CS-BIGS 1(1): 54-60
http://www.bentley.edu/csbigs/vol1-1/lawson.pdf

Energy Savings Prediction: a Case Study

John Lawson
Brigham Young University, USA

This case study illustrates the prediction of energy savings at a chemical plant after energy saving improvements were implemented at the plant. The study shows the danger in fitting prediction equations by automatic procedures, and how over fitting an equation can actually increase the error of prediction. Errors of prediction estimated from the data used to find the equation are shown to be optimistic. Data splitting is demonstrated as a method to validate a prediction equation, and a more reliable prediction equation than the one found by the automatic procedure was developed along with a more realistic estimate of prediction error.

Keywords*: Least-Squares Regression, Overfitting, Data-Splitting, Mean Square Prediction Error.*

Introduction

This case study illustrates the use of empirical modeling augmented by some fundamental knowledge to estimate the cost savings from energy reduction measures implemented in a chemical plant. The company name and detailed variable descriptions have been masked, and the data have been coded to protect company confidentiality. However, this article should be useful to those seeking examples of the way in which empirical modeling can be used effectively in industrial settings and the danger of over fitting an empirical model by including unnecessary terms. The case study will illustrate the use of regression analysis techniques at an intermediate level. The data set contains fewer observations than the number generally recommended in text books for conducting the type of analysis shown. In actual practice, however, it is not uncommon to encounter this situation. When only sparse information is available, it is even more critical to use sound methods of data analysis, as illustrated in this case study.

Energy consumption in a chemical plant varies from month to month on the basis of production rates, quality of raw materials, ambient temperature and so forth. To estimate the savings that are realized during a period of time after energy reducing plant improvements have been implemented, there is a need to predict what the energy consumption would have been during the same period of time if the improvements had not been made. Otherwise the energy savings attributable to the improvements can not be separated from the reduction in energy use due to changes in the plant operating environment. Predictions of energy consumption can be made by building a predictive model. Predictive models can be fundamental, based upon first principles; empirical, based on available data; or a blend of the two.

A fundamental predictive model would be based on first principles of mass and energy transfer from the fundamental physics and chemistry of the process under study. The advantage to this type of model is that the relationships are already known and proven, and the model will be more reliable for extrapolating to conditions not covered in the available data. The disadvantage is that even for relatively simple processes the fundamental equations

that describe the system may be unmanageably complex.

An empirical predictive model is based on fitting equations to observed data using statistical principles such as the method of least squares. The models developed in this way are much simpler and easier to work with, but the disadvantage is that the accuracy of the predictive model depends upon the quality of the data available. Correlation among input variables can lead to fitted relationships that contradict fundamental principles, and the empirical models may only be useful for prediction within the range of conditions represented in the data.

By blending ideas from fundamental knowledge empirically fit models can be improved. It may be possible to determine from fundamental knowledge the form of the empirical relationship between an input and output variable (i.e., positive, negative, linear or curvilinear). Also, when fitted relationships between input and output variables in an empirical model contradict fundamental knowledge, they should be dropped from the model, regardless of their statistical significance.

Objective and Description of the Data

A company was contracted to make energy reducing improvements at a chemical plant that produced two products. Billing for the work was to be based on a percentage of the resulting energy savings. The objective of the current study was to develop a prediction model for energy consumption that could be used to estimate the energy savings.

The first 12 months of data in Table 1 are the data that were available to develop an empirical prediction model. The last seven months of data in this same table were collected after the energy savings improvements had been implemented in the plant. Only 12 months of data were available, because the plant (as currently constituted) startup was only 12 months before the energy savings improvements began. In order to align the variables shown in Table 1, data had to be aggregated by month and no finer breakdown (i.e., weekly data) was possible. The dependent variable in Table 1 was the monthly energy consumed (coded units) in the plant, and potential predictor variables were: X_1 = ambient temperature expressed in degree days (again coded), X_2 = amount of recycled steam (in coded units) used in the plant, X_3 = down time on product 1 line (when no energy was consumed), X_4 = amount of product 1 produced in the month, X_5 amount of product 2 produced in the month. From basic fundamental knowledge it was known that higher values for X_1, X_4 and X_5 should increase energy consumption, and higher values of X_2 and X_3 should reduce energy consumption. Whether the relationships were linear or curvilinear would require more detailed consideration of the physics and chemistry of the process and was not done in this case.

