Print

Summary: Lesson 2: Regression Diagnostics and Remedial Measures

This summary contains topic summaries, syntax, and sample programs.

Topic Summaries

To go to the movie where you learned a task or concept, select a link.

Regression Model Diagnostics

For a linear regression analysis to be valid, four assumptions must be met: linearity in the parameters, normality of the errors, constant variance of the errors, and independence of the errors. If your model meets these assumptions, plots of the residuals versus the predicted values are a random scatter about a zero reference line, as expected, and you can trust subsequent results.

The independence assumption is that the error terms (ε) are independent of each other. That is, there is no correlation between any one observation in the data set and another. Knowing how your data is gathered helps you evaluate the assumption of independence.

Data collected over time might be serially correlated. Serial correlation in the errors (also known as autocorrelation) means correlation between consecutive errors or errors separated by some other number of periods. To evaluate if the error terms for time series data are correlated, you can plot residuals versus time or another ordering component to examine whether there seems to be any positive or negative autocorrelations.

You can use the SGPLOT procedure, or the ODS graphics output from the REG and UNIVARIATE procedures, to examine a histogram of the residuals with a normal curve overlaid, a normal probability plot of the residuals, or a normal quantile plot of the residuals. You can also use PROC UNIVARIATE to produce the formal tests of normality.

PROC SGPLOT;

PROC REG;

PROC UNIVARIATE;

To evaluate the assumption of constant variance of the error terms, also known as homoscedasticity, you can use the REG procedure and ODS Graphics to produce plots of the residuals versus predicted values and plots of the residuals versus independent variables.

You can also perform a test for heteroscedasticity by using the SPEC= option in the MODEL statement in PROC REG. However, the SPEC test is a multiple null hypothesis test, so rejection of the null hypothesis in this case may not be due solely to heteroscedasticity.

MODEL *dependent=model-effects* **SPEC;**

In addition, you can compute the Spearman rank correlation coefficient, which measures the correlation between the size of the ordered predicted values and the absolute value of their associated residuals. (Carroll and Ruppert 1988) The Spearman rank correlation coefficient is available as an option in PROC CORR.

PROC CORR;

To evaluate model fit, you can assess the diagnostic plots created by ODS Graphics in the output of PROC

REG. For models that are fit with PROC GLMSELECT, you can use the OUTDESIGN= option to create an input data set for PROC REG. These plots include:

- plots of residuals and studentized residuals versus predicted values
- "residual-fit spread" (or R-F) plots
- plots of the observed values versus the predicted values
- partial regression leverage plots

PROC GLMSELECT DATA= *SAS-data-set* **OUTDESIGN=;**

You can use plots of residuals versus the predicted values to visually assess the goodness of fit of the model. The Residual-Fit Spread Plot provides a visual summary of the amount of variability accounted for by a model. The plot consists of two panels. The left panel shows the quantile plot of the predicted values minus their mean. The right panel is a quantile plot of the residuals.

Using centered data (fit minus mean value) instead of raw values enables you to compare the spread of variables that have different means. Plots of the observed values versus the predicted values also provide a visual tool for examining how close the fitted (or observed) values are to the predicted values.

A partial regression leverage plot is the plot of the residuals for the response variable against the residuals for a selected predictor. The slope of the linear regression line in the partial regression leverage plot is the regression coefficient for that predictor variable in the full model. (Sall 1990) This plot helps you evaluate whether you have specified the relationship between the response and the predictor variables correctly.

In addition to assessing diagnostic plots in order to evaluate the fit of your model, you can examine model-fitting statistics. These statistics include R-square, Adjusted R-square (the higher the value, the better), Akaike's Information Criterion (AIC) or Corrected Akaike's Information criterion (AICC), Schwarz's Bayesian criterion (the smaller the value, the better), and Mallows' C_p. The model with a C_p value less than or equal to *p*, the number of parameters including the intercept, will better fit the data.

f your data includes multiple observations (or replicates) for each value of the combination of the independent variables, then you can use the LACKFIT option in the MODEL statement in PROC REG to perform a lack-of-fit test for the regression model.

MODEL *dependent=model-effects* **LACKFIT;**

To evaluate multicollinearity, you can use the CORR procedure to compute the correlation statistics that measure the linear relationship between pairs of the independent variables. Variance inflation factors can help determine the presence of multicollinearity. You use the VIF option in the MODEL statement in PROC REG to calculate these factors. You use the COLLIN and COLLINOINT options in the MODEL statement in PROC REG to identify the sets of variables involved in multicollinearity.

MODEL *dependent=model-effects* **VIF;**

MODEL *dependent=model-effects* **COLLIN COLLINOINT;**

An outlier is a data point that differs more than expected from the general trend of the data. Outliers have residuals that are considerably larger in absolute value than the residuals of other data points, such as two or three standard deviations from the mean. An influential observation is an observation that is so far away from the rest of the data that it singlehandedly exerts influence on the slope of the regression line.

