MA478 Generalized Linear Models (Spring 2024) **Name:** ____________________________

Midterm - 150 points

READ THESE INSTRUCTIONS CAREFULLY BEFORE YOU BEGIN.

1. This exam consists of this cover page and 8 pages of questions (total of 9 pages) worth a total of 150 points. You will have 75 minutes to complete this exam.
2. You are authorized to use your course notes and R/Rstudio (blank scripts only). You may NOT use any resources that are not on the authorized reference list, including computers, phones, the Internet, your textbook, and your classmates.
3. All work written on this exam will be graded unless it is clearly marked through. To receive full credit for your answer, you must show ALL mathematical work and provide explanations within the context of the associated research question.
4. Clearly indicate your final answer for questions that require calculations and **round all numbers to at least three significant digits**.
5. Use a blank continuation sheet and clearly identify that the problem is continued both on the exam and on the continuation sheet. Use one continuation sheet per problem continued. Be sure to put your name on each continuation sheet.
6. Cadets are **not** authorized to discuss the content, structure, or any other information about this exam until this exam has been released from academic security. Discussion includes all forms of written, electronic, and verbal communication.
7. Honor Acknowledgement Statement: Sign and date the statement below when you have finished the exam and are ready to submit it for grading.

"I did not use any sources nor did I receive any assistance while completing this exam. I will not discuss this exam with anyone until it is released from academic security on ____________________ at __________ hours."

____________________ ____________________ ____________________

Printed Name of Cadet Signature of Cadet Time and Date Signed

Question	1	2	3	Total
Points	55	70	25	150
Total				

Part 1 (55 pts)

> The Donner Party were a group of emigrants moving to start a new life in California. But between 1846 and 1847, 45 out of the 87 people on the wagon train would die from sickness, starvation, murder, and cannibalism. You conduct an analysis on the data and get the following output:

```r
library(tidyverse)
library(faraway)

donner_dat <- read.table("https://dnett.github.io/S510/Donner.txt",header=T)

donner_dat <- donner_dat %>% mutate(survive=ifelse(status=="DIED",0,1))

our_glm <- glm(survive~sex+age,data=donner_dat,
               family="binomial")


summary(our_glm)
```

```
##
## Call:
## glm(formula = survive ~ sex + age, family = "binomial", data = donner_dat)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.23041    1.38686   2.329   0.0198 *
## sexMALE     -1.59729    0.75547  -2.114   0.0345 *
## age         -0.07820    0.03728  -2.097   0.0359 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

P1.1 (20) Write the complete estimated regression equation of the model using the summary output ensuring you have properly identified the function, linear predictor and distribution of the data.

You next run the model:

```r
our_glm2 <- glm(survive~sex+age+sex:age,data=donner_dat,family="binomial")


summary(our_glm2)
```

```
##
## Call:
## glm(formula = survive ~ sex + age + sex:age, family = "binomial",
##     data = donner_dat)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.24638    3.20517   2.261   0.0238 *
## sexMALE     -6.92805    3.39887  -2.038   0.0415 *
## age         -0.19407    0.08742  -2.220   0.0264 *
## sexMALE:age  0.16160    0.09426   1.714   0.0865 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 47.346  on 41  degrees of freedom
## AIC: 55.346
##
## Number of Fisher Scoring iterations: 5
```

P1.2 (10) Based on all the information given, which model would you prefer and why? To answer this question, perform a statistical test, ensure you give the test statistic and the distribution of that test statistic.

P1.3 (5) According to `our_glm` how does the person's age impact the odds that they survived?

P1.4 (10) Assuming you wanted to compare `our_glm` with a model that was not nested explain how you could do this WITHOUT relying on AIC or BIC.

P1.5 (10) Your officemate states that you should conduct a goodness of fit test by testing the deviance of the model and runs the test:

```r
1-pchisq(51.256,41)
```

```
## [1] 0.1308893
```

Is this a correct approach? If no, provide an alternative approach. If it is a correct approach provide a conclusion. Note here if you provide an alternative approach you do not have to actually carry out the test.

Part 2 (70 pts)

You are interested in exploring factors that impact the number of burglaries in Chicago so you collect data on 552 different city blocks and count the number of burglaries that occur over a month. You also collect data on the percent of the population that is unemployed and the average salaries on the block. In this class we have discussed at least four different models that could be used to analyze this data. In particular, you could use a negative binomial distribution, a Poisson distribution, a Quasi-Poisson, or a zero inflated Poisson.

P2.1 (30) Discuss how you would go about picking between these four models. Give examples of when each of them would be appropriate.

P2.2 (15) Your friend decides to fit the following model, write out the model that they are fitting and explain what issues they may have fitting this model:

```
chi_df <- read.csv("chi_burg.csv")

chi_mod <- glm(burglaries ~ unemployment + wealth, offset=log(population),
            family=poisson(link="identity"),
            data=chi_df)
```

P2.3(15) You fit the model below and run the below lines of code to assess your model. Explain what you are checking in the model, what your findings suggest, and what you would do next.

```r
chi_mod <- glm(burglaries ~ unemployment + wealth, offset=log(population),
            family=poisson,
            data=chi_df)

sum(residuals(chi_mod,type="pearson")^2)/chi_mod$df.residual
```

```
## [1] 1.332808
```

```r
1-pchisq(deviance(chi_mod),df.residual(chi_mod))
```

```
## [1] 1.746496e-10
```

P2.3(10) Your roommate has heard about quasi-Poisson models and decides to a quasi-Poisson model to the data, they argue that they can use AIC to compare their model to your Poisson regression model. Are they correct? Why/why not?

P3 (25 Pts) For the distribution listed below:

- Show that it is part of the exponential dispersion family
- Identify the canonical parameter θ
- Show that the expected value is μ and find the variance function (in terms of μ)

$$f(y|\mu, \lambda) = \left(\frac{\lambda}{2\pi y^3}\right)^{1/2} \exp\left(\frac{\lambda}{2\mu^2}\frac{(y-\mu)^2}{y}\right)$$

HINT: Let $\phi = \frac{1}{\lambda}$ and $b(\theta) = -\sqrt{-2\theta}$