August 2021 MS Comprehensive Exam

- Independent work: All work on the comprehensive exams is expected to be done independently. If you have questions about the exam, please email the statistics faculty list serve (math-statfac@sympa.montana.edu) for clarifications.
- Resources: For this exam you may use any textbooks or course materials. You are also free to use any publicly available materials that you might come across in a research setting; however, all writing must be original and all reference materials must be cited including code or text passages from course notes.
- Computer Code and Reproducibility: Please turn in all relevant computer code to reproduce your results; a reproducible document is a requirement. Either work in R Markdown or provide both SAS/python/Julia code and output.
- Time: This exam is due on August 18th at 10AM. We would anticipate that you would spend approximately 1 day on part I, 1 day on part II, 1 day on part III, and 2 days on part IV.
- Format: Only label the document with your GID. Make sure no directories or other artifacts contain identifying information. Submit a single output file containing responses to parts I, II, and III with the file name of GIDXXXX.pdf. Mathematical derivations can be completed by hand and scanned to include in this document. For the report in part IV submit a document with the file name of GIDXXXX_report.pdf. Also please include relevant source code as .RMD file, or other, to recreate your results. All files should be submitted via email to Jane Crawford (jane.crawford@montana.edu).
- Advice: Be sure to adequately justify your answers and choices and appropriately reference any sources used.
- Bayesian / Frequentist Paradigm: This exam makes no requirements for using Bayesian or Frequentist procedures, but when using Bayesian procedures make sure to clearly describe prior distributions. For either paradigm make sure to appropriately interpret results, e.g. confidence intervals vs. credible intervals.

Part I:

Throughout Part I of the exam, when applicable, please cite all theorems used in your solutions and provide a brief 'in words' explanation of why you are using them. Since many results in mathematical statistics can easily be found on the internet, your work must demonstrate a clear understanding of the steps leading up to an answer, in sufficient detail. Recall that all reference materials must be cited, including code or text passages from course notes.

Question 1

Assume $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables with probability density function

$$f_X(x|\theta) = e^{-(x-\theta)} I_{(\theta,\infty)}(x),$$

where $\theta > 0$ and $I_A(x)$ is an indicator variable for $x \in A$. Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ be the order statistics, and define $X_{(0)} = 0$, and $Z_i = X_{(i)} - X_{(i-1)}$ for $i = 1, \ldots, n$.

A. Give an example of a plausible real-world situation where the above model could apply, and justify why this is an appropriate model for your chosen example. Explain what the X_i's and Z_i's represent, and provide an interpretation of the parameter θ in context of your example.

B. Show that $X_{(1)}$, the first order statistic, is a sufficient statistic for θ.

C. Find the maximum likelihood estimator (MLE) for θ.

D. Derive the probability distribution of the MLE.

E. Is the MLE unbiased for θ? consistent for θ? Prove your claims.

F. Choose **one** of questions i. or ii. below. You may attempt both questions, but turn in only one.

i. Show that $Z_1, \ldots, Z_n$ are mutually independent and that $2(n+1-i)Z_i \sim \chi^2_2$.
ii. Let $Y = 2\sum_{i=1}^{n}(X_i - X_{(1)})$. What is the probability distribution of Y? Is Y independent of $X_{(1)}$? Justify your claim.

Question 2

Consider a Bernoulli trials setting with unknown probability of success $\theta \in (0, 1)$. Under one sampling method, a fixed number of independent Bernoulli trials, n, are performed, and the number of successes out of the n trials, X, is recorded. The probability mass function for the *Binomial* random variable X is

$$f_X(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} I_{\{0,1,\ldots,n\}}(x).$$

On the other hand, in inverse sampling, independent trials are performed until the kth success occurs, and the number of trials needed for k successes, Y, is recorded. Then the *Negative Binomial* random variable Y has mass function

$$f_Y(y|\theta) = \binom{y-1}{k-1} \theta^k (1-\theta)^{y-k} I_{\{k,k+1,\ldots,\}}(y).$$

Inferences about θ can be based on either a binomial experiment where n is fixed and X is observed, or based on a negative binomial experiment for k is fixed and Y is observed.

A. Explain why these are the correct mass functions for random variables X and Y. Be as precise as possible, especially arguing why the combinatorics are correct for each sampling method.

B. Recall that when $k = 1$, $Y \sim Geometric(\theta)$. Give a good *intuitive* argument about how the Negative Binomial random variable with parameters k and θ can be characterized as the sum of k independent, identically distributed (iid) $Geometric(\theta)$ random variables in the context of sampling iid $Bernoulli(\theta)$ random variables.

C. Suppose the two experiments are performed with $n = 100$ (Binomial) and $k = 25$ (Negative Binomial). Assume that $x = 25$ and $y = 100$ are observed.

i. Derive a large sample maximum likelihood-based confidence interval for the odds

$$\gamma = \frac{\theta}{1-\theta}$$

using the *Binomial data only*. Justify all steps.

ii. How would maximum likelihood-based inferences be different if you had used the Negative Binomial data?

iii. Now suppose you have a $Beta(a, b)$ prior on θ, and that the above data have been observed. How would your posterior inferences about θ compare in a Bayesian analysis depending on your choice of experimental data to analyze? In other words, assume the experimental data would be analyzed separately for the two experiments and explain how the posterior inferences would compare.

D. Now suppose you have a situation where there are two independent Binomial experiments with two binomial parameters θ_1 and θ_2. The Binomial counts are X_1 and X_2 based on sample sizes n_1 and n_2, respectively. (We are no longer considering the Negative Binomial.)

i. Derive a level $\alpha = 0.05$ likelihood ratio test for testing $H_0 : \theta_1 = \theta_2$ versus $H_a : \theta_1 \neq \theta_2$.

ii. Suppose we are interested in making inferences about the odds ratio:

$$\lambda = \frac{\theta_1/(1-\theta_1)}{\theta_2/(1-\theta_2)}.$$

Using large sample maximum likelihood theory, derive an explicit large sample confidence interval formula for the log odds ratio, and from that, obtain a large sample confidence interval formula for the actual odds ratio.

Part I:

Throughout Part I of the exam, when applicable, please cite all theorems used in your solutions and provide a brief 'in words' explanation of why you are using them. Since many results in mathematical statistics can easily be found on the internet, your work must demonstrate a clear understanding of the steps leading up to an answer, in sufficient detail. Recall that all reference materials must be cited, including code or text passages from course notes.

Question 1

Assume $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables with probability density function

$$f_X(x|\theta) = e^{-(x-\theta)} I_{(\theta,\infty)}(x),$$

where $\theta > 0$ and $I_A(x)$ is an indicator variable for $x \in A$. Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ be the order statistics, and define $X_{(0)} = 0$, and $Z_i = X_{(i)} - X_{(i-1)}$ for $i = 1, \ldots, n$.

A. Give an example of a plausible real-world situation where the above model could apply, and justify why this is an appropriate model for your chosen example. Explain what the X_i's and Z_i's represent, and provide an interpretation of the parameter θ in context of your example.

X_i = time to event after time θ

Z_i = time between an event and the one before it

This could be a model for time-to-event data, but starting after time θ. An example could be n machines, each with an independent time-to-failure after the machines needed to be serviced at time θ. X_i would model the time-to-failure for machine i. Z_i would model the time between the i^{th} failure and the one preceding it. θ would represent the time point where the machines should be serviced.

Justify

In Casella & Berger, they discuss the exponential distribution as a special case of the gamma distribution on page 101, and this pdf is a $Y \sim Exp(1)$ distribution with the transformation:

$$g(y) = y + \theta \quad \Rightarrow \quad g^{-1}(x) = x - \theta$$

or, the "shifted exponential" distribution.

In Casella & Berger, they discuss the exponential distribution is used to model lifetimes of things (like machines!).

B. Show that $X_{(1)}$, the first order statistic, is a sufficient statistic for θ.

$f_x(x|\theta) = \exp(-(x-\theta))\, I_{(\theta,\infty)}(x)$ defn

$f_x(\underline{x}|\theta) = \prod_{i=1}^{n} \exp(-(x_i-\theta))\, I_{(\theta,\infty)}(x_i)$ since x_i's iid.

