Case 11 - Cell Phone Service:
Multiple Regression – Two Predictors

Marlene Smith, University of Colorado Denver
Business School

Cell Phone Service[1]:
Multiple Regression - Two Predictors

Background

What impacts cellular phone call performance? Physical obstructions, such as buildings and mountains, can interfere with radio signal signals. More recent scientific studies suggest that radio waves can also be influenced by the effect that wind has on vegetation. Wind-induced movement of large trees, for example, can create a moving barrier between a radio transmitter and the signal destination. The result is called wind-induced fading, which predicts that higher wind speeds will result in degraded cell phone performance.

Knowing the impact of wind-induced fading has important business implications for Reliable Wireless, a mobile phone company in a Midwestern city in the U.S. If wind-induced fading is impacting the quality of cell phone service to customers, this must be considered when deciding how best to improve service. Although Reliable Wireless has no control over climatic conditions, decisions *can* be made about things such as data call capacity, number and configuration of base stations, band spectrums, and choice of wireless technologies. Knowledge of the impact of wind-induced fading on call performance should factor into the overall decisions about capital expenditures on equipment and technology.

The Task

What do the data have to say about the impact of wind-induced fading and barometric pressure on cell phone call performance at Reliable Wireless?

The Data Cell Phone Service.jmp

The data table contains hourly data on weather conditions and call performance for the Reliable Wireless network from about August to mid-October of a recent year. The call performance data comes from Reliable Wireless databases and the climate information from the local airport. Among the variables in the data set are **Hour** and:

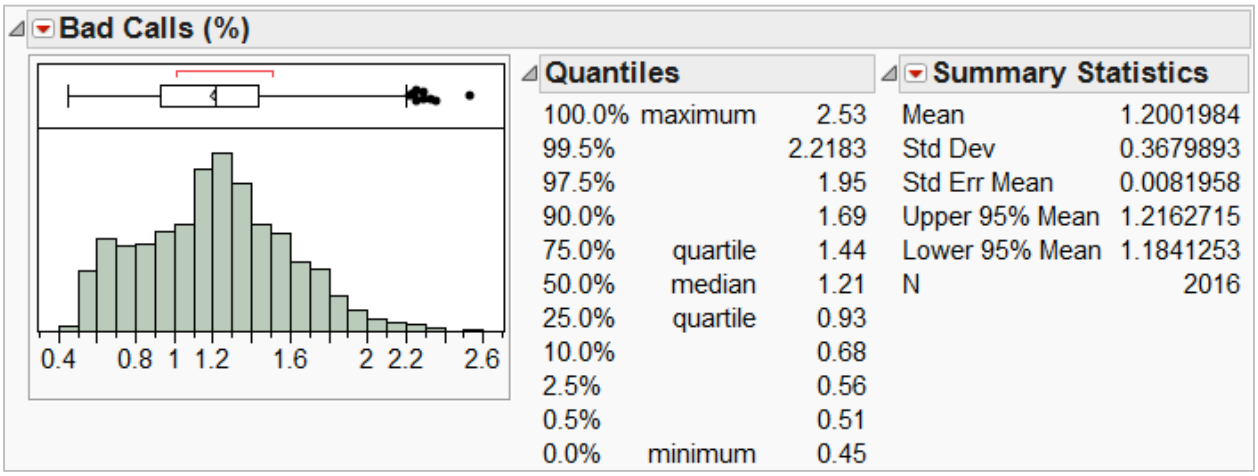
Pressure	Barometric pressure in inches of mercury at the end of the hour
Wind Speed	Wind speed measured in miles per hour at the end of the hour
Bad Calls (%)	Percentage of bad calls relative to the total number of calls made that hour. Total calls is a measure of all calls terminated (normally or otherwise) that hour. Bad calls is the sum of the number of failed (unsuccessful connection) or dropped (interrupted connection or delayed transmission) calls that hour. The word “calls” is used generically to include voice, text and email transmissions.

[1]These data are real, although the company name and its location have been disguised.

Analysis

The average percentage of bad calls over this time period is 1.2% (Exhibit 1) with standard deviation of 0.37%. The largest percentage of bad calls in any one hour was 2.53%.

