UNIVERSITY OF TORONTO
Faculty of Arts and Science

APRIL 2017 EXAMINATIONS

STA303H1S / STA1002H1S

Duration - 3 hours

Examination Aids: Scientific Calculator

STA 303/1002
Winter 2017
Final Exam
April 27, 2017
Time Limit: 3h

Last Name (Print): ____________________

First Name: ____________________

Student Number: ____________________

Check one: STA303 ☐ STA1002 ☐

This exam contains 17 pages (including this cover page) and 6 problems. Check to see if any pages are missing. Enter all requested information on the top of this page.

- You may *not* use your books or notes on this exam. You may use a scientific calculator and the formulae and tables at the end of the exam.
- MLE stands for Maximum Likelihood Estimate.
- REML stands for Restricted/Residual Maximum Likelihood
- You are required to show your work on each problem on this exam, except for the problems containing missing output in R. Please carry all possible precision through a numerical question, and give your final answer to four (4) decimals, unless they are trailing zeroes.
- You may use a benchmark of 5% for all inference, unless otherwise indicated. Round DF down to the nearest integer if not available on the table.
- When quoting effects, please give the magnitude, direction and evidence of the effect.
- Do not write in the table to the right.

Problem	Points	Score
1	15	
2	10	
3	20	
4	20	
5	10	
6	25	
Total:	100	

1. (15 points) Short questions.

   (a) (2 points) True or false? Explain your reasoning: If the outcome variable is quantitative and all explanatory variables take values 0 or 1, a logistic regression model is most appropriate.

   (b) (4 points) In the class we learned that $\hat{\beta}_{LS}$ is the solution that minimizes $SSE = \sum_i^n (Y_i - X_i\beta)^2$, $\hat{\beta}_{ridge}$ is the solution that minimizes $SSE_R = SSE + \lambda \sum_{j=1}^p \beta_j^2$ and $\hat{\beta}_{LASSO}$ minimizes $SSE_L = SSE + \lambda \sum_{j=1}^p |\beta_j|$. Now identify which figure (A or B) below is for LASSO, and explain why? What are these ellipse contours? And what is the solution of $\hat{\beta}$ inside the contours?

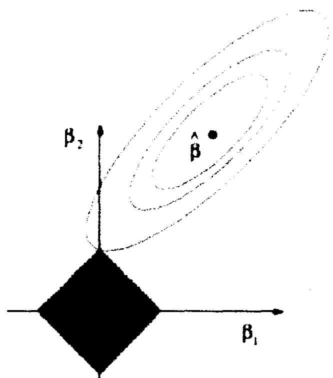

Figure A

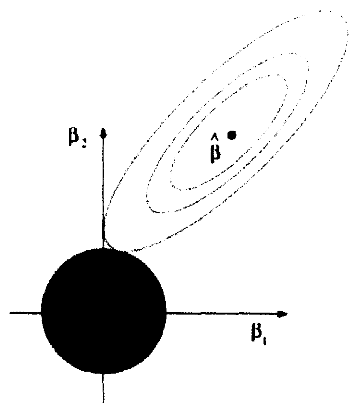

Figure B

(c) (3 points) True or false? Explain your reasoning: In a greenhouse experiment with several predictors, the response variable is the number of seeds that germinate out of 60 planted with each treatment combination. A Poisson regression model is most appropriate for this data.

(d) (2 points) True or false? Explain your reasoning: The same data is fit with two models using exactly the same predictors. The first model uses standard logistic regression (with **glm(...,family=binomial)**) while the second model accounts for overdispersion (with **glm(...,family=quasibinomial)**). The estimated coefficients for the predictors in the two models will be identical.

(e) (2 points) Both deviance and Pearson X^2 statistic are measures of discrepancy between observed and fitted values, and they can be used to compare nested models.

(f) (2 points) For one-way random effect model ($Y_{ij} = \mu + \tau_i + \epsilon_{ij}$ where $\tau_i \sim_{iid} N(0, \sigma_\mu^2) \perp \epsilon_{ij} \sim_{iid} N(0, \sigma^2)$), observations Y_{ij} are not independent.

