Print

Demo: Performing a One-Way ANOVA Using PROC GLM

Filename: **st102d02.sas**

In this demonstration, we use PROC GLM to run an analysis of variance to test whether the average SalePrice differs among the houses with different heating qualities. Before we can trust the results from our ANOVA, such as the p-values and confidence intervals, we check the assumptions of our model using Levene's test of homogeneity of variances to assess constant variance. We check normality and independence through residual plots such as histograms, Q-Q plots, residual versus predicted values, and residual versus predictor plots.

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

1. Open program st102d02.sas.

```
/*st102d02.sas*/
ods graphics;

proc glm data=STAT1.ameshousing3 plots=diagnostics;
   class Heating_QC;
   model SalePrice=Heating_QC;
   means Heating_QC / hovtest=levene;
   format Heating_QC $Heating_QC.;
   title "One-Way ANOVA with Heating Quality as Predictor";
run;
quit;

title;
```

The PROC GLM statement specifies the ameshousing3 data set, and include the PLOTS=diagnostics option to produce a panel display of the diagnostic plots. The CLASS statement identifies the categorical predictor variable, Heating_QC. The MODEL statement identifies the dependent and independent variables as indicated in the ANOVA model, SalePrice=Heating_QC.

The MEANS statement computes the unadjusted means, or arithmetic means, of the dependent variable SalePrice for each level of the specified effect, Heating_QC. We also use the MEANS statement to test the assumption of equal variances by including the HOVTEST=levene option. This option performs Levene's test for homogeneity of variances by default. The null hypothesis is that all group variances are equal. If the resulting p-value of Levene's test is greater than some critical value, typically0.05, we fail to reject the null hypothesis and conclude that group variances are not statistically different.

The FORMAT statement is included to display descriptive labels for Heating_QC instead of the actual data values, TA, EX, Fa, and Gd. A QUIT statement is used because PROC GLM supports RUN-group processing, which means that the procedure stays active until SAS encounters a PROC, DATA, or

QUIT statement. RUN-group processing enables you to submit additional statements, followed by another RUN statement, without resubmitting the PROC statement.

2. Submit the program.

3. Review the output.

   In the GLM Procedure output, the Class Levels table specifies the number of levels and the formatted values of the class variable.

   The Number of Observations table indicates the number of observations that were read and the number used. These values are the same because there are no missing values for any variable in the model. If any row has missing data for a predictor or response variable, SAS drops that row from the analysis.

   In the Analysis of Variance output for SalePrice, the Overall ANOVA table includes all the information that's necessary to evaluate the null hypothesis. It reports the model and error sum of squares, the degrees of freedom, the mean squared errors, the F value, and the p-value. Because the p-value here, <.0001, is less than 0.05, you reject the null hypothesis of no difference between the means. Evidence suggests that at least one sale price mean is different for the four levels of heating quality.

   As discussed previously, the R-square value is often interpreted as the proportion of variance accounted for by the model. Therefore, you might say in this model, Heating_QC explains about 16% of the variability of SalePrice. The root MSE is simply the square root of the mean squared error from the ANOVA table above. Recall that its an estimate of the standard deviation for all treatment groups.

   The coefficient of variation is represented as a percent of the mean, so the root MSE divided by the SalePrice mean, all times 100. Its a unitless measure that's useful in comparing the variability of two sets of data with different units of measurement. The SalePrice mean is the overall mean ignoring the type of heating quality.

   The Type I sum of squares specify the sums of squares accounted for by adding effects into the model sequentially. However, for a one-way analysis of variance, only a single effect is included in the model. Therefore, the values in this table are an exact duplicate of the model line in the ANOVA table above.

   Before we can trust the p-value for our model, we need to assess the assumptions, so lets look at the diagnostic plots. The plot in the upper left panel shows the residuals plotted against the fitted values from the ANOVA model. Essentially, were looking for a random scatter within each group. Any patterns or trends in this plot can indicate model misspecification.

   To check the normality assumption, look at the residual histogram and Q-Q plot, which are at the bottom left and middle left, respectively. The histogram is approximately symmetric. The data values in the Q-Q plot stay close to the diagonal reference line and both plots give support to the assumption of normally distributed errors.

   The other plots can be used to further assess assumptions and also identify possible outliers. The default plot that was created with this code is a box plot.

   In the box plots, potential outliers are evident in all groups except for Fair, but the variability appear similar for all four levels of Heating_QC.

   Near the end of the tabular output, you can check the assumption of equal variances.

   The output in the Levene's Test for Homogeneity of SalePrice Variance table is the result of the HOVTEST option in the MEANS statement. The null hypothesis is that the variances are equal for all Heating_QC groups. The p-value of 0.6305 is not smaller than your alpha level of 0.05, and therefore, you do not reject the null hypothesis. Evidence suggests that the variances within each group of heating quality are not statistically different.

   Because we've met the model assumptions of independence, normal residuals, and constant variance, we can trust the results of our analysis and conclude that there are statistically significant differences

in SalePrice among houses with different heating qualities. At this point, we can conclude that among houses with Excellent, Good, Average/Typical, and Fair heating quality, at least one of these groups is different. Which group or groups are different? Well answer this question in the next section when we use ANOVA post hoc tests to conduct multiple comparisons.

Close