**Andrew Hillard**

**Data Analysis 2**

**Homework 3**

**Problem 1 Continued…**

(g) The data is fitted for a negative binomial and log link function. The fitted model is given below.

Both the deviance divided by the degrees of freedom, and the Pearson Chi-Square divided by the degrees of freedom are close to 1, indicating that the data is not overly dispersed. The model is better than the “null model” since and . Therefore, the likelihood ratio test is equal to . Since the likelihood ratio test for this comparison is distributed as a , the p-value Therefore, the “full model,” given above, is significant and a better fit to the data than the “null model.” As such, the model is significant and is useful to predict the number of days of absence, the goal of the experiment.

**Problem 2**

1. The model is given below.

refers to the i’th treatment of chemical concentration applied to the seeds. There are four possible chemical concentration treatments in this experiment.

j refers to the j’th treatment of temperature regime applied to the seeds. There are four possible temperature regimes in this experiment.

The mean number of seeds out of 200 that germinates when treated with the i and j treatments.

= The number of seeds in a given treatment group.

The intercept of the model, which can be interpreted as the overall mean number of germinated seeds out of 50.

= The main effect of the i’th chemical treatment.

The main effect of the j’th temperature treatment.

The interaction effect for the i’th chemical and j’th temperature treatment.

1 if the i’th treatment has been applied. Otherwise, -1.

1 if the j’th treatment has been applied. Otherwise, -1.

1. The difference between the deviance of model 4, the full model, and model 5, the reduced model, is equal to 92.5 and is distributed as a chi-squared distribution with 57-48 = 9 degrees of freedom. The p-value is very close to zero. Therefore, model 5 is significant and the interaction term is needed in the model.
2. The deviance divided by the degrees of freedom equals 1.15. This is close to 1 and over dispersion is probably not occurring. Therefore, the model seems to fit.
3. & (e) See next page of written work.

**Problem 3**

1. I created scatterplots for mpg01 against all possible predictor variables. The only scatterplots that showed a visible difference in the spread of data between mpg01 = 0 and mpg01 = 1 were displacement, horsepower, weight, and acceleration. Therefore, we will use these four predictor variables to start modeling the data.
2. The response variable is binary and the response distribution can be modeled using a Binomial distribution. We will use the canonical link, or logit link. Based on part a, the model is first built using the four predictor variables: displacement, horsepower, weight, and acceleration. The model is significantly better than the intercept only model since the ratio likelihood test equals , with a p-value of approximately zero. The Type 3 chi-squared test for the acceleration coefficient produces a p-value of 0.7308, indicating acceleration is not a significant coefficient. However, since acceleration is related to weight, acceleration may be interacting with weight, resulting in its insignificance. When the model is fitted with displacement, horsepower, weight, acceleration, and a weight acceleration interaction, the ratio likelihood test between the model with and without the weight acceleration interaction gives a p-value of 0.0253. The ratio likelihood test between the full model (including the variables displacement, horsepower, weight, acceleration, and a weight acceleration interaction) versus the reduced model (including the variables displacement, horsepower, weight, acceleration) gives a likelihood ratio value equal , p-value of 0.082. The p-value is close to 5%. While acceleration is not a significant parameter, the acceleration plus the weight acceleration interaction is very close to significant. Therefore, we will keep the model that has the following variables: displacement, horsepower, weight, acceleration, and a weight acceleration interaction. The fitted model is given below.

The model assumes a binomial distribution and a logit link.

The confusion matrix is given below and assumes a cutoff of 0.5.

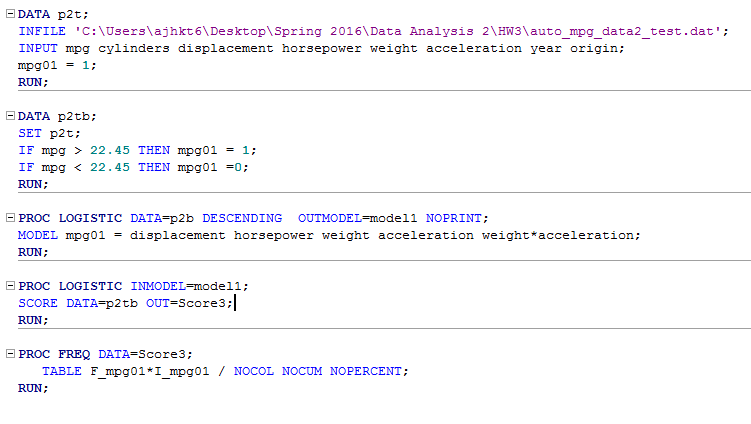
|  |  |  |
| --- | --- | --- |
|  | Predicted | |
| Truth | 1 | 0 |
| 1 | 91.8% | 8.6% |
| 0 | 12.3% | 87.1% |

The sensitivity and specificity rates are high and the false negative and false positive rates are low, indicating that the model is good at determining whether the miles per gallon will be above or below the median for the training set.

1. When I score the test data and assume a cutoff of 0.5, the confusion matrix has high sensitivity and specificity rates. The false negative and false positive rates are still low. Overall, the confusion matrix for the training and test data are very similar. Therefore, the model and overall quality of the prediction is good.

|  |  |  |
| --- | --- | --- |
|  | Predicted | |
| Truth | 1 | 0 |
| 1 | 92.31% | 7.69% |
| 0 | 16.67% | 83.33% |

The R Code for the scoring of the test data and the creation of the confusion matrix is given below.



**Problem 4**

1. The below model is fit using the data.

The mean for the i’th macrohabitat.

The overall mean of all macrohabitats, also known as the intercept term in the model.

The effect of the i’th macrohabitat.

Indicator variables for i’th treatment. Equal to 1 if i of is equal to treatment number. Otherwise, equal to -1.

The fitted model is given below.

Macrohabitat 1 and Macrohabitat 2 are significant in the above model, indicated by their low p-values. Macrohabitat 3 is not significant and it provides no more information than the intercept term, indicated by its high p-value. Therefore, Macrohabitat 3 and 4 are not significantly different than one another.

The model appears to be over dispersed since the deviance divided by the degrees of freedom is 9.8390 and the dispersion parameter for Poisson is 1.

1. The zero-inflated Poisson portion of the model is given below.

The zero model portion of the model is given below, where are the indicator variables for the gear type.

The zero-inflated Poisson model has a full log likelihood of -671.8967, deviance of 1343.7934, and AIC of 1361.7934. The Poisson model from part (a) had a full log likelihood of -992.4129, deviance of 1672.6277, and AIC of 1992.8258. The log likelihood ratio test between the two models is equal to with a chi-squared distribution, 5 degrees of freedom. The p-value of the log likelihood ratio test is approximately 0. The model is therefore significantly better at explaining the data. In addition, the AIC is much lower in the zero-inflated Poisson model. It’s interesting to note that in both models X3 is insignificant when you run the Wald test.

The portion of the model that uses gear type as the indicator variable, the zero model portion, could be refined by dropping insignificant variables, which include , , and potentially the intercept.