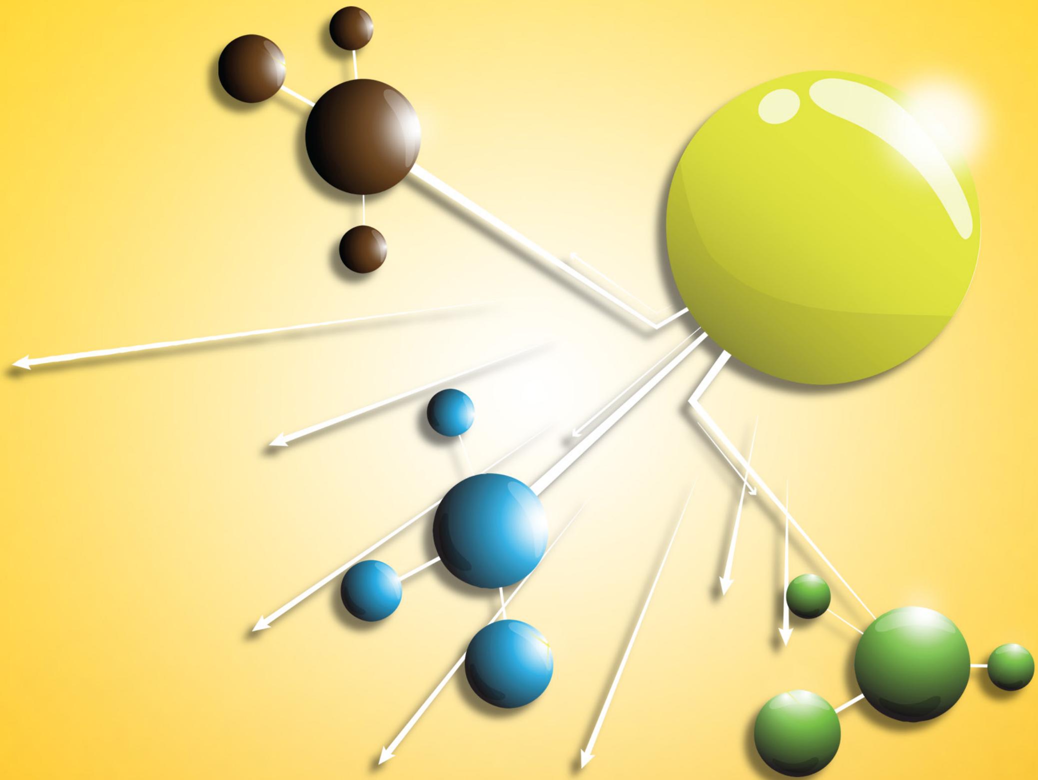
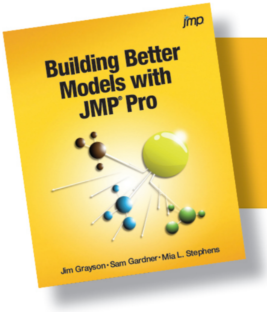
Building Better Models with JMP® Pro

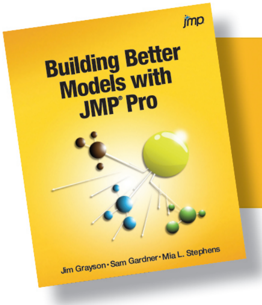


Jim Grayson • Sam Gardner • Mia L. Stephens



From *Building Better Models with JMP® Pro.*
Full book available for purchase here.

Contents



4 Multiple Linear Regression

In the News

These days our entire lives revolve around predictions. Government departments project the cost of health exchanges, the rate of economic growth, next year's crop yields, the future birth rate and the arms buildup of unfriendly countries. Websites and retailers anticipate what we want to find and buy; oil companies gauge the best sites for drilling; pharmaceutical companies assess the probable efficacy of molecules on a disease; while, in the background, the bobble-heads on television incessantly spew out largely irrelevant and inaccurate forecasts. In the meantime, we busy ourselves with personal projections. How long will our commute take? When will the turkey be golden? How much will the price of a stock rise? What will the future value be of a law degree? (Michael Moritz."Are We All Being Fooled by Big Data?" January 3, 2013. Accessed at http://linkd.in/1zFuuq2.)

Representative Business Problems

Multiple linear regression is perhaps the most widely used and well-known statistical modeling tool. A straight-forward extension of simple linear regression, multiple regression is used to predict the average response value based on values of multiple predictors, or factors. For example, multiple regression can be used for:

- Identifying and optimizing critical to quality characteristics, with the goal of developing low cost and high quality products
- Predicting customer spending based on demographic information and historical buying patterns
- Developing pricing strategies based on product mix and consumer characteristics
- Establishing housing prices
- Determining the optimal timing for traffic lights to minimize traffic delays
- Predicting future product success from the results of a pilot study or trial

Preview of End Result

Suppose a model is developed to predict the amount of money a patron will spend on food per day while attending a popular professional golf tournament. A study is conducted and a model is developed based on household income and the average cost of food items sold:

Golf Tournament Daily Spend ($/day) = $25 + 0.08 * Annual Household Income (in thousands) – 0.21 * Average Cost of Food Items

So, if the average cost of food items is $10 and a patron's household income is $100,000, the predicted spending on food for that patron is:

$25 + (0.08 * 100) + (-0.21 * 10) = $30.90 per day.

Looking Inside the Black Box: How the Algorithm Works

Consider points on a scatterplot, where x is some predictor variable and y is some response. For example, the response in the previous section is money spent per day, and one of the x or predictor variables is household income. Now think about drawing a line that best represents or fits these data points. This line is our linear model fit to the data. The line provides a predicted value of y for each value of x.

The equation for this line can be expressed as:

$$\mathrm{y} = b_0 + b_1 x$$

where b_0 is the y-intercept and b_1 is the slope. The y-intercept is the predicted value of y if the value of x is zero, and slope represents the change in y for every unit increase in x. Since we are fitting a model using data, this equation is generally expressed as:

$$\hat{y} = b_0 + b_1 x$$

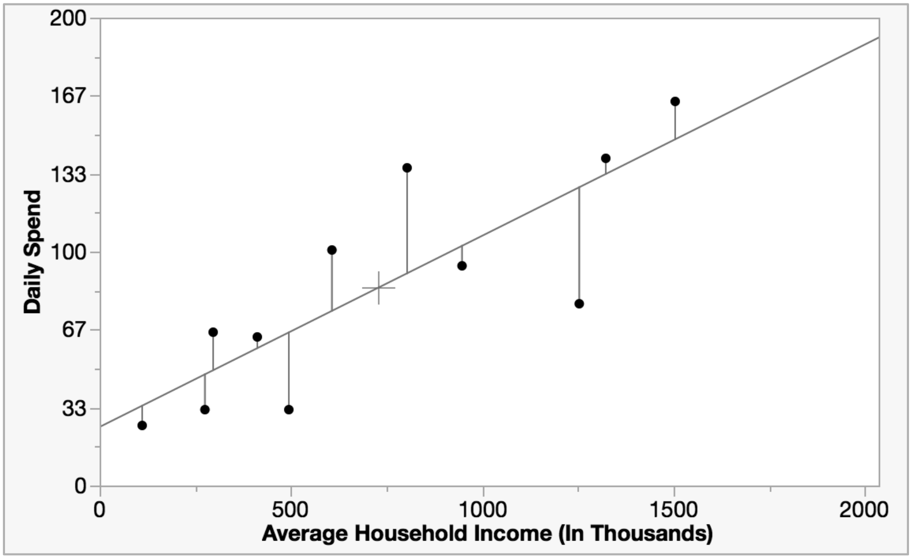
where $\hat{y}$ represents the fitted value of y. This line doesn't fit the data perfectly. There is generally some difference between each response value and the line that we have drawn. This difference is called a residual, and it tells us how far the predicted value is from the observed value. This is illustrated in Figure 4.1. Each point is an observed value (the actual Daily Spend) for a given income level, and the vertical line tells us how far off each point is from the Daily Spend predicted by our linear model.

In fitting a line to the data, we are attempting to model the true unknown relationship between our predictor and our response. That is, we are trying to model reality. Our line is actually an estimate for the model. It represents the true unknown relationship between y and x, which is written as:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Here, y is the response at a given value of x, β_0 is the true y-intercept, β_1 is the true slope, and ϵ represents random error.

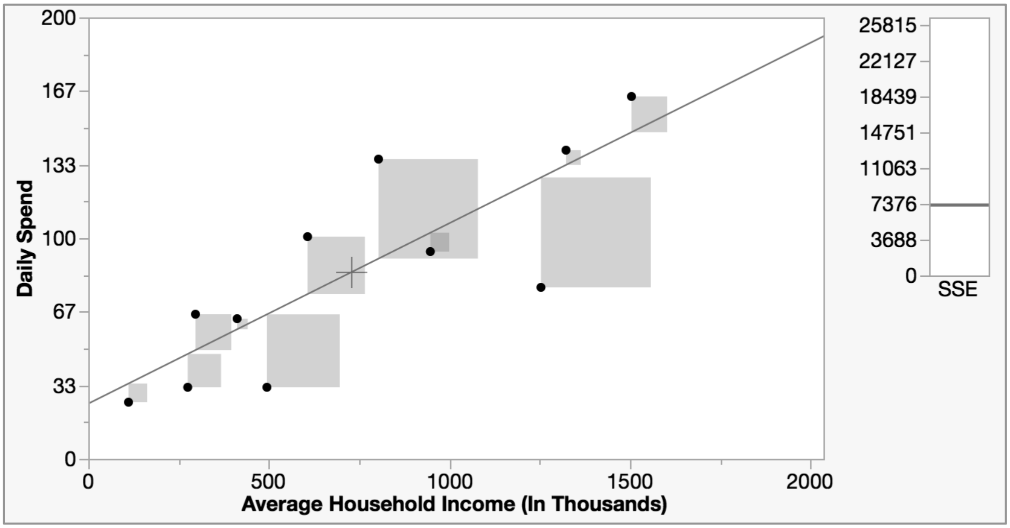
Figure 4.1: Fitting a Line



We estimate this true unknown model by drawing a line through our sample data that best fits the data. How do we measure "best"? Intuitively, the line that produces the smallest residual values is the best fit to our data. Because some residuals are positive and some are negative, we use an algorithm that finds the line with the smallest *sum of squared residuals*. This algorithm is aptly named the *Method of Least Squares*, and the line with the lowest sum of squared residuals, also called sum of squared errors, is referred to as the least squares regression line.

A visual representation of squared residuals is shown in Figure 4.2. Larger residuals have larger squared residuals, represented by pink squares. Smaller residuals have smaller squares. Intuitively, the least squares regression line is the line that results in the smallest squares in terms of total area.

Figure 4.2: Squared Residuals



Simple linear regression, described above, involves one response and one predictor variable. In many situations, there are several variables that can be used to predict the response. In this case, we use the term multiple linear regression, and the formula is a straightforward extension of the simple regression model:

$$y = \beta_0 + \beta_i x_i + \cdots + \beta_k x_k + \epsilon$$

Each predictor has a corresponding slope, or coefficient, and the response is dependent upon the values of each predictor variable. In the original example, our model had two predictors, Annual Household Income (in thousands) and Average Cost of Food Items for Sale.

> ***Note**: Figures 4.1 and 4.2 were created with the* **Demonstrate Regression** *script for illustration purposes. The script is found in JMP (starting with JMP 12) under* **Help > Sample Data > Teaching Scripts > Interactive Teaching Modules**. *Click the* **Help** *button in the script for information on how to use the script to explore fitted lines, residuals, and sums of squares.*

Example 1: Housing Prices

A real estate company that manages properties around a ski resort in the United States wishes to improve its method for pricing homes. Sample data is obtained on a number of measures, including size of the home and property, location, age of the house, and a strength-of-market indicator.

The Data **HousingPrices.jmp**

The data set contains information on about 45 residential properties near a popular North American ski resort sold during a recent 12-month period. The data set is a representative sample of the full set of properties sold during that time period (example provided by Marlene Smith, University of Colorado at Denver). The variables in the data set are:

Price: Selling price of the property (in thousands of dollars)

Beds: Number of bedrooms in the house

Baths: Number of bathrooms in the house

Square Feet: Size of the house in square feet

Miles to Resort: Miles from the property to the downtown resort area

Miles to Base: Miles from the property to the base of the ski resort's facing mountain

Acres: Lot size in number of acres

Cars: Number of cars that will fit into the garage

Years Old: Age of the house at the time it was listed in years

DoM: Number of days the house was on the market before it was sold

Applying the Business Analytics Process

Define the Problem

The real estate company wants to develop a model to predict the selling price of a home based on the data collected. The resulting pricing model will be used to determine initial asking prices for homes in the company's portfolio.

