Model Building Process Part 2: Factor Assumptions

Authored by: Sarah Burke, PhD

17 July 2018

Revised 6 September 2018

The goal of the STAT COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.afit.edu/STAT.

STAT Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Introduction 2
Assumption: Input Variables are Independent 2
Consequences of Multicollinearity 2
Methods to Assess Multicollinearity 3
Dealing with Multicollinearity 6
Assumption: The Model is Correct 6
Informal Method to Assess Model Fit 7
Lack of Fit Test 7
Conclusion 8
References 8

Revision 1, 6 Sep 2018: Formatting and minor typographical/grammatical edits.

Introduction

This document is the second part in a series on the steps of the (statistical) model building process. Part 1 (Burke, 2017) discussed methods to assess whether the error assumptions in a linear regression model had been satisfied and suggested several remedial measures. The purpose of this best practice is to highlight methods and metrics to assess the remaining model assumptions. Specifically, this paper shows how to assess whether the inputs into the model are independent and if the model itself is adequate.

Keywords: linear regression, multicollinearity, lack of fit, variance inflation factor

Assumption: Input Variables are Independent

When building a linear regression model, apart from the assumptions on the error terms (which were discussed in Part 1 of this series [Burke, 2017]), we also assume the input variables (factors) are independent of each other. *Multicollinearity* occurs when factors are correlated with one another. Note this correlation does not relate to how the factors are correlated with the response, only with each other. For example, when estimating miles per gallon (mpg) of a car, potential factors impacting mpg might include the number of cylinders, horsepower, weight, model year, and acceleration associated with the car. Horsepower and weight, however, are positively correlated with each other; as weight of the car increases, horsepower tends to increase (or vice versa; recall that correlation does not imply causation!). A regression model that estimates mpg will not require both of these terms because they are so closely associated with each other.

From a design of experiments (DOE) viewpoint, an important design property to evaluate when building a design is the degree of aliasing or confounding. When two terms are aliased (confounded) with each other, it is impossible to distinguish which term caused the change in the response. When there is severe aliasing/confounding, multicollinearity can become an issue when analyzing the data from that design. An *orthogonal designed experiment* is one in which the factors and model terms are uncorrelated with each other. In a well-designed experiment, the values of the factors are controlled to ensure that the factors are uncorrelated with each other. When the factors are orthogonal, the estimated parameter for one model term will be the same value whether another model term is included in the model or not. This property makes model-fitting straightforward because there is no aliasing, meaning there is no ambiguity on which factor has the true effect on the response and the order in which terms are added or removed from the model does not matter. In an observational study, this attribute is often not present in the data precisely because the factor levels are not controlled.

Consequences of Multicollinearity

The repercussions of multicollinearity can be severe, particularly with data from an observational study, so an assessment of the degree of multicollinearity should be done. When the factors are highly correlated with each other, the variance estimates of the model coefficients are inflated resulting in an

unstable linear model fit. This can result in an increase in the Type II error (i.e., terms are deemed not statistically significant when they really are). Interpretation of the model coefficients is also no longer straightforward because the value of the model parameters depends on which other model terms are included. This means the order in which terms are added or removed from the model now matters. In addition, the model coefficients can have the wrong magnitude and the wrong sign, leading to incorrect conclusions (Silvestrini and Burke, 2018).

Methods to Assess Multicollinearity

Several methods to informally and formally assess multicollinearity include graphing, studying the sign and magnitude of the model coefficients for unusual or unexpected results, assessing the correlation among the model coefficient estimates, and calculating the *variance inflation factor* (VIF).

One informal way to detect multicollinearity is to examine a scatterplot matrix of the factor values. Visually discernable patterns in this plot are indicators of multicollinearity. Figure 1a shows a scatterplot matrix of data for different cars from an observational study done to estimate miles per gallon (mpg). This dataset was obtained from the University of California, Irvine (UCI) Data Mining Repository (Dua and Taniskidou, 2017). If these variables were independent, then there should be no discernable pattern in this plot. However, we see in particular that displacement is linearly correlated with horsepower, weight, and acceleration. In addition, horsepower is correlated with weight and acceleration. In contrast, Figure 1b shows a scatterplot matrix for a designed experiment (a factorial design with center points) with six factors, demonstrating orthogonal factors and no multicollinearity.