The goal was to develop a prediction model from data in

Table 1. Data for Energy Prediction

Month	Energy Consumed	X_1	X_2	X_3	X_4	X_5	
1	12432	233.5	314.82	860	170.45	140.67	
2	12322.8	260.75	260.57	960	252.49	123.94	
3	14340.6	339.25	125.15	1000	201.31	112.62	
4	8176.7	285	122.27	980	238.79	23.46	
5	12937.2	195.5	100.82	1000	227.79	120.06	
6	12641.9	125.25	62.16	960	119.8	137.31	
7	14262.7	71	17.6	970	121.03	151.04	Base
8	10325.1	28	201.2	120	104.78	109.86	period
9	12028.3	1.25	211.85	940	189.86	144.69	
10	10703.3	15.25	141.78	940	124.02	152.25	
11	13053	59	153.92	850	144.09	147.55	
12	14444.2	120.5	63.08	760	180.62	148.07	
13	15230.8	238	167.16	930	184.75	157.71	
14	14506.3	252.5	110.79	940	189.1	154.44	Period after
15	13751.1	288.25	5.28	950	215.25	134.34	energy-saving
16	13516.7	311.25	105.92	930	158.8	163.32	improvements
17	10636.6	281.75	75.2	630	109.14	149.87	were made
18	11161.4	190	168.72	850	137.99	149.87	
19	10896.6	124.5	96.27	960	206.02	158.04	

the base period, and use that model to estimate what the energy consumption would have been in the period after energy saving improvements were made, if the energy saving improvements had not been made.

A general rule to follow, when attempting to fit a prediction model by regression from an exploratory study such as this, is that there should be 5 to 10 observations for every variable in the pool of candidates, see [Kutner, Nachtsheim and Neter (2004) p. 346, and Nair et. al. (1995) p.105]. In this study there were 7 variables but only 12 observations, far fewer than generally recommended. However this case is still useful to illustrate the way a prediction equation should be developed if more data were available, and to illustrate that sometimes in practice predictions must be made to form the basis for action, even when the information is considered to be inadequate.

Empirical Model Developed by Contract Company

The initial attempt to develop a prediction model was made by the engineers from the company that was contracted to make the energy saving improvements. They sought to develop a simple linear prediction model of the form

$$y_i = \beta_0 + \sum_{i=1}^{5} \beta_i X_i \,. \qquad (1)$$

The β coefficients, determined from the data in the base period, were found by maximizing

$$R^2 = \left(\sum_i (y_i - \bar{y})^2 - \sum_i (y_i - \hat{y}_i)^2 \right) \Big/ \sum_i (y_i - \bar{y})^2 \qquad (2)$$

with respect to choice of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5$ using a numerical optimization procedure. Here $\bar{y} = \sum_{i=1}^{12} y_i \big/ 12$, and $\hat{y} = \hat{\beta}_0 + \sum_{i=1}^{5} \hat{\beta}_i X_i$. This is equivalent to the method of least squares since $SST = \sum_i (y_i - \bar{y})^2$ is constant. The resulting prediction equation was:

$$\hat{y} = 4592.706 + 10.722X_1 - 10.918X_2 - 2.145X_3 + 10.503X_4 + 62.009X_5 \,,$$

and the maximized value of R^2 was 0.9012. Figure 1 shows a graph of the actual (points) and predicted values (connected by lines) from this equation in the base period.

The engineering group at the contract company felt that they had a good prediction model due to the fact that the proportion of variation explained by the prediction equation (R^2) was greater than 0.90, the coefficients matched

