STA303 Summer 2019 Test 2

First Name:_____

Last Name:_____

Student Number:_____

U of T Email:_____

READ THE FOLLOWING INSTRUCTIONS CAREFULLY

- This exam booklet contains 12 single-sided pages.
- This exam booklet contains **2** short-answer questions.
- This exam is being marked using crowdmark. We will scan the fronts of the pages only. Nothing written on the backs of pages will be read or marked.
- Write all final answers in the exam booklet, on the front of the page where the question appears.
- **Use the backs of the pages and the pages at the end for rough work. Nothing written on these pages will be scanned or marked.**
- Aids permitted: non-programmable calculator.
- If you don't know an answer, explain clearly anything that you do know. Do not write nonsense; this will not achieve part marks. The TAs may mark questions harder if there is lots of unrelated information present.
- Good luck!

1. **Penalized Regression: theory**. Consider the linear regression scenario: we observe independent pairs (y_i, x_i), $i = 1, \ldots, n$, $y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$. We postulate a linear model,

$$\begin{aligned} y_i &= x_i^T \beta + \epsilon_i \\ \epsilon_i &\overset{iid}{\sim} \text{Normal}(0, \sigma^2) \end{aligned} \tag{0.1}$$

The goal of inference is to produce point and interval estimates for $\beta \in \mathbb{R}^p$.

(a) (2 marks): write the model in vector form. Define all quantities needed to do this, and make sure to state the full joint distribution of the vector $\boldsymbol{\epsilon}$.

(b) (4 marks) The log-likelihood for β is

$$\ell(\beta) = c - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) \tag{0.2}$$

Find the maximum likelihood estimator for β.

(c) (2 marks) The MLE won't exist if $p > n$. Say in one sentence why this is the case, referencing the formula you derived in (b).

(d) (6 marks) Consider the penalized likelihood optimization problem,

$$\widehat{\beta} = \text{argmax}_{\beta} \left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) - \frac{\lambda}{2}\|\beta\|_2^2 \right)$$

where $\|\beta\|_2^2 = \beta^T\beta = \sum_{j=1}^{p} \beta_j^2$ and $\lambda \geq 0$. Find $\widehat{\beta}$ and explain why it always exists, even if $p > n$.

2. **Linear Mixed Models: application**. Consider the following fictional dataset of a bank's credit card portfolio, containing monthly spends of all their customers. The variables are

- `id`: customer id, uniquely identifies each customer in the dataset.
- `month`: Month of the year.
- `limit`: customer's credit limit, total amount of funds available. The same across months for each customer.
- `spend`: dollar amount that each customer spent in each month.
- `log spend`: log(spend).

