STA 214
Fall 2022

Exam 1

Dr. Evans
Exam 1

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

You are contacted by the US Small Business Administration (SBA), a government agency dedicated to helping support small businesses. The SBA provides loans to small businesses, but some businesses *default* on their loan (i.e., fail to pay it back). Researchers at the SBA are interested in predicting whether a business will default on the loan, and they have collected a random sample of 998 different loans. For each loan, we have the following variables:

- *Default*: Whether the business defaulted (0 = no, 1 = yes)
- *Amount*: The amount (in US dollars) of the original loan
- *NewBusiness*: Whether the business receiving the loan is a new business (1) or an existing business (0)

Research question: Researchers at the SBA want you to investigate the following question: *Is there a relationship between loan amount and whether a business defaults on the loan, after accounting for the age of the business (whether it is new or not).*

---

1. (3 pts) We will begin by investigating the relationship between loan amount and loan default (we will add NewBusiness to the model later). Choose *one* of the following population models for the relationship between loan amount and loan default, and explain your choice. (Let Y_i denote whether the business defaulted on the loan, and assume that the shape/linearity assumption is satisfied).

   (a) $Y_i = \beta_0 + \beta_1 \ Amount_i + \varepsilon_i$

   (b) $Y_i \sim N(\pi_i, \sigma^2) \qquad \pi_i = \beta_0 + \beta_1 \ Amount_i$

   (c) $Y_i \sim Bernoulli(\pi_i) \qquad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \ Amount_i + \varepsilon_i$

   (d) $Y_i \sim Bernoulli(\pi_i) \qquad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \ Amount_i$

   (e) $Y_i \sim Bernoulli(\pi_i) \qquad \pi_i = \beta_0 + \beta_1 \ Amount_i$

We fit the model from Question 1, and obtain the following output:

Model 1:

```
              Estimate Std. Error
(Intercept) -1.160e-01  2.558e-01
Amount      -5.713e-06  1.010e-06
---

    Null deviance: 923.58  on 997  degrees of freedom
Residual deviance: 889.82  on 996  degrees of freedom
```

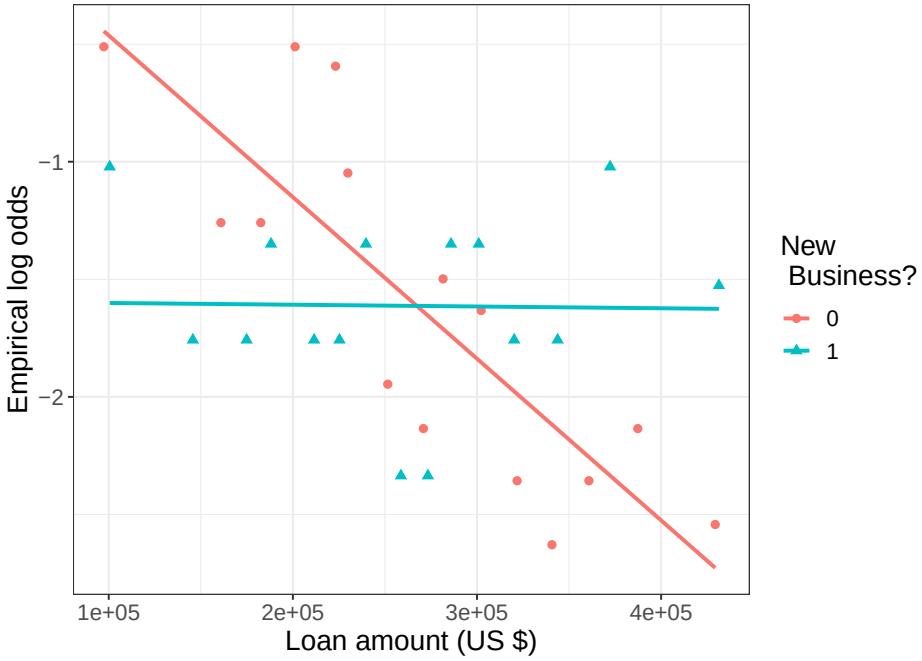
---

2. (3 pts) Based on the output above, write down the equation of the fitted model.

3. (3 pts) Interpret the slope of the fitted model in terms of the odds.

4. (3 pts) What is the predicted probability that a business will default on a loan of $200,000?

To answer the research question, we need to include the NewBusiness variable to the model. Below is an empirical log odds plot showing the relationship between loan amount and whether the business defaulted, with a separate line fit for new and existing businesses.



---

5. (4 pts) Using the empirical log odds plot, write down a population model that allows us to investigate the research question. Use appropriate notation, and explain your reasoning for choosing the model.

You fit the model from Question 5, and two other models as well. The output for these three models is shown below, in no particular order:

Model 2:

```
glm(formula = Default ~ NewBusiness, family = binomial, data = sba_data)

            Estimate Std. Error
(Intercept) -1.50729    0.09408
NewBusiness -0.21249    0.20397
---

    Null deviance: 923.58  on 997  degrees of freedom
Residual deviance: 922.46  on 996  degrees of freedom
```

Model 3:

```
glm(formula = Default ~ Amount + NewBusiness, family = binomial,
    data = sba_data)

              Estimate Std. Error
(Intercept) -3.026e-02  2.630e-01
Amount      -5.796e-06  1.012e-06
NewBusiness -2.870e-01  2.076e-01
---

    Null deviance: 923.58  on 997  degrees of freedom
Residual deviance: 887.84  on 995  degrees of freedom
```

Model 4:

```
glm(formula = Default ~ Amount * NewBusiness, family = binomial,
    data = sba_data)

                     Estimate Std. Error
(Intercept)         3.257e-01  2.918e-01
Amount             -7.286e-06  1.166e-06
NewBusiness        -1.963e+00  6.443e-01
Amount:NewBusiness  6.961e-06  2.440e-06
---

    Null deviance: 923.58  on 997  degrees of freedom
Residual deviance: 879.84  on 994  degrees of freedom
```

6. (4 pts) The SBA wants to identify loans for which the predicted probability of default is at least 0.3. For *existing* businesses, which loan amounts correspond to a predicted probability of at least 0.3? Use the fitted model corresponding to your population model from Question 5.

7. (3 pts) Now let's address the original research question. We want to test whether there is a relationship between loan amount and whether a business defaults on the loan, after accounting for whether the business is new. Using your model from Question 5, write down null and alternative hypotheses, in terms of one or more model parameters, for this research question.

8. (3 pts) Which test (Wald, Likelihood ratio, or both) could you use to test your hypotheses from Question 7? Explain your answer.

9. (4 pts) Using the output from Model 1 and/or Model 2 and/or Model 3 and/or Model 4, calculate a test statistic for your hypotheses in Question 7. What distribution will you compare your test statistic with to calculate a p-value (give the name and degrees of freedom, if applicable, for the distribution)?

Part II: Maximum likelihood estimation

(This question is unrelated to the SBA data or the previous questions)

Suppose we are told that we have a certain random variable Y, which can take three possible values: $Y = -1$, 0, or 1. We are told that the probability for each outcome is

- $P(Y = 0) = \pi_0$
- $P(Y = -1) = 5\pi_0$
- $P(Y = 1) = 1 - 6\pi_0$

where the parameter π_0 is unknown. We observe data $-1, 0, 0, 1, 1, -1, 0$, and we want to estimate π_0 from the data.

---

10. (6 pts) Using the method of maximum likelihood, solve for the maximum likelihood estimate of π_0. Show your work.

You are done!!! Whooo!!!