How To: Deal with Multicollinearity When Fitting a Regression Model Using STATGRAPHICS Centurion

by

Dr. Neil W. Polhemus

July 25, 2005

Introduction

Multiple regression models are widely used to quantify the relationship between a response variable *Y* and multiple predictor variables. The models take the general form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$$

where

β_0 = a constant

$\beta_1, \beta_2, \ldots, \beta_p$ = regression parameters

ε = random error

There are usually two important reasons for fitting such models:

1. predicting values of Y given known values for the X's.
2. improving one's understanding of how the X variables affect Y.

In the latter case, interest centers on the value of the β's, which link the predictors to the response.

If the data used to fit the regression model come from a designed experiment, there should be little trouble interpreting the coefficients. Good experimental designs are structured to give estimates of the β's which are precise, uncorrelated with each other, and scaled to assist in interpretation. However, if the data come from an undesigned experiment, such as records culled from plant operations, it may be difficult to make sense of the fitted model. This is because predictor variables often change together, as when one increases one ingredient whenever another is decreased. Correlation amongst predictor variables is referred to as *multicollinearity*, which can result in very imprecise regression coefficients that are difficult to interpret.

This guide will examine the problem of multicollinearity and show how a technique called *ridge regression* can be used to help obtain more meaningful regression coefficients.

Sample Data

As an example, we will consider data describing a set of 93 automobiles, compiled by Robin Lock of the Mathematics Department at St. Lawrence University. The data were downloaded from the Journal of Statistical Education (JSE) Data Archive at www.amstat.org/publications/jse/jse_data_archive.html. The file *howto4.sf6* contains the data, several records of which are shown below:

Make	*Model*	*Horsepower*	*Width*	*Weight*	*Gallons*
Acura	Integra	140	5.667	2.705	3.226
Acura	Legend	200	5.917	3.560	4.000
Audi	90	172	5.583	3.375	3.846
Audi	100	172	5.833	3.405	3.846
BMW	535i	208	5.750	3.640	3.333
Buick	Century	110	5.750	2.880	3.226
Buick	LeSabre	170	6.167	3.470	3.571
Buick	Roadmaster	180	6.500	4.105	4.000
Buick	Riviera	170	6.083	3.495	3.704
Cadillac	DeVille	200	6.083	3.620	4.000

Figure 1: First 10 Rows of Sample Data

The columns in the file are:

1. Make of the automobile in that row.
2. Model of the automobile.
3. Maximum horsepower (X_1).
4. Width of the automobile in feet (X_2).
5. Weight of the automobile in 1000's of pounds (X_3).
6. Number of gallons required to travel 100 miles in highway driving (Y).

Note that the Y variable (*Gallons*) has been transformed from the usual metric of miles per gallon to that of gallons per 100 miles. This transformation is commonly used when building models for automobile fuel economy, since the observed relationships with the predictor variables are much more linear than when using miles per gallon.

Step 1: Plot the Data

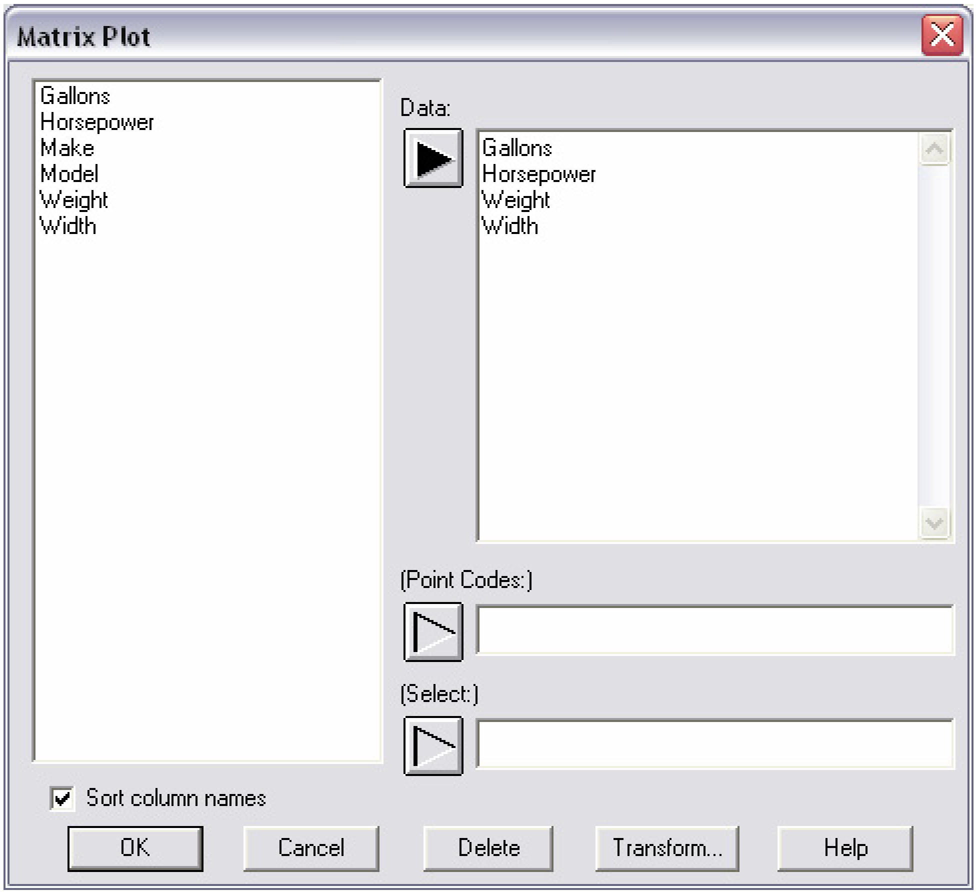
The first step in analyzing any new set of data is to plot it. For multivariate data, a matrix plot is often very useful.

Procedure: Matrix Plot

To create a matrix plot in STATGRAPHICS Centurion:

- If using the Classic menu, select: *Plot – Scatterplots – Matrix Plot.*
- If using the Six Sigma menu, select: *Measure – Scatterplots – Matrix Plot.*

On the data input dialog box, indicate the name of the column containing the data:



Data:
Gallons
Horsepower
Weight
Width

☑ Sort column names

Figure 2: Data Input Dialog Box for Matrix Plot

The resulting plot displays a matrix of X-Y scatterplots, including each pair of variables:

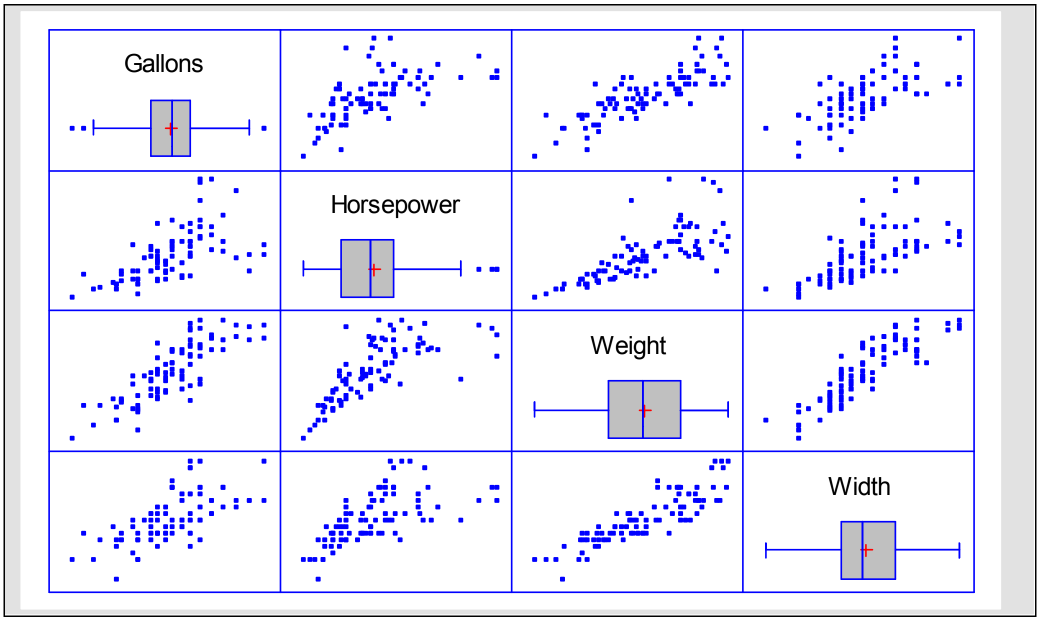
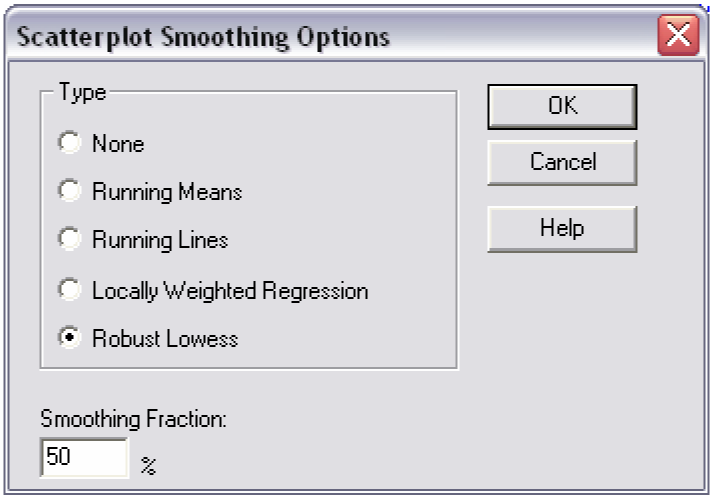


Figure 3: Matrix Plot of Automobile Data

Each variable is plotted on the vertical axis of all plots in its row and on the horizontal axis of all plots in its column. A box-and-whisker plot is also included for each variable.

