Final Project Instructions

- Output for the project will be written in RMarkdown and submitted as a .rmd file
- Submit the .rmd file and any additional scripts in a .zip file
- Provide well-documented commentary on your code throughout your project & generally follow good programming practices
- Ensure your final report looks professional, is designed well, etc.
- Hide your code in your final Markdown document
- Projects are due December 10, 2025 at 5pm EST. Feel free to submit your project early!

Part 1: Background

Generalized linear models (GLM) is a term in statistics that describes a large family of regression models that are extensions of the ordinary linear regression model. GLMs characterize a linear relationship between response and predictor variables where the underlying distribution of the response variable is not normally distributed. In order to use GLMs appropriately, you must know the underlying distribution of the Y variable. We will explore the impact of model misspecification - that is, what happens if we use the wrong distribution.

We will simulation data from a Poisson distribution. The pdf of a Poisson distribution is given by:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

We will examine results from Poisson and Normal regressions. Poisson regression, for example, uses the following formula:

$$\ln(\lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

We can use our estimated $\hat{\beta}$s to estimate the true Poisson parameter (which we set as part of our simulation), λ, and assess the bias of our model: Bias = $\hat{\lambda}$ - λ. We will do this for both GLM models.

We will also consider the estimated variance of our $\hat{\beta}$s. We will consider two different variance formulas:

- The naïve variance estimator: This estimator is our traditional variance estimate. The matrix formula is $Var_{naïve}(\hat{\beta}) = \sigma^2(X^TX)^{-1}$
- The robust variance estimator: This estimator is more robust to extreme values and heteroskedasticity. The matrix formula is: $Var_{robust}(\hat{\beta}) = (X^TX)^{-1}X^T\Omega X(X^TX)^{-1}$ where Ω is a diagonal matrix of square residuals.

The above formulas are so you can understand the mathematical underpinnings of this project, but we will use R functions to do all of this for us! Please do not hard code any complex formulas.

Part 2: Simulation Study

Create a simulation study to assess the impact of model misspecification on model bias and variance, including comparing the naïve and robust variances. Create a function to generate your simulation data and summarize your results (use multiple functions as needed). Carefully consider how to organize both your code scripts and your results.

Use the following conditions:

1. Sample sizes: N = 20, 200
2. True Poisson parameter: $\lambda = 4, 10$
3. Run each simulation 2,000 times

I have created your binary predictor variables for you. Read in the appropriate one for each respective sample size. There is only one explanatory variable for simplicity. The y variables are what you are simulating.

Using your simulated data, calculate your regression models: Poisson and normal.

```
fit_pois = glm(y ~ x, family = poisson, data = dat)
fit_norm = glm(y ~ x, family = gaussian, data = dat)
```

Part2a: The $\hat{\beta}$s

You will use your $\hat{\beta}$s to estimate $\hat{\lambda}$ and then calculate the mean bias for each model. You can easily do to this in R by using the predict function. You will need to set up your design matrix first:

```
#set up design matrix
new_data <- data.frame(intercept = 1, x1 = x)
```

The code below will take your fitted models and return a $\hat{\lambda}$ for each model. You can then calculate the bias of your $\hat{\lambda}$s.

```
#calculate lambda hats
pois_lambda_hat = predict(fit_pois, newdata=new_data, type = "response")
norm_lambda_hat = predict(fit_norm, newdata=new_data, type = "response")
```

Part2b: The Variances

You will need to obtain both variance estimators from your fitted models. The code below demonstrates how to obtain both variances from the Poisson model. You can use the same functions for the Normal models.

```
library(sandwich) #needed to get the robust sandwich variance estimator
nvar_pois = vcov(fit_pois)[2,2]
rvar_pois = sandwich(fit_pois)[2,2]
```

Part 3: Results

Your summary of results should include:

1. A table summarizing your results on the bias of $\hat{\lambda}$ under the various scenarios
2. A table summarizing the naïve and robust variances under the various scenarios
3. A plot or plots showing the 95% confidence intervals for the beta coefficients using both the naïve and robust variances (i.e., two CIs per distribution). For simplicity, use the normal distribution for the CIs:

   $$\hat{\beta} \pm 1.96 \times \text{SE}(\hat{\beta})$$

   Recall you can obtain the $\text{SE}(\hat{\beta})$ by taking the square root of the variance.

   - The y-axis should be the $\hat{\beta}$s, LBs, and UBs
   - The x-axis should be the two probability distributions used (categorical x-axis)
   - Facet on λ and N as needed
   - Use color to improve the readability of your plot
   - Provide a title, axis labels, etc. to make your plot informative

Part 4: Conclusions

Based on your simulation results, what conclusions can you draw? Consider questions like: What impact does misspecification have on the bias of $\hat{\lambda}$? How do the two variance estimators differ? How does sample size impact these results?