Exhibit 1 Distribution of **Bad Calls (%)**



Quantiles

100.0%	maximum	2.53
99.5%		2.2183
97.5%		1.95
90.0%		1.69
75.0%	quartile	1.44
50.0%	median	1.21
25.0%	quartile	0.93
10.0%		0.68
2.5%		0.56
0.5%		0.51
0.0%	minimum	0.45

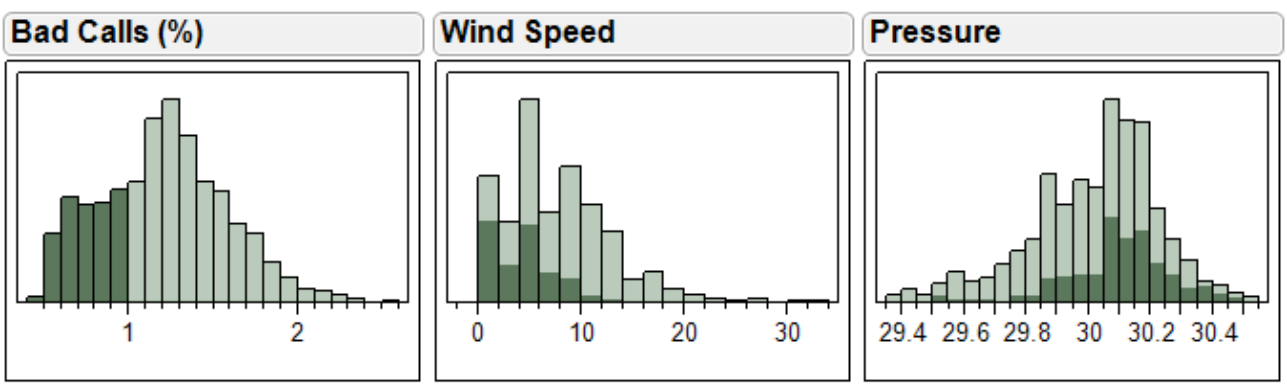
Summary Statistics

Mean	1.2001984
Std Dev	0.3679893
Std Err Mean	0.0081958
Upper 95% Mean	1.2162715
Lower 95% Mean	1.1841253
N	2016

(Analyze > Distribution, select Stack under the top red triangle for a horizontal layout)

Dynamic plot-linking indicates that lower values of **Bad Calls (%)** are generally associated with lower values of **Wind Speed** and higher values of **Pressure**.

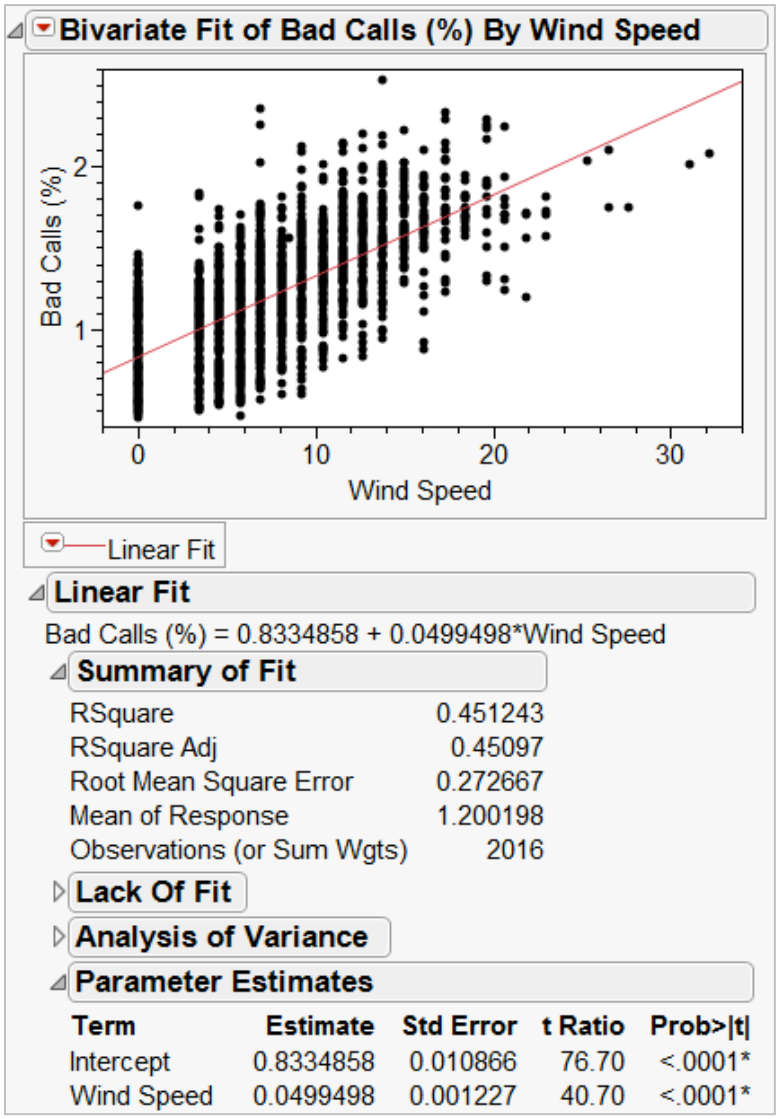
Exhibit 2 Distributions of **Bad Calls (%)**, **Wind Speed** and **Pressure** with Dynamic Linking



