Demo: Exploring Associations Using PROC SGSCATTER

Filename: **st102d01.sas**

In this demonstration, we want to get a general idea of the associations between the response variable SalePrice and predictor variables in our data. We also want to see the general shape of each association. We'll use PROC SGSCATTER to produce a variety of scatter plots. We'll start by generating a scatter plot to see whether there's an association between SalePrice and Above Ground Living Area.

PROC SGSCATTER DATA=*SAS-data-set* <*options*>;
    PLOT *plot-request(s)* **<** */ options*>;
RUN;

1. Open program st102d01.sas.

```
/*st102d01.sas*/  /*Part A*/
proc sgscatter data=STAT1.ameshousing3;
    plot SalePrice*Gr_Liv_Area / reg;
    title "Associations of Above Grade Living Area with Sale Price";
run;


/*st102d01.sas*/  /*Part B*/
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
         Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;

/*PROC SGSCATTER is used to explore relationships among continuous variables*/
/*using scatter plots*/
options nolabel;
proc sgscatter data=STAT1.ameshousing3;
    plot SalePrice*(&interval) / reg;
    title "Associations of Interval Variables with Sale Price";
run;


/*st102d01.sas*/  /*Part C*/
proc sgplot data=STAT1.ameshousing3;
    vbox SalePrice / category=Central_Air
                     connect=mean;
    title "Sale Price Differences across Central Air";
run;


/*st102d01.sas*/  /*Part D*/
%let categorical=House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces
         Season_Sold Garage_Type_2 Foundation_2 Heating_QC
         Masonry_Veneer Lot_Shape_2 Central_Air;

/*PROC SGPLOT is used here with the VBAR statement to produce vertical bar charts*/
/*PROC SGPLOT can only produce one plot at a time and so the macro is written to*/
/*produce one plot for each member in the list of the &categorical macro variable.*/
/*
      Macro Usage:
      %box(DSN = ,
           Response = ,
           CharVar = )
```

```
*/
%macro box(dsn      = ,
           response = ,
           Charvar  = );
%let i = 1 ;
%do %while(%scan(&charvar,&i,%str( )) ^= %str()) ;
    %let var = %scan(&charvar,&i,%str( ));
    proc sgplot data=&dsn;
        vbox &response / category=&var
                         grouporder=ascending
                         connect=mean;
        title "&response across Levels of &var";
    run;
    %let i = %eval(&i + 1 ) ;
%end ;
%mend box;

%box(dsn      = STAT1.ameshousing3,
     response = SalePrice,
     charvar  = &categorical);

title;
options label;
```

In Part A, the PLOT statement specifies the Y-axis variable, SalePrice, followed by an asterisk, and then the X-axis variable, Above Ground Living Area. The keyword reg after the forward slash tells SAS to include a regression line fit to the scatter plot.

2. Submit this step.

3. Review the output.

   The scatter plot seems to exhibit a linear association between SalePrice and Above Ground Living Area, as a non-horizontal line can be drawn depicting the general trend. Notice that there seems to be more variability in SalePrice as the sizes of the homes increase.

   Let's go back to our program and create additional scatter plots using several other predictor variables. An advantage of using PROC SGSCATTER is that you can create a panel of multiple scatter plots using a single PROC step.

4. In Part B of the program, the PLOT statement references the interval macro variable in parentheses. This will produce a scatter plot of SalePrice on the Y axis by all the interval variables on the X axis. Note that, if we wanted to, we could list multiple variables on the Y axis by putting them in parentheses as well. The OPTIONS NOLABEL statement tells SAS not to use labels for the variables.

5. Submit this step.

6. Review the output.

   In the SGScatter panel, all the plots exhibit some degree of linear association between each predictor and SalePrice. Most scatter plots demonstrate a positive linear association, meaning as one variable increases, the other tends to increase as well. However, SalePrice and Age_Sold are inversely related, meaning as the age of the home increases, the SalePrice tends to decrease.

Close