$= \exp\left(-\sum_{i=1}^{n}(x_i-\theta)\right) \prod_{i=1}^{n} I_{(\theta,\infty)}(x_i)$ simplifying

$= \exp\left(n\theta - \sum_{i=1}^{n} x_i\right) I(\theta < \text{all } x_i < \infty)$ $\prod_{i=1}^{n} I_{(\theta,\infty)}(x_i)$ indicates all x_i need to be $\theta < x_i < \infty$

$= \exp(n\theta)\exp\left(-\sum_{i=1}^{n} x_i\right) I_{(0,x_{(1)})}(\theta)$ if all $x_i > \theta$, then $x_{(1)} > \theta$ is equivalent

$= \exp(n\theta)\exp\left(-\sum_{i=1}^{n} x_i\right) I_{(0,x_{(1)})}(\theta)$ rewriting for bounds to not be dep. on θ

By Thm. 6.2.13: For a function $T(\underline{x})$,

$\frac{f(\underline{x}|\theta)}{f(\underline{y}|\theta)}$ is constant as a function of θ iff $T(\underline{x}) = T(\underline{y})$,

then $T(\underline{x})$ is a minimal suff. stat. for θ

$$\frac{f(\underline{x}|\theta)}{f(\underline{y}|\theta)} = \frac{\exp(n\theta)\exp\left(-\sum_{i=1}^{n} x_i\right) I_{(0,x_{(1)})}(\theta)}{\exp(n\theta)\exp\left(-\sum_{i=1}^{n} y_i\right) I_{(0,y_{(1)})}(\theta)}$$

$$= \frac{\exp\left(-\sum_{i=1}^{n} x_i\right) I_{(0,x_{(1)})}(\theta)}{\exp\left(-\sum_{i=1}^{n} y_i\right) I_{(0,y_{(1)})}(\theta)}$$

is constant as a fxn of θ only when $x_{(1)} = y_{(1)}$, which implies $T(\underline{x}) = x_{(1)}$ is a M.S.S.

This was similar to Q2a HW4 of S02, so I followed that for this problem

C. Find the maximum likelihood estimator (MLE) for θ.

Defn 7.2.4: $\hat{\theta}_{MLE}$ is the parameter value where $L(\theta|\underline{x})$ attains its maximum for fixed $\underline{x}$

From B:

$$L(\theta|\underline{x}) = f(\underline{x}|\theta) = \exp(n\theta)\exp\left(-\sum_{i=1}^{n} x_i\right) I_{(0,x_{(1)})}(\theta)$$

$$\hat{\theta}_{MLE} = \sup_{\theta} L(\theta|\underline{x}) = \sup_{\theta}\left[\exp(n\theta)\exp\left(-\sum_{i=1}^{n} x_i\right) I_{(0,x_{(1)})}(\theta)\right]$$

θ must be $0 < \theta < x_{(1)}$, so the value of θ that will maximize $L(\theta|\underline{x})$ is $x_{(1)}$, since $\exp\left(n\theta - \sum_{i=1}^{n} x_i\right)$ is strictly increasing as θ increases, since $n \cdot \theta > 0$.

$$\hat{\theta}_{MLE} = x_{(1)}$$

This was an example on 7.1-7.2 notes, p. 34

D. Derive the probability distribution of the MLE.

$$F_{x_{(1)}}(x) = P(x_{(1)} \le x) = 1 - P(x_{(1)} > x) = 1 - P(\text{all } x_i > x)$$

$$= P(x_1 > x, x_2 > x, \ldots, x_n > x) \quad \text{since } x_i\text{'s indep.}$$

$$= 1 - [P(x > x)]^n \quad \text{since } x_i\text{'s identically dist}$$

$$= 1 - [1 - F_x(x)]^n$$

$$f_{x_{(1)}}(x) = \frac{d}{dx} F_{x_{(1)}}(x) = n f_x(x)[1 - F_x(x)]^{n-1}$$

pdf of $\hat{\theta}_{MLE}$:

$$f_{x_{(1)}}(x) = \begin{cases} n e^{-(x-\theta)}\left[1 - (1 - \exp(-(x-\theta)))\right]^{n-1} & x > \theta \\ 0 & \text{o.w.} \end{cases}$$

$$= \begin{cases} n\exp(-(x-\theta))[\exp(-(x-\theta))]^{n-1} & x > \theta \\ 0 & \text{o.w.} \end{cases}$$

$$f_{X_{(1)}}(x) = \begin{cases} n\exp(-(x-\theta))^n & x > \theta \\ 0 & \text{o.w.} \end{cases}$$

Let $Y = X - \theta$ $\quad f_Y(y) = n(e^{-y})^n I_{(0,\infty)}(y) = ne^{-ny} I_{(0,\infty)}(y)$

$Y \sim \text{EXP}(\text{rate} = n)$ or $Y \sim \text{EXP}(\text{scale} = 1/n)$

$\hat{\theta}_{MLE} \sim$ Shifted Exponential (rate $= n$, shift $= \theta$)

or $\hat{\theta}_{MLE} \sim$ Shifted EXP(scale $= 1/n$, shift $= \theta$)

E. Is the MLE unbiased for θ? consistent for θ? Prove your claims.

Def 7.3.2: The bias of an estimator W of parameter θ is $E_\theta(W) - \theta$

$\text{Bias}(\hat{\theta}_{MLE}) = E_\theta(\hat{\theta}_{MLE}) - \theta$

$$E(\hat{\theta}_{MLE}) = \int_\theta^\infty n\exp(-(x-\theta))^n \cdot x\, dx = \theta + \frac{1}{n}$$

$\text{Bias}(\hat{\theta}_{MLE}) = E(\hat{\theta}_{MLE}) - \theta = \theta + 1/n - \theta = 1/n$

$\hat{\theta}_{MLE}$ is not unbiased for θ

Def 10.1.1: A sequence of estimators $W_n = W_n(X_1, \ldots, X_n)$ is a consistent sequence of estimators of θ if, for every $\varepsilon > 0$ and $\theta \in \Theta$,

$\lim_{n\to\infty} P_\theta(|W_n - \theta| < \varepsilon) = 1$, or equivalently,

$\lim_{n\to\infty} P_\theta(|W_n - \theta| \geq \varepsilon) = 0$

Thm 10.1.3: If W_n is a sequence of estimators of θ satisfying

i) $\lim_{n\to\infty} \text{Var}_\theta(W_n) = 0$

ii) $\lim_{n\to\infty} \text{Bias}_\theta(W_n) = 0$

for every $\theta \in \Theta$

then W_n is a consistent sequence of estimators of θ

i) $\text{Var}(\hat{\theta}_{MLE}) = \int_{\theta}^{\infty} x^2 \, n \exp(-n(x-\theta)) \, dx = \frac{n\theta(n\theta+2)+2}{n^2}$

$= \frac{n^2\theta^2 + 2n\theta + 2}{n^2} = \theta^2 + \frac{2\theta}{n} + \frac{2}{n^2}$

$\lim_{n\to\infty} \text{Var}(\hat{\theta}_{MLE}) = \lim_{n\to\infty} \left[\theta^2 + \frac{2\theta}{n} + \frac{2}{n^2}\right] = \theta^2 \neq 0$

ii) $\lim_{n\to\infty} \text{Bias}(\hat{\theta}_{MLE}) = \lim_{n\to\infty} 1/n = 0$

$\Rightarrow \hat{\theta}_{MLE}$ is <u>not</u> an unbiased or consistent estimator of θ since $\text{Bias}(\hat{\theta}_{MLE}) = 1/n \neq 0$ and $\lim_{n\to\infty} \text{Var}(\hat{\theta}_{MLE}) = \theta^2 \neq 0$

F. Choose **one** of questions i. or ii. below. You may attempt both questions, but turn in only one.

i. Show that $Z_1, \ldots, Z_n$ are mutually independent and that $2(n+1-i)Z_i \sim \chi^2_2$.

ii. Let $Y = 2\sum_{i=1}^{n}(X_i - X_{(1)})$. What is the probability distribution of Y? Is Y independent of $X_{(1)}$? Justify your claim.

$X_i = \text{EXP}(1,\theta) I_{(\theta,\infty)}(x)$

$X_{(1)} = \text{EXP}(n,\theta) I_{(\theta,\infty)}(x)$

ii) $Y = 2\sum_{i=1}^{n}(X_i - X_{(1)})$

$Y = 2\left[\sum_{i=1}^{n} X_i - nX_{(1)}\right] = 2\left[\underbrace{\sum_{i=1}^{n}(X_i-\theta)}_{\sim \text{EXP}(n)} - \underbrace{n(X_{(1)}-\theta)}_{\sim \text{EXP}(1)}\right] \sim 2[\text{EXP}(n-1)]$ since sums of EXP(θ) RV is EXP(nθ)

$= \text{Gamma}(n-1, 2)$ $\quad \text{Gamma}\left(\frac{2n-2}{2}, 2\right) \sim \chi^2(2n-2)$

$2\sum_{i=1}^{n}(X_i - X_{(1)}) = 2\sum_{i=1}^{n}\left[(X_i-\theta) - (X_{(1)}-\theta)\right]$

$(X_i - \theta) \sim \text{EXP}(1) \rightarrow$ indep of θ

$(X_{(1)} - \theta) \sim \text{EXP}(n) \rightarrow$ indep of θ

$\Rightarrow 2\sum_{i=1}^{n}(X_i - X_{(1)})$ is an ancillary statistic for θ

Since $X_{(1)}$ is a c.s.s. for θ, $Y = 2\sum_{i=1}^{n}(X_i - X_{(1)})$ is independent of $X_{(1)}$ by Basu's Thm.

