**Lab 4**

The data used in this lab was provided by the instructor, and the data can be seen in the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Female Crab Width (cm) | Age Category | | | |
| Young | | Old | |
| Number Observed | Number Having Satellites | Number Observed | Number Having Satellites |
| 23 | 14 | 5 | 12 | 1 |
| 24 | 14 | 4 | 15 | 2 |
| 25 | 28 | 17 | 24 | 10 |
| 26 | 39 | 21 | 35 | 11 |
| 27 | 22 | 15 | 25 | 15 |
| 28 | 24 | 20 | 22 | 17 |
| 29 | 18 | 15 | 17 | 14 |
| 30 | 14 | 14 | 14 | 12 |

1. Construct and solve five logistic GLMs for different combinations of individual and interacting factors.

Multiple Logistic Regression is an approach to perform analysis using more than one independent variable, in our case age categorization and width of female horseshoe crabs. Similar to logistic regression with a single predictor, multiple logistic regression produces a GLM in the form of:

The model can be designed to account for individual levels of variables, as well as joint effect of the variables analyzed.

For the purposes of this assignment, α, β, and X will relate to the widths (in cm) of the female crabs; while γ and Y will refer to the age category of the crab (young or old). Because Y is binary, two equations for each model will be displayed, and γ will be treated as a constant that is either included or disregarded based on the age category. In doing this, some of the models will appear as though they are only a function of width; in reality we are analyzing the effects of width at different levels of age.

* 1. **Model 1**

Model 1 was created analyzing width only (i.e. assuming age has no effect), and takes the following form:

.

|  |  |  |  |
| --- | --- | --- | --- |
| Coefficients | Value | P-Value | Significance Level |
| α | -13.463 | 8.8E-13 | 0.001 |
| βwidth | 0.52193 | 7.284E-13 | 0.001 |
| Null Deviance | (Degrees of Freedom): | 88.907 | (15) |
| Residual Deviance | (Degrees of Freedom): | 21.045 | (14) |
| AIC: | | 72.249 | |

Table 1: Tabular data relating to Model 1.

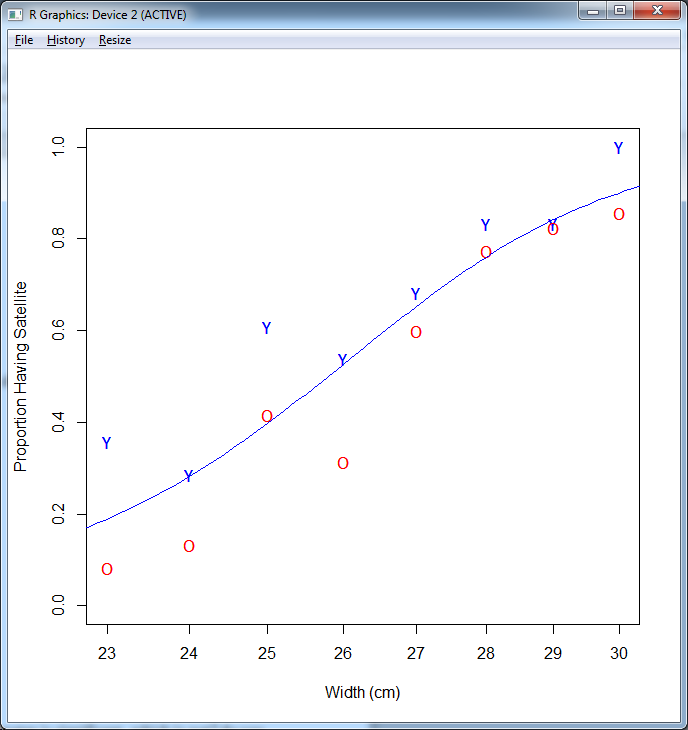


Figure 1: A graphical representation of Model 1 against a plot of the sample data.

This model does seem to do a good job of predicting the variation in the data and has significant parameter estimations, but it seems that a model that includes an age parameter might do an even better job explaining the variation in the data.

* 1. **Model 2**

Model 2 was created using only age category as its only predictor (i.e. assumes width has no effect) which, separating old and young crabs, might be expressed as:

Young Crabs:

Old Crabs:

|  |  |  |  |
| --- | --- | --- | --- |
| Coefficients | Value | P-Value | Significance Level |
| α | 9.882E-17 | 1.000 | NONE |
| γ | 0.5824 | 2.617 | 0.01 |
| Null Deviance | (Degrees of Freedom): | 88.907 | (15) |
| Residual Deviance | (Degrees of Freedom): | 81.987 | (14) |
| AIC: | | 133.19 | |

Table 2: Tabular data relating to Model 2.

