Fitting a Gamma Regression Model

Recall the **cars2** data set that you analyzed earlier in this course. The plot of the Residuals versus Predicted Values shown here appeared in model diagnostics in one of the linear regression models generated by PROC REG. We modeled **Price** as the response variable, and **HWYMPG**, **HWYMPG^2** and **Horsepower** as the predictor variables. We suspected heterogeneity of variance, since there's a fanning to the right. The spread on this plot becomes larger as the prices of the cars become greater. This violates the assumption of constant variance. In lesson 2, we discussed remediation methods to deal with this problem. The solutions considered included transformation, choice of a different distribution, and the use of different procedures to model the nonconstant variance. In that lesson we fit a model using PROC GLIMMIX and a lognormal distribution. Now we're going to try the gamma distribution.

Let's begin by using PROC UNIVARIATE to examine whether the gamma distribution describes the distribution of **Price** adequately.

We're using an ODS SELECT statement to specify which output we'd like to see, in this case, the histogram, a table of parameter estimates, and the tests of Goodness of Fit.

In the HISTOGRAM statement, we specify the gamma distribution, along with options to control its appearance. The three parameters, alpha, sigma and theta will be estimated from the data. The line will be colored blue, with a width of 2, and we've also specified details regarding the axes and midpoints.

```
ods select Histogram ParameterEstimates GoodnessofFit;
proc univariate data=mydata.cars2;
   var price;
   histogram /gamma(alpha=est sigma=est theta=est color=blue w=2)
                    vaxis=0 to 14 by 2 midpoints=8 to 50 by 2;
   title 'Testing Gamma Distributions';
run;
```

Let's run the code and look at the histogram. It looks like the gamma distribution fits the response variable **Price** quite well. The table of Goodness of Fit tests shows the statistics for the gamma distribution. The nonsignificant *p*-values for the tests for the gamma distribution indicate that you do not have enough evidence to reject the null hypothesis that the variable **Price** follows a gamma distribution. Thus, the histogram looks consistent with the test.

Now we'll use PROC GLIMMIX to fit a gamma regression model to the **cars2** data set. We use the PLOTS= option to turn on some diagnostics.

We want to be able to compare the results of this model with those that were fit using the normal distribution. As in the previous analyses, our response variable is **Price**. The predictor variables are **HWYMPG**, **HWYMPG^2** and **HORSEPOWER**. For the distribution, we specify DIST= gamma.

Although the canonical link function for the gamma distribution is the inverse link, we'll use the log link. Log is the most commonly used link function for gamma-distributed variables. We turn on the SOLUTION option to request a table of parameter estimates.

To identify the outlying observations, we use the OUTPUT statement to create a data set, **check1**, that will contain the studentized residuals and predicted values, on both the original scale of the data and on the log scale. The option ILINK tells SAS to apply the inverse link, that is, to produce statistics on the original scale of the data. The NOILINK option leaves the statistics on the log scale, that is, on the linked scale.

```
proc glimmix data=mydata.cars2 plots=studentpanel(unpack);
   model price=hwympg hwympg2 horsepower / dist=gamma link=log solution;
   id model hwympg hwympg2 horsepower price;
   output out=check1 student
        pred(ilink)=   Pred stderr(ilink)=Stderr lcl(ilink)=LCL
                       ucl(ilink)=UCL
        pred(noilink)= XB stderr(noilink)=StderrXB lcl(noilink)=LCLXB
                       ucl(noilink)=UCLXB;
   title 'Cars Data Set - Gamma Distribution with Log Link';
run;
```

Let's run the code and compare the Fit Statistics and Parameter Estimates to those from other models. The AIC statistic for this model is 457.25 and the BIC is 469.22.