To assist in visualizing the relationships, press the *Smooth/Rotate* button on the analysis toolbar. When the dialog box appears, request a *Robust LOWESS* smoother:


Scatterplot Smoothing Options

Type

○ None
○ Running Means
○ Running Lines
○ Locally Weighted Regression
● Robust Lowess

Smoothing Fraction: 50 %

Figure 4: Dialog Box for Scatterplot Smoothing

LOWESS stands for *Locally Weighted Scatterplot Smoothing* and is a technique that can be applied to any X-Y scatterplot to help visualize the relationship between the variables plotted on each axis. It is made "robust" or resistant to outliers by smoothing the data twice, down-weighting points far from the first smooth when the second smooth is performed. In this case, it is applied to each plot separately:

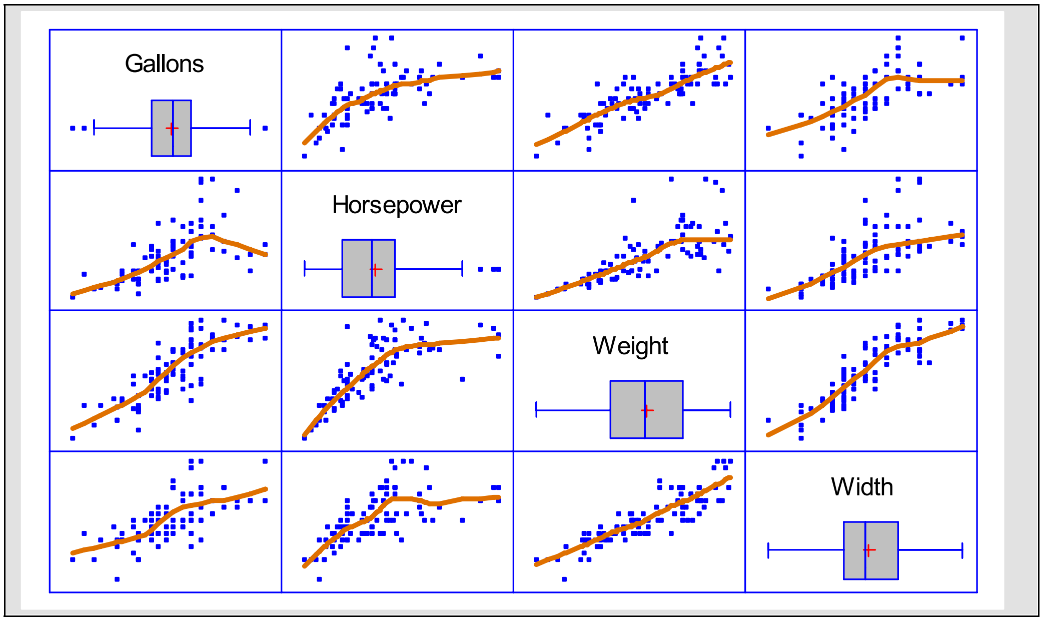


Figure 5: Matrix Plot with LOWESS Smoothes

The top row shows the relationships between *Gallons* and the three predictor variables. In each case, there is a strong positive relationship, with *Gallons* increasing with increased *Horsepower*, *Weight* or *Width*. There is also, however, also a very strong positive relationship between each pair of predictor variables. It is the strong relationship amongst the X variables that is called multicollinearity and causes problems when fitting multiple regression models.

Step 2: Calculate the Correlations

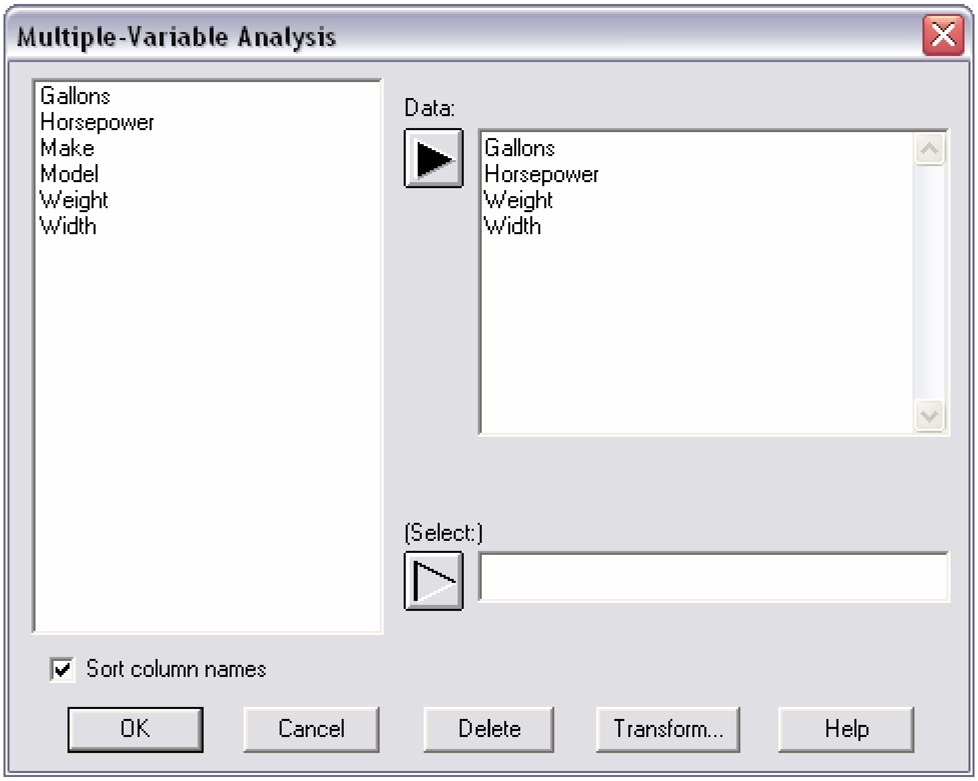
It is also helpful, before fitting a multiple regression model, to calculate the correlations amongst pairs of variables. Correlation coefficients, usually referred to as *r*, measure the strength of the linear correlation between two variables on a scale of -1 to +1. Perfect positive correlation yields a +1. Perfect negative correlation yields a -1. If there is no relationship, r will be close to 0.

Procedure: Multiple Variable Analysis

To calculate correlation coefficients in STATGRAPHICS Centurion:

- If using the Classic menu, select: *Describe – Numeric Data – Multiple-Variable Analysis.*
- If using the Six Sigma menu, select: *Analyze – Variable Data – Multivariate Methods – Multiple-Variable Analysis.*

Complete the data input dialog box as shown below:



