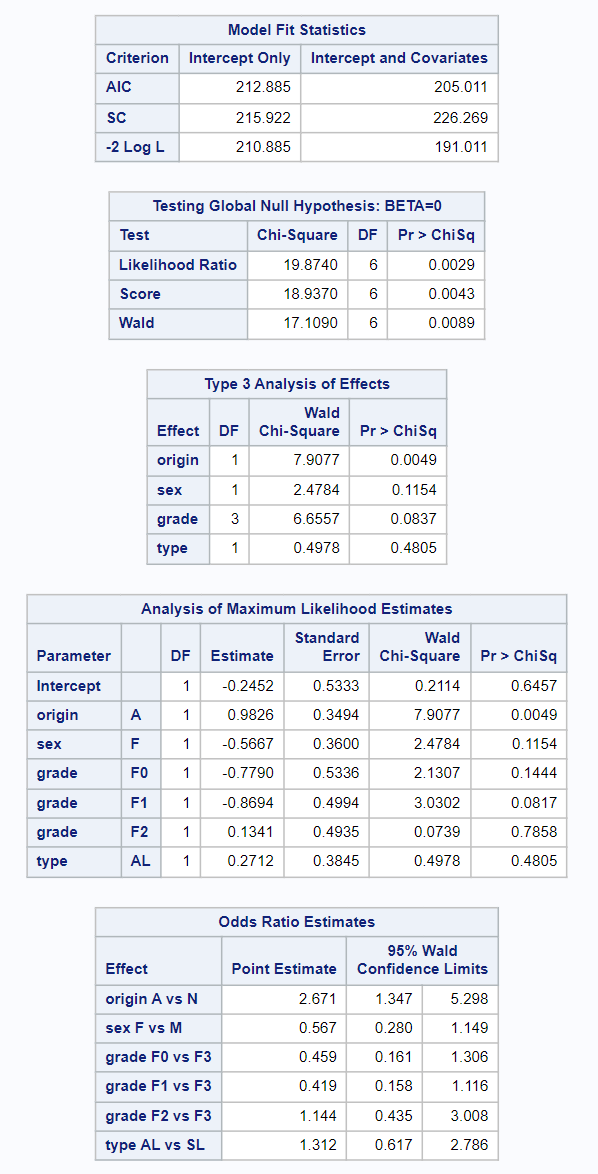
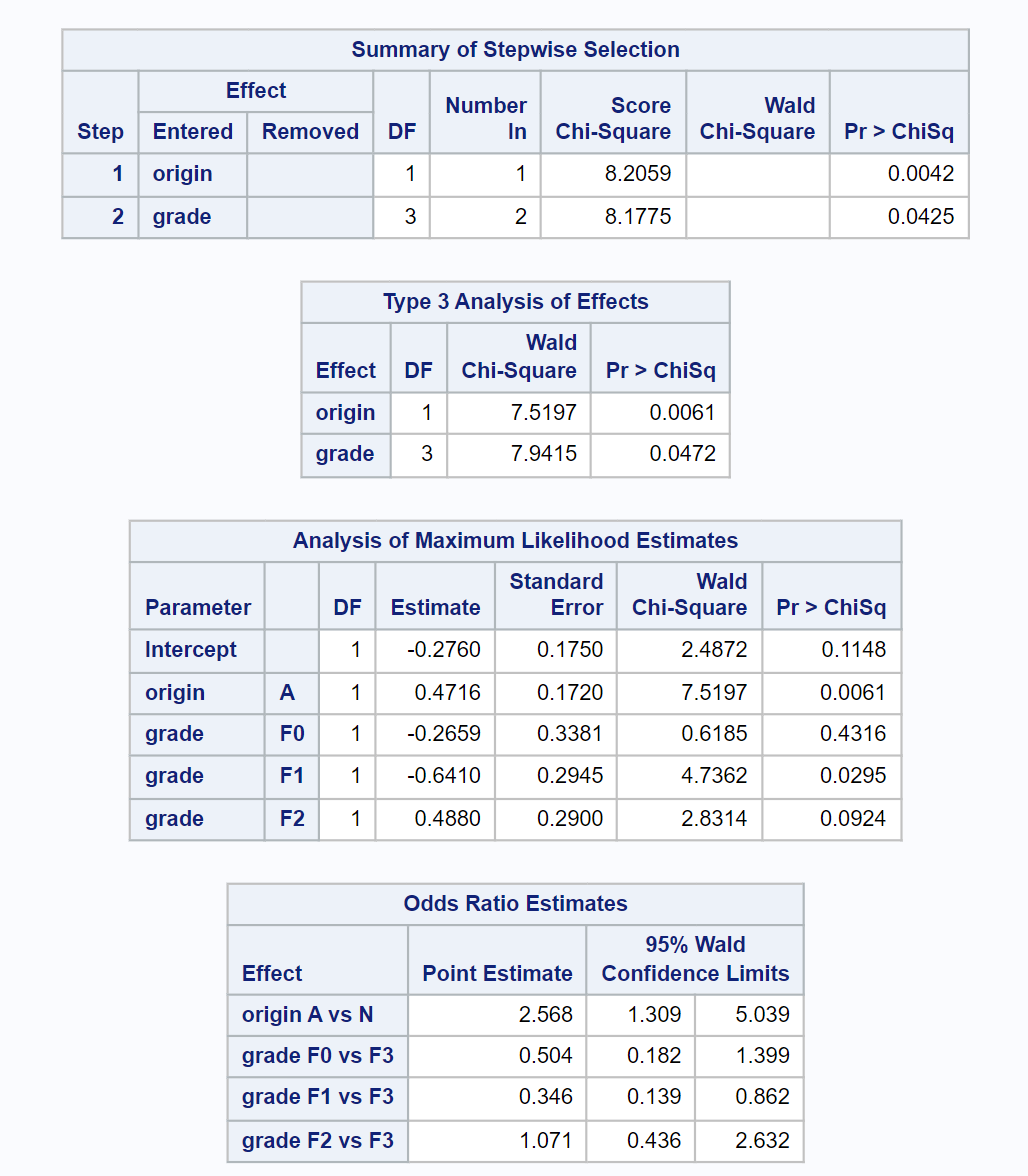
HW 5