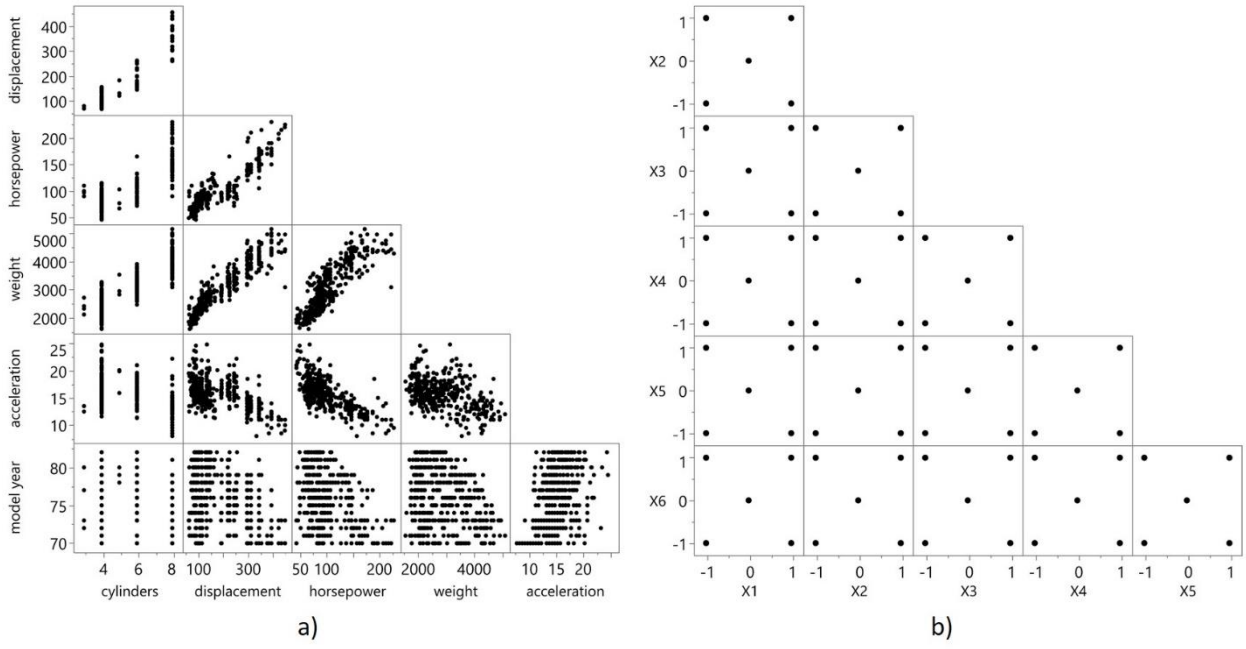


Figure 1. Scatterplot matrix for (a) mpg observational dataset and (b) full factorial design

Subject matter expertise can also provide an informal assessment of the presence of multicollinearity. For example, if scientific principles or prior information indicate an estimated regression coefficient should have a specific sign, then seeing the opposite sign may indicate multicollinearity is present.

A more formal method to assess multicollinearity is to calculate the correlation coefficients between all pairs of model coefficient estimates. Correlation coefficients range from 0 to 1 in magnitude; the larger the value, the more linearly correlated two terms are. Correlations greater than 0.7 in magnitude are signals that multicollinearity is an issue in the dataset. Figure 2a shows the model term correlation coefficients for a specified model of the mpg dataset. As expected, several of the model coefficients are highly correlated with each other, including the two-factor interactions (e.g., the correlation coefficient between the main effects of weight and horsepower is -0.8282). This is further evidence that multicollinearity is a concern for this observational dataset. Figure 2b shows the pairwise correlations for the full factorial design shown in Figure 1b. As designed, the correlations between all terms in the model (including the interactions) are zero, indicating the factors are uncorrelated.

