UNIVERSITY OF TORONTO
Faculty of Arts and Science

AUGUST 2017 EXAMINATIONS

STA303H1S / STA1002HS

Duration - 3 hours

Examination Aids: Non-Programmable Calculator

Name:________________________ **Student #:**________________________

- **Show all your work.** Answers, correct or not, unsupported by calculations or explanation will not earn any marks.
- **Clearly state and define any variables. You must state and justify any distributions you use in your calculations.** These are part of the problem solving process and are worth marks.
- **Organize your work** in a reasonably neat and coherent way, in the space provided. Work scattered all over the page that cannot be understood will not earn full marks.
- You may use the backs of pages to do rough work, but **ONLY work written in the provided space will be graded.**
- This exam has a total of 21 pages including this cover page. Check to see if any pages are missing.
- This is a closed book test. You are allowed a **non-programmable** calculator on this test.
- Do not write in the table below.

Question	Grade	Out Of
1		17
2		15
3		18
4		20
5		20
6		10
Total		100

Question 1. Bacterial Canker (17 Marks)

You have been hired by a lumber farm to assess the effectiveness of a newly proposed treatment for bacterial cankers in trees. They have provided you with a dataset containing 24 observations of the following:

- `afflicted`: the number of afflicted trees on the plot of land
- `acreage`: the area of the plot of land in acres
- `treatment`: whether or no the plot of land received treatment

Each observation corresponds to a distinct plot of land. The total number of trees on each plot of land was too large for the company to bother counting. They tell you that it scales roughly with the area of the land.

The **main question of interest** is: *Does the new treatment have an effect on the rate of bacterial canker occurances?*

Your first instinct is to fit the following Poisson model:

```
q3_model1 <- glm(afflicted ~ factor(treatment)
                 + offset(log(acreage)), family = poisson())
summary(q3_model1)

##
## Call:
## glm(formula = afflicted ~ factor(treatment) + offset(log(acreage)),
##     family = poisson())
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -25.77  -20.27  -11.65   18.13   33.34
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         2.86098    0.01011  283.05   <2e-16 ***
## factor(treatment)1 -0.95346    0.02097  -45.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 13518  on 23  degrees of freedom
## Residual deviance: 11129  on 22  degrees of freedom
## AIC: 11307
##
## Number of Fisher Scoring iterations: 5
```

1.a) (3 Marks) Express the model (`q3_model1`) mathematically. Be sure to state the distibution of the response, the link function and the linear predictor. Define all parameters you use.

1.b) (3 Marks) Is it plausible that the data was generated by the model (`q3_model1`)? Inspection of which deviance(s) lead(s) you to this conclusion, and why?

1.c) (3 Marks) Based on your answer to the previous part, would it be conservative to use a p-value quoted from the summary output from (`q3_model1`) to form a conclusion regarding the main question of interest? If so, form such a conclusion at significance level 5%. If not, briefly explain why not.

Suppose you fit the following overdispersed Poisson model instead:

```
q3_model2 <- glm(afflicted ~ factor(treatment)
                 + offset(log(acreage)), family = quasipoisson())
summary(q3_model2)
```

```
##
## Call:
## glm(formula = afflicted ~ factor(treatment) + offset(log(acreage)),
##     family = quasipoisson())
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -25.77  -20.27  -11.65   18.13   33.34
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.8610     0.2315  12.357 2.26e-11 ***
## factor(treatment)1  -0.9535     0.4802  -1.985   0.0597 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 524.7228)
##
##     Null deviance: 13518  on 23  degrees of freedom
## Residual deviance: 11129  on 22  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

You also compute the following (possible) p-values

```
1-pchisq(13443-11053, df=23-22)

## [1] 0

1-pchisq((13443-11053)/518.056, df=23-22)

## [1] 0.03172306

1-pf(13443-11053, df1=23-22, df2=22)

## [1] 0

1-pf(13443-11053, df1=22, df2=23-22)

## [1] 0.01613526

1-pf((13443-11053)/518.056, df1=23-22, df2=22)

## [1] 0.04299446

1-pf((13443-11053)/518.056, df1=22, df2=23-22)

