MA478 Generalized Linear Models (Spring 2024)
Midterm - 150 points

Name: Joshua Wang

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 06MAR at 1700 hours."

Joshua Wang		0509UU MAR2024
Printed Name of Cadet	Signature of Cadet	Time and Date Signed

Question	1	2	3	Total
Points	55	70	25	150
Total	53	62	22	137

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.

$Y_i \sim \text{Binomial}(87, p_i)$ i- our person in the dataset ungrouped

$\log\left(\frac{p_i}{1-p_i}\right) = 3.23041 - 1.59729 x_1 - 0.07820 x_2$

$Y_i \sim Bin(1, p_i)$ or $Ber(p_i)$

$x_1 = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$

x_2 = age in years

+18

You next run the model:

```
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.

small mod large mod
1 - pchisq(51.256 - 47.346, 1) = 0.048

test stat: 51.256 - 47.346 = 3.91

distribution: χ^2_1

I would prefer the smaller model our_glm, at the 0.01 significance level as at this significance level I would fail to reject the null hypothesis that smaller model is better.

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

A one unit increase in age is associated with a 0.078 decrease in the log odds of survival, or a exp(-0.078) change in odds or 0.924 decrease in odds of survival.

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.

You could split the dataset into a train and test set and look at the F1 score of the train test samples to compare models

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.

If the given test is on our_glm, then you need to test the residual deviance, 51.256 and the residual df for the model which is 42. Then the goodness of fit test for our glm would be 1-pchisq(51.256, 42), to get p-value.

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.

I would start off with using a poisson distribution model, as it is the simplest and most interpretable out of the four models. I would then conduct a goodness of fit test. If the poisson model fails the goodness of fit test, I would then look at the data for excess outliers and the general distribution of the response variable. I would also look at the poisson's model's estimation of ϕ. If the model does not pass the GOF test and the value of $\phi > 1$ may the model is overdispersed, I would consider / try a NB distribution.

NB would give me more room for variance, as this can the value does not have to be equal to the mean. Additionally in the context of the problem, it is reasonable that the λ parameter of the poisson model can form a gamma distribution.

Based on the problem I would not immediately jump to a ZIP model unless I see a lot of zero values in the response. Otherwise than, as of now I cannot conceptualize a reasonable mechanism to explain why there would be many zeros, and w/o that I would avoid a ZIP.

Finally, I would only use the quasi-poisson test if the NB model failed poorly and does not change very much when NB compared to the poisson model. I would use quasi-poisson as a last resort to report my CI again to try to fit the data accurately. I would just as I cannot rely on MLE estimates for my β.

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)
```

$$\log\left(\frac{\text{burglaries}}{\text{population}}\right) = \beta_0 + \beta_1(\text{unemployment}) + \beta_2(\text{wealth})$$

this is log-link

$$\log(\text{burglaries}) = \beta_0 + \beta_1(\text{unemp}) + \beta_2(\text{wealth}) + \log(\text{population})$$

The issue of fitting this model is that the offset way to measure the burglaries per population does not make sense with respect to the problem.

+7

identity is

$$\lambda = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \log(P \cdot p)$$

issue is $\beta \in (-\infty, \infty)$ but

$$\lambda \in (0, \infty)$$

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
```

$\phi \approx 1$

GOF test

Ho: Model is not a bad fit for the data,

I see that the model does not pass the GOF test, however it's estimate of ϕ (how our relationship between mean and variance) is close to 1 may that the data is not massively over dispersed.

What I would do next is try to train a new model with more variables to see a better fit for the data. Possibly I could add population as a covariate rather than an offset. I would then test to see if the fit is better.

However because ϕ estimate is slightly over 1, it is only slightly overdispersed, while I would not particularly consider a new model such as a quasi or NB. I would try them if there is no more covariates to consider.

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?

You cannot use AIC to compare a quasi-Poisson to a poisson model. The formula for AIC involves calculating the likelihood of each model. However you cannot calculate a likelihood for a quasi-poisson the same way you would for a Poisson model. You need to calculate quasi-likelihood which cannot be compared to a Poisson model's normal likelihood.

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

$$f(y_i|\theta_i,\phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right)$$

$$f(y,\mu,\lambda) = \exp\left(\frac{1}{2}\log\left(\frac{\lambda}{2\pi y^3}\right) + \left[\frac{\lambda}{2\mu^2}\frac{(y-\mu)^2}{y}\right]\right)$$

$$\frac{1}{2}\log(\lambda) - \frac{1}{2}\log(2\pi y^3) + \frac{\lambda}{2\mu^2}\left[\frac{y^2}{y} + \frac{2y\mu}{y} - \frac{\mu^2}{y}\right]$$

$b(\theta) = -\sqrt{-2\theta}$ $\quad \theta = -\frac{1}{2}\mu^2$ $\quad \phi = \frac{1}{\lambda}$ $\quad \theta = \frac{1}{2\mu^2}$

$b'(\theta) = E[Y]$

$\frac{d}{d\theta}\left[-\sqrt{-2\theta}\right]$

$\frac{1}{2}\frac{1}{d\theta}\left(-(-2\theta)^{-\frac{1}{2}}\right)$

$-\frac{1}{2}(-2\theta)^{-\frac{3}{2}} \times -2$

$b'(\theta) = (-2\theta)^{-\frac{3}{2}}$

$= \left[-2\left(-\frac{1}{2}\mu^2\right)\right]^{-\frac{3}{2}}$ Follows

$= (\mu^2)^{-\frac{3}{2}} = \frac{1}{\mu^3}$? μ^3

$\frac{\lambda y}{2\mu^2} + \frac{2\mu\lambda}{2\mu^2} - \frac{\lambda\mu^2}{2\mu^2 y} \rightarrow$

$\frac{\lambda y}{2\mu^2} + \frac{\lambda}{\mu} - \frac{\lambda}{2y}$

$\frac{\frac{y}{2\mu^2} + \frac{1}{\mu} - \frac{1}{2y}}{\frac{1}{\lambda}}$

$\phi = \frac{1}{\lambda}$ ✓ close

$\frac{1}{\mu}$

$\frac{y}{2\mu^2} + \frac{1}{\mu}$ $\quad$ $\frac{1}{\lambda}$

$\frac{1}{2\mu^2}y = \left(-\frac{1}{\mu}\right)$ $\quad$ $\frac{1}{2}$

$b(\theta) = -\frac{1}{\mu}$

$L(\theta) = -\frac{1}{\mu}$

$\theta = \frac{1}{2\mu^2}$

$b(\theta) = -\sqrt{-2\left(\frac{1}{2}\mu^2\right)}$

$-\sqrt{-\mu^2}$ $\quad (\mu^2)^{\frac{1}{2}}$