Prepare for Analysis

We begin by getting to know our data. As we saw in Chapter 3, we explore the distributions for each of our variables using **Analyze > Distribution**. We investigate relationships between the response and potential predictor variables using **Analyze > Multivariate Methods > Multivariate**.

Note that in this example, and throughout the modeling chapters, our focus is on particular modeling techniques. In each example, we use only a handful of methods, discussed in Chapter 3 to provide you with some familiarity with the data. However, we recommend that you use the graphical and numeric tools for exploring variables that were introduced, follow the suggestions for data preparation, and have a good understanding of each data set prior to modeling.

Exploring One Variable at a Time

Distribution output for all of the variables is shown in Figure 4.3a and Figure 4.3b. For each of these continuous variables, we see a histogram, a box plot, and various summary statistics.

Notice that the homes in this sample range in price from $160,000 to $690,000. Many of the homes have three or four bedrooms, two or three baths, are under 2,000 square feet on average, and are within twenty miles of the resort.

Figure 4.3a: A First Look at the Data – First 5 Housing Price Variables

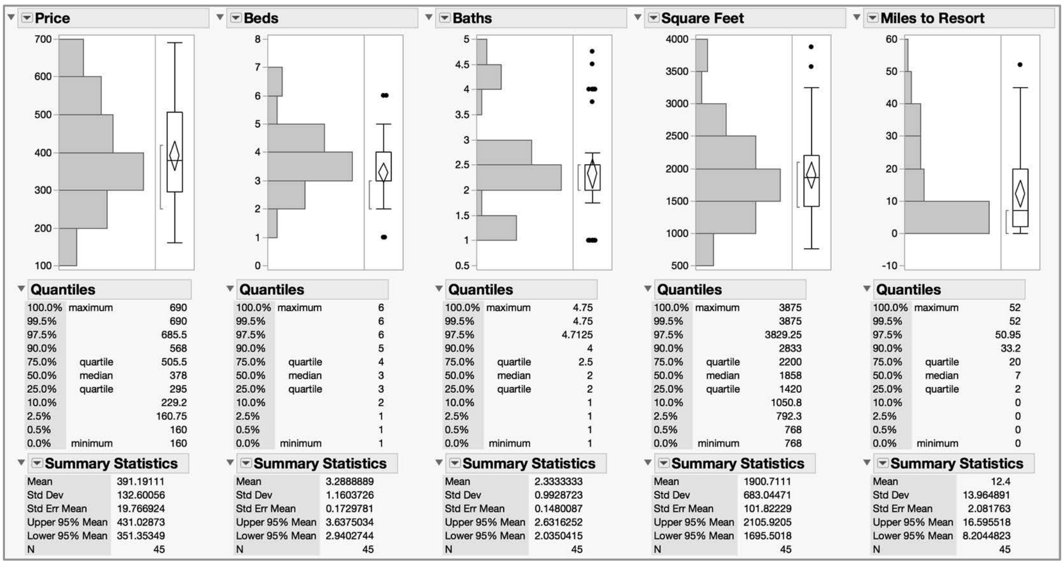
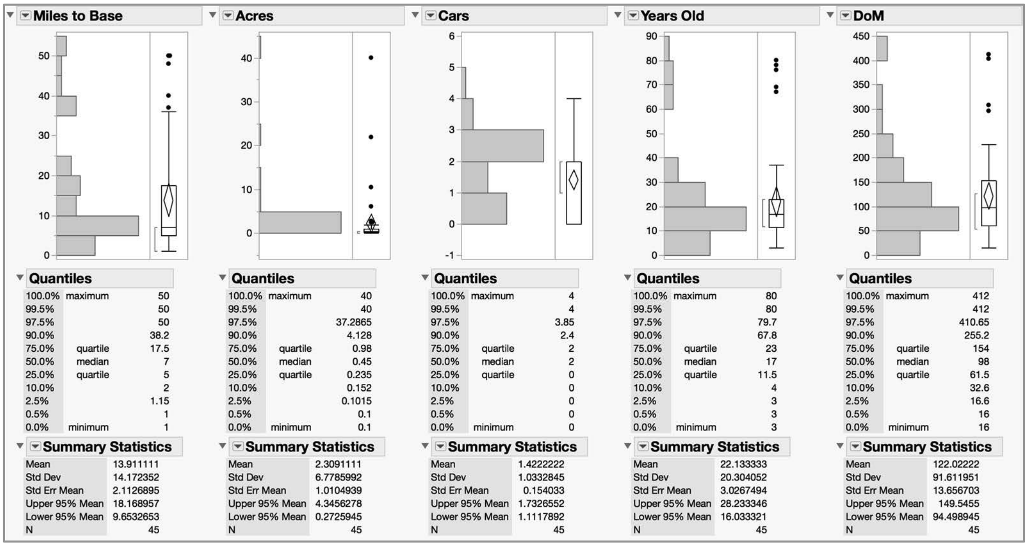
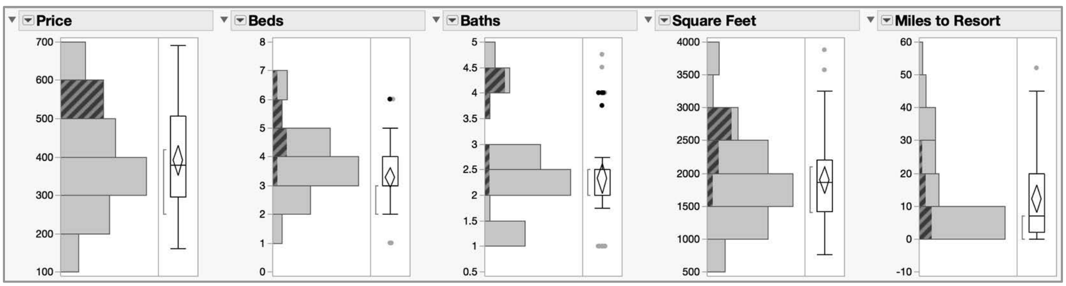


Figure 4.3b: A First Look at the Data – Last 5 Housing Price Variables



When we click on the $500-600,000 bin in the **Price** histogram, the values for these homes are also selected (shaded) in the other graphs. As one might expect, these more expensive homes tend to be on the larger side and are closer to the resort. For example, houses in this price range tend to have four or five bedrooms and three or four baths, are generally between 2000 and 3000 square feet, and are, for the most part, within twenty miles of the resort area.

Figure 4.4: Exploring Relationships



The histograms also tell us about the shapes of the distributions, and whether there are any patterns, unusual observations or potential outliers that may cause concern. **Miles to Resort** (in Figure 4.4) appears to be right-skewed, while the other four variables appear to be more symmetric. While the data set is small, there don't appear to be any unusually large or small values for any of the variables in Figure 4.4.

Exploring Many Variables at a Time

The **Multivariate** platform in JMP can be used for a more formal exploration of the relationships between the predictor variable and potential response variables. The **Correlations** table (Figure 4.5) shows correlations between all of the variables. Strong positive correlations are shown in blue in JMP, strong negative correlations are shown in red, and weak correlations are gray.

We're most interested in the correlation between **Price** and the other variables. We see strong positive correlations between **Price** and the number of beds, baths, and square feet, which are measures of house size, and negative correlations with miles from the base and miles from the resort, which are both distance measures.

We're also interested in understanding potential relationships between the predictor variables. For example, we can see strong correlations between each of the size measures (**Beds**, **Baths**, and **Square Feet**) and between the two distance measures (**Miles to Resort** and **Miles to Base**).

Figure 4.5: Correlations between Variables

Correlations

	Price	Beds	Baths	Square Feet	Miles to Resort	Miles to Base	Acres	Cars	Years Old	DoM
Price	1.0000	0.6753	0.8001	0.6970	-0.5391	-0.6332	0.0251	0.4523	-0.3551	0.2298
Beds	0.6753	1.0000	0.7332	0.7282	-0.3509	-0.4241	-0.1473	0.1045	-0.3403	-0.0971
Baths	0.8001	0.7332	1.0000	0.7901	-0.3745	-0.4880	-0.1930	0.4357	-0.3267	0.1205
Square Feet	0.6970	0.7282	0.7901	1.0000	-0.1895	-0.2972	-0.1456	0.3728	-0.3037	-0.0110
Miles to Resort	-0.5391	-0.3509	-0.3745	-0.1895	1.0000	0.9480	0.2958	-0.1584	0.1082	-0.2219
Miles to Base	-0.6332	-0.4241	-0.4880	-0.2972	0.9480	1.0000	0.2634	-0.2612	0.2211	-0.1841
Acres	0.0251	-0.1473	-0.1930	-0.1456	0.2958	0.2634	1.0000	0.1474	-0.0295	0.3288
Cars	0.4523	0.1045	0.4357	0.3728	-0.1584	-0.2612	0.1474	1.0000	-0.2714	0.3137
Years Old	-0.3551	-0.3403	-0.3267	-0.3037	0.1082	0.2211	-0.0295	-0.2714	1.0000	0.0077
DoM	0.2298	-0.0971	0.1205	-0.0110	-0.2219	-0.1841	0.3288	0.3137	0.0077	1.0000

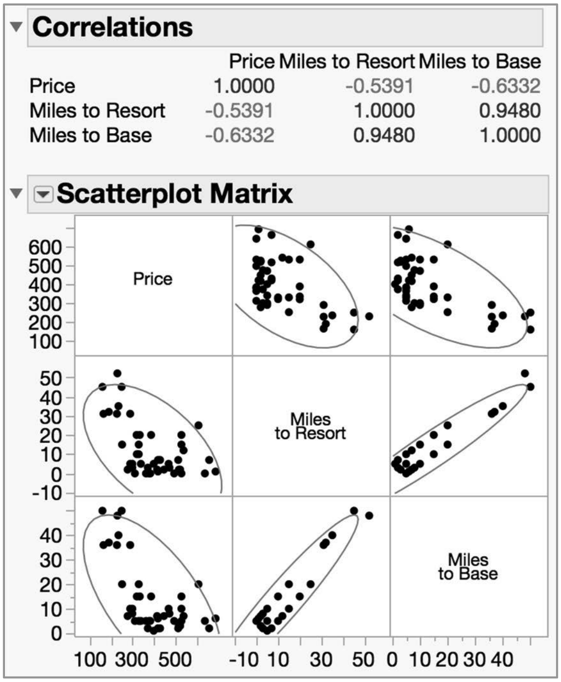
These relationships can be examined visually with the scatterplot matrix, which displays all of the two-way scatterplots between each pair of variables. Figure 4.6 shows the correlations and scatterplot matrix for three of the variables: **Price**, **Miles to Resort**, and **Miles to Base.** In the first row of the matrix, the *y*-axis for each of the graphs is **Price**, and the *x*-axis corresponds to the variable on the diagonal. So, for example, the two scatterplots in the first row display the relationship between **Price** and **Miles to Resort** and between **Price** and **Miles to Base**.

In each scatterplot, the correlation is displayed graphically as a density ellipse. The tighter (less circular) the ellipse, the stronger the correlation. The direction of the ellipse indicates whether the correlation is positive (the ellipse slopes up) or negative (the ellipse slopes down).

Looking at the variables, individually and together, helps us understand our data and potential relationships. We are starting to get a sense of the data and the variables that might need to be included in the model.

Note that other graphical tools, such as **Fit Y by X** (under the **Analyze** menu) and **Graph Builder** (under the **Graph** menu) can also be useful in exploring potential bivariate and multivariate relationships. We urge you to explore this data set on your own using all of these tools.