Multiple-Variable Analysis

Gallons
Horsepower
Make
Model
Weight
Width

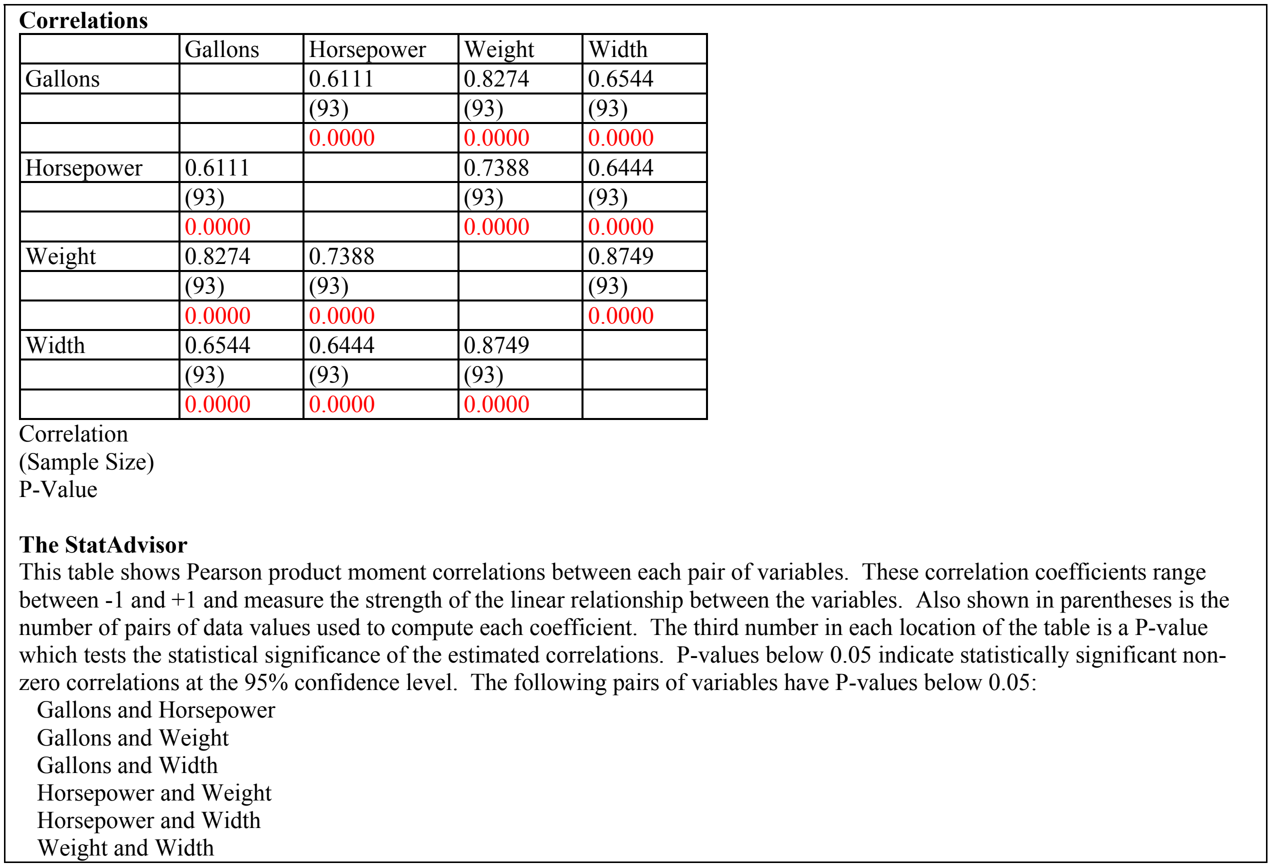
Data:
Gallons
Horsepower
Weight
Width

(Select:)

☑ Sort column names

Figure 6: Data Input Dialog Box for Multiple-Variable Analysis

Part of the default output of the procedures is a table of correlation coefficients:



Correlations

	Gallons	Horsepower	Weight	Width
Gallons		0.6111	0.8274	0.6544
		(93)	(93)	(93)
		0.0000	0.0000	0.0000
Horsepower	0.6111		0.7388	0.6444
	(93)		(93)	(93)
	0.0000		0.0000	0.0000
Weight	0.8274	0.7388		0.8749
	(93)	(93)		(93)
	0.0000	0.0000		0.0000
Width	0.6544	0.6444	0.8749	
	(93)	(93)	(93)	
	0.0000	0.0000	0.0000	

Correlation
(Sample Size)
P-Value

The StatAdvisor

This table shows Pearson product moment correlations between each pair of variables. These correlation coefficients range between -1 and +1 and measure the strength of the linear relationship between the variables. Also shown in parentheses is the number of pairs of data values used to compute each coefficient. The third number in each location of the table is a P-value which tests the statistical significance of the estimated correlations. P-values below 0.05 indicate statistically significant non-zero correlations at the 95% confidence level. The following pairs of variables have P-values below 0.05:

Gallons and Horsepower
Gallons and Weight
Gallons and Width
Horsepower and Weight
Horsepower and Width
Weight and Width

Figure 7: Table of Correlation Coefficients

The table shows the correlation coefficient of each pair of variables, the sample size, and a P-Value. If the P-Value is less than 0.05, then there is a statistically significant correlation between the pair of variables at the 5% significance level.

The strongest correlation (0.8749) is between *Weight* and *Width*, since wide cars also tend to be heavy. *Weight* has a particularly high correlation with *Gallons*, although the correlations of *Horsepower* and *Width* with *Gallons* are also positive and statistically significant.

Step 3: Fit a Multiple Regression Model

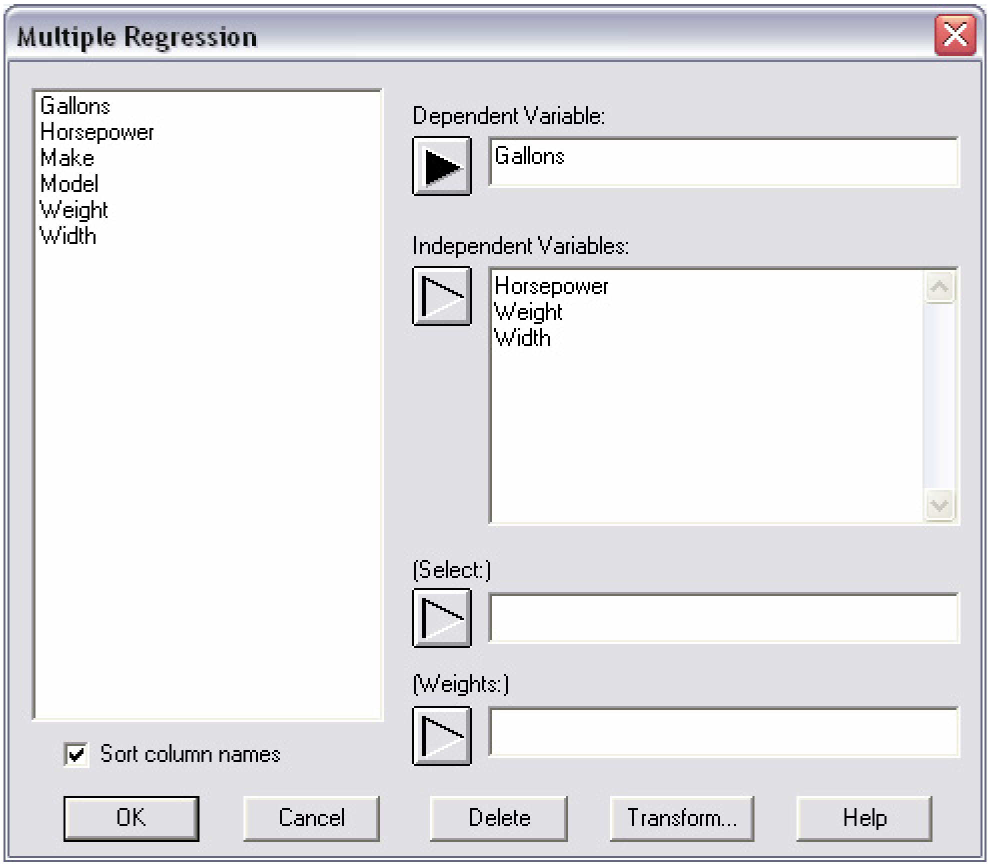
We are now ready to fit the model regression model described earlier.

Procedure: Multiple Regression

To do so:

- If using the Classic menu, select: *Relate –Multiple Factors – Multiple Regression.*
- If using the Six Sigma menu, select: *Improve – Regression Analysis – Multiple Factors – Multiple Regression.*

The dialog input dialog box should be completed as shown below:



Dependent Variable: Gallons

Independent Variables: Horsepower, Weight, Width

(Select:)

(Weights:)

☑ Sort column names

Figure 8: Data Input Dialog Box for Multiple Regression

The top half of the *Analysis Summary* shows the estimated model coefficients:

Multiple Regression – Gallons

Dependent variable: Gallons (per 100 miles)
Independent variables:
    Horsepower (maximum)
    Weight (in 1000 lbs.)
    Width (feet)

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	3.41209	0.986747	3.45792	0.0008
Horsepower	-0.0000211616	0.000963515	-0.0219629	0.9825
Weight	1.09047	0.13508	8.0728	0.0000
Width	-0.556709	0.222944	-2.49708	0.0144

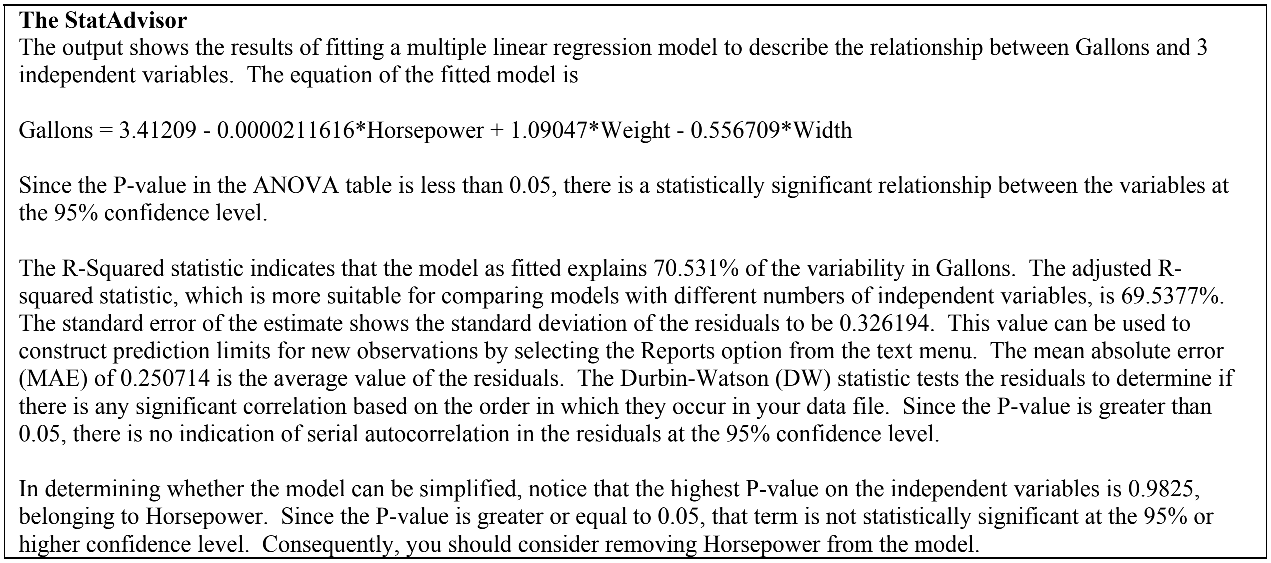
Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	22.6651	3	7.55504	71.00	0.0000
Residual	9.46984	89	0.106403		
Total (Corr.)	32.135	92			

