**Lab 3**

1. Horseshoe Crab Study
   1. Construct GLMs

Generalized Linear Models

The logit function is a sigmoid function, and can be useful in characterizing changes in probability due to the fact that a non-linear function can represent reality better than a linear one. While logistic regression allows us to linearize a logit model in order to be able to more easily see how changes in the independent variable affect the probability of observing a success in the following manner:

[1]

Which is actually the logarithm of the odds ratio:

[2]

The probability of a success, , can be found by:

[3]

It should be noted that the sigmoid function [3] is valuable for producing predicted probabilities, while the linearized model produces results in terms of logarithms of odds ratios. This is why the sigmoid function is more useful than the linear function, because a scale of logarithms of odds ratios are not as intuitive as a scale of probability values are.

A related general linearized model is the probit model. The probit model is also sigmoid shaped, and transforms probabilities to normalized z-scores. Probit regression (whose formula is not expressed here) produces a GLM similarly to logistic regression:

[4]

Both logit and probit models were produced for the given horseshoe crab data, wherein width of a female with an attached male was related to whether or not that female crab had at least one satellite male. These models were computed using R and its inherent smoothing methods to produce the following:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Coefficient | Value | Significance Level |
| Logit | α | -11.79311 | 0.001 |
| β | 0.47186 | 0.001 |
| Probit | α | -7.1806 | 0.001 |
| β | 0.2873 | 0.001 |

Or, written explicity, where X is the width of the female crab in centimeters:

Logit(π(having a satellite)) = -11.79311 + .47186 \* X

Probit(π(having a satellite)) = -7.1806 + .2873 \* X

According to both models, as the width of a female crab increases, so does the probability that she has a satellite male crab.

* 1. Plot the observed sample proportions and the predicted values from the *logit* and *probit* models on a graph.

Using R, graphs of predicted values were produced using the previously found models:

The Logit Model



Figure 1. A plot of the data with the found logistic regression model superimposed on it.

The Probit Model



Figure 2. The plot of the data with the probit regression curve superimposed.

Looking at Figures 1 and 2 it is apparent that the logit and probit models both fit the data well, and also fit similarly to one another.

* 1. Find the probability of having at least one satellite for a female of width 25.5 cm according to each model.

These values were computed using the predict( ) function in R, but the logit result certainly could have been computed by hand.

|  |  |  |
| --- | --- | --- |
| Model | Female Width (cm) | Predicted Probability of Having at Least One Satellite |
| Logit | 25.5 | 0.55952 |
| Probit | 25.5 | 0.55779 |

The two models produce similar probabilities, and both models return a probability that indicates slightly more than half of female crabs 25.5 cm wide will have at least one male satellite. It should be noted that these values did not come from the linear models, but rather the sigmoid ones. The logit result corresponds with a value equal to the logarithm of a certain odds ratio for the given input in the linear model, and the probit result corresponds with a certain normalized z score (~0.145) that would have been produced by its linearized function.

1. Use different numerical representation of snoring levels to construct *logit* and *probit* models for snoring data. Find such snoring levels that your models have the best fit to the data (by trial-and-error). Compare the best levels and discuss.

The Logit Model



Figure 3. The graph of a logit model fitted to the provided snoring data, using a “trial and error” approach to estimate quantified values for the various snoring levels.

The Probit Model



Figure 4. The graph of a probit model fitted to the provided snoring data, using a “trial and error” approach to estimate quantified values for the various snoring levels.

|  |  |  |
| --- | --- | --- |
| Snoring Level | Values Used in Logit Model | Values Used in Probit Model |
| Never | 0 | 0 |
| Occasional | 3 | 3 |
| Nearly Every Night | 4.3 | 4.3 |
| Every Night | 4.9 | 4.9 |

The table above displays the values assigned to the different snoring categories. Note that the same values were used for both the logit and probit models. This is because differing numbers really did not produce any advantage in terms of fitting. It was to be expected that using the same snoring values would produce the best results, as “In practice, probit and logistic regression models provide similar fits” (Agresti, 2007). While a simple guess and check method was used, I believe these fittings to be superior because the lines pass through (or nearly through) two of the four points, and the line passes between the remaining two, a path that I would assert minimizes the difference between the prediction line and the points on the line.

