STAT 505: Final Exam
Name:

Please turn in the exam to GitHub and include the R Markdown code and a PDF or Word 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.

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.

Short Answer Questions (16 points)

For questions in this section, keep your answers concise. You are welcome to use a combination of prose, math, and pseudocode, but your responses should be well thought out and defended.

1. (4 points)

Detail the process for conducting a posterior predictive check and then describe how they can be used for assessing model fit.

2. (4 points)

Make an argument (to a collaborator) for centering and/or standardizing continuous predictors.

3. (4 points)

Why should inferences about sampling units be characterized as differences in predictors *between* units rather than differences in predictors *within* a sampling unit?

4. (4 points)

Consider the distribution of an expected outcome given a set of predictors and the distribution of a new observation given a set of predictors. Describe how the point estimate and uncertainty would differ (or not) for the two situations.

Code Interpretation (16 points)

For this question, we will use a subset of a dataset that contains Indian recipes.

```
indian_food <- read_csv('https://raw.githubusercontent.com/stat408/final_exam/master/indian.csv') %>%
  filter(course != 'starter') %>%
  select(-ingredients, -diet, -region) %>%
  mutate(flavor_profile = factor(flavor_profile), course = factor(course))
```

```
## Parsed with column specification:
## cols(
##   name = col_character(),
##   ingredients = col_character(),
##   diet = col_character(),
##   prep_time = col_double(),
##   cook_time = col_double(),
##   flavor_profile = col_character(),
##   course = col_character(),
##   region = col_character()
## )
```

```
summary(indian_food)
```

```
##      name              prep_time        cook_time      flavor_profile
##  Length:179         Min.   :  5.00   Min.   :  5.00   spicy:106
##  Class :character   1st Qu.: 10.00   1st Qu.: 25.00   sweet: 73
##  Mode  :character   Median : 10.00   Median : 30.00
##                     Mean   : 34.76   Mean   : 41.32
##                     3rd Qu.: 20.00   3rd Qu.: 45.00
##                     Max.   :500.00   Max.   :720.00
##          course
##  dessert    :70
##  main course:81
##  snack      :28
##
##
##
```

1. (4 points)

Using the following model specification right out the complete linear model and define all of the coefficients in the model.

```
model_specification <- formula(cook_time ~ prep_time + flavor_profile + course)
model_specification
```

```
## cook_time ~ prep_time + flavor_profile + course
```

2. (4 points)

Interpret the results.

```
lm(model_specification, data = indian_food) %>% summary()
```

```
##
```

```
## Call:
## lm(formula = model_specification, data = indian_food)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -60.43 -16.52  -6.99   3.76 673.48
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          37.12660   32.91436   1.128   0.2609
## prep_time             0.09342    0.05583   1.673   0.0961 .
## flavor_profilesweet   8.46149   32.33476   0.262   0.7939
## coursemain course    -1.81583   32.37908  -0.056   0.9553
## coursesnack         -10.76795   34.77966  -0.310   0.7572
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.81 on 174 degrees of freedom
## Multiple R-squared:  0.02858,    Adjusted R-squared:  0.006252
## F-statistic:  1.28 on 4 and 174 DF,  p-value: 0.2797
```

3. (4 points)

Interpret the results.

```
stan_glm(model_specification, data = indian_food, refresh = 0)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      cook_time ~ prep_time + flavor_profile + course
##  observations: 179
##  predictors:   5
## ------
##                      Median MAD_SD
## (Intercept)          36.2   34.1
## prep_time             0.1    0.1
## flavor_profilesweet  9.2   33.3
## coursemain course   -0.7   33.5
## coursesnack         -9.6   35.6
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 55.0    3.0
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

4. (4 points)

What is the problem with trying to fit this model?

```
lm(cook_time ~ prep_time + flavor_profile + course + flavor_profile:course,
   data = indian_food) %>% summary()
```

```
##
## Call:
## lm(formula = cook_time ~ prep_time + flavor_profile + course +
##     flavor_profile:course, data = indian_food)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -60.43 -16.52  -6.99   3.76 673.48
##
## Coefficients: (2 not defined because of singularities)
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           37.12660   32.91436   1.128   0.2609
## prep_time                              0.09342    0.05583   1.673   0.0961 .
## flavor_profilesweet                    8.46149   32.33476   0.262   0.7939
## coursemain course                     -1.81583   32.37908  -0.056   0.9553
## coursesnack                          -10.76795   34.77966  -0.310   0.7572
## flavor_profilesweet:coursemain course        NA         NA      NA       NA
## flavor_profilesweet:coursesnack              NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.81 on 174 degrees of freedom
## Multiple R-squared:  0.02858,    Adjusted R-squared:  0.006252
## F-statistic:  1.28 on 4 and 174 DF,  p-value: 0.2797
```

Simulation Question (12 points)

1. (4 points)

Consider the code below. Create a figure of x and y. What linear regression assumption does this data violate?

```
set.seed(11112020)
n <- 500
x <- seq(1,20, length.out = n)
beta <- c(1, .1)
sigma <- sqrt(x)
y <- rnorm(n, beta[1] + beta[2] * x, sd = sigma)
```

2. (4 points)

How well does a linear regression model recover the point estimates of beta? Justify your answer (simulation may be useful).

```
lm(y ~x) %>% summary()
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7022  -1.9748  -0.1912   2.1047  13.6252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.17914    0.30742   3.836 0.000141 ***
## x            0.07876    0.02594   3.036 0.002519 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.188 on 498 degrees of freedom
## Multiple R-squared:  0.01818,    Adjusted R-squared:  0.01621
## F-statistic:  9.22 on 1 and 498 DF,  p-value: 0.002519
```

2. (4 points)

How well does a linear regression model capture the uncertainty in a predictions for y conditional on

- x = 1
- x = 10
- x = 20

Data Analysis (Scaled to be worth 26 points)

Using the Indian recipe dataset fit a logistic regression model to model the probability of a dish being classified as a main course. Write your results in a short report (shorter than the projects). Turn this document in separately. Including figures and tables, I am setting a four page maximum using standard PDF output settings in RMD. This will require careful selection and sizing of tables and figures. The page limit does not apply to references or code in the appendix.

Report generalities	**Points**
Spelling, grammar, writing clarity, paragraphs, section labels	/8
Citations/Acknowledgments for papers and packages used	/4
Code in appendix	/4

Introduction + Data Overview	**Points**
Research question	/4
Variables with units and descriptive statistics	/4
Data Viz: Figure Clarity (Titles, Labels, and Captions)	/4

Statistical Procedures	**Points**
Define model to fit with complete notation (including priors)	/8
Defense of model choice	/4

Results + Discussion	**Points**
Discuss Results in the context of the research question	/4
Summarize estimates from final model including uncertainty	/8