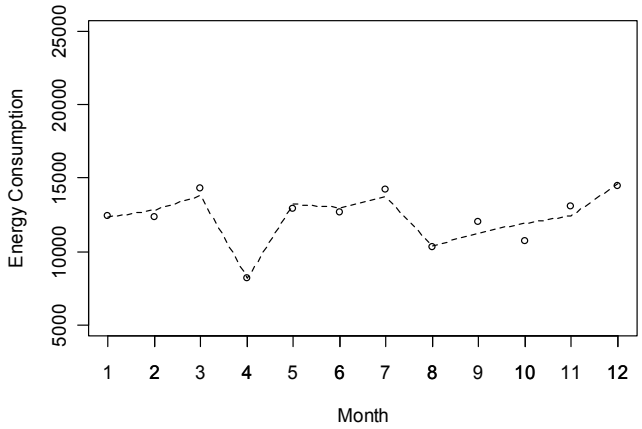

Figure 1. Graph of Actual (o) and Predicted Values (---) in the Base Period

what would be expected from the basic fundamental knowledge, and the fact that the predictions, shown in Figure 1, seemed to follow the trend of the energy consumption data over time. Table 2 shows the actual energy consumption and predictions from the model for the seven months after energy improvements had been implemented. The difference of actual and predicted is the estimated energy savings resulting from the improvements, and the cumulative total is the savings predicted for the entire 7 month period. The average absolute error in prediction for data in the base period divided by the average energy consumption in the base period was 3.56%. This was reported as the measure of how accurately they believed they could predict energy consumption in future months.

Table 2: Actual Energy Consumption and Predictions from the 5 Variable Model in the Period after Energy Savings were Implemented

Month	Actual	Predicted	Predicted Savings
13	15,230.8	15,044.5	-186.3
14	14,506.3	15,636.9	1130.6
15	13,751.1	16,179.0	2427.9
16	13,516.7	16,573.9	3057.2
17	10,636.6	15,880.9	5244.3
18	11,161.4	13,707.2	2545.8
19	10,896.6	14,781.0	3884.4

Cumulative energy savings predicted for 7-month period = 18,108.8.

However, there is always danger in fitting an empirical prediction model using an automatic procedure such as the numerical optimization procedure. Problems with the data such as: extraordinary data points, non-linearity in the relationship between energy consumption and one or more predictor variables [Henderson and Velleman (1981)], multicollinearity between predictor variables

[Faraway (2004)], serial correlation of the response [Kutner, Nachtsheim and Neter (2004)], energy consumption over time, can seriously reduce the value of a prediction model fit by the method of least squares. These problems can be avoided by integrating graphical methods with the model fitting in order to expose outliers and diagnose other problems in the data. Also, without testing the statistical significance of model coefficients, there is a danger of over fitting the model, or including terms which are not significant. This over fitting can actually increase the prediction error [Kutner, Nachtsheim and Neter (2004)]. By determining the error of prediction from the data that were used to fit the prediction model, as described above, the error will be underestimated and overly optimistic. The next section describes measures taken to find an improved prediction equation for the data, and a more realistic estimate of the error of prediction.

Fitting an Improved Model

Extraordinary data points or outliers can cause serious bias in regression coefficients determined by the method of least squares. By examination of the distribution of the response the independent variables and the residuals (or differences in actual response values and predictions from the model), the outliers can usually be detected. Using the first 12 lines of data in Table 1, no outliers were discovered that had a large effect on the model coefficients. This could have been due to the fact that each line of data represented an aggregate or monthly total, which smoothed out high frequency variation.

Non-linearity of the relationship between the response and predictor variables can be examined through partial residual plots [Faraway (2005)]. Partial residual plots for variables X_1 through X_5 revealed strictly linear relationships confirming the form of the simple equation (1) as appropriate for the prediction equation.

The variance inflation factors (VIF) were calculated [Faraway (2005)] to determine if there was a problem with multicollinearity among the predictor variables. The VIF coefficients ranged from 1.4 to 2.92. Usually a VIF coefficient greater than 10 indicates there is a problem with collinearity of the predictors.

A test for serial correlation in the data was made using the Durbin-Watson statistic [Kutner, Nachtsheim and Netter (2004)]. It was insignificant, indicating the data are essentially independent from month to month. Again this is probably due to the monthly aggregation.
Normality of the errors of prediction was checked using a normal probability plot and the Wilk-Shapiro (1965) test statistic. Based on these, the normality assumption appeared justified. However, the purported 3.56% error in prediction was suspect due to the fact that the statistical significance of the coefficients in the prediction equation had not been checked and the estimate of prediction error was made with the same data used to fit the equation.

Table 3 shows a summary of the prediction model fit to the data using the lm function in the statistical package R. The coefficients are the same as those obtained with the numerical solver, but in addition to finding the coefficients the lm function in R produces t-statistics to test their statistical significance. The coefficients for X_3 and X_4 are not marked by an asterisk indicating that these two coefficients are not statistically significantly different than zero at the 0.05 significance level. Even though the sign (-, +) on these two coefficients agrees with the basic fundamental knowledge, it would be better to exclude them from the model to improve the accuracy of the prediction equation.