Figure 4.6: Visually Examining Correlations with a Scatterplot Matrix


Correlations

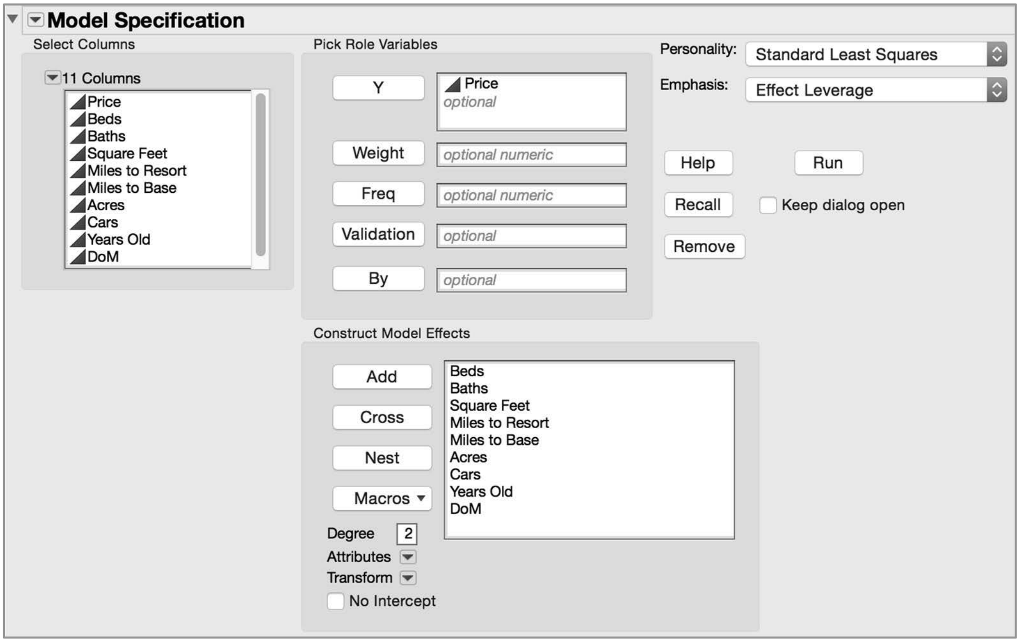
	Price	Miles to Resort	Miles to Base
Price	1.0000	-0.5391	-0.6332
Miles to Resort	-0.5391	1.0000	0.9480
Miles to Base	-0.6332	0.9480	1.0000

Build the Model

Our goal is to develop a model to predict the selling price of a home based on available data. Multiple linear regression is one of the core methods that can be used to develop a model to predict a continuous response from multiple predictor variables.

We begin by fitting a model using **Price** as the **Y** (response variable) and all of the potential factors as model effects using **Analyze > Fit Model**, as shown in Figure 4.7. Click **Run** to run the model.

Figure 4.7: Fit Model Dialog Window



The results are shown in Figure 4.8. The **Effect Summary** table shows each of the terms in the model, sorted in ascending order of the *p*-value.

The **Actual by Predicted Plot** provides a graphical indication of the overall significance of the model. The closer the data points are to the diagonal line, the better our model does at explaining the variation in the response. (The solid diagonal line on this plot is the line where the *actual* value equals the *predicted* value.)

The **Summary of Fit** table provides key statistics, such as **RSquare** and **RSquare Adj** (adjusted R Square). RSquare indicates the percent of variation in the data that is explained by our model (0.81, or 81%). Because RSquare can be inflated simply by adding additional predictors to the model, the adjusted RSquare is sometimes used instead of RSquare for comparing models with more than one predictor (the “adjustment” applied to the **RSquare Adj** is based on the number of terms in the model).

The **Analysis of Variance** table indicates that the overall model is statistically significant. The *p*-value, reported as **Prob > F**, is < .0001.

The **Parameter Estimates** table provides coefficients for our model, along with *p*-values for each of the terms in the model.

Other output (effects tests, a residual plot, and leverage plots) is also provided by default, and additional options are available under the top red triangle.

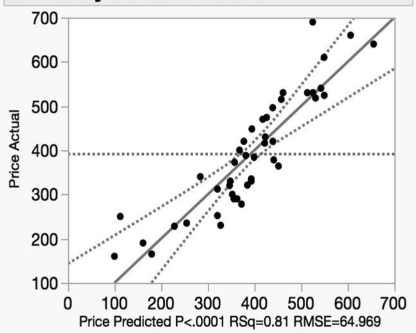
Figure 4.8: Fitted Model

Effect Summary

Source	LogWorth	PValue
Acres	1.880	0.01318
Baths	1.846	0.01424
Square Feet	0.982	0.10425
Miles to Base	0.319	0.48023
Miles to Resort	0.285	0.51915
Beds	0.265	0.54336
Cars	0.209	0.61769
Years Old	0.207	0.62093
DoM	0.092	0.80958

Remove Add Edit FDR

Actual by Predicted Plot



Summary of Fit

RSquare	0.809043
RSquare Adj	0.759939
Root Mean Square Error	64.96897
Mean of Response	391.1911
Observations (or Sum Wgts)	45

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	9	625914.08	69546.0	16.4763
Error	35	147733.84	4221.0	Prob > F
C. Total	44	773647.92		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	179.45965	51.25743	3.50	0.0013*
Beds	9.7668087	15.9136	0.61	0.5434
Baths	53.275621	20.65062	2.58	0.0142*
Square Feet	0.0445899	0.026733	1.67	0.1043
Miles to Resort	-1.698122	2.60755	-0.65	0.5191
Miles to Base	-1.878499	2.632587	-0.71	0.4802
Acres	4.6547176	1.782149	2.61	0.0132*
Cars	6.3837972	12.67613	0.50	0.6177
Years Old	-0.273392	0.547926	-0.50	0.6209
DoM	0.0324497	0.133651	0.24	0.8096

Effect Tests

Residual by Predicted Plot

In Figure 4.8, we see that only two of the predictors (**Baths** and **Acres**) are significant at the 0.05 level, given that all of the other variables are in the model. Does this mean that none of the other predictors are important in predicting housing prices? We'll want to reduce this model to include only those variables that, in combination, do a good job of predicting the response. This is known as the principle of *parsimony:* the simplest model that can predict the response well is often the best model. Given that we have nine predictor variables, there can be many possible models to predict **Price**, from very simple models to more complex. In general, our goal is to find a concise model that makes sense, fits the data, and predicts the response well (Hansen, 2001). But, since significance is dependent upon which other variables are in the model, it is difficult to determine which terms to keep in the model and which to remove.

For illustration, we click **Remove** at the bottom of the **Effect Summary** table to remove **Baths** from the model (top, in Figure 4.9). **Acres** is no longer significant at $\alpha = 0.05$, but **Square Feet** is now significant (bottom, in Figure 4.9).

Figure 4.9: Effect Summary Table with and without Baths

Effect Summary

Source	LogWorth	PValue
Acres	1.880	0.01318
Baths	1.846	0.01424
Square Feet	0.982	0.10425
Miles to Base	0.319	0.48023
Miles to Resort	0.285	0.51915
Beds	0.265	0.54336
Cars	0.209	0.61769
Years Old	0.207	0.62093
DoM	0.092	0.80958

Remove Add Edit FDR

Effect Summary

Source	LogWorth	PValue
Square Feet	2.067	0.00856
Acres	1.179	0.06629
Beds	1.062	0.08675
Miles to Base	0.667	0.21549
Cars	0.619	0.24055
DoM	0.398	0.39962
Years Old	0.156	0.69775
Miles to Resort	0.056	0.87910

Remove Add Edit Undo FDR

Part of the issue is that some of the variables are correlated with other variables in the model. Recall the correlation and scatterplot for **Miles to Resort** and **Miles to Base** (Figure 4.6). There is a very strong correlation between these two predictor variables. This means that they are somewhat redundant to one another. In fact, the resort and the base are in nearly the same geographic location.

A Bit About Multicollinearity

When two or more predictors are correlated with one another, the term *multicollinearity* is used. If multicollinearity is severe, then it is difficult to determine which of the correlated predictors are most important. In addition, the coefficients and standard errors for these coefficients may be inflated and the coefficients may have signs that don't make sense.

A measure of multicollinearity is the *VIF* statistic, or *Variance Inflation Factor*. The VIF for a predictor, VIF_j, is calculated using the following formula:

$$VIF_j = \frac{1}{1 - RSquare_{X_j}}$$

For each predictor, X_j, a regression model is fit using X_j as the response and all of the other X variables to predict X_j. The RSquare for that model fit $(RSquare_{X_j})$ is calculated, and is then used to calculate the VIF_j. An $RSquare_{X_j}$ of 0.9 results in a VIF_j of 10, while an $RSquare_{X_j}$ of 0.99 results in a VIF_j of 100.

If the VIF is 1.0, then each of the predictor variables is completely independent of the other predictor variables. But if the VIF is large (say, greater than 10), then the multicollinearity is a problem that should be addressed (Neter, 1996). In some cases, this can be resolved by removing a redundant term from the model. In more severe cases, simply removing a term will not address the issue. In these cases, variable reduction techniques such as Principal Components Analysis (PCA), Partial Least Squares (PLS), tree-based methods (covered in Chapter 6), and generalized regression methods (Chapter 9) are recommended.

In Figure 4.10 (on the left), we see VIFs for the original model. To display VIFs, right-click on the **Parameter Estimates** table and select **Columns > VIF**. The VIFs for most of the predictors are relatively small (< 5), while the VIFs for **Miles to Resort** and **Miles to Base** are both over 10, indicating that multicollinearity is a problem. Since we've learned that these two variables are largely redundant to one another, it makes sense to re-fit the model with only one of these variables. Subject matter knowledge can be used to determine which variable to remove. On the right in Figure 4.9, we see the results after removing **Miles to Resort** from the model. Notice that all of the VIFs are now low. In addition, **Miles to Base** is now significant, and the coefficient and the standard error for **Miles to Base** have both changed substantially!

Figure 4.10: Variance Inflation Factor, VIF

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	179.45965	51.25743	3.50	0.0013*	.
Beds	9.7668087	15.9136	0.61	0.5434	3.5544519
Baths	53.275621	20.65062	2.58	0.0142*	4.3822145
Square Feet	0.0445899	0.026733	1.67	0.1043	3.4757465
Miles to Resort	-1.698122	2.60755	-0.65	0.5191	13.822325
Miles to Base	-1.878499	2.632587	-0.71	0.4802	14.510754
Acres	4.6547176	1.782149	2.61	0.0132*	1.5212784
Cars	6.3837972	12.67613	0.50	0.6177	1.7883544
Years Old	-0.273392	0.547926	-0.50	0.6209	1.2901804
DoM	0.0324497	0.133651	0.24	0.8096	1.5627487

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	181.44396	50.75587	3.57	0.0010*	.
Beds	11.370069	15.59576	0.73	0.4707	3.4693829
Baths	50.724242	20.11275	2.52	0.0162*	4.224488
Square Feet	0.0430186	0.026411	1.63	0.1121	3.4474345
Miles to Base	-3.493138	0.877935	-3.98	0.0003*	1.6400357
Acres	4.3625859	1.710918	2.55	0.0152*	1.4248944
Cars	5.4694553	12.49696	0.44	0.6642	1.766415
Years Old	-0.192613	0.529416	-0.36	0.7181	1.2240612
DoM	0.0592831	0.12612	0.47	0.6412	1.414216

To further assist in refining our model, after removing **Miles to Resort**, we'll rely on an automated variable selection approach, *stepwise regression*. We proceed with *stepwise regression* to identify the best subset of significant factors. **Stepwise** is a method, or a **Personality**, available in the **Fit Model** dialog (Figure 4.11).

Note that stepwise regression does not address multicollinearity. If correlated terms are used as inputs, stepwise may not result in the "best" model because variable selection will be determined by which variables are selected first.