Question 2

Consider a Bernoulli trials setting with unknown probability of success $\theta \in (0,1)$. Under one sampling method, a fixed number of independent Bernoulli trials, n, are performed, and the number of successes out of the n trials, X, is recorded. The probability mass function for the *Binomial* random variable X is

$$f_X(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}I_{\{0,1,\ldots,n\}}(x).$$

On the other hand, in inverse sampling, independent trials are performed until the kth success occurs, and the number of trials needed for k successes, Y, is recorded. Then the *Negative Binomial* random variable Y has mass function

$$f_Y(y|\theta) = \binom{y-1}{k-1}\theta^k(1-\theta)^{y-k}I_{\{k,k+1,\ldots,\}}(y).$$

Inferences about θ can be based on either a binomial experiment where n is fixed and X is observed, or based on a negative binomial experiment for k is fixed and Y is observed.

A. Explain why these are the correct mass functions for random variables X and Y. Be as precise as possible, especially arguing why the combinatorics are correct for each sampling method.

For a fixed n number of trials, there are $\binom{n}{x}$ ways to arrange x successes, with each success have equal probability θ of success and each of the n-x failures have equal probability $1-\theta$. X can be anything from 0 to n successes total.

For k total successes in y trials, there are $\binom{y-1}{k-1}$ ways to arrange the first k-1 successes in y-1 trials since the last trial must be a success. Each of the k successes has equal probability θ and each of the y-k failures has equal probability $1-\theta$. There must be at least k trials, but there is no limit to the total number of trials.

B. Recall that when $k = 1$, $Y \sim Geometric(\theta)$. Give a good *intuitive* argument about how the Negative Binomial random variable with parameters k and θ can be characterized as the sum of k independent, identically distributed (iid) $Geometric(\theta)$ random variables in the context of sampling iid $Bernoulli(\theta)$ random variables.

Instead of stopping at the first success, the exact same Geometric process is repeated k times. The probability of success doesnt change, so the Negative Binomial is just k replications of a Geometric with the same θ parameter.

C. Suppose the two experiments are performed with $n = 100$ (Binomial) and $k = 25$ (Negative Binomial). Assume that $x = 25$ and $y = 100$ are observed.

i. Derive a large sample maximum likelihood-based confidence interval for the odds

$$\gamma = \frac{\theta}{1-\theta}$$

using the *Binomial data only.* Justify all steps.

ii. How would maximum likelihood-based inferences be different if you had used the Negative Binomial data?

iii. Now suppose you have a $Beta(a, b)$ prior on θ, and that the above data have been observed. How would your posterior inferences about θ compare in a Bayesian analysis depending on your choice of experimental data to analyze? In other words, assume the experimental data would be analyzed separately for the two experiments and explain how the posterior inferences would compare.

Goal: find MLE large sample CI

1) Find MLE (Ex. 7.2.7, Thm 7.2.10)

2) Estimate variance by Delta method assuming asymptotic Normal (Ex. 5.5.22, 10.1.14)

3) Find CI by $\hat{\gamma} - z_{\alpha/2}\sqrt{Var(\hat{\gamma})} \leq \gamma \leq \hat{\gamma} + z_{\alpha/2}\sqrt{Var(\hat{\gamma})}$ (Ex. 10.4.1)

1) $\hat{\gamma}_{MLE} = \frac{\hat{\theta}_{MLE}}{1-\hat{\theta}_{MLE}}$ by invariance property of MLE's Thm 7.2.10

$\hat{\theta}_{MLE} = \sup_{\theta} L(\theta|\underline{x})$ defn of MLE

$L(\theta|\underline{x}) = \prod_{i=1}^{N} \binom{n}{x_i} \theta^{x_i}(1-\theta)^{n-x_i} I_{\{0,\dots,n\}}(x_i)$ N = number of experiment replications

$= \prod_{i=1}^{N}\binom{n}{x_i}\, \theta^{\sum_{i=1}^{N} x_i}(1-\theta)^{Nn-\sum_{i=1}^{N} x_i} \prod_{i=1}^{N} I_{\{0,\dots n\}}(x_i)$ By defn of Likelihood

$\log L(\theta|\underline{x}) = \log\left(\prod_{i=1}^{N}\binom{n}{x_i}\right) + \sum_{i=1}^{N} x_i \log\theta + \left(Nn - \sum_{i=1}^{N} x_i\right)\log(1-\theta)$

$+ \log\left(\prod_{i=1}^{N} I_{\{0,\dots,n\}}(x_i)\right)$ Take ln of both sides

-value that maximizes $L(\theta|\underline{x})$ also maximizes $\log L(\theta|\underline{x})$

$\partial/\partial\theta \log L(\theta|\underline{x}) = \frac{\sum_{i=1}^{N} x_i}{\theta} - \frac{Nn - \sum_{i=1}^{N} x_i}{1-\theta} \overset{\text{set}}{=} 0$ Take $\partial/\partial\theta$ and set equal to 0

$\frac{\sum_{i=1}^{N} x_i}{\theta} = \frac{Nn - \sum_{i=1}^{N} x_i}{1-\theta} \Rightarrow \frac{1-\theta}{\theta} = \frac{Nn - \sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i}$ solve for θ

$\hat{\theta}_{MLE} = \frac{\sum_{i=1}^{N} x_i}{nN} = \frac{25}{100} = 0.25$ plug in known values $n=100$ $N=1$ $\sum_{i=1}^{N} x_i = 25$

$\hat{\gamma}_{MLE} = \frac{0.25}{1-0.25} = 0.33$ solve for $\hat{\gamma}_{MLE} = \frac{\hat{\theta}_{MLE}}{1-\hat{\theta}_{MLE}}$

2) Ex. 5.5.22, 10.1.14 goes through estimating the variance of γ

$$\widehat{Var}(\hat{\gamma}_{MLE}) = Var\left(\frac{\hat{\theta}_{MLE}}{1-\hat{\theta}_{MLE}}\right) \approx \frac{\hat{\theta}_{MLE}}{n(1-\hat{\theta}_{MLE})^3}$$

pf: Thm 10.1.12

Let $\hat{\theta}$ be MLE of θ, $T(\theta)$ to be a continuous fn of θ

then $\sqrt{n}(T(\hat{\theta}) - T(\theta)) \to N(0, v(\theta))$

where $v(\theta)$ is the CRLB

Let $T(\hat{\theta}) = \frac{\hat{\theta}_{MLE}}{1-\hat{\theta}_{MLE}}$, $T(\theta) = \frac{\theta}{1-\theta}$

$$v(\theta) = CRLB\left(\frac{\theta}{1-\theta}\right) = \frac{\left[\frac{\partial}{\partial\theta}\left(\frac{\theta}{1-\theta}\right)\right]^2}{-n E\left(\frac{\partial^2}{\partial^2\theta} \log\left(\frac{\theta}{1-\theta}\right)\right)}$$

$$= \frac{\left[\left(\frac{1}{1-\theta}\right)^2\right]^2}{-n\,\partial^2/\partial^2\theta\left[\log(\theta) - \log(1-\theta)\right]} = \frac{\frac{1}{(1-\theta)^4}}{\frac{n}{\theta(1-\theta)}} = \frac{\theta}{n(1-\theta)^3}\Big|_{\theta=\hat{\theta}_{MLE}}$$

$$= \frac{\hat{\theta}_{MLE}}{n(1-\hat{\theta}_{MLE})}$$

Also: $\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{D} N(0,1)$

by Slutsky's Thm: $\frac{\hat{\gamma} - \gamma}{\sqrt{\widehat{Var}(\hat{\gamma})}/\sqrt{n}} \to N(0,1)$

3) $100(1-\alpha)\%$ CI for γ:

$$\hat{\gamma}_{MLE} - z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\gamma}_{MLE})} \le \gamma \le \hat{\gamma}_{MLE} + z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\gamma}_{MLE})}$$

$$= \left(\frac{\hat{\theta}_{MLE}}{1-\hat{\theta}_{MLE}} - z_{\alpha/2}\sqrt{\frac{\hat{\theta}_{MLE}}{(1-\hat{\theta}_{MLE})^3 \cdot n}},\ \frac{\hat{\theta}_{MLE}}{1-\hat{\theta}_{MLE}} + z_{\alpha/2}\sqrt{\frac{\hat{\theta}_{MLE}}{(1-\hat{\theta}_{MLE})\cdot n}}\right)$$

$$= \left(0.33 - z_{\alpha/2}(0.076978),\ 0.33 + z_{\alpha/2}(0.076978)\right)$$

ii) How would inferences be different using NB data?

I would use a different estimate of $\hat{\theta}_{MLE}$ everywhere where I used $\hat{\theta}_{MLE}$, but the same general CI still works.

