Statistical Methods II
Midterm II
In-Class

Name: Key Date: March 31, 2011

This in-class examination covers the most important things we covered during the second third of the course. It is worth 15 points; the take-home is worth 85 points. Please make sure your answers are correct and make sense.

Good luck!

Section: Links

For each of the following distributions, name the *canonical* link function.

[1] 1. Poisson log

[1] 2. Binomial logit

[1] 3. Gaussian identity

SECTION: TESTS

For each of the following problems, I describe the variables in a set of data. From the descriptions, select the appropriate distribution family.

[2] 4. There are two variables. The independent variable is a count of the number of times a person has had a heart attack. It ranges from 0 to 9 in this sample. The dependent variable is the cost of the person's monthly insurance premiums. It is effectively continuous and ranges from \$400 to \$3400 in this sample. What is the appropriate distribution family in this problem?

Best: Poisson, as discrete
Acceptable: Gaussian, as effectively continuous
Bad: Binomial

[2] 5. There are two variables. The independent variable is the age of tree according to a new age test. It is a continuous variable that ranges from 10 to 25 years in this sample. The dependent variable is the true age of the tree. It is a continuous variable that ranges from 10 to 20 years in this sample. What is the appropriate distribution family in this problem?

Best: Gaussian, as continuous
Poor: Poisson, as "count of years"
Bad: Binomial

[2] 6. There are two variables. The independent variable is the patient's score on a revolutionary Schizophrenia Screening Quiz (SSQ). It is discrete and ranges from 1 to 12. The dependent variable is an indicator variable of whether the patient actually has schizophrenia. As it is an indicator variable, it only takes on values of 0 and 1. In our sample, 45% of the patients actually had schizophrenia. What is the appropriate distribution family in this problem?

Best: Binomial
Bad: Gaussian & Poisson

[2] 7. There are two variables. The independent variable is the level of religiosity in the US state. It is a discrete variable taking on the values 1, 2, 3, 4, and 5. The dependent variable is the poverty rate in the US state. It is an effectively continuous variable that ranges from 8.6 to 20.1. What is the appropriate distribution family in this problem?

Best: Poisson as discrete
Acceptable: Gaussian as "effectively" continuous
Bad: Binomial

R FUNCTIONS

[4] 8. Back to Problem 6 (above): My research hypothesis is that those scoring higher on the SSQ tend to have a higher probability of having schizophrenia. If we assume the patient's SSQ score is contained in the variable `score` and the paatient's schizophrenia status is contained in the variable `schizo`, answer the following questions.

(1) What is the null hypothesis? H_0: Score (schizo) $\leq$ Score (non-schizo)

(2) What is the alternative hypothesis?

H_A: Score (schizo) $>$ Score (non-schizo)

(3) Since we will be using GLMs to determine if the effect exists, we need to know the canonical link function. Assuming your answer to Problem 6 is correct, what is the canonical link?

logit

(4) Finally, let us assume there is a need for a `data=ssq` parameter in our R function. With this, what is the actual function call (what you have to type) to get R to estimate the appropriate model?

glm(schizo ~ score, data = ssq, family = binomial (link = logit))