2. (10 points) Answer the following two questions with detail.

   (a) (4 points) Assume we have an independent sample of size n from Poisson(λ_i) with the following probability density function

$$P(Y_i = y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots$$

   Show the deviance is

$$D = -2\{\ell(\hat{\beta}) - \ell_S(\tilde{\beta})\} = 2\sum_{i=1}^{n}\Big(y_i\frac{y_i}{\hat{\lambda}_i} - (y_i - \hat{\lambda}_i)\Big), \text{ where } \hat{\lambda}_i = \exp\{X_i\hat{\beta}\}$$

(b) (6 points) State the one-way ANOVA factor effect model and its assumptions. Be sure to define all terms in the model. Using this model to illustrate why we have identifiability issues (or estimation problem) in ANOVA, list two constraints that can be imposed to the model to fix the problem.

3. (20 points) Analysis of the following salary data set

Females		Males	
Salary	years	Salary	years
80	5	78	3
50	3	43	1
30	2	103	5
20	1	48	2
60	4	80	4

```
> with(salarydata, tapply(salary,sex,var)) # sample variance in gender groups
    F     M
570.0 616.3

                        ## Two-sample t-test
> t.test(salary~sex,var.equal=T)

Two Sample t-test
data:  salary by sex
t = -1.4542, df = 8, p-value = 0.184
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -57.91995  13.11995
sample estimates:
mean in group F mean in group M
           48.0            70.4

                        ## Model 1
> mod1 = lm(salary~sex,data=salarydata)
> anova(mod1)

Response: salary
          Df Sum Sq Mean Sq F value Pr(>F)
sex        1 1254.4   [A]     [B]     [C]
Residuals [D]   [E]    [F]

                        ## Model 2
> mod2=lm(salary~years*sex,data=salarydata)
> anova(mod2)

Response: salary
          Df Sum Sq Mean Sq  F value    Pr(>F)
years      1 4560.2  4560.2 148.0584 1.874e-05 ***
sex        1 1254.4  1254.4  40.7273 0.0006961 ***
years:sex  1    0.2     0.2   0.0065 0.9383948
Residuals  6  184.8    30.8
```

```
                          ## Model 3
> mod3=lm(salary~years+sex,data=salarydata)
> anova(mod3)

Response: salary
          Df Sum Sq Mean Sq F value    Pr(>F)
years      1 4560.2  4560.2 172.548 3.458e-06 ***
sex        1 1254.4  1254.4  47.464 0.0002335 ***
Residuals  7  185.0    26.4

                          ## Model comparison
> anova(mod1,mod3,mod2,test="Chisq")

Model 1: salary ~ sex
Model 2: salary ~ years + sex
Model 3: salary ~ years * sex
  Res.Df    RSS Df Sum of Sq Pr(>Chi)
1      8 4745.2
2      7  185.0  1    4560.2   <2e-16 ***
3      6  184.8  1       0.2   0.9358
```

(a) (5 points) Write down the H_0 and H_a for the two-sample t-test. Is it ok to assume equal variance in the two-sample t-test (explain your reasoning)? Based on t-test result, what conclusion do you have?

(b) (2 points) True or false (no explanation): the two-sample t-test is equivalent to the one-way ANOVA analysis in this case.

(c) (6 points) Some values in model 1 output have been replaced with letters. Fill in those values.

(A)=__________ (B)=__________ (C)=__________

(D)=__________ (E)=__________ (F)=__________

(d) (3 points) In this data set, we could use years of working as covariate and run ANCOVA analysis. But for ANCOVA analysis, except all the assumptions for ANOVA, we have one more crucial assumption, what is it and which model can be used to evaluate this assumption? Is this assumption violated?

(e) (2 points) Which model gives the correct ANCOVA output? Taking into account of years of working, how significant is it on testing no gender gap in pay (be sure to report the p-value)?