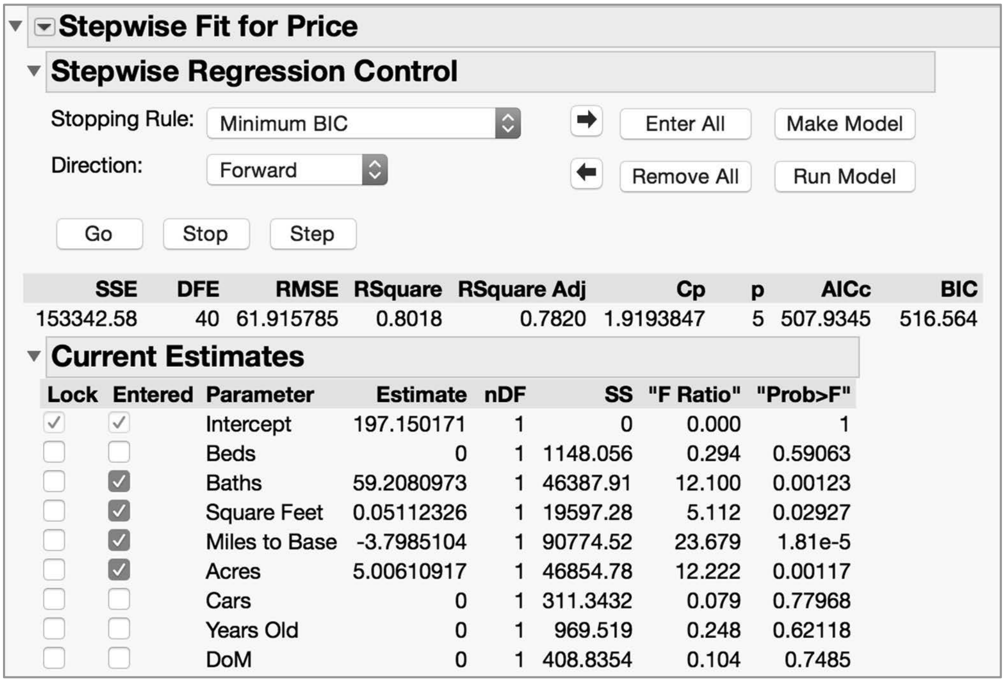
Figure 4.11: Fit Model Stepwise Dialog

Stepwise regression provides a number of stopping rules for selecting the best subset of variables for the model. The default rule is **Minimum BIC**, or minimum *Bayesian Information Criterion*. The **Direction**, which is set to **Forward** by default, indicates that variables will be added to the model one at a time. After you click **Go**, the model with the smallest BIC statistic is selected.

Another common rule, which works in a similar manner, is **Minimum AICc** (*Akaike's Information Criterion*, with a correction for small sample sizes). Both of these rules attempt to explain the relationship between the predictors and the response, without building models that are overly complex in terms of the number of predictors. Since different criteria are used to determine when to stop adding terms to the model, these stopping rules may lead to different "best" models (Burnham, 2002).

We will develop a model using each criterion and then compare results. First, we use the default **Minimum BIC** criterion, and click **Go** to start the selection process. We identify four factors for the model: **Baths**, **Square Feet**, **Acres**, and **Miles to Base** (see Figure 4.12).

Figure 4.12: Stepwise Regression Variable Selection Using BIC



Stepwise Fit for Price

Stepwise Regression Control

Stopping Rule: Minimum BIC

Direction: Forward

Enter All | Make Model | Remove All | Run Model | Go | Stop | Step

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
153342.58	40	61.915785	0.8018	0.7820	1.9193847	5	507.9345	516.564

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
✓	✓	Intercept	197.150171	1	0	0.000	1
		Beds	0	1	1148.056	0.294	0.59063
	✓	Baths	59.2080973	1	46387.91	12.100	0.00123
	✓	Square Feet	0.05112326	1	19597.28	5.112	0.02927
	✓	Miles to Base	-3.7985104	1	90774.52	23.679	1.81e-5
	✓	Acres	5.00610917	1	46854.78	12.222	0.00117
		Cars	0	1	311.3432	0.079	0.77968
		Years Old	0	1	969.519	0.248	0.62118
		DoM	0	1	408.8354	0.104	0.7485

Next, we change the stopping rule to **Minimum AICc**, click **Remove All** to clear the estimates from the BIC model, and then click **Go**. In this example, BIC and AICc yield the same set of factors (AICc output not shown). This isn't always the case.

We have identified a common set of factors for the model using two different stopping criteria. To run this regression model, select **Make Model**. Then in the **Fit Model** dialog, click **Run** (or, simply select **Run Model** from within the **Stepwise** platform).

Recall that our original model (Figure 4.8), with nine predictors, had only two significant terms and an adjusted R Square of 0.76. The results of fitting this reduced model are shown in Figure 4.13. As expected, this model is significant (Prob > F < .0001), and the four terms in the model are also significant. The adjusted R square is 0.782, which is slightly higher than our original model.

Figure 4.13: Revised Fitted Model

Response Price

Whole Model

Effect Summary

Actual by Predicted Plot

Summary of Fit

RSquare	0.801793
RSquare Adj	0.781972
Root Mean Square Error	61.91579
Mean of Response	391.1911
Observations (or Sum Wgts)	45

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	620305.34	155076	40.4523
Error	40	153342.58	3834	Prob > F
C. Total	44	773647.92		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	197.15017	34.72253	5.68	<.0001*
Baths	59.208097	17.0208	3.48	0.0012*
Square Feet	0.0511233	0.022611	2.26	0.0293*
Miles to Base	-3.79851	0.780608	-4.87	<.0001*
Acres	5.0061092	1.43194	3.50	0.0012*

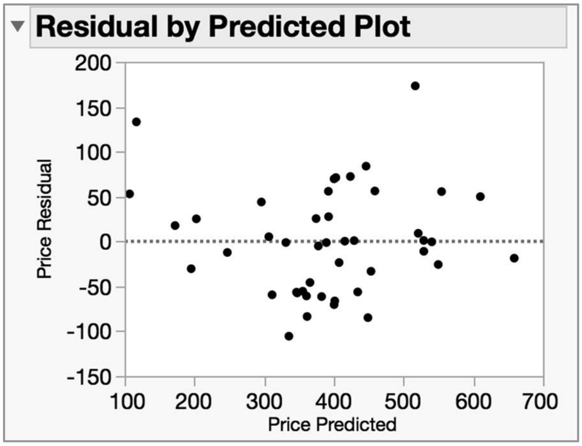
Before using this model, we need to check a few key assumptions. Namely, that our model errors are independent, have equal variance, and are normally distributed. Another key assumption is that the relationship between our response and the predictors is linear (i.e., that there isn't an underlying non-linear relationship).

The variation in the *residuals*, which is another word for the errors, shows us the variation in the response that could not be explained by the model that we have fit. Plots of residuals can be used to check that our assumptions about the model errors were correct. The default residual plot in JMP shows the residuals for each point plotted against predicted values. If our model assumptions are met, the points should be randomly scattered about the center line (zero), with no obvious pattern (just a cloud of seemingly random points). Other residual plots are also available (under the **red triangle > Row Diagnostics**), and the residuals can be saved (using **red triangle > Save Columns > Residuals**) and evaluated using **Distribution** or the **Graph Builder**.

The **Residual by Predicted** plot (see Figure 4.14) shows a somewhat curved pattern. That is, the largest residuals are at the lower and higher predicted values, while the smallest residuals are in the middle. This subtle pattern could be due to a term that is missing from the model. For example, the model may fit better if an interaction or quadratic (squared) term is added, or there may be an important variable that we've missed altogether. A *quadratic term* is used to explain curvature in the relationship between the factor and the response. An *interaction term* is used if the relationship between one factor and the response depends on the setting of another factor. We'll revisit interactions and quadratic terms in an exercise.

The pattern that we see in the residuals may also be due to outliers or influential observations, which can tilt or warp the regression model.

Figure 4.14: Examining Residuals

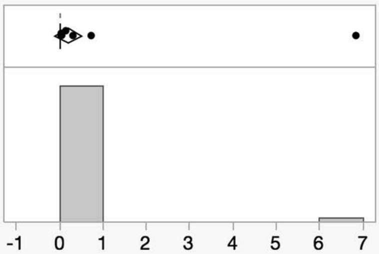


A statistic that helps us determine whether particular points are influencing the model is *Cook's D*, or *Cook's Distance*. Cook's D values for each observation can be saved to the data table, and then plotted using the **Analyze > Distribution** platform (see the plot of Cook's D values in Figure 4.15). To save Cook's D values from the Fit Model output window, click the red triangle and select **Save Columns > Cook's D Influence**.

A high Cook's D value for a particular observation indicates that the model predictions with and without that observation are different. What is considered high? A general rule of thumb is that any Cook's D value >1 is worthy of investigation (Cook, 1982). Observation #7, with a Cook's D value over 6, has a large influence on our model.

Figure 4.15: Cook's D Values

Cook's D Influence Price



Quantiles

100.0%	maximum	6.8564296677
99.5%		6.8564296677
97.5%		5.9387810987
90.0%		0.0958283162
75.0%	quartile	0.0156540769
50.0%	median	0.0064905051
25.0%	quartile	0.0018892103
10.0%		5.5676731e-6
2.5%		3.1578066e-7
0.5%		6.6482284e-8
0.0%	minimum	6.6482284e-8

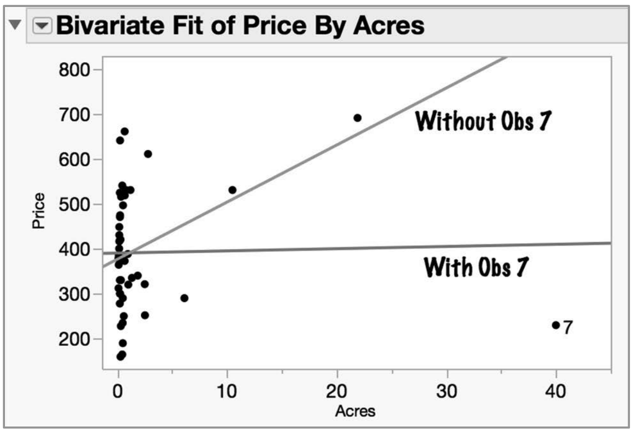
Summary Statistics

Mean	0.1880731
Std Dev	1.0235765
Std Err Mean	0.1525858
Upper 95% Mean	0.4955895
Lower 95% Mean	-0.119443
N	45

To illustrate how an influential point can impact model, see Figure 4.16. We use **Analyze > Fit Y by X**, with **Price** as **Y, Response** and **Acres** as **X, Factor**, and fit a line (select **Fit Line** from the red triangle). Then, we exclude observation 7, and again select **Fit Line**. The resulting regression lines are labeled (using the **Annotate** tool from the toolbar in JMP).

Note the difference in the slopes for fitted regression lines with and without observation 7 included in the model! Clearly, these two models will result in different predicted values, particularly for properties with higher acreage.

Figure 4.16: Illustration of Influential Point



Upon investigation, we find that this property is actually a 40-acre farm rather than a residential property. The focus of this pricing study is residential properties. Since this property is not of direct interest in this study and is influencing our predictions, we exclude and hide observation 7 (use **Rows > Hide and Exclude**) and proceed without it.

Since the data set is small and stepwise was performed using observation 7, the procedure may result in a different reduced model if run without this observation. We return to stepwise, and find that the same reduced model is produced without this observation (Figure 4.17).

Compared to the model in Figure 4.13, this model has a higher R Square Adjusted (0.84), and most of the *p*-values for the terms in the model are lower. In particular, the *p*-value for **Acres** has dropped from 0.0012 to <.0001 (and the coefficient is now much larger). In addition, the residuals now appear more randomly scattered about the center line, with no obvious patterns, evidence of curvature, or lack of constant variance.

Figure 4.17: Model and Residuals after Removing Influential Point

Summary of Fit

RSquare	0.854004
RSquare Adj	0.83903
Root Mean Square Error	52.88352
Mean of Response	394.8545
Observations (or Sum Wgts)	44

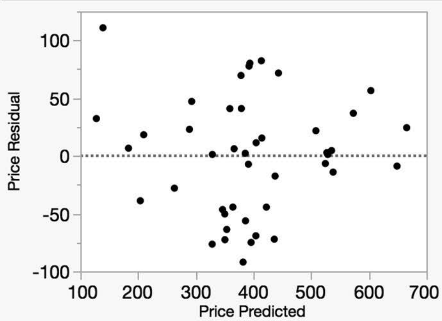
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	638004.83	159501	57.0326
Error	39	109070.00	2797	Prob > F
C. Total	43	747074.83		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	175.46632	30.1538	5.82	<.0001*
Baths	63.61719	14.57999	4.36	<.0001*
Square Feet	0.0472992	0.019337	2.45	0.0191*
Miles to Base	-3.025002	0.694498	-4.36	<.0001*
Acres	12.462947	2.237932	5.57	<.0001*

Residual by Predicted Plot



Given that the houses are scattered around the geographic area surrounding the resort and there is no obvious clustering of points in the residual plot, we have some additional comfort that the independence assumption is also met.

For additional confirmation that the normality assumption has been met, we can save the residuals to the data table (click the red triangle and select **Save Columns > Residuals**), and then use a histogram and normal quantile plot to check normality (use **Distribution**, and select **Normal Quantile Plot** from the red triangle). We leave it to the reader to confirm that the normality assumption has indeed been met.