## [1] 0.353904
```

1.d) (3 Marks) What restrictive property of the poisson distribution is relaxed by the quasipoisson model?

1.e) (3 Marks) Quote a p-value directly from the summary output from (`q3_model2`) to form a conclusion regarding the main question of interest at significance level 5%.

1.f) (3 Marks) Among possible deviance-based tests applied to `q3_model2`, which is most appropriate for answering the question of interest: an F-test or a Chi-Squared Test? Circle the code/p-value on the previous page that is useful for this test. Form a conclusion regarding the main question of interest at significance level 5%.

Question 2. Diet of Cows (15 Marks)

Suppose you are given the following longitudinal model for the weights of cows, used to compare the effects of two competing diets:

$$\begin{aligned}
Y_{ijk} &= \beta_0 + \beta_X X_{ij}^{\texttt{Diet}} + \beta_T T_{ijk} + \beta_{XT} X_{ij}^{\texttt{Diet}} T_{ijk} + U_i + V_{ij} + \epsilon_{ijk} \\
U_i &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_U^2) \\
V_{ij} &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_V^2) \\
\epsilon_{ij} &\overset{\text{ind}}{\sim} \mathcal{N}_{m_{ij}}\left(0, A^{(ij)}\right) \\
A^{(ij)}_{k_1,k_2} &= \sigma^2 \exp\left(\frac{|T_{ijk_1} - T_{ijk_2}|}{\omega}\right)
\end{aligned}$$

Where

- Y_{ijk} is the kth weight measurement on cow ij from farm i
- $X_{ij}^{\texttt{Diet}}$ is a dummy variable encoding for the diet, 1 for the new diet and 0 for the old diet.
- T_{ijk} is the age of the cow (in months) at the time of observation ijk.
- U_i is a random effect for farm i
- V_{ij} is a random effect for cow ij on farm i
- ϵ_{ijk} is a temporally correlated residual.
- m_{ij} is the number of observations on cow ij from farm i.

For the new diet to be considered effective, it should have a positive effect that accumulates as the cow ages.

Questions follow on the next page.

2.a) (2 Marks) What are the parameters of this model?

2.b) (2 Marks) State a hypothesis test you would perform to assess whether the new diet is effective. Write the null hypothesis in terms of model parameters. Name a procedure you would use to perform this test.

2.c) (2 Marks) What is the marginal distribution of a single response observation, Y_{ijk}? (Don't forget to state the mean and variance)

2.d) (3 Marks) What is the joint distribution of a pair of response observations at different times on the same cow, (Y_{ijk_1}, Y_{ijk_2})? (Don't forget to state the mean and variance and covariance)

2.e) (2 Marks) Holding other parameters fixed, how does increasing ω affect the modelled covariance between observations on the same cow?

2.f) (2 Marks) Holding other parameters fixed, how does increasing ω affect the covariance between observations on different cows?

2.g) (2 Marks) State a hypothesis you would test to determine if the between cow-variation is greater than the between-farm variation. How could we peform this test, in theory?

Question 3. Lung Cancer (18 Marks)

A large health insurance provider has hired you to determine what patient and physician factors are important in determining whether a patient's lung cancer goes into remission after treatment, as part of a larger study of treatment outcomes and quality of life in patients with lung cancer. A variety of variables were collected on patients, who are nested within doctors, who are in turn nested within hospitals. There is also a doctor level variable, Experience, that we will use in our model. Each patient corresponds to a single observation.

The descriptions of a few of the variables on the dataset follow:

- remission: binary response variable, 0 for did not go into remission, 1 for went into remission
- IL6: blood concentration of the IL6 protein (Interleukin 6) (continuous)
- CRP: blood concentration of the CRP (C-reactive protein) (continuous)
- LengthofStay: the length of the patient's hospital stay, in days (integer)
- CancerStage: factor with 4 levels "I", "II", "III", "IV", higher is worse – cancer has progressed further.
- Experience: the amount of experience of the doctor in years (integer)
- DID: Doctor ID, unique identifier for the doctor.
- HID: Hospital ID, unique identifier for the hospital.

The following model was fit:

```
q5_model1 <- glmer(remission ~ IL6 + CRP + CancerStage
                    + LengthofStay + Experience
                    + (1 | DID) + (1 | HID),
                  data = hdp,
                  family = binomial(),
                  control = glmerControl(optimizer = "bobyqa"))
```

Questions follow on the next page.

3.a) (3 Marks) Express the model (`q5_model1`) mathematically. Be sure to state the distibution of the response, the link function, the linear predictor, and the distributions of any random effects. Define all parameters you use.

Question 3. Continued...

The next few questions refer to the summary output for model q5_model1.