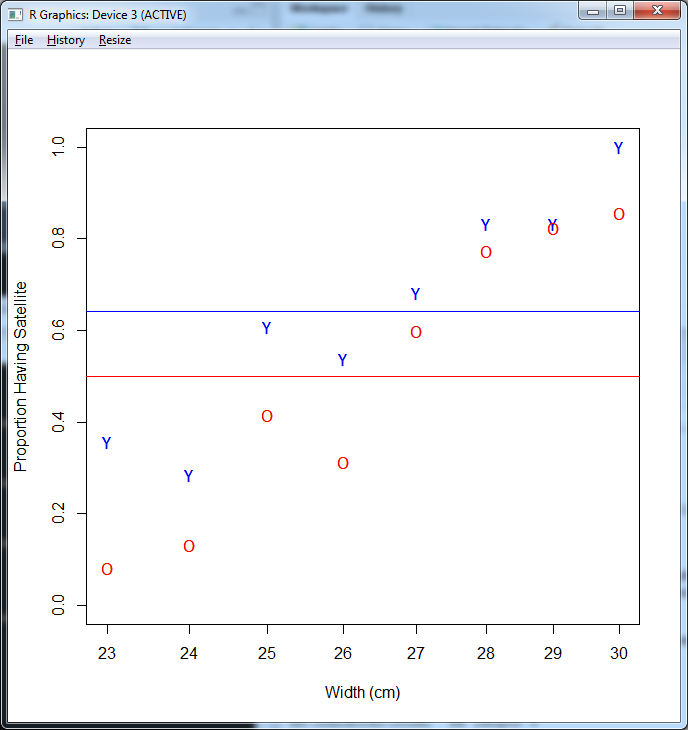


Figure 2: A graphical representation of Model 2 against a plot of the sample data.

This model leaves a lot to be desired, poorly predicting much of anything. While the effect of age does appear to be significant, the common intercept is not, making the utility of this model questionable.

* 1. **Model 3**

Model 3 considers the separate effects of age and width, it assumes that width has the same effect regardless of age, but that age still has an effect as well, resulting in differing intercepts, but the same slope:

Young Crabs:

Old Crabs:

|  |  |  |  |
| --- | --- | --- | --- |
| Coefficients | Value | P-Value | Significance Level |
| α | -14.33322 | 2.23E-13 | 0.001 |
| ageCatY | 0.075451 | 0.00265 | 0.01 |
| width | 0.54056 | 1.86E-13 | 0.001 |
| Null Deviance | (Degrees of Freedom): | 88.907 | (15) |
| Residual Deviance | (Degrees of Freedom): | 11.771 | (13) |
| AIC: | | 64.975 | |

Table 3: Tabular data relating to Model 3.

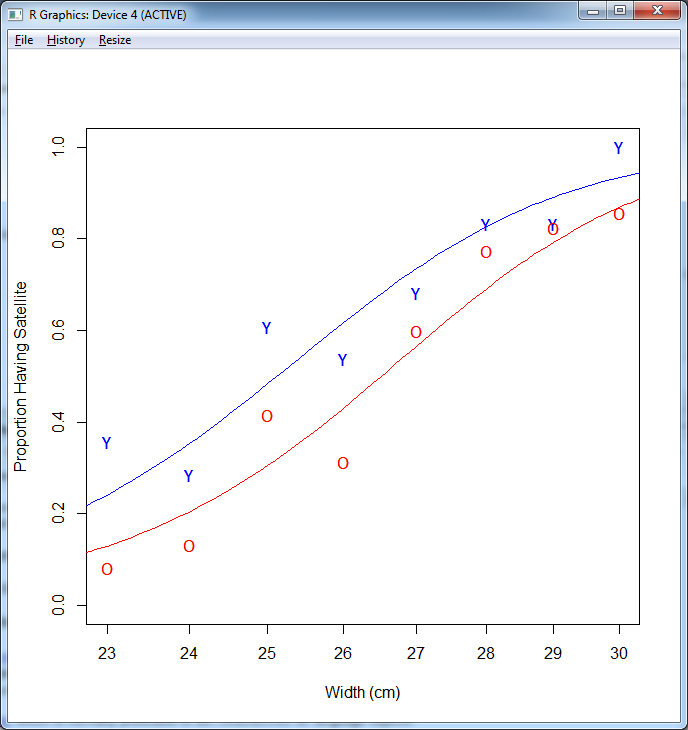


Figure 3: A graphical representation of Model 3 against a plot of the sample data.