## Problem 1

### a.)

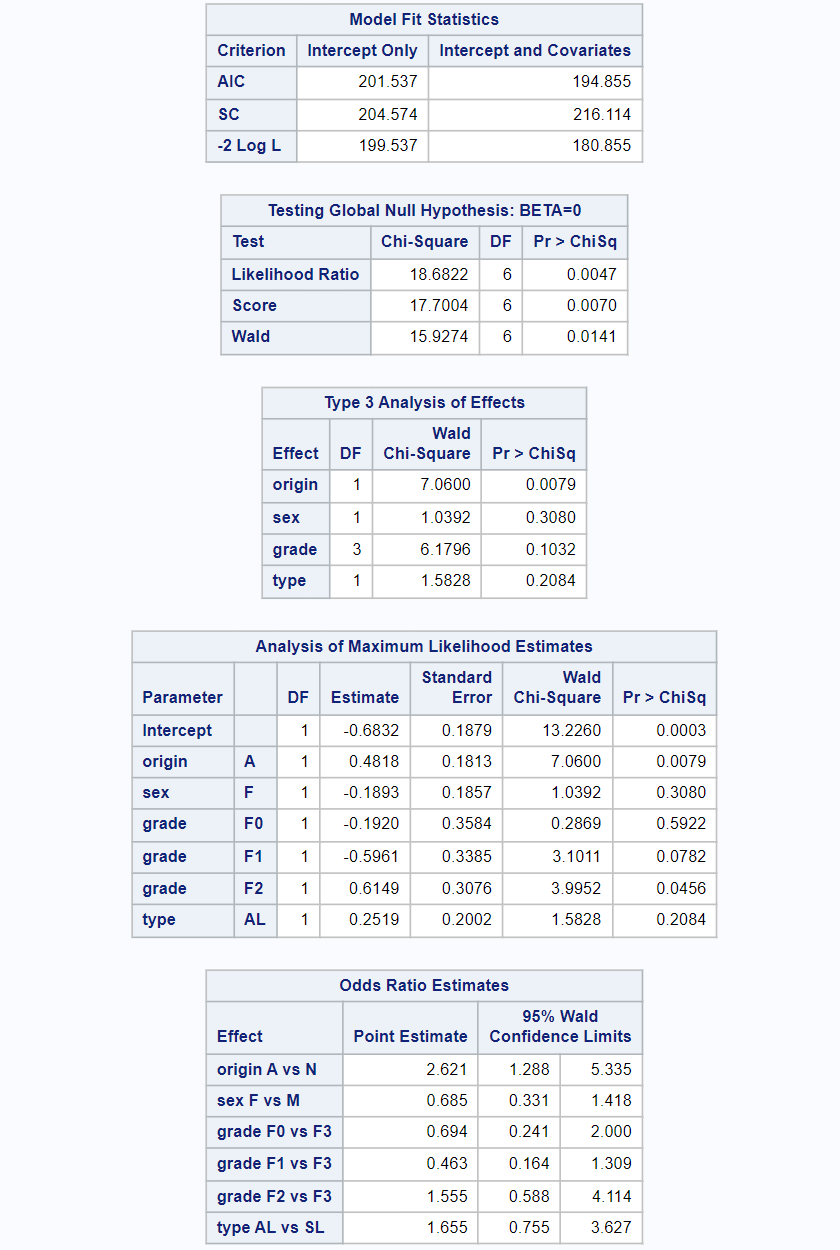
  
We see from the output of the logistic regression procedure that at the 0.05 or 5% significance level the entire model itself is significant when predicting whether a student has more than 14 days of absence, we see this from the model fit statistics and Testing the global hypothesis. However, when looking at the individual predictor variables, the only variable that appears significant at the 5% level is the Origin variable. A potential problem with this model is that most of the predictors and the intercept are not significant in predicting the response. Also note that the SC increases in the full model compared to the AIC with decreases, this is due to the SC having a more substantial penalty term for the number of predictors.

### b.)

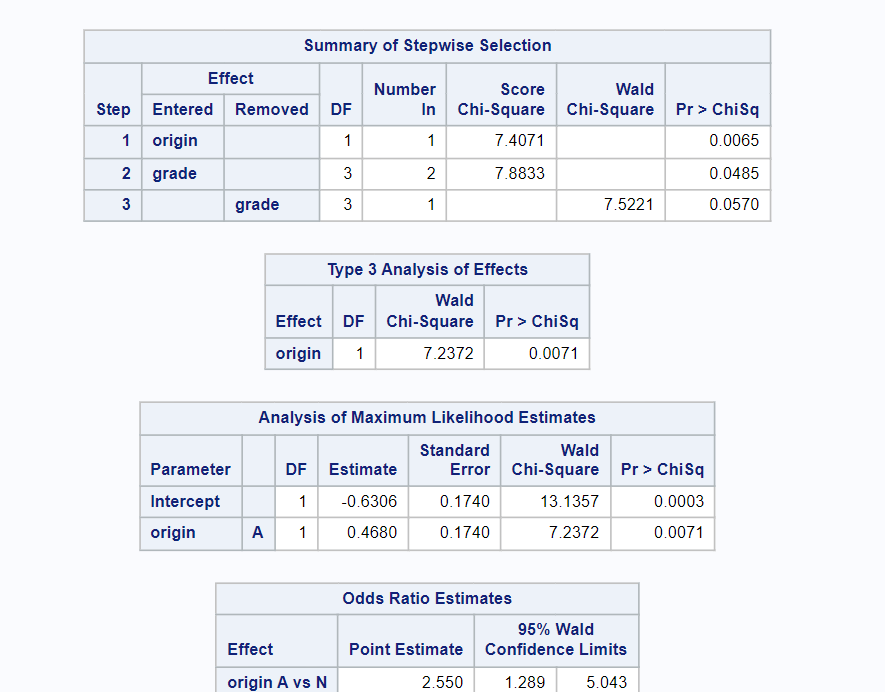
  
When using stepwise selection, we see that at the default significance level of 0.05 we get a model with two predictors, Origin and Grade. This is interesting because previously in the full model, grade was not a significant predictor at the 0.05 significance level. So this model tells us that the significant predictors are whether a student is aboriginal or not and the grade of the student is significant in predicting the odds of having more than 14 absent days. According to our model, while holding all other variables constant, it is estimated that students that are Aboriginal have about 2.568 times the odds of having greater than 14 absent days when compared to students who are not Aboriginal. The for the grade variable, while holding all other variables constant, students in grade F0 have an estimated 50.4% the odds having more than 14 absent days when compared to students in grade F3. Students in grade F1 have an estimated 34.6% the odds of greater than 14 absences when compared to F3 students. Finally F2 students have 1.071 times the odds of greater than 14 absences when compared to F3 students. Note that this model treats the grade F3 as the reference level.

