to check multicollinearity for panel data provided the appropriate model is Fixed effects model?

Reply

Paul Allison says:

April 20, 2020 at 1:27 pm
It’s definitely desirable to check for multicollinearity.

Reply

167. *vera_chen* says:

April 28, 2020 at 10:25 am
Thank you for this very insightful article! It offers a really helpful and comprehensive overview of multicollinearity issues.
However, my question concerns multicollinearity in longitudinal research. I would expect a rather high correlation between t1 and t2 of the same variable. But how should I deal with multicollinearity between variables across measurement points, which are not the repeated measures?
Thank you very much!

Reply

Paul Allison says:

May 1, 2020 at 3:51 pm
All depends on the model you’re trying to estimate.

Reply

168. *Anthony Kityo* says:

May 20, 2020 at 3:16 am
Dear Paul,

I am using Proc surveylogistic to estimate odds ratios and CI for the association between IV(4 categories) and a binary outcome. I have 6 categorical covariates.

1. Should I be concerned about multicollinearity in the covariates?

2. You suggested using Proc reg to assess multicollinearity. In this case, what is the dependent variable. In other words how do I build this model?

3.

Reply

Paul Allison says:

May 20, 2020 at 8:26 am
1. Yes
2. The dependent variable doesn’t matter. Just use your dichotomous outcome.

Reply

169. *JL* says:

June 10, 2020 at 10:56 pm
Hi Dr Allison,

Thank you for this interesting and useful resource. I am reviewing a manuscript and the authors have attempted to perform an analysis which I believe is incorrect, due to the perfect collinearity of the two variables of interest.

The authors attempted to examine the association of nutrient A from say bread with a disease outcome, and claim that they wish to take away the effect of nutrient B from bread in their Cox model by including nutrient B as a covariate. My understanding is that this is incorrect (or even impossible) as nutrients A and B are basically derivatives of the intake of bread, which means they have the same variance structure and thus perfectly collinear (R^2 = 1). It is akin to putting the same variable in the model twice. Is my interpretation correct?

Thank you.

Reply

Paul Allison says:

June 11, 2020 at 8:24 am
If these two variables are, indeed, perfectly collinear, then there's no way that you can estimate the effect of one controlling for the other. Most regression programs will just boot one of them out.

Reply

170. *Laurel Watson* says:

June 26, 2020 at 6:52 pm
Thank you for all of you helpful guidance! I have a question for you about multicollinearity among interaction terms. I am conducting a hierarchical linear regression with four steps. Each variable is continuous. In the fourth step of my model, I have entered four interaction terms and there appears to be multicollinearity among two of the interaction terms. In spss, I received an error message indicating that "For models with dependent variable y, some variables have impossible tolerances. The original correlation matrix may not be positive definite. Pairwise deletion may be inappropriate." The fourth step, the step that includes the four different interaction terms is significant, and 2/4 interactions are significant. And, I have mean centered the variables before calculating the interaction terms. I am wondering if this is a situation in which I should be concerned about multicollinearity. Thank you so much for your feedback and guidance–much appreciated!

Reply

Paul Allison says:

July 21, 2020 at 8:23 am
Well, multicollinearity might reduce your power. But I would be inclined to believe the results in your fourth step.

Reply

171. *Christine* says:

July 19, 2020 at 4:23 pm
Hi Dr. Allison,

I have a question regarding point 3.

I have 4 dummies for 5 prime minister terms, and 3 of them have VIFs around 10 (I am already using the largest category as reference). All of them are not significant according to the results of the regression as well as in the results of a Wald test. Are the high VIFs influencing their significance in any way? In other words, are their lack of significance legitimate despite the high vifs?

Thank you so much

Reply

Paul Allison says:

July 21, 2020 at 7:48 am
Are there any other variables that have high VIF? Have you done a joint test that all four have coefficients of 0?

Reply

Christine says:

July 24, 2020 at 10:36 pm
GDP growth and term duration have high VIFs (in the 10+ range). They are all stationary, and when I used their residuals in the regression, their VIFs are still high (10ish) but the VIFs for the Prime minister dummy variables become lower (but still larger than a 5). What do you mean as a joint test? I ran the Wald test by using the test command in Stata for all 4 dummies after the regression; is that sufficient?

Thank you very much

Reply

Paul Allison says:

July 30, 2020 at 12:40 pm
The Wald test should be sufficient. But if the dummy variables are highly correlated with GDP growth and/or term duration, you may have very little power to test your hypothesis.

Reply

172. *Tom* says:

July 28, 2020 at 10:35 am
Dear Paul,

Thank you for this post. I have been reading your post (and a lot of the answers) with great interest.

Do the takeaways from your post also apply to perfect multicollinearity?

I have a situation in which I use a pooled cross section at the firm level, with about 15 countries and 2 time periods. I am using Elections as a dummy-IV. In addition, I am using country-year fixed effects.

The country-year dummies by themselves, and especially combined with the IV cause some perfect multi-collinearity, leading R to remove about 5 country-years (i.e. US 2010, Canada 2007 etc..) and Stata occasionally throws out my IV.

I am wondering how this affects my results, especially my IV of course.

Removing countries does not seem to do any good.

Tom

Reply

Paul Allison says:

July 30, 2020 at 12:33 pm
No, I would not extend my arguments to perfect multicollinearity. What your data are telling you is there is insufficient variation within county-years to estimate the effects of your IVs. Stata will throw out whichever collinear variables come last in the model. So I recommend putting your country-year variables first. If Stata throws out your IVs, then you can’t really estimate their effects in a fixed effects model.

Reply

173. *Brad* says:

September 10, 2020 at 10:07 pm
Hi Paul,

Nice post, and it is amazing that you have continued to keep up with responses for 8 years (to the day) since your original post! Awesome!

I am noticing a number of comments seeking additional clarification about the 1st point you make here, which you provide already in some of your responses. Though you provide several citations in your responses to other comments, I did not catch one for this first point. Is there a favorite reference you have that elaborates on this first point?

Thanks!

Reply

Paul Allison says:

September 22, 2020 at 8:34 am
Wish I could find a good citation, but I haven't been successful.

Reply

174. *Tomer Karni* says:

September 22, 2020 at 5:18 am
Dear Professor Allison,

Thank you very much for this helpful article.

I have a dilemma in my research regarding multicollinearity and I wonder whether the 2nd situation you referred in the article can answer it.
I went through all the comments and did not find a similar question.

I'm using forward stepwise procedure on a generalized linear model analysis (both logistic and poisson) with 16 predictor variables with which I'm trying to explain a binary dependent variable and also a continuous dependent variable, separately.

My multicollinearity issue is with some of my predictors that are rainfall data of a given year (all continuous predictors). Some are rainfall data of different seasons of that year and one predictor is the total rainfall, consisting of the sum of all the others.
So, autumn rainfall (x),winter rainfall (z) and spring rainfall (a) make out total rainfall (x+z+a).
Since x+z+a is not entirely analogical to xz in your article I was not sure if I can refer to the 2nd situation you described. Will it be safe for me to ignore multicollinearity in my analysis?

I would appreciate it greatly if you could spare the time to address this issue.

Thank you so much,

Tomer

Reply

Paul Allison says:

September 22, 2020 at 8:28 am
You have perfect multicollinearity and this cannot be ignored. It's simply not possible to estimate a model with all four of these variables. I would estimate a model with the seasonal rainfall variables. If the coefficients for two or three of these are not significantly different, then combine them.

Reply

Leave a Reply