Correlation, Variance Inflation and Multicollinearity in Regression Model

Richard TAY

School of Business IT & Logistics, RMIT University, Melbourne, Australia
E-mail: rtay888@gmail.com

Abstract: Since observational data are often used and variables in real life are often correlated, correlations among the variables are common in transportation research. In practice, this problem is often addressed by examining the correlations among the explanatory variables and estimating the variance inflation factors. More importantly, it is a common practice to exclude variables that are highly correlated from the regression model. This study provides some simulated cases to demonstrate that including highly correlated variables, as measured by correlation coefficients and the variance inflation factor in the estimation models, will not necessarily create significant estimation problems, at least in terms of coefficient estimates. Therefore, depending on the purpose of the research, researchers and practitioners should not automatically exclude any variables from their regression models simply because they are highly correlated with other explanatory variables. Some discussions on possible strategies to deal with multicollinearity are also provided.

Keywords: Correlation, Variance Inflation, Multicollinearity, Regression Model

1. INTRODUCTION

Observational data are often used in transportation research, especially safety and security research, because of the practical and ethical difficulties that may place the study subjects in danger in any risky experiments. Without the ability to exercise some controls over the data generating process, traffic collision, violation or emergency evacuation data, as well as naturalistic or self-reported pedestrian or risky driving behavior data, are often correlated. The correlation problem is even more serious when dealing with a large set of explanatory factors. Many of the explanatory factors are often correlated in real life. Some examples include speed limit and road classification, weather and road surface condition, etc.

Although the statistical theories related to multicollinearity are well-established and widely known in the areas of econometrics and statistics (Gujarati, 2004; Greene, 2003), this issue is not very well understood in practice, including in transportation and road safety (Washington et al, 2011; Fitzpatrick et al, 2001; Chen & Chen, 2016; Berhanu, 2004; Amoh-Gyimah, 2016). Some consequences resulting from multicollinearity include high standard errors of the estimated parameters of the collinear variables, large sampling variability which means estimated parameters based on different samples could vary widely, and inaccurate interpretation of the effects of explanatory variables because the change of one variable would inevitably lead to the change of another variable (Washington et al, 2011, Shi et al, 2016).

In practice, this problem is often addressed by examining the correlations among the explanatory variables (Fitzpatrick et al, 2001; Chen & Chen, 2016; Berhanu, 2004; Amoh-Gyimah, 2016; Oh et al, 2009). Typical correlation values that have been suggested and used as the threshold range from 0.6 to 0.8. Additionally, the variance inflation factor is often recommended as another selection criterion, with typical threshold values of 5-10

(Kutner et al, 2004; Heiberger & Holland, 2015; Jung et al, 2011; Khan et al, 2012, 2013). Although these techniques are useful tools for detecting potential multicollinearity, they serve only as an indication of potential multicollinearity; correlation is not collinearity.

More importantly, there is much confusion in the transportation literature, including the road safety literature, on what to do if correlations among explanatory variables are detected. One common practice is to automatically exclude variables that are highly correlated from the regression model (Fitzpatrick et al, 2001; Chen & Chen, 2016; Berhanu, 2004; Amoh-Gyimah, 2016; Jung et al, 2011; Khan et al, 2012; Mohammadi et al, 2014). Theoretically, this practice should not be encouraged. Excluding a correlated variable from the model may result in the error term being correlated with the exploratory variables included in the model, which will result in biased and inconsistent estimates.

The objective of this study is to demonstrate some problems that may arise from using common techniques, such as correlation among explanatory variables and variance inflation factor, to detect multicollinearity problems and the danger of automatically excluding correlated variables from regression models. It will add to the transportation and road safety literature by extending our understanding of the issues related to correlation and multicollinearity, using simulated cases where the true models are known.

2. SIMULATED CASE STUDIES

An appropriate place to start examining the problems associated with multicollinearity is the standard multiple regression or ordinary least square regression which is widely used to analyze many road safety measures, including traffic collision rates, violation rates, speed, time-to-collision, and social costs associated with traffic collisions, as well as being used in estimating other transportation models.

