Print

Demo: Examining Residual Plots Using PROC REG

Filename: **st105d01.sas**

In this demonstration, we use PROC REG to create residual plots and other diagnostic plots. We use these plots to check our model assumptions and to check for outliers. First, to assess our model overall, we'll produce the eight default plots for fit diagnostics.

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

1. Open program st105d01.sas.

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
         Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/*st105d01.sas*/  /*Part A*/
ods graphics on;
proc reg data=STAT1.ameshousing3;
    CONTINUOUS: model SalePrice
                  = &interval;
    title 'SalePrice Model - Plots of Diagnostic Statistics';
run;
quit;


/*st105d01.sas*/  /*Part B*/
proc reg data=STAT1.ameshousing3
         plots(only)=(QQ RESIDUALBYPREDICTED RESIDUALS);
    CONTINUOUS: model SalePrice
                  = &interval;
    title 'SalePrice Model - Plots of Diagnostic Statistics';
run;
quit;
```

In Part A, the PROC REG statement specifies the data set ameshousing3, and the MODEL statement specifies SalePrice as the response variable and all the variables in the interval macro variable as the predictor variables. We've added an optional label of CONTINUOUS to the MODEL statement to label the output. Notice that the label must be followed by a colon.

2. Submit this step.

3. Review the output.

Scroll to the Diagnostic Plots. In the Diagnostic Panel, the first plot is the plot of residuals versus

predicted values. Looking at this plot, we're able to verify the equal variance assumption. We can also verify the independence assumption and check the adequacy of the model. Remember that we want to see a random scatter, with no patterns of our residuals above and below the 0 reference line. And the plot shows just that. We can conclude that the errors have constant variance. There's also no indication of correlated residuals, so we've met the independence assumption as well.

The Residuals versus Quantile plot is a normal quantile plot of the residuals. Using this plot, we can verify that the errors are normally distributed. The residuals follow the normal reference line pretty closely.

In the lower left corner, a histogram shows the normality of the residuals. Notice that a normal density curve is overlaid on the residual histogram to help detect departures from normality. Considering both the QQ plot and the histogram, we can conclude that the errors are normally distributed.

The plot of SalePrice versus Predicted Values of SalePrice shows data points spread along the 45-degree reference line, which indicates good model fit. There's a reasonably close match between the actual values and the predictions based on this model.

The last plot in the Diagnostic Panels is called a residual-fit or RF plot. It consists of side-by-side quantile plots of the centered fit and the residuals. The Fit (minus) Mean picture on the left shows the predicted or fitted values minus the overall mean. You can check to determine whether the vertical spread of the residuals in the plot on the right is greater than the spread of the centered fit in the plot on the left. The vertical spread of the residuals seems less than the vertical spread of the centered fit, so the model is fine. In other words, after accounting for the predictors in the model, relatively little residual variation remains.

The three remaining plots in this panel can be used to diagnose possible outliers. We'll discuss Rstudent residuals and Cook's D in the subsequent section.

The Residual Plots panel, the first panel includes the plots of the residuals versus the values of each of the interval predictor variables. They show no obvious trends or patterns in the residuals. Recall that independence of residual errors, or no trends, is an assumption for linear regression, as is constant variance across all levels of all predictor variables and across all levels of the predicted values. None of the variables contribute to possible violations in the assumptions.

Notice that when you visually inspect residual plots, the distinction of whether a pattern exists is a matter of discretion. If there's any question about the presence of a pattern, you should further investigate for possible causes of the pattern.

4. Let's go back to the code. We want to run PROC REG again, but request only specific plots. In Part B, we've added the PLOTS=ONLY option and requested the QQ plot to assess the normality of the residual error, RESIDUALBYPREDICTED to request a plot of residuals by predicted values, and RESIDUALS to request a panel of plots of residuals by the predictor variables in the model. The rest of the program is the same. This produces separate full-size plots for the QQ and the residual by predicted plots. If we wanted separate full-sized plots for the eight residual by predictor graphs, we could add DIAGNOSTICS(UNPACK) to the plot options.

5. Submit the PROC REG step in Part B.

6. Review the output.

   The Diagnostic Plots section contains the full-sized versions of the plots that we just saw. Full-size plots are easier to copy and paste into documents and presentations, if needed.

   Consider the Q-Q plot. If the residuals are normally distributed, the plot should appear to be a straight, diagonal line. This plot shows little deviation from the expected pattern. Thus, you can conclude that the

residuals do not significantly violate the normality assumption. If the residuals did violate the normality assumption, then a transformation of the response variable or a different model might be warranted.

Close