Print

Demo: Fitting a Multiple Linear Regression Model Using PROC REG

Filename: **st103d03.sas**

In this demonstration, we use PROC REG to run a linear regression model with two predictor variables. Then we use PROC GLM to fit the same model again to show a few additional plots that are not available in PROC REG. We'll save the results of our analyses in an item store, and then use PROC PLM to perform additional analysis.

PROC REG DATA=*SAS-data-set* <*options*>;
    MODEL *dependent-variables* = *regressors* < / *options*>;
RUN;

PROC GLM DATA=*SAS-data-set* <*options*>;
    MODEL *dependent-variables* = *independent-effects*;
    STORE <OUT=> *item-store-name* < / **LABEL=***'label'*>;
RUN;

PROC PLM RESTORE=*item-store-specification* <*options*>;
    EFFECTPLOT <*plot-type* <*(plot-definition-options)*>> </*option(s)*>;
RUN;

1. Open program st103d03.sas.

```
/*st103d03.sas*/  /*Part A*/
ods graphics on;

proc reg data=STAT1.ameshousing3 ;
    model SalePrice=Basement_Area Lot_Area;
    title "Model with Basement Area and Lot Area";
run;
quit;

/*st103d03.sas*/  /*Part B*/
proc glm data=STAT1.ameshousing3
         plots(only)=(contourfit);
    model SalePrice=Basement_Area Lot_Area;
    store out=multiple;
    title "Model with Basement Area and Gross Living Area";
run;
quit;

/*st103d03.sas*/  /*Part C*/
proc plm restore=multiple plots=all;
    effectplot contour (y=Basement_Area x=Lot_Area);
    effectplot slicefit(x=Lot_Area sliceby=Basement_Area=250 to 1000 by 250);
run;

title;
```

In the PROC REG step, the MODEL statement specifies SalePrice as the response variable, and

Basement_Area and Lot_Area as predictors.

2. Submit the PROC REG step in Part A.

3. Review the output.
   The Analysis of Variance table shows that this model is statistically significant at the 0.05 alpha level.

   In the Fit Statistics table, the R-square of 0.4802, indicates that 48% of the variability in SalePrice can be explained by both Basement_Area and Lot_Area. Recall from a previous model that Lot_Area alone explained only 6.42%. Is the R-square higher because the new model is better, or simply because the model has more predictors? To find out, compare the adjusted R-square values.

   The simpler model with only Lot_Area had an adjusted R-square of 0.061. The adjusted R-square for the multiple regression is higher, at 0.4767. The higher adjusted R-square indicates that adding Basement_Area improved the model enough to warrant the additional model complexity.

   Let's look at the Parameter Estimates tables. Our earlier analysis showed that the correlation between Lot_Area and SalePrice was statistically significant. With Basement_Area added to the model, the Lot_Area estimate is notably different than it was in the simple linear regression model (2.87 in the simple regression model and 0.80 in this model), and its p-value no longer shows statistical significance.

   The reason is that in the two-predictor model, the parameter estimate for each predictor variable is adjusted for the presence of the other variable in the model. Basement_Area is a significant predictor of SalePrice even after controlling for Lot_Area. But Lot_Area is not a significant predictor of SalePrice after controlling for Basement_Area. This means that Lot_Area and Basement_Area are correlated, and Lot_Area does not explain significant variation in SalePrice over and above Basement_Area. So, when the model accounts for the effect of Basement_Area, the effect of Lot_Area no longer shows statistical significance.

   If these were our only predictors, we'd consider removing Lot_Area from the model, but we might decide to add other predictors instead. The additional predictors might change the p-values for Lot_Area and Basement_Area, so it's best to wait to see the full model before discarding non-significant terms.

   Now, let's use the Fit Diagnostics graphical output to verify our statistical assumptions. The residuals plotted against predicted values give us a relatively random scatter around 0. They provide evidence that we have constant variance.

   In the Q-Q plot, the residuals fall along the diagonal line, and they look approximately normal in the histogram. This indicates that there are no problems with an assumption of normally distributed error.

   Next, in the Residual Plots, we see the residuals plotted against the predictor variables. Patterns in these plots are indications of an inadequate model. The residuals show no pattern, although lot size does show a few outliers.

4. Let's go back to the code. In the second step, we'll run the same model in PROC GLM, requesting a contour plot and an item store named multiple.

5. Submit the PROC GLM step in Part B.

6. Review the output.
   In the ANOVA results, we see that the values in the Fit Statistics table are the same as in PROC REG. PROC GLM doesn't report an adjusted R-square value.

   The Solution, or parameter estimates table gives the same results (within rounding error) as in PROC REG.

   We can use this Contour Fit Plot with the overlaid scatter plot to see how well the model predicts observed values. The plot shows predicted values of SalePrice as gradations of the background color from blue, representing low values, to red, representing high values. The dots are similarly colored, and represent the actual data. Observations that are perfectly fit would show the same color within the circle as outside the circle. The lines on the graph help you read the actual predictions at even intervals.

   For example, this point near the upper right represents an observation with a basement area of approximately 1,500 square feet, a lot size of approximately 17,000 square feet, and a predicted value of more than $180,000 for sale price. However, the dot's color shows that its observed sale price is actually closer to $160,000.

7. Let's go back to the code. In the last step, we use PROC PLM to process the item store created by PROC GLM and create additional plots. The EFFECTPLOT statement produces a display of the fitted model and provides options for changing and enhancing the displays. The EFFECTPLOT option, CONTOUR, requests a contour plot of predicted values against two continuous predictors. We want Basement_Area plotted on the Y axis, and Lot_Area on the X axis.

   The SLICEFIT option displays a curve of predicted values versus a continuous variable grouped by the levels of another effect. We want to see the Lot_Area effect at different values of Basement_Area. with tick marks ranging from 250 to 1000, in increments of 250.

8. Submit the PROC PLM step in Part C.

9. Review the output.
   Notice that the lines in the Contour Fit Plot are oriented differently than the plot from PROC GLM. The item store doesn't contain the original data, so PROC PLM can show only the predicted values, not the individual observed values. Clearly, the PROC GLM contour fit plot is more useful, but if you don't have access to the original data, and you can run PROC PLM on the item store, this plot gives you an idea of the relationship between the predictor variables and predicted values.

   The last plot, a Sliced Fit Plot, is another way to display the results of a two-predictor regression model. This plot displays SalePrice by Lot_Area, categorized by Basement_Area. The regression lines represent the slices of Basement_Area that we specified in the code. As you can see, you have several options for visualizing and communicating the results of your analyses.

   Okay. We've created a multiple regression model with two predictors, so what's next? We can test the remainder of our predictors in a larger multiple regression model. With 11 predictors, there are many possible models that we could explore. As we've seen, the significance and parameter estimates for each predictor can change depending on which other predictors are included in the model.

   So how do we decide which model is best to go forward with? Ultimately, it will be decided by our specific research goal and our subject-matter knowledge. There are tools that we can use to limit the possible models to a manageable number of candidates. In the next lesson, we'll see some commonly used approaches to model selection.

Close