Table 3. Summary of lm fit in statistical package R

```
Call:
lm(formula =
        EnergyC ~ X1 + X2 + X3 + X4 + X5, data = base)

Residuals:
Min       1Q     Median     3Q     Max
-1232.1  -332.2   -12.0   529.1   785.3

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept)4592.706   1655.711   2.774 0.032258 *
X1           10.722      3.169   3.383 0.014795 *
X2          -10.918      3.226  -3.385 0.014773 *
X3           -2.145      1.351  -1.588 0.163475
X4           10.503      7.960   1.319 0.235119
X5           62.009      9.219   6.726 0.000525 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01
                '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 783.1 on 6 degrees of freedom
Multiple R-Squared: 0.9012,
Adjusted R-squared: 0.8189
F-statistic: 10.95 on 5 and 6 DF,  p-value: 0.005637
```

The best means of validating a prediction equation, and estimating the error in prediction, is to test the equation on newly collected data. That was impossible for the present study, since data from the plant after the 12 month base period would not be representative of the way the plant was operated in the base period. An alternative is to split the data at hand. When data are collected over time it is often useful to split the data at some point in time and use the earlier data to fit the model and the later data to validate the model [Kutner, Nachtsheim and Neter(2004)]. For this reason the data in the base period were split into a training sample and a validation sample as shown in Table 4.

If more data were available in this study, another approach to validation would be to actually repeat the model building process using each half of the data to see

if the same subset of variables with similar coefficients were identified in both.

Comparison of the two Models

Table 5 shows a comparison of the model coefficients determined from the entire base period and from the training set only. Even though all coefficients have the same sign they would be predicted to have from the basic fundamental knowledge, it can be seen that the coefficients from the five variable model (especially the coefficient for X_3) change quite a bit when data from only the training set are used, while the coefficients in the three variable model are relatively consistent. This is another indication that X_3 is unnecessary in the model.

Next the equations were compared with respect to their predictive ability. The five variable model and the three variable model were fit to the data in the training set, and then these two models were used to predict the data in the validation set. The box-plots in Figure 2 show a comparison of the errors in prediction (actual values – predicted values) in the validation set from the two models fit with the data in the training set. Here we can see that there is much more variability in the errors of prediction for the five variable model, thus verifying the danger warned of above. This can be quantified using the mean squared prediction error.

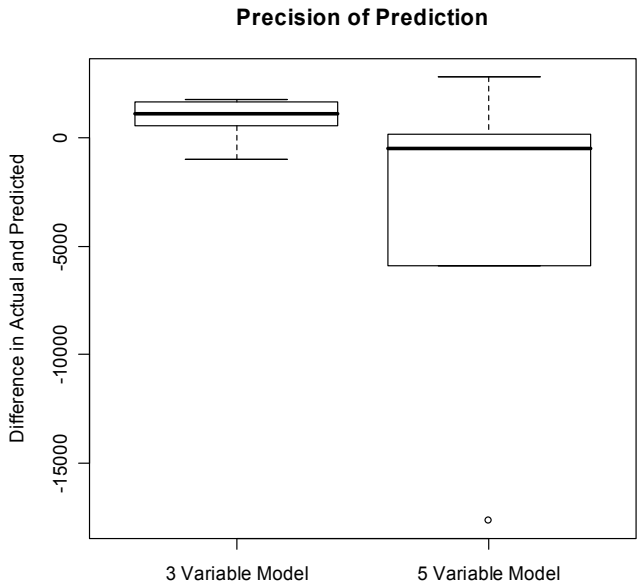

Figure 2. Box Plots of Actual – Predicted for Validation Set

The mean squared prediction error *MSPR* given by $MSPR = \sum_{i=1}^{n^*}(y_i - \hat{y}_i)^2 \Big/ n^*$ is a useful measure for validating the prediction model with n^* additional data points, and $CV = 100 \times \sqrt{MSPR}/\bar{y}$ is a useful measure for quantifying the percent prediction error [Kutner, Nachtsheim and Netter(2004)]. The CV for the 5 variable model fit to the data in the training set and used to

