Consequences of Multicolinearity in Multiple Regression

Jeremy Meyer

December 13, 2018

1 Introduction

One of the assumptions for multiple linear regression is for the covariates to have no (or low) multicolinearity. Multicolinearity, sometimes called colinearity, happens when two or more predictor variables are highly correlated with each other. This can cause problems such as a reduction in statistical power of the beta tests for significance. A motivating example will be introduced with correlated predictors and the consequences of multicolinearity on the significance of the betas will be explored by a simulation study.

2 Motivating Example

Consider bridge build time data with construction time (in days) as the response. The dataset consists of 45 different bridge completion times with 5 covariates: deck area (ft^2), construction cost ($1000), number of structural drawings, length of bridge (ft) and number of spans. Predicting the time it takes to complete the bridge can be helpful for planning and budgeting purposes. The data was obtained online[1] and all variables were transformed to the log sale to fix the skewness. Consider the following model:

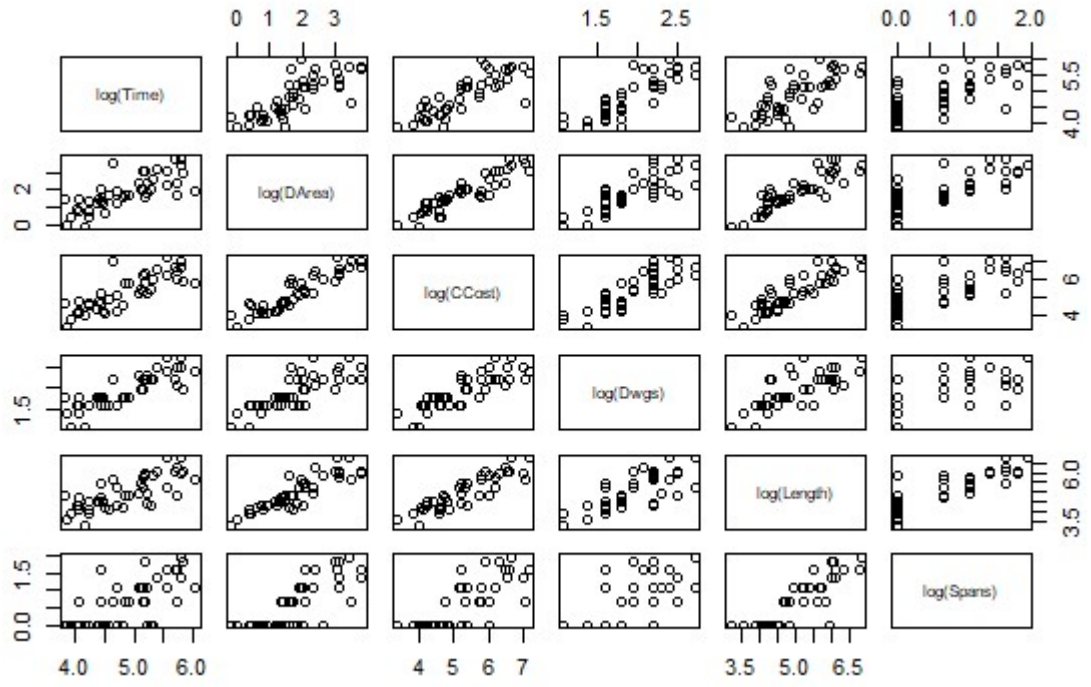
$$\begin{aligned}\log(time) =& \beta_0 + \beta_1 \log(DArea) + \beta_2 \log(CCost) + \beta_3 \log(Dwgs) + \beta_4 \log(Length) + \beta_5 \log(Spans) + \epsilon \\ & where \quad \epsilon \sim N(0, \sigma^2 \mathbf{I})\end{aligned} \tag{1}$$

The first sign of trouble in the model can be seen from the pairs plots (Figure 1 next page) for each of the predictors. Note that most the predictors have a strong association with each other. This can also be seen in the predictor variable correlations (Table 1). All of the correlations are high; in fact, they are all above 0.6!