```
summary(q5_model1)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## remission ~ IL6 + CRP + CancerStage + LengthofStay + Experience +
##     (1 | DID) + (1 | HID)
##    Data: hdp
## Control: glmerControl(optimizer = "bobyqa")
##
##      AIC      BIC   logLik deviance df.resid
##   7407.8   7478.3  -3693.9   7387.8     8515
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7440 -0.4426 -0.2024  0.3986  7.1354
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  DID    (Intercept) 3.6519   1.9110
##  HID    (Intercept) 0.2368   0.4866
## Number of obs: 8525, groups:  DID, 407; HID, 35
##
## Fixed effects:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.04132    0.51492  -3.964 7.36e-05 ***
## IL6            -0.05689    0.01124  -5.061 4.17e-07 ***
## CRP            -0.02142    0.00998  -2.146 0.031878 *
## CancerStageII  -0.41303    0.07385  -5.593 2.23e-08 ***
## CancerStageIII -1.00053    0.09595 -10.427  < 2e-16 ***
## CancerStageIV  -2.32816    0.15532 -14.990  < 2e-16 ***
## LengthofStay   -0.12092    0.03284  -3.682 0.000231 ***
## Experience      0.12035    0.02623   4.587 4.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) IL6    CRP    CncSII CnSIII CncSIV LngthS
## IL6         -0.084
## CRP         -0.088  0.002
## CancerStgII  0.014  0.006  0.005
## CancrStgIII  0.057  0.009  0.015  0.493
## CancerStgIV  0.065  0.031  0.014  0.333  0.318
## LengthofSty -0.300  0.012 -0.020 -0.264 -0.337 -0.289
## Experience  -0.902 -0.006 -0.002 -0.004 -0.009 -0.014 -0.010
```

3.b) (3 Marks) Which appears larger: the between hospital variation or the between doctor variation?

3.c) (3 Marks) Provide an estimate of the odds-ratio of remission between patients treated by a docter at the 97.5th percentile versus patients treated by a docter at the 2.5th percentile (percentiles of the doctor random effect, the Z-score needed is 1.96).

3.d) (3 Marks) Provide an estimate of the odds-ratio of remission between patients treated at hospitals at the 97.5th percentile versus patients treated at hospitals at the 2.5th percentile (percentiles of the hospital random effect, the Z-score needed is 1.96).

3.e) (3 Marks) Provide an estimate of the odds-ratio of remission between patients treated while in cancer stage I versus patients treated while in cancer stage II. Provide a 95% confidence interval as well (the Z-score needed is 1.96).

3.f) (3 Marks) Provide an estimate of the odds-ratio of remission between patients treated while in cancer stage II versus patients treated while in cancer stage III. Provide a 95% confidence interval as well (the Z-score needed is 1.96).

Question 4. Health Insurance (20 Marks)

Suppose that we are hired by a health insurance company to *determine whether alcohol consumption impacts claims amounts.* If so, the insurance company will incorporate alcohol usage information into their risk-rating system and it will affect the premiums of the future insured.

The response variable is `TotalClaims`, which is defined as the inflation-adjusted yearly total claims for each insured, for each year. Suppose that the policy has no upper limit, so the response is not bounded above. Suppose also that yearly claims are always strictly positive. The client informs you that it is known that claims distributions are always right-skewed.

The available covariates are:

- `AlcoholUsage`: encodes the average number of standard units of alcohol consumed per week, levels are `"None"`, `"1/week"`, `"2to5/week"`, `"6to14/week"`, `"15+/week"`
- `Age`: encodes the age of the insured in units of years
- `BMIcat`: encodes the categorisation of the body mass index of the insured, levels are `under`, `normal`, `over`, `obese`

It is suspected that, if `AlcoholUsage` has an effect, that the effect will not be constant across levels of `BMIcat`

We also have the following additional information regarding the observations on the dataset.

- `InsuredID`: A unique identifier associated with each policyholder.
- `County`: A unique identifier associated with the county where the insured lives.
- `Year`: the calendar year associated with the observation