Satisfied with our final model, we can now use this model to predict home prices. The terms in our model, and their coefficients, are given in the **Parameter Estimates** table (left, in Figure 4.18). Each estimate tells us how much the predicted selling price changes with a change in the value of the predictor.

The parameter estimates can be rewritten as a formula, or prediction expression (right, in Figure 4.18). To calculate the selling price of a home, we simply need to plug in the number of baths, the square feet, the miles to base, and the acres into the formula.

Figure 4.18: Parameter Estimates and Prediction Expression

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	175.46632	30.1538	5.82	<.0001*
Baths	63.61719	14.57999	4.36	<.0001*
Square Feet	0.0472992	0.019337	2.45	0.0191*
Miles to Base	-3.025002	0.694498	-4.36	<.0001*
Acres	12.462947	2.237932	5.57	<.0001*

Prediction Expression

175.466319174148
+ 63.6171900658549 * Baths
+ 0.04729922796845 * Square Feet
+ -3.025002137271 * Miles to Base
+ 12.4629468282909 * Acres

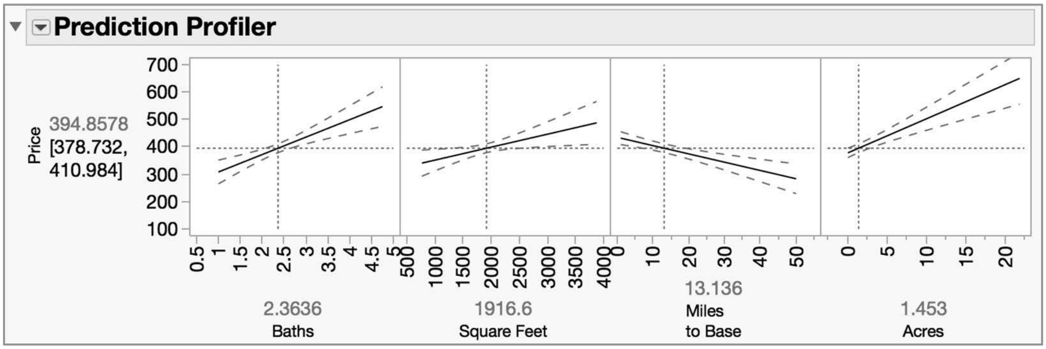
To display the prediction expression, click the red triangle and select **Estimates > Show Prediction Expression**. To save this formula to the data table, click the red triangle and select **Save Columns > Prediction Formula**.

The model can also be explored graphically using the **Prediction Profiler** (Figure 4.19). To access the profiler, select **Factor Profiling > Profiler** from the red triangle.

The profiler shows the predicted response (on the far left) at specified values of each of the predictor values (given at the bottom). The initial values for the predictors are predictor averages, and vertical red lines are drawn at these values. The starting value for the response is also the overall average (the mean **Price** in this example), and the bracketed values are the 95% confidence interval for the average. The confidence interval can be used to determine a *margin of error* for the prediction. The margin of error is the *half-width* of the confidence interval, or half the range of the interval. In this example, the margin of error is approximately (410.984 - 378.732)/2 = $16.125K.

Drag the vertical lines for a predictor to change the value for that predictor. The slopes of the lines for each predictor indicate whether predicted **Price** will increase or decrease if the predictor value increases, assuming that the other predictor values are held constant.

Figure 4.19: Using Prediction Profiler



Summary

In this example, we have created a regression model for home selling prices using historical data and have assessed model assumptions to ensure that the model makes sense. If we are satisfied with our model's suitability and performance, we can put the model to use to predict selling prices of new homes entering the market in this geographic region. For example, our model tells us that the predicted selling price of a 2500 square foot home with 3 baths that is 13 miles to the base and sits on one acre is just over $457,704, with a margin of error of approximately $21.5K. (You can confirm this by entering these values into the **Prediction Profiler**.)

Of course, we might ask if this is the best possible model to predict selling price. Can we build a better model? Our margin of error is relatively large, and our standard deviation (reported as Root Mean Square Error in Figure 4.17) is just under $53K. Can we develop a model that provides more precise predictions? What if we built a model using information on a larger sample of houses? What if we include additional information for each of the houses sold, such as recent renovations, measures of home quality, or the time of year that the home was first put on the market? Would this lead to a better model?

We should also keep in mind that all models need to be updated periodically. Housing prices change over time, so the model to predict housing prices should be updated to stay current and reflect these changes.

Example 2: Bank Revenues

A bank wants to understand how customer banking habits contribute to revenues and profitability. The bank has the customer age and bank account information, such as whether the customer has a savings account, if the customer has received bank loans, and other indicators of account activity.

The Data **BankRevenue.jmp**

The data set contains information on 7420 bank customers:

Rev_Total: Total revenue generated by the customer over a 6-month period.

Bal_Total: Total of all account balances, across all accounts held by the customer.

Offer: An indicator of whether the customer has received a special promotional offer in the previous one-month period. Offer=1 if the offer was received, Offer=0 if it was not.

AGE: The customer's age.

CHQ: Indicator of debit card account activity. CHQ=0 is low (or zero) account activity, CHQ=1 is greater account activity.

CARD: Indicator of credit card account activity. CARD=0 is low or zero account activity, CARD=1 is greater account activity.

SAV1: Indicator of primary savings account activity. SAV1=0 is low or zero account activity, SAV1=1 is greater activity.

LOAN: Indicator of personal loan account activity. LOAN=0 is low or zero account activity, LOAN=1 is greater activity.

MORT: Indicator of mortgage account tier. MORT=0 is lower tier and less important to the bank's portfolio. MORT=1 is higher tier and indicates the account is more important to the bank's portfolio.

INSUR: Indicator of insurance account activity. INSUR=0 is low or zero account activity, INSUR=1 is greater activity.

PENS: Indicator or retirement savings (pension) account tier. PENS=0 is lower balance and less important to bank's portfolio. PENS=1 is higher tier and of more importance to the bank's portfolio.

Check: Indicator of checking account activity. Check=0 is low or zero account activity, Check=1 is greater activity.

CD: Indicator of certificate of deposit account tier. CD=0 is lower tier and of less importance to the bank's portfolio. CD=1 is higher tier and of more importance to the bank's portfolio.

MM: Indicator of money market account activity. MM=0 is low or zero account activity, MM=1 is greater activity.

Savings: Indicator of savings accounts (other than primary) activity. Savings=0 is low or zero account activity, Savings=1 is greater activity.

AccountAge: Number of years as a customer of the bank.

Applying the Business Analytics Process

Define the Problem

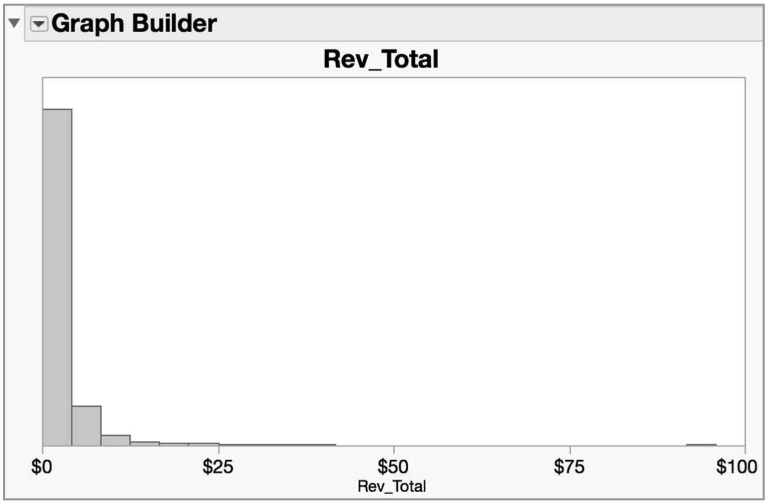
We want to build a model that allows the bank to predict profitability for a given customer. A surrogate for a customer's profitability that is available in our data set is the **Total Revenue** a customer generates through their accounts and transactions. The resulting model will be used to forecast bank revenues and guide the bank in future marketing campaigns.

Prepare for Modeling

We begin by looking at the variable of interest, total revenue (**Rev_Total**) using **Graph > Graph Builder**. **Rev_Total** is highly skewed, which is fairly typical of financial data (Figure 4.20).

> ***Note***: *To explore the underlying shape of the distribution, select the* ***Grabber*** *(hand) tool from your toolbar, click on the graph and drag up and down.*

Figure 4.20: Distribution of Total Revenue



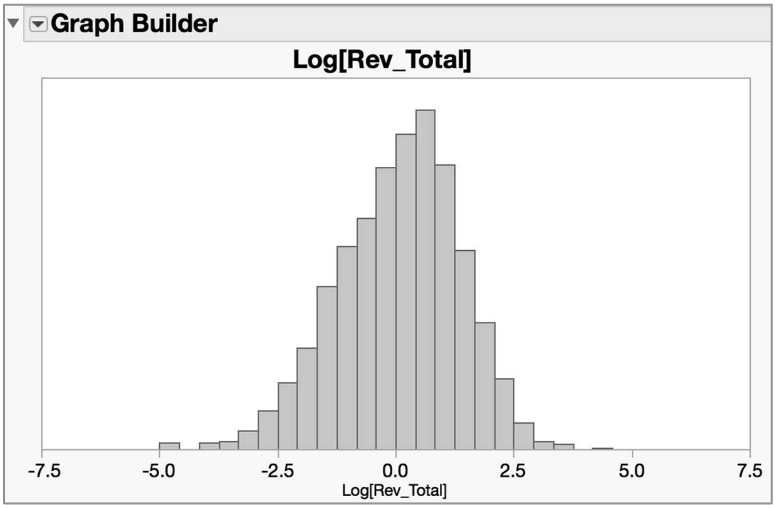
In regression situations, highly skewed data can result in a poorly fitting model. A transformation that can often be used to normalize highly skewed data, where all of the values are positive, is a log (natural logarithm) transformation (see Ramsey and Shafer, 2002, page 68).

We apply a log transformation to the **Rev_Total** variable directly in the **Graph Builder** and reexamine the distribution (Figure 4.21). To apply this transformation, right-click on the variable in the variable selection list, and select **Transform > Log**. Then, to save the transformation to the data table, right-click on **Log(Rev_Total)** and select **Add to Data Table**.

This transformation gives us a much less skewed and more symmetric distribution, so we use **Log(Rev_Total)** for the rest of our analysis.

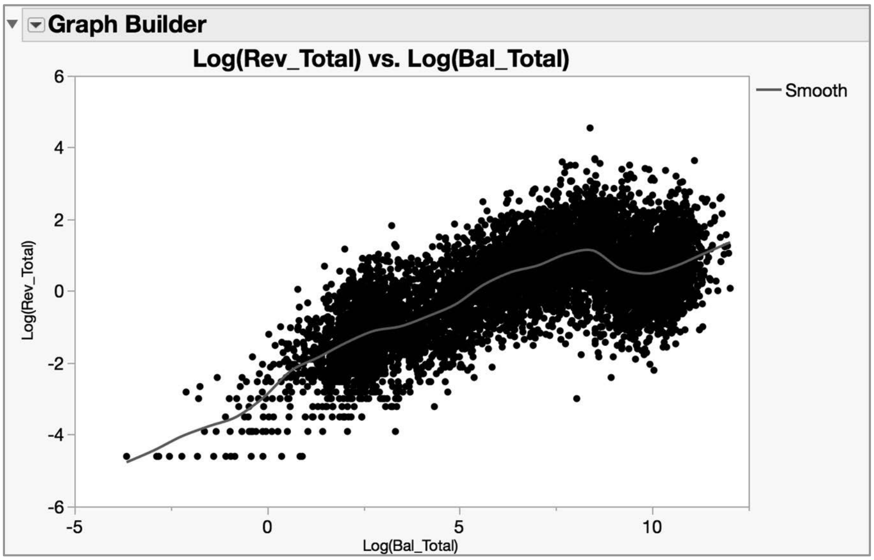
A similar examination of the total account balance (**Bal_Total**), which also has a skewed distribution, leads to using the **Log(Bal_Total)** in our analyses.

Figure 4.21: Transformed Total Revenue Using Log Transformation



The relationship between the log total revenue and log total account balance is shown in the scatterplot in Figure 4.22. The relationship appears nearly linear at lower account balances—higher account balances generally have higher revenues. But the relationship seems to change at the higher account balances.

