STA 712 Exam 1

Due: Friday, October 7, 12:00pm (noon) on Canvas.

Instructions: You have until the due date to complete this exam. There are no other time restrictions.

Submit your work as a single PDF. For this exam, you may include written work by scanning it and incorporating it into the PDF. Include all R code needed to reproduce your results in your submission.

For this exam, you may:

- Use any resources from the course (the textbook, the course website, class notes, previous assignments, etc.)
- Use the internet for R help
- Email me, or come to office hours, with specific questions (I may be somewhat less helpful than for regular assignments)

You may *not*:

- Discuss the exam with anyone else
- Use the internet or other textbooks (except for R help and to access the course materials)

Part I: Modeling an Exponential response

In this section, you will model a response variable $Y_i \sim Exponential(\lambda_i)$. Recall that the pdf for an exponential distribution with parameter λ_i is $f(Y_i; \lambda_i) = \frac{1}{\lambda_i} e^{-Y_i/\lambda_i}$. The mean is $\mathbb{E}[Y_i] = \lambda_i$, and the variance is $Var(Y_i) = \lambda_i^2$. You may use these facts without proof.

1. Suppose that Y_i is a response variable of interest, and we use the following model for Y_i:

$$Y_i \sim Exponential(\lambda_i)$$
$$\frac{1}{\lambda_i} = \beta^T X_i,$$

   where $X_i = (1, X_{i,1}, ..., X_{i,k})^T \in \mathbb{R}^{k+1}$, and $\beta = (\beta_0, ..., \beta_k) \in \mathbb{R}^{k+1}$ is the vector of regression coefficients. We observe data $(X_1, Y_1), ..., (X_n, Y_n)$.

   Calculate the score $U(\beta)$ and the Fisher information $\mathcal{I}(\beta)$.

2. Now we apply the model from Question 1 to real data. A factory is interested in the relationship between the amount of stress applied to a piece of steel, and the time it takes until that steel breaks. We use the following model:

$$time_i \sim Exponential(\lambda_i)$$
$$\frac{1}{\lambda_i} = \beta_0 + \beta_1 stress_i$$

   The raw data contains $n = 40$ observations $(stress_1,\ time_1), ..., (stress_{40},\ time_{40})$.

   You can load the data into R by

```
steel <- read.csv("https://sta712-f22.github.io/exams/steel.csv")
```

   In this question, we want to begin Fisher scoring to estimate β_0 and β_1 using the observed data. Our starting values are $\beta^{(0)} = (1, 0)^T$.

   (a) Using the results of Question 1, calculate $U(\beta^{(0)})$ and $\mathcal{I}(\beta^{(0)})$.

   (b) Use Fisher scoring to calculate the first updated estimates, $\beta^{(1)}$. (You only need to do one iteration of scoring here; you do not need to continue until convergence).

Part II: Logistic regression with earthquake data

In the second part of this exam, you will work with a dataset from DrivenData, an online data competition site that hosts competitions aimed at improving education, health, safety, and general well being for individuals around the world.

Our data come from the 2015 Gorkha earthquake in Nepal. After the earthquake, a large scale survey was conducted to determine the amount of damage the earthquake caused for homes, businesses and other structures. This is one of the largest post-disaster surveys in the world, and researchers are interested in which building characteristics are associated with earthquake damage.

You will work with a subset of the earthquake data, consisting of 211774 buildings, containing the following variables:

- `Damage`: whether the building sustained any damage (1) or not (0)
- `Age`: the age of the building (in years)
- `Surface`: a categorical variable recording the surface condition of the land around the building. There are three different levels: `n`, `o`, and `t`. (The researchers who collected the data anonymized the level names to protect inhabitants' privacy).

You can load the data into R by

```
earthquake <- read.csv("https://sta712-f22.github.io/exams/earthquake_small.csv")
```

You will work with the following logistic regression model (you may assume all assumptions are met; no transformations or diagnostics are needed):

$$Damage_i \sim Bernoulli(p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Age_i + \beta_2 SurfaceO_i + \beta_3 SurfaceT_i + \beta_4 Age_i \cdot SurfaceO_i + \beta_5 Age_i \cdot SurfaceT_i$$

where $SurfaceO$ and $SurfaceT$ are indicator variables for whether surface is o or t, respectively.

3. (a) Fit the logistic regression model in R, and interpret the estimated slope $\widehat{\beta}_1$ in terms of the *odds* of damage.
   (b) Calculate the estimated probability of damage for a 50 year old building with surface condition = t.
   (c) Calculate a 99% confidence interval for the probability in (b).
4. The researchers want to know whether the relationship between Age and the probability of damage is the same for buildings in all three surface conditions. Use a hypothesis test to address the researchers' question; you should state the hypotheses in terms of one or more model parameters, calculate a test statistic and p-value, and make a conclusion.
5. Create and interpret a 95% confidence interval for the difference in *odds* between surface `o` and surface `t`, for a 25 year old building.