Test #2

STAT 875

Spring 2015

Complete the problems below. Make sure to fully explain all answers and show your work to receive full credit!

1. (31 total points) Vansteelandt et al. (2000) examine a data set involving the HIV status (hiv, 1 = positive, 0 = negative) of pregnant Kenyan women. The following explanatory variables are available in the data set:

|  |  |
| --- | --- |
| **Variable name** | **Description** |
| age | Age in years |
| education | Highest attained education level: 1 = no schooling, 2 = primary school, 3 = secondary school, and 4 = higher |
| parity | Number of children |

The data can be obtained from the graded materials web page of my course website. Below is how I read the data into R:

> set1 <- read.csv(file = "C:\\data\\Vansteelandt.csv")

> head(set1)

age education parity hiv

1 10 3 0 0

2 15 2 0 0

3 15 2 1 0

4 15 2 0 0

5 15 1 0 0

6 16 2 1 0

Complete the following using this data.

* 1. (6 points) Estimate the logistic regression model with age, parity, and their corresponding interaction in the model.

* 1. (7 points) Calculate the convergence criteria value used by R to determine that convergence was obtained at the last iteration. Why did R indicate convergence was obtained?
  2. (12 points) Fully interpret the effect parity has on HIV through using 95% profile likelihood ratio intervals for odds ratios. Use c = 1.

* 1. (6 points) Suppose you would like to include education as a categorical explanatory variable in the model. Because this variable is given in a numerical format, R will by default include the variable in the model in a numerical manner with only one term (say, β4education). Show how to make R treat this as a categorical explanatory variable in the model.

1. (27 total points) The alligator.csv file contains a listing of an alligator’s length (length) and the primary type of food found in the alligator’s stomach (food). The food categories are fish (F), invertebrate (I), and other (O). The data can be obtained from the graded materials web page of my course website. Below is how I read in the data:

> gator <- read.csv(file = "C:\\data\\alligator.csv")

> head(gator)

length food

1 1.24 I

2 1.30 I

3 1.30 I

4 1.32 F

5 1.32 F

6 1.40 F

Use length as an explanatory variable to estimate the primary type of food consumed. Complete the following.

* 1. (6 points) Estimate the multinomial regression model using length and the square of length as explanatory variables in the model.
  2. (8 points) Perform one likelihood ratio test that evaluates the importance of the squared term in the previous model. Make sure to fully state all hypotheses and use α = 0.05.

* 1. (6 points) Suppose an ordering of the food is other (O) < fish (F) < invertebrate (I). Estimate the proportional odds regression model than includes ONLY length as an explanatory variable.
  2. (7 points) Fully interpret the effect length has on food source through using one 95% profile likelihood ratio interval for an odds ratio obtained from the model in part c). Use c = 1.

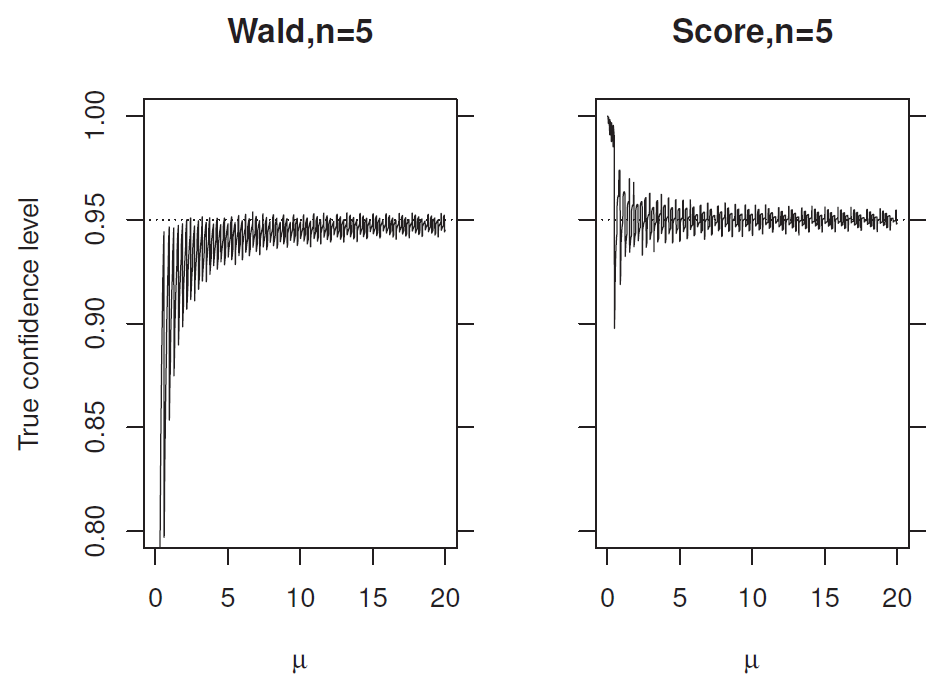
1. (42 total points) Answering the following questions:
   1. (7 points) Since becoming the father of two young children, your instructor has seen first-hand how statistics can be used to better understand children and their behavior. For example, late on the night of January 22, 2013, your instructor tweeted the following regarding his 8-month-old son Keegan:

I think the number of times that I put a pacifier back into Keegan's mouth each night is distributed Poisson(mu=4).

If you are unaware of the habits of babies, pacifiers are very important to many of them. When a baby wakes up a night without a pacifier in their mouth, they will often start crying. This will wake up their parents, leading a parent to place a pacifier back into the baby’s mouth.

In the context of this tweet, what should be done to determine if the Poisson distribution with μ = 4 is appropriate?

* 1. (12 points) Name and describe the three components of a generalized linear model.
  2. (7 points) When n = 5 and Yi has a Poisson distribution for i = 1, …, n, the two plots below display the true confidence levels of 95% Wald and score intervals. Which of these two intervals is generally better for this setting? Explain.



* 1. (8 points) Specifically describe why a proportional odds regression model has “proportional” and “odds” in its name.
  2. (8 points) What is the purpose of an offset when modeling a Poisson response variable? Fully explain your answer.

1. (2 points extra credit) What is the name of one person who was thanked in my book?