R-squared = 70.531 percent
R-squared (adjusted for d.f.) = 69.5377 percent
Standard Error of Est. = 0.326194
Mean absolute error = 0.250714
Durbin-Watson statistic = 1.84851 (P=0.2159)
Lag 1 residual autocorrelation = 0.0741008

Figure 9: Top Half of Multiple Regression Analysis Summary

The second half of the *Analysis Summary* output is the StatAdvisor, which displays the equation for the fitted model:



```
The StatAdvisor
The output shows the results of fitting a multiple linear regression model to describe the relationship between Gallons and 3
independent variables.  The equation of the fitted model is

Gallons = 3.41209 - 0.0000211616*Horsepower + 1.09047*Weight - 0.556709*Width

Since the P-value in the ANOVA table is less than 0.05, there is a statistically significant relationship between the variables at
the 95% confidence level.

The R-Squared statistic indicates that the model as fitted explains 70.531% of the variability in Gallons.  The adjusted R-
squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 69.5377%.
The standard error of the estimate shows the standard deviation of the residuals to be 0.326194.  This value can be used to
construct prediction limits for new observations by selecting the Reports option from the text menu.  The mean absolute error
(MAE) of 0.250714 is the average value of the residuals.  The Durbin-Watson (DW) statistic tests the residuals to determine if
there is any significant correlation based on the order in which they occur in your data file.  Since the P-value is greater than
0.05, there is no indication of serial autocorrelation in the residuals at the 95% confidence level.

In determining whether the model can be simplified, notice that the highest P-value on the independent variables is 0.9825,
belonging to Horsepower.  Since the P-value is greater or equal to 0.05, that term is not statistically significant at the 95% or
higher confidence level.  Consequently, you should consider removing Horsepower from the model.
```

Figure 10: Bottom Half of Multiple Regression Analysis Summary

Several items are of particular interest:

1. The model as fit represents *Gallons* as a linear function of the three predictor variables. Although all 3 predictors showed a strong positive correlation with *Gallons*, two of the three regression coefficients are negative. The model would seem to imply that increasing either *Horsepower* or *Width* would reduce the *Gallons* needed, which is at best counterintuitive.

2. The P-Values for the regression coefficients in Figure 8 shows that the coefficient on Horsepower is not significantly different from 0. Standard interpretation would be that it could be dropped from the model without significantly hurting the fit. In fact, if you press the *Input Dialog* button on the analysis toolbar and remove *Horsepower* from the model, you get the following fit:

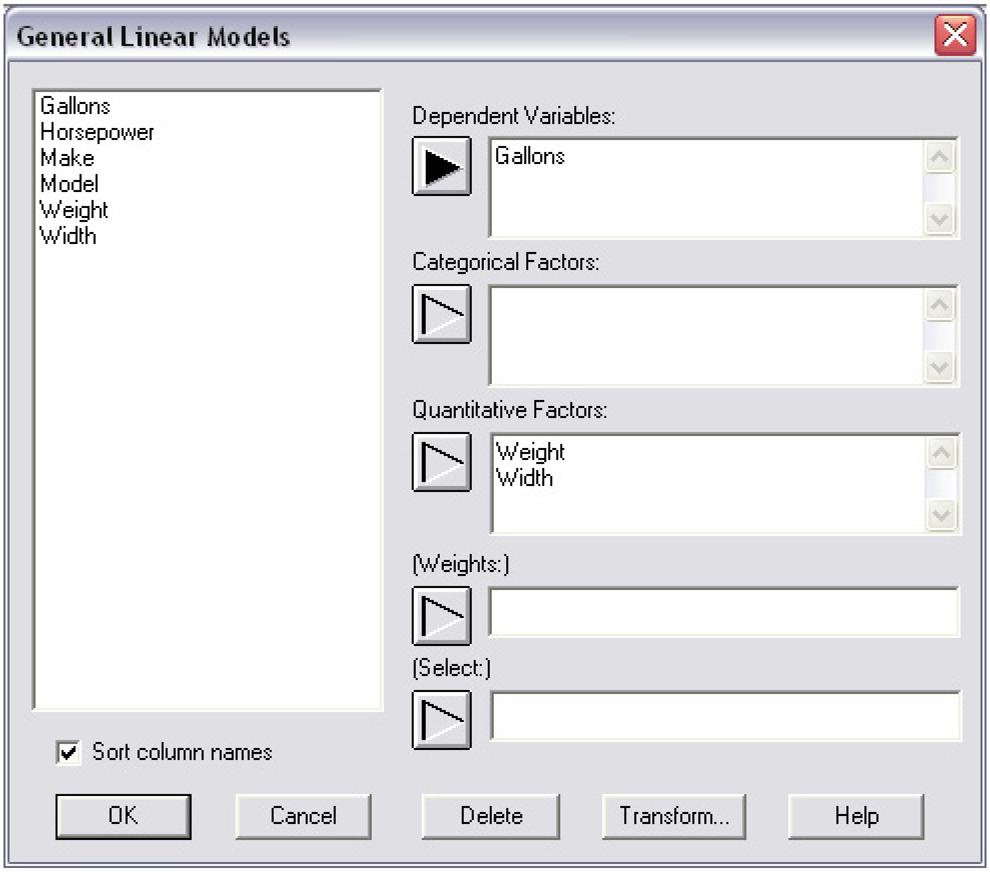
   Gallons = 3.41318 + 1.08907*Weight - 0.556679*Width

The *Adjusted R-Squared* statistic, which measures the proportion of variability in Y that has been explained by the model, actually rises from 69.54% to 69.88% when *Horsepower* is removed. It thus seems quite clear that whatever information *Horsepower* provides as a predictor variable for *Gallons* is redundant given the information in *Weight* and *Width*. That is all well and good if one is constructing a model for predicting the fuel used by similar automobiles, but it lends little to our understanding of the real relationships between the variables.

Procedure: General Linear Models
It is instructive to plot the model relating *Weight* and *Width* to *Gallons*. This is most easily done by:

- If using the Classic menu, select: *Relate –Multiple Factors – General Linear Models.*
- If using the Six Sigma menu, select: *Improve – Analysis of Variance – General Linear Models.*

To reproduce the analysis described above, specify *Weight* and *Width* as *Quantitative Factors*:



General Linear Models

Gallons
Horsepower
Make
Model
Weight
Width

Dependent Variables: Gallons

Categorical Factors:

Quantitative Factors: Weight
Width

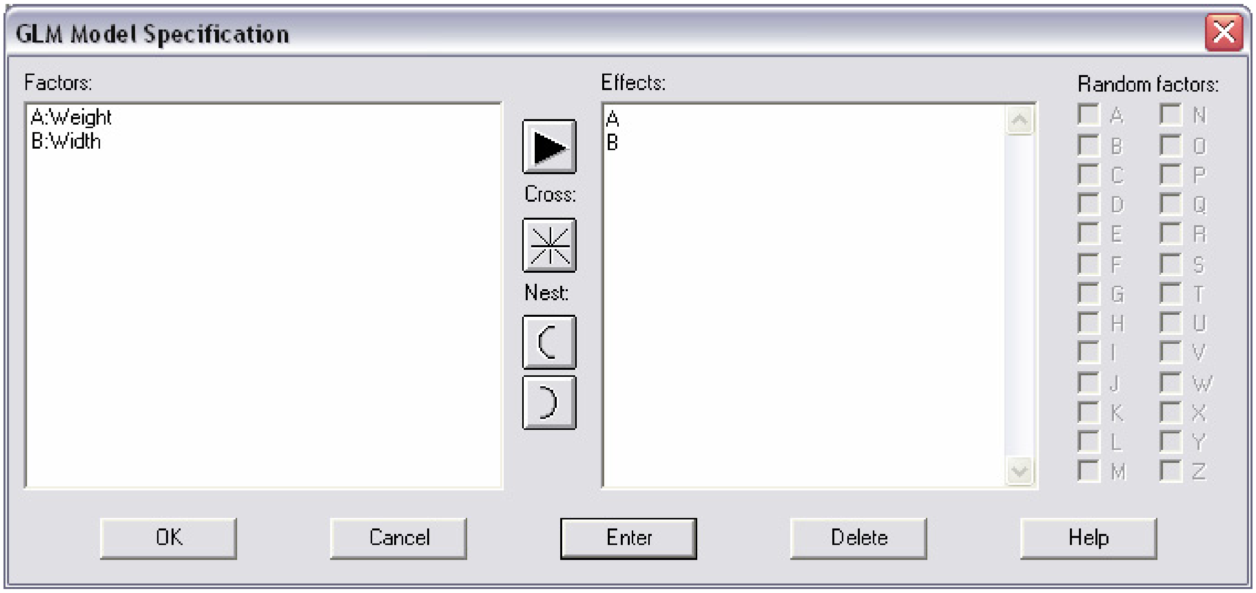
(Weights:)

