Print

Practice 5.3 (Level 1): Conducting a Linear Regression Analysis

Task

In this practice, you conduct a linear regression analysis to make predictions based on the data in **mydata.softdrinks**.

A soft drink bottler is analyzing vending machine service routes in his distribution system. He is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. The service activities include stocking the machine with beverage products and minor maintenance as required. The industrial engineer responsible for the study suggested that the two most important variables affecting the delivery time are the number of cases of product stocked and the distance walked by the route driver. The engineer collected 24 observations on delivery time (in minutes), number of cases, and distance walked (in feet).

Reminder: Make sure you've defined the **mydata** library.

1. Write a PROC PRINT step to print the data set. Write a PROC SGPLOT step to create a histogram of the data on the original scale by using **Time** in the HISTOGRAM statement. Overlay the kernel density and a normal density. Do the times seem to be normally distributed?

```
proc print data=mydata.softdrinks;
   title 'Softdrink Data';
run;

proc sgplot data=mydata.softdrinks;
   histogram Time;
   density Time;
   density Time / type=kernel;
   title 'Softdrink Data - Original Scale';
run;
```

   As shown in the results, the data do not appear to be normally distributed. Because it is skewed to the right and has only positive values, the gamma distribution might be a good fit.

2. Create a histogram of the data on the log-tranformed scale by using **LogTime** in the HISTOGRAM statement in PROC SGPLOT. Overlay the kernel density and a normal density. What do you conclude about the distribution of the log-transformed times?

```
proc sgplot data=mydata.softdrinks;
   histogram LogTime;
   density LogTime;
   density LogTime / type=kernel;
   title 'Softdrink Data - Log Scale';
run;
```

   As shown in the results, the log-transformed data seeem to more closely resemble the normal distribution than the original data.

3. Use PROC GLIMMIX to model **Time** as a function of **Cases**, **Cases** squared, **Distance**, and **Distance** squared. In the MODEL statement, use the options DIST=GAMMA and LINK=LOG.

```
proc glimmix data=mydata.softdrinks;
   model Time=Cases Cases*Cases Distance Distance*Distance /
```

```
          solution dist=gamma link=log;
title 'Softdrink Data - Gamma Regression with Log Link';
run;
```

As shown in the results, the **Distance*Distance** term has an estimate of essentially 0 and a standard error of 0, so model reduction begins when you remove this term.

Note: For the gamma distribution, the scale parameter reported in PROC GLIMMIX is the reciprocal of the scale parameter reported in PROC GENMOD.

4. Eliminate any terms that are not significant to obtain a final "candidate" model.
    - Use the PLOTS=STUDENTPANEL (UNPACK) option in the PROC GLIMMIX statement. This creates a plot of the studentized residuals versus the linear predictor.
    - Use the OUTPUT OUT=**gamma_predicted** statement to save the predicted values in the data set **gamma_predicted**.
    - If you want to be able to identify particular observations in any of the graphs, enable IMAGEMAP using the ODS GRAPHICS statement.

   After looking at the plots of the studentized residuals, do you think the model fits well?

```
ods graphics / imagemap=on;
proc glimmix data=mydata.softdrinks plots=studentpanel(unpack);
   model Time=Cases Cases*Cases Distance /
         dist=gamma link=log solution;
   output out=gamma_predicted pred(ilink)=pred;
title 'Softdrink Data - Final Model';
run;
```

Examine the results. As shown in the Parameter Estimates table, after you remove **Distance*Distance**, all terms in the model have reasonable estimates and are significant. As shown in the plots, the residuals seem to be a random scatter about zero with no apparent patterns. There are two data points with studentized residuals above 2 that you might want to investigate.

5. Use the **gamma_predicted** data set and PROC SGPLOT with the REG statement to plot the predicted times against the observed times. Use the DATALABEL=Distance option in the REG statement.

```
title 'Softdrink Data - Gamma Regression with Log Link - Final Model';
proc sgplot data=gamma_predicted noautolegend;
   reg y=pred  x=Time / datalabel=distance;
   title2 'Predicted Time versus Observed Time';
run;
title;
```

Examine the results. The predicted values for **Time** seem to align closely to the observed values, except for one data point, where the technician walked 770 feet to service the account. The model predicts the time needed to service the account for this observation.

Hide Solution

Copyright © 2017 SAS Institute Inc., Cary, NC, USA. All rights reserved.

Close