$\hat{\theta}_{NB}$ = MLE of θ based on y

$$L(\theta|y) = f(y|\theta) = \prod_{i=1}^{n}\binom{y-1}{k-1}\theta^k(1-\theta)^{y-k} I_{\{k,\ldots\}}(y) \xrightarrow{n=1}$$

$$= \binom{y-1}{k-1}\theta^k(1-\theta)^{y-k} I_{\{k,\ldots\}}(y)$$

$$\log L(\theta|y) = \log\binom{y-1}{k-1} + k\log\theta + (y-k)\log(1-\theta) + \log(I_{\{k,\ldots\}}(y))$$

$$\partial/\partial\theta \log L(\theta|y) = \frac{k}{\theta} - \frac{y-k}{1-\theta} \overset{\text{set}}{=} 0$$

$$\frac{k}{\theta} = \frac{y-k}{1-\theta} \Rightarrow \frac{1-\theta}{\theta} = \frac{y-k}{k}$$

$$\hat{\theta}_{NB,MLE} = k/y = \frac{25}{100} = 0.25$$

so nothing would change in this case since $\hat{\theta}_{BIN} = \hat{\theta}_{NB}$?

iii. Now suppose you have a $Beta(a, b)$ prior on θ, and that the above data have been observed. How would your posterior inferences about θ compare in a Bayesian analysis depending on your choice of experimental data to analyze? In other words, assume the experimental data would be analyzed separately for the two experiments and explain how the posterior inferences would compare.

Def 7.2.7 Ex. 7.2.14

$$\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{m(x)} \propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \sim Beta(x+\alpha, n-x+\beta)$$

$$\hat{\theta}_{Bayes} = \frac{x+\alpha}{\alpha+\beta+n}$$

$$\pi(\theta|y) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^k(1-\theta)^{y-k}$$

$$\propto \theta^{k+\alpha-1}(1-\theta)^{y-k+\beta-1} \sim Beta(\alpha+k, y-k+\beta)$$

$$\hat{\theta}_{Bayes,NB} = \frac{\alpha+k}{y-k+\beta}$$

I should be able to use both $\hat{\theta}_{Bayes}$ in place of $\hat{\theta}_{MLE}$ in the inferences above depending on which exp. I ran

D. Now suppose you have a situation where there are two independent Binomial experiments with two binomial parameters θ_1 and θ_2. The Binomial counts are X_1 and X_2 based on sample sizes n_1 and n_2, respectively. (We are no longer considering the Negative Binomial.)

i. Derive a level $\alpha = 0.05$ likelihood ratio test for testing $H_0 : \theta_1 = \theta_2$ versus $H_a : \theta_1 \neq \theta_2$.

ii. Suppose we are interested in making inferences about the odds ratio:

$$\lambda = \frac{\theta_1/(1-\theta_1)}{\theta_2/(1-\theta_2)}.$$

Using large sample maximum likelihood theory, derive an explicit large sample confidence interval formula for the log odds ratio, and from that, obtain a large sample confidence interval formula for the actual odds ratio.

i) Test $H_0: \theta_1 = \theta_2 = \theta_0$

$H_A: \theta_1 \neq \theta_2$

$\hat{\theta}_0 = \frac{X_1 + X_2}{n_1 + n_2}$ $\hat{\theta}_1 = \frac{X_1}{n_1}$ $\hat{\theta}_2 = \frac{X_2}{n_2}$ from MLE

LRT: $\lambda(\underline{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta | \underline{x})}{\sup_{\theta \in \Theta} L(\theta | \underline{x})} = \frac{L(\hat{\theta}_0 | \underline{x})}{L(\hat{\theta} | \underline{x})}$ Def. 8.2.1

$$= \frac{\binom{n_1}{x_1} \theta_0^{x_1} (1-\theta_0)^{n_1 - x_1} I_{\{0,\ldots,n_1\}}(x_1) \binom{n_2}{x_2} \theta_0^{x_2} (1-\theta_0)^{n_2 - x_2} I_{\{0,\ldots,n_2\}}(x_2)}{\binom{n_1}{x_1} \theta_1^{x_1} (1-\theta_1)^{n_1 - x_1} I_{\{0,\ldots,n_1\}}(x_1) \binom{n_2}{x_2} \theta_2^{x_2} (1-\theta_2)^{n_2 - x_2} I_{\{0,\ldots,n_2\}}(x_2)}$$

$$= \frac{\theta_0^{x_1 + x_2} (1-\theta_0)^{n_1 + n_2 - x_1 - x_2}}{\theta_1^{x_1} \theta_2^{x_2} (1-\theta_1)^{n_1 - x_1} (1-\theta_2)^{n_2 - x_2}} = \frac{\binom{x_1 + x_2}{n_1 + n_2}^{x_1 + x_2} \left(1 - \frac{x_1 + x_2}{n_1 + n_2}\right)^{n_1 + n_2 - x_1 - x_2}}{\left(\frac{x_1}{n_1}\right)^{x_1} \left(\frac{x_2}{n_2}\right)^{x_2} \left(1 - \frac{x_1}{n_1}\right)^{n_1 - x_1} \left(1 - \frac{x_2}{n_2}\right)^{n_2 - x_2}}$$

Reject if $\lambda(\underline{x}) \leq c$

Thm 10.3.6: As $n \to \infty$, $-2 \log(\lambda(\underline{x})) \sim \chi^2_{(4-3)}$

$$\phi(\underline{x}) \begin{cases} 1 & -2 \log \lambda(\underline{x}) \leq \chi^2_{1, 0.05} \\ 0 & \text{else} \end{cases}$$

where $\lambda(\underline{x}) = \frac{\binom{x_1 + x_2}{n_1 + n_2}^{x_1 + x_2} \left(1 - \frac{x_1 + x_2}{n_1 + n_2}\right)^{n_1 + n_2 - x_1 - x_2}}{\left(\frac{x_1}{n_1}\right)^{x_1} \left(\frac{x_2}{n_2}\right)^{x_2} \left(1 - \frac{x_1}{n_1}\right)^{n_1 - x_1} \left(1 - \frac{x_2}{n_2}\right)^{n_2 - x_2}}$

ii) $\lambda = \dfrac{\theta_1/(1-\theta_1)}{\theta_2/(1-\theta_2)}$ derive CI for $\log(\lambda)$ and λ

$$\log(\lambda) = \log\left(\frac{\theta_1}{1-\theta_1}\right) - \log\left(\theta_2/1-\theta_2\right)$$

$\hat{\theta}_1 = \dfrac{x_1}{n_1}$ $\hat{\theta}_2 = \dfrac{x_2}{n_2}$ by MLE

$$Var(\log\lambda) = Var\left(\log\left(\frac{\theta_1}{1-\theta_1}\right)\right) + Var\left(\log\left(\frac{\theta_2}{1-\theta_2}\right)\right)$$

since θ_1, θ_2 independent

Using Delta Method:

$Y_n = \hat{\theta}_1$ $g(\theta) = \log\left(\frac{\theta_1}{1-\theta_1}\right)$ $g'(\theta_1) = \frac{1}{\theta_1(1-\theta_1)}$

$g(Y_n) = \log\left(\frac{\hat{\theta}_1}{1-\hat{\theta}_1}\right)$

If $\sqrt{n}(\hat{\theta}_1 - \theta_1) \xrightarrow{D} N(0, \theta_1(1-\theta_1))$ then

$\sqrt{n}\left(\log\left(\frac{\hat{\theta}_1}{1-\hat{\theta}_1}\right) - \log\left(\frac{\theta_1}{1-\theta_1}\right)\right) \xrightarrow{D} N(0, \theta_1(1-\theta_1)g'(\theta_1))$

$$\Rightarrow \log\left(\frac{\hat{\theta}_1}{1-\hat{\theta}_1}\right) \sim AN\left(\log\left(\frac{\theta_1}{1-\theta_1}\right), \frac{1}{n\theta_1(1-\theta_1)}\right)$$

$$\rightarrow \frac{\sqrt{n}(\hat{\theta}_1 - \theta_1)}{\theta_1(1-\theta_1)} \xrightarrow{D} N(0,1)$$ by CLT since θ_1 are iid Bernoulli trials

$$\Rightarrow \sqrt{n}(\hat{\theta}_1 - \theta_1) \xrightarrow{D} N(0, \theta_1(1-\theta_1))$$ by Slutsky's Thm

$$Var(\log\lambda) = Var\left(\log\left(\frac{\theta_1}{1-\theta_1}\right)\right) + Var\left(\log\left(\frac{\theta_2}{1-\theta_2}\right)\right)$$

$$\approx \frac{1}{n_1\theta_1(1-\theta_1)} + \frac{1}{n_2\theta_2(1-\theta_2)}$$

