STAT 505: Midterm Exam
Name:

1. **Format**: Submit the exam to GitHub and include the R Markdown code and a PDF file with output. Please verify that all of the code has compiled and the graphics look like you think they should on your files, you are welcome to upload image files directly if they look distorted in the output file.
2. **Advice**: Be sure to adequately justify your answers and appropriately reference any sources used. Even if you are not able to answer a question completely, do your best to provide an answer and discuss solutions that you tried. For full credit, include your code and graphics for each question and create neat output by using options like `kable()` for tables and writing results in line with R commands.
3. **Computer Code / Reproducibility:** Please turn in all relevant computer code to reproduce your results; a reproducible document is a requirement. Include all relevant code and output needed to answer each question and write an answer to each question. Even if the answer seems obvious from the output, make sure to state it in your narrative as well.
4. **Resources and Citations:** While the exam is open book and you can use any resources from class or freely available on the internet, this is strictly an individual endeavor and **you should not discuss the problems with anyone outside the course instructor including class members.** All resources, including websites, should be acknowledged.
5. **Exam Questions:** If clarification on questions is required, please email the course instructor: andrew.hoegh@montana.edu.
6. **A note on sharing / reusing code:** This is a huge volume of code is available on the web to solve any number of problems. For this exam you are allowed to make use of any online resources (e.g., StackOverflow) but you must explicitly cite where you obtained any code you directly use (or use as inspiration). Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism. All communication with classmates is explicitly forbidden.

Academic Honesty Statement

Include the following statement at the beginning of your submission.

> I, ___ (your full name here) ___, hereby state that I have not communicated with or gained information in any way from my classmates or anyone other than the course instructor during this exam, and that all work is my own.

In the event that you have inadvertently violated the above statement, you should not sign above and instead discuss the situation with the course instructor.

Synthetic Data Question (24 points)

Consider data generated from a study with repeated measures where individual participants have multiple responses. For instance, assume we are interested in modeling the time spent skiing hully gully for children and adults. Hully gully is a narrow gully at Bridger Bowl - the best comparison would be trying to ski through a hallway in Wilson during passing time between classes.

Formally, we will generate data from a model that includes coefficients for both the age and the specific individual. This model, is known as a random effects model, is specified as:

$$\begin{aligned} y_{ij} =& \beta_0 + \beta_1 \times x_{i,group=child} + \gamma_i + \epsilon_{ij} \\ \gamma_i \sim& N(0, \tau^2) \\ \epsilon_{ij} \sim& N(0, \sigma^2) \end{aligned}$$

where y_{ij} is the time for the j^{th} run for the i^{th} skier, β_0 is the expected time for an adult, β_1 is the expected time difference for a child, $x_{i,group=child}$ is a binary indicator for whether the i^{th} skier is a child, γ_i is a random effect associated with the i^{th} skier, and ϵ_{ij} is a random error.

In particular we will simulate times for 12 skiers to ski through Hully Gully, where each skier takes 5 runs.

```r
set.seed(11042022)
beta <- c(100,-20)
gamma <- rnorm(n = num_skiers, mean = 0, sd = 20 )
epsilon <- rnorm(n = num_skiers * num_runs, mean = 0, sd = 5)
ski_times <- tibble(skier_id = factor(rep(1:(num_skiers), each = num_runs)),
                    skier_run = rep(1:num_runs, num_skiers),
                    group = rep(c('adult','kid'), each = num_runs * num_skiers / 2),
                    x_child = rep(c(0,1), each = num_runs * num_skiers / 2)) %>%
  mutate(time = beta[1] + beta[2] * x_child + rep(gamma, each = num_runs) + epsilon)
```

1. (4 points)

Use the `ski_times` dataset and create a figure to compare the times for adults and kids. For full credit include informative labels, titles, and captions **AND** distinguish points for each skier (using color and/or symbols).

2. (4 points)

Consider a naive model, `lm_naive`, that doesn't control for individuals.

```r
lm_naive <- lm(time ~ group, data = ski_times)
```

Assess the residuals and comment on whether this model framework satisfies the independence assumption of linear models. Provide visual support of your claim.

3. (4 points)

An alternative model to the naive model, fit in question 2, is to average the responses for individuals. Using the `ski_times_reduced` dataset interpret the model parameters and compare with what you'd expect from the simulation.

```
ski_times_reduced <- ski_times %>% group_by(skier_id, group, x_child) %>%
  summarize(time = mean(time), .groups = 'drop')

lm_reduced <- lm(time ~ group, data = ski_times_reduced)
```

4. (4 points)

State and assess the the assumptions for `lm_reduced`.