As a simple case study, we start off by generating a sample of 1000 observations with three normally distributed explanatory variables. X_1 is independent and not correlated with any of the other explanatory variables, whereas X_2 and X_3 are highly correlated. Note that the probability distribution of the explanatory variables is not critical and any other distribution may be used. In addition, we generate a normally distributed random term (ε) with zero mean. The descriptive statistics of the variables in the sample are shown in Table 1.

Table 1. Descriptive statistics

Variables	Minimum	Maximum	Mean	Std Dev
X_1	66.05	92.31	79.91	4.09
X_2	82.38	119.20	100.02	5.12
X_3	95.20	129.83	110.67	5.19
ε	-29.34	27.89	-0.24	8.10

The bivariate correlation among the variables in the sample is shown in Table 2. The results confirm that X_1 is independent whereas X_2 and X_3 are very highly correlated. The correlation coefficient of 0.952 is much higher than the commonly recommended thresholds of 0.6-0.8.

Table 2. Correlation coefficients and p-values

	X_1	X_2	X_3
X_1	1.000	-0.010 (0.741)	-0.024 (0.456)
X_2	-0.010 (0.741)	1.000	0.952 (<0.001)
X_3	-0.045 (0.441)	0.952 (<0.001)	1.000

In addition to simple correlation analysis, it is a common practice to compute the variance inflation factor (VIF). Note that tolerance which is the inverse of the variance inflation factor is sometimes computed instead.

$$VIF = 1/(1-R_x^2) \quad (1)$$

where R_x^2 is the R-square of the regression of the explanatory under investigation on the other explanatory variables. To obtain the VIF for X_3, we run a regression of X_3 on X_1 & X_2. The R-square is estimated as 0.906 and the VIF is computed as 10.64, which is larger than the widely used thresholds of 5-10.

Hence, by both the correlation coefficient and VIF criteria, the explanatory variables (X_2 and X_3) are highly correlated and there is a very high chance of multicollinearity, whereas X_1 is not correlated with the other explanatory variables.

2.1 Case 1

As the base case (Model 1a), we begin by estimating a simple multiple regression model with one independent and two highly correlated variables. First, we compute a dependent variable, Y_1, using the following equation:

$$Y_1 = 100 + 5X_1 + 3X_2 - 2X_3 + \varepsilon \quad (2)$$

We then run the ordinary least square regression using the simulated data and the results are shown in Table 3. It should be highlighted that although X_2 and X_3 are very highly correlated (r = 0.952, VIF = 10.64), the correlation does not result in any estimation problem. All the coefficient estimates are statistically significant and all the estimated coefficients are very close to the true coefficients.

Table 3. OLS results with full model

Variable	**Coefficient**	**Standard Error**	**t-statistic**	**p-value**
Model 1a (Base or Full Model)				
Constant	96.903	7.560	12.818	< 0.001
X_1	5.032	0.063	80.287	< 0.001
X_2	2.961	0.163	18.144	< 0.001
X_3	-1.961	0.161	-12.160	< 0.001

This simple example demonstrates that having highly correlated variables does not *necessarily* imply that there will be an estimation problem. In fact, the OLS estimates obtained in this case are very good (the coefficients are estimated correctly and identified as statistically significant) despite the very high correlation between two of the three variables.

2.2 Case 2

To illustrate what will happen to the OLS estimates if one of the correlated variables is excluded from the model, we rerun the model by omitting one variable each time. The estimations results are reported in Table 4. In the Model 2a where the uncorrelated variable is omitted, the estimation results are still fairly good. The coefficients of the two correlated variables are close to their true values although the estimate of the intercept term is significantly biased. Note that the estimate of the intercept term in this case is close to the estimate of the true intercept plus the coefficient of X_1 times the mean of X_1.

Table 4. OLS results without one of the explanatory variable

Variable	**Coefficient**	**Standard Error**	**t-statistic**	**p-value**
Model 2a (Uncorrelated variable excluded)				
Constant	511.505	15.085	33.909	< 0.001
X_2	3.474	0.445	7.799	< 0.001
X_3	-2.538	0.440	-5.766	< 0.001
Model 3a (One correlated variable excluded)				
Constant	66.001	7.626	8.654	< 0.001
X_1	5.066	0.067	75.537	< 0.001
X_2	1.072	0.054	19.996	< 0.001
Model 4a (Other correlated variable excluded)				
Constant	81.176	8.658	9.376	< 0.001
X_1	5.076	0.072	70.310	< 0.001
X_3	0.824	0.057	14.447	< 0.001

On the other hand, as shown in Table 4, in Model 3a and Model 4a, where one of the correlated variables is excluded, the estimates of the coefficients of the remaining correlated explanatory variables are very biased. In fact, the estimate of the remaining highly correlated variable in both Model 3a and 4a are the combined effect of X_2 and X_3. The main advantage of omitting the correlated variable is the much smaller standard error (≈ 0.05) estimated compared to the full model (≈ 0.16).