## Problem 2

### a.)

  
We see from the output that under the 0.05 significance level, the entire model is significant in predicting the odds for students having greater than 16 days of absence. When looking at the individual predictors, we see that the only predictor that seems significant is origin at the 0.05 significance level. Additionally, the intercept coefficient term is also significant in this model, which was not the case in the model for greater14. Its also interesting to note that the grade level F2 has a just barely significant coefficient at the 0.05 significance level. Again, an issue with this model is that we have a bunch of predictors that are not significant.

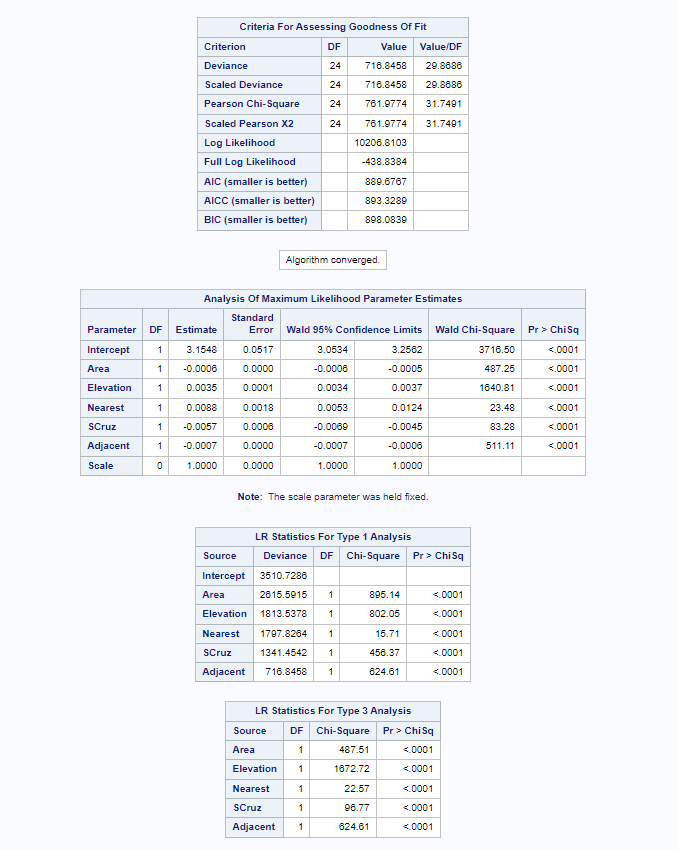
### b.)

  
After running stepwise selection with the default significance level, origin was first added to the model, then grade was added, but in the very next step grade was removed from the model. We see that for greater than 16 days of absence, the predictor origin is significant when looking at the odds of a student having more than 16 days of absence. The intercept is also significant in this model. Looking at the odds ratio estimates, we see the students that are Aboriginal have a 2.550 times odds of having more than 16 days of absence when compared to students who are not Aboriginal.

The difference between this model and the greater14 model is that we have a significant intercept term and we don't include the grade predictor. The greater14 model also estimated a slightly higher odds ratio between Aboriginal and non Aboriginal students for the odds of being absent for an amount of days.

## Problem 3

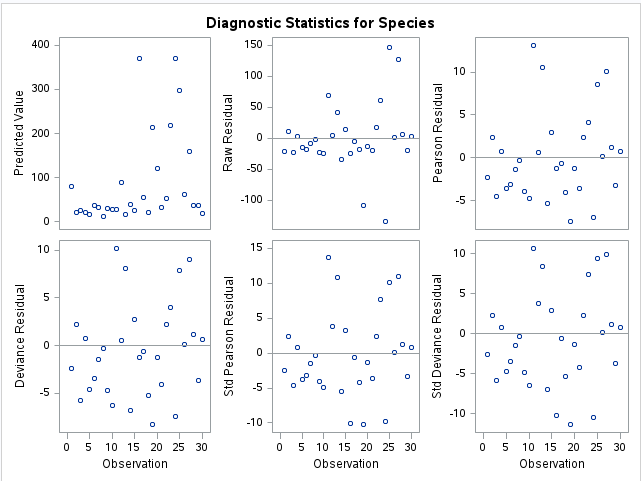
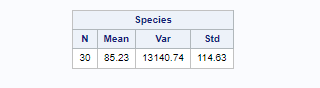
### a.)

