STA 214
Fall 2022

Exam 2

Dr. Evans
Exam 2

Last Name: _____ **First Name:** _____

I hereby state that I have not communicated with or gained information in any way from other students or any outside resource during this exam. I agree to abide by the rules stated below, and to abide by the Wake Forest Honor Code. All work is my own. I understand that any violation of this agreement will be reported to the Honor Council and will result, at minimum, in a 0 on this exam.

Signature : _____

All work on this exam must be your own.

1. You have 50 minutes to complete the exam.
2. Show all your work on the open ended questions in order to get partial credit. No credit will be given for open ended questions where no work is shown, even if the answer is correct.
3. You are allowed a calculator, however you may not share a calculator with another student during the exam. The calculator must be only a calculator, and may not be connected to the internet.
4. You are allowed to ask clarification questions to me, but you may not ask anyone else.
5. You are **not** allowed a cell phone, even if you intend to use it as a calculator or for checking the time. You are **not** allowed a music device or headphones, notes, books, or other resources.
6. You may **not** communicate with anyone other than myself during the exam.
7. Write clearly and be clear. Make it easy to find your answers.

Good luck!

The Data

Your friend runs a bookstore in the local mall, and is interested in predicting how many books they will sell, and which types of books will be most popular with their customers. We have data on how many books were purchased from the store in the last 30 days. Each row in the data represents one book, with the following variables recorded:

- title: The title of the book
- author: The author of the book
- rating: An average score the book has received on Amazon.
- purchases: The number of copies of the book purchased in the last 30 days.
- price: The price of the book in US. Dollars.
- publisher: The company that published the book.
- page_count: The number of pages in the book.
- age: how many years ago the book was first published
- genre: the book's genre (Fiction, Fantasy, Mystery, Business, General Interest, Comics and Graphic Novels, or Other).

Research question 1: When your friend gets new books in the store, they need to categorize the new books by genre. However, determining the genre of a book can be challenging. To save time, your friend asks you to build a regression model to predict a book's genre from its rating, purchases, price, page count, and age.

---

1. (3 pts) Which type of model (logistic regression, multinomial regression, Poisson regression, quasi-Poisson regression, negative binomial regression, or a ZIP model) should you use when your response Y_i is genre? Explain your reasoning, and write down the distribution for this response variable (Step 1 of the parametric modeling process).

You fit a model with genre as the response and rating, purchases, price, page count, and age as explanatory variables. You may assume all assumptions are met, and no transformations are necessary. Here is the R output for your fitted model:

Model 1:

```
Coefficients:
                   (Intercept)         age    page_count         price     rating     purchases
General              2.815334  0.02238879  0.0009943770 -0.0124937221 -0.9970760 -2.782382e-04
Fantasy            -11.344759  0.02605114  0.0011837241 -0.0208921547  2.0684096  4.299147e-04
ComicsGraphicNovel   6.448461 -0.17070209 -0.0137132907  0.0007489826 -1.0034271  3.977844e-04
Business             2.786146  0.03669941 -0.0023766062  0.0054251860 -1.1447918  4.135807e-04
Mystery              6.183348  0.08923819  0.0005447292 -0.0252044648 -1.9463305  1.974567e-05
Fiction             -1.080619 -0.10781100  0.0008696791 -0.0136541880  0.1746037  2.242896e-04

Std. Errors:
                   (Intercept)        age   page_count       price    rating    purchases
General             0.02785340 0.05086547 0.0006411061 0.008569520 0.1322861 0.0004067197
Fantasy             0.02786519 0.05769481 0.0006192579 0.011082734 0.1337849 0.0001912607
ComicsGraphicNovel  0.02610947 0.08844962 0.0030997710 0.010129917 0.1335700 0.0002817135
Business            0.03667764 0.06449447 0.0018710676 0.006760399 0.1804858 0.0001781087
Mystery             0.04182181 0.05320238 0.0011846022 0.013857287 0.1891294 0.0003244738
Fiction             0.02271761 0.05872766 0.0006151048 0.007621892 0.1107418 0.0002179117
```

---

2. (2 pts) Which genre category is being treated as the reference category in this model?

3. (2 pts) Interpret the estimated coefficient 0.0224 (first row, second column of the `Coefficients` output).

4. (4 pts) Calculate the estimated relative risk of Fantasy vs. Mystery for a 3 year old book that is 250 pages long, costs $15, received a rating of 4.5 on Amazon, and was purchased 100 times in the last 30 days.

To assess the ability of your model to predict book genre, you create the following confusion matrix:

	actual						
predicted	Other	General	Fantasy	ComicsGraphicNovel	Business	Mystery	Fiction
Other	97	18	16	9	11	12	27
General	0	0	0	0	0	0	0
Fantasy	2	0	2	0	0	0	1
ComicsGraphicNovel	7	0	0	13	2	0	2
Business	2	0	0	0	2	0	0
Mystery	1	2	0	0	0	0	0
Fiction	0	1	0	0	0	0	1

---

5. (5 pts) Is the model doing a good job predicting genre? Your answer should include at least one summary measure of the confusion matrix, and should discuss which genres we predict well, and which genres we predict poorly.

