Performing Model Diagnostics – Part 2

We ended the previous demonstration by using PROC CORR to compute the Spearman Rank Correlation. Now, let's return to the previously generated PROC REG results. To assess model fit, let's examine the plots of residuals versus predictor variables, the R-F plot, and the plot of the observed values versus the predicted values.

First, let's examine the plots of the residuals versus predictor variables. The variables **s_Hwympg** and **s_Hwympg^2** show a linear pattern for large values of the predictor variable. This indicates that the model does not fit well for cars with higher gas efficiency. These two plots also confirm the nonconstant variance. We can also see evidence of the nonconstant variance in the plot of the residuals by **Horsepower**.

Now, let's look at the Residual-Fit Spread Plot. On the left is the spread of the fitted values minus the Mean. On the right side, we have the spread of the residuals. On the left there is a spread of about -10 to 25, so a total range of about 35. In the right panel, there is a spread from approximately -10 to 15, for a range of approximately 25.

Because the spread of the residual distribution is smaller than that of the fitted values, and the model has an R square of 0.7192, the results indicate that the model explains most of the variation in the data.

Another way to look at model fit is to examine the plot of Observed versus Predicted values. If the values were identical (that is, if the model were an exact fit), then all of the points would be on the line. Points tightly clustered around the line indicate that our model fits the data well. As the prices get larger, however, you do see a bit more scatter (or variability), confirming the trend we saw elsewhere. There is greater variability for the cars with higher prices. As we saw in the plot of the residuals versus the predicted values, the Dodge Stealth, Mercedes-Benz 190E, and the Audi 100 lie the farthest from the fitted line.

Now we'll look at the statistics for identifying influential observations. Two of the quantities that we can use to detect influential observations are leverage, which measures the distance of a point from the cloud of data points, and then also the RSTUDENT residuals, which measures the change in the residuals when you subtract an observation. SAS gives us reference lines that correspond to the cutoffs that determine whether an observation is an outlier. For a RSTUDENT residual being above two or below negative two denotes that that observation is influential. For leverage, SAS puts the demarcation line at 0.1.

The points in red in the top left portion of the plot are influential on the RSTUDENT residual, but not on the leverage. The points in green in the middle are influential on the leverage, but not on the RSTUDENT residual. The points in blue on the far left are not influential according to either statistic. There is only one observation (the Dodge Stealth) that is influential based on both statistics.

Let's look at the DFFITS, which measures the impact on predicted value of deleting an observation. SAS again provides reference lines for the cutoff values. Again, SAS labels the observations that are influential.

Remember that SAS computes a DFBETA for each parameter estimate in the model. This tells us whether an observation is influential on the parameter estimate for the intercept, the partial slope for **Hwympg**, the partial slope for **Hwympg^2**, and the partial slope for **Horsepower**. We do see quite a number of observations in these plots that are influential.

Let's look at the partial residual plots to look for outliers and influential points. Most of these plots seem fine. We do see points that could be outliers on the plot for **Hwympg^2**.

From looking at all of these graphs we can see that there are influential observations. Now we should go back to the data set and the model fit itself. It might be helpful to output these influential observations to a data set. Then we can consider each influential observation, based on subject matter knowledge. Perhaps we're missing a variable that would help explain this outcome.

Now we'll do a first pass at identifying observations that are influential on either the RSTUDENT residual or leverage. We'll be using two macro variables to specify the number of parameters and the number of observations. Remember that the cutoff value for leverage was 2 times *p* divided by *n*. We're using the macro variables in the computation for the cutoff value.

```
%let numparms = 4;
%let numobs = 81;
data influence;
```

```
   set check;
   absrstud=abs(rstudent);
   if absrstud ge 2 then output;
   else if leverage ge (2*&numparms /&numobs) then output;
run;

proc print data=influence;
   var manufacturer model price hwympg horsepower;
run;
```

The DATA step creates a data set named **influence**. We read in the **check** data set. We compute a variable **Absrstud**, which is the absolute value of the RSTUDENT residual. If the value is outside the cutoff value, then SAS sends it to the output data set. Also, SAS will look for observations that are influential based on the leverage statistic and output those as well.

Let's run this code. The results display nine observations that appear to be influential based on LEVERAGE or RSTUDENT statistics. Notice that all of the observations were flagged by at least two, and in some cases, by three or four, influence statistics. We should check the data for these observations to ensure that no transcription or data entry errors occurred. If the data are erroneous, we'll correct the errors and re-analyze the data.

It's also possible that the model is not adequate. Notice that most of these cars fall at the high or low end of the range for **Price**. There might be another variable, such as one indicating whether the car is a luxury, midrange, or economy car, that would be important in explaining these unusual observations. Another possibility is that the observation, though valid, might be unusual. If we had a larger sample size, there might be more observations like the unusual ones. We might need to collect more data to confirm the relationship suggested by the influential observation.

In general, we should not exclude data. Often, unusual observations contain important information. If there is good reason to exclude some observations, we should document why, including a description of the types of observations that we exclude and why. We also need to document the limitations of our conclusions, given these exclusions, as part of a report or presentation.

Close