Print

Demo: Performing a Two-Way ANOVA Using PROC GLM

Filename: **st103d01.sas**

In the Ames Housing example, we want to consider the effect that heating system quality and season sold have on home sale prices. We'll start by exploring the data using the MEANS and SGPLOT procedures.

```
PROC MEANS DATA=SAS-data-set <statistic-keyword(s)>;
    CLASS variable(s) < / option(s)>;
    VAR variable(s);
RUN;
```

```
PROC SGPLOT DATA=SAS-data-set <option(s)>;
    VLINE category-variable < / option(s)>;
RUN;
```

```
PROC GLM DATA=SAS-data-set <options>;
    CLASS variable(s);
    MODEL dependent-variable = independent-effects < / options>;
    LSMEANS effects < / options>;
RUN;
```

1. Open program st103d01.sas.

```
/*st103d01.sas*/  /*Part A*/
ods graphics off;
proc means data=STAT1.ameshousing3
           mean var std nway;
    class Season_Sold Heating_QC;
    var SalePrice;
    format Season_Sold Season.;
    title 'Selected Descriptive Statistics';
run;

/*st103d01.sas*/  /*Part B*/
proc sgplot data=STAT1.ameshousing3;
    vline Season_Sold / group=Heating_QC
                        stat=mean
                        response=SalePrice
                        markers;
    format Season_Sold season.;
run;
```

```
/*st103d01.sas*/  /*Part C*/
ods graphics on;

proc glm data=STAT1.ameshousing3 order=internal;
    class Season_Sold Heating_QC;
    model SalePrice = Heating_QC Season_Sold;
    lsmeans Season_Sold / diff adjust=tukey;
    format Season_Sold season.;
    title "Model with Heating Quality and Season as Predictors";
run;
quit;

title;
```

In Part A, the PROC MEANS step requests summary statistics for the variables of interest. The analysis variable, SalePrice, is named in the VAR statement, and the classification variables, Season_Sold and Heating_QC, are listed in the CLASS statement. The NWAY option requests the combination of all variables named in the CLASS statement. The FORMAT statement applies the Season format to the Season_Sold variable to display Winter, Spring, Summer, and Fall, instead of the corresponding numeric values, 1, 2, 3, and 4.

We use the SGPLOT procedure in Part B to plot the mean SalePrice by Season_Sold in a vertical line chart with the bars grouped by Heating_QC. The MARKERS option adds data point markers to the chart.

2. Submit Parts A and B to run both steps.

3. Review the output.

   In the Summary Statistics generated by the MEANS procedure, the mean sale price is lowest for houses with fair heating systems. The table of means also shows that few houses with fair heating are sold regardless of season. For example, only one fair-heating-quality house was sold in the fall, which is why there is no standard deviation or variance for that mean. We can't be as confident about estimated means that are based on small samples sizes. Looking at the graph produced by the SGPlot procedure, the season that a home sold doesn't seem to affect the sale price very much, except where the heating system is fair. For those homes, the mean sale price seems markedly lower in the colder seasons.

   How does this exploratory plot help us plan our analysis? Well, we see that the effect of the heating quality on sale price seems to depend on the season the house is sold. This indicates a possible interaction effect. Well use PROC GLM to first test only the main effects of Season_Sold and Heating_QC. Later, we'll incorporate the interaction suggested by our plot.

4. In the PROC GLM step, the ORDER=INTERNAL option tells SAS to use the order of the variable values stored internally, rather than the order of the formatted values. The internal values for Season_Sold are 1, 2, 3, and 4, so by including this option, the seasons will appear in the order Winter, Spring, Summer, and Fall, instead of alphabetical order. In the MODEL statement, SalePrice is the dependent variable, and Heating_QC and Season_Sold are the factors or model effects. In PROC GLM, the order of variables in the CLASS statement determines the look of the graph. The first variable labels the X axis, and the second variable is represented by the color-coded lines. The LSMEANS statement requests a Tukey-adjusted analysis of the difference across all seasons.

5. Submit the PROC GLM step in Part C.

6. Review the output.

We're testing to see whether all means are equal for each predictor variable. In the Analysis of Variance table, the degrees of freedom is 6, because Season_Sold and Heating_QC each accounts for three degrees of freedom, the number of levels minus 1 for each variable. The statistically significant p-value indicates not all means are equal for each predictor variable, but it doesn't indicate which mean values are significantly different. We can determine which means differ by looking at the table showing tests of individual factors. In the Fit Statistics table, the R-square value, 0.171954, indicates approximately 17% of the variability in SalePrice is explained by the two categorical predictors.

Next we'll consider the Type I and Type III Model ANOVA tables. In the Type I table, each effect is tested sequentially, and adjusts for all preceding listed effects. In other words, the order of the effects matters. The model specification determines the order in this table. The test of Heating_QC is an unadjusted test, because there are no other terms above it, whereas the Season_Sold test adjusts for the Heating_QC, which appears before it. The test for Season_Sold asks whether the season can explain the leftover variation in SalePrice after heating quality has explained as much of the sales price variation as possible. Typically, only Type III sums of squares tables are interpreted and reported for ANOVA. Type I sums of squares are more useful in say, polynomial regression models when we want to understand how high-order terms sequentially benefit the model.

Unlike Type I sums of squares, the Type III values are not generally additive, and the values do not necessarily sum to the model sums of squares. In the Type III table, all listed effects are adjusted for all other effects in the table, so order is not important. The Type III sums of squares for a variable, also called the partial sums of squares, is the increase in the model sum of squares due to adding the variable to a model which already contains all the other variables.

Judging from the p-values in the Type III sums of squares table, there seems to be no significant differences across levels of Season_Sold, with a p-value of 0.1768, but there are significant differences across the Heating_QC variable, with a p-value less than .0001. That is, even after you control for the effects of Season_Sold, the heating quality variable still explains significant differences in SalePrice.

The interaction plot for SalePrice differs from the exploratory plot because PROC GLM imposes a main effects model on the data given our model specification. In other words, the effect of each variable is not permitted to differ at different levels of the other variable. That constraint can be relaxed by adding an interaction term, as you'll see in the next demonstration.

Before we move on, this interaction plot illustrates an important point. In this plot, it seems that there is no interaction between heating quality and season sold. These results look completely reasonable, and we might be tempted to stop here. But we know, based on the earlier SGPLOT graphic, that this model is not adequate. We've already seen that the effect of season changes with the level of heating quality. So remember, its always a good idea to plot your data before you fit a model.

Recall that the LSMEANS statement requested a Tukey-adjusted analysis of the difference across all seasons. There are no significant differences in SalePrice means among the four levels of Season_Sold. The p-values range from 0.176 to 0.987. All are well above the typical significance threshold of 0.05.

Next, we could request comparisons of means for the different heating qualities, but the automatically generated interaction plot and the model that it represents don't match what we learned from our exploratory data analysis. The plot we created before generating the model showed that the effect of heating quality seems to change with the season. To allow for this in our model, we need to add and test an interaction effect. Let's take a moment to discuss interactions, and then come back and add an interaction effect to our model.

Close