Final Exam

Instructions

- **Due Date:** This exam is due by 5pm on April 28th.
- **Open Book:** This is an open book exam. You may consult your course materials, textbooks, and class notes. However, you are not allowed to collaborate with your classmates or seek help from others.
- **No collaboration:** Collaboration with other students or any form of academic dishonesty is strictly prohibited. Submitting work that is not your own or using unauthorized resources will result in severe penalties, including a failing grade on the exam and potential disciplinary action.
- **Exam Format:** There are four questions, with a total of 100 points possible. You are required to answer all questions.
- **Submission Format:** *Use this markdown file to complete the exam.* You should submit both (i) the compiled pdf and (ii) the markdown code used to generate the output. Both should be emailed to me. *Your markdown file must compile to generate exactly the output in the pdf, and should run all of your code!*
- **Show your work:** For questions that involve calculations, derivations, or proofs, you must show your work to receive full credit. Clearly explain your reasoning and methodology for each step. For questions that involve statistical analysis, working code must be submitted. Partial credit may be awarded for incomplete or partially correct answers, depending on the quality of your work.
- **Quality of Writing:** Your answers should be well-organized, clear, and concise. Use proper grammar, punctuation, and spelling. Points may be deducted for poor presentation or lack of clarity.
- **Answers should use the solution environment.** An example of this is given below.

Solution:

Here is a solution, with some `R` code.

```
set.seed(1111)
mean(rnorm(1E5)^2)
```

```
## [1] 1.0034
```

Good luck!

Question 1

In this problem, we will study some limitations of the nonparametric bootstrap. Consider $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}(0, \theta)$ where θ is the parameter of interest.

a. *Find the maximum likelihood estimator of θ under this model.*

b. Taking $n = 1000$, simulate $k = 200$ synthetic datasets with $\theta = 1$. Using $B = 200$ resamples, compute a *quantile interval* for θ for each of the $k = 200$ samples using the nonparametric bootstrap (i.e., use the quantiles of the bootstrap resamples of $\widehat{\theta}$ directly to make a confidence interval). In addition to this, for subsequent parts, return the bootstrap standard error and the estimate of θ on the data. *What is the coverage of the quantile interval? Why does this occur?*

c. What is the coverage of the interval $\widehat{\theta} \pm 2\,\widehat{\text{SE}}(\widehat{\theta})$ with the standard error estimated according to the nonparametric bootstrap? *Critique the performance of this interval.*

d. *Find the asymptotic distribution of $\widehat{\theta}$. Specifically, show that $n(1 - \frac{\widehat{\theta}}{\theta}) \to \text{Exponential}(1)$.*

e. Note that, in part (d), we have shown that the rate of convergence is actually n^{-1} as opposed to the familiar (parametric) rate of $n^{-1/2}$. When rates other than the parametric rate $n^{-1/2}$ occur, we generally can't have much confidence in the intervals output by an unmodified bootstrap procedure.

Compare the coverage of the intervals you have computed so far to one that makes use of the asymptotic distribution given above (i.e., using the limiting exponential distribution to derive a confidence interval). That is, use the fact that

$$\Pr\left\{L \leq n\left(1 - \frac{\widehat{\theta}}{\theta}\right) \leq U\right\} = 1 - \alpha$$

where L and U are chosen to be appropriate quantiles of the Exponential(1) distribution to form your confidence interval, and evaluate the performance of this interval.

f. Generate the following dataset

```
set.seed(11111)
X <- runif(1000)
```

and use the bootstrap to estimate the distribution of $Z = n(1 - \widehat{\theta}/\theta)$. *Plot the cdf of this distribution and compare it with the exponential cdf. How well does the bootstrap do at estimating the cdf of Z?*

Question 2

The following dataset is taken from the book *Categorical Data Analysis* by Alan Agresti. He describes the dataset as follows:

> This dataset concerns a study of female horseshoe crabs on an island in the Gulf of Mexico. During spawning season, the females migrate to a shore to breed, with a male attached to her posterior spine, and she burrows into the sand and lays clusters of eggs. During spawning, other male crabs may group around the pair and may also fertilize the eggs. These male crabs that cluster around the female crab are called *satellites*.

