UNIVERSITY OF TORONTO
Faculty of Arts and Science
August 2019 Examinations
STA303H1S
Methods of Data Analysis II
Duration: 3 hours
Aids Allowed: Non-programmable calculator

First Name:____

Last Name:____

Student Number:____

U of T Email:____

- Fill out your name and student number on the top of this page.
- Do not begin writing the actual exam until the announcements have ended and the Exam Facilitator has started the exam.
- As a student, you help create a fair and inclusive writing environment. If you possess an unauthorized aid during an exam, you may be charged with an academic offence.
- Turn off and place all cell phones, smart watches, electronic devices, and unauthorized study materials in your bag under your desk. If it is left in your pocket, it may be an academic offence.
- When you are done your exam, raise your hand for someone to come and collect your exam. Do not collect your bag and jacket before your exam is handed in.
- If you are feeling ill and unable to finish your exam, please bring it to the attention of an Exam Facilitator so it can be recorded before leaving the exam hall.
- In the event of a fire alarm, do not check your cell phone when escorted outside.

Question	Out of	Question	Out of
1	20	4	20
2	20	5	20
3	20		
Total	100		

1. **Exponential Families (20 marks)**. A random variable Y follows an $\mathrm{IG}(\mu, 1)$ distribution with $\mu > 0$ if the density of Y is given by

$$f_Y(x) = \sqrt{\frac{1}{2\pi x^3}} \exp\left(-\frac{(x-\mu)^2}{2\mu^2 x}\right), \quad x > 0.$$

Let (Y_i, X_i) be independent response/covariate measurements, with $X_i \in \mathbb{R}^p$ and $Y_i \sim \mathrm{IG}(\mu, 1)$. We propose a generalized linear model for Y_i,

$$\begin{aligned} Y_i &\sim \mathrm{IG}(\mu_i, 1) \\ \theta_i &= \frac{1}{\mu_i^2} = x_i^T \beta \end{aligned}$$

where θ_i is the canonical parameter and $\beta \in \mathbb{R}^p$ the parameter to be estimated.

(a) (5 marks): Write down what it means for a distribution to be in the exponential family, and show that $\mathrm{IG}(\mu, 1)$ is such a distribution by writing down the likelihood for this model in exponential family form. Explicitly specify the dispersion parameter and whichever functions ($b(\cdot)$, $c(\cdot,\cdot)$, etc.) you're using.

(b) (10 marks) Write down the link function, linearized response, and variance function for this model.

(c) (5 marks) Derive the score equations to which the maximum likelihood estimator $\widehat{\beta}$ is the solution.

2. **Bayesian Theory (20 marks)**. In this question, we assume that our data follows a standard regression model. That is,

$$y = X\beta + \epsilon,$$

where X is a known $n \times p$ matrix of full rank, $y \in \mathbb{R}^n$ is a vector of observations, $\beta \in \mathbb{R}^p$ is a vector of parameters to be estimated, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$.

(a) (5 marks): As a statistical analyst working at a bank, you've decided to build a Bayesian regression model to predict future losses that a portfolio will incur as a function of some covariates. However, your business executives are skeptical; they know little beyond the basics of frequentist statistics and linear/logistic regression, and they get anxious when they hear advanced words like "Bayesian" which they don't understand. Write a short paragraph explaining how Bayesian statistics differs from frequentist statistics (2-3 sentences), and how this difference affects the intepretation of regression models (2-3 sentences). Your paragraph should be tailored to your executives (i.e., write in plain language; don't use any terminology or notation that they wouldn't understand).

(b) (5 marks): Recall that in the penalized regression version of ordinary least squares, called ridge regression, the estimated coefficients minimize the penalized residual sum of squares

$$\widehat{\beta} = \operatorname{argmin}_\beta \left((y - X\beta)^T (y - X\beta) + \lambda ||\beta||_2^2 \right).$$

Derive an explicit form for $\widehat{\beta}$. (Hint: this was done in lecture and on Test 2.)

(c) (2 marks): Circle **true** or **false**: a larger choice of λ will usually lead to smaller coefficient estimates $\widehat{\beta}$.

(d) (8 marks): Assume we put a $\mathcal{N}(0, \sigma^2/\lambda)$ prior distribution independently on each β_j. Use Bayes' theorem to prove that the mean of the posterior distribution of β given the data (X, y) coincides with the estimator you derived in part (b). (Hint: by completing the square in β, show that the posterior density is proportional to $\exp(\frac{1}{2}(\beta - Za)^T Z^{-1}(\beta - Za))$, where $Z = (X^T X + \lambda I)^{-1}$ and $a = X^T y$.)

