CPSC 541 Exam 2

Possible extra questions

Question 1

Let's say you fit a simple linear model and you have one observation that falls exactly on your prediction line. What would happen if you removed that point? Make sure you addess:

- The residual sums of squares
- The mean squared error
- The parameter estimates (β_0 and β_1)

There are multiple ways to address this problem, and I will accept anything from empirical-style answers to mathematical proofs (if you really want to write out a proof for some reason).

Question 2

Here we have output from a logistic model based on a familiar dataset. The model is:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \cdot ringcat + \beta_2 \cdot shell + \beta_3 \cdot shell \cdot ringcat$$

Note that `ring_cat` is a categorical variable with 2 levels. `ring_cat8+` is all abalones with ring counts of 8 or higher and this category ***is not*** the reference level. Assume also that `mature` has two levels: not mature (0) and mature (1). Please predict the ***probability*** that an abalone is mature if it has 4 rings and a shell weight of 0.25.

Then, please simulate 10 new observations for this prediction level.

Please include your math (or code) for the prediction. Please include your code for the simulation.

This problem is solvable without running the model, however, you can always run the model if you like. I promise that I only made 1 (hopefully obvious) change from the `logistic_and_simulations.rmd` file (available on compass).

```
##
## Call:
## glm(formula = mature ~ ring_cat * shell, family = binomial(),
##     data = data3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5566  -0.5744   0.3407   0.6393   2.3167
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.6920     0.1862 -14.455  < 2e-16 ***
## ring_cat8+         1.2320     0.2243   5.492 3.96e-08 ***
## shell             15.8324     1.6528   9.579  < 2e-16 ***
## ring_cat8+:shell  -3.9501     1.7456  -2.263   0.0236 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5243.4  on 4174  degrees of freedom
```

```
## Residual deviance: 3534.0  on 4171  degrees of freedom
## AIC: 3542
##
## Number of Fisher Scoring iterations: 5
```

Question 3

You're planning an experiment and need to estimate how much data you need (n) to take to achieve 80% power. The model is:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

After performing a quantitative literature review you have a high degree of certainty that

$$\epsilon \sim (0, 10^2)$$

and

$$\beta_0 = 3$$

What you're interested in is β_1 (we'll assume that σ and β_0 are exact as given). The hypotheses are:

$$H_0 : \beta_1 \leq 0$$
$$H_A : \beta_1 > 0$$

The experiment needs to be powerful enough to reject H_0 80% of the time under the assumption that β_1 is at least equal to 4.

The predictor variable x can take on values between 0 and 100.

Pseudocode for this simulation is acceptable if it's complete and in the right order. Full code and graph would be optimal. Make sure that you explain what you're doing at each step and why.