Figure 4.22: Relationship between Log(Rev_Total) and Log(Bal_Total)

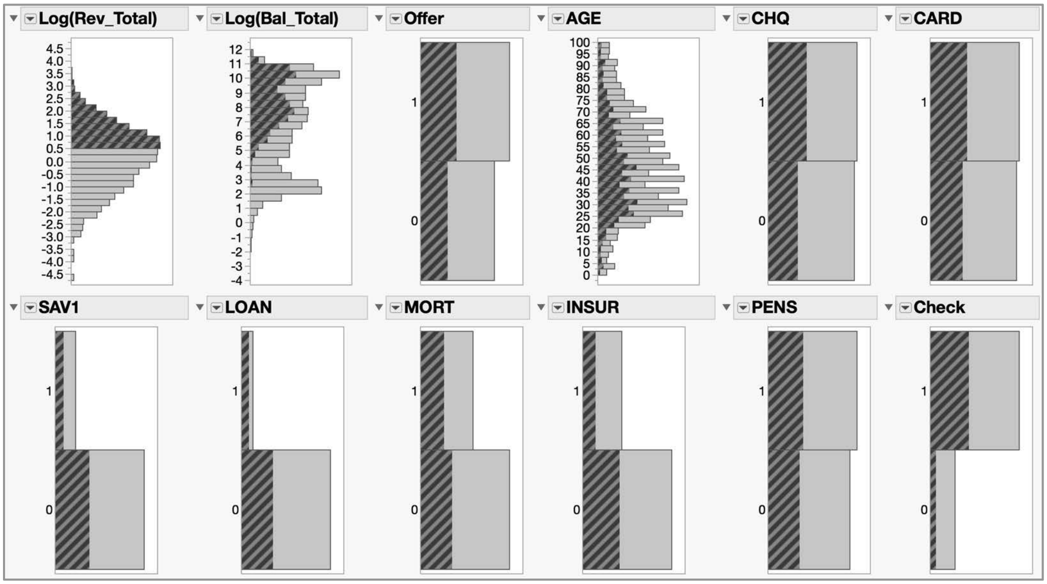


Now we examine the other variables. We can see their distributions and also their relationship to **Log(Rev_Total)**. Many of the variables are categorical, with two-levels. Higher revenue values are selected in Figure 4.23, and we can see this selection across the other variables in our data set. (The **Arrange In Rows** option under the red triangle was used to generate Figure 4.23; not all variables are displayed.) Other than total account balance, **Log(Bal_Total)**, there is no variable that stands out as being strongly related to revenue.

As we have discussed, other graphical and analytic tools can be used to understand the data and explore potential relationships, such as **Fit Y by X** and **Graph Builder**. In addition, the **Data Filter** (under the **Rows** menu) and **Column Switcher** (under the **red triangle > Scripts** in any output window) are dynamic tools that allow you to dive deeper into your data to explore variables of interest and potential relationships. Again, we encourage you to explore the data using these tools on your own. See Chapter 3 for discussion and illustration of different exploratory tools.

> ***Note****: Recall that within JMP there are a number of preferences that can be set (under* ***File > Preferences*** *or* ***JMP > Preferences*** *on a Mac), and all JMP output is customizable with your mouse and keystrokes. Going forward, we periodically resize graphs and change axis scaling to better fit content on the page, and change marker sizes or colors to improve interpretability. We also turn off shaded table headings in output to provide a cleaner display (within* ***Preferences, Styles > Report Tables****).*

Figure 4.23: Relationships between Transformed Variables and Other Variables



Build the Model

We now build a regression model to predict **Log(Rev_Total)** using **Fit Model** and the 15 potential predictor variables. There are some immediate signs of trouble (Figure 4.24). At the top of the **Fit Least Squares** window, we see some unexpected output, *Singularity Details*. This means that there are linear dependencies between the predictor variables. The first row of this table, **LOAN[0] = CD[0]** , indicates that JMP can't tell the difference between these two variables, **LOAN** and **CD**. The second line indicates that JMP can't tell the difference between **INSUR**, **MM**, and **Savings**.

The cause of this problem is illustrated in the **Distribution** output in Figure 4.25. The distributions of these three variables are identical. Every time **LOAN** = 1 (a customer has high loan activity), **MM** and **Savings** are also 1 (money market and savings activity are also high). The variables within each grouping are completely redundant to one another!

The result of this problem is seen in the parameter estimates table. JMP can't estimate all of these coefficients, indicating that the estimates for **LOAN** and **INSUR** are *biased*, and the estimates for **CD**, **MM**, and **Savings** are *zeroed*. JMP can estimate some of the parameters for the redundant variables (these estimates are biased), but not all (these are zeroed). Whether the variables appear as biased or zeroed depends entirely on the order in which they were entered into the model—those entered first into the model are displayed as biased.

Figure 4.24: Fit Least Squares with Singularity

Response Log(Rev_Total)

Singularity Details

LOAN[0] = CD[0]
INSUR[0] = MM[0] = Savings[0]

Effect Summary

Summary of Fit

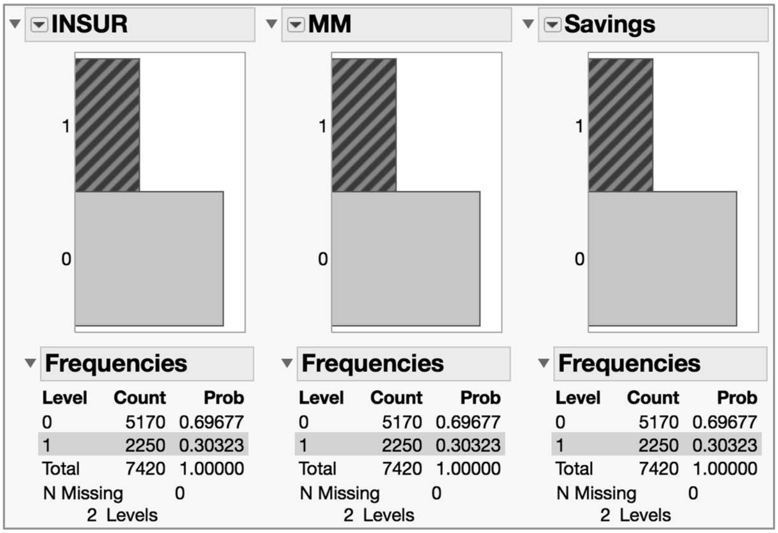
Analysis of Variance

Lack Of Fit

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept		0.3113431	0.049687	6.27	<.0001*
Bal_Total		2.3813e-5	1.27e-6	18.74	<.0001*
Offer[0]		-0.106738	0.027105	-3.94	<.0001*
AGE		-0.002446	0.000645	-3.79	0.0002*
CHQ[0]		-0.017754	0.014937	-1.19	0.2346
CARD[0]		-0.002709	0.040624	-0.07	0.9468
SAV1[0]		-0.010189	0.017976	-0.57	0.5709
LOAN[0]	Biased	-0.599056	0.023225	-25.79	<.0001*
MORT[0]		0.0162212	0.023909	0.68	0.4975
INSUR[0]	Biased	-0.091974	0.019419	-4.74	<.0001*
PENS[0]		-0.00232	0.013493	-0.17	0.8635
Check[0]		-0.310457	0.039308	-7.90	<.0001*
CD[0]	Zeroed	0	0	.	.
MM[0]	Zeroed	0	0	.	.
Savings[0]	Zeroed	0	0	.	.
AccountAge		-0.005122	0.003674	-1.39	0.1633

Figure 4.25: Distributions of INSUR, MM, and Savings



We refit the model without the redundant variables. In JMP 12, this can be done using the **Remove** button at the bottom of the **Effect Summary** table. We keep **LOAN** (and eliminate **CD**), and keep **INSUR** (eliminating **MM** and **Savings**). Note that this was an arbitrary decision: subject matter knowledge should guide the decision as to which redundant variables to remove (and which variables to keep in the model). As we remove each variable (or term), the **Singularity Details** table updates, along with all of the other statistical output. JMP is now able to estimate coefficients for each of the parameters (Figure 4.26).

Figure 4.26: Fit Least Squares Parameter Estimates without Singularity, Showing VIFs

Parameter Estimates

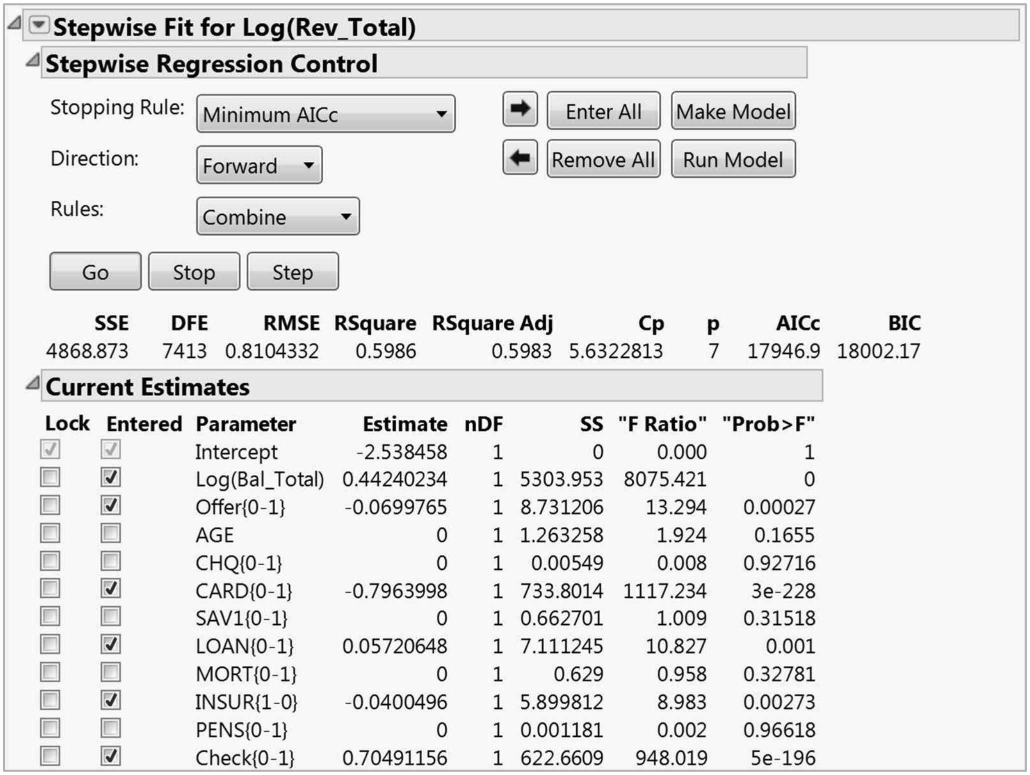
Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-2.531352	0.044361	-57.06	<.0001*	.
Log(Bal_Total)	0.4421894	0.004931	89.68	<.0001*	2.539337
Offer[0]	-0.069263	0.019212	-3.61	0.0003*	4.1313113
AGE	-0.00057	0.000455	-1.25	0.2103	1.0364354
CHQ[0]	-0.004403	0.010521	-0.42	0.6756	1.2500319
CARD[0]	-0.783241	0.027512	-28.47	<.0001*	8.5478082
SAV1[0]	0.0109566	0.012739	0.86	0.3898	1.1167727
LOAN[0]	0.0587778	0.018132	3.24	0.0012*	1.4725238
MORT[0]	0.0160545	0.016929	0.95	0.3430	3.0203293
INSUR[0]	0.0371717	0.013774	2.70	0.0070*	1.8111374
PENS[0]	-0.000349	0.009562	-0.04	0.9709	1.0309563
Check[0]	0.6864921	0.028624	23.98	<.0001*	6.3844217

A quick check of the VIFs indicates that multicollinearity is not a serious issue (Figure 4.26).

Since we have 11 remaining potential predictor variables, we use again stepwise regression to help with variable selection. We use the **Stepwise Personality** in the **Fit Model** platform, with **Log(Rev_Total)** as our Y and the other variables as model effects. For this example, we use the **Minimum AICc** stopping rule.

Stepwise selects six variables for the model. These are checked under **Current Estimates** in Figure 4.27. Note that when using AICc (or BIC), the resulting models may include terms that are not significant. This is because both AICc and BIC build models based on *important effects* (effects that explain the relationship between the response and the predictors) rather than searching for *significant effects* (see Burhnam, 2002). However, in this example, all six selected variables have low p-values.