3. **Linear Mixed Effects (20 marks)**. Recall from lecture the Rat Growth data: data on the weights of each of 30 rats was measured for 5 consecutive weeks. We are interested in modelling growth as a function of week, accounting for the fact that weight measurements on the same rat will be correlated. We model these data using a linear mixed effects model. Here is the data: `y` is the response (weight, units not specified), `rat` is an ID for which rat it is, and `week` is an ID for which week it is.

```
## Observations: 150
## Variables: 3
## $ rat  <fct> 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, ...
## $ week <int> 0, 1, 2, 3, 4, 0, 1, 2, 3, 4, 0, 1, 2, 3, 4, 0, 1, 2, 3, ...
## $ y    <dbl> 151, 199, 246, 283, 320, 145, 199, 249, 293, 354, 147, 21...
```

```
## ===============RAT MODEL 1===============
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ 1 + (1 | rat) + week
##    Data: rat
##
## REML criterion at convergence: 1127.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.7919 -0.4897  0.1287  0.5794  2.4702
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  rat      (Intercept) 191.86   13.851
##  Residual              64.29    8.018
## Number of obs: 150, groups:  rat, 30
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 156.0533     2.7715   56.31
## week         43.2667     0.4629   93.46
##
## Correlation of Fixed Effects:
##      (Intr)
## week -0.334
```

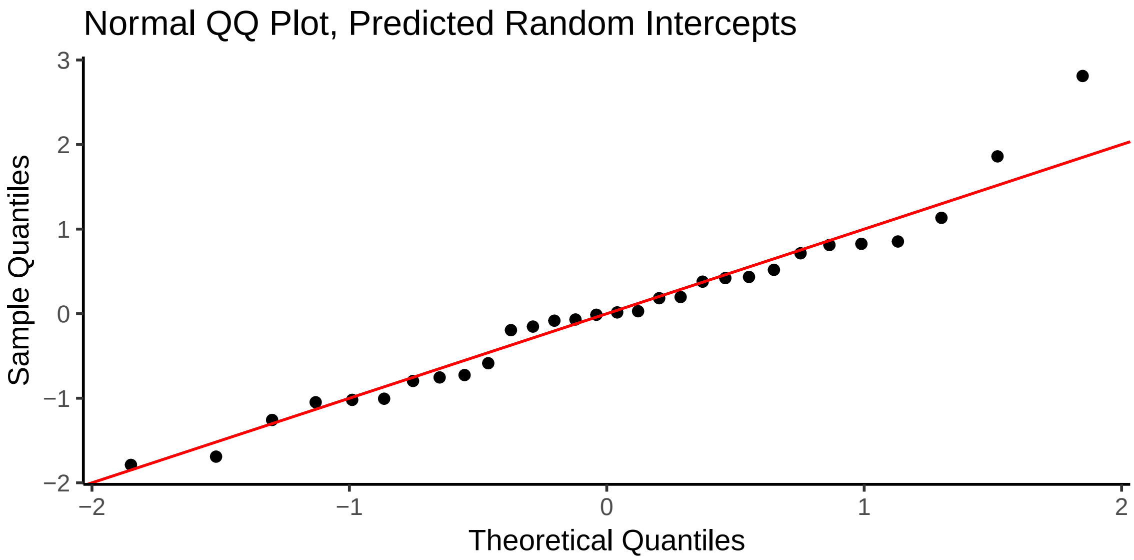
```
## ===============RAT MODEL 2===============
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ 1 + (week | rat) + week
##    Data: rat
##
## REML criterion at convergence: 1084.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.7275 -0.5670  0.1213  0.5562  2.3688
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  rat      (Intercept) 119.53   10.933
##           week         12.49    3.535   0.18
##  Residual              33.84    5.817
## Number of obs: 150, groups:  rat, 30
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 156.0533     2.1590   72.28
## week         43.2667     0.7275   59.47
##
## Correlation of Fixed Effects:
##      (Intr)
## week 0.007
```

```
## ===============RAT MODEL 3===============
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: y ~ 1 + (1 | rat) + week
##    Data: rat
##
##      AIC      BIC   logLik deviance df.resid
##   1139.2   1151.2   -565.6   1131.2      146
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8057 -0.4932  0.1276  0.5779  2.4817
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  rat      (Intercept) 185.14   13.607
##  Residual              63.76    7.985
## Number of obs: 150, groups:  rat, 30
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  156.053      2.729   57.19
## week          43.267      0.461   93.85
##
## Correlation of Fixed Effects:
##      (Intr)
## week -0.338
```



(a) (4 marks): Write down the full statistical model for each of the three models, clearly defining all terms including all parameters and distributions (if you use the same term(s) in multiple models, you only need define them once).