Large sample CI: 10.4.1

$$\log(\hat{\lambda}) - z_{\alpha/2}\sqrt{\frac{1}{n_1}\hat{\theta}_1(1-\hat{\theta}_1) + \frac{1}{n_2}\hat{\theta}_2(1-\hat{\theta}_2)} \le \log(\lambda) \le \log(\hat{\lambda}) + z_{\alpha/2}\sqrt{\cdots}$$

$$= \left(\log(\hat{\lambda}) - z_{\alpha/2}\sqrt{\frac{1}{x_1}\left(1-\frac{x_1}{n_1}\right) + \frac{1}{x_2}\left(1-\frac{x_2}{n_2}\right)}\right.,$$
$$\left.\log(\hat{\lambda}) + z_{\alpha/2}\sqrt{\frac{1}{x_1}\left(1-\frac{x_1}{n_1}\right) + \frac{1}{x_2}\left(1-\frac{x_2}{n_2}\right)}\right)$$

I dont have methematical backing for why we can do this, but in regression when we model log-odds, we get our estimates in log scale and we often exponentiate to get out of log scale. I figure we can exponentiate this CI's endpoints and get the CI for the odds

$\Rightarrow$ Large sample CI for λ:

$$\left(\hat{\lambda} - \exp\left(z_{\alpha/2}\sqrt{\frac{1}{x_1}\left(1-\frac{x_1}{n_1}\right) + \frac{1}{x_2}\left(1-\frac{x_2}{n_2}\right)}\right),\right.$$
$$\left.\hat{\lambda} + \exp\left(z_{\alpha/2}\sqrt{\frac{1}{x_1}\left(1-\frac{x_1}{n_1}\right) + \frac{1}{x_2}\left(1-\frac{x_2}{n_2}\right)}\right)\right)$$

August 2021 MS Comprehensive Exam

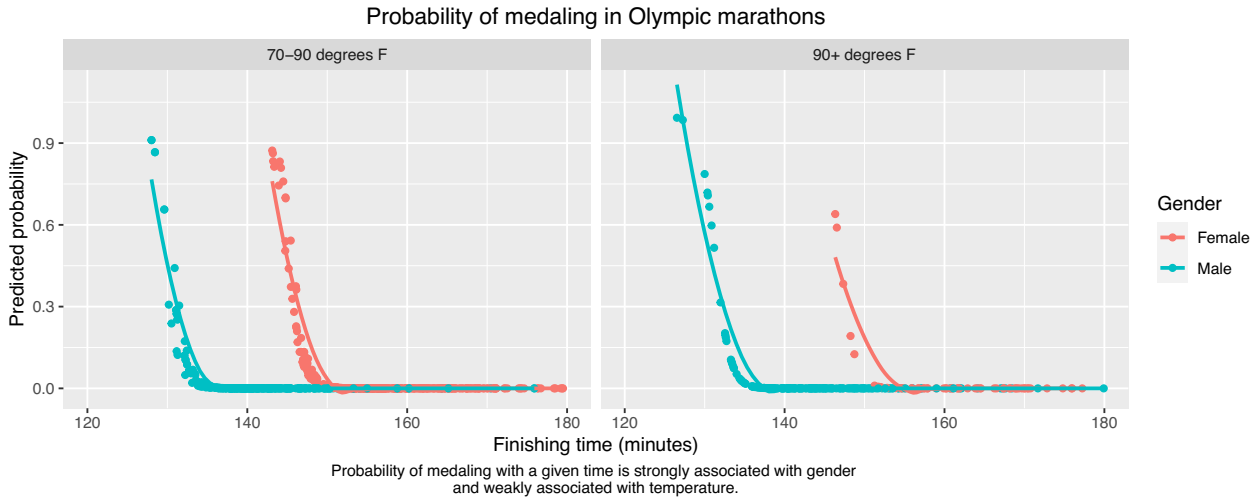
GID9475

Part II:

Question 1

For this question use the following dataset on Olympic marathon results which contains finishers from the last six Olympic games (1996, 2000, 2004, 2008, 2012, and 2016). The goal is to use a logistic regression model to predict whether a marathon time (in seconds) would result in winning a medal when considering sex of the athlete and temperature on the race day.

A. Data Visualization. Create one or, a maximum of two, data visualizations to explore the data with respect to understanding factors related to winning a medal. For each figure include meaningful titles, labels, and potentially annotation and also interpret the results with a succinct written summary.



The probability of medaling in an Olympic marathon based on a given time is high (above 50% probability) at around 130 minutes for men and around 145 minutes for women when the temperature is 70-90 degrees F, and around 130 minutes for men and 150 minutes for women when the temperature is 90 degrees F or above. Female runners that have a high probability of medaling tend to be around 15-20 minutes slower on average. It appears that there could be an interaction effect between gender and temperature. The women who medaled on the highest temperature days had slower times compared to more moderate temperature days, and men stayed approximately the same at all temperatures.

B. Model Specification. Using matrix algebra, write out the full statistical model to address this question. Include any model assumptions. This model should match what you will use in the next question (Part c).

$X_{n\times 5} = \begin{pmatrix} \text{Int.} & \text{Seconds} & \text{Gender} & \text{Temp} & \text{G:T} \\ 1 & \vdots & \vdots & \vdots & \vdots \\ 1 & & & & \end{pmatrix}$ $y_{n\times 1} = \begin{pmatrix} \text{Medal?} \\ 0 \\ \vdots \\ 1 \end{pmatrix}$

$y_i \sim \text{Bernoulli}(\pi_i)$

$\beta_{5\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \end{pmatrix}$

$\text{logit}(\pi) = X\beta$

$(n\times 5)(5\times 1) = n\times 1$

C. Model Fitting. Fit the best model to predict winning a medal and include your code in the document. Describe you model choice criteria and defend your final model.

```
model = glm(data = marathon2,
            formula = Medal ~ `Seconds after 2 hrs` + Gender * `Temp above 70 degrees`,
            family = binomial(link = "logit")) # AIC = 111.36
# model = glm(data = marathon2,
# formula = Medal ~ Seconds + Gender + Temp,
# family = binomial(link = "logit")) # AIC = 115.22
summary(model)
```

```
##
## Call:
## glm(formula = Medal ~ `Seconds after 2 hrs` + Gender * `Temp above 70 degrees`,
##     family = binomial(link = "logit"), data = marathon2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.66903  -0.02876  -0.00286  -0.00004   2.52266
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        26.66784    4.22403   6.313 2.73e-10
## `Seconds after 2 hrs`              -0.01830    0.00277  -6.607 3.91e-11
## GenderMale                        -15.09298    2.45883  -6.138 8.34e-10
## `Temp above 70 degrees`             0.18437    0.05566   3.312 0.000925
## GenderMale:`Temp above 70 degrees` -0.18063    0.07633  -2.366 0.017957
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 308.63  on 980  degrees of freedom
## Residual deviance: 101.36  on 976  degrees of freedom
## AIC: 111.36
##
## Number of Fisher Scoring iterations: 10
```

A logistic regression model is used to model the log-odds of a participant winning a medal. The included predictors are the time that a participant finished the race (centered at 2 hours), the gender of the participant, the temperature when the race was ran, and an interaction term between the gender and temperature. In exploratory data analysis in part A, there was a difference between how temperature affected men and women in the plot, so two candidate models containing and excluding the interaction term were compared, and the model with the interaction term had a lower AIC compared to the model without.

D. Model Summary. Summarize the model fit in Part C using both written summaries and data visualization. If you use a Bayesian procedure, prior distributions should be clearly stated.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	26.67	4.22	6.31	0.00
`Seconds after 2 hrs`	-0.02	0.00	-6.61	0.00
GenderMale	-15.09	2.46	-6.14	0.00
`Temp above 70 degrees`	0.18	0.06	3.31	0.00
GenderMale:`Temp above 70 degrees`	-0.18	0.08	-2.37	0.02

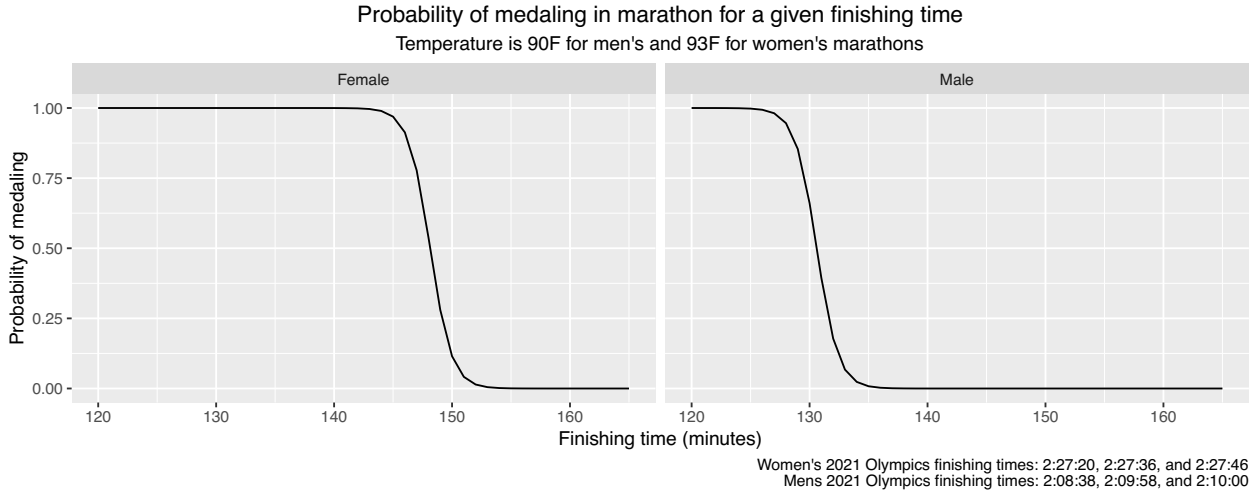
The main predictor to predict if a medal was won is the number of seconds after 2 hours the finishing time.

