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

```
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.

Here we can write:

$$
\begin{aligned}
i &= \text{person} \\
Z_i &= 1 \text{ if survived, } 0 \text{ if died} \\
Z_i &\sim Bern(p_i) \\
\eta_i &= \log(\frac{p_i}{1-p_i}) \\
\eta_i &= \beta_0 + \beta_1 x_{male} + \beta_2 x_{age}
\end{aligned}
$$

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.

As the models are nested we can use the change in deviance divided by the change in degrees of freedom (likelihood ratio test) to compare. Here our test statistic has 1 degree of freedom as we only have one additional parameter in our second model.

$$\chi^2 = \frac{\Delta \text{Deviance}}{\Delta \text{DF}}$$
$$\chi^2 \sim \chi^2_1$$

In our case this is equal to:

```r
chi_sq = 51.256-47.346
1-pchisq(chi_sq,1)
```

```
## [1] 0.0479996
```

Therefore, we likely would conclude that the second model fits the data better, hence we should include the interaction term.

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

This is relatively straightforward, we need to exponentiate the coefficient for age which is:

```
exp(-0.0782)
```

```
## [1] 0.9247795
```

Meaning, given a person's gender, for every year older an individual is their odds of surviving change by a factor of 0.925

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.

A few different ways to answer this, but I think the most straightforward would be to create a training set and a test set from the data, fit both models to the training set, then find the AUC, or percent misclassified, on the test set.

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:

```
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.

Unfortunately I put the wrong degrees of freedom in this, so if you identified that the df was wrong I gave you full credit. HOWEVER, the point of this is that we cannot rely on comparing our model to the saturated model using deviance for ungrouped data. For this question it would be better to conduct the Hosmer Lemeshow test or to create groups somehow in the data.

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.

Let's start with the Poisson. Which is a very good place to start.... We would use the Poisson if our data were counts and we did not have any grouping. In general, we can think of the Poisson as the Binomial model when n gets really big but p is small. This seems reasonable in the case of burglaries as we don't think fo the number of houses out of XXX having been burgled. However, the drawback to the Poisson is that the mean and variance have a 1 to 1 relationship. The Poisson should probably be our default starting point for this example.

The next most complex model is probably the negative binomial distribution. We can think of the negative binomial as arising when the rate parameter of the Poisson has a gamma distribution. This means that we would expect the mechanism of interest, or λ, to manifest differently in different observations even if the covariates are the same. This model also allows for over dispersion and can be considered when our Poisson does not fit the data well.

The ZIP model has assumes there are two mechanisms that could potentially exist in our observations. Here we can think of either zeros arising because the area does not get burgled, or perhaps a zero arrises because a place is burgled but not reported to the police. This might be a reasonable asssumption and we can fit this model if the number of zeros in our dataset far exceed the number we would expect under a Poisson regression.

Likely the most complex model is the quasi-Poisson which allows for a free ϕ parameter changing the variance to mean ration of the Poisson. This would likely be appropriate if our data are overdispersed and we cannot think of any other way to fix it. The draw back for a Q-P is we don't necessarily know the distribution our data came from so we cannot generated new data from the model. Also, the confidence intervals for our covariates will likely be larger than the CIs from the Poisson.

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)
```

This model assumes that there is a linear relationship between unemployment, wealth and the expected number of burglaries. While on the surface this might make sense, we see some issues if we write out the model:

$$\begin{aligned}
y_{ij} &= \text{Number of Burglaries at city block i during month j} \\
y_{i,j} &\sim Po(\lambda_{i,j}) \\
\lambda_{i,j} &= \beta_0 + \beta_1 x_{unemp,i} + \beta_2 x_{wealth,i} + \log(Pop_i)
\end{aligned}$$

The first thing we note that is that having a log offset doesn't really make sense in this model. Recall that the log offset came from assuming we had a log link and were interested in the number of events per person (which might still be of interest here, but we probably shouldn't structure the model like this). This is a relatively minor point. The major point, though, is that λ must be a positive number but there's nothing constraining our β terms to be positive. So therefore, we could get estimates of β that result in a negative λ.

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.

```
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
```

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

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

Looking at the output we see two things. The first is that our estimate of ϕ actually isn't too bad (it's relatively close to 1) but we fail our goodness of fit test. Therefore, we likely shouldn't jump to a Quasi-Poisson or a model accounting for over dispersion as we don't yet have evidence that over-dispersion is an issue. So, first we should consider if we are missing any other covariates in our model that may be of value in explaining the variance in our data. This likely is the case here as burglaries occur for a wide range of reasons outside of unemployment and wealth. We also would want to look at the Pearson residuals to see if we have any extreme outliers. Finally we would want to check the number of zeros and see if this is an issue compared to what we would expect.

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?

Recall that AIC is based on likelihood. As quasi-Poisson are fit via quasi-likelihood we cannot compute AIC. Nor could we compute a likelihood as we don't really know what distribution the data arose from.

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 $b(\theta) = -\sqrt{-2\theta}$

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

From here we let:

$$\begin{aligned} \phi &= \frac{1}{\lambda} \\ \theta &= \frac{-1}{2\mu^2} \\ b(\theta) &= -\sqrt{-2\theta} \\ c(\phi, y) &= \frac{1}{2}\log\lambda - \frac{1}{2}\log(2\pi y^3) + \frac{\lambda}{2y} \end{aligned}$$

Therefore, we have:

$$f(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

As desired. Note from above this implies $\mu = \sqrt{-\frac{1}{2\theta}}$. The canonical parameter is $\theta = \frac{-1}{2\mu^2}$.

To find the expected value we compute $b'(\theta)$ which is:

$$\begin{aligned} b'(\theta) &= \frac{1}{\sqrt{-2\theta}} \\ &= \frac{1}{\sqrt{-2\frac{-1}{2\mu^2}}} \\ &= \mu \end{aligned}$$

To find the variance function we need $b''\theta$ which is:

$$\begin{aligned} b''(\theta) &= (-2\theta)^{-3/2} \\ &= (-2\frac{-1}{2\mu^2})^{-3/2} \\ &= (\mu^2)^{3/2} \\ &= \mu^3 \end{aligned}$$

Therefore, the variance to mean ratio implied by this distribution is μ^3