(b) (2 marks): Is it better to include `rat` as a random effect or a fixed effect (with 30 levels)? Answer "random effect" or "fixed effect" and provide one good reason for your answer.

(c) (2 marks): Is it better to include `week` as a random effect or a fixed effect? Answer "random effect" or "fixed effect" and provide one good reason for your answer.

(d) (1 mark): For the model "RAT MODEL 1", what is the proportion of total variance explained by `rat`?

(e) (3 marks): Consider the models "RAT MODEL 1" and "RAT MODEL 2". The latter contains one more term than the former; hence we wish to compare them using a procedure for nested model comparison (i.e. a likelihood ratio test). Can we do this? Answer "Yes" or "No" and briefly explain why or why not.

(f) (2 marks): Consider the plot titled "Normal QQ Plot, Predicted Random Intercepts", which is a normal QQ-plot of the predicted random intercepts from "RAT MODEL 1". What model assumption is being tested here?

(g) (2 marks): Consider the models "RAT MODEL 1" and "RAT MODEL 3". Why are the estimated variances different?

(h) (2 marks): If we wanted to predict the weight in a given week after birth for a given `rat` from the 30 included in the dataset using "RAT MODEL 1", what would be the prediction equation? You may denote the predicted random effects for the i^{th} rat as $\widetilde{b}_{i1}, \widetilde{b}_{i2}$ and use these in your answer without stating their formula.

(i) (2 marks): If we wanted to predict the weight in a given week after birth for a new `rat` that was not one of the 30 included in the dataset using "RAT MODEL 1", what would be the prediction equation?

4. **GLMMs and INLA (20 marks)**. As part of your new job as an agriculturist, you have been provided the `nitrofen` data, which consists of counts of living offspring of zooplankton exposed to various concentrations of nitrofen, a herbicide. Specifically, 50 zooplankton were split into 10 groups of 5 each, and exposed to different concentrations of nitrofen. Each animal then gave birth to three broods, and the number of live offspring in each brood was recorded. The data is as follows:

```
## Observations: 150
## Variables: 4
## $ conc  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ id    <int> 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 7,...
## $ brood <chr> "1", "2", "3", "1", "2", "3", "1", "2", "3", "1", "2", "...
## $ live  <dbl> 3, 14, 10, 5, 12, 15, 6, 11, 17, 6, 12, 15, 6, 15, 15, 5...
```

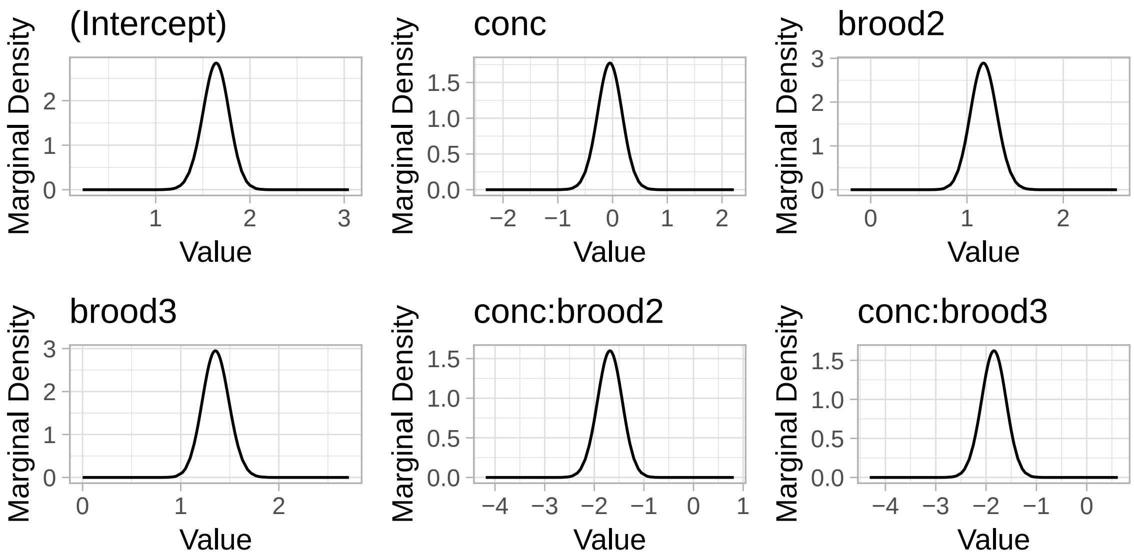
The counts are naturally grouped, with counts from the same zooplankton expected to be more similar than counts from different zooplankton. We wish to model the linked mean with one random-effect for each zooplankton, and linear terms corresponding to the available covariates and their interaction. After scaling the concentration by dividing by 300 (the maximum value is 310), you fit a model using INLA:

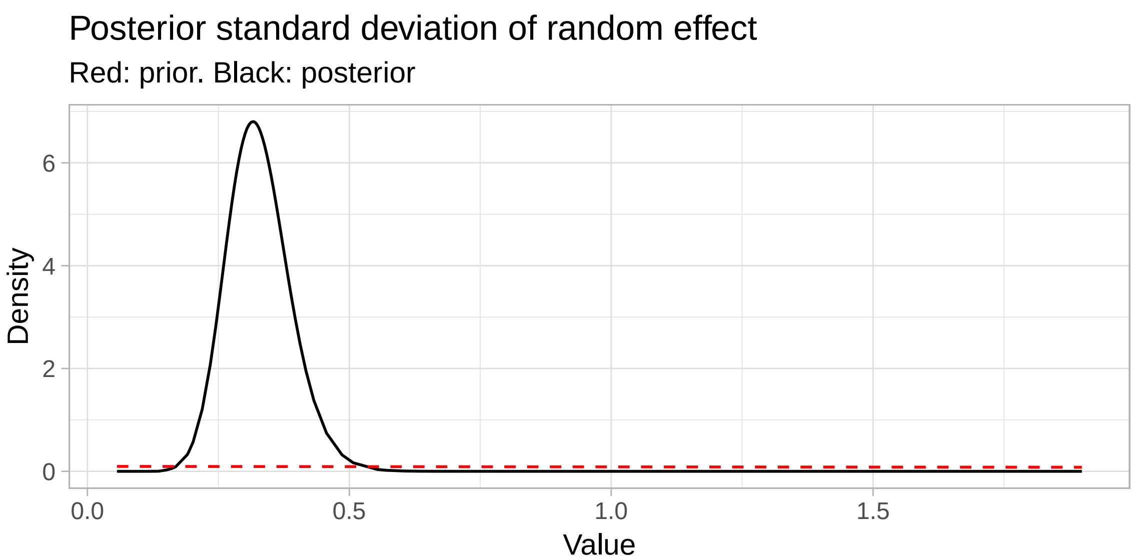
```
## Model formula: live ~ conc * brood +
## f(id,model = 'iid',
## prior = 'pc.prec',param = c(3,.75))
```

```
##
## Call:
## c("inla(formula = live ~ conc * brood + f(id, model = \"iid\", prior = \"pc.prec\", ",  "    param =
##
## Time used:
##  Pre-processing    Running inla Post-processing           Total
##          1.7597          0.2016          0.0937          2.0550
##
## Fixed effects:
##                 mean     sd 0.025quant 0.5quant 0.975quant    mode kld
## (Intercept)  1.6376 0.1414     1.3554   1.6391     1.9117  1.6420   0
## conc        -0.0509 0.2268    -0.5001  -0.0499     0.3919 -0.0479   0
## brood2       1.1747 0.1382     0.9076   1.1733     1.4496  1.1706   0
## brood3       1.3576 0.1355     1.0961   1.3561     1.6278  1.3531   0
## conc:brood2 -1.6885 0.2498    -2.1832  -1.6870    -1.2023 -1.6841   0
## conc:brood3 -1.8479 0.2461    -2.3357  -1.8464    -1.3692 -1.8433   0
##
## Random effects:
## Name   Model
##  id   IID model
##
## Model hyperparameters:
##                    mean    sd 0.025quant 0.5quant 0.975quant  mode
## Precision for id 10.39 4.181      4.798    9.569      20.81 8.271
##
## Expected number of effective parameters(std dev): 37.77(3.744)
## Number of equivalent replicates : 3.972
```

```
##
## Marginal log-Likelihood:  -431.88
```

```
##                 mean         sd    q0.025      q0.5    q0.975      mode
## SD for id 0.3269879 0.06006813 0.2194122 0.3231583 0.4557495 0.3163052
```





