Stat 439: Homework 5

Due Thur 3/31/22 by 11pm in Gradescope

Your name here

Instructions

You are strongly encouraged to use R Markdown to complete your Homework assignments, starting with this file as a template and Knitting to pdf. Submit your homework to Gradescope as a single pdf file.

Part I

Consider the data described in Agresti Exercise 3.11. The wafers were also classified by thickness of silicon coating (0 = low, 1 = high). The first five imperfection counts reported for each treatment refer to low thickness, and the last five refer to high thickness. The code below will input the data into R.

```
Y <- c(8,7,6,6,3,4,7,2,3,4,
         9,9,8,14,8,13,11,5,7,6)
Trt <- c(rep("A",10), rep("B",10))
Thick <- rep(c(rep(0,5), rep(1,5)), 2)
dat <- data.frame(Y, Trt, Thick)
head(dat)
```

```
##   Y Trt Thick
## 1 8   A     0
## 2 7   A     0
## 3 6   A     0
## 4 6   A     0
## 5 3   A     0
## 6 4   A     1
```

a. Complete Agresti Exercises 3.11 and 3.12.
b. Researchers would like to investigate the effects of treatment type and of thickness of coating on the production of silicon wafers for computer chips. Choose an appropriate model to analyze these data, carry out the analysis, then present your analysis in narrative/report form. Your analysis should include:
    - at least one appropriate plot that illuminates the relationships between the response and the predictors, including a description of the plot,
    - justification of your model choice,
    - a printed summary of the fitted model, and
    - confidence intervals and interpretations of parameters relative to the research question.

Part II

Problem 1

The 1976 U.S. Panel Study of Income Dynamics measured a number of variables on a sample of 753 married women. This data set is built into `R` in the `carData` library. Load the data and then look at its help file for variable definitions.

```r
data(Mroz)
Mroz$lfp <- as.numeric(Mroz$lfp == "yes")  # Code response as 0/1
```

We are going to consider three predictor variables: wife's college attendance (`wc`), wife's age in years (`age`), and number of children 5 years old or younger (`k5`) categorized as 0, 1, or ≥ 2. First, create a new variable `k5f`:

```r
Mroz$k5f <- cut(Mroz$k5, c(-1,0.5,1.5,3.5), labels = c("Zero","One","Two_or_more"))
```

Consider the following model:

```r
mod1 <- glm(lfp ~ wc + k5f + age, family = binomial, data = Mroz)
summary(mod1)
```

```
##
## Call:
## glm(formula = lfp ~ wc + k5f + age, family = binomial, data = Mroz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0303  -1.1388   0.7251   1.0152   1.9514
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       2.98193    0.51394   5.802 6.55e-09 ***
## wcyes             0.81116    0.18406   4.407 1.05e-05 ***
## k5fOne           -1.53064    0.24068  -6.360 2.02e-10 ***
## k5fTwo_or_more   -2.65782    0.47900  -5.549 2.88e-08 ***
## age              -0.06026    0.01122  -5.370 7.89e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.75  on 752  degrees of freedom
## Residual deviance:  943.75  on 748  degrees of freedom
## AIC: 953.75
##
## Number of Fisher Scoring iterations: 4
```

a. Compute the estimated odds ratio and a corresponding 95% confidence interval comparing two groups of women similar in age and education, where one has two or more children 5 years old or younger and the other only one child 5 years old or younger. *Hint*: This will require you to use the variance-covariance matrix of the estimated coefficients to calculate the standard error needed for the confidence interval. It may be helpful to write a function to calculate such an interval, as you will need to use the same method for later questions.
b. Use the `anova` function to perform a likelihood ratio test of the null hypothesis that both true coefficients of `k5fOne` and `k5fTwo_or_more` are equal to zero. State your conclusion in context.
c. Now suppose it is believed that the effect of age may change depending upon the number of children under age 5 a woman has. Modify `mod1` to allow for the effect of age to change depending upon the number of children under age 5 a woman has. Name this model `mod2` and print its summary.
d. Using `mod2`, estimate the effect of age and report a 95% confidence interval for:
   - women with no children under age 5,
   - women with 1 child under age 5, and
   - women with 2 or more children under age 5.

e. Use the `anova` function to perform a likelihood ratio test of the null hypothesis that the effect of age does not vary with number of children less than 5. Carefully state (using model equations) the null and alternative hypothesis of this test, provide the test statistic, and a conclusion in context.

Problem 2

As an example of field observation in evidence of theories of sexual selection, S. J. Arnold and M. J. Wad ("On the measurement of natural and sexual selection: applications," *Evolution* (1984) Vol. 38: pp. 720-734) presented data on the size and number of mates observed in 38 male bullfrogs. Some output (from R) of a Poisson regression of number of mates on body size (in millimeters) is shown below. Body size is the size of the male bullfrogs. Note that some of the output has been omitted.

```
Coefficients:
              Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   -8.11831     2.58813  -3.137    0.00171
Bodysize       0.05723     0.01847

Null deviance:  39.956 on 37 degrees of freedom
Residual deviance:  28.003 on 36 degrees of freedom
```

a. The fitted Poisson loglinear model has the form

$$\log(\hat{\mu}(x)) = \hat{\alpha} + \hat{\beta}x.$$

Explain what $\hat{\mu}(x)$ and x represent in context of the data. Using the R output above, what are the values for $\hat{\alpha}$ and $\hat{\beta}$?

b. Use $\hat{\beta}$ to interpret the effect of body size on the mean number of mates and give an approximate 95% confidence interval for this effect. (Read carefully here; think about what is being modeled.)

c. Give an approximate 95% confidence interval for the effect of a 10mm increase in body size on the mean number of mates. Interpret this interval in terms of the problem.

Part II: Cite Sources

Write the sources you used to complete this assignment at the end of your homework submission, adhering to the "Guidance on Citing Sources" bullet points in the collaboration policy section on our course syllabus.