Table 1: Correlations between predictor variables

Correlation	log(DArea)	log(CCost)	log(Dwgs)	log(Length)	log(Spans)
log(DArea)	1.000	0.909	0.801	0.884	0.782
log(CCost)	0.909	1.000	0.831	0.890	0.775
log(Dwgs)	0.801	0.831	1.000	0.752	0.630
log(Length)	0.884	0.890	0.752	1.000	0.858
log(Spans)	0.782	0.775	0.630	0.858	1.000

Figure 1: Pairs plots for response/predictor variables



Due to the correlated predictors carrying similar information, the model has a hard time determining what factors have a significant impact on `log(Time)`. This can be seen from the results of R's `summary()` function on the model. Note that the F-statistic comparing the full model against an intercept only model is highly significant (1.043e-11), but only one of the covariates is showing significance. Another issue is that 2 of the coefficient estimates are negative, which is clearly not the case from the pairs plots against the response variable (Figure 1).

Figure 2: R `summary()` output

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.28590    0.61926   3.691 0.000681 ***
log(DArea)  -0.04564    0.12675  -0.360 0.720705
log(CCost)   0.19609    0.14445   1.358 0.182426
log(Dwgs)    0.85879    0.22362   3.840 0.000440 ***
log(Length) -0.03844    0.15487  -0.248 0.805296
log(Spans)   0.23119    0.14068   1.643 0.108349
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 0.3139 on 39 degrees of freedom
Multiple R-squared:  0.7762,    Adjusted R-squared:  0.7475
F-statistic: 27.05 on 5 and 39 DF,  p-value: 1.043e-11
```

One reason many predictors can lose their significance is because their variance is inflated. Variance inflation factors (VIFs) measure this and are common diagnostic for checking collinearity.

VIFs are calculated as such:

$$VIF_j = \frac{1}{1 - R_j^2} \tag{2}$$

Where R_j^2 is the coefficient of determination found by fitting the j^{th} predictor against all the other X's. Thus, higher correlations result in a higher VIF. These are calculated for each of the predictor variables. The VIFs using the bridge data are displayed in Table 2.

Table 2: Variance Inflation Factors

log(DArea)	log(CCost)	log(Dwgs)	log(Length)	log(Spans)
7.164	8.483	3.409	8.014	3.878

One way to interpret these is that the variance of `log(Length)` is 8.014 times larger than if the data were orthogonal, or equivalently, $R^2_{log(length)} = 0$. Typically, VIF cutoff values as high as 5 or 10 are used for the assumption of low colinearity to hold. In the bridge data example, the assumption would be violated depending on what cutoff is used. In the simulation study, we will explore different VIF cutoff values and measure the consequences of such on the significance of the beta coefficients.

3 Simulation Study

In this simulation, we will generate new data with VIF levels close to 3, 5, 7.5, 10, 15, 20, and 100. To make our simulation consistent with the motivating example, a new model of 5 predictors will then be fit with the correlated data, but only 1 predictor will be significant. We will then make power curves for the significant beta parameter and analyze the effects of various VIF levels.

3.1 Generating Correlated Data