Note that the sum of the estimates of X_2 in model 3a and X_3 in model 4a are close to 1 because of the "true" values of β_2 and β_3 used in the simulation. It can easily be shown that the estimated coefficients in theory will be combination of these two coefficients ($\beta_2 = 3$ and $\beta_3 = -2$) and their correlation coefficient (ρ_{23} which is 0.925):

$$\widehat{\beta_2} = \beta_2 + \rho_{23}\beta_3 \quad (3a)$$

$$\widehat{\beta_3} = \beta_3 + \rho_{23}\beta_2. \quad (3b)$$

2.3 Case 3

In the third case study, we will examine what will happen if we include a correlated variable that is theoretically unimportant. To do this, we compute a new dependent variable, Y_2, using the following equation:

$$Y_2 = 100 + 5X_1 + 3X_2 + \varepsilon \quad (4)$$

We then run the ordinary least square regression again using Y_2 as the dependent variable and the results are shown in Table 5. As evident from the results shown in the table, the OLS estimation correctly identifies the unimportant variable as statistically insignificant and correctly identifies the important variables as statistically significant. Moreover, the coefficient estimates are close to their true values. Note that the main disadvantage of including an unimportant variable is the inflated standard error (≈ 0.16 compared to ≈ 0.05).

Table 5. OLS results with additional correlated explanatory variable

Variable	**Coefficient**	**Standard Error**	**t-statistic**	**p-value**
Model 5a (2nd Base Model)				
Constant	97.510	7.116	13.702	< 0.001
X_1	5.031	0.063	80.395	< 0.001
X2	2.998	0.050	59.915	< 0.001
Model 6a (With extra correlated variable)				
Constant	96.903	7.560	12.818	< 0.001
X_1	5.032	0.063	80.287	< 0.001
X_2	2.961	0.163	18.144	< 0.001
X_3	0.039	0.161	0.239	0.811

In summary, adding a highly correlated but theoretically unimportant variable does not create a significant estimation problem, in terms of estimating the coefficients although the standard errors are slightly larger and the t-statistics are slightly smaller. If the main purpose of the research is to identify the explanatory factors, then this is not as critical a problem compared to excluding a correlated variable, especially if the sample size is quite large.

2.4 Effects of sample size

One of the main problems with multicollinearity is the inflated variance and smaller than expected t-statistics. Since t-statistics is directly affected by sample size, it will be valuable to see if the results are robust when the sample size is reduced. In the above three case studies, the sample used (N=1000) is quite large. To examine if the results will be substantially affected by the sample, we randomly select 30 observations from the above samples and rerun the analyses. The descriptive statistics of the smaller sample is shown in Table 6.

Table 6. Descriptive statistics

Variables	Minimum	Maximum	Mean	Std Dev
X_1	73.499	85.409	78.296	3.150
X_2	88.115	111.977	101.882	4.697
X_3	99.450	122.950	112.720	5.067
ε	-14.581	19.433	1.231	8.315

The bivariate correlation among the variables in the sample is shown in Table 7. The results confirm that X_1 is independent whereas X_2 and X_3 are very highly correlated.

Table 7. Correlation coefficients and p-values

	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$
$\mathbf{X_1}$	1.000	0.189 (0.317)	0.282 (0.131)
$\mathbf{X_2}$	0.189 (0.317)	1.000	0.951 (<0.001)
$\mathbf{X_3}$	0.282 (0.131)	0.952 (<0.001)	1.000

Again, in addition to simple correlation analysis, the variance inflation factor (VIF) is computed by running a regression of X_3 on X_1 & X_2. The R-square is estimated as 0.915 and the VIF is computed as 11.76, which is larger than the widely used thresholds of 5-10. Hence, by both the correlation coefficient and VIF criteria, the explanatory variables (X_2 and X_3) are highly correlated and there is a very high chance of multicollinearity.