We have multiple observations per policyholder, one for each policy year. We have multiple policyholders within each county, but not every county where the insurance company operates is represented in the dataset. Policyholders each live in exactly one county. It is believed that there is a temporal correlation in the claim amounts for each insured individual in excess of what may be explained by individual level indiosyncracies alone.

On the following page(s), propose a model for this dataset by providing a mathematical formulation for your model. Be sure to state the distibution of the response, the link function, the linear predictor, and the distribution(s) of any random effects. Define all terms you use. Explain your modelling choices (e.g. why did you select this distribution for the response). Explain how you would use your model to answer the main question of interest. What `R` function might you use to fit your model?

Question 4. Continued...

Question 4. Continued...

Question 5. Concepts (20 Marks)

5.a) (4 Marks) What is an odds-ratio and what is a probability ratio? Which one is easier to estimate from logistic regression output? For the following model, what is the odds ratio / probability ratio (whichever you said is easier) associated with a unit change in X_1?

$$Y_i \sim \text{Ber}(p_i) \qquad \text{logit}(p_i) = \beta_0 + \beta_1 X_1$$

5.b) (4 Marks) Why is an offset term useful in Poisson regression when we want to model event rates?

5.c) (4 Marks) Consider the model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + U_i + V_i X_{ij} + \epsilon_{ij} \qquad \begin{pmatrix} U_i \\ V_i \end{pmatrix} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \begin{bmatrix} \sigma_U^2 & \rho\sigma_U\sigma_V \\ \rho\sigma_U\sigma_V & \sigma_V^2 \end{bmatrix}\right) \qquad \epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

where β_0 and β_1 are fixed effects regression coefficients, X_{ij} is a deterministic regression covariate, U_i and V_i are correlated random effects, σ_U^2 and σ_V^2 are random effects variances, and ρ is the correlation coefficient of the U and V random effect. Circle the hypotheses for which a Likelihood Ratio Test is asymptotically valid.

- $H_0 : \sigma^2 = 0$
- $H_0 : \rho = 0$
- $H_0 : \rho = 1$
- $H_0 : \sigma_U^2 = \sigma_V^2$

5.d) (4 Marks) What is the approximate distribution of

$$\frac{\text{Null Deviance} - \text{Residual Deviance}}{(p-1)\hat{\phi}}$$

for an overdispersed Binomial model, where the Deviances are the values that would be output by R, and $\hat{\phi}$ is estimated from the data using the formula $\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(Y_i-\hat{\mu}_i)^2}{\hat{\mu}_i(1-\hat{\mu}_i)}$? (Don't forget to state the degrees of freedom!)

5.e) (4 Marks) Consider the following 3 models:

$$(Y_i|U_i) \sim \text{Binomial}(n_i, p_i) \qquad \text{logit}(p_i) = X_i\beta + U_i \qquad U_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_U^2) \qquad (1)$$

$$Y_i \sim \text{Binomial}(n_i, p_i) \qquad \text{logit}(p_i) = X_i\beta \qquad (2)$$

$$(Y_i|U_i) \sim \mathcal{N}(\mu_i, \sigma^2) \qquad \mu_i = X_i\beta + U_i \qquad U_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_U^2) \qquad (3)$$

One of the above models requires solving an intractable integral to evaluate the likelihood. Which one? Write down the likelihood function of the parameters in terms of this integral. (Don't try to evaluate the integral, just write it down).

Question 6. Weighted Least Squares Formula (10 marks)

Suppose that the vector $\mathbf{Y}$ is multivariate normal with **known** variance-covariance matrix, Σ. Suppose that there is some known matrix, $\mathbf{X}$, and some unknown vector of parameters β such that

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\beta, \Sigma)$$

so that the multivariate density of the vector $\mathbf{Y}$ is

$$f(\vec{y};\beta) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\vec{y} - \mathbf{X}\beta)^{\mathrm{T}}\Sigma^{-1}(\vec{y} - \mathbf{X}\beta)\right)$$

Derive the Maximum Likelihood Estimator for β in terms of $\mathbf{X}$ and $\mathbf{Y}$. Your answer should be the familiar weighted least squares formula with weight matrix, W, replaced by Σ^{-1}. Show all your work for full credit.

Question 6. Continued...