Figure 4.27: Stepwise Regression Dialog with Model Variables Selected



Stepwise Fit for Log(Rev_Total)

Stepwise Regression Control

Stopping Rule: Minimum AICc

Direction: Forward

Rules: Combine

Enter All | Make Model | Remove All | Run Model

Go | Stop | Step

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
4868.873	7413	0.8104332	0.5986	0.5983	5.6322813	7	17946.9	18002.17

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	-2.538458	1	0	0.000	1
☐	☑	Log(Bal_Total)	0.44240234	1	5303.953	8075.421	0
☐	☑	Offer{0-1}	-0.0699765	1	8.731206	13.294	0.00027
☐	☐	AGE	0	1	1.263258	1.924	0.1655
☐	☐	CHQ{0-1}	0	1	0.00549	0.008	0.92716
☐	☑	CARD{0-1}	-0.7963998	1	733.8014	1117.234	3e-228
☐	☐	SAV1{0-1}	0	1	0.662701	1.009	0.31518
☐	☑	LOAN{0-1}	0.05720648	1	7.111245	10.827	0.001
☐	☐	MORT{0-1}	0	1	0.629	0.958	0.32781
☐	☑	INSUR{1-0}	-0.0400496	1	5.899812	8.983	0.00273
☐	☐	PENS{0-1}	0	1	0.001181	0.002	0.96618
☐	☑	Check{0-1}	0.70491156	1	622.6609	948.019	5e-196

We now run this model, and explore the results (Figure 4.28). As expected, the overall model is significant with a p-value < .0001, as are all of the terms in the model. The R Square is 0.5986, indicating that our model explains nearly 60% of the variation in the response.

Figure 4.28: Model Results, Reduced Model

Summary of Fit

RSquare	0.598624
RSquare Adj	0.598299
Root Mean Square Error	0.810433
Mean of Response	0.059558
Observations (or Sum Wgts)	7420

Analysis of Variance

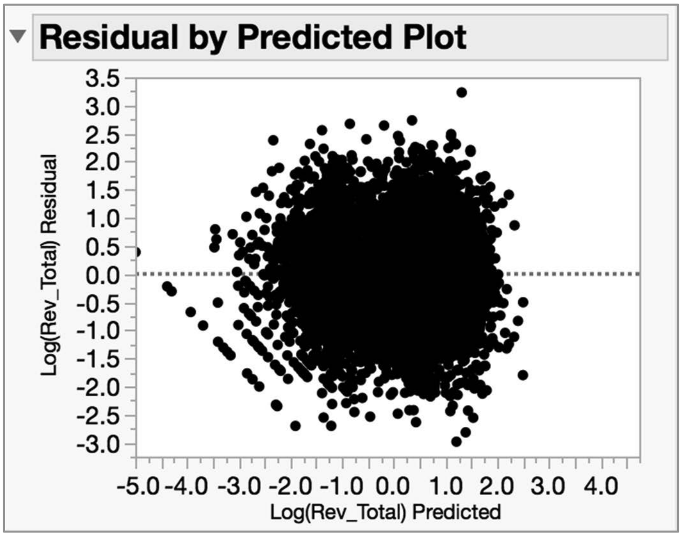
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	7261.582	1210.26	1842.661
Error	7413	4868.873	0.66	**Prob > F**
C. Total	7419	12130.455		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-2.538458	0.034835	-72.87	<.0001*
Log(Bal_Total)	0.4424023	0.004923	89.86	<.0001*
Offer[0]	-0.069977	0.019193	-3.65	0.0003*
CARD[0]	-0.7964	0.023826	-33.43	<.0001*
LOAN[0]	0.0572065	0.017386	3.29	0.0010*
INSUR[0]	0.0400496	0.013363	3.00	0.0027*
Check[0]	0.7049116	0.022894	30.79	<.0001*

Before interpreting the results of the regression model, we check that the regression assumptions are met. Since the data are from over 7400 different customers, we have some assurance that the independence assumption is met. The default residual versus predicted value plot (Figure 4.29) shows some diagonal striations in the lower left corner.

Figure 4.29: Residual versus Predicteds

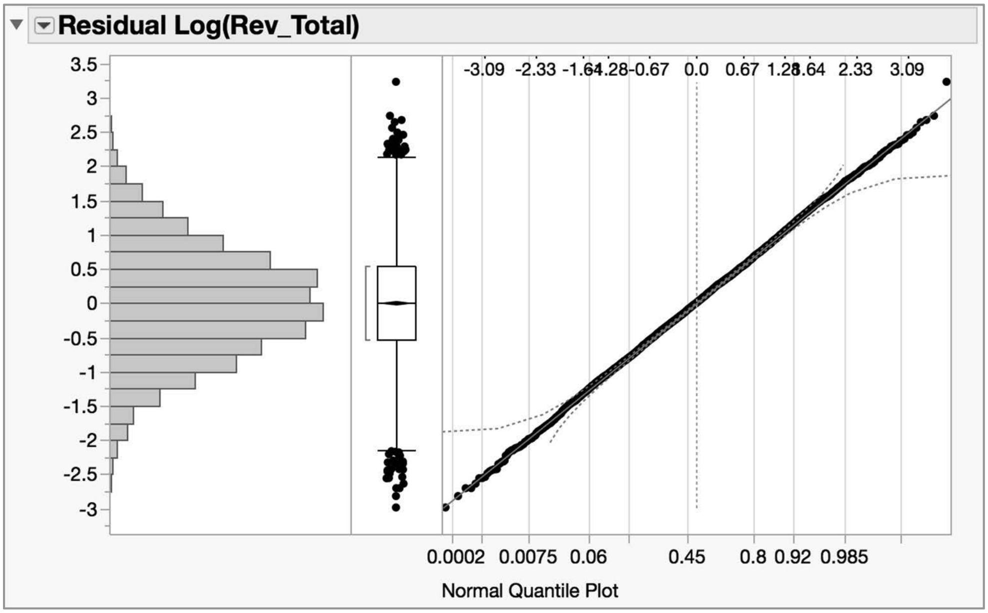


To explore these values, we use the **lasso tool** on the toolbar to select the observations, activate the data table, and then use the **F7** function key to scroll through these selected observations in the data table. The first strip on the left corresponds to revenue $0.01, and the second is revenue $0.02. This is the result of the fact that there are many customers who generate little, if any, revenue for the bank.

Otherwise, points appear randomly scattered around the center line (zero), and the residual plot shows no obvious evidence of unusual patterns.

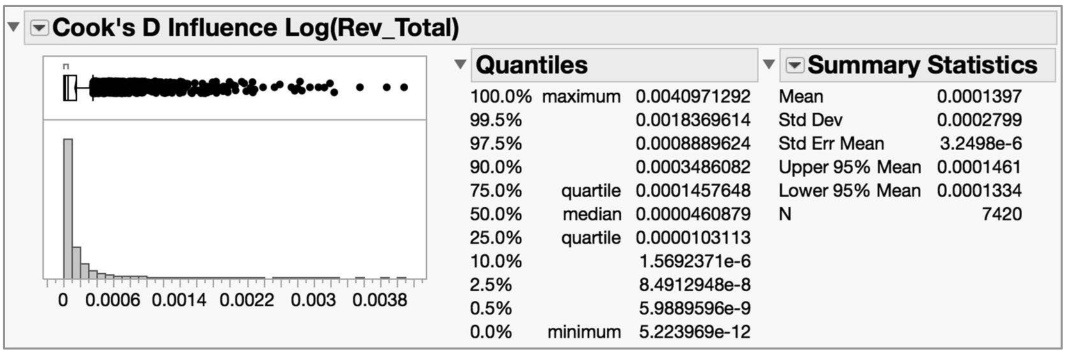
For further exploration of the regression assumptions, we save the residuals to the data table (under the red triangle, select **Save Columns > Residuals**), and use the **Distribution** platform to generate a histogram with a normal quantile plot (Figure 4.30). These plots provide evidence that the normality assumption has been met.

Figure 4.30: Distribution of Residuals with Normal Quantile Plot



We also see (in Figure 4.30) that there are no serious outliers. A quick peek at Cook's D values (Figure 4.31) confirms that there are no highly influential observations. No single point is exerting too much influence over our model.

Figure 4.31: Checking Assumptions with Cook's D



After investigating residuals and looking at Cook's D values, we have confidence that the regression assumptions have been satisfied. Our final model, shown in Figure 4.32, includes the following variables:

- The total account balance (**Log(Bal_Total)**)
- Whether the customer received a promotional offer (**Offer**)
- Credit card activity (**CARD**)
- Personal loan account activity (**LOAN**),
- Insurance account activity (**INSUR**)
- Checking account activity (**Check**)

All of the significant variables except **Log(Bal_Total)** are binary categorical variables. For the continuous predictor, **Log(Bal_Total)**, the coefficient in the parameter estimates table (top in Figure 4.32) indicates how the revenues change as the account balance changes. The positive coefficient indicates that revenues increase on average as account balances increase. The coefficient value itself is a little difficult to interpret because it reflects the transformation of **Rev_Total** to **Log(Rev_Total)**.

For each of the two-level categorical predictors, the parameter estimates show how the average response changes at the low level of each predictor. For example, the coefficient for **CARD[0]** is negative 0.7964. This indicates that log revenues are 0.7964 lower on average if credit card activity is low, and 0.7964 higher on average if credit card activity is high. The coefficients for **LOAN**, **Check**, and **INSUR** are all positive, indicating that low activity in these three accounts leads to higher revenues.

Note: When fitting regression models in JMP, two-level categorical predictors are automatically transformed into coded indicator variables using a -1/+1 coding scheme. The parameter estimate is reported for the lowest level or value of the predictor. In this example, **CARD** is a nominal predictor with levels with 0 and 1. The term in the reduced model is represented as **CARD[0]**, and the parameter estimate is -0.7964 (see Figure 4.32). The estimate for **CARD[1]**, which is not reported, is +0.7964. To display both estimates, select **Expanded Estimates** from the **top red triangle > Estimates**.

Many statistical software packages require dummy coding of categorical predictors, using a 0/1 "dummy" or "indicator" coding scheme. This is done prior to fitting the model, and results in different parameter estimates and a different interpretation of the estimates. For example, the parameter estimate for **CARD**, using 0/1 dummy coding, is 1.5928 instead of -0.7964. The sign is different, and the estimate is exactly twice the magnitude. To confirm this, change the modeling type for **CARD** to **Continuous** (to tell JMP to use dummy coding) and refit the reduced model shown in Figure 4.28. Note that, although the parameter estimates are different, the two coding schemes produce identical model predictions.

To view the indicator-coded version of the parameter estimates in the **Fit Least Squares** output, select **Indicator Parameterization Estimates** from the **top red triangle > Estimates**. Further details of how JMP transforms categorical factors can be found in the Statistical Details section of the book Fitting Linear Models (under **Help > Books**).

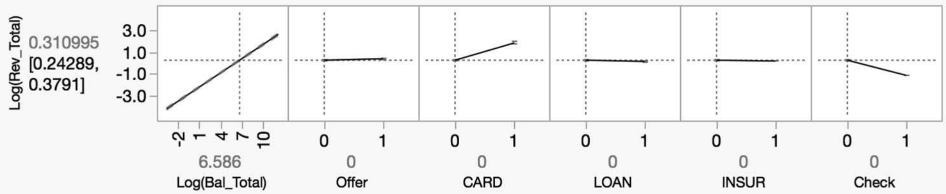
The prediction profiler (bottom of Figure 4.32) can help us see the impact of changes in values of the predictor variables on **Log(Rev_Total)**.

