STAT 505: Final Exam
Name:

Please turn in 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.

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 group mates or class members.** All resources, including websites, should be acknowledged.

Logistic Regression (16 points)

With generalized linear models, the specified functional form (linearity and additivity) of the model is still an important assumption.

1. (4 points)

Given the model, specified as

$$y_i \sim Bernoulli(p_i) \tag{1}$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i \tag{2}$$

create a figure and write a short summary describing how the linearity assumption relates to x_i and p_i.

2. (4 points)

Suppose the true generative model is

$$y_i \sim Bernoulli(p_i) \tag{3}$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2. \tag{4}$$

Data can be simulated from this model using the following code.

```
set.seed(12062021)
n <- 500
beta <- c(2, 1, -1)
x <- runif(n, -3, 3)
X <- cbind( rep(1,n), x, x^2)
p <- plogis(X%*%beta)
```

```
y <- rbinom(n, 1, p)
quad_logistic <- tibble(x = x, x_sq = x^2 , y = y)
```

Create an exploratory data visualization to explore the relationship between x and y. Add a title and caption that comments on the functional form between x and y.

3. (4 points)

Using the data simulated in question 2, fit two separate logistic regression equations. The first model will use just (`y ~ x`) and the second model uses (`y ~ x + x_sq`). Use a model selection tool to choose the best model. Justify your results.

4. (4 points)

Create a plot of the fitted model results for both models in part 3 (ideally including the data). Compare those to the true known model fit and comment on the differences.

Data Analysis: Used Volkswagon prices

This question will use a dataset that contains sales price of used Volkswagon (VW) vehicles in Great Britain.

The dataset has been filtered to include the following information:

- `price`: sales price in British pounds

Predictor Variables:

- `model`: type of car, (Golf, Passat, and Tiguan)
- `year`: model year the car was produced
- `mileage`: miles on used vehicle at time of sale
- `fuelType`: how vehicle is powered (petrol, diesel, hybrid)

1. (8 points)

For each predictor variable, create a figure that shows the relationship between that variable and price. Each figure should contain an informative caption. Figures can be paneled together.

2. (4 points)

Create at least two figures to explore possible interactions between predictor variables.

3. (4 points)

Consider the following model output. Why are there `NA` values in the table? How problematic are those `NA` values?

```
lm(price ~ model * fuelType, data = vw) %>% summary()
```

```
##
## Call:
## lm(formula = price ~ model * fuelType, data = vw)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17761.4  -4101.8   -569.7   3506.8  25255.2
##
## Coefficients: (1 not defined because of singularities)
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  14588.1      126.7 115.175  < 2e-16 ***
## modelPassat                    145.7      251.7   0.579    0.563
## modelTiguan                   6895.1      201.2  34.276  < 2e-16 ***
## fuelTypeHybrid                7761.1      626.7  12.384  < 2e-16 ***
## fuelTypePetrol                3476.6      167.8  20.717  < 2e-16 ***
## modelPassat:fuelTypeHybrid    5127.4     1002.6   5.114 3.23e-07 ***
## modelTiguan:fuelTypeHybrid       NA         NA      NA       NA
## modelPassat:fuelTypePetrol    2826.0      531.9   5.313 1.11e-07 ***
## modelTiguan:fuelTypePetrol   -2979.1      365.1  -8.159 3.92e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5725 on 7484 degrees of freedom
## Multiple R-squared:  0.2021, Adjusted R-squared:  0.2014
## F-statistic: 270.8 on 7 and 7484 DF,  p-value: < 2.2e-16
```

4. (8 points)

Assume that the your goal is to build the best model to predict the sales price of used VW vehicles. Write a few paragraphs (that could slot into a statistical report) that address the statistical procedures. In particular:

- Define model to fit with complete notation
- Defense of model choice

5. (8 points)

Building on the last question, write a few paragraphs (that could slot into a statistical report) that address the statistical results. In particular:

- Discuss model results in the context of the purchasing a used vehicle
- Summarize estimates from final model including uncertainty

5. (4 points)

Assume that your (non-statistician) roommate has accepted a job in Great Britain. They have narrowed their vehicle choices to

- a 2017, VW Golf Hybrid, with 20,000 miles and
- a 2019, VW Passat Hybrid, with 25,000 miles

Provide a range that they should expect to pay for each of these vehicles.