To better understand these patterns, we next construct two simple regression models.

A simple linear regression model (Exhibit 3) indicates that **Wind Speed** is statistically related to percentage of bad calls (the p-value is < 0.0001).

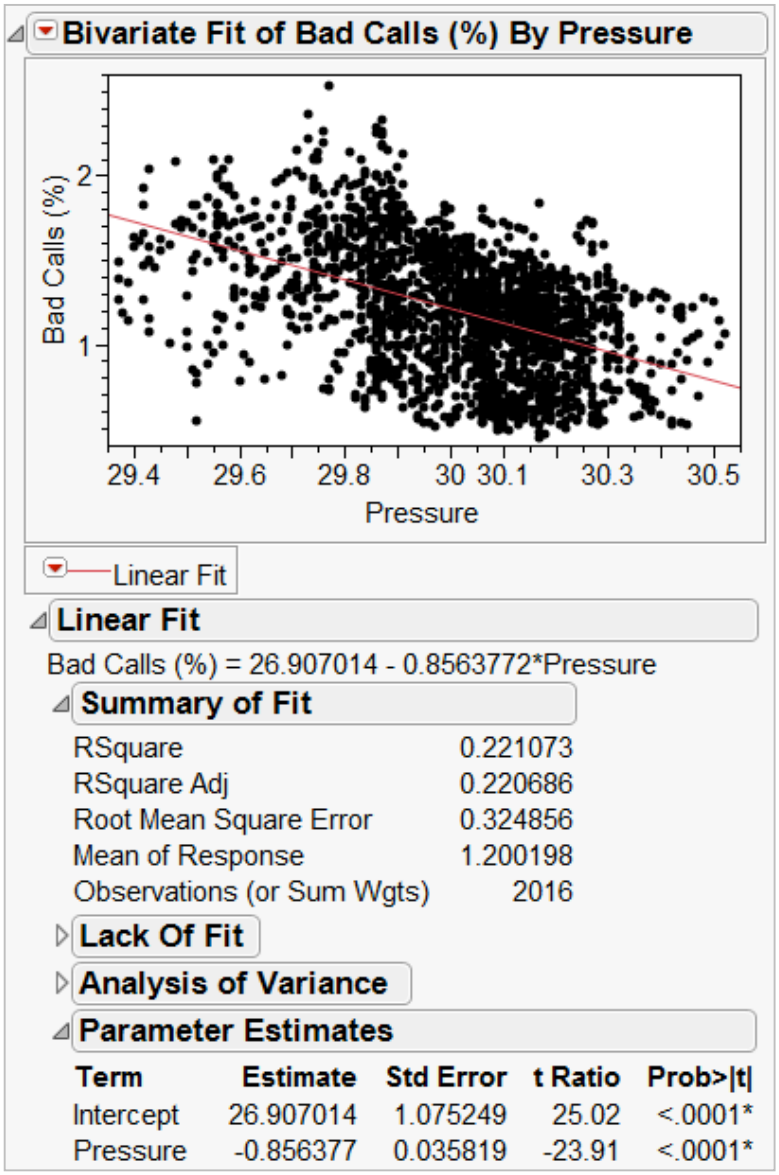
Exhibit 3 Regression with **Bad Calls (%)** and **Wind Speed**



*(Analyze > Fit Y by X. Use **Bad Calls (%)** as Y, Response and **Wind Speed** as X, Factor. Under the red triangle select Fit Line.)*

Exhibit 4 shows that barometric pressure is indeed related to percentage of bad calls, since **Pressure** is a statistically significant predictor of **Bad Calls (%)**:

Exhibit 4 Regression with **Bad Calls (%)** and **Pressure**



The table below summarizes the results from Exhibits 3 and 4.