For each 1 second increase after 2 hours, the predicted odds of medal being won decrease by around 2% ($\pm$ 0.3%) with all else remaining the same. The gender of the participant also makes a big difference, since female medal winners run on average around 17.6 minutes slower than male medal winners. The predicted effect due to gender was around 13.7 minutes ($\pm$ 2.46 *mins*) worth by itself. There was an effect due to the temperature of around 10 seconds ($\pm$ 3 *seconds*) worth for every 1 degree increase above 70 degrees, but this effect was nearly completely canceled out for the men by the interaction term. It appears that, all else held the same, female runners are strongly affected by increases in temperature and men are affected much less. On a 94 degree F day, the effect due to the high temperatures for male runners is predicted to be around 5 seconds ($\pm$ 73 *seconds*) worth, but for female runners the effect is estimated to be around 241 seconds ($\pm$ 73 *seconds*) worth.

This effect could be an artifact of the data, because there are no Olympics competitions where women ran in temperatures between 79 and 94 degrees in the data, so there could be an unmeasured extraneous factor on the only day where the female runners ran in the extreme heat. I would be hesitant to make any strong conclusions about the temperature predictor in this analysis.

E. Predictions for 2021. Due to projected high temperatures in Tokyo, the host city of the 2021 Olympic Games, both the men's and women's races were moved 500 miles to Sapparo and the races were started earlier in the morning than previously scheduled. Even so, the daily high temperatures were 93 degrees for the women's race and 90 degrees for the men's race. Using these temperatures, plot the probability of winning a medal for times between 2 hours and 2 3/4 hours (including every minute, so 120 minutes, 121 minutes...) for both male and female competitors.

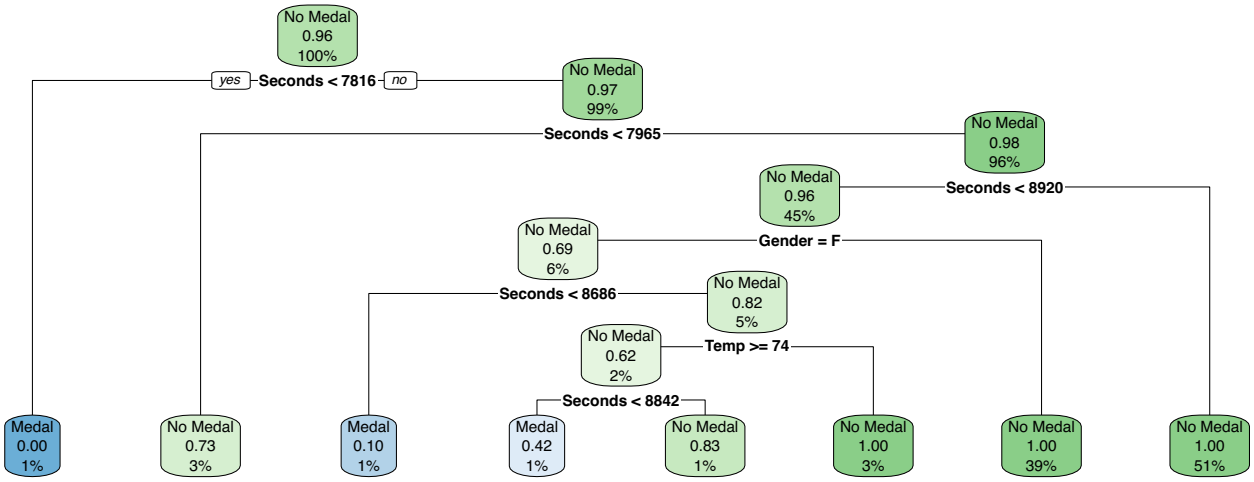
Consider that the 2021 medal winning times were 02:27:20, 02:27:36, and 02:27:46 for the women's race and 02:08:38, 02:09:58, and 02:10:00 for the men's race. Discuss your predictions in relation to the actual outcome and discuss whether you have any concerns about extending the results from the past Olympics to 2021 and potential differences due to COVID-19 or other factors.



Based on this model, the probability that medals were won with the women's 2021 medalists' times are estimated to be 60.2, 64.4, and 70.8 percent, and the men's 2021 medalists' times are estimated to be 66.0, 66.9, and 89.7 percent likely to get a medal. Those seems like reasonably high probabilities that I'm not concerned with the model being obviously not useful. It seems like COVID probably had little to no effect on this race based on the given data - if I suspected COVID had a large effect on finishing times, I would have expected to see competitors medaling with times that have low probability of medaling in pre-COVID competitions. If that were the case, it could be that top competitors weren't able to train as effectively, or weren't able to compete due to COVID.

If we had data on the year, and we saw a yearly trend of lower times being needed to medal (which is what I would expect to see), that would be better able to answer the question of how COVID affected the times required to likely win the marathon. With the given data, there's no better way to answer that question.

F. Decision Tree. Interpret the classification tree in the figure below and then compare and contrast your results from Parts C, D, and E with the classification tree.



The first split on 7816 seconds (130 minutes and 16 seconds) indicates that everyone who has ever run a marathon in under that time has medaled. The second split is if a time is between 130:16 and 132:45, there is a 26% chance of a that time winning a medal. The third split indicates that time over 8920 seconds (148:40) has never won a medal, which are just over half of the observations in the data set. The fourth split indicates that a male competitor has never medaled with a time over 132:45. All the remaining observations are female competitors with times under 148:40. The fifth split indicates that 9/10 women who finished under 144:46 have medal. The penultimate split showed that if the temperature is less than 74 degrees F, a time over 148:40 has never won a medal in the women's competition. The last split shows that if the temperature is over 74 degrees, 58% of the times between 144:46 and 147:22 have won a medal in the women's competition, or if between 147:22 and 148:40, 16.7% of the times won a medal if the temperature is over 74 degrees.

Tree based models have a lot of nested structure that is harder to capture in a linear model. This model is kind of similar to the generalized linear model I used, in that it uses the same predictors and there is an "interaction" between gender and temperature in both. My model explicitly uses an interaction term, and this tree contains that same information by having a split on gender as as the fourth split and temperature as the sixth split.

The main difference between the models is that the tree model basically separates out the men and women by first splitting out all the male medal winners before capturing any of the female medal winners. I considered doing something similar when I was modeling by fitting two GLM's for the two genders. The differences between the times in male medal winners and female medal winners are so different that it could be a good strategy if you're not interested in measuring the difference between men and women medal winning times.

It's hard to say which model is better, or if one is better. They're kind of just different - you could use hold out data (like the 2021 Olympic times) to see which classifies better, but that's about all the comparison I know that you could do.

G. Additional Thoughts. This dataset does not include additional information that could be useful, such as athlete and year. How would including this information change your model, the assessment of the model assumptions, and/or scope of inference?

I think the year could be a nice predictor for this model, which may capture some of the trend in times lowering over the years that may be missing from the current data set. That would help this model be useful for more than the next couple of Olympic races.

If information about the athletes, like age, past Olympic history, and nationality were included, a more complicated hierarchical model structure could be used that would incorporate that information and would give potentially better estimates for future competitions. That would also eliminate some concerns about

some of the rows containing the same athletes, so they're not completely independent observations from each other.

I pointed out concerns earlier about looking too deeply at the temperature coefficients. More data with different temperatures on race day would be the easiest way to alleviate those concerns.