You work for the bank, in your first job out of undergrad as an analyst. Your job is to present to the suit-people a summary of what factors are associated with increases in customers' credit card spending. Your manager fit a linear mixed model, and asked you to write it up. Review the below analysis and then answer the stated questions. Be brief in your answers; the hard part of this question is reading the analysis and figuring out what's going on.

```
glimpse(credit)
```

```
## Observations: 600
## Variables: 5
## $ id        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2...
## $ month     <chr> "Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Au...
## $ limit     <dbl> 10000, 10000, 10000, 10000, 10000, 10000, 10000, 100...
## $ log_spend <dbl> 4.551041, 5.730030, 5.497011, 4.762456, 6.752824, 5....
## $ spend     <dbl> 94.731009, 307.978435, 243.961738, 117.033027, 856.4...
```

```
# How many subjects?
credit %>% pull(id) %>% unique() %>% length()
```

```
## [1] 50
```

```
# How many months (lol)?
credit %>% pull(month) %>% unique() %>% length()
```

```
## [1] 12
```

```
# Average spend per month:
credit %>% group_by(month) %>% summarize(avgspend = mean(spend),sdspend = sd(spend))
```

```
## # A tibble: 12 x 3
##    month avgspend sdspend
##    <chr>    <dbl>   <dbl>
##  1 Apr       789.   1412.
##  2 Aug      3398.  12387.
##  3 Dec      3630.  11451.
##  4 Feb      2753.   6640.
##  5 Jan       688.   1211.
##  6 Jul      1459.   3264.
##  7 Jun      1413.   6400.
##  8 Mar      1421.   4085.
##  9 May      1262.   3175.
```

```
## 10 Nov       626.   1260.
## 11 Oct      1232.   3568.
## 12 Sep       879.   1891.
```

```
# Average spend for a few selected customers:
credit %>%
  group_by(id) %>%
  summarize(avgspend = mean(spend),sdspend = sd(spend)) %>%
  inner_join(tibble(id = sample(1:50,size = 10,replace = FALSE)),by = "id")
```

```
## # A tibble: 10 x 3
##       id avgspend sdspend
##    <int>    <dbl>   <dbl>
##  1     4     475.   1331.
##  2    12     646.    634.
##  3    23     218.    186.
##  4    36     460.    525.
##  5    37    7931.  25075.
##  6    38    2638.   7893.
##  7    40     902.   1349.
##  8    41     335.    406.
##  9    46     437.   1000.
## 10    48    1090.   1996.
```

```
# Fit the model. Rescale credit limit so it's on the same scale as the others
spendmodel <- lme4::lmer(
  log_spend ~ I(limit/5000) + month + (1|id),
  data = credit,
  REML = TRUE
)
summary(spendmodel)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log_spend ~ I(limit/5000) + month + (1 | id)
##    Data: credit
##
## REML criterion at convergence: 2378.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.82264 -0.64711 -0.02505  0.64104  3.02619
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  id       (Intercept) 0.9597   0.9796
##  Residual             2.6853   1.6387
## Number of obs: 600, groups:  id, 50
##
## Fixed effects:
```

```
##                Estimate Std. Error t value
## (Intercept)     4.77269    0.49398   9.662
## I(limit/5000)   0.23294    0.19698   1.183
## monthAug        1.28291    0.32774   3.914
## monthDec        1.43526    0.32774   4.379
## monthFeb        0.71186    0.32774   2.172
## monthJan       -0.29753    0.32774  -0.908
## monthJul        0.67186    0.32774   2.050
## monthJun       -0.21299    0.32774  -0.650
## monthMar        0.01190    0.32774   0.036
## monthMay        0.07585    0.32774   0.231
## monthNov       -0.15053    0.32774  -0.459
## monthOct       -0.04606    0.32774  -0.141
## monthSep       -0.23552    0.32774  -0.719
##
## Correlation matrix not shown by default, as p = 13 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

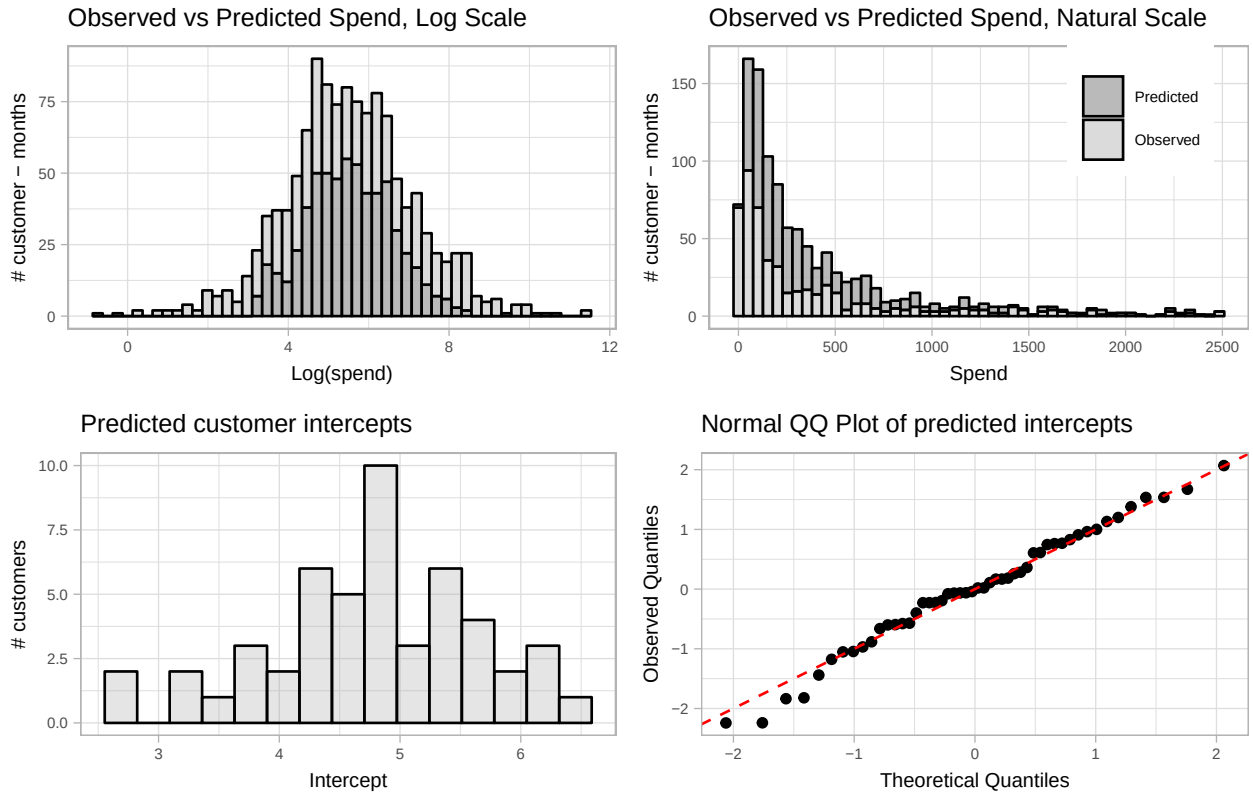
```
# Plot posterior expected spend
# coef(spendmodel) gives each subject's predicted intercept and the regression
# coefficients:
head(coef(spendmodel)$id)
```

```
##   (Intercept) I(limit/5000) monthAug monthDec  monthFeb   monthJan
## 1    4.919360     0.2329424 1.282909 1.435255 0.7118552 -0.2975269
## 2    5.438042     0.2329424 1.282909 1.435255 0.7118552 -0.2975269
## 3    4.274551     0.2329424 1.282909 1.435255 0.7118552 -0.2975269
## 4    3.515786     0.2329424 1.282909 1.435255 0.7118552 -0.2975269
## 5    4.736337     0.2329424 1.282909 1.435255 0.7118552 -0.2975269
## 6    4.575990     0.2329424 1.282909 1.435255 0.7118552 -0.2975269
##    monthJul   monthJun   monthMar   monthMay   monthNov    monthOct
## 1 0.6718601 -0.2129914 0.01189649 0.07585315 -0.1505346 -0.04605987
## 2 0.6718601 -0.2129914 0.01189649 0.07585315 -0.1505346 -0.04605987
## 3 0.6718601 -0.2129914 0.01189649 0.07585315 -0.1505346 -0.04605987
## 4 0.6718601 -0.2129914 0.01189649 0.07585315 -0.1505346 -0.04605987
## 5 0.6718601 -0.2129914 0.01189649 0.07585315 -0.1505346 -0.04605987
## 6 0.6718601 -0.2129914 0.01189649 0.07585315 -0.1505346 -0.04605987
##     monthSep
## 1 -0.2355249
## 2 -0.2355249
## 3 -0.2355249
## 4 -0.2355249
## 5 -0.2355249
## 6 -0.2355249
```

```
# ...but I'm smart and will use the predict() method.
predvals <- credit %>%
```

```
  mutate(pred_log_spend = predict(spendmodel),
         pred_spend = exp(pred_log_spend))

# Plot predicted and actual, on log and natural scale
predplot_log <- predvals %>%
  tidyr::gather(type,val,log_spend,pred_log_spend) %>%
  ggplot(aes(x = val,fill = type)) +
  theme_light() +
  geom_histogram(colour = "black",alpha = .8,bins = 50) +
  labs(title = "Observed vs Predicted Spend, Log Scale",
       x = "Log(spend)",
       y = "# customer - months",
       fill = "") +
  scale_fill_manual(labels = c("log_spend" = "Observed",
                               "pred_log_spend" = "Predicted"),
                    values = c("log_spend" = "lightgrey","pred_log_spend" = "darkgrey"))

predplot_natural <- predvals %>%
  tidyr::gather(type,val,spend,pred_spend) %>%
  filter(val < 2500) %>%
  ggplot(aes(x = val,fill = type)) +
  theme_light() +
  geom_histogram(colour = "black",alpha = .8,bins = 50) +
  labs(title = "Observed vs Predicted Spend, Natural Scale",
       x = "Spend",
       y = "# customer - months",
       fill = "") +
  scale_fill_manual(labels = c("spend" = "Observed",
                               "pred_spend" = "Predicted"),
                    values = c("spend" = "lightgrey","pred_spend" = "darkgrey"))

# Plot the predicted intercepts only
intercept_plot <- coef(spendmodel)$id %>%
  as_tibble() %>%
  dplyr::select(intercept = `(Intercept)`) %>%
  ggplot(aes(x = intercept)) +
  theme_light() +
  geom_histogram(bins = 15,colour = "black",fill = "darkgrey",alpha = .3) +
  labs(title = "Predicted customer intercepts",
       x = "Intercept",
       y = "# customers")

# Normal QQ plot of predicted intercepts
qqplot_intercepts <- coef(spendmodel)$id %>%
  as_tibble() %>%
  dplyr::select(intercept = `(Intercept)`) %>%
```

```
  arrange(intercept) %>%
  mutate_at("intercept",funs( (. - mean(.)) / sd(.))) %>%
  mutate(q = qnorm(seq(1:50) / (1 + 50))) %>%
  ggplot(aes(x = q,y = intercept)) +
  theme_light() +
  geom_point() +
  geom_abline(slope = 1,intercept = 0,linetype = "dashed",colour = "red") +
  labs(title = "Normal QQ Plot of predicted intercepts",
       x = "Theoretical Quantiles",
       y = "Observed Quantiles")

cowplot::plot_grid(
  predplot_log + guides(fill = FALSE) + theme(text = element_text(size = 8)),
  predplot_natural + theme(legend.position = c(.8,.8),text = element_text(size = 8)),
  intercept_plot + theme(text = element_text(size = 8)),
  qqplot_intercepts + theme(text = element_text(size = 8)),
  nrow = 2
)
```



(a) (6 marks) Write down the full statistical model that I fit, clearly defining all terms including all parameters and distributions.

(b) (4 marks) What is the proportion of variance explained by `id`? State this number and interpret it in the context of the problem — does spending vary more within, or between customers?

(c) (5 marks) State all model assumptions (even if you're repeating part of your answer from (a)). Comment on whether they are satisfied, or if I haven't given you enough information to tell.

(d) (6 marks) Write 3 - 5 sentences describing the spending behaviour of this bank's customers to the business folks. Reference specific numbers from the output. For full marks, also briefly comment on any limitations of this analysis. Anything more than 5 sentences is definitely too much here.

THIS PAGE IS FOR ROUGH WORK. NOTHING ON THIS PAGE WILL BE MARKED.

THIS PAGE IS FOR ROUGH WORK. NOTHING ON THIS PAGE WILL BE MARKED.