	Regression Equation	RMSE	R-squared
Exhibit 3	**Bad Calls (%)** = 0.8 + 0.05*__Wind Speed__	0.27	0.451
Exhibit 4	**Bad Calls (%)** = 26.9 - 0.86*__Pressure__	0.32	0.211

The **Wind Speed** model estimates a 0.05% increase in bad calls per mile per hour of wind speed, and the model explains about 45.1% of the variation in percentage of bad calls. The **Pressure** model estimates a decrease in the rate of bad calls of 0.86% per inch of mercury, and about 22.1% of the variation in percentage of bad calls is associated with variation in pressure.

Each of these results, though, comes from *simple linear regression* models in which each model contains only one predictor—either wind speed or barometric pressure.

What would happen if we included *both* predictor variables in one regression model? Given the results shown in Exhibits 3 and 4, it would be natural to think that a *two-predictor model* would generate these results:

Bad Calls (%) = ? + 0.05***Wind Speed** – 0.86***Pressure**

R-squared = 0.451 + 0.221 = 0.672

After all, since 45.1% of the variation in **Bad Calls (%)** is associated with variation in wind speed, and 22.1% of the variation in **Bad Calls (%)** is associated with variation in barometric pressure, then the combined influence should be 67.2%. Right?

Exhibit 5, the two-predictor model, suggests otherwise.

Exhibit 5 Regression with both **Pressure** and **Wind Speed**

Response Bad Calls (%)

Summary of Fit

RSquare	0.480295
RSquare Adj	0.479779
Root Mean Square Error	0.265417
Mean of Response	1.200198
Observations (or Sum Wgts)	2016

Analysis of Variance

Lack Of Fit

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	11.486495	1.004297	11.44	<.0001*
Pressure	-0.353204	0.033296	-10.61	<.0001*
Wind Speed	0.0430728	0.001359	31.69	<.0001*

*(Analyze > Fit Model; select **Bad Calls (%)** as Y and **Pressure** and **Wind Speed** as a model effects, and hit Run. Some of the output has been hidden.)*

Note that both slope estimates have declined in absolute value. Note too that the R-squared value in Exhibit 5 is 0.48, quite a bit lower than the suspected 0.672. Thus, the regression results for the two-predictor model are, in actuality,

Bad Calls (%) = 11.49 + 0.04***Wind Speed** – 0.35***Pressure**

R-squared = 0.48

Why is the two-predictor regression model different from what we might have expected?

Exhibit 6 displays the correlations among the three variables. As expected, **Bad Calls (%)** is positively correlated with **Wind Speed** (r = 0.672) and negatively correlated with barometric **Pressure** (r = 0.47). More importantly, though, is the correlation *between the two predictor variables*, **Wind Speed** and **Pressure** (r = -0.477): higher barometric pressure is related to lower wind speed, a well-known atmospheric relationship.

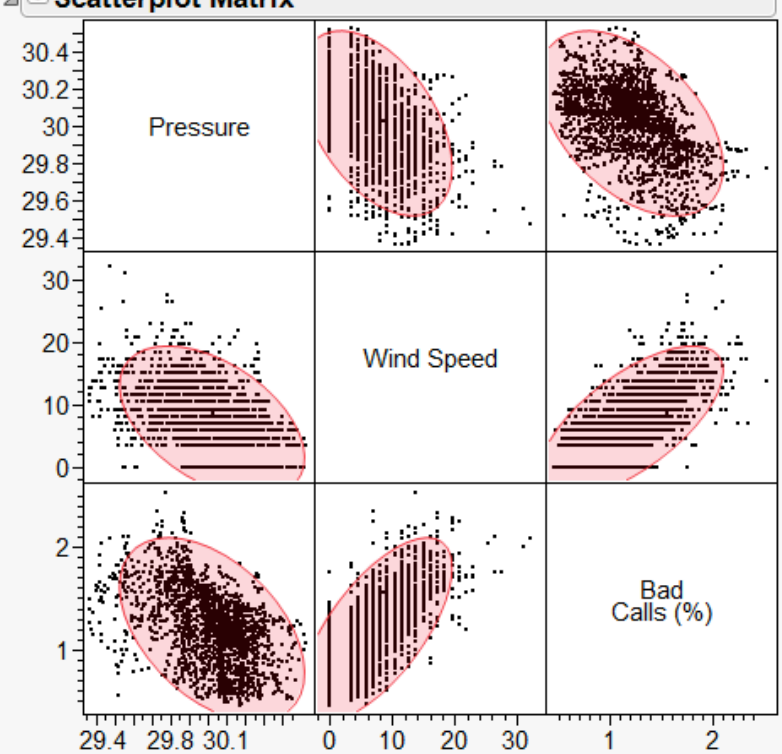
Exhibit 6 Correlations and Scatterplot Matrix