(f) (2 points) Which model (1,2, 3) do you prefer and why?

4. (15 points) A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution (rank: factor variable with 4 levels, value 1 is the highest prestige and 4 is lowest), affect admission into graduate school. The response variable, admit/don't admit, is a binary variable.

```
                          ## Summary output of model
> summary(q4mod)
glm(formula = admit ~ gpa + gre + rank, family = binomial, data = adm)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.989979   1.139951  -3.500 0.000465 ***
gpa          0.804038   0.331819   2.423 0.015388 *
gre          0.002264   0.001094   2.070 0.038465 *
rank2       -0.675443   0.316490  -2.134 0.032829 *
rank3       -1.340204   0.345306  -3.881 0.000104 ***
rank4       -1.551464   0.417832  -3.713 0.000205 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 458.52  on 394  degrees of freedom
AIC: 470.52

                          ## var-covariance matrix of beta.hat
> print(vcov(q4mod),digits=3) # var-cov matrix of the estimates of coefficients
            (Intercept)       gpa         gre       rank2      rank3      rank4
(Intercept)    1.299488 -0.303660 -0.00030122 -0.08447562 -0.0486435 -0.0894313
gpa           -0.303660  0.110104 -0.00012418  0.00452051 -0.0094686  0.0035683
gre           -0.000301 -0.000124  0.00000120 -0.00000169  0.0000186  0.0000118
rank2         -0.084476  0.004521 -0.00000169  0.10016571  0.0695664  0.0701272
rank3         -0.048644 -0.009469  0.00001861  0.06956636  0.1192365  0.0697416
rank4         -0.089431  0.003568  0.00001184  0.07012724  0.0697416  0.1745833

                          ## sum of squared Pearson residuals
> sum(residuals(q4mod, type = "pearson")^2) # sum of squared Pearson residuals
[1] 397.4902
                          ## model selection
> anova(q4mod,test="Chisq")
Analysis of Deviance Table

Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                   399     499.98
gpa   1  13.0089       398     486.97   0.00031 ***
gre   1   6.6236       397     480.34   0.01006 *
rank  3  21.8265       394     458.52 7.088e-05 ***
```

(a) (2 points) Based on R output, find the estimate of the dispersion parameter using Pearson test statistic. Do we have overdispersion or underdispersion?

(b) (6 points) Based on the R output, obtain the estimated odds ratio (OR) of admitted of subjects whose university rank is 4, and those whose university rank is 2 (Show steps for full mark on calculation). What conclusion do you have based on the estimate of OR? i.e. how to interpret the OR.

(c) (4 points) Find the variance estimate of the log(OR) where the OR is defined in (b).

(d) (4 points) Based on the R output, obtain a 95% confidence interval for the OR in (b). You could use the following quantile from N(0,1) distribution.

$$P(Z < 1.96) = 0.975, P(Z < 1.645) = 0.95, Z \sim N(0, 1)$$

(e) (4 points) Applying likelihood ratio test based on deviance to compare the following 3 models

M1: glm(formula = admit ~ gpa , family = binomial, data = adm)

M2: glm(formula = admit ~gpa+gre , family = binomial, data = adm)

M3: glm(formula = admit ~ gpa+gre+rank , family = binomial, data = adm)

Based on R output: from M1 to M3, what is the deviance drop and the associated d.f. to the change in deviance?

5. (10 points) From the data from previous question, we have the following 2-way contingency table for the rank and admission status.