Question 2

Continuing with the Olympic Marathon theme, assume that the distribution for Marathon times for Men's Olympic competitors can be modeled as:

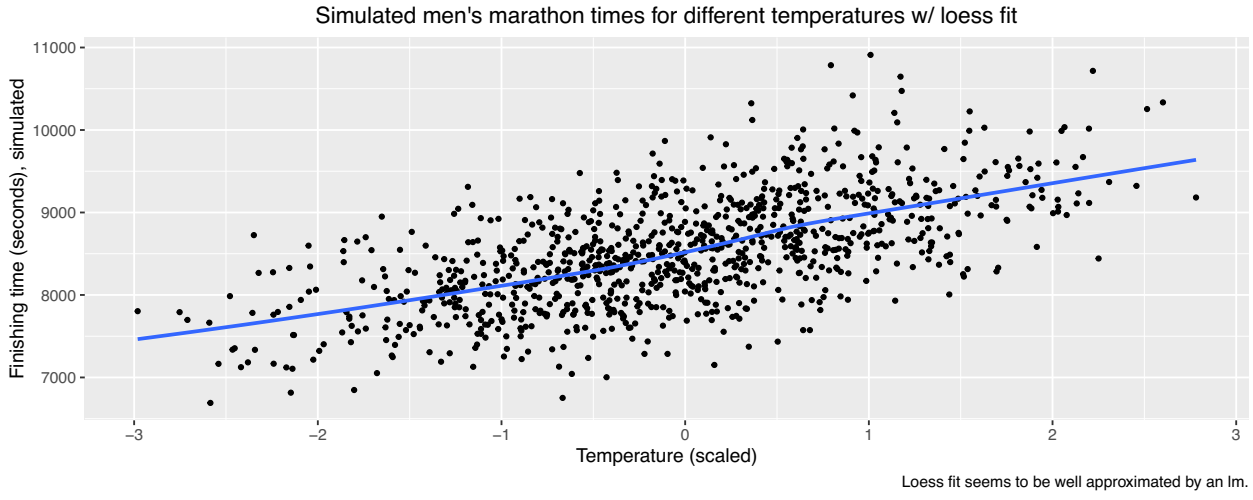
$$\text{Time} \sim \text{LogNormal}(9.05 + .05 * \text{Scaled Temp}, .06),$$

where the mean and standard deviation are, again, on the log scale and scaled temperature is $\frac{temp - mean(temp)}{sd(temp)}$.

A. Simulate a dataset from this model for a range of scaled temperature values (these should roughly be between -2 and 2) and create a data visualization to display the results. Explain and/or document your simulation code.

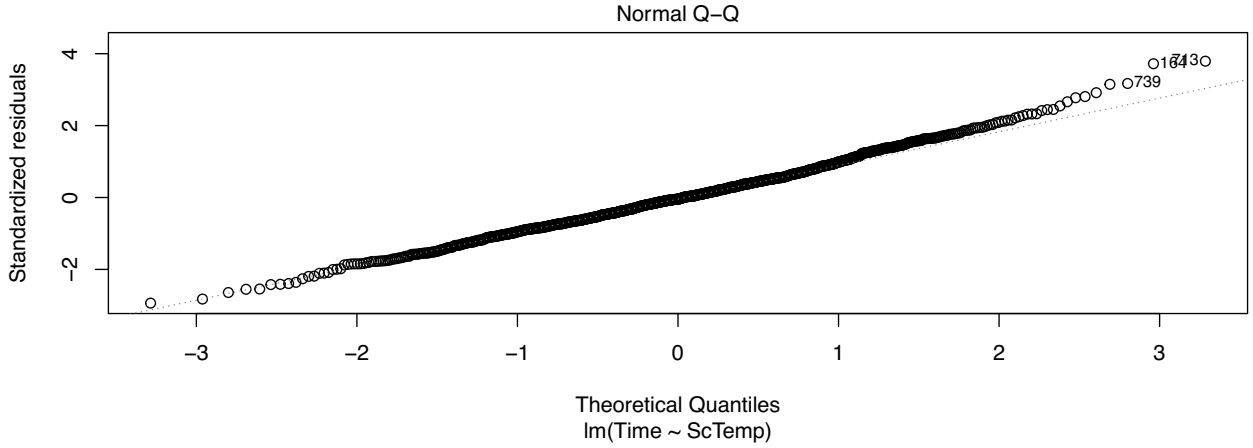
```
# How large of dataset - seems like the same size as the real data makes sense
n = nrow(marathon)
# Temp in the data set appears to be approximately normal
# To get values that are mostly between -2 and 2 a SD of 1 makes sense
# Simulate Times based on the model above
set.seed(8162021)
dat2 = tibble(ScTemp = rnorm(n, 0, 1) ,
              Time = rlnorm(n, 9.05 + 0.05*ScTemp, 0.06))

# Make visualization of Scaled Temp by Time with a loess fit
ggplot(data = dat2, aes(x = ScTemp, y = Time)) +
  geom_point(size = 1) +
  geom_smooth(formula = y~x, se = F, method = "loess") +
  labs(x = "Temperature (scaled)",
       y = "Finishing time (seconds), simulated",
       title = "Simulated men's marathon times for different temperatures w/ loess fit",
       caption = "Loess fit seems to be well approximated by an lm.") +
  theme(plot.title = element_text(hjust = 0.5))
```



B. Using your simulated dataset, fit a standard regression model (using the assumption of normal errors) with marathon time as the response variable and scaled temperature as a predictor. Assess the assumptions of this model and report your results. Discuss how the results compare to the known simulation values, including an interpretation of the parameter coefficients.

```
##
## Call:
## lm(formula = Time ~ ScTemp, data = dat2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1513.33  -348.98   -24.22   301.98  1952.54
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8541.02      16.48   518.3   <2e-16
## ScTemp        413.16      16.52    25.0   <2e-16
##
## Residual standard error: 515.5 on 979 degrees of freedom
## Multiple R-squared:  0.3897, Adjusted R-squared:  0.3891
## F-statistic: 625.1 on 1 and 979 DF,  p-value: < 2.2e-16
```



This linear model assumes that the observations are distributed iid $Time \sim N(\beta_0 + \beta 1_1 * Temp, \sigma^2)$. These normality assumptions are violated because I simulated the data from a lognormal distribution, so the residuals are not normally distributed nor will they be homoskedastic. This is shown in the normal q-q plot.

The normal q-q plot assumes that the residuals are normally distributed and plots the standardized residuals against that assumed normal distribution. The data should follow the trend line if the residuals are normally distributed. In the plot, the residuals tail off at the upper end, indicating that the upper tail are longer in the data than expected by the normal distribution assumption. This would be expected when fitting assumed normal data to a lognormal distribution, since a lognormal distribution has a much longer upper tail than a normal distribution.

In the standard scale, the estimate for the intercept, $\hat{\beta}_0 = 8541$, is the estimated marathon time in seconds when the temperature is the mean temperature in the data. The slope parameter, $\hat{\beta}_1 = 413$, is the estimated change in the marathon times due to a 1 sd increase in temperature (in seconds). If we transform the estimated $\hat{\beta}_0$ parameter to the log scale like the parameters in the lnorm function, then the estimate of 9.05 matches the parameter used to simulate the data. Unfortunately it's not possible to do that for the $\hat{\beta}_1$ parameter that estimates the slope due to temperature because the log makes it not an additive relationship (I think?).

$Y \sim N(\beta_0 + \beta_1 T, \sigma^2) = \beta_0 + \beta_1 T + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$

$X = \exp(Y) = \exp(\beta_0 + \beta_1 T + \varepsilon_i) = \exp(\beta_0 + \beta_1 T) + \varepsilon_i^*$

$\varepsilon_i^* \sim LN(0, \sigma^{*2})$

$X = \exp(\beta_0)\exp(\beta_1 T) + \varepsilon_i^*$

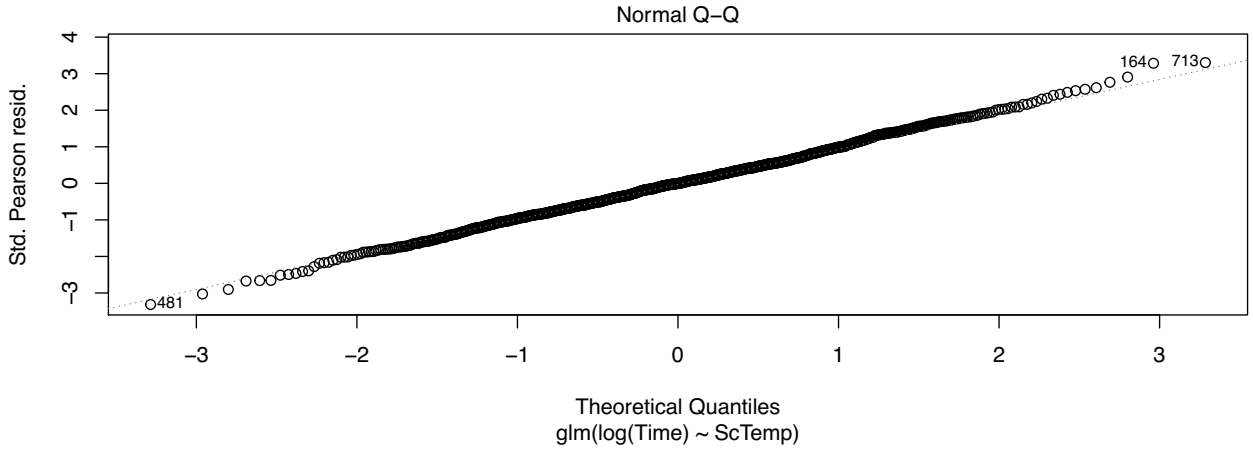
$\widehat{\exp(\beta_0)} = 8451$ $\quad \widehat{\exp(\beta_1)} = 451$?

$\beta_0 = 9.05$ $\quad \beta_1 = 0.05$

?

