STA303H1S/STA1002HS Assignment 1
Due on July 24th, 2021 at 11:59 PM EDT on Quercus
All relevant work must be shown for credit.

Note: For any question, if you are using `R`, all `R` code and output must be included in your answers. You should assume that the reader is not familiar with `R` output, so please explain all your findings, quoting to relevant values from your output. Please note that academic integrity is fundamental to learning and scholarship. You may discuss questions with other students; however, the work you submit should be your own. If I feel suspicious of any assignment (e.g. if your work does not appear to be consistent with what we have discussed in class), I will not mark the assignment. Instead, I will ask you to present your work to me via Zoom/Bb Collaborate and your grade will be assigned based on your presentation. Assignments can be hand-writen but the `R` code and output should be printed.

1. Let $(Y_1, Y_2, \ldots, Y_K) \sim \text{Multinomial}(n, \pi_1, \pi_2, \ldots, \pi_K)$. Then:

   (a) Calculate the moment-generating function (This will be a multivariate MGF.) [3 marks]

   (b) Using the MGF from part (a),

      i. Show that $\mathbb{E}(Y_j) = n\pi_j$. [2 marks]

      ii. Show that $\text{Var}(Y_j) = n\pi_j(1 - \pi_j)$. [4 marks]

      iii. Show that $\text{Cov}(Y_i, Y_j) = -n\pi_i\pi_j$. [6 marks]

   (c) Assume that $(Y_1, Y_2, \ldots, Y_C) \sim \text{Multinomial}(n, \pi_1, \pi_2, \ldots, \pi_C)$. Show that the correlation coefficient between Y_i and Y_j is

   $$\text{Cor}(Y_i, Y_j) = \frac{-\pi_i\pi_j}{\sqrt{\pi_i(1-\pi_i)\pi_j(1-\pi_j)}}$$

   [3 marks]

   (d) Show that for $C = 2$ we have $\text{Cor}(Y_1, Y_2) = -1$. Explain why. [2 marks]

   Hint: You can use the multivariate MGF. Recall that if $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_K)$ is a multivariate random variable, then the MGF is defined as

   $$M_{\mathbf{Y}}(t_1, \ldots, t_K) = \mathbb{E}(\exp(t_1 Y_1 + \cdots + t_K Y_K))$$

   Then use partial derivatives on t_i, t_j to achieve the results.

2. Let Y_1 and Y_2 be independent Poisson random variables with parameters μ_1 and μ_2, respectively. Find the conditional distribution of Y_1 given $Y_1 + Y_2 = n$. That is, calculate $\text{P}(Y_1 = k | Y_1 + Y_2 = n)$. [5 marks]

3. Let $Y \sim \text{Bin}(n = 30, \pi = 0.9)$. Then Y can be interpreted as the number of successes in a sample of size $n = 30$ from a Binomial distribution with probability of success $\pi = 0.9$.

   (a) Suppose the observed number of successes after 30 trials is $y = 27$. Calculate the Wald and score (Wilson) 95% confidence intervals. [5 marks]

(b) Simulate $N = 100,000$ observations of Y using `R` function `rbinom()`. Calculate the Wald and score 95% confidence intervals for each of the observations. This means that you must calculate 100,000 confidence intervals of each type. Calculate the proportion of the Wald intervals which contain 0.9 (the true value of π). Then calculate the proportion of score intervals which contain 0.9. Compare the results and comment on your findings. Which do you think is a more reliable CI? [10 marks]

Note: `R` cannot generate random numbers. It only generates "pseudo" random numbers. Thus, a seed must be provided to reproduce the results. You should fix the seed in `R` using the `set.seed()` command, using your student ID as the seed. Thus, you must start the code with `set.seed(yourstudentID)`. If you do not provide the seed you will lose 3 marks.

4. Same as the previous question, let $Y \sim \text{Bin}(n = 30, \pi)$ and $y = 27$. This time we do not know the true value of π.

   (a) Find the likelihood function $L(\pi)$ and the log-likelihood function $\ell(\pi)$. [2 marks]

   (b) Using `R`, find the maximum likelihood estimate of π and plot $L(\pi)$ and $\ell(\pi)$ over the values of π. [5 marks]

   (c) Test $H_0 : \pi_0 = 0.5$ versus $H_a : \pi \neq 0.5$ using the likelihood ratio test. [3 marks]

   (d) Using `R`, calculate the 95% likelihood ratio confidence interval for π. [5 marks]

5. (a) Perform the following simulation (for this question, please set the seed to your student ID once again).

   – Generate 500 random values from each of $X_1 \sim \text{Uniform}[-10, 10]$, $X_2 \sim N(0, 4)$ and $X_3 \sim \text{Bernoulli}(0.7)$. You should therefore have 1,500 random values in total.
   – Set $\boldsymbol{\beta} = (-0.8, 0.1, 0.2, 0.3)$.
   – Simulate $Y_i \sim \text{Poisson}(\mu_i)$ where $\mu_i = \exp\left(\sum_j x_{ij}\beta_j\right)$.

   [10 marks]

   (b) Estimate the β coefficients using the iteratively re-weighted least squares (IRLS) method by writing your own function (or modifying the one I presented in class). Explain the procedure and state the W matrix as mentioned in Lecture 4. Compare the results with `glm` code in `R`. [15 marks]