You want the regression model to be representative of all of the sample observations, not an artifact of a few. Consequently, it is important to identify these influential data points and assess their impact on the model. Keep in mind that an outlier might or might not be an influential observation, and vice versa.

Influential observations might be difficult to detect from simple scatter plots in a multiple regression setting. Several statistics are designed to assist in identifying influential observations.

The leverage statistic measures how far an observation is from the cloud of observed data points.

Cook's distance, or the Cook's D statistic, is the most common measure of the influence of an observation. The Cook's D statistic measures the distance between the set of parameter estimates with that observation deleted from your regression analysis and the set of parameter estimates with all the observations in your regression analysis.

DFFITS measures the impact that each observation has on its own predicted value. There are two versions of the rule of thumb for DFFITS. The general cutoff value is 2. The more precise cutoff is 2 times the square root of p divided by n, where *p* is the number of terms in the model, including the intercept, and *n* is the sample size.

The DFBETAS statistic not only helps you to identify an influential observation, it also tells you which predictor variable is being influenced. DFBETAS, which stands for difference in betas, measure the change in each parameter estimate when an observation is deleted from the analysis.

The Covariance Ratio measures the change in the precision of the parameter estimates when an observation is deleted from the model. It is calculated as the ratio of the determinant of the covariance matrix with the i^{th} observation deleted, to the determinant of the covariance matrix with all the observations included.

You can use ODS Graphics in PROC REG to create plots of values of RSTUDENT, LEVERAGE, Cook's D, DFFITS, DFBETAS, and COVRATIO, as well as plots of the studentized residuals plotted against the predicted values and against the leverage statistics.

Remedial Measures

Let's summarize the steps involved in building a regression model. The first step is to perform a preliminary analysis. Next, you check for multicollinearity among the variables that you identified in step 1 by using the VIF statistic, condition indices, and variation proportions.

In the third step, you use the information gathered in steps 1 and 2, together with model selection options in PROC GLMSELECT to identify one or more candidate models. Fourth, you need to check and validate your assumptions by examining plots of the residuals versus predicted values and performing other statistical tests, including tests for normality of the residuals, constant variance, and independent observations.

If the testing you perform in step 4 indicates a problem, then in step 5 you revise your model and generate a new model. The final step is to evaluate the model's predictive capability with data that was not used to build the model.

You know that when the normality assumption is violated, it affects the test results, for example, tests of significance and confidence intervals of the parameter estimates. One way to remediate this problem is to transform the dependent variable to normalize the distribution.

You can fit a generalized linear model using the GENMOD or GLIMMIX procedures with the distribution and link options.

If your data violates the constant variance assumption, you can request tests using both the usual covariance matrix and the heteroscedasticity-consistent covariance matrix. In the MODEL statement of PROC REG, you specify the ACOV, HCC, or WHITE option.

MODEL *dependent=model-effects* **ACOV HCC WHITE;**

You can also transform the dependent variable to stabilize the variance, or use different procedures to model the nonconstant variance. These procedures include:

- PROC GENMOD or PROC GLIMMIX with the appropriate distribution function, using the DIST= option
- PROC MIXED with the GROUP= option, which allows you to define an effect specifying heterogeneity in the covariance structure, the TYPE= option, or the power-of-mean models, and
- PROC SURVEYREG for survey data.

You can also use a weighted least squares regression model when variances are not constant.

You know that when the independence assumption is violated, it can affect the standard errors of the parameter estimates, confidence intervals, and significance tests for the parameters. You can use several modeling tools to account for correlated observations. If the data to be analyzed is time series data, you use SAS/ETS

procedures like PROC AUTOREG or PROC ARIMA.

PROC AUTOREG;

PROC ARIMA;

To model correlations arising with data that include repeated measures from each subject, you can use PROC MIXED, PROC GENMOD, or PROC GLIMMIX. To model the data gathered from a complex survey design, you should use PROC SURVEYREG.

When a plot of the residuals versus the predicted values exhibits a discernable pattern, you know that this indicates misspecification of the model.

When the relationship between the dependent variable and one or more predictor variables does not follow a linear relationship, you might consider transforming the predictor variables to obtain the linearity. (Neter, Wasserman, and Kutner 1990)

When the parametric form of the relationship is difficult or impossible to define, you might want to fit a local regression model using PROC LOESS. The idea of local regression is that near any chosen value of X, the regression function can be well approximated by low-degree polynomials.

PROC LOESS;

You know that multicollinearity is often caused by the choice of model, such as when two highly correlated predictor variables are used in the regression equation. In these situations, you can lessen the impact of multicollinearity by respecifying the regression equation.

One approach to model respecification is to redefine the predictor variables. Another widely used approach to model respecification is to eliminate redundant predictor variables. Variable elimination is often a highly effective technique.

You know that ordinary least squares estimators provide unbiased estimates of parameters and that the estimates have minimum variance among all unbiased estimators. But in the presence of multicollinearity, the minimum variance of the parameter estimates might be unacceptably large.