Our goal is to learn the factors that lead to a female crab having satellites, and how many if so. First, we load the dataset:

```
data(crabs)
crabs <- crabs %>% mutate(has_satell = ifelse(satell > 0, 1, 0))
crabs %>% head() %>% knitr::kable()
```

color	spine	width	satell	weight	has_satell
2	3	28.3	8	3.05	1
3	3	22.5	0	1.55	0
1	1	26.0	9	2.30	1
3	3	24.8	0	2.10	0
3	3	26.0	4	2.60	1
2	3	23.8	0	2.10	0

There are two variables of interest here: `satell`, which is the number of satellites a female crab has, and `has_satell`, which is just a binary indicator that `satell > 0`. Then, there are other features of the female crab: her color (from `1` to `4`, which is an ordinal variable going from medium-light color to dark color), her weight in kilograms, her width in centimeters, and spine condition (going from `1` if both spines are good, `2` if one is worn or broker, and `3` if both worn or broken).

a. Build a model for `satell` as a function of `color`, `spine`, `width`, and `weight` using a GLM of your choice. Justify the form of your model and how you treat the different predictors as input into the model. Interpret the coefficients of the fitted model. Which variables seem to be the most important in terms of a female crab attracting satellites?

b. *Criticize the model you fit.* Does the GLM you chose to fit seem appropriate? What issues might there be with it? Use whatever tools for answering this question that you deem fit.

c. *Based on your answer to part (b), fit a model that fixes the issue you identified, and discuss how this affects the substantive conclusions you drew.* I'm not looking for any answer in particular, as there are multiple different ways you could have answered part (b).

d. A common feature of count data is *zero-inflation*, in which there are more 0s in the data than would be expected from (say) a Poisson model. Rather than looking at `satell`, let us now focus on the binary variable `has_satell`.

Fit a binary logistic regression model with ***`has_satell`*** *as a response. Then, compare the predicted probability of* ***`has_satell == 0`*** *to the predicted probability of* ***`satell == 0`*** *from the model you fit in part (a). Which model does a better job of predicting 0s? Does there appear to be zero-inflation?* **Hint:** recall that, for a Poisson model, the probability of $Y_i = 0$ is $e^{-\mu_i}$.

e. Perform an analysis of deviance to assess the importance of `color` in predicting `has_satell`.

Question 3

This question concerns the `muscatine` dataset in the `geepack` package. We first load the dataset:

```
muscatine <- geepack::muscatine
muscatine %>% head() %>% knitr::kable()
```

id	gender	base_age	age	occasion	obese	numobese
1	M	6	6	1	yes	1
1	M	6	8	2	yes	1
1	M	6	10	3	yes	1
2	M	6	6	1	yes	1
2	M	6	8	2	yes	1
2	M	6	10	3	yes	1

The documentation describes this dataset as follows:

> The data are from the Muscatine Coronary Risk Factor (MCRF) study, a longitudinal survey of school-age children in Muscatine, Iowa. The MCRF study had the goal of examining the development and persistence of risk factors for coronary disease in children. In the MCRF study, weight and height measurements of five cohorts of children, initially aged 5-7, 7-9, 9-11, 11-13, and 13-15 years, were obtained biennially from 1977 to 1981. Data were collected on 4856 boys and girls. On the basis of a comparison of their weight to age-gender specific norms, children were classified as obese or not obese.

For the purpose of this question, our interest is in understanding how the risk of obesity varies across age and gender. The variables we are interested in are:

- `id`: the id of the child (rows with the same `id` correspond to the same child).
- `gender`: male or female.
- `base_age`: the age of the child at the start of the study.
- `age`: the age of the child at the time of measurement.
- `numobese`: binary indicator, 1 if obese and 0 otherwise.

a. Fit a generalized linear mixed effect model (possible software options include `stan_glmer` from the `rstanarm` package or `glmer` from the `lme4` package) that accounts in some way for the fact that measurements are taken on the *same* individuals over time by using a random intercept. Assume a linear trend in `age`. Based on this result, does there appear to be evidence that the risk of obesity varies across gender? (**Note:** the model fitting process may be more stable if you center and scale `age` prior to fitting the model.)

