Stat 526 **Assignment #4**

Analysis of Binomial and Count Data

READING - Faraway Chapter 3, 4, and 5

0. (5 pts) Name

1. (13 pts) The data set esoph contains a case-control study of esophageal cancer. The three ordered factors are age (6 levels), tobacco consumption (4 levels), and alcohol consumption (4 levels).

a. Fit a binomial GLM that includes all two-factor interactions and use AIC to select the best model. State this model.

b. Convert your factors to numerical representation using the unclass function (e.g., you can centralize the numeric predictors by

```
age=unclass(agegp)-3.5; alc=unclass(alcgp)-2.5; tob=unclass(tobgp)-2.5;
```

). Then construct a simplified Binomial regression model using polynomials of these representations. [HINT: The the factors in the dataset are ordinal, thus R creates orthogonal polynomial rather than binary dummy variables. You can use the significant test results in part (a) to determine up to which polynomial degree you should use for the numerical regression.]

c. Commment on the fit of this model. Include plots to support your summary.

d. Construct a 95% confidence interval for the effect of moving one category higher in tobacco consumption.

e. Construct a 95% confidence interval for the effect of moving from the 45-54 age group to the 55-64 age group.[HINT: You may want to use the vcov(model) function to obtain the covariance matrix of $\hat{\beta}$]

2. (12 pts) Faraway Chapter 3 Exercise #4

3. (15 pts) Simulation study. Let's now look at the two goodness of fit measures under the binomial distribution. For each setting, you will generate $N = 1000$ data sets, fit each using the correct binomial logistic regression model, and save the deviance and Pearson X^2 measures. You will then compare the distribution of each measure against the appropriate asymptotic χ^2 density.

You will use the following model in all simulations: $y|x \sim \text{Binomial}(m, p)$ where $x \sim N(0, 1)$ and $\text{logit}(p) = 0.35 + x$. Consider the following nine settings:

1. $m = 5$,trials and $n = 50$ cases
2. $m = 15$ trials and $n = 50$ cases
3. $m = 45$ trials and $n = 50$ cases
4. $m = 5$ trials and $n = 200$ cases
5. $m = 15$ trials and $n = 200$ cases
6. $m = 45$ trials and $n = 200$ cases

7. $m = 5$ trials and $n = 800$ cases
8. $m = 15$ trials and $n = 800$ cases
9. $m = 45$ trials and $n = 800$ cases

For each setting and goodness-of-fit measure, create a histogram and overlay the appropriate χ^2 density. Summarize what you find, making sure to address whether the distributions of deviance and Pearson X^2 better fit the χ^2 density as the number of trials and/or sample size increases.

4. (5 pts) An experiment analyzes imperfection rates for two processes used to fabricate silicon wafers for computer chips. For treatment A applied to 10 wafers, the numbers of imperfections are 8, 7, 6, 6, 3, 4, 7, 2, 3, 4. Treatment B applied to 10 other wafers has 9, 9, 8, 14, 8, 13, 11, 5, 7, 6 imperfections. The following model was fit to the data: **R code and output:**

```
Call: glm(formula = imperfections ~ treatment, family = poisson, data = wafers)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.6094     0.1414  11.380  < 2e-16 ***
treatmentB    0.5878     0.1764   3.332 0.000861 ***
...
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 27.857  on 19  degrees of freedom
Residual deviance: 16.268  on 18  degrees of freedom
AIC: 94.349
```

a. State the model and the assumptions. Denote the expected number of imperfections in treatment A as μ_A, and the expected number of imperfections in treatment B as μ_B. Provide the numeric estimate of $\frac{\mu_B}{\mu_A}$ based on the model fit above. Interpret the estimate.

b. Test $H_0 : \frac{\mu_B}{\mu_A} = 1$ against $H_a : \frac{\mu_B}{\mu_A} \neq 1$ using both the Wald test and the likelihood ratio test.

c Construct a 95% confidence interval for $\frac{\mu_B}{\mu_A}$