This model appears to be a good predictor, has significant parameters, and the lowest residual deviance and AIC thus far. It is a good candidate for what may be the best model of the five.

* 1. **Model 4**

Model 4 considers the joint effects of age and width, also referred to as the interaction of age and width, the intercept is the same, but the sloped differ:

Young Crabs:

Old Crabs:

|  |  |  |  |
| --- | --- | --- | --- |
| Coefficients | Value | P-Value | Significance Level |
| α | -13.94177 | 4.96E-13 | 0.001 |
| βOld | 0.52614 | 4.99E-13 | 0.001 |
| βYoung | 0.55419 | 8.87E-14 | 0.001 |
| Null Deviance | (Degrees of Freedom): | 88.907 | (15) |
| Residual Deviance | (Degrees of Freedom): | 12.182 | (13) |
| AIC: | | 65.387 | |

Table 4: Tabular Data relating to Model 4.

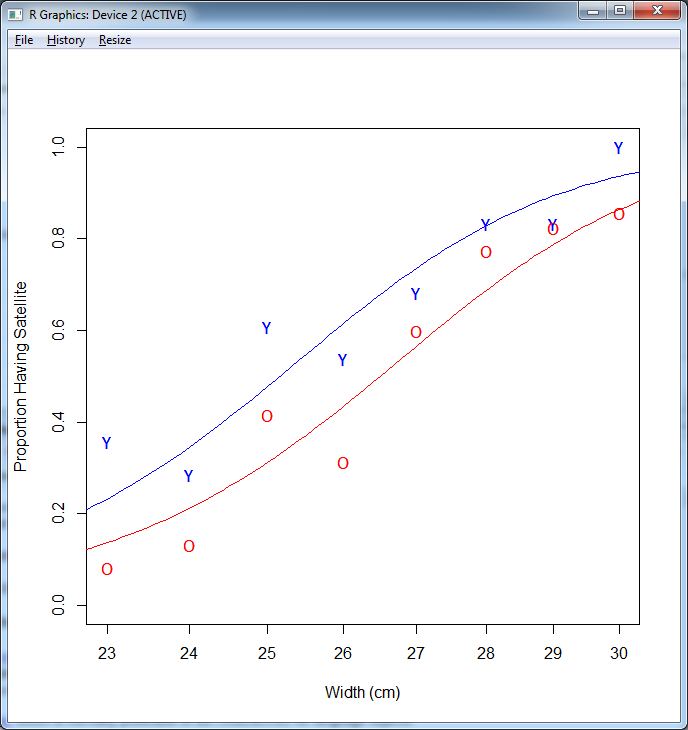


Figure 4: A graphical representation of Model 4 against a plot of the sample data.

This model also appears to be effective, but just doesn’t quite stand up to model 3 as it has a slightly higher AIC and residual deviance. Its collective parameter significance is superior, but model 3’s was very acceptable already.

* 1. **Model 5**

Model 5 also considered the joint effects of age and width, but this time with an age-dependent slope and intercept:

Young Crabs:

Old Crabs:

|  |  |  |  |
| --- | --- | --- | --- |
| Coefficients | Value | P-Value | Significance Level |
| α | -16.2929 | 1.98E-08 | 0.001 |
| γ | 4.4998 | 0.247 | N |
| β | 0.6145 | 1.91E-08 | 0.001 |
| βYoung | -0.1427 | 0.334 | N |
| Null Deviance | (Degrees of Freedom): | 88.907 | (15) |
| Residual Deviance | (Degrees of Freedom): | 10.832 | (12) |
| AIC: | | 66.036 | |

Table 5: Tabular data relating to Model 5.