There are many different ways to generate explanatory data with various VIF levels. To make the simulation simpler and less computational, we will assume that all predictors have the same variance ($\sigma^2 = 1$) and correlation (ρ) with each other. Since there were $n = 45$ observations in our dataset, we will take 45 draws from a multivariate normal of dimension $p = 5$ and mean $\mathbf{0}$. We will call this correlated dataset X and the i^{th} column X_i. Thus, X will have 5 columns (one for each predictor) and 45 rows that correspond to each draw. The multivariate normal model used for the rows in X is shown below:

$$\text{Each row in } X \sim N(\mathbf{0}, \mathbf{V}), \quad \text{where } \mathbf{V} = \begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{bmatrix}$$

Getting the correlation coefficient that yields a specific VIF is not as straightforward as backsolving for R_j in equation (2). Since there are multiple predictors, it has to be computed numerically. This was done by tweaking ρ in the covariance matrix until the mean VIF of

2000 different X samples was close to the desired value. The VIFs between samples of each ρ had a standard deviation of about 20% the mean VIF, so the higher the VIF, the more varied the sample VIFs were. The approximate ρ for VIF = 3, 5, 7.5, 10, 15, 20 and 100 are displayed in Table 3.

Table 3: Corresponding ρ for each VIF

VIF	0	3	5	7.5	10	15	20	100
ρ	.000	.690	.815	.877	.907	.938	.954	.991

After generating the data, consider the following model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon, \quad \epsilon \sim N(0, \sigma^2 \mathbf{I}) \tag{3}$$

To generate Y, we will fix the the true values of $\beta_0, \beta_2, \beta_3, \beta_4, \beta_5$ to zero, change β_1 for the power curve, and plug in the correlated X_is. The model will then be fitted with both the simulated Y and X, and an ANOVA test will be done using the full model against the reduced model without β_1. The resulting F statistic will be evaluated for significance at $\alpha = 0.05$. This process was repeated 1,000 times with different Xs and Ys. The reason new Xs were generated each time was due to the randomness of generating a dataset with a specific VIF.

3.2 Power Curves and other results

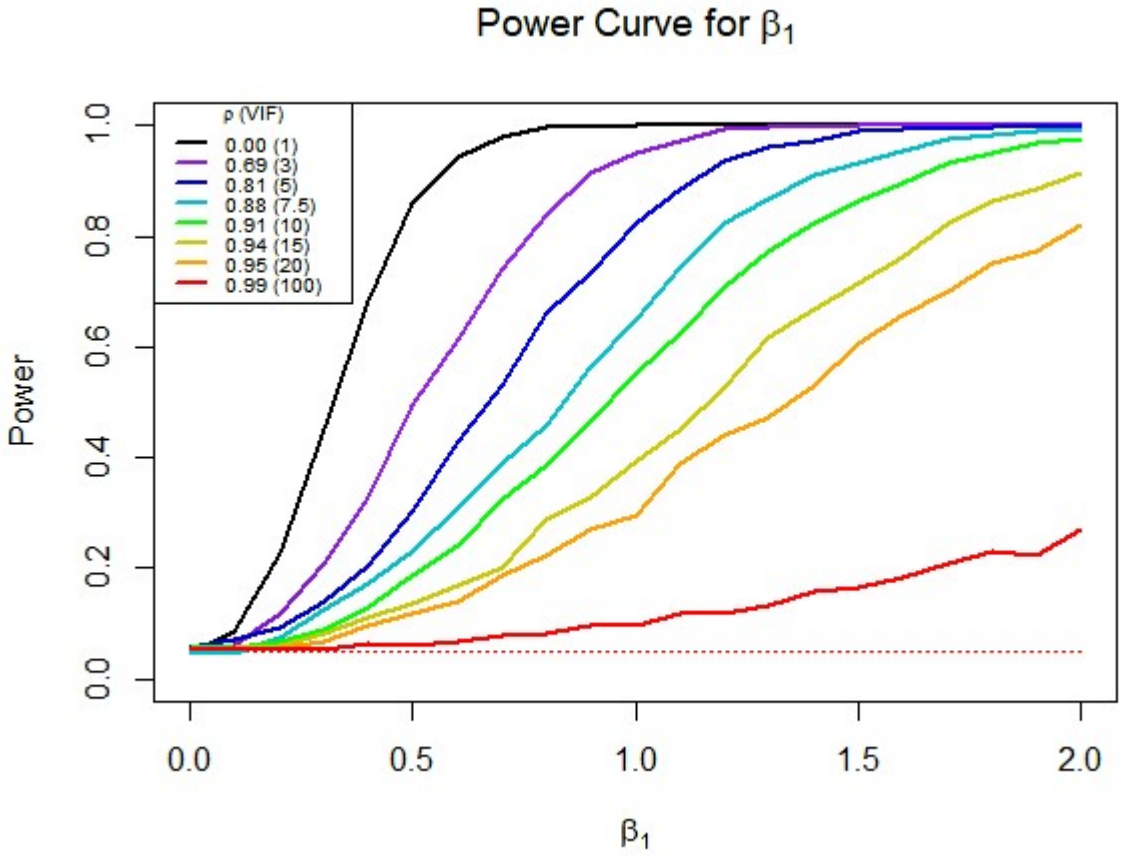
To calculate the power curve for one VIF level, we counted the proportion of models that were significant at many β_1 values. This was done for each VIF level of interest, with multiple lines in the power curve graph corresponding to different VIFs (see next page, Figure 3). While only the positive β_1 values are shown, due to symmetry, the graph is mirrored for negative values of beta.

A table of values for the graph is included below. When $\beta_1 = 0$ or when the null hypothesis is true, the power matches the probability of type 1 error (α) for every value of VIF, which is as expected. In general, every 1 unit increase in VIF has a smaller change on the curve as the VIF gets higher. Having a VIF of 10 or lower makes it possible with these parameters to detect a difference in β_1 of 1 more than half the time, but higher VIFs will really struggle detecting smaller differences.

Table 4: Power for different VIFs (columns) and β_1 (rows):

	1	3	5	7.5	10	15	20	100
0	0.0445	0.0530	0.0570	0.0475	0.0555	0.0530	0.0520	0.0525