Now your friend is interested in predicting which books will sell next month.

Research question 2: Your friend asks you to build a regression model to predict the number of copies sold, using genre, age, page count, price, and rating.

To answer this research question, you fit a Poisson regression model, producing the following output (you may assume all regression assumptions are met):

Model 2:

```
glm(formula = purchases ~ genre + age + page_count + price +
    rating, family = poisson, data = books)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               7.057e+00  4.331e-02 162.942   <2e-16 ***
genreGeneral             -2.350e-01  1.412e-02 -16.640   <2e-16 ***
genreFantasy              1.329e+00  8.599e-03 154.541   <2e-16 ***
genreComicsGraphicNovel   7.473e-01  1.227e-02  60.911   <2e-16 ***
genreBusiness             1.295e+00  8.829e-03 146.641   <2e-16 ***
genreMystery             -1.262e-01  1.497e-02  -8.433   <2e-16 ***
genreFiction              6.458e-01  1.008e-02  64.072   <2e-16 ***
age                       1.087e-01  4.697e-04 231.408   <2e-16 ***
page_count                4.567e-04  4.972e-06  91.849   <2e-16 ***
price                     1.263e-03  9.761e-05  12.934   <2e-16 ***
rating                   -5.858e-01  9.225e-03 -63.504   <2e-16 ***

    Null deviance: 338094  on 227  degrees of freedom
Residual deviance: 241976  on 217  degrees of freedom
```

6. (3 pts) Perform a goodness of fit test for the fitted Poisson regression model. Report the test statistic, the distribution you will use to calculate a p-value, and the degrees of freedom for that distribution. *You do not need to actually calculate the p-value!*

7. (2 pts) One reason your Poisson model may not be a good fit to the data is overdispersion. You estimate the dispersion parameter, and your estimate is $\widehat{\phi} = 1956.3$. Interpret the estimated dispersion parameter 1956.3. What does this number represent, and do you think there is overdispersion in the data?

8. (3 pts) To account for overdispersion, we can fit a quasi-Poisson model instead of a Poisson model. The output for the quasi-Poisson model is shown below, but the estimated coefficient and standard error for `rating` are missing. Fill in the missing coefficient and standard error.

```
Coefficients:
                          Estimate Std. Error
(Intercept)               7.0570182  1.9155904
genreGeneral             -0.2350004  0.6246327
genreFantasy              1.3289646  0.3803484
genreComicsGraphicNovel   0.7473223  0.5426593
genreBusiness             1.2947517  0.3905206
genreMystery             -0.1262199  0.6620000
genreFiction              0.6458380  0.4458275
age                       0.1086846  0.0207732
page_count                0.0004567  0.0002199
price                     0.0012625  0.0043174
rating
```

9. (4 pts) Calculate a 95% confidence interval for the change in the average number of books sold associated with an increase of 1 unit in `rating`, holding other variables constant. (The critical value is $z^* = 1.96$)

Finally, your friend is interested in predicting how many books a visitor to the mall will purchase from your friend's store. To answer this question, they collect a new set of data. Each visitor to the mall is asked, as they leave the mall, how many books they purchased from the bookstore. They are also asked about their other shopping habits, and the following variables are recorded. Sadly, your friend did not think to ask the customers whether they actually visited the bookstore!

- books: the number of books purchased at the bookstore
- cost: the total amount of money the visitor spent at the mall that day (in US dollars)
- coffee: whether the customer bought a coffee from the Starbucks in the mall (0 = no, 1 = yes)
- clothing: whether the customer purchased any clothing from the mall (0 = no, 1 = yes)
- movie: whether the customer watched a movie at the mall's movie theater (0 = no, 1 = yes)

Research question 3: Your friend asks you to build a regression model to predict the number of books purchased from their bookstore, using the total amount of money the visitor spent at the mall, whether the visitor bought coffee, whether the visitor bought clothing, and whether the visitor saw a movie.

---

10. (2 pts) You propose using a ZIP model for this data. Explain to your friend why a ZIP model is appropriate here, and what the latent variable Z_i represents. (In other words, there are two groups in the data – what does $Z_i = 0$ and $Z_i = 1$ indicate?)

To address the research question, you fit a ZIP model with `books` as the response, and the other variables as explanatory variables. You may assume that all regression assumptions are met. The output of your fitted model is shown below:

```
zeroinfl(formula = books ~ cost + coffee + clothing + movie |
          cost + coffee + clothing + movie, data = mall)

Count model coefficients (poisson with log link):
                        Estimate
(Intercept)             -1.510
cost                     0.011
coffee                   0.340
clothing                 0.207
movie                   -0.025


Zero-inflation model coefficients (binomial with logit link):
                        Estimate
(Intercept)              0.769
cost                     0.006
coffee                   1.202
clothing                 0.317
movie                   -0.503
```

---

11. (4 pts) What is the probability of purchasing *at least* one book from your friend's bookstore, for a visitor who spent \$20 at the mall and bought a coffee, but did not buy clothing and did not see a movie?

You are done!!! Whooo!!!