Table 4. Split Base Period Data

Month	Energy Consumed	X_1	X_2	X_3	X_4	X_5	
1	12432	233.5	314.82	860	170.45	140.67	
2	12322.8	260.75	260.57	960	252.49	123.94	
3	14340.6	339.25	125.15	1000	201.31	112.62	
4	8176.7	285	122.27	980	238.79	23.46	Training
5	12937.2	195.5	100.82	1000	227.79	120.06	set
6	12641.9	125.25	62.16	960	119.8	137.31	
7	14262.7	71	17.6	970	121.03	151.04	
8	10325.1	28	201.2	120	104.78	109.86	
9	12028.3	1.25	211.85	940	189.86	144.69	
10	10703.3	15.25	141.78	940	124.02	152.25	Validation
11	13053	59	153.92	850	144.09	147.55	set
12	14444.2	120.5	63.08	760	180.62	148.07	

Table 5. Comparison of Prediction Model Coefficients Fit with All Base Data to Training Set Alone

		Coefficients					
		Intercept	X_1	X_2	X_3	X_4	X_5
5 Variable	All Base Data	4592.706	10.722	-10.918	-2.145	10.503	62.009
Model	Training Set	25808.94	19.245	-26.204	-27.170	20.723	72.844
3 Variable	All Base Data	5155.915	10.582	-8.091	-	-	54.124
Model	Training Set	3817.138	14.834	-10.846	-	-	60.405

predict the data in the validation set is 69.5% while the *CV* for the 3 variable model is only 10.6%. This reduction in prediction error can also be seen in Figure 3 which shows the data (points) for the training set, validation set and the months after improvements were made. The lines on the figure join the predicted values from each model. It can be seen that the predictions for both models are very close to the data in the training set that were used to estimate the model coefficients. However, the predictions for the 3 variable model are much closer to the data in the validation set than are the predictions in from the 5-variable model. The predictions from the 5 variable model appear to be too high in months 8 and 12 in the validation set as well as month 17 and over all in the period after improvements were made. This is due to the fact that these three observations have the lowest values for X_3 in the data set. The coefficient for X_3 in the 5-variable model, fit to the data in the training set, has a large negative coefficient, and thus high predictions resulted for those three data points where X_3 was low.

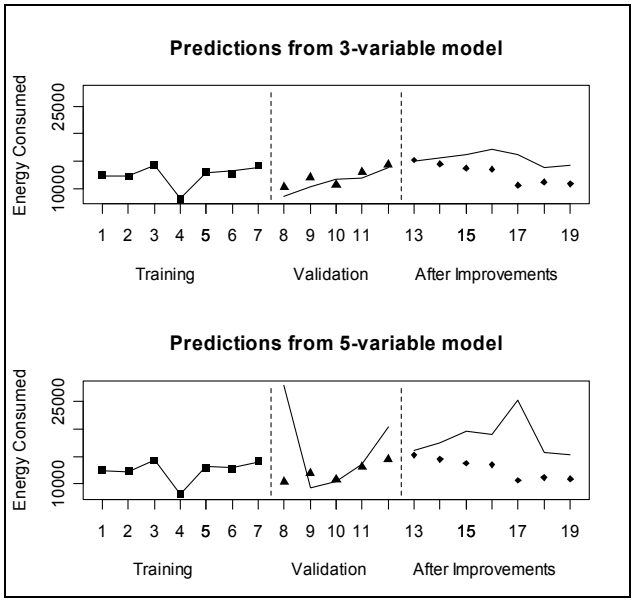

Figure 3. Data in Training Set Prediction Set and After Improvements and Predictions from 3-Variable Model and 5-Variable Model Fit to the Data in the Training Set

Table 6 shows the actual energy consumption and predictions from the 3-variable model for the seven months after energy improvements had been implemented. The difference of actual and predicted is the estimated energy savings resulting from the improvements, and cumulative total is the savings predicted for the entire 7 month period. The estimated energy savings using this model are lower than the same predictions from the 5-variable model shown in Table 2.

