Practice 2.1 (Level 1): Performing Model Diagnostics on the Revised Model

Task

In Lesson 1, two models were created for the cars scenario: one with **Price** as the dependent variable (which was created in a demonstration) and a revised model with **LogPrice** as the dependent variable (which was created in practice 1.4). The revised model (with **LogPrice**) resulted in different predictor variables than the **Price** model. In this practice, you perform model diagnostics on that revised model. The data is stored in the **mydata.cars4** data set.

Note: Before you do this practice, you must run the code for practice 1.4 (in Lesson 1) in the same SAS session.

Reminder: Make sure you've defined the **mydata** library.

1. Use PROC GLMSELECT to fit the model with **LogPrice** as the dependent variable, and **Citympg**, **Citympg^2**, **EngineSize**, **Horsepower**, **Horsepower^2**, and **Weight** as the independent variables. Include the OUTDESIGN= option with ADDINPUTVARS to create a data set for performing the diagnostics in PROC REG.

   In PROC REG, use the appropriate options in the MODEL statement to evaluate multicollinearity, influential observations, and the constant variance assumption. To reference the independent variables, use the &_GLSMOD macro variable.

   Generate the necessary plots to check the assumptions of regression.

```
ods output ParameterEstimates=param;

proc glmselect data=mydata.cars4 outdesign(addinputvars)=d_carslog;
   effect p_City=polynomial(Citympg / degree=2
       standardize(method=moments)=center);
   effect p_hp=polynomial(Horsepower / degree=2
       standardize(method=moments)=center);
   model LogPrice = p_City EngineSize p_hp Weight / selection=none;
run;

proc reg data=d_carslog  plots (label)=all;
   model LogPrice = &_GLSMOD
     / vif collin collinoint influence spec partial;
   output out=check r=Residual p=Pred rstudent=rstudent h=leverage;
   id model;
run;
quit;

data check;
   set check;
   abserror=abs(residual);
run;

proc corr data=check spearman nosimple;
   var abserror pred;
run;
```

2. Is multicollinearity a problem for this model? Does this model meet the assumptions for linear regression?

   Examine the results.

   In the PROC REG results, note the following:

- In the Parameter Estimates table, the VIF for **Weight** is 14.39. This indicates moderate collinearity between **Weight** and one or more other variables in the model.
- The Collinearity Diagnostics table suggests that moderate multicollinearity might exist between **Weight** and the intercept. You might or might not want to remove the variable **Weight**, depending on the objectives of the study and other considerations.
- In the Collinearity Diagnostics (intercept adjusted) table, you can see that when the intercept is adjusted out of the model, there seems to be no apparent collinearity among the predictor variables.
- In the Test of First and Second Moment Specification table, the SPEC test indicates that there is not enough evidence to reject the null hypothesis that the model is correctly specified, that the errors are independent of the predictor variables, and that the variances are constant. However, based on the following warning about the SPEC test that appears in the log, you should use caution when interpreting the results: `WARNING: The average covariance matrix for the SPEC test has been deemed singular which violates an assumption of the test. Use caution when interpreting the results of the test.`

In the PROC CORR results, the Spearman correlation coefficient is lower for this model than for the original model with **Price** as the dependent variable (0.41820 versus 0.60274). However, the significant *p*-value for this test (0.0001) indicates that the variance is not stabilized for this model, which contradicts the results of the SPEC test. Remember that the results of the SPEC test are questionable because an assumption for the test was violated.

The Residual by Predicted plot looks better than the one for the model with **Price** as the outcome variable. However, the variance still appears to be smaller at the lower range of the predicted values than at the higher range. This corroborates the results of the Spearman correlation coefficient test.

Neither the histogram of the residuals nor the normal quantile plot indicates any problems with the normality assumption of the error terms. This model seems to meet the assumptions of normality and independence for linear regression, but does not appear to meet the assumption of constant variance.

3. To assess the model fit, examine the R-F plot of the observed values versus the predicted value, and the plots of residuals versus the independent variables. What are your conclusions?

Examine the results.

Because the spread of the residual plot is smaller than that of the fit-mean plot, the model explains most of the variation in the data.

The plot of the observed values versus the predicted values has a fairly tight distribution around the 45-degree reference line. This indicates a good fit. The fit for observation 24 (the Dodge Stealth) seems to improve in this graph.
However, in the following graphs of the residuals versus the predictors, the model does not seem to fit the Dodge Stealth well. Consequently, the potential improvement seen here might be due solely to the change in scale from **Price** to **LogPrice**.

The residual plots for **Citympg** and **Citympg^2** for this model look similar to the residual plots for **Hwympg** and **Hwympg^2** for the model with **Price** as the dependent variable. Also seen is the linear pattern for the three data points with (centered) values of **Citympg** above 10 mpg. The residual plot for **EngineSize** shows no apparent patterns. Only three observations have values of **Horsepower** above 100, so they are prominent in the residual plots of **Horsepower** and **Horsepower^2**. The residual plot for **Weight** exhibits no apparent pattern. The outlying point with the large negative residual in all the graphs is observation 24, the Dodge Stealth.

4. Examine the plots of the Cook's D, DFFITS, and DFBETA statistics, the plot of RSTUDENT versus LEVERAGE, and the partial leverage plots to identify any outlying or influential observations.

Examine the results.

The plots of Cook's D and DFFITS for the model with **LogPrice** as the independent variable flag more influential observations than the model with **Price** as the outcome variable. The Dodge Stealth is flagged as highly influential by both measures.

The RStudent versus Leverage plot for this model identifies observations that are influential due to RStudent and have high leverage. The Dodge Stealth and Mercedes-Benz 190E are still flagged as influential for this model. The Audi 100 and Lincoln Continental were influential due to RStudent for the original model, but not here. The SAAB 900 is flagged as influential for this model, but not for the original one. For this model of **LogPrice**, two more points are potentially more influential than for the model with **Price** as the dependent variable.

The Dodge Stealth exhibits influence on the parameter estimates for **Intercept**, **EngineSize**, **Horsepower^2**, and **Weight**. You can output these influential observations that were flagged by the influence statistics to a data set to further examine them.

Although these partial leverage plots for the model for **LogPrice** look better than the ones for **Price**, several points still appear to be influential or suffering from lack of fit.

5. Output the potentially influential observations to the new data set, **influence**. What are your conclusions?

```
%let numparms = 7;  /* # of predictor variables + 1 */
%let numobs = 81;   /* # of observations */
%let idvars = Manufacturer Model Price;
/* relevant identification variable(s) */

data influence;
   set check;
   absrstud=abs(rstudent);
   if absrstud ge 2 then output;
   else if leverage ge (2*&numparms /&numobs) then output;
run;

proc print data=influence;
   var &idvars;
run;
```

Based on the results, three more observations are either outlying or more influential for this model with **LogPrice** as the dependent variable than for the previous model for **Price**.

Hide Solution

Close