(Select:)

☑ Sort column names

Figure 11: Data Input Dialog Box for General Linear Models

The default entries on the second dialog box correspond to a multiple linear regression model:



GLM Model Specification

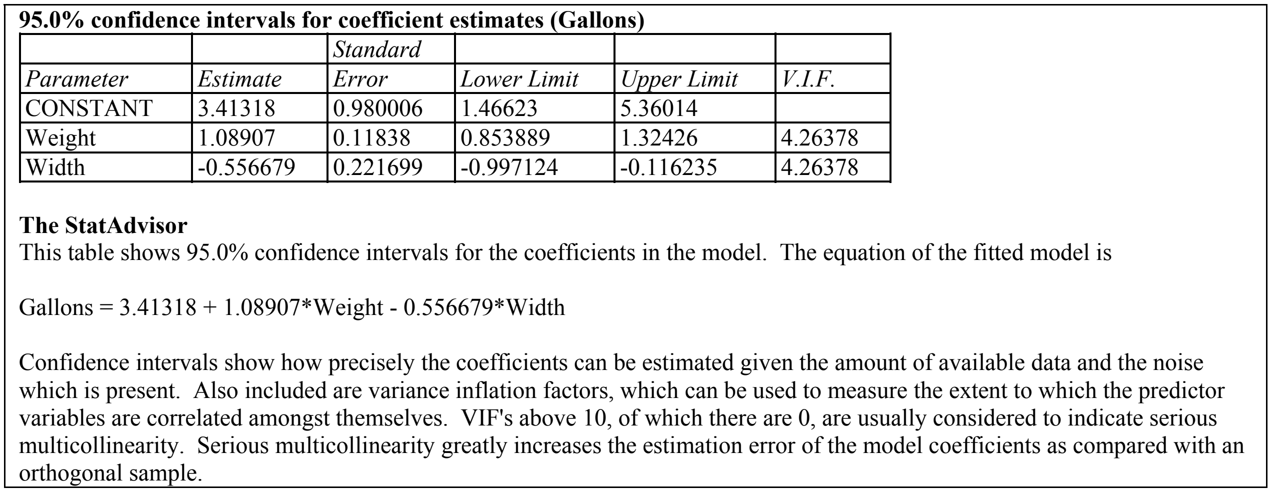
Factors:
A:Weight
B:Width

Effects:
A
B

Random factors:
□ A □ B □ C □ D □ E □ F □ G □ H □ I □ J □ K □ L □ M
□ N □ O □ P □ Q □ R □ S □ T □ U □ V □ W □ X □ Y □ Z

Figure 12: Model Specification Dialog Box for General Linear Models

After pressing OK, the procedure will fit the same model as fit by the *Multiple Regression* procedure earlier. It provides some additional information, however. If you press the *Tables* button on the analysis toolbar and select *Model Coefficients*, the following table will be displayed:



95.0% confidence intervals for coefficient estimates (Gallons)

		Standard			
Parameter	*Estimate*	*Error*	*Lower Limit*	*Upper Limit*	*V.I.F.*
CONSTANT	3.41318	0.980006	1.46623	5.36014	
Weight	1.08907	0.11838	0.853889	1.32426	4.26378
Width	-0.556679	0.221699	-0.997124	-0.116235	4.26378

The StatAdvisor

This table shows 95.0% confidence intervals for the coefficients in the model. The equation of the fitted model is

Gallons = 3.41318 + 1.08907*Weight - 0.556679*Width

Confidence intervals show how precisely the coefficients can be estimated given the amount of available data and the noise which is present. Also included are variance inflation factors, which can be used to measure the extent to which the predictor variables are correlated amongst themselves. VIF's above 10, of which there are 0, are usually considered to indicate serious multicollinearity. Serious multicollinearity greatly increases the estimation error of the model coefficients as compared with an orthogonal sample.

Figure 13: Table of Estimated Model Coefficients

The rightmost column contains *Variance Inflation Factors*, which quantify how much the variability of the estimated model coefficients has been increased compared to a dataset in which the factors were selected so as to be orthogonal to each other. In this case, the variance of each coefficient is inflated by a factor of approximately 4.25.

If you press the Graphs button on the analysis toolbar, you will now see an option to create a *Surface Plot*. The first rendition of the plot is not very helpful, since it views the surface from a poor angle:

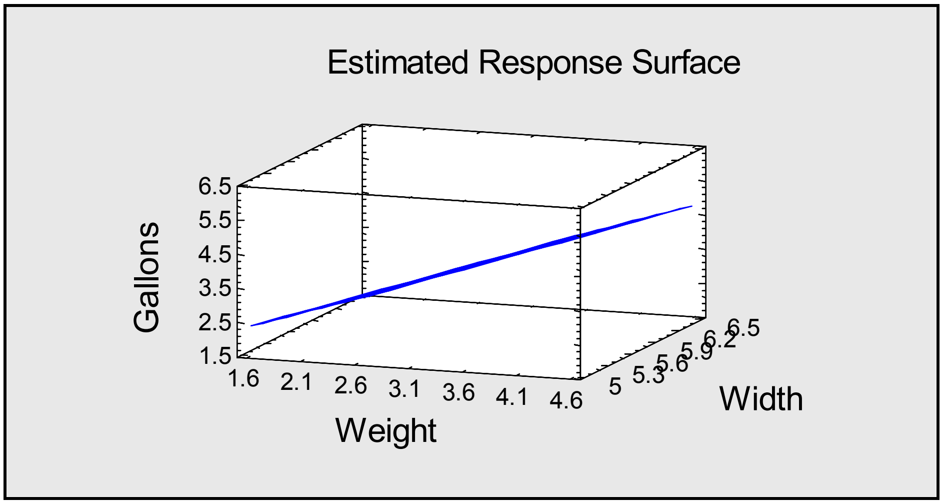
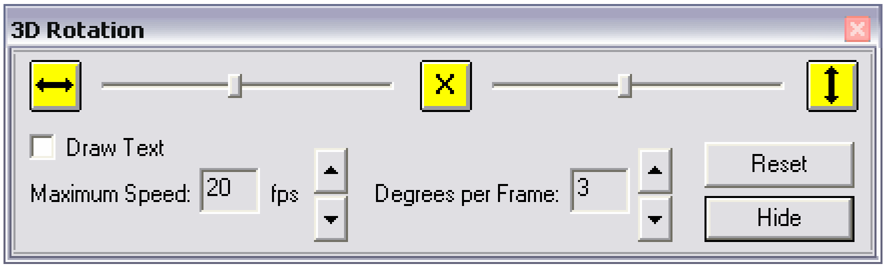


Figure 14: Plot of Fitted Model from Initial Viewpoint

About all you can see is that it is a plane. If you press the *Smooth/Rotate* button on the analysis toolbar, however, a floating rotation dialog bar will appear:



□ Draw Text

Maximum Speed: 20 fps

Degrees per Frame: 3

Figure 15: Rotation Dialog Bar

You can use the arrow keys to rotate the plot in real time, or you can click on a slider and then use the cursor keys to rotate the plot. A much better viewpoint is shown below:

