

Pearson Group TAC #1

Molly Frendo, Chris Sloan, Andrea Zellner

Title: What in the World is VIF?

YouTube link to video: <http://www.youtube.com/watch?v=rUNzRBKje70>

For this installment of the Think Aloud Clip, our group decided that we would like to explore the variance inflation factor (VIF) and the tolerance statistic. We were interested in clarifying when to worry about these numbers and what's considered an acceptable level. The book *Discovering Statistics with SPSS* by Andy Field cited two other textbooks in the discussion of VIF. By most accounts, VIFs around 1 are considered good, whereas a VIF above 10 is of concern. Where, we wondered, did these numbers come from? We tracked down these texts and used them to discern the genesis of this idea about the VIF values.

In order to track down these citations, a trip to the MSU Math Library was in order. (Side note: the MSU Math Library is kind of awesome.) Andrea spent some quality time in the regression section reading indexes to track down the mystery of the VIF cut-off. This led her to a 1970 paper called "Ridge Regression in Practice" from the *American Statistician*. Citation:

Ridge Regression in Practice

Donald W. Marquardt and Ronald D. Snee

The *American Statistician*

Vol. 29, No. 1 (Feb., 1975), pp. 3-20

Published by: [American Statistical Association](http://www.amstat.org/)

Article Stable URL: <http://www.jstor.org/stable/2683673>

From that article: "As multiple correlation of any predictor with the other predictors approaches unity, the corresponding VIF becomes infinite." They also supported the VIF of 1 and the cut-off of 10.

Additionally, it was interesting to read a paper on regression that did not rely on statistical analysis software. The researchers compared VIFs on standardized and unstandardized data and noted that standardizing the data significantly reduced the VIFs. It is unclear how that fits into the theoretical history of regression, but it was certainly intriguing. The other interesting discussion in the article was how the researchers recommended splitting data into an 'estimation' and 'prediction' data. They used the estimation data to build the model and the prediction data to test the model, which apparently is still done today and is a nice way to test the model if you have big enough sample sizes. It was an interesting glimpse into how one might run regressions without computing software: they used to hand draw their graphs.

The concept of Carmen Sandiego fit in with our assignment because, like the show, we were trying to solve the mystery of what the variance inflation factor is and what various VIF levels

indicated about a regression model. The premise of the show that aired in the mid-'80s hinged on solving a mystery that was based on learning geography and there was a popular computer game for kids based on the premise. It was something that all three of us remembered fondly.

And that theme music, now that's hard to get out of your head. Recently, Facebook launched an online version of the game, allowing us to use the power of screen capture software to manipulate the images for our assignment. The solution to our VIF mystery really did take us around the world. Chris lives in Utah; Andrea and Molly live in Michigan. Andy Field, a major source for our inquiry, teaches in England. Finally Andrea's trip to the math library really did take her to a place she never knew existed. We think that we tried to have fun with the topic, but at the forefront of our collaboration was the need to explain the concept clearly and in enough detail. We all learned quite a bit about VIF and multicollinearity. We also enjoyed ourselves too.

For more information on Carmen Sandiego, see

[http://en.wikipedia.org/wiki/Where\\_in\\_the\\_World\\_Is\\_Carmen\\_Sandiego%3F](http://en.wikipedia.org/wiki/Where_in_the_World_Is_Carmen_Sandiego%3F)