Print

Practice 1.1 (Level 1): Fitting a Simple Polynomial Regression Model

Task

In this practice, you fit a simple polynomial regression model.

An analyst for a cafeteria chain wanted to investigate whether the sales of coffee are related to the number of self-service coffee dispensers in a cafeteria line. Fourteen cafeterias that are similar in terms of volume of business, type of clientele, and location were chosen for the study. The number of self-service dispensers was assigned randomly at each cafeteria and the sales in hundreds of gallons of coffee were recorded. The data is stored in the data set **mydata.cafeteria**.

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

1. Write a PROC SGSCATTER step to plot the variables **Sales** versus **Dispensers**. Submit the step. How is **Sales** related to **Dispensers**?

   ```
   title "Scatter Plot of mydata.cafeteria Data Set";
   proc sgscatter data=mydata.cafeteria;
      plot Sales * Dispensers;
   run;
   quit;
   ```

   As shown in the results, as the number of dispensers increases, the sales increase as well. The relationship appears to be slightly curvilinear.

2. Write a PROC REG step to fit a simple linear regression model to the data. Add the PLOTS (ONLY UNPACK)=DIAGNOSTICS options in the PROC REG statement to get the diagnostics panel of plots. Submit the code.

   Examine the plot of the residuals versus the predicted values and the plot of the observed versus the predicted values. Does your model seem to fit the data well?

   ```
   proc reg data=mydata.cafeteria plots(only unpack)=diagnostics;
      model Sales=Dispensers;
   run;
   ```

   As shown in the tables at the top of the results, the model is significant with a *p*-value of <0.0001. The variable **Dispensers** has a positive slope of 55.71. As the number of dispensers increases by 1, the sales of coffee are expected to increase by 55.71 in hundreds of gallons. The R-square value is 0.974. The model explains about 97% of variation in **Sales**.

   The plot of the residuals versus the predicted values shows a curvilinear relationship between the residuals and the predicted values. The linear model might not fit your data well.

   In the plot of the observed values versus the predicted values, the regression model seems to be close to the observed data, but the data points do not seem to be randomly dispersed around the regression line. This is another indication that the linear model might not fit your data well.

   A good next step is to try fitting a quadratic model to your data.

3. To fit a quadratic model, write a PROC GLMSELECT step and include the EFFECT statement. Use the OUTDESIGN option to output the design matrix to a data set named **d_disp**. Is the model significant?

Add a PROC REG step with the **&_GLSMOD** macro variable to request Type I tests and the DIAGNOSTICS panel of plots.

Add a PROC SGPLOT step. Use the REG statement with the DEGREE=2 option to create a scatter plot of the data with a second-degree regression overlaid.

Submit the code and examine the results to answer the following questions:

- Is the model significant?
- Look at the plot of the residuals versus the predicted values, the plot of the observed versus the predicted values, and the normal quantile plot for residuals. Do you think the quadratic model fits your data better than the linear model?

```
proc glmselect data=mydata.cafeteria outdesign=d_disp;
   effect q_disp=polynomial(Dispensers / degree=2);
   model Sales = q_disp / selection=none;
run;

proc reg data=d_disp plots(only unpack) = diagnostics;
   model Sales=&_GLSMOD / scorr1(tests);
run;
quit;

proc sgplot data=mydata.cafeteria;
   reg y=Sales x=Dispensers / degree=2;
run;
```

The results indicate the following:

- At the top of the PROC GLMSELECT, the tables show that the model is significant. The adjusted R-square value increases to 0.9955.

- In the PROC REG results, the Type 1 test in the Parameter Estimates table indicates that both **Dispensers** and **Dispensers^2** are significant factors. The residual plots show a random scatter around the reference line. The plot of the observed values versus the predicted values indicates a better model fit than a linear model. From the Q-Q plot, the residuals seem to be normally distributed.

- In the PROC SGPLOT output, the scatter plot with the overlaid quadratic model indicates that this model fits the data better than the linear model.

Hide Solution

Close