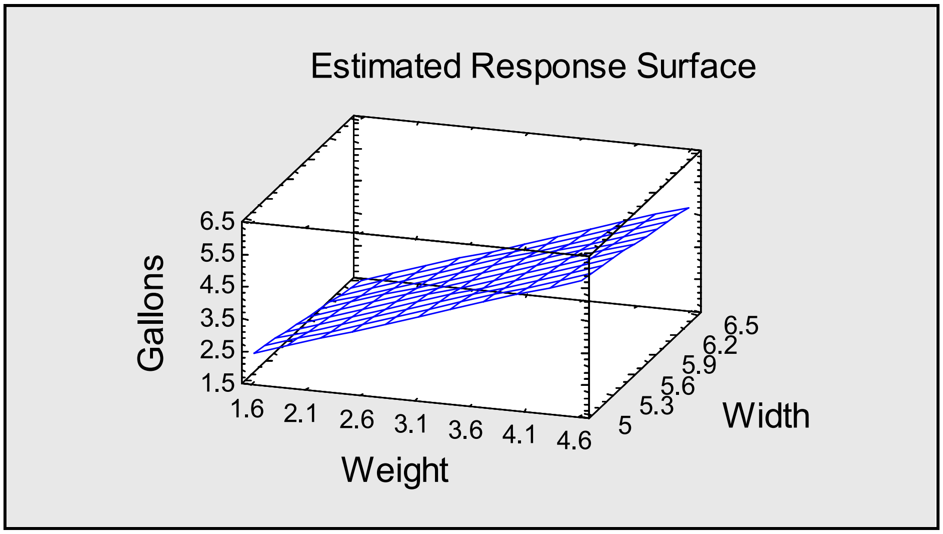
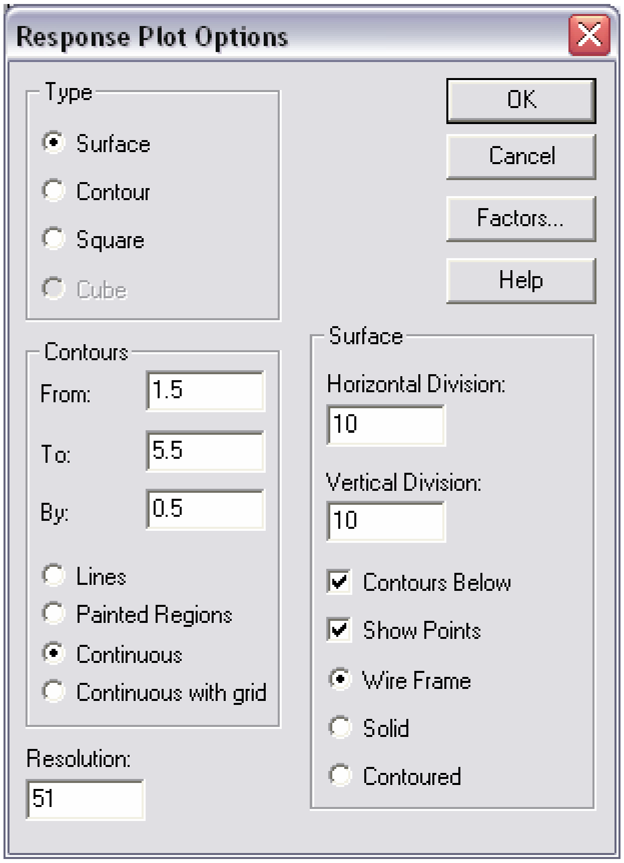


Figure 16: Plot of Fitted Model from Better Viewpoint

Note the strong positive slope in the direction of increasing *Weight* and the negative slope with increasing *Width*.

If you now press the *Pane Options* button on the analysis toolbar, you can add additional features to this plot:


Response Plot Options

Type

● Surface
○ Contour
○ Square
○ Cube

Contours

From: 1.5
To: 5.5
By: 0.5

○ Lines
○ Painted Regions
● Continuous
○ Continuous with grid

Resolution: 51

Surface

Horizontal Division: 10
Vertical Division: 10

☑ Contours Below
☑ Show Points
● Wire Frame
○ Solid
○ Contoured

Figure 17: Response Plot Pane Options

In the above dialog box, we have:

1. Requested *Contours Below*, which draws a contour map in the base of the cube.
2. Asked to *Show Points*. This option draws each data value as a point symbol and drops a vertical line to the surface.

The modified plot is shown below:

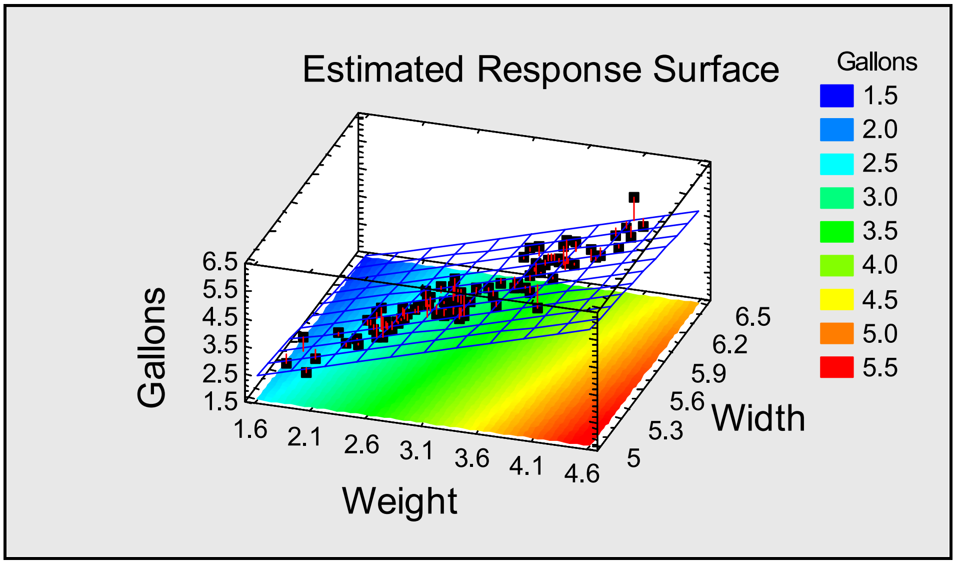


Figure 15: Response Plot With Added Features

The primary feature to notice is that all of the points lie close to a diagonal line extending from lower left corner to upper right corner. As *Weight* and *Width* increase together, so does *Gallons*. Many surfaces could give nearly as good a representation of the observed responses. Such surfaces could be obtained by pulling up on the back left corner and pushing down on the front right. This includes surfaces with positive slopes in the direction of increasing *Width*. In summary, the slopes in the directions of increasing *Weight* and *Width* are very poorly determined because of the high correlation between the predictors variables.

Step 4: Perform a Ridge Regression

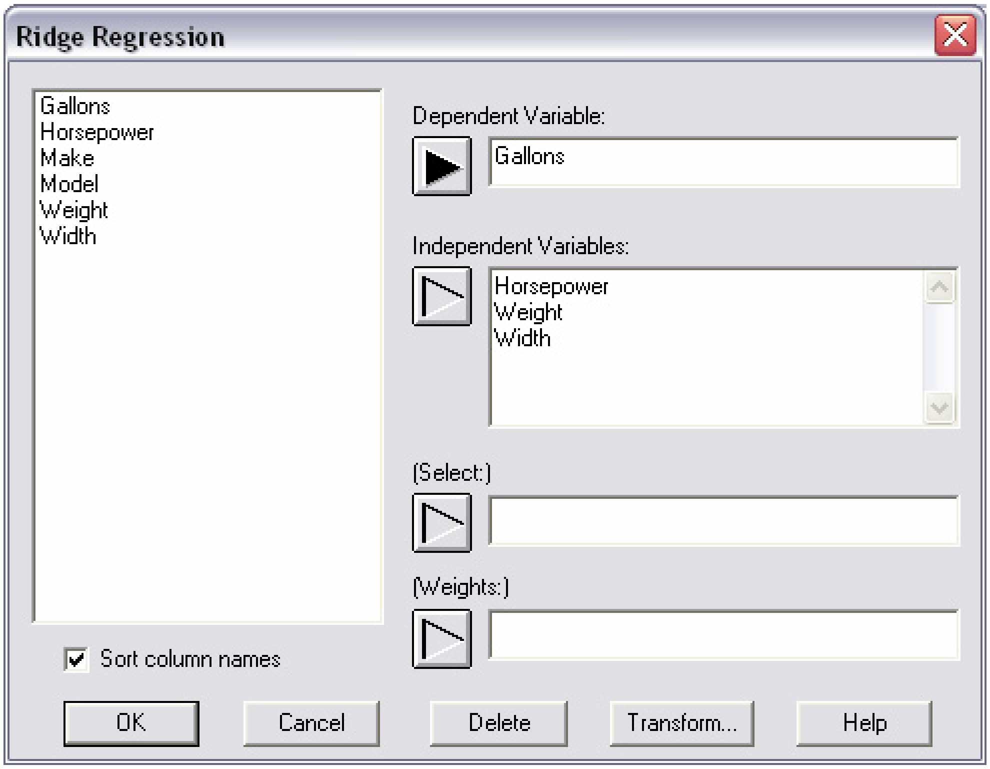
The *Multiple Regression* and *General Linear Models* procedures use the method of least squares to fit the regression model. That is, they find the model that miminizes the sum of the squared residuals, where a residual is defined as the difference between an observed value of Y and the value predicted by the fitted model. If the residuals are independent and follow a Gaussian distribution with constant variance, this leads to estimates of the model coefficients that have the smallest variance among all *unbiased* estimators.

In cases of high multicollinearity, there are often *biased* estimators with considerably less variance than those obtained via least squares. Ridge regression is a method by which one can systematically introduce a small amount of bias and obtain more precise estimates. Often, the resulting models make more sense intuitively.