  
I started with including all the possible predictors in the model & then ran the GEN mod procedure. Since all the terms were significant, I decided not to do any more variable selection and keep the full model with all the terms.

### b.)

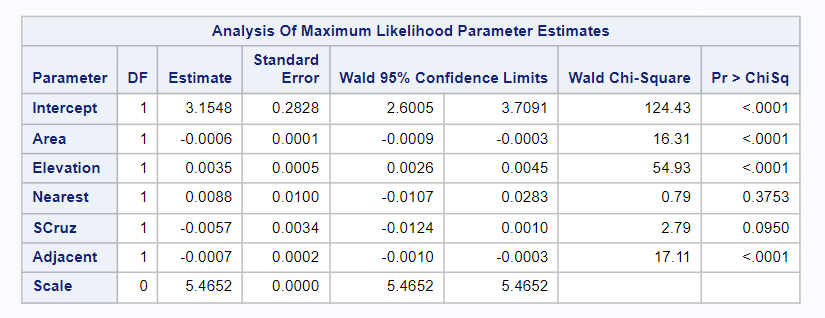
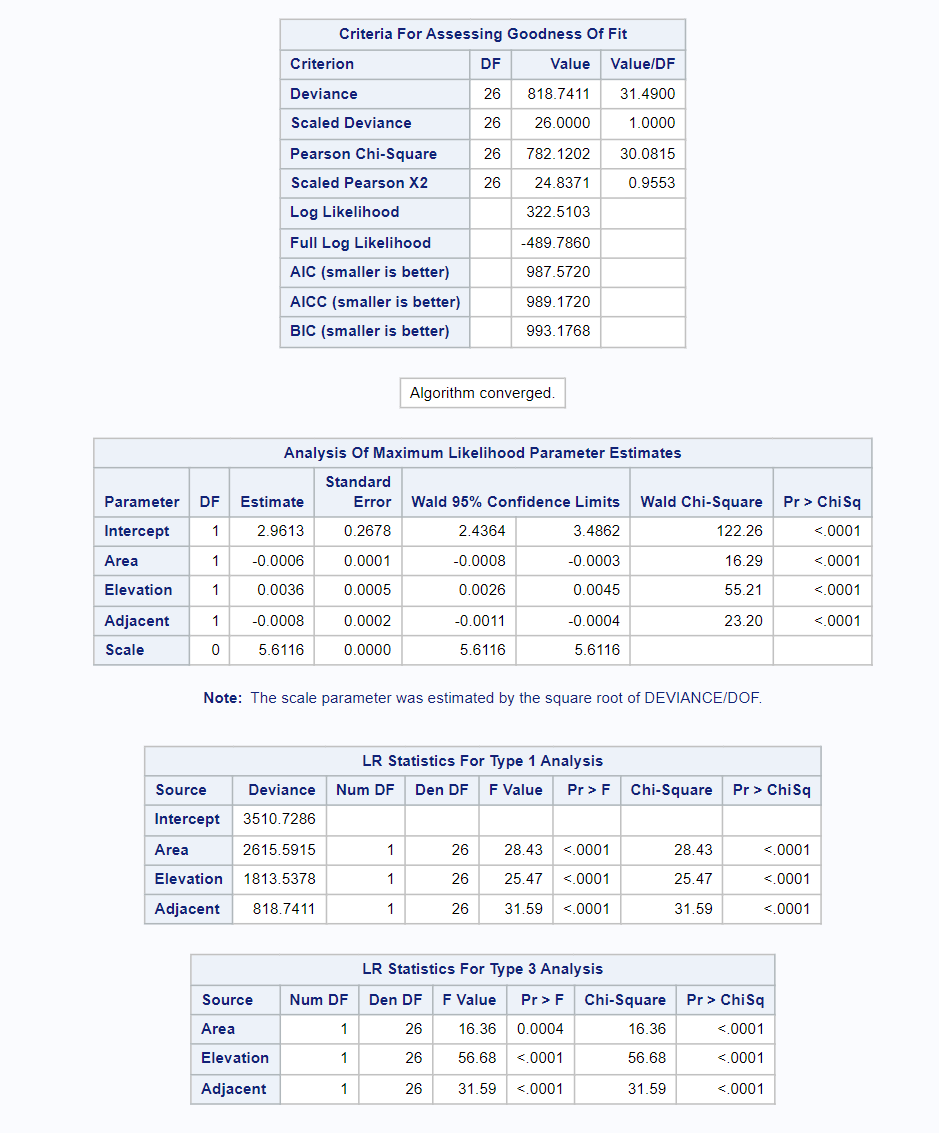
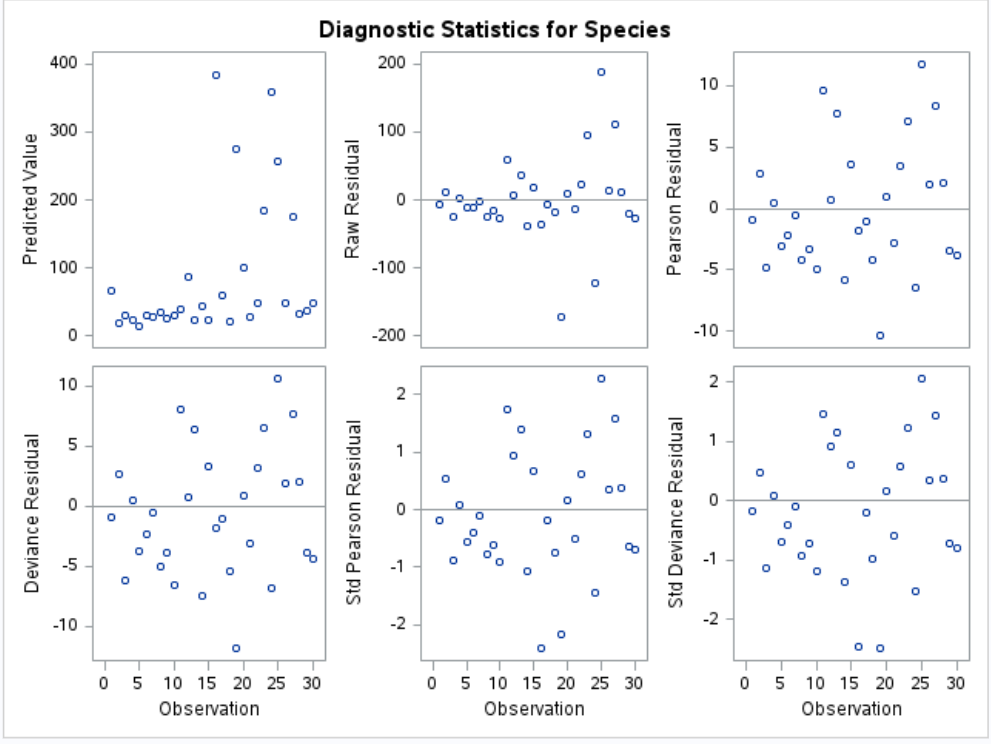
From this model we see that Area, Elevation, Nearest, SCruz, and Adjacent are all significant. Since we are using the log link, we have to interpret the model with that in consideration. So for Elevation and Nearest, as these predictors increase by a unit, the log of the number of species also increases by 0.0035 and 0.0088 respectively. While for Area, SCruz, and Adjacent, as these predictors increase by a unit, the log number of species decrease by -0.0006, -0.0057 and -0.0007 respectively.