Correlation of Estimates

Corr

	Intercept	Disp	HP	Weight	Acc	MY	Disp*HP	Disp*Weight	Acc*MY
Intercept	1.0000	-0.0641	0.1616	-0.1503	-0.9816	-0.9938	-0.2957	0.2412	0.9765
Disp	-0.0641	1.0000	0.1368	-0.0641	0.0256	0.0168	-0.1246	-0.3433	-0.0115
HP	0.1616	0.1368	1.0000	-0.8282	-0.2047	-0.2130	-0.9053	0.7069	0.2546
Weight	-0.1503	-0.0641	-0.8282	1.0000	0.1523	0.1561	0.8461	-0.8610	-0.1966
Acc	-0.9816	0.0256	-0.2047	0.1523	1.0000	0.9830	0.3158	-0.2285	-0.9969
MY	-0.9938	0.0168	-0.2130	0.1561	0.9830	1.0000	0.3201	-0.2255	-0.9843
Disp*HP	-0.2957	-0.1246	-0.9053	0.8461	0.3158	0.3201	1.0000	-0.8329	-0.3498
Disp*Weight	0.2412	-0.3433	0.7069	-0.8610	-0.2285	-0.2255	-0.8329	1.0000	0.2562
Acc*MY	0.9765	-0.0115	0.2546	-0.1966	-0.9969	-0.9843	-0.3498	0.2562	1.0000

a)

Correlation of Estimates

Corr

	Intercept	X1	X2	X3	X4	X5	X1*X2	X1*X3	X1*X5
Intercept	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
X1	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
X2	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
X3	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
X4	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
X5	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
X1*X2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
X1*X3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
X1*X5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

b)

Figure 2. Correlation matrix of model estimates for (a) mpg dataset and (b) full factorial design

Because of the issues that arise due to multicollinearity, it is critical to use a color map on correlations to assess the level of multicollinearity in a designed experiment. If the design has a large amount of aliasing, then it will be difficult to build an acceptable linear regression model. Not only will it be challenging to resolve which terms affect the response, but the inflated variances of the model

estimates will affect the power of the design. The matrix of correlation values is displayed in the color map on correlations when evaluating a design. As a comparison, Figure 3 shows the color maps for the mpg dataset and the full factorial design. Recall that the ideal plot has a red line across the diagonal with blue in the off diagonal (i.e., the full factorial design in Figure 3b). Severe amounts of aliasing is present in the mpg dataset as observed in Figure 3a.

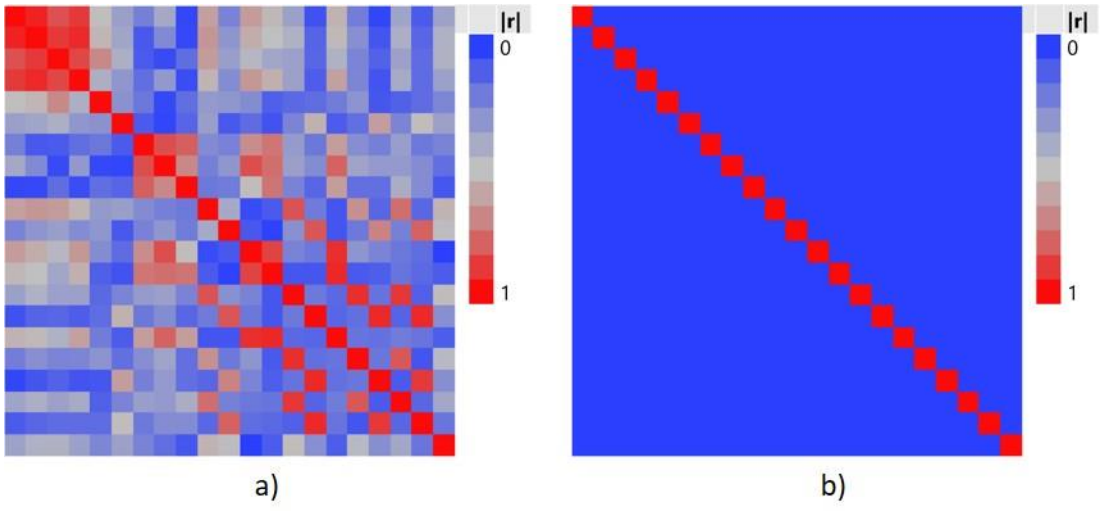


Figure 3. Color map on correlations for (a) mpg dataset and (b) full factorial design

