STAT 505: Final 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 (18 points)

A common assumption that we've talked about, in great detail, throughout the course is that the functional form of the model adequately captures the relationship between the predictors and the data.

This set of questions will explore a few model scenarios for logistic regression by looking at graphical displays of data, fitting models, and evaluating model diagnostics.

Consider three models.

1. Logistic Regression

$$\begin{aligned} y_i &\sim Bernoulli(p_i) \\ \log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 x_i, \\ \text{where } \beta_0 = -5, &\text{ and } \beta_1 = .25 \end{aligned}$$

2. Logistic Regression with logarithmic relationship

$$\begin{aligned} y_i &\sim Bernoulli(p_i) \\ \log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 \log(x_i), \\ \text{where } \beta_0 = -10, &\text{ and } \beta_1 = 4 \end{aligned}$$

3. Logistic Regression with quadratic relationship

$$\begin{aligned} y_i &\sim Bernoulli(p_i) \\ \log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2, \\ \text{where } \beta_0 = -5, \beta_1 = .5 &\text{ and } \beta_2 = -.01 \end{aligned}$$

1. (2 points)

With logistic regression, the linearity assumption is between Xβ and the log-odds of y. Generally we don't know p_i, as this is what we are modeling, so it is more difficult to visualize this relationship. With this question, we will simulate data, where we know p_i, in order to visualize this relationship.

Using the model parameters specified above and n and X specified below. Plot the relationship between log-odds and x for each of the three models. Note, I've started the first model for you.

```
n <- 5000
x <- seq(.01, 50, length.out = n)

beta <- c(-5, .25)
p1 <- invlogit(beta[1] + beta[2] * x)
log_odds <- logit(p1)
y1 <- rbinom(n,1,p1)
```

2. (4 points)

We don't generally know p, but rather we are typically given y and x. Simulate binary responses (y) from the three models and plot the y vs x with a LOESS curve, the true model curve, and the curve estimated from $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$ (`glm(y ~ x, family = binomial)`), note this will be the incorrect model specification for models 2 and 3.

3. (4 points)

For all models, fit $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$ (`glm(y ~ x, family = binomial)`) and create residual diagnostic plots. Comment on whether you see any structure in the residual diagnostic figures *and* if this matches with your expectation.

4. (4 points)

Now fit the correctly specified model for each of the three scenarios, create residual diagnostic plots, and comment on whether you see any structure in the residual diagnostic figures *and* if this matches with your expectation.

5. (4 points)

Based on what you've seen in the previous questions, write a short paragraph summarizing the implications of violating the functional form assumption in logistic regression. In that paragraph address the following question: "What, in particular, would concern you about using one of the misspecified models in Question 3?"

Modeling Questions (22 points)

For this question, we will use a similar dataset to that of Project 2. Instead of the binary variable focused on whether a house costs more than $1,000,000 dollars we have actual price along with zipcode.

```
KingCo <- read_csv("https://raw.githubusercontent.com/STAT505/FinalExam/main/KingCo_wPrice.csv")
```

Part 1. (11 points)

For this question we will explore how zipcode impacts the average housing price. (Note: for Part 1, you only need to consider zipcode and price, but Part 2 will allow you to make use of all predictors.)

A. (3 points) Create a figure to display housing prices across zipcodes.

B. (4 points) Fit an ANOVA model to address how zipcode impacts the average housing price. Interpret the coefficients in the model and address how zipcode impacts average housing prices.

C. (4 points) List the assumptions of this model and indicate whether you have any concerns with the model assumptions.

Part 2. (11 points)

Now we will consider all variables in the dataset with the goal of producing the model that best predicts housing prices. Our interest in this case is more predictive than explanatory.

A. (3 points) Write out the statistical model that you've selected with formal and complete notation.

B. (4 points) Justify this model - why did you select this specification as opposed to others?

C. (4 points) Use your model to predict the housing price, along with uncertainty bounds, for a home with:

1. 3 bedrooms, 2 bathrooms, 2000 sqft_living, 70,000 sqft_lot, 1 waterfront, 98039 zipcode
2. 3 bedrooms, 2 bathrooms, 2000 sqft_living, 70,000 sqft_lot, 1 waterfront, 98032 zipcode
3. 1 bedrooms, 1.5 bathrooms, 1000 sqft_living, 6000 sqft_lot, 0 waterfront, 98102 zipcode
4. 4 bedrooms, 2.5 bathrooms, 2600 sqft_living, 50,000 sqft_lot, 0 waterfront, 98014 zipcode