0.5	0.8600	0.4935	0.3025	0.2315	0.1865	0.1375	0.1190	0.0600
1	0.9995	0.9510	0.8220	0.6515	0.5530	0.3940	0.2940	0.0965
1.5	1.0000	1.0000	0.9905	0.9330	0.8620	0.7145	0.6070	0.1630
2	1.0000	1.0000	1.0000	0.9945	0.9770	0.9135	0.8205	0.2680

Figure 3: Power curves for the different VIF levels



4 Advice for statistical practice

The effect of colinear predictors on the significance of the betas is clear: they decrease the power of the tests, especially for detecting smaller differences. As stated before, this is because the model has a hard time picking up the significant variable if all the other non-significant predictors carry similar information. As a result, the variances are inflated, which can cause some problems with the estimates and standard errors of the betas.

Non-significant covariates should be examined carefully, as they could very well be significant, but the tests might not have enough power to detect them. One consequence of this is that non-significant estimates can also switch signs, which is something to be mindful of. One way to increase the power is to simply collect more data because that will give the model more information to work with. Another option would be to use AIC/BIC selection techniques to eliminate redundant predictors from the data, although these run the risk of producing biased estimates due to data omission. Additionally, creating a designed experiment with orthogonal predictors gets rid of colinearity altogether, but this option isn't always feasible. However, since the model has a hard time detecting significance, when the betas are significant, there's a good chance the true values are significant.

As for what makes a good cutoff value for VIF, that really depends on what level of certainty is desired for detecting differences in the coefficients. For example, in the simulation,

if we want to be confident in detecting a difference of 2 in β_1, then a VIF of 10 or even 15 might be sufficient. If more precise detections are necessary, a VIF of 5 might be too high. In general, every 1 unit increase in VIF has a smaller change on the curve as the VIF gets higher, so there is no clear cutoff value. Choosing the VIF cutoff depends on the application and the consequences of a type II error.

4.1 Advice for Motivating Example

The bridge data probably does not have a $\sigma^2 = 1$ and the Xs do not have a mean of zero like we used in the simulation study. We will not get the exact same power curve. However, the results are comparable because the data can be standardized. The shape of the curve will be similar and so higher VIF values will have less power. In the bridge example, high VIFs can be problematic if the goal of our analysis is to determine which covariate has a significant effect on bridge build time. In that case, a VIF cutoff of 5 may be best and the analysis should be done on predictors with less colinearity. However, if prediction or interpolation is the main concern, then a VIF cutoff of 10 may be fine.

References

[1] Tryphos (1998), Methods for business analysis and forecasting: text & cases, Wiley, New York p.130-131 available at `http://gattonweb.uky.edu/sheather/book/docs/datasets/bridge.txt`.