A formal metric for multicollinearity is the variance inflation factor (VIF). VIF is calculated for each model term as:

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

where R_j^2 is the coefficient of determination (R^2) from a model where the factor or model term x_j is used as the response and all of the other factors in the original model are used as independent variables to predict values of x_j. When there is a strong relationship among factors, this R_j^2 value will be large, resulting in a large value of VIF (Silvestrini and Burke, 2018). The VIF is 1 when the predictor is not linearly related to the other independent factors since R_j^2 is zero when the factors are orthogonal to each other. *A VIF greater than 10 is often an indication that severe multicollinearity is present in the data*. Note that if a factor has a perfect linear association with another model term, the VIF is infinity. Figure 4 shows the parameter estimates table for specified models for the mpg dataset and the full factorial design. The last column in each of these tables shows the VIF values for each model term. Figure 4a shows that all of the VIF values are greater than 10, another indication that multicollinearity is an issue in this dataset. Figure 4b shows VIF values of 1, as expected, for the model parameters from the full factorial design.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	5.4008076	0.731475	7.38	<.0001*	.
Disp	-0.002269	0.000306	-7.40	<.0001*	33.928629
HP	-0.004143	0.001064	-3.89	0.0001*	55.37964
Weight	-0.000262	4.753e-5	-5.51	<.0001*	53.775532
Acc	-0.209883	0.045608	-4.60	<.0001*	522.34767
MY	-0.012217	0.009562	-1.28	0.2021	40.933007
Disp*HP	6.6077e-6	3.074e-6	2.15	0.0322*	146.50991
Disp*Weight	3.4883e-7	1.399e-7	2.49	0.0131*	177.03055
Acc*MY	0.0026636	0.000599	4.44	<.0001*	633.44281

a)

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	10.00705	0.133706	74.84	<.0001*	.
X1	1.1116048	0.134746	8.25	<.0001*	1
X2	-2.197748	0.134746	-16.31	<.0001*	1
X3	0.6921473	0.134746	5.14	<.0001*	1
X4	1.5443999	0.134746	11.46	<.0001*	1
X5	1.3738659	0.134746	10.20	<.0001*	1
X1*X2	0.6745169	0.134746	5.01	<.0001*	1
X1*X3	0.4104374	0.134746	3.05	0.0035*	1
X1*X5	-1.120494	0.134746	-8.32	<.0001*	1

b)

Figure 4. Variance inflation factor values for models (a) mpg dataset and (b) full factorial design

Dealing with Multicollinearity

What should you do if your dataset has multicollinearity? Centering the data for the predictor variables can reduce multicollinearity among first- and second-order terms. Centered data is simply the value minus the mean for that factor (Kutner et al., 2004). Alternative analysis methods such as principal component analysis (PCA), ridge regression, or Least Absolute Shrinkage and Selection Operator (LASSO) are more often used to account for multicollinearity (Silvestrini and Burke, 2018; Kutner et al., 2004). PCA creates one or more linear combinations of the input variables that are uncorrelated and that explain a large portion of the variability in the data. These linear combinations then become the "factors" used in a linear regression model to model the response. Because these new variables have been constructed so that they are orthogonal to each other, multicollinearity is no longer an issue in the model. *The downside is that interpretation becomes much harder*. If predicting the response is more important than understanding which variables are associated with changes in the response, PCA is a reasonable approach. Ridge regression and LASSO are also alternative models to a traditional linear regression model. These methods are called penalized likelihood methods and produce biased regression coefficients with smaller standard errors. The advantages in reducing the variability of the parameter estimates often outweighs the disadvantages of the biased model parameters. Ridge regression and LASSO are techniques often used in observational studies to reduce the number of potential predictors in the model. These methods can be used to reduce the pool of variables to consider in a model of the response. For more information on ridge regression and LASSO, see Silvestrini and Burke (2018).

Assumption: The Model is Correct

The final assumption when fitting a linear regression model is that the fitted model is correct or adequate. This means all terms affecting the response are included in the model and we have captured the appropriate relationship between the factors and the response. This model is typically assessed using simple plots of residuals, and when possible, a statistical test for lack of fit.