Multivariate

Correlations

	Pressure	Wind Speed	Bad Calls (%)
Pressure	1.0000	-0.4769	-0.4702
Wind Speed	-0.4769	1.0000	0.6717
Bad Calls (%)	-0.4702	0.6717	1.0000

Scatterplot Matrix



*(Analyze > Multivariate Methods > Multivariate; select **Pressure**, **Wind Speed** and **Bad Calls (%)** as Y, Columns, and hit OK. Under the lower red triangle, select Shaded Ellipses.)*

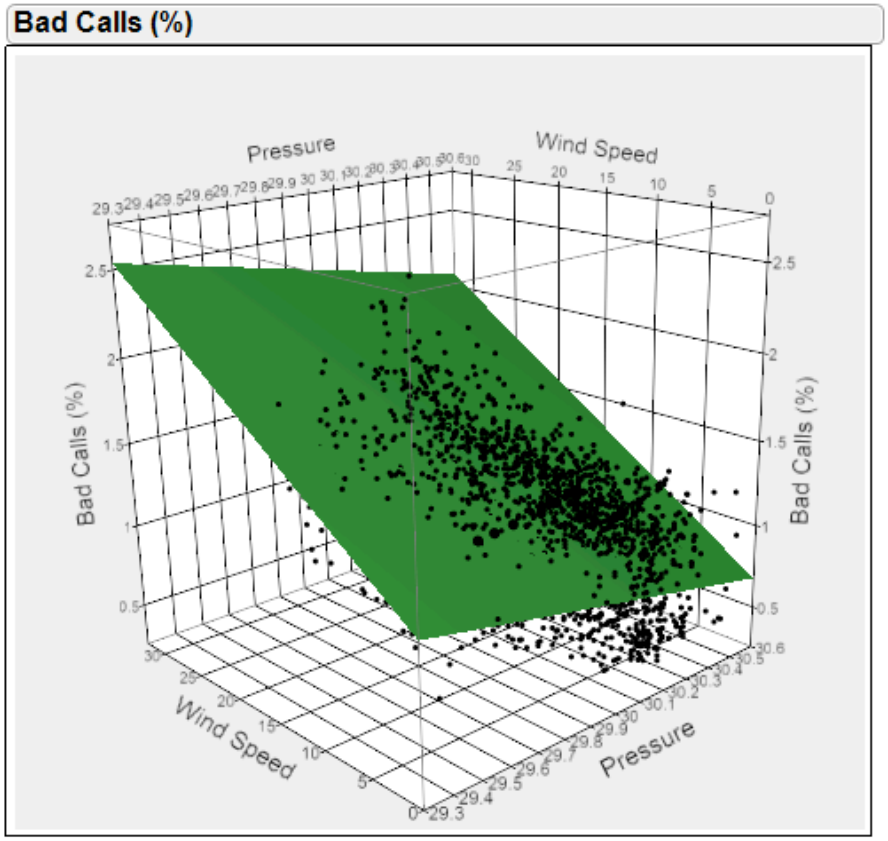
In essence, the simple linear regression models in Exhibits 3 and 4 show:

- the relationship between percentage of bad calls and wind speed *without regard* to the influence of the pressure (Exhibit 3), and
- the relationship between percentage of bad calls and pressure *without regard* to the influence of the wind speed (Exhibit 4),

whereas Exhibit 5 indicates the *combined* influence of pressure *and* wind speed on call performance, including any interrelationship between the two predictor variables.

A dynamic graph for visualizing the relationships between the two predictors and the response is a three-dimensional scatterplot (Exhibit 7).

Exhibit 7 Three-Dimensional Scatterplot and Surface Profile of Model



(In the Fit Model analysis window (see Exhibit 5), select Factor Profiling > Surface Profiler from the top red triangle. To display the data points, select Actual under Appearance.)

In regression involving one predictor, we produce a two-dimensional scatterplot to represent the relationship between the two variables, and the regression model is a line. With two-predictor models, the model is actually *plane* in three-dimensions.