Rank\Admission	No	Yes	Row sum
1	28	33	**61**
2	97	54	**151**
3	93	28	**121**
4	95	12	**107**
Colum sum	**313**	**127**	***Total=440***

```
> rank = as.factor(rep(c(1,2,3,4),2))
> admit=as.factor(rep(c(0,1),each=4))
> y0=c(28,97,93,95); y1=c(33,54,28,12)
> rankFac=as.factor(c(1,2,3,4))
>
> adm2=data.frame(rank,admit,y=c(y0,y1))
> fitp = glm(y~rank*admit,family=poisson,data=adm2)
> fitb = glm(cbind(y1,y0)~rankFac,family=binomial)

                    ## ANOVA output of Poisson model

> anova(fitp,test="LRT")
Analysis of Deviance Table
Model: poisson, link: log
Response: y

Terms added sequentially (first to last)

           Df Deviance Resid. Df Resid. Dev        Pr(>Chi)
NULL                           7    163.765
rank        3   40.889         4    122.876 0.000000006903 ***
admit       1   81.154         3     41.722      < 2.2e-16 ***
rank:admit  3   41.722         0      0.000 0.000000004596 ***

                    ## ANOVA output of Logit model

> anova(fitb,test="LRT")
Analysis of Deviance Table
Model: binomial, link: logit
Response: cbind(y1, y0)

Terms added sequentially (first to last)

        Df Deviance Resid. Df Resid. Dev        Pr(>Chi)
NULL                        3     41.722
rankFac [A]    [B]          0      0.000         [C] ***
```

(a) (3 points) Some values have been replaced with letters in the ANOVA output of Logit model. Fill in those values. You do not need to show any work for this part.

(A)= (B)= (C)=

(b) (5 points) State the independence (log-linear) model for the two-way contingency table and assumptions. Compute the fitted values in the top right count of the table, i.e. (rank=1, admission=Yes cholesterol) according to the model.

(c) (2 points) Based on R output, either from the Poisson model or Logit model, is it reasonable to conclude that the row and column variables are independent? Explain your reasoning.

6. (25 points) Researchers study the performance of nurse practitioners in only three **specialities**(paediatrics, obstetrics and diabetes). They randomly selected 3 **cities**, and recorded competency scores of 4 nurses randomly selected within each speciality and each city. The **scores** are on a continuous scale, and the values are summarized below.

	City 1		City 2		City 3		Mean
Diabetes	71.50	49.80	80.20	76.10	48.70	54.40	61.4
	55.10	75.40	44.20	50.50	60.10	70.80	
Obstetrics	80.10	76.20	71.30	73.40	90.10	65.60	77.85833
	70.30	89.50	76.90	87.20	74.60	79.10	
Pediatrics	91.70	74.90	86.30	88.10	82.30	78.70	83.79167
	88.20	79.50	92.00	69.50	89.80	84.50	
	75.18333		74.64167		73.22500		74.35

```
> anova(lm(score~spec*city,data=data6))

Response: score
          Df Sum Sq Mean Sq F value    Pr(>F)
spec       2 3229.9 1614.94 15.0243 4.112e-05 ***
city       2   24.5   12.27  0.1142    0.8925
spec:city  4   34.5    8.63  0.0803    0.9877
Residuals 27 2902.2  107.49
```

(a) (6 points) State the linear mixed effect model that is appropriate for this data set, and the assumptions (be sure to define all terms in the model).

(b) (3 points) Assume α_i is the fixed effect of treatment group i. Provide the estimates of the fixed effects of the model with the zero-sum constraint ($\sum_i^3 \alpha_i = 0$), that is, find the estimate of $\alpha_i, i = 1, 2, 3$.

(c) (3 points) Provide the estimates of the fixed effects of the model with the reference constraint ($\alpha_1 = 0$)

(d) (7 points) Based on the given R output and formula in last page, estimate and interpret the three variance components of the model.

(e) (6 points) If we decide to exclude **city** (both main effects and interactions) from the model. Use the new model to test whether there is a difference between the specialities. State the new model (use the notation you have in Q6-(a), so you don't need to define all the terms). What is the observed test statistic on testing whether there is a difference between the specialities and what's your conclusion?

You might use the following quartile from F distribution

(df1,df2)	(2,27)	(3,27)	(2,33)	(3,33)
$F_{0.95,df1,df2}$	3.3541	2.9603	3.2849	2.8915

Some formulae:

One-way Analysis

$$SST = \sum_i^a \sum_j^b (Y_{ij} - \bar{Y}_{..})^2, \quad SS_{trmt} = \sum_