5. (4 points)

Create a figure that displays the `ski_times_reduced` data along with the model fit for `lm_reduced`. Similar to what we've done in class, this figure should include the model parameters, and uncertainty in them.

6. (4 points)

Compare and contrast the output from these two models. Which do you think is more appropriate? Why?

```
summary(lm_reduced)
```

```
##
## Call:
## lm(formula = time ~ group, data = ski_times_reduced)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.484  -6.353   2.945  14.388  20.539
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  102.715      8.115  12.657 1.77e-07 ***
## groupkid     -25.690     11.477  -2.238   0.0491 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.88 on 10 degrees of freedom
## Multiple R-squared:  0.3338, Adjusted R-squared:  0.2672
## F-statistic: 5.011 on 1 and 10 DF,  p-value: 0.04913
```

```
summary(lm_naive)
```

```
##
## Call:
## lm(formula = time ~ group, data = ski_times)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.618  -6.820   2.814  13.384  26.078
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  102.715      3.441  29.850  < 2e-16 ***
## groupkid     -25.690      4.866  -5.279 2.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.85 on 58 degrees of freedom
## Multiple R-squared:  0.3245, Adjusted R-squared:  0.3129
## F-statistic: 27.87 on 1 and 58 DF,  p-value: 2.028e-06
```

Modeling Questions (26 points)

For this dataset we will use a dataset containing information from a Garmin fitness tracker.

```
Garmin <- read_csv("https://raw.githubusercontent.com/STAT505/2022midterm/main/Garmin_clean.csv")
```

There are five variables in this datasest:

- *heart_rate*: Numeric response for instantaneous heart rate in beats per minute
- *speed_past_60_seconds*: Numeric response for average speed over the past 60 seconds, in km / hour
- *avg_temp*: Numeric response for average temperature, in celsius, for the entire run.
- *slope*: Categorical response for whether the last 60 seconds had been `flat`, `uphill`, or `downhill`.
- *walk*: Binary response for whether the last 60 seconds has been exclusively running (`run`) or if any walking (`walk`) had occurred. **continue here**

1. (4 points)

Make a figure, or more likely a panel of figures, to explore the relationship between `heart_rate` and the other four variables. Note `grid.arrange()` allows you to combine `ggplot2()` graphics.

Include a summary paragraph to discuss the potential relationships between each predictor and `heart_rate`.

2. (2 points)

Based on your figure from Question 1, which variables do you believe are helpful to explain heart rate?

3. (20 points)

For this set of questions, use information from Question 1 and Question 2 and fit a model to explain `heart_rate()`. Note when considering models, the additivity and linearity assumption would encourage you to assess the presence of interactions.

a. (4 points) Using complete notation, write out the model you selected.

b. (4 points) With an eye on model assumptions, justify the model you selected.

c. (4 points) Include one critique of your model.

d. (4 points) Human heart-rates are generally in the range between 45 beats per minute and 200 beats per minute. Using your model, create distributional predictions for heart rate for the following scenarios:

- `speed_past_60_seconds = 0`, `avg_temp = 15`, `slope = up`, `walk = run`
- `speed_past_60_seconds = 22`, `avg_temp = 15`, `slope = up`, `walk = run`

Hint you can use `posterior_predict()` for `stan_glm` models or `predict(interval = 'prediction')` for `lm()` models.

e. (4 points) Summarize the findings from your model - the audience should be an avid runner with minimal background in statistical analysis.

Simulation Question (8 points)

This section will focus on a 2-way ANOVA model with an interaction term. Formally this model can be written as

$$y_i = \beta_0 + \beta_1 \times x_{1i,I(group=2)} + \beta_2 \times x_{2i,I(group=red)} + \beta_3 \times x_{1i,I(group=2)} x_{2i,I(group=red)} + \epsilon_i$$

where $x_{1i,I(group=2)}$ is an indicator variable so the i^{th} observation of x_1 is group 2, $x_{2i,I(group=red)}$ is an indicator variable so the i^{th} observation of x_2 is group "red", $\epsilon_i \sim N(0, \sigma^2)$

1. (4 points)

Provide detailed descriptions of $\beta_0, \beta_1, \beta_2, \beta_3$, and σ^2.

2. (4 points)

Write code to simulate data from a model with two categorical variables. You can consider x_1 to have two categories `1` and `2` and x_2 has two categories: `blue` and `red`. Create a plot that visualizes the interaction.