Summary

Statistical Insights

This case looks at regression models containing two predictor variables. The term "regression model" takes on different meanings depending on whether that model contains one or two predictor variables.

- **One Predictor**. The "regression model" is the *line* that best fits the scatter of points in 2-D. A *scatterplot* is used to visualize the relationship between a response variable and one predictor variable. Mathematically, we would write the regression model as:

$$\hat{Y} = b_0 + b_1 X_1$$

- **Two Predictors**. The "regression model" is the *plane* that best fits the cloud of points in 3-D. A *surface profile* is used to visualize the relationship between a response variable and two predictor variables. The generic mathematical representation is:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

A multiple regression model (one with two or more predictor variables) combines the influence of each predictor variable on the response variable after accounting for the linear influence of the other predictor variables in the model. We'll see more about this in the exercise.

In this particular data set, the R-squared statistic from the two-predictor model (Exhibit 5) was smaller than the sum of the R-squared statistics from the two simple regression models (Exhibits 3 and 4). That need not always be so. *Suppression* is the term used to describe the opposite outcome, although it occurs less often in practice.

This case demonstrates that inclusion of all relevant factors into a regression model is important when trying to get a complete picture of relationships among many variables. Thus, the next step of the statistical analysis would be to include predictors in addition to weather that might also influence cell call performance.

Managerial Implications

There is some evidence to suggest that interrupted cell phone service is related in part to climate factors outside of the control of management. Also outside of management's control is the location of the caller; wind-induced fading won't be much of a problem for customers in areas with limited foliage. If indeed wind-induced fading is in play, we might see less interrupted service during winter and spring months when deciduous trees are without leaves. Management might ask that an expanded study be conducted to determine whether there are seasonal influences at work.

JMP Features and Hints

To estimate regression models with *one* predictor variable, use Fit Y by X. Regression models with one predictor can also be explored using Graph > Graph Builder. To estimate regression models with *two or more* predictor variables, use Fit Model. As always, explore your variables using the Distribution platform and Graph Builder before beginning any formal analysis.

Many analysis options are available Fit Model, including leverage plots. A leverage plot shows whether an individual predictor is statistically significant; it displays the influence of a predictor variable on Y *after accounting for the linear influence of the other predictor*.

Exercise 1

Estimate a regression model for **Bad Calls (%)** with **Wind Speed**, **Pressure**, and **Hour** as the predictors.

1. With **Wind Speed** and **Pressure** in the model, is **Hour** a significant predictor of **Bad Calls (%)**?
2. Compare the RMSE and R-Squared values for this model to the values for the two-predictor model. Does this three-predictor model do a better job of explaining variation in **Bad Calls (%)** than the two-predictor model?

Exercise 2

This exercise is designed to show more formally what is meant by multiple regression accounting for the linear influence of other predictor variables in the model.

1. Estimate a regression model in which **Bad Calls (%)** as the response variable and **Wind Speed** is the predictor. From this regression model, save the *residuals* to the JMP worksheet (red triangle next to Linear Fit). Call these residuals **Step 1 Resids**.
2. Estimate a regression model in which **Pressure** is the response variable and **Wind Speed** is the predictor variable. From this regression model, save the *residuals* to the JMP worksheet; name them **Step 2 Resids**.
3. Estimate a regression model in which the **Step 1 Resids** serve as the response variable and the **Step 2 Resids** serve as the predictor variable.

Explain any similarity you observe between the slope from the regression in step 3 and the slopes from the two-predictor regression model shown in Exhibit 6.

Next, do this:

4. Estimate a regression model in which **Bad Calls (%)** is the response variable and **Pressure** is the predictor. From this regression model, save the *residuals* to the JMP worksheet.
5. Estimate a regression model in which **Wind Speed** is the response variable and **Pressure** is the predictor variable. From this regression model, save the *residuals* to the JMP worksheet.
6. Estimate a regression model in which the residuals from step 5 serve as the response variable and the residuals from step 5 serve as the predictor variable.

SAS Institute Inc. World Headquarters +1 919 677 8000

JMP is a software solution from SAS. To learn more about SAS, visit **www.sas.com**
For JMP sales in the US and Canada, call 877 594 6567 or go to **www.jmp.com**

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries.
® indicates USA registration. Other brand and product names are trademarks of their respective companies. S81971.1111