Figure 4.32: Exploring the Reduced Model with the Prediction Profiler

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-2.538458	0.034835	-72.87	<.0001*
Log(Bal_Total)	0.4424023	0.004923	89.86	<.0001*
Offer[0]	-0.069977	0.019193	-3.65	0.0003*
CARD[0]	-0.7964	0.023826	-33.43	<.0001*
LOAN[0]	0.0572065	0.017386	3.29	0.0010*
INSUR[0]	0.0400496	0.013363	3.00	0.0027*
Check[0]	0.7049116	0.022894	30.79	<.0001*

Prediction Profiler



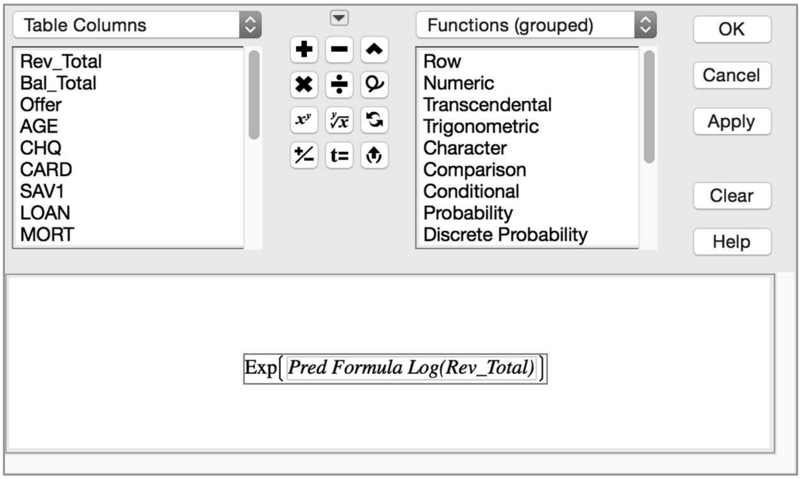
Clearly, **Log(Bal_Total)** has a large positive effect on the response. Three predictors, **Offer**, **LOAN**, and **INSUR**, while significant, have a relatively small effect on the response.

To show the predicted values for each bank customer, the prediction equation (the formula) can be saved to the data table (red triangle, **Save Columns > Prediction Formula**). Unfortunately, these are the log predicted values, which are difficult to interpret.

The inverse transformation (in this case the *exponential*, or Exp function) can be used to see the predicted values on the original scale. To apply this transformation, create a new column in the data table (we've named this column **Pred Rev_Total).** Then, right-click on the column and select **Formula** to open the **Formula Editor**, and use the **Transcendental > Exp** function from the **Functions (grouped)** list (see Figure 4.33).

Note that this formula can be created using a shortcut. Simply right-click on the saved prediction formula column, and select **New Formula Column > Transform > Exp**. JMP will create the new column with the stored formula shown in Figure 4.33.

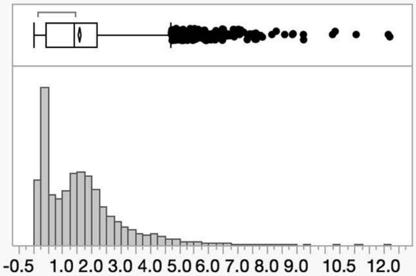
Figure 4.33: Transforming Predicted Log(Rev_Total) to Predicted Rev_Total



Now, we can explore the distribution of these values using **Distribution** or **Graph Builder** (Figure 4.34).

Figure 4.34: Distribution of Predicted Rev_Total

Pred Rev_Total



Quantiles

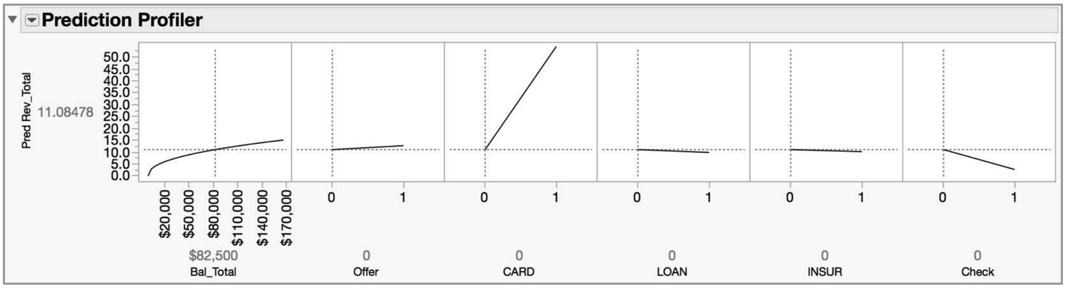
100.0%	maximum	12.218379622
99.5%		6.5200938162
97.5%		4.9030345728
90.0%		3.2604062067
75.0%	quartile	2.1543797924
50.0%	median	1.4049795287
25.0%	quartile	0.4411949596
10.0%		0.2678533829
2.5%		0.1509048307
0.5%		0.0667292746
0.0%	minimum	0.0067823167

Summary Statistics

Mean	1.580706
Std Dev	1.3093443
Std Err Mean	0.0152003
Upper 95% Mean	1.6105029
Lower 95% Mean	1.5509091
N	7420

We can also explore the formula itself using **Graph > Profiler** (Figure 4.35). Select the transformed prediction formula as the **Y, Prediction Formula**, and check the **Expand Intermediate Formulas** box to drill down to the original saved prediction formula. Now, we can readily see and explore the impact of changes to each of the variables on the predicted revenues in the original scale.

Figure 4.35: Prediction Profiler for Predicted Rev_Total



Summary

It is clear that high account balance customers, and those who use their credit card frequently, generate more revenue. What is curious is that high checking account usage seems to indicate lower revenue, and that customers with higher activity on the loan and insurance accounts have lower predicted revenue on average.

Was the promotional offer was successful? That is, did it lead to increased revenue? For a customer maintaining an account balance of $82,500, with low credit card, loan, insurance, and checking account activity, the promotional offer increased revenues from $11.08 to $12.75 on average. If this same customer had high credit card activity instead of low, the predicted revenue increased from $54.5 to $62.7. However, this analysis does not determine return on investment. Further information would need to be gathered to determine the cost of the promotional offer program and to examine the increased revenue relative to that cost. All of these insights lead to more questions, with new business problems to solve.

Exercises

Exercise 4.1: In this exercise, we use the **HousingPrices.jmp** data. In this chapter, we built a predictive model for **Price**, but limited model terms to main effects (that is, the predictors themselves). This is due, in part, to the fact that our data set is very small. However, other possible model effects include interactions and squared terms (quadratic effects).

Consider a model with all of the original predictors, plus one two-factor interaction and one quadratic term. Pick one interaction and one quadratic term that you think might be significant in predicting house prices.

Build a model using all of the original model effects and these two new terms.

1. Add all of the terms to the model.
2. Add the interaction term. Select the two terms from the **Select Columns** list and click **Cross**.
3. Add the squared term. Select the term in both the **Select Columns** list and the **Model Effects** list and click **Cross**.

Using this model, repeat the analysis illustrated in Example 1.

Questions:

a. Why did you pick the particular interaction and quadratic effect?
b. Are either of these two new terms significant?
c. Do they improve our model predictions?
d. Can you think of other predictors or terms, either in the data set or not contained in the data, that might improve the ability of our model to predict house prices?

Exercise 4.2: In this exercise we use the **BankRevenue.jmp** data.

Fit a full model to **Log(Rev_Total)** using **Log(Bal_Total)** and the other variables as model effects (using main effects only). Note, you may need to re-create these columns. Use the Minimum BIC stopping rule and stepwise regression to build your model.

a. Compare your reduced model to that obtained using Minimum AICc in this chapter. Describe the differences in terms of the variables in the model and key statistics (adjusted R Square, RMSE, and other statistics provided).
b. Which is the "better" model? Why? Does one model do a better job of predicting the response than the other? Explain

Exercise 4.3: Continue with the **BankRevenue.JMP** data.

Instead of fitting a model using the transformed variables, fit a model using the original (untransformed) variables. Use **Rev_Total** as the response, and **Bal_Total** and the other variables as model effects. Use stepwise and your preferred stopping rule to build the model.

a. What are the model assumptions?
b. Use the tools covered in this chapter to check model assumptions. Which tools should you use to check these assumptions? Explain how each tool helps check assumptions.
c. Explain why the model assumptions are or are not met.
d. Does it make sense to use this model to make predictions? Why or why not?

Exercise 4.4: Use the **BostonHousing.jmp** data set from the **Sample Data Directory** for this exercise. The response of interest, **mvalue**, is the median value of homes for towns in the Boston area in the 1970s.

a. Use the tools introduced in Chapter 3 to explore the data and prepare for modeling. Are there any potential data quality issues (other than the fact that the data are from the 1970s)? Determine what actions, if any, should be taken to address data quality issues that you identify.
b. Fit a model to **mvalue** using only **chas** and **rooms**. Recall that **rooms** is the number of rooms (rooms) and **chas** is a dummy variable (**chas=**1 indicates the town tracks the Charles River).
    i. Write down the equation for this model.
    ii. Interpret the coefficients for **chas[0]** and rooms.

iii. What is the predicted **mvalue** for a home that tracks the Charles River and has 6 rooms?

c. Fit a model to **mvalue** using all of the other variables as model effects. Use the Minimum BIC stopping rule and stepwise regression to build your model. How many terms are in the final model? Which terms are not included in the model?

d. Check model assumptions. Are model assumptions met? Explain.

e. How would a realtor, selling homes in the Boston area (in the same time period), use this model? How would a potential home buyer use this model?

References

Burnham, K. P., and D. R. Anderson. 2002. *Model Selection and Multimodel Inference*, Second Edition. Springer: New York.

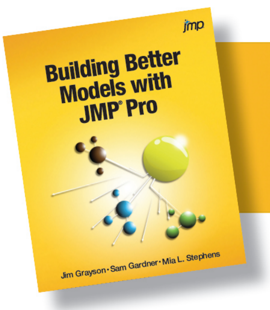
Cook, R. D., and S. Weisberg. 1983. *Residuals and Influence in Regression*, New York, NY: Chapman & Hall.

Dormann, C. F., J. Elith, S. Bacher, et al. 2012. *Collinearity: A review of methods to deal with it and a simulation study evaluating their performance*. Available at http://bit.ly/1p9ml70.

Hansen, M. H., and B. Yu. 2001. "Model Selection and the Principle of Minimum Description Length." *Journal of the American Statistical Association*, Vol. 96, No. 454, Review Paper. Available at http://bit.ly/1n9f3AU; accessed 8/28/2014.

Neter, J., M. Kutner, W. Wasserman, and C. Nachtsheim. 1996. *Applied Linear Statistical Models*, 4th ed. Irwin.

Ramsey, F., and D. Shafer. *The Statistical Sleuth*, 2nd ed. Cengage Learning.



Index

D

E

F

G

H

I

J

W

Z

About These Authors

Jim Grayson, PhD, is a Professor of Management Science and Operations Management in the Hull College of Business Administration at Georgia Regents University. He currently teaches undergraduate and MBA courses in operations management and business analytics. Previously, Jim held managerial positions at Texas Instruments in quality and reliability assurance, supplier and subcontractor management, and software quality. He has a PhD in management science with an information systems minor from the University of North Texas, an MBA in marketing from the University of North Texas, and a BS from the United States Military Academy at West Point.

Sam Gardner is a Senior Research Scientist at Elanco, a business division of Eli Lilly and Company. He currently works as a statistician supporting manufacturing development for a variety of animal health products. He is recognized by the American Statistical Association as an Accredited Professional Statistician. He has an MS in mathematics from Creighton University and an MS in statistics from the University of Kentucky. He graduated from Purdue University with BS degrees in mathematics and chemistry.

Gardner started his professional career as a military officer in the US Air Force, where he had roles that focused on modeling and simulation, operational flight test planning and analysis, and research and development. He also taught statistics at the Air Force Institute of Technology. After leaving the military, he began his work in the pharmaceutical industry as a statistician supporting the development and manufacturing of active pharmaceutical ingredients, and later transitioned to a role as a chemist in pharmaceutical manufacturing. He also worked as a statistician in pharmaceutical marketing with a focus on using statistical modeling to help solve business problems related to sales and marketing effectiveness. An avid user of statistical software, he spent several years working for JMP®, a division of SAS®, where he worked as a product expert and seminar speaker, with a focus on data visualization, applied statistics, and modern statistical modeling.

Mia L. Stephens is an Academic Ambassador with JMP®, a division of SAS®. Prior to joining SAS®, she split her time between industrial consulting and teaching statistics at the University of New Hampshire. A coauthor of *Visual Six Sigma: Making Data Analysis Lean* and *JMP® Start Statistics: A Guide to Statistics and Data Analysis Using JMP®, Fifth Edition*, she has developed courses and training materials, taught, and consulted within a variety of manufacturing and service industries. Stephens holds an MS in statistics from the University of New Hampshire and is located in York Harbor, Maine.

Learn more about these authors by visiting their author pages, where you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more:
http://support.sas.com/grayson
http://support.sas.com/gardner
http://support.sas.com/stephens