Demo: Producing Correlation Statistics and Scatter Plots Using PROC CORR

Filename: **st102d04.sas**

In this demonstration, we use PROC CORR to produce correlation statistics and scatter plots for our data. Our goal is to identify, both visually and numerically, which predictors are linearly associated with SalePrice, as well as the strength of the relationship. By default, PROC CORR produces Pearson correlation coefficients and corresponding p-values.

PROC CORR DATA=*SAS-data-set* <*options*>;
    VAR *variables*;
    WITH *variables*;
    ID *variables*;
RUN;

1. Open program st102d04.sas.

```
/*st102d04.sas*/  /*Part A*/
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
         Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;

ods graphics / reset=all imagemap;
proc corr data=STAT1.AmesHousing3 rank
          plots(only)=scatter(nvar=all ellipse=none);
   var &interval;
   with SalePrice;
   id PID;
   title "Correlations and Scatter Plots with SalePrice";
run;

title;

/*st102d04.sas*/  /*Part B*/
ods graphics off;
proc corr data=STAT1.AmesHousing3
          nosimple
          best=3;
   var &interval;
   title "Correlations and Scatter Plot Matrix of Predictors";
run;

title;
```

The PROC CORR statement specifies AmesHousing3 as the data set. To rank-order the absolute value of the correlations from highest to lowest, we're using the RANK option. To request individual scatter plots, we specify the PLOTS=SCATTER option. After the keyword SCATTER, there are two more options. NVAR=ALL specifies that all the variables listed in the VAR statement be displayed in

the plots, and ELLIPSE=NONE suppresses the drawing of ellipses on scatter plots. The VAR statement specifies the continuous variables that we want correlations for. By default, SAS produces correlations for each pair of variables in the VAR statement, but we'll use the WITH statement to correlate each continuous variable with SalePrice.

The IMAGEMAP option in the ODS GRAPHICS statement enables tooltips to be used in HTML output. Tooltips enable you to identify data points by moving the cursor over observations in a plot. In PROC CORR, the variables used in the tooltips are the X-axis and Y-axis variables, the observation number, and any variable in the ID statement, which in this case, is the variable PID.

2. Submit Part A of this program.

3. Review the output.

   By default, the CORR Procedure generates a table of Variable Information that lists the variables that were analyzed. It also displays a Simple Statistics table with descriptive statistics for each variable, including the mean, standard deviation, and minimum and maximum values.

   The Pearson Correlations table displays the correlation coefficients and p-values for the correlation of SalePrice with each of the predictor variables. Notice that the table is ranked by the absolute correlation coefficient. Basement_Area has the strongest linear association with the response variable with a correlation coefficient of about 0.69. Therefore, Basement_Area would be the best single predictor of SalePrice in a simple linear regression. The p-value is small, which indicates that the population correlation coefficient, ρ, is likely different from 0. The second largest magnitude correlation coefficient is Above Ground Living Area at about 0.65, and so on.

   Next we'll consider the scatter plots. In the SalePrice by Gr_Liv_Atea, the Above Ground Living Area seems to exhibit a noticeably positive linear association with SalePrice. Of course, the scatter plot of SalePrice by Basement_Area also shows a positive linear relationship. Notice that there are several houses that have basements with a size of zero square feet. These are houses without basements, not missing values. This mixture of data can affect the correlation coefficient. You would need to take this into account if you build a model with Basement_Area as a predictor variable. You can move the cursor over the observation to display the coordinate values, observation number, and ID variable values.

   The scatter plots with Deck_Porch_Area and Lot_Area show the variables have weak correlations with SalePrice, because a horizontal line could be an adequate line of best fit.

   As expected, SalePrice and the age of the house when sold, Age_Sold, have a negative linear relationship. The older the house, the less the home tends to sell for.

   The scatter plots with the total number of bedrooms (Bedroom_AbvGr) and bathrooms (Total_Bathroom) have few continuous values and could be analyzed as classification variables. These plots are basically displaying the distribution of SalePrice at each level, similarly to a box plot. However, the scatter plot with total number of bathrooms seems to exhibit a positive linear relationship, as the center of the distributions tends to increase as the number of bathrooms increases. Overall, the correlation and scatter plot analyses indicate that several variables might be good predictors for SalePrice.

   When you prepare to conduct a regression analysis, it's always a good practice to examine the correlations among the potential predictor variables. This is because strong correlations among predictors included in the same model can cause a variety of problems, like multicollinearity.

4. In Part B of this program, we want to produce a correlation matrix to help us compare the relationships between predictor variables. The correlation matrix shows correlations and p-values for all combinations of the predictor variables. Here we'll limit our attention to the strongest three correlations with each predictor.

   In this PROC CORR statement, we're using the NOSIMPLE option to suppress the printing of the simple descriptive statistics for each variable. The BEST= option prints the n highest correlation coefficients for each variable, so in this case, the three strongest correlations.

5. Submit this step.

6. Review the output.

   In the results, notice that the Variables Information table is still listed, but the table of simple statistics is gone.

   The Pearson Correlations table indicates that there are moderately strong correlations between Total_Bathroom and Age_Sold, -0.52889, between Total_Bathroom and Basement_Area, 0.48500, and between Bedroom_AbvGr and above Gr_Liv_Area, 0.48431.

   If some of these potential predictors were highly correlated, we might omit some from the multiple regression models that we'll produce later in the course. Strong correlations among sets of predictors, also known as multicollinearity, can cause a variety of problems for statistical models. Correlation analysis has the potential to reveal multicollinearity problems, but additional methods to detect it are necessary. Bivariate correlations in the range shown above are not causes for concern.

Close