We begin our reanalyses with the base case (Model 1b) with one independent and two highly correlated variables. The estimates are shown in Table 8. It should be highlighted again that although X_2 and X_3 are very highly correlated (r = 0.951), the correlation does not result in any significant estimation problems, in terms of the coefficient estimates. Both the coefficient estimates for X_2 and X_3 are statistically significant and all the estimated coefficients are relatively close to the true coefficients.

Table 8. Estimation results for smaller sample

Variable	Coefficient	Standard Error	t-statistic	p-value
Model 1b (Base or Full Model)				
Constant	198.959	45.490	4.373	< 0.001
X_1	4.371	0.514	8.501	< 0.001
X_2	2.822	1.066	2.646	0.014
X_3	-2.270	1.012	-2.243	0.034
Model 2b (Uncorrelated variable excluded)				
Constant	460.948	63.828	7.222	< 0.001
X_2	0.410	1.961	0.209	0.836
X_3	0.622	1.818	0.342	0.735
Model 3b (One correlated variable excluded)				
Constant	205.659	48.664	4.226	< 0.001

X_1	3.983	0.519	7.672	< 0.001
X_2	0.544	0.348	1.561	0.130
Model 4b (Other correlated variable excluded)				
Constant	227.381	48.872	4.653	< 0.001
X_1	4.009	0.548	7.315	< 0.001
X_3	0.281	0.341	0.806	0.417
Model 5b (2nd Base Model)				
Constant	199.755	44.604	4.478	< 0.001
X_1	4.325	0.476	9.089	< 0.001
X_2	2.552	0.319	7.995	< 0.001
Model 6b (With extra correlated variable)				
Constant	198.959	45.490	4.374	< 0.001
X_1	4.371	0.514	8.501	< 0.001
X_2	2.822	1.066	2.646	0.014
X_3	-0.270	1.012	-0.266	0.792

This example again demonstrates that having highly correlated variables does not *necessarily* imply that there will be an estimation problem when the correct model is specified, even when the sample size is not very large. The OLS estimates of coefficients obtained are very good despite the very high correlation between two of the three variables and an extremely high VIF. Again, the main disadvantage is the much larger standard errors (≈1.0) compared to the reduced model (≈ 0.3).

In Model 2b, where the uncorrelated variable is omitted, the estimated results are quite biased. The coefficients estimated are quite different from the true values although they are statistically insignificant. This result is surprising and in contrast to the results obtained with a large sample. In Models 3b and 4b where one of the correlated variables is excluded, the estimates of the coefficients of the uncorrelated variable are still reasonably good but the estimates of the remaining correlated explanatory variables are very biased.

In Model 6b, we include a correlated variable to Model 6a that is theoretically unimportant to examine its effect on the estimates. Again, the results in Table 8 show that the OLS estimation correctly identifies the unimportant variable as statistically insignificant and correctly identifies the important variables as statistically significant. Moreover, all the coefficient estimates are relatively close to their true values.

3. DISCUSSION

The results from Model 1a demonstrate that having highly correlated independent variables does not *necessarily* imply that there will be a significant estimation problem. In fact, the OLS estimates obtained in this case are very good despite the very high correlation between two of the three variables. The main problem with multicollinearity is that it results in inflated variances and lower values for the t-statistics. Therefore, if the main purpose of the model is to obtain an unbiased estimate of the coefficients and all the t-statistics of the theoretically important variables are significant, then there is no need to compute the correlation coefficients or variance inflation factors to check for multicollinearity.

In the Model 2a where the uncorrelated variable is omitted, the estimation results are still fairly good. The coefficients of the two correlated variables are close to their true values

although the estimate of the intercept term is significantly biased. This result indicates that we can exclude an explanatory variable that is uncorrelated with the other explanatory variables without a serious impact on the estimates of the coefficients of the other explanatory variables, if the sample size is very large. However, the estimate of the intercept will be biased. If the correct estimate for the intercept is not critical for the purpose of the research, then this approach will not create a serious problem.