Informal Method to Assess Model Fit

Initial graphical analysis of the factors and the response can often suggest the types of model terms required (e.g., main effects or quadratic terms). Once you have fit a model and are evaluating it, you can utilize some of the residual plots discussed in Part 1 of this series (Burke, 2017) to assess the model fit. For example, when graphing the model residuals by each factor, look for any visually discernable patterns. Patterns in these types of plots indicate that something (e.g., a higher order model term) is missing from the model. For example, if the residuals have a parabolic shape, this is an indication that a quadratic term should be included in the model. Follow-on testing may be required to be able to fit additional or higher order model terms like a quadratic.

When fitting a model using data from an observational study, not all input variables may be initially included in the model. If a plot of the residuals versus an input variable not currently in the model has a visually discernable pattern, then we should consider adding that variable into the model.

Lack of Fit Test

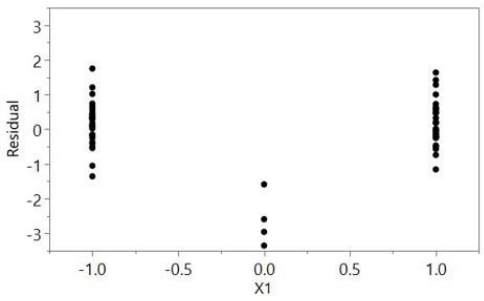
A statistical lack of fit test provides a more formal approach to assess model fit. The lack of fit test is typically utilized for data from a designed experiment because the test itself requires replicated test points at one or more levels of the factors. This does not always happen in data from an observational study, but it can be enforced in a designed experiment. The lack of fit test uses a sum of squares for lack of fit defined as:

$$SS_{LOF} = \sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2 \qquad (2)$$

where m is the number of levels of the factor, n_i is the total observations at the ith level of factor x_i, $\bar{y}_i$ is the average response at the ith level of factor x_i, and $\hat{y}_i$ is the estimated response at the ith level of factor x_i (Kutner et al., 2004). When the lack of fit sum of squares can be calculated, it is used in a hypothesis test where the hypotheses are stated as:

H_0: The current model is adequate
H_a: The current model is inadequate

The alternative hypothesis typically means important terms (such as interaction effects or higher order terms) have been excluded from the model. In the lack of fit test, when p-values are greater than 0.05 we have confidence in a good fit as this results in failing to reject H_0, demonstrating the model is sufficient to explain the variability in the response. Consider the following example using data from the factorial design originally shown in Figure 1b. To assess the model fit, Figure 5a shows a plot of the model residuals by factor X1. Because there is a parabolic pattern in this plot, a quadratic term is likely missing from the model. The lack of fit test for the model in Figure 5b has a small p-value (<0.0001), further indicating that the model is not sufficient. Additional testing is necessary to determine which factor is contributing to the curvature in the response.



a)

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	24	39.900317	1.66251	4.0559
Pure Error	35	14.346402	0.40990	Prob > F
Total Error	59	54.246719		<.0001*
				Max RSq
				0.9809

b)

Figure 5. Assessing lack of fit using (a) graphical analysis of residuals and (b) lack of fit test

Conclusion

In addition to checking the error assumptions when fitting a linear regression model, assessing the presence of multicollinearity and model fit is essential to ensure you draw the correct conclusions. Multicollinearity can make determining what factor truly impacted the response impossible. A well-designed experiment can eliminate multicollinearity and provide results that are accurate and easier to interpret. The methods discussed in this best practice are readily accessible within statistical software packages such as JMP and R.

References

Burke, Sarah. "Model Building Process Part 1: Checking Model Assumptions V. 1.1". Scientific Test and Analysis Techniques Center of Excellence (STAT COE), 31 July 2017.

Dua, D. and Karra Taniskidou, E. *UCI Machine Learning Repository*, 2017, archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li. *Applied linear statistical models*., Chicago: Irwin, 2004.

Silvestrini, R. and Burke, S. *Linear Regression Analysis with JMP and R*. ASQ Quality Press, 2018.