You can use biased regression techniques such as ridge regression and principal component regression to obtain biased estimators of regression coefficients.

In ridge regression, you reduce the variances of the parameter estimates by considering a matrix, X'X + *k*I, where *k* is a small positive quantity referred to as a shrinkage parameter. To perform ridge regression, you use the RIDGE= option in the MODEL statement of PROC REG to specify the range of values of *k*. To obtain the ridge traces, you use the RIDGEPLOT option in the PLOT statement.

MODEL *dependent=model-effects* **RIDGE=;**

PLOT <*yvariable*xvariable*> <=*symbol*> <*...yvariable*xvariable*> <=*symbol*> **RIDGEPLOT=;**

Principal component regression is another biased regression technique. This approach combats multicollinearity by using less than the full set of principal components in the model. Instead of dropping individual variables from the model, you drop linear combinations of independent variables.

To compute parameter estimates using all but the last *m* principal components, you use the PCOMIT= option in the MODEL statement in PROC REG.

MODEL *dependent=model-effects* **PCOMIT=;**

In polynomial regression models, you can sometimes overcome the effects of multicollinearity by centering the independent variables. Finally, you can limit the influence of outliers by performing robust regression analysis using PROC ROBUSTREG. The main purpose of robust regression is to detect outliers and provide resistant, that is, stable, results in the presence of outliers.

PROC ROBUSTREG;

When your data exhibits nonconstant variance, you model the response using a probability distribution that accommodates that nonconstant variance (heteroscedasticity).

One such distribution, which is used frequently to analyze cost or price data, is the lognormal distribution. A variable is said to follow the lognormal distribution when its logarithm follows a normal distribution.

You can use a generalized linear model to enable the distribution of the response variable to be something other than the normal distribution. In SAS, you can fit these models using PROC GENMOD or PROC GLIMMIX.

The lognormal distribution is available only in PROC GLIMMIX. The GLIMMIX procedure fits statistical models to data with correlations or nonconstant variance, as well as data in which the response variable is not necessarily normally distributed.

In the PROC GLIMMIX statement, you specify the data set to be modelled and various options. In the EFFECT statement, you construct new effects for the model using predictor variables in the input data set. In the MODEL statement, you specify dependent variable and fixed effects. You use DIST=option to specify the built in probability distribution of the data. The output data set contains predicted values and residual diagnostics, computed after fitting the model.

PROC GLIMMIX DATA=*SAS-data-set <options>***;**
  EFFECT *effect-name=effect-type<(effect-options)>***;**
  MODEL *response<(response-options)>=<fixed-effects>*
    < DIST=*keyword options***>;**
  OUTPUT OUT= *SAS-data-set keyword=name(s)***;**
RUN;

To fit a lognormal distribution, PROC GLIMMIX applies a log transformation to the response variable and models that transformed response using a normal distribution.

To make interpretation easier, statisticians usually prefer to obtain predicted means on the original scale of the data. The formula for the mean of a lognormal distribution becomes:

$$E[Y] \approx \exp\left(X\hat{\beta} + \frac{\hat{\sigma}^2}{2}\right)$$

Sample Programs

Performing Model Diagnostics

```
ods graphics / imagemap=on;

proc glmselect data=mydata.cars outdesign(addinputvars)=d_carfinal;
   effect q_hwympg = polynomial(hwympg / degree=2
      standardize(method=moments)=center);
   model price = q_hwympg horsepower / selection=none;
run;

proc reg data=d_carfinal plots(unpack label)=all;
```

```
   model price = &_GLSMOD / vif collin collinoint influence spec partial;
   id model;
   output out=check r=residual p=pred rstudent=rstudent h=leverage;
run;
quit;

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

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

%let numparms = 4;
%let numobs = 81;
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 manufacturer model price hwympg horsepower;
run;

proc print data=influence;
   var manufacturer model price hwympg horsepower;
run;
```

Fitting a Lognormal Regression Model

```
title 'Lognormal model for CARS data set';

ods output ParameterEstimates=params;
proc glimmix data=mydata.cars;
   effect q_hwympg = polynomial(hwympg / degree=2
          standardize(method=moments)=center);
   model price = q_hwympg horsepower / dist=lognormal solution;
   output out=out pred=pred resid=resid;
   id model price;
run;

data check3;
   set out;
   abserror=abs(resid);
run;

proc corr data=check3 spearman nosimple;
   var abserror pred;
run;

data _null_;
   set params;
   if Effect='Scale' then call symput('var',Estimate);
run;

data back;
   set check3;
   Estimate = exp(pred + &var/2);
   Difference = Price-Estimate;
run;
```

```
proc sgplot data=back;
   scatter x=Estimate y=Difference / datalabel=model;
   xaxis min=0 max=60;
   yaxis min=-30 max=30;
   refline 0;
run;

proc reg data=d_carfinal;
   model price = &_GLSMOD / hcc hccmethod=3;
run;
quit;
```


Close