On the other hand, as shown in Table 4, in Model 3a and Model 4a, where one of the correlated variables is excluded, the estimates of the coefficients of the remaining correlated explanatory variables are very biased. In fact, the coefficient of the remaining explanatory variable in the Model 4a actually has an opposite sign. This result indicates that not only is the estimate biased, it is predicting an opposite effect. This is a very serious mistake if the correct estimate of the coefficients is important for the purpose of the research. Therefore, important correlated explanatory variables should not automatically be excluded from the model even though they are highly correlated with another explanatory variable. This conclusion is in contrast to the popular recommendation to exclude variables that are highly correlated.

In summary, if we perform correlation analyses and/or compute the variance inflation factor, and then using the results to decide which variables to include in the estimation model, it is recommended that you include all correlated variables that are theoretically important, if the correct estimation of the coefficients is critical for the purpose of the research. This recommendation is contrary to the common practice of excluding correlated variables in the model. Also, although excluding any uncorrelated variables does not appear to result in serious estimation errors, except for the intercept term, it is also not recommended; particularly in non-linear models since the variables can still be correlated with other explanatory variables in a non-linear manner.

The results of the third model also have some interesting practical implications. They provide some possible avenues to look for explanations if an estimated coefficient has an unexpected sign. An estimated coefficient may have an unexpected sign if a highly and positively correlated variable is excluded from the model and this excluded variable has a larger and opposite effect on the dependent variable. Alternatively, it will also have an unexpected sign if a highly and negatively correlated variable is excluded from the model and this excluded variable has a similar but larger impact on the dependent variable.

In Model 2b, where the uncorrelated variable is omitted, the estimated coefficients are very biased. The coefficients estimated are quite different from the true values although they are statistically insignificant. This result is surprising and in contrast to the results obtained with a large sample. Since X_1 is not highly correlated with X_2 and X_3, its omission should not have created significant problems, at least in terms of getting unbiased estimates, since X_2 and X_3 should not be correlated with the error term. In fact, the VIF for X_1 is estimated at only 1.17. However, because of the small sample size (we use a smaller sample of the same data and check the descriptive statistics, correlation and VIF), apparently even low correlations could create an estimation problem. In this case, neither the correlation analysis nor the VIF is useful in detecting the problem.

In summary, the six models in the three case studies demonstrate that including highly correlated variables, as measured by correlation coefficients and variance inflation factors, in the estimation models do not always create significant estimation problems, even in small samples. This is true whether the correlated variable is theoretically important or not. However, excluding any important correlated variable will create a serious estimation problem. The estimate of the remaining correlated variable will be biased. The main advantage of omitting the correlated variable is a much more precise estimate (smaller standard error) of the coefficients.

Therefore, if the main use of the model is to predict the values of the dependent variable, then multicollinearity may not be as serious a problem (Gujarati, 2004). However, if the main purpose of the model is to correctly identify the significant independent variables and their effects, then multicollinearity is a problem and the omission of correlated variables is not recommended. On the other hand, if the main purpose of the research is to have precise estimates (smaller confidence intervals), then omitting correlated variables may be preferred.

If the decision is to omit the correlated variables to get more precise estimates with smaller standard errors and confidence intervals, and identification of significant explanatory variables is relatively important, then caution needs to be exercised when interpreting the coefficients estimated. In practice, observational variables are often proxies for multiple attributes which may not be adequately captured by a single label on the variable or by a single variable. Hence, this issue may be implicitly present in many regression models. However, if an explicit decision is made to omit variables that are known to be correlated, then it should be made clear that the estimated coefficients represent the combined effects of both the included and excluded variables.

Some practical methods to reduce the problem include getting more data. As illustrated in this study, correlations among explanatory variables are a more serious problem in small samples compared to large samples. Collecting more data also may reduce the correlations among variables, especially if the data are collected from different sources or collected using different methods. Other approaches to deal with multicollinearity include the use of extraneous or prior information, and transforming the data (Gujarati, 2004). Note that transforming the data (e.g. taking logarithms) can only reduce linear correlations among the explanatory variables and may not be suitable if the dependent and independent variables are supposed to be linearly related in theory because transforming the variables may then lead to misspecification errors.