### c.)

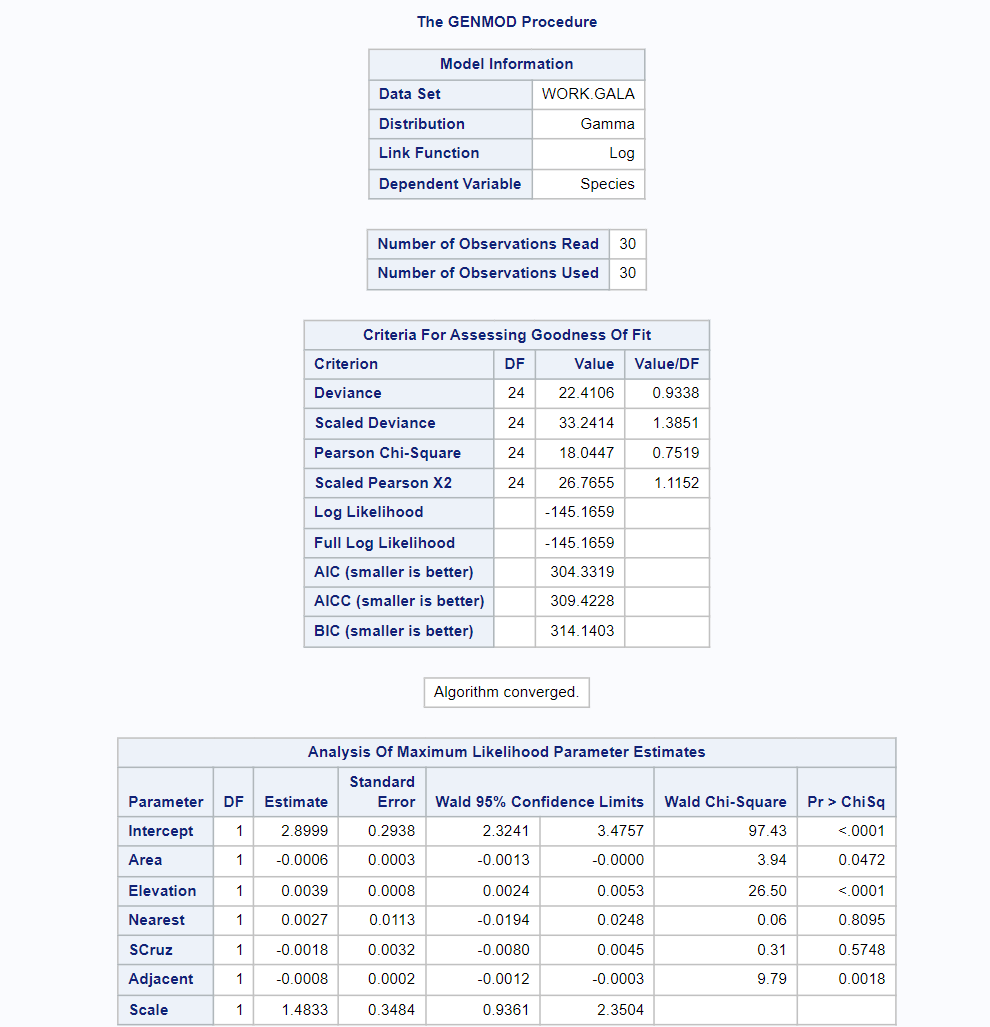
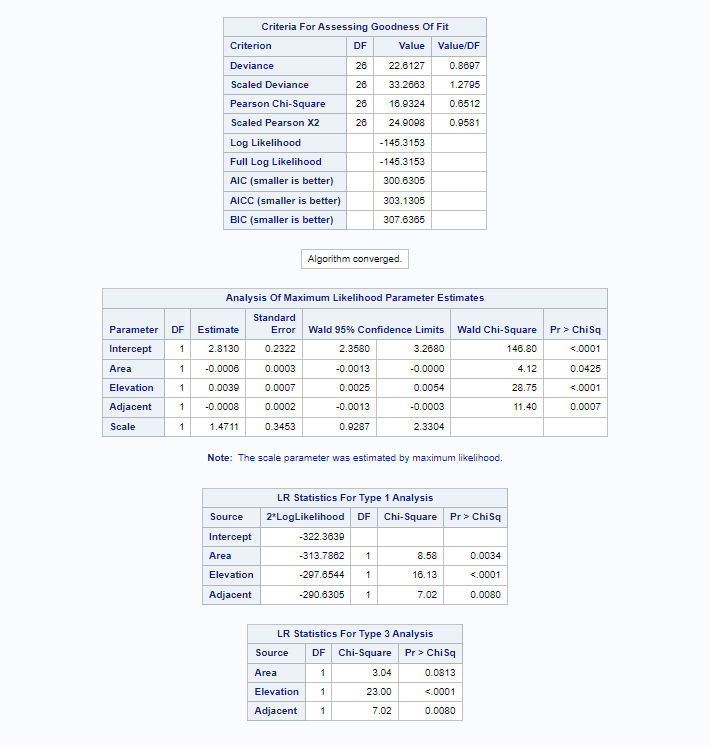
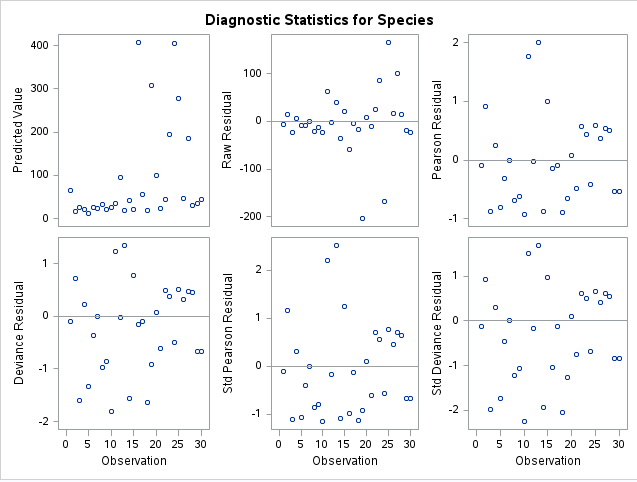
  
