STATISTICS 475: Applied Discrete Data Analysis
MIDTERM 1
Feb 24, 2016

NAME: _____

ID Number: _____

INSTRUCTIONS: **DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO!**

- Answer all parts of all problems in the space provided.
- Place final answers in the blanks when they are provided.
- **Show work!**
  - For hypothesis tests you need to show all parts of the test
  - For confidence intervals you need to explain what it is that you expect the interval to cover
- If you need more space (I don't think you should), indicate clearly where the problem continues.
- Use $\alpha = 0.05$ unless told otherwise

Score: _____+_____+_____+_____+_____+_____=_____/28

PROBLEM 1 (8 pts)
A medical group wanted to see whether praying for recovery could help hospitalized patients recover from a certain serious illness. They asked patients early in their hospitalization whether they had been praying for recovery, and also recorded whether the patient recovered. The results are summarized below.

	Recovery	No recovery	Total
Prayer	9	11	20
No Prayer	3	6	9
Total	11	18	29

1. (4 pts) Test whether there is a difference in probability of recovery between patients who pray and those who don't.

let π_1 = probability of recovery of prayer group
π_2 = probability of recovery of no prayer group

$H_0: \pi_1 = \pi_2$, $H_A: \pi_1 \neq \pi_2$

$\hat{\pi}_1 = \frac{9}{20}$, $\pi_2 = 3/9 = 1/3$, $\bar{\pi} = 12/29$

Score:

$$z_0 = \frac{\frac{9}{20} - \frac{1}{3}}{\frac{12}{29}(1-\frac{12}{29})(\frac{1}{20}+\frac{1}{9})} = \frac{7/60}{12/435} = 2.98 > 1.96$$

so, we cannot reject H_0

2. (3 pts) Find a 95% Wald confidence interval for the ratio of the odds of recovery with prayer to the odds of recovery without prayer.

$$\widehat{OR} = \frac{9(9-3)}{3(20-9)} = \frac{18}{11}$$

$$\log(\widehat{OR}) \pm 1.96\sqrt{\frac{1}{9}+\frac{1}{11}+\frac{1}{3}+\frac{1}{6}}$$

$$= \log(\widehat{OR}) \pm 1.642$$

$$= [3.278, -5.636 \times 10^{-3}]$$

$$\exp[3.278, -5.636 \times 10^{-3}] = [0.994, 26.523]$$

3. (1 pt) Do the results of the previous two parts agree regarding their conclusions about the relationship between prayer and recovery? Explain.

We failed to reject $\pi_1 = \pi_2$, and the confidence interval does contain $OR = 1$, so it is possible to say they agree. However the C.I. is really wide in this case, so the true OR may vary.

PROBLEM 2 (4 pts)
Suppose that when a married couple drives somewhere together, the male does the driving 70% of the time. Among cars containing a married couple,

1. (1 pt) What is the expected ratio of cars being driven by the husband to cars being driven by the wife?

   70% and 30%

2. (1 pts) What is the statistical term for this quantity? π

3. (1 pts) Suppose I observe 10 cars with married couples. Assuming that the cars are independent of one another, what is the probability that all of them have male drivers?

   $\left(\frac{7}{10}\right)^{10} = 0.028$

4. (1 pts) Suppose I take a larger sample of cars. In addition to who was driving, I observe things like type of car, age of car, state of residence, age of the couple, and so on. Describe how this information could be used to improve the claim that the male does 70% of the driving. (One or two sentences are enough.)

   If we use model approach, more variables included improves the overall fit of the model. So better we can improve our claim.

PROBLEM 3 (2 pts)
Why do we usually not use an identity link in a generalized linear model with a binomial random component? Give two reasons.

1. Linear model is not bounded by 0 and 1

2. Y_i is not randomly distributed.

3. potentially we have changing variance

PROBLEM 4 (6 pts)
Consider a problem where we have a binary response, Y, with four explanatory variables: A is categorical with 3 levels, B is categorical with 5 levels, and C and D are numerical. We want to fit a logistic regression model with all main effects and the interactions between A and B.

1. (1 pts) Using these same variable names, write the `glm()` code to fit this model.

   `mod.fit <-` glm(Y ~ A + B + C + D, family = binomial(link = logit), data = ...)

2. (1 pt) How many parameters will be fit, including the intercept? 8

3. (2 pts) The table of coefficients from `summary(mod.fit)` looks like this:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.702359   0.141176   xxxxx   xxxxxx
                         etc.
```

   (Of course, there would be more rows than just the one for "`(Intercept)`".)
   Note the two columns "`z value`" and "`Pr (>|z|)`". Explain exactly what these two columns represent.

   Z value contains the test statistics for significance test of each parameter

   Pr(>|z|) is p value of the test.

4. (1 pt) Suppose we wanted to do a LR test to see whether we can get rid of all of the numerical explanatory variables without adversely affecting the model fit. How many degrees of freedom would the test have? 2

   (a) (1 pt) Could I answer this question based on the coefficient table from `summary(mod.fit)`? Explain.

   No, summary(glm.fit) reports the df of model fit.

BONUS: (1 pt) Who wrote Chapter 2 of the class textbook?

Chris

PROBLEM 5 (8 pts)
Consider a problem where we have a binary response with three explanatory variables: A is numerical, B is binary, which we treat as numerical 0/1, and C is categorical with 4 levels. We fit a logistic regression model with all three main effects and the interactions between A and B and between B and C.

From the model fit, we determine that the coefficient order is as given in the top row of the table below, where "Int" means intercept, "AB" and "BC" represent interaction terms, and the numbers after "C" correspond to the levels of C.

We want to find LR confidence intervals for a variety of different quantities as indicated below the table. We plan to use `mcprofile()` to do this, so that we need to know the coefficients to put into the matrix for the function. For each quantity, determine the coefficients that need to be attached to the corresponding parameters in order to estimate that quantity. Enter them into the table in their corresponding line. You may leave zeroes blank if you wish.

Part	Int	A	B	C2	C3	C4	AB	BC2	BC3	BC4
1										
2										
3										
4										

1. (2 pts) The probability of success when $A = 50$, $B = 1$, and C is at the first level

2. (2 pts) The odds of success when $A = 12$, $B = 0$, and C is at the second level

3. (2 pts) The odds ratio representing the effect of a 1-unit increase in A, assuming that $B = 0$ and C is at the third level

4. (2 pts) The ratio of odds ratios that would be used to test whether the effect of B is the same at the fourth and first levels of C, holding A constant.