In the Type III Tests of Fixed Effects table, the *p*-values are very similar to those calculated by the linear model using the normal distribution. All the parameter estimates are significant at the alpha level of 0.05. The equation for this model is as follows:

- log(E(**Price**)) = 2.1190 – 0.0433 * **Hwympg** + 0.0016 * **Hwympg**^{2} + 0.0050 * **Horsepower** or
- E(**Price**) = $e^{2.1190 – 0.0433 * \textbf{Hwympg} + 0.0016 * \textbf{Hwympg}^2 + 0.0050 * \textbf{Horsepower}}$

Let's also review the plots of the residuals versus the predicted values. By using the gamma distribution we wanted to correct the heterogeneity of variance issue. With the exception of two data points, this plot does display a more random scatter around the reference line.

Remember that our program output the studentized residuals to a data set, **check1**. We can run PROC PRINT to view the contents of **check1**.

```
proc print data=check1 (obs=5);
   var Model Hwympg Hwympg2 Horsepower Price pred stderr lcl ucl xb
       stderrxb lclxb uclxb student;
title2 'Predicted Values';
run;
```

Let's review the results, in particular the studentized residuals and linear predictors. Note the presence of several possible outliers in the plot of studentized residuals versus predictors.

Now we can use the PRINT procedure to look for outliers or influential observations in the **check1** data set. We use a WHERE statement to specify a rule to include any studentized residual greater than or equal to 2, as well as those less than or equal to negative 2.

```
proc print data=check1;
   where student ge 2 | student le -2;
   var model student price pred hwympg horsepower;
title2 'Outlying Student Residuals';
run;
```

Let's run the program. The results list four data points that have extreme residuals, only one of which is extremely large (the residual for the 190E). This indicates that the model does not fit these four cars well.

Now we'll fit the gamma regression model using PROC GLIMMIX again, but this time we'll use the identity link. The syntax is essentially the same. We're turning on the plots. The model statement is the same except for the link function. We've changed the link function. so that we are modelling directly on the original scale of the data. This time we're outputting statistics to the **check2** data set - the studentized residuals and predicted values on the original scale of the data.

```
proc glimmix data=mydata.cars2 plots=studentpanel (unpack);
   model price = hwympg hwympg2 horsepower / dist=gamma link=id solution;
   id model hwympg hwympg2 horsepower price;
   output out=check2 student=Student pred(ilink)=Pred;
   title 'Cars Data Set - Gamma Distribution with Identity Link';
run;
```

Let's run the code and review the Fit Statistics. An indication that the model seems to fit the data better than the previous model is that the AIC and BIC statistics are smaller for this model (447.19 versus 457.25 for AIC and 459.17 versus 469.24 for BIC).

We'll also look at the parameter estimates. Remember that these are on the original scale of the data. Therefore, the estimated equation for this model is

- E(**Price**) = 1.5568 □ 0.5171 * **Hwympg** + 0.0294 * **Hwympg**^{2} + 0.1177 * **Horsepower**

Compared with the OLS regression model that you fit in the previous chapter,

- E(**Price**) = 4.0395 □ 0.8041 * **Hwympg** + 0.0435 * **Hwympg**^{2} + 0.0973 * **Horsepower**

the gamma regression model with an identity link provides different parameter estimates. However, this model accounts for heteroscedasticity.

Let's look again at the plot of Studentized Residuals by Predicted Values. It certainly looks better than our original plot. The variance seems to be more stable. Also, the model has both positive and negative residuals in the high price ranges, whereas the gamma model with log link showed only negative residuals in the high price ranges.

Comparing these two models, the model with the identity link does fit a little better. Another advantage is that the parameter estimates and predicted values are already on the original scale of the data.

Let's run a PROC PRINT program again to review the contents of the **check2** data set.

```
proc print data=check2;
   where student ge 2 | student le -2;
   var model student price pred hwympg horsepower;
title2 'Outlying Student Residuals';
run;
title1;   title2;
```

The results show that the model has only three potential outliers, indicating that this model (the gamma model with identity link) might fit the data better than the gamma model with log link.

Notice that we are still using the variables **HWYMPG**, **HWYMPG^2**, and **HORESEPOWER** in our model. These predictors were selected using the variable selection methods within the OLS regression model. You should complete the entire modeling cycle to choose the best candidate model for the gamma regression with either the log or identity link.

Close