Procedure: Ridge Regression
To perform a ridge regression using STATGRAPHICS Centurion:

- If using the Classic menu, select: *Relate – Multiple Factors – Ridge Regression.*
- If using the Six Sigma menu, select: *Improve – Analysis of Variance – Ridge Regression.*

The data input dialog box is shown below:


Ridge Regression

Gallons
Horsepower
Make
Model
Weight
Width

☑ Sort column names

Dependent Variable:
Gallons

Independent Variables:
Horsepower
Weight
Width

(Select:)

(Weights:)

Figure 19: Data Input Dialog Box for Ridge Regression

Note that we have restored the original three predictor variables.

Two plots are of particular interest here. The first plots the variance inflation factors as a function of the "ridge parameter":

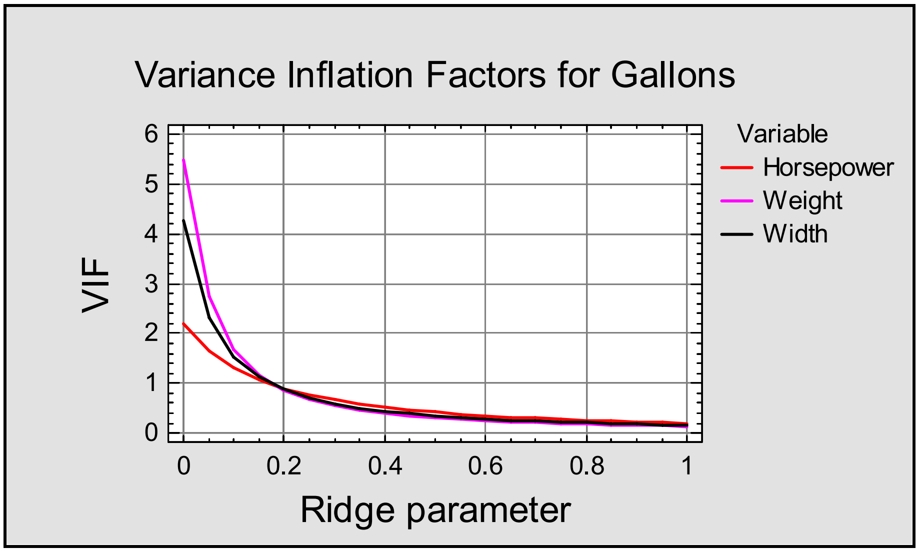


Figure 20: Plot of Variance Inflation Factors

The ridge parameter controls how much bias is introduced into the estimated model coefficients, with the value 0 corresponding to the unbiased least squares estimators. Note that the VIF decreases dramatically as the ridge parameter moves away from 0, falling to 1.0 at about 0.17. At the same time, the *Ridge Trace* shows the change in the standardized model coefficients (scaled to be independent of the units in which the X variables are measured):

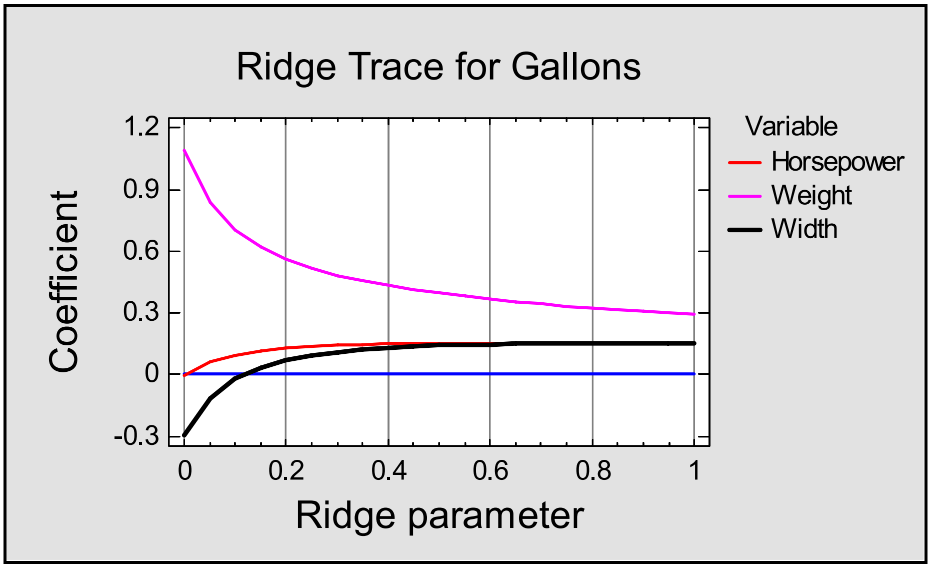


Figure 21: Plot of Model Coefficients

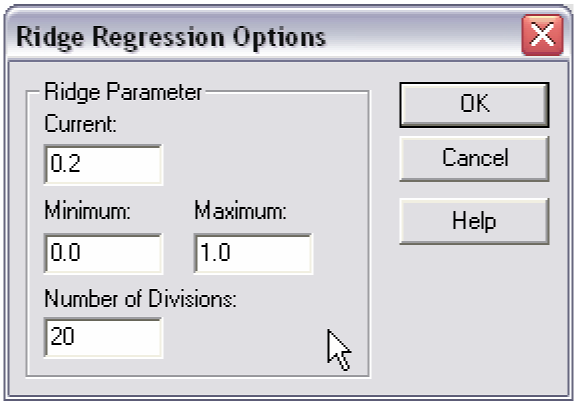
Notice that:

1. The estimated coefficient on *Weight* falls quickly as the ridge parameter moves away from 0.

2. The estimated coefficient for *Horsepower*, which was originally negligible, takes on a much more significant positive value.

3. The estimated coefficient for *Width* switches from negative to positive.

Generally, it is desirable to select a value for the ridge parameter where the VIF has become sufficiently small and the model coefficients have stopped changing dramatically. This appears to be the case in the neighborhood of 0.2. To select a final value for the ridge parameter, press the *Analysis Options* button on the analysis toolbar to display:


Ridge Regression Options

Ridge Parameter

Current: 0.2

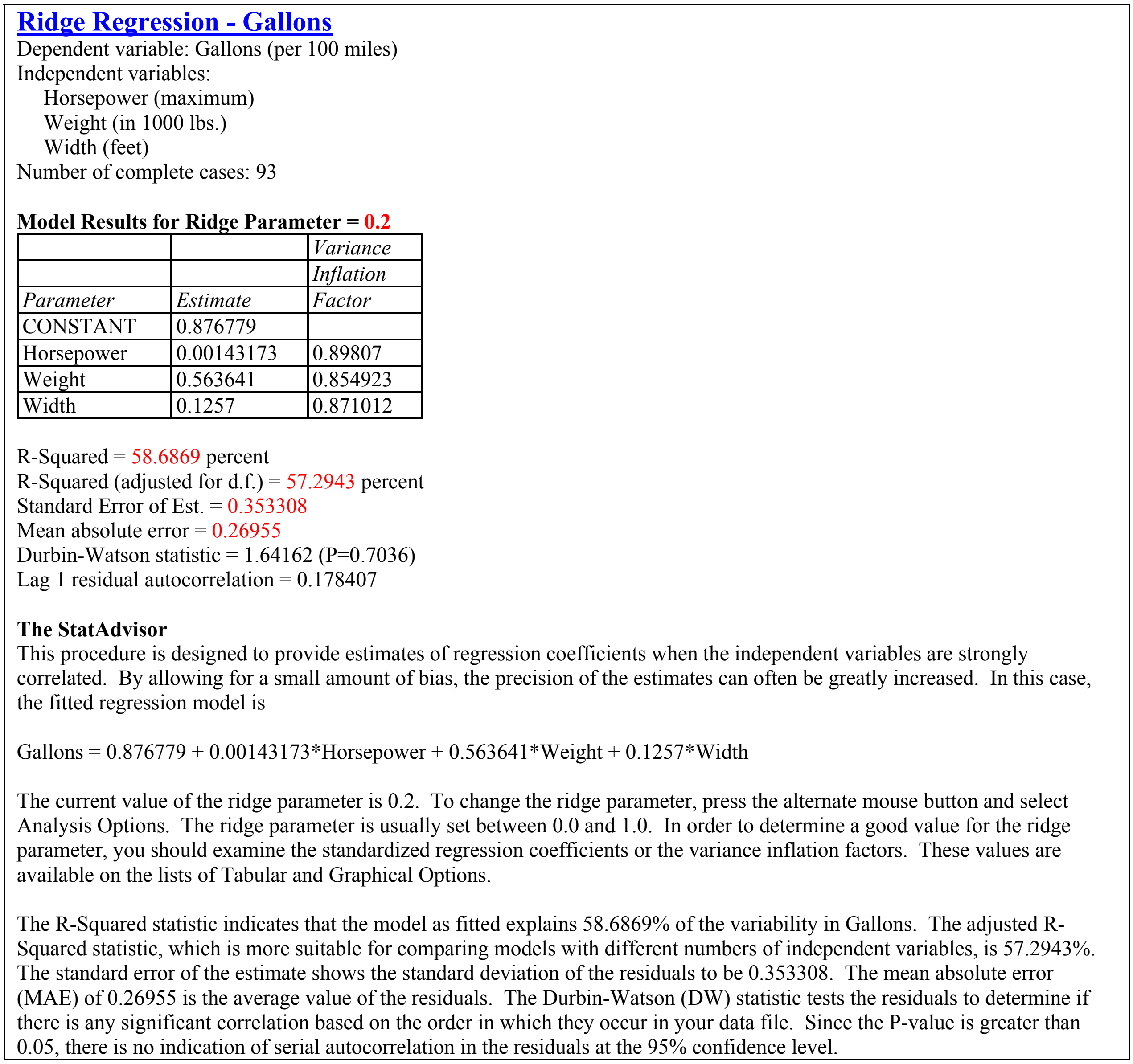
Minimum: 0.0

Maximum: 1.0

Number of Divisions: 20

Figure 22: Ridge Regression Analysis Options

The A*nalysis Summary* will then display the equation of the fitted model:



Ridge Regression - Gallons

Dependent variable: Gallons (per 100 miles)
Independent variables:
    Horsepower (maximum)
    Weight (in 1000 lbs.)
    Width (feet)
Number of complete cases: 93

Model Results for Ridge Parameter = 0.2

		Variance
		Inflation
Parameter	*Estimate*	*Factor*
CONSTANT	0.876779	
Horsepower	0.00143173	0.89807
Weight	0.563641	0.854923
Width	0.1257	0.871012

R-Squared = 58.6869 percent
R-Squared (adjusted for d.f.) = 57.2943 percent
Standard Error of Est. = 0.353308
Mean absolute error = 0.26955
Durbin-Watson statistic = 1.64162 (P=0.7036)
Lag 1 residual autocorrelation = 0.178407

The StatAdvisor

This procedure is designed to provide estimates of regression coefficients when the independent variables are strongly correlated. By allowing for a small amount of bias, the precision of the estimates can often be greatly increased. In this case, the fitted regression model is

Gallons = 0.876779 + 0.00143173*Horsepower + 0.563641*Weight + 0.1257*Width

The current value of the ridge parameter is 0.2. To change the ridge parameter, press the alternate mouse button and select Analysis Options. The ridge parameter is usually set between 0.0 and 1.0. In order to determine a good value for the ridge parameter, you should examine the standardized regression coefficients or the variance inflation factors. These values are available on the lists of Tabular and Graphical Options.

The R-Squared statistic indicates that the model as fitted explains 58.6869% of the variability in Gallons. The adjusted R-Squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 57.2943%. The standard error of the estimate shows the standard deviation of the residuals to be 0.353308. The mean absolute error (MAE) of 0.26955 is the average value of the residuals. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the P-value is greater than 0.05, there is no indication of serial autocorrelation in the residuals at the 95% confidence level.

Figure 23: Ridge Regression Analysis Summary

Note that all of the coefficients in the fitted model are now positive, and the variance inflation factors are all less than 1. We now have a predictive model that makes much more intuitive sense.

To visualize the effect of ridge regression, it is best to go back to the model involving only *Weight* and *Width*. A similar analysis using a ridge parameter of 0.2 yields the following model:

$$\mathit{Gallons} = 0.636895 + 0.624469*\mathit{Weight} + 0.170479*\mathit{Width}$$

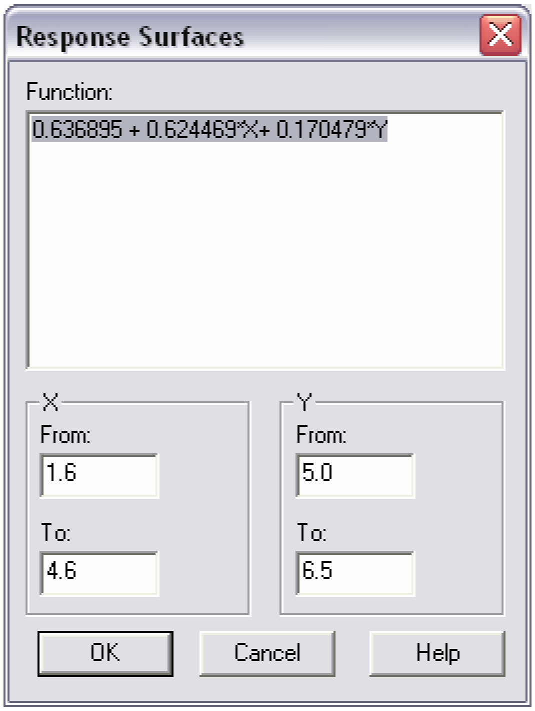
Again, the model coefficients are all positive.

Procedure: Surface Plot

To compare the models generated by *Multiple Regression* and *Ridge Regression*, the *Surface Plot* procedure can be used to plot each model, which can then be copied to the StatGallery. To create a surface plot:

- If using the Classic menu, select: *Plot – Surface and Contour Plots.*
- If using the Six Sigma menu, select: *Tools – Surface and Contour Plots.*

Complete the dialog box as shown below, substituting the coefficients for each model:



Function: 0.636895 + 0.624469*X+ 0.170479*Y

X From: 1.6 To: 4.6

Y From: 5.0 To: 6.5

Figure 24: Response Surfaces Dialog Box

In the equation, enter an X in place of *Weight* and a Y in place of *Width*, since that is the general form expected by this procedure. Also, be sure to set the scales. Then generate each plot and place them side-by-side in the StatGallery as shown below:

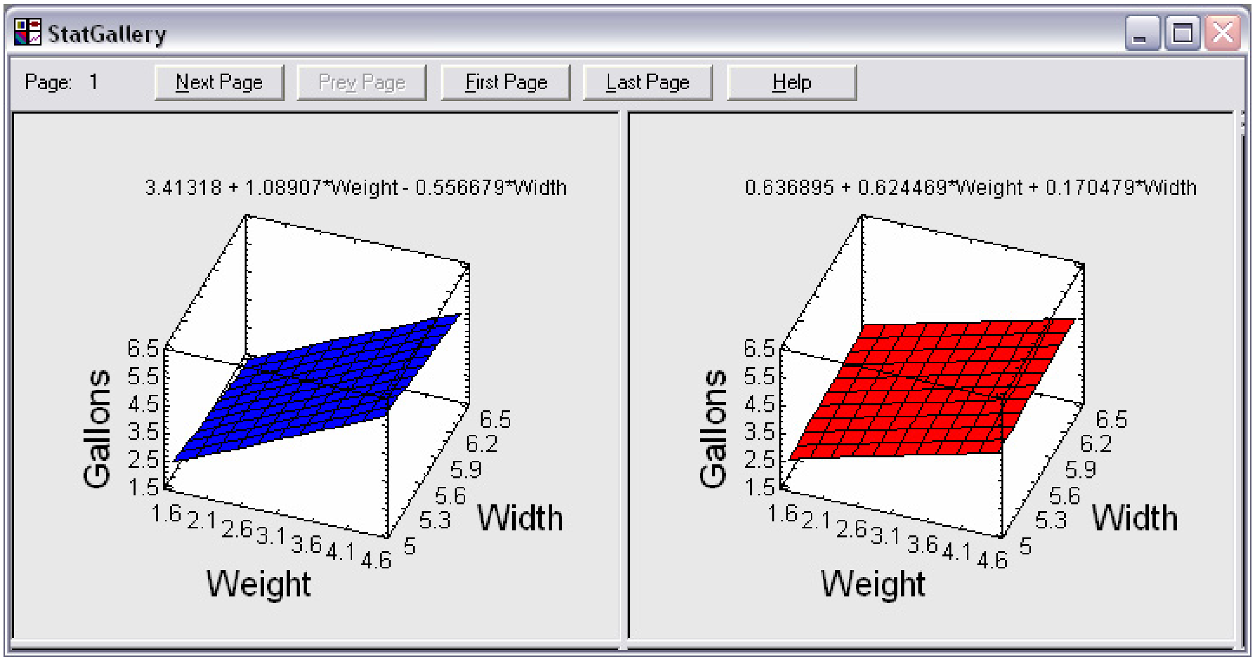


Figure 25: Display of Both Models in the StatGallery

Both models are essentially the same along a diagonal running from lower left to upper right, along which all of the observations were located. The model fit by the *Ridge Regression* procedure (on the right) is much flatter than the other model, however. It does not extend to as extreme values at the upper left and lower right, places where no data exist. If we needed to extrapolate away from the diagonal, we are likely to get much better results with the ridge regression model than with the model determined from least squares.

Conclusion

When collecting data on which to base a multiple regression model, it is not always possible to design the data collection to avoid correlations between predictor variables. In such cases, ridge regression provides a way to obtain precise, meaningful model coefficients. By allowing a small amount of bias in the coefficient estimates, the variability of those estimates can often be reduced dramatically. The resulting models may well give a better understanding of the true relationships in the data.

Note: The author welcomes comments about this guide. Please address your responses to neil@statgraphics.com.