C. Using your simulated dataset, fit a more appropriate regression model that accounts for the known functional form from the simulation. Assess the assumptions of this model and report your results. Discuss how the results compare to the known simulation values, including an interpretation of the parameter coefficients.

```
##
## Call:
## glm(formula = log(Time) ~ ScTemp, family = gaussian(link = "identity"),
##     data = dat2)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.199675  -0.040638  -0.000309   0.037102   0.198838
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.049660   0.001926 4697.89   <2e-16
## ScTemp      0.048538   0.001932   25.13   <2e-16
##
## (Dispersion parameter for gaussian family taken to be 0.003630635)
##
##     Null deviance: 5.8470  on 980  degrees of freedom
## Residual deviance: 3.5544  on 979  degrees of freedom
## AIC: -2723.6
##
## Number of Fisher Scoring iterations: 2
```



You could use a SLR model to model the log of the response with respect to the temperature: $log(Time_i) = \beta_0 + \beta_1 * Temp + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$. This has the same assumptions as before, but since the $log(Y) \leq 0$ is undefined, Y values must be greater than 0. The q-q plot above shows the assumption that, after the log transformation, the data are normally distributed.

By definition, if $Y \sim N(X\beta, \sigma^2)$, *then* $log(Y) = X \sim lognormal(X\beta, \sigma^2)$, so taking the log of the Y values should transform into the correct distribution.

The coefficients can be interpreted as the intercept and slope for the log(Time), and both $\hat{\beta}_0 = 9.05$ and $\hat{\beta}_1 = 0.049$ agree closely with the known simulation values of 9.05 and 0.05.

Part III:

Question 1

The Olympic Marathon record of 2:06:32 was set in 2008 by Samuel Wanjiru of Kenya; however, Eliud Kipchoge recently broke the 2 hour barrier in the marathon.

Assume that Men's Olympic Marathon times, in seconds, can be adequately simulated by a log normal distribution (see code below) with (log) mean of 9.05 and (log) sd of .06.

```
## $start.arg
## $start.arg$meanlog
## [1] 9.052331
##
## $start.arg$sdlog
## [1] 0.06482685
##
##
## $fix.arg
## NULL
```

Assume there are 50 competitors in the race. For Parts A-C, use simulation to estimate the desired probability.

A. What is the probability of someone breaking the 2 hour barrier?

```
## [1] 0.004261
```

There is around a 0.4% chance of someone breaking the 2 hour barrier based on the parameters and assuming the distribution models it well.

B. What is the probability of two athletes breaking the 2 hour barrier in the same race?

```
## [1] 0.0008364
```

```
## [1] 0.0007132
```

The question is unclear if it is asking about exactly two or more than two athletes breaking the 2 hour barrier. There is approximately a 0.08% chance of there being more than two, and around 0.07% chance of exactly two athletes breaking the 2 hour barrier.

C. What is the probability of someone breaking the 2 hour barrier in any of the next 10 olympics?

```
## [1] 0.8824
```

Presumably both of these ways work (independence or something? They give the same answer) but there is approximately an 88% chance of someone breaking the 2 hour barrier in one of the next 10 Olympics.

Question 2

The findings in Part II: Question 2 are focused on the expected result, or the conditional mean, from Olympic competitors. However, athletes hoping to win a medal would be more interested in the 1st, 2nd, and 3rd places (or the 1st, 2nd, and 3rd percentiles in this scenario).

Using an OLS framework, the estimate for the regression coefficient vector $\underline{\beta}_{OLS}$ can be calculated as

$$\underline{\hat{\beta}}_{OLS} = \text{argmin}_{\beta} \sum (y_i - X\beta)^2$$

If rather than the conditional mean, we are interested in the conditional median, then the regression problem can be viewed through the lens of quantile regression, where

$$\underline{\hat{\beta}}_{q=.5} = \text{argmin}_\beta \sum |y_i - X\beta|$$

More generally, quantile regression can be formalized as

$$\underline{\hat{\beta}}_q = \text{argmin}_\beta \sum [(y_i - X\beta) \times (q - I(\{y_i - X\beta\} < 0))],$$

where q is a quantile ($\in (0, 1)$), $I()$ is an indicator function, and the term $(q - I(\{y_i - X\beta\} < 0))$ effectively weights positive and negative residuals.

A.

Use the Men's Olympic Marathon times (excluding temperature for now) in the regression model $y = \beta_0 + \epsilon$ to find the the conditional median ($\beta_{0,q=.5}$) and the conditional 1st percentile ($\beta_{0,q=.01}$)

In addition to presenting your results, write a short paragraph describing your approach.

```
mens_results <- marathon %>% filter(Gender == 'M') %>% dplyr::select(Seconds) %>% pull()

quantreg::rq(formula = mens_results~1, tau = c(0.5, 0.01)) %>% summary(se = "boot")
```

```
##
## Call: quantreg::rq(formula = mens_results ~ 1, tau = c(0.5, 0.01))
##
## tau: [1] 0.01
##
## Coefficients:
##             Value      Std. Error t value    Pr(>|t|)
## (Intercept) 7707.00000   48.88883  157.64338    0.00000
##
## Call: quantreg::rq(formula = mens_results ~ 1, tau = c(0.5, 0.01))
##
## tau: [1] 0.5
##
## Coefficients:
##             Value      Std. Error t value    Pr(>|t|)
## (Intercept) 8427.00000   20.49635  411.14642    0.00000
```

I considered trying to learn the calculus required to minimize the expression given in the question, or finding someone on the internet that has done that before, but the question doesn't ask us to do that. I'm guessing that in the case where β and X are vectors, this math problem simplifies significantly. I found the standard package "quantreg" created by Roger Koenker that does quantile regression. This package algorithmically solves for the $\underline{\hat{\beta}}_q$ that minimizes the weighted sum of absolute residuals.

Using that package with the rq() function, I found the top 1% of male marathon competitors finish in an estimated $\hat{\beta}_{0,q=.01} = 7707 \pm 50.7$ seconds, and the median runners finish in $\hat{\beta}_{0,q=.5} = 8427 \pm 17.9$ seconds.

B.

Write a short paragraph to describe how you would find $\underline{\beta}_q = (\beta_{0,q}, \beta_{1,q})$ for the model $y = \beta_0 + \beta_1 \times Temp + \epsilon$.

Very easily. I would use the rq() function in the quantreg package and fit a linear model with that, as such:

```
# Assuming wanting median and 1st percentile times
marathon %>%
  filter(Gender == "M") %>%
```

```
  dplyr::select(Seconds, Temp) %>%
  quantreg::rq(data = ., formula = Seconds ~ Temp, tau = c(0.5, 0.01)) %>% summary(se = "boot")
```

```
##
## Call: quantreg::rq(formula = Seconds ~ Temp, tau = c(0.5, 0.01), data = .)
##
## tau: [1] 0.01
##
## Coefficients:
##             Value      Std. Error t value    Pr(>|t|)
## (Intercept) 8216.90909  634.87069   12.94265    0.00000
## Temp          -6.45455    8.02423   -0.80438    0.42155
##
## Call: quantreg::rq(formula = Seconds ~ Temp, tau = c(0.5, 0.01), data = .)
##
## tau: [1] 0.5
##
## Coefficients:
##             Value      Std. Error t value    Pr(>|t|)
## (Intercept) 7986.00000  302.87799   26.36705    0.00000
## Temp           5.40000    3.73230    1.44683    0.14855
```

To be more specific, if I assume I have an analytical solution to $\hat{\beta}_q$, it should be reasonably simple to extend it to the case when X is an $n \times 2$ matrix and β is a 2×1 vector.

The estimated time for Olympic marathon competitors in the 1st percentile is $8217 - 6.45 * Tempature$ seconds, which is very interesting because the slope due to the temperature is negative, implying that the top 1% of runners actually run faster when it's warmer out based on these data. For median Olympic competitors, the estimated time is $7986 + 5.40 * Temperature$, so the average competitor slows down when the temperature is higher. For both the 1st and 50th percentile Olympic competitors, the effect on finishing time due to the temperature is is not large compared to the estimated standard error.

Part 1 sources:

Text:
Statistical Inference, Casella & Berger

Notes:
501 & 502 course notes, Hancock

Web:

https://math.stackexchange.com/questions/2542269/the-difference-of-two-order-statistics-of-exponential-distribution

https://math.stackexchange.com/questions/2764443/for-a-random-sample-from-the-distribution-fx-e-x-theta-x-theta-sh/2793341#2793341

.https://math.stackexchange.com/questions/3206984/basus-theorem-to-show-independence?rq=1

https://math.stackexchange.com/questions/509816/order-statistics-of-n-i-i-d-exponential-random-variables

Part 2 & 3 sources:

Text:

Regression and Other Stories, Gelman, Hill & Vehtari
Data Analysis Using Regression and Multilevel/Hierarchical Model, Gelman & Hill

Notes:

505 & 506 course notes, Hoegh

Stat 405 course notes, Bergen