Stat 526 **Assignment #3**

Analysis of Binary Data

READING - Faraway Chapter 2

0. (5 pts) Name

1. (5 pts) Consider a logit and probit model for binary data with one predictor. Show that if the predictor variable equals the negative value of the intercept (β_0) divided by the slope (β_1) then the probability is 1/2. This value of the predictor is often denoted the $LD50$. Find the $LD50$ for the complementary log-log link, i.e., $g(p) = \log(-\log(1-p))$.

2. (5 pts) Explain why $\sum Y_i = \sum \hat{p}_i$ in logistic regression.

3. (10 pts) Faraway Chapter 2 Exercise #5.

4. (15 pts) Simulation study. For this problem, you will demonstrate that the deviance and Pearson χ^2 goodness of fit measures are not accurate when there are no replicates.

For each setting, you will generate $N = 1000$ data sets, fit each using the correct binary 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: $y|x \sim$ Bernoulli(p) where $x \sim N(0,1)$ and logit(p)$= 0.35 + x$. Consider the following three settings:

1. $n = 50$ cases
2. $n = 200$ cases
3. $n = 800$ 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 sample size increases.

5. (10 pts) (Old Qual Exam) Results from a Copenhagen housing condition survey are compiled in an R data frame `housing1` consisting of the following components:

Infl Influence of renters on management: Low, Medium, High.
Type Type of rental property: Tower, Atrium, Apartment, Terrace.
Cont Contact between renters: Low, High.
Sat Highly satisfied or not: two columns of counts.

A model is fitted to the data using the following commands.

```
fit <- glm(Sat~Infl+Type+Cont,family=binomial,data=housing1)
```

A portion of the results from `summary(fit)` follow:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.6551     0.1374  -4.768 1.86e-06 ***
InflMedium      0.5362     0.1213   4.421 9.81e-06 ***
InflHigh        1.3039     0.1387   9.401  < 2e-16 ***
TypeApartment  -0.5285     0.1295  -4.081 4.49e-05 ***
TypeAtrium     -0.4872     0.1728  -2.820  0.00480 **
TypeTerrace    -1.1107     0.1765  -6.294 3.10e-10 ***
ContHigh        0.3130     0.1077   2.905  0.00367 **

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 166.179  on 23  degrees of freedom
Residual deviance:  27.294  on 17  degrees of freedom
AIC: 146.55
```

1. Write the assumptions of the model, and the expression of the log-likelihood.
2. According to the fitted model, what percentage of renters, who have low influence on management, live in apartment, and have high contact between neighbors, are highly satisfied?
3. Do people who live in apartments have a significantly different probability of satisfaction than people who live in atriums? The correlation between the respective coefficients is 0.494.
4. Estimate the odds ratio of high satisfaction for groups with high contact among neighbors over groups with low contact among neighbors, using a 95% confidence interval.