The residual plots don't appear to have any issues when standardized, we also know that the response variable is a positive count variable, so the assumption of the Poisson distribution could be reasonable, however, the Poisson distribution has mean and variance equal to each other.  
  
This is not the case in this data set, and the variance is far greater than the mean which suggests overdispersion, meaning we should instead use an over dispersed Poisson model as we will see in the next Problem.

## Problem 4

### a.)

  
  
Again I started with all the predictors, however this time, Nearest and Adjacent were not significant predictors so I removed them one at a time until I was left with only significant predictors. The predictors that were left are Area, Elevation and Adjacent. Again we are using a log link here, so we must consider than when interpreting the coefficient estimates. We see that as Elevation increases one unit, the log of the number of species is estimated to increase by 0.0036. While for Area and Adjacent, as they increase one unit, the estimated log of the number of species changes by -0.0006 and -0.0008 respectively.  
  
Again we see no apparent issues with the residual plots, expect for the raw residual plot which appears to have a slight "fan" shape. As seen in the last exercise, the variance of the response is greater than the mean of the response, which signifies overdispersion making the standard log link Poisson model not very reasonable. However, since we used the over dispersed Poisson model, this model is a reasonable one for this data set, since the Species variable is a positive count variable.

### b.)

  
  
I again started with all variables in the model, and then removed the non significant ones, one at a time using a significance level of 0.05. After this I was left with the same 3 predictors as the previous model, Area Elevation and Adjacent as significant at the 0.05 significance level. Since the Gamma uses a reciprocal log link function, we need to interpret the coefficient with that in consideration. So for the predictor elevation, as Elevation increases by a unit, it is estimated that the reciprocal log of the number of species increases by 0.0039. While for Area and Adjacent, as these predictors increase by one unit, the estimated reciprocal log of the number of species changes by -0.0006 and -0.0008 respectively.  
  
We see that the for the standardized residuals, we have a few large positive residuals above 2 while most of the negative residuals are just smaller than -1 or between -1 and 0. The assumptions of the model seem reasonable, since we can find a dispersion parameter to ensure that the variance is the square of the mean times this dispersion parameter. Also all the number of species are greater than 0 so that holds with the assumption of the Gamma distribution.

### c.)

The over dispersed Poisson and Gamma are similar in the fact that they have the same 3 predictors as being significant while the regular Poisson model had all 5 predictors as significant. For this data, I think the over dispersed Poisson model is the most reasonable for 2 reasons. Since the Species variable is over dispersed it makes sense to use this model instead of the standard Poisson model. Additionally, this model is easier to interpret than the Gamma model, since the Gamma model uses a reciprocal log link function rather than just a log link function which makes the over dispersed Poisson model simpler to understand as we can quickly do an exponential function of a prediction to recover the original response variable.