Table 6: Actual Energy Consumption and Predictions from the 3 Variable Model in the Period after Energy Savings were Implemented

Month	Actual	Predicted	Predicted Savings
13	15,230.8	14,857.7	-373.1
14	14,506.3	15,290.2	783.95
15	13,751.1	15,434.4	1683.3
16	13516.7	16,432.0	2915.2
17	10,636.6	15,640.4	5003.8
18	11,161.4	13,912.8	2751.4
19	10,896.6	14,248.1	3351.5

Cumulative energy savings predicted for 7-month period = 16,116.0.

Discussion and Conclusions

The conclusion of this study is that care must be taken when developing empirical prediction equations from data to prevent over fitting. Including variables in the equations just because the sign of their coefficient confirms basic fundamental knowledge may be a bad idea. If the coefficients for these variables are not statistically significant, including them in the prediction equations may actually increase the error of prediction as was illustrated in this case study. Splitting the data used to fit the model allows an analyst to fit the model with part of the data and test the model with the other half. This can help to get a more realistic estimate of the error of prediction. Prediction error estimated from the data used to fit the prediction equation will always be biased low.

After splitting the data in this study, the 3-variable model was found to be more appropriate because all of the coefficients in the model are statistically significant. Using the expanded 5-variable model fit to all the data in the base period (12 months), a cumulative energy savings of 18,103 units was predicted for months 13 to 19 after energy savings improvements had been implemented, and the error of prediction was thought to be 3.56% per month. But the predicted energy savings are probably a little too high and the estimate of the error of prediction is too optimistic. Using the 3-variable model fit to all data in the base period, the cumulative energy savings for months 13 to 19 was predicted to be 16,116 units and the error of prediction determined from the validation set (i.e., 10.6% per month) is probably a lot more realistic estimate. The comparison of the predictions for months 13 to 19 from the models fit to all the data in the base period is illustrated in Figure 4.

All the calculations described in this paper to fit the equations, make predictions from the equations and make plots were accomplished using the statistical com-

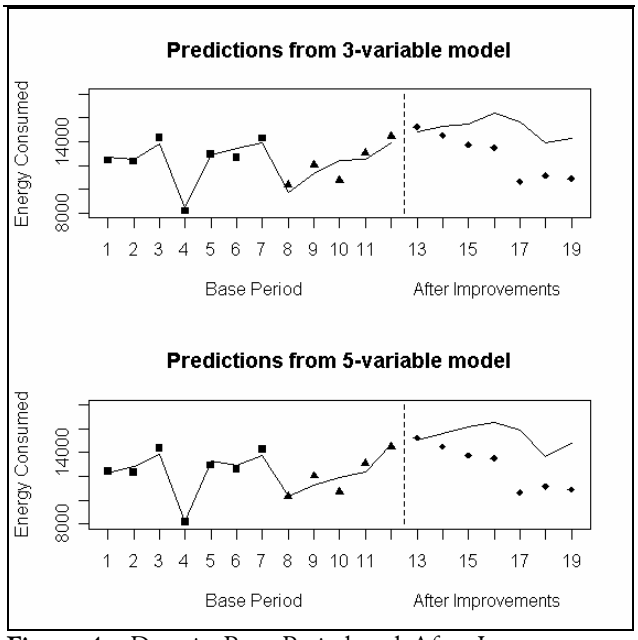

Figure 4. Data in Base Period and After Improvements and Predictions from 3-Variable Model and 5-Variable Model Fit to the Data in the Base Period

puting package R. R is open-source software and may be obtained free of charge. Versions of R for various platforms can be obtained from the R-project at www.r.project.org. The scripts use to fit the equations and make the predictions and plots in this paper are available with this article.

REFERENCES

Faraway, J.J. 2004. *Linear Models with R*. Chapman & Hall/CRC, Boca Raton, FL.

Hair, J.F., R.E. Anderson, R.L. Tatham and W.C. Black. 1995. *Multivariate Data Analysis with Readings*, 4th Edition. Prentice Hall, Upper Saddle River, NJ.

Henderson, H.V. and P.F. Velleman. 1981. Building Multiple Regression Models Interactively, *Biometrics*, 37: 391-411.

Kutner, M.H., C.J. Nachtsheim and J. Neter. 2004. *Applied Linear Regression Models*. McGraw-Hill Irwin, Boston, MA.

Shapiro, S.S. and M.B. Wilk, M.B. 1965. An Analysis of Variance Test for Normality (Complete Samples), *Biometrika*, 52: 591-611.

Correspondence: lawson@byu.edu