Print

Demo: Performing Simple Linear Regression Using PROC REG

Filename: **st102d05.sas**

When we performed exploratory data analysis, we found a significant Pearson correlation between SalePrice and several continuous variables in the ameshousing3 data set. Let's use PROC REG to build a simple linear regression model using Lot_Area as the predictor variable in order to determine how exactly Lot_Area and SalePrice are linearly related.

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

1. Open program st102d05.sas.

   

   ```
   /*st102d05.sas*/
   ods graphics;

   proc reg data=STAT1.ameshousing3;
       model SalePrice=Lot_Area;
       title "Simple Regression with Lot Area as Regressor";
   run;
   quit;

   title;
   ```

   The PROC REG statement specifies the data set ameshousing3, and the MODEL statement specifies the model that we're analyzing, SalePrice=Lot_Area.

2. Submit the program.

3. Review the output.

   In the REG procedure output, the Number of Observations table shows that the number of observations that were read and the number used are the same. This indicates there are no missing values for SalePrice and Lot_Area.

   Next, the Analysis of Variance table shows how the total variability in SalePrice can be partitioned to test the null hypothesis that the slope for Lot_Area is equal to 0.

   The ANOVA table in regression is equivalent to the ANOVA table from analysis of variance. It provides the model, error, and total sums of squares. It provides the degrees of freedom for each source of variability, and it also calculates the mean squares that are used to compute the F value. Recall that the mean squares are calculated as the sum of squares divided by their corresponding degrees of freedom, and dividing the mean square model by the mean square error computes the F value.

   In regression, the degrees of freedom are calculated as the number of parameters minus 1,

in this case, 2 - 1 = 1 for the model degrees of freedom, and the number of observations used minus the number of model parameters is 300 - 2 = 298. The total degrees of freedom are the same as before, n-1.

Finally, the ANOVA table reports the p-value to evaluate the null hypothesis, and in this case, it's highly significant. Thus, you can conclude that the simple linear regression model fits the data better than the baseline model. Evidence suggests that there's a significant linear relationship between SalePrice and Lot_Area, because the slope for Lot_Area is significantly different from zero.

The the third part of the PROC REG output, Fit Statistics table, displays summary measures of fit for the model. The root MSE is 36456 and is, of course, just the square root of the mean square error in the Analysis of Variance table. The root MSE is a measure of the error standard deviation. The dependent mean is 137525, which is the overall mean of SalePrice. The coefficient of variation is 26.50882. This is the size of the error standard deviation divided by the dependent mean. This statistic is used less often than the R-square and the adjusted R-square, and typically in specialized situations. The coefficient of determination is also referred to as the R-square value. Recall that it's the proportion of variability in the response variable explained by the regression model. In this example, the value is 0.0642, which means that Lot_Area explains 6% of the total variation in SalePrice.

The R-square is also just the squared value of the bivariate Pearson correlation coefficient that we saw in a previous demonstration between Lot_Area and SalePrice, 0.25335. The adjusted R-square is adjusted for the number of parameters in the model. This statistic is useful for comparing models with different numbers of predictors.

The Parameter Estimates table specifies the individual pieces of your model equation based on your data, whereas the Analysis of Variance table provides the overall fit for the model. The Parameter Estimates table also provides significance tests for each model parameter.

The parameter estimate for the intercept, $\widehat{\beta}_0$, is 113740, and the parameter estimate for the slope of Lot_Area, $\widehat{\beta}_1$, is 2.86770. So the regression equation is SalePrice=113740 2.86770*Lot_Area. The model indicates that each additional square foot of lot area is associated with an approximately $2.87 higher sale price.

The p-values for each parameter estimate tests the null hypothesis that the parameter estimate equals zero. Typically, we're not interested in the test of the intercept=0, because only the slope defines the nature of the linear association between the response and the predictor. The t test value is calculated by dividing the parameter estimate by the corresponding standard error estimate. Pr > |t| is the p-value associated with the test statistic. It tests whether the parameter is different from 0. For this example, the slope for the predictor variable is statistically different from 0.

Notice that, in simple linear regression, the t value for the slope is equivalent to the square root of the F value from the ANOVA table, and the p-values are identical. This will not be the case when more predictors are added to the model. Note that, extrapolation of the model beyond the range of your predictor variables is inappropriate. You can't assume that the relationship maintains in areas that were not sampled from.

The parameter estimates table also shows that the intercept parameter is not equal to 0. However, the test for the intercept parameter has practical significance only when the range of values for the predictor variable includes 0. In this example, the test could not have practical significance because SalePrice=0, or giving away a house for free, is not within the

range of observed values.

The diagnostics panel and the residuals by Lot_Area graph provide graphics to verify our model assumptions. For normality, the histogram of residuals looks bell-shaped and the dots on the Q-Q plot essentially fall on a straight line. Both indicate no deviations from normality. Non-constant variance can often be detected in residual plots when the residuals are close to zero and then expand to larger magnitudes.

The Fit PLot produced by ODS Graphics shows the predicted regression line superimposed over a scatter plot of the data.

To assess the level of precision around the mean estimates of SalePrice, you can produce confidence intervals around the means. This is represented in the shaded area in the plot. A 95% confidence interval for the mean states that you're 95% confident your interval contains the population mean of Y for a particular X. Confidence intervals become wider as you move away from the mean of the independent variable. This reflects the fact that your estimates become more variable as you move away from the means of X and Y.

Suppose that the mean SalePrice at a fixed value of Lot_Area is not the focus. If you're interested in making a prediction for a future single observation, you need a prediction interval. This is represented by the area between the broken lines in the plot. A 95% prediction interval is one that you are 95% confident contains a new observation if you were to actually sample another observation. Prediction intervals are wider than confidence intervals, because single observations have more variability than means.

Close