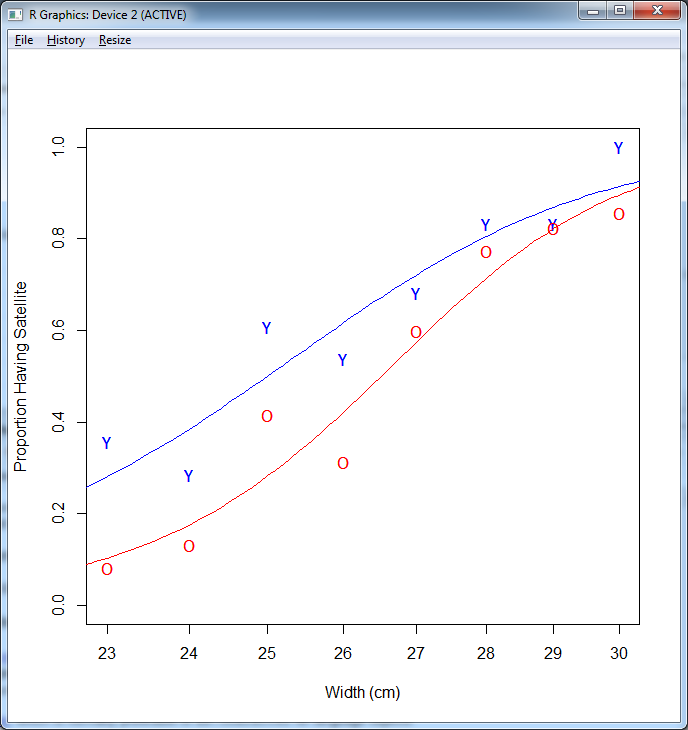


Figure 5: A graphical representation of Model 5 against a plot of the sample data.

Model 5 appears to be a good predictor as well, but it includes parameters without significance and has a higher AIC than model 3. Collinearity is a definite concern with this model, as the age dependent slope and intercept parameters are insignificant, indicating that they may have the same effect. In this case, adding more parameters did not improve the quality of our model.

1. Discuss parameter significance (which parameter is significant, which is not? do you see collinearity effect?) and compare the model quality using the graphs, deviance and AIC.

Parameter significance is a test of whether or not a coefficient can reasonably be assumed to have any bearing on the outcome, in our case the probability/proportion of female crabs with satellites. When evaluating parameter significance, we are testing the null hypothesis that the coefficient of said parameter is not equal to zero (for example, H0: βwidth = 0, HA: βwidth ≠ 0). If we fail to reject the null hypothesis in any case, we should understand that this parameter estimate is likely flawed, and therefore not trustworthy; we may decide to drop that parameter. As can be seen in these models, with few exceptions, most parameters were significant at the 0.001 level, the exceptions being cases of collinearity, discussed later.

Related to parameter significance, is the appearance of collinearity or multicollinearity. When the coefficients of two parameters are insignificant, this should indicate the possibility that collinearity exists. When parameters are collinear, they are highly linearly related, and in a sense predict each other. While the model will still be a good predictor for the data set it was created from, but are not robust and not very generally applicable. Some methods for remedying collinearity might be dropping one of the parameters or collecting more data.

One way to compare models to one another, providing a method for determining if a model is superior to another, is by comparing the deviance of a model, or how far the model departs from the observed data.

The null deviance (the deviance from the null hypothesis that the variables are independent from one another) is of relatively little value, especially when we do not include variables that are expected to be independent. Including variables/parameters in a model that are not meant to predict anything would be preposterous.

However, the residual deviance, which is the deviance of the model from the data in the sample used to generate the model. While residual deviance is not an indicator of how universally accurate a model may be (it is an indicator relative only to the sample data used to create the model), it certainly can be used to guide decisions on whether one model or another should be used.

The Akaike Information Criterion (AIC) can also considered when evaluating how well a model may or may not fit the data used to create that model. It is, in effect, a measure of how much information is estimated to be lost when a model is used to describe reality. The AIC can be found using the expression:

,

where L is the maximized likelihood function for the model, and k is the number of parameters in the model. It is important to note that the AIC can only be used to compare models created using the same data set; the AIC is not a metric of how accurate a model truly is.

1. Choose the best model and use it to predict the proportion of old female crabs with width 28 cm and having satellites.

I choose model 3 as the best model based on the fact that it has the lowest AIC (64.975), as well as a very low residual deviance of 11.771 (only model 5 has a lower residual deviance, 10.832). Furthermore, model 5, despite having the lowest residual deviance and a low AIC, has two parameters that are not significant, while model 3’s parameters are all significant at the 0.01 level or better.

Using model 3 (the model that assumes the effect of width does not change with age, but that age does have its own effect), we find that the expected proportion of older female crabs that are 28 cm wide that have satellites will be **0.6905258**.