4. CONCLUSION

Since observational data are often used and variables in real life are often correlated, correlations among the explanatory variables are common in transportation and road safety research. In practice, this problem is often addressed by examining the correlations among the explanatory variables and estimating the variance inflation factors. More importantly, it is common to exclude variables that are highly correlated from the regression model to avoid the multicollinearity problem in estimation.

This study provides some simulated cases which demonstrate that including highly correlated variables, as measured by correlation coefficients and variance inflation factors, in the estimation models does not *always* create significant problems in the estimation of the coefficients for the explanatory variables. This is true whether the correlated variable is theoretically important or not. However, excluding any important correlated variable will create a serious estimation problem. The estimate of the remaining correlated variable will be biased. The results are true whether the sample size is large or small. Therefore, researchers and practitioners should not *automatically* exclude any variables from their regression models simply because they are highly correlated with other explanatory variables.

It should be noted again that the main issue with including correlated variables is the inflated variance which will affect the t-statistics and the corresponding significance tests. Therefore, if the main use of the model is to predict the values of the dependent variable and not to conduct hypothesis tests, then multicollinearity may not be as serious a problem (Gujarati, 2004). However, if the main purpose of the model is to correctly identify the

significant independent variables and their effects, then multicollinearity is a problem. Some practical methods to reduce the problem include getting more data, using extraneous or prior information or transforming the data (Gujarati, 2004).

It should be noted that there are other factors or parameters that may affect the results. For example, the t-statistic is affected by not only the sample size but also the means and variances of the relevant variables as well. Hence, more studies should be done to investigate the effects of these influences to get a better understanding of the multicollinearity problem, and the usefulness of requiring correlation analysis and the computation of variance inflation factors under different circumstances.

5. REFERENCES

Amoh-Gyimah, R., Saberi, M., Sarvi, M. (2016) Macroscopic modeling of pedestrian and bicycle crashes: Across-comparison of estimation method. *Accident Analysis and Prevention,* 93, 147-159

Berhanu, G. (2004) Models relating traffic safety with road environment and traffic flows on arterial roads in Addis Ababa. *Accident Analysis and Prevention,* 36, 697-704

Chen, P., Chen, Q. (2016) Built environment effects on cyclist injury severity in automobile-involved bicycle crashes. *Accident Analysis and Prevention,* 86, 239–246

Fitzpatrick, K., Carlson, P., Brewer, M., Wooldridge, M. (2001) Design factors that affect driver speed on suburban streets. *Transportation Research Record,* 1751, 18-25

Greene, W. (2003) *Econometric Analysis*, 5th edition, Prentice Hall, New Jersey

Gujarati, D. (2004) *Basic Econometrics*, 4th edition, MacMillan, New York

Heiberger, R., Holland, B. (2015) *Statistical analysis and data display: an intermediate course with examples in* R., 2nd ed. Springer, New York

Jung, S., Qin, X., Noyce, D. (2011) Modeling Highway Safety and Simulation in Rainy Weather. *Transportation Research Record,* 2237, 2011, pp.134-143

Khan, G., Bill, A., Chitturi, M., Noyce, D. (2012) Horizontal Curves, Signs, and Safety. *Transportation Research Record,* 2279, 2012, pp.124-131

Khan, G., Bill, A., Chitturi, M. and Noyce, D. (2013) Safety Evaluation of Horizontal Curves on Rural Undivided Roads, *Transportation Research Record*, 2386, 147-157

Kutner, M., Nachtsheim, C., Neter, J., Li, W. (2004) *Applied linear statistical models*, 5th ed., McGraw-Hill, New York

Mohammadi, M., Samaranayake, V., Bham, G. (2014). Safety Effect of Missouri's Strategic Highway Safety Plan: Missouri's Blueprint for Safer Roadways, *Transportation Research Record*, 2465, 33-39

Oh, J., Washington, S., Lee, D. (2009) Expected Safety Performance of Rural Signalized Intersections in South Korea. *Transportation Research Record,* 2114, 72-82

Shi, Q, Abdel-Aty, M., Lee, J. (2016) A Bayesian ridge regression analysis of congestion's impact on urban expressway safety. *Accident Analysis and Prevention,* 88, 124-137

Washington, S., Karlaftis, M., Mannering, F. (2011) *Statistical and Econometric Methods for Transportation Data Analysis*, 2nd edition, Chapman & Hall/CRC, Boca Raton