b. The model fit in **a.** assumes a rather unrealistic relationship between `age` and the probability of obesity. Instead of doing this, model the relationship between `age` and `numobese` in a flexible fashion (using splines is an option here, although not the only one). *Does this modified model lead to a different conclusion regarding the importance of `gender`? What if `gender` is allowed to interact with `age`?*

c. Do you think that a single random intercept for each individual is likely to be a good assumption for this data? Try fitting the model with both a random intercept and a random slope (i.e., for individual j add a term $a_j + t\,b_j$). What impact does this have on your results? (**Note:** try your best to fit this model, but don't worry too much if it turns out to be too difficult/expensive.)

d. Suppose that I wanted to compute $\Pr(\texttt{numobese} = 1 \mid \texttt{age}, \texttt{gender})$ for all combinations of `age` and `gender` in the dataset (just for the sake of comparing the obesity risk over time for both genders). Without actually doing this, state how we could go about doing this from the output of the model. Is it possible to do this using only the output of `summary(my_fitted_glmm)`? Is it *easy* to do this with the output?

e. Now, use a GEE to accomplish the goals outlined in part **d.** Your answer should include the following.

- A discussion of the correlation structure you chose for the model.

- A discussion of the pros and cons of the GEE relative to the GLMMs you have fit, both in general and as it pertains to the task at hand.
- A plot of the probability of obesity for each age and gender, with appropriate error bars. (Note: if using the `geeglm` function, you can compute the standard error of the linear predictor at x as $\sqrt{x^\top V x}$ where $V =$ `vcov(myfit)`.)
- A discussion of the results and how the results of the fitted GEE differ (or don't) from the GLMMs you fit.

Question 4

The `weather` dataset contains two variables from 147 weather stations in the American Pacific northwest:

- `pressure`: the difference between the forecast pressure and the actual pressure reading at that station (in Pascals).
- `temperature`: the difference between the forecast temperature and the actual temperature reading at that station (in Celsius).

There are also latitude and longitude coordinates for each station. To load the data, run

```
weather_path <- str_c("https://raw.githubusercontent.com/",
                      "jgscott/SDS383D/master/data/weather.csv")
weather <- read_csv(weather_path)
```

We will use models of the form

$$Y_i \sim \text{Normal}\{\mu(\texttt{lat}_i, \texttt{lon}_i), \sigma^2\}$$

with a Gaussian process prior for $\mu(\cdot, \cdot)$. It may be helpful to reuse the code you wrote to fit Gaussian process models from the Chapter 6 notes.

a. Fit a Gaussian process (using whatever software you want) for both `pressure` and `temperature` using the squared exponential covariance

$$K(x, x') = \tau^2 \exp\left\{-\frac{(\texttt{long} - \texttt{long}')^2 + (\texttt{lat} - \texttt{lat}')^2}{2h^2}\right\}.$$

Use an appropriate technique for choosing the hyperparameters. What values of (σ, τ, h) do you estimate for the two models?

b. Visualize the fitted Gaussian process on a fine grid of latitude and longitude variables. Consider using the `filled.contour`, `contourplot`, or `geom_contour_filled` functions to aide in your visualization. Comment on qualitative features that pop out of your visualizations.

c. One consideration in covariance modeling is whether all the variables should share the same bandwidth or should use separate bandwidths. Specifically, we might set

$$K(x, x') = \tau^2 \exp\left\{-\frac{(\texttt{lat} - \texttt{lat}')^2}{2h_{\texttt{lat}}^2} - \frac{(\texttt{long} - \texttt{long}')^2}{2h_{\texttt{long}}^2}\right\}$$

Repeat part **a.** and part **b.** using this modified kernel, estimating both $h_{\texttt{lat}}$ and $h_{\texttt{long}}$ rather than a single bandwidth. Then answer the following questions:

- Does the decision to use different bandwidths lead to different substantive results for the two contour plots? Explain.
- Look at the negative marginal log-likelihoods of the four models you fit. Do the models with separate bandwidth parameters seem to perform much better than the models with a single bandwidth parameter? Explain.