Exploring the Data

In this demonstration, we explore the data to get a sense of the relationships between **Price** and the other continuous variables in the **cars** data set.

First we use PROC SGSCATTER to generate plots of **Price** versus the other variables. Notice that all of the variables to be plotted against **Price** are listed in parentheses.

Let's submit the code.

```
proc sgscatter data=mydata.cars;
   plot price*(citympg hwympg cylinders enginesize horsepower fueltank
        luggage weight);
run;
```

We can learn a lot by looking at these scatterplots:

- Both **Citympg** and **Hwympg** appear to have a negative relationship with **Price**, but the relationships do not appear to be linear. A quadratic term might be appropriate for these two variables.
- **Cylinders** might not be a true continuous variable because it only takes on a few values. However, the values are ordinal, so it could be used as a numeric variable in a regression.
- Each of the three variables **EngineSize**, **Horsepower**, and **FuelTank** appears to have a positive relationship with **Price**. The relationship between **FuelTank** and **Price** might be curvilinear.
- However, there does not appear to be a relationship between **Luggage** and **Price**.
- **Weight** seems to have a positive relationship with **Price** and the relationship might be curvilinear.

The patterns in several of the plots—for example, **Price** versus **Weight** and **Price** versus **EngineSize**--might point toward nonconstant variance and weaker associations.

Let's create a second set of scatterplots to examine more closely the relationships between **Price** and the variables where curvature is seen. In the PLOT statement in the new PROC SGSCATTER step, notice that the list of variables in parentheses has been reduced to **Citympg**, **Hwympg**, **FuelTank**, and **Weight**. It might be useful to add a smooth curve to the plots. One way to do this is to fit a penalized B-spline curve by adding the PBSPLINE option to the PLOT statement. In addition, in the ODS GRAPHICS statement, we can use the IMAGEMAP=ON option to turn on data-tips generation for HTML output. A data tip is explanatory text that appears when you move your mouse pointer over a data point in a graph in an HTML page, as you'll see in a minute.

We run this code.

```
ods graphics / imagemap=on;

proc sgscatter data=mydata.cars;
   plot price*(citympg hwympg fueltank weight) / pbspline;
run;
```

With the curves added to the plots, you can see that **Citympg** and **Hwympg** seem to exhibit a quadratic relationship with **Price**, so quadratic terms for these two variables will be added to the model. The curve in the scatter plot of **Price** versus **FuelTank** shows a positive curvilinear relationship up to the values of **FuelTank** at approximately 20 gallons. Then the curve turns downward. This might be due to the two lower points in the graph. Because tooltips are turned on in the graphical output, you can see that these two points have a value of 23 for **FuelTank**. The top data point has a price of 23.7 and the lower data point has a price of 18.8. (These two points might be influential and are examined later in the course.) For a polynomial model, it might be appropriate to start with cubic and quadratic terms for **FuelTank** in the model.

In the plot of **Price** versus **Weight**, notice the increasing spread toward the right. The apparent curvature for **Weight** might be evidence of increasing variability in the data, or it might just be a curvilinear relationship. You might choose to include both the linear and quadratic terms of **Weight**, or because the curvature is slight, only include the linear term. For now, we can proceed with the regression and then check the assumption of homogeneity of variances during model diagnostics. (Model diagnostics are discussed in a later lesson.)

Now we want to generate the correlations between the variables using PROC CORR. The VAR statement specifies the response variable and all of the predictor variables. This will produce correlation results for **Price** and each predictor

variable as well as between predictor variables. In this PROC SGSCATTER step, the MATRIX statement requests a matrix plot of the independent variables.

We submit this code.

```
proc corr data=mydata.cars nosimple;
   var price citympg hwympg cylinders enginesize horsepower fueltank
       luggage weight;
run;

proc sgscatter data=mydata.cars;
   matrix citympg hwympg cylinders enginesize horsepower fueltank
          luggage weight;
run;
```

In the PROC CORR results, we see quite a few significant correlations. **Horsepower** has the strongest correlation with **Price** (0.81667), so we conclude that this variable would be one of the best variables to include in a regression model. However, recall that the Pearson correlation statistic measures the linear relationship between variables. **Citympg**, **Hwympg**, and **FuelTank** appeared to have a relationship with **Price** that was not linear. A correlation analysis does not reveal these relationships. It might be that these variables are better predictors of **Price** if we consider the nature of their relationships. Also many of the variables that are potential independent variables are highly correlated with one another. Correlation of the independent variables can cause model instability. This should be taken into consideration when you develop the model. The scatter plots of the independent variables might help you visualize these relationships.

Most of the independent variables are related to one another, so the majority of the plots indicate strong linear or curvilinear patterns. For example, look at the strong linear relationship that is evident between **Citympg** and **Hwympg**, or the strong curvilinear relationship among **Citympg** and **EngineSize**, **Horsepower**, **FuelTank**, and **Weight**. Because most of the plots indicate strong relationships, it is easier to look for plots with the weakest indicated patterns. The plots that seem to indicate the weakest relationships are those for **Luggage** versus the other predictor variables. Even some of those plots, such as the plots of **Luggage** versus **FuelTank** and **Luggage** versus **Weight** and **EngineSize**, indicate that some relationship might be present. Notice that many of the relationships among the independent variables are curvilinear.

Close