The full heirarchical model takes the form

$$
\begin{aligned}
Y_{ij} &\sim \text{Poisson}(\lambda_{ij}), i = 1, \ldots, 50; j = 1, \ldots, 3 \\
\log \lambda_{ij} = \eta_{ij} &= \beta_0 + \beta_1 x_i + \beta_2 b_j + \beta_3 x_i \times b_j + u_i \\
u_i &\sim \text{Normal}(0, \sigma_u^2)
\end{aligned} \tag{0.1}
$$

where

1. $i = 1, \ldots, 50$ represents the i^{th} zooplankton,
2. $j = 1, \ldots, 3$ represents the j^{th} brood for each zooplankton,

3. Y_{ij} is the count of live offspring in the j^{th} brood of the i^{th} zooplankton,
4. x_i is the concentration of nitrofen to which the i^{th} zooplankton was exposed,
5. b_j is an indicator variable representing a fixed intercept for the j^{th} brood,
6. u_i is a random effect designed to capture correlation in offspring counts across broods from the same zooplankton.

(a) (3 marks): Unfortunately, the business executives from Problem 2a have also switched industries from banking to agriculture, and they're clueless about mixed effects models. Explain the interpretation of the sole hyperparameter in the model to them in one sentence.

(b) (3 marks): You used a PC prior for the random effect standard deviation. Based on the observed standard deviation of flea-means (a very rough proxy for the scale of parameter σ_u), you decided that there's about a 75% chance that $\sigma_u > 3$. Write down the corresponding parameters for `pc.prec`.

(c) (8 marks): Based on the model, what happens to the average numbers of offspring for each brood at zero concentration of the herbicide? How about when concentration increases to 1? Provide point estimates with 95% confidence intervals on the natural scale.

(d) (4 marks): INLA reports summary statistics of the posterior of the log precision, which is $-2\log(\sigma_u)$, rather than that of σ_u (which is what we prefer). Provide a 95% confidence interval for σ_u.

(e) (2 marks): Does the individual variation in `id` outweigh the brood and concentration effects? Answer "Yes" or "No" and justify based on the INLA output.

5. **Logistic Regression Mystery (20 marks)**. (Note: this question is challenging. Manage your time wisely.) The `mystery` data contains six observations as follows:

```
## Observations: 6
## Variables: 2
## $ x <dbl> -4.757, -3.941, -4.413, 4.453, 3.349, 4.899
## $ y <dbl> 1, 1, 1, 0, 0, 0
```

Since `y` is binary and `x` is continuous, you try to fit a logistic regression model to predict `y` using R. To your surprise, R gives you a warning (which has been suppressed) and outputs this model summary:

```
##
## Call:
## glm(formula = y ~ x, family = "binomial", data = mystery)
##
## Deviance Residuals:
##          1           2           3           4           5           6
##  9.207e-07   1.240e-05   2.756e-06  -2.110e-08  -1.315e-05  -2.110e-08
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.828  46931.107       0        1
## x              -6.373  12805.672       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8.3178e+00  on 5  degrees of freedom
## Residual deviance: 3.3510e-10  on 4  degrees of freedom
## AIC: 4
##
## Number of Fisher Scoring iterations: 23
```

Look at those coefficient estimates! Look at those standard errors! Clearly something strange has happened.

(a) (5 marks): Coefficients and standard errors notwithstanding, does the model fit the data well? Answer "Yes" or "No" and justify based on the residual deviance.

(b) (5 marks): Recall that the log-likelihood for binary logistic regression takes the general form

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \cdot \log\left(\frac{1}{1+e^{-x_i^T\beta}}\right) - (1-y_i) \cdot \log\left(1 - \frac{1}{1+e^{-x_i^T\beta}}\right)\right].$$

Use this to explain mathematically why the coefficient estimates have such a large magnitude and why the estimation procedure failed. (Hint: start by splitting the above into two sums, one for the $y_i = 0$ data and one for the $y_i = 1$ data. What happens when you try to maximize $\ell(\beta)$?)

(c) (5 marks): Recall that R computes approximate standard errors for the coefficient estimates using

$$\widehat{\mathrm{Var}}(\widehat{\beta}) = (X^T W X)^{-1}$$

, where the weight matrix W is diagonal with entries $W_{ii} = \widehat{p_i} \cdot (1 - \widehat{p_i})$. Use this to explain mathematically why the standard errors are so large.

(d) (5 marks): Use your knowledge of Bayesian statistics and/or penalized likelihood to propose an alternative regression approach that does *not* involve removing any variables or modifying the data, and show that your idea keeps the magnitude of the coefficient estimates from blowing up. (Hint: such an approach has already come up elsewhere on this exam. No need to derive an explicit expression for an estimator of β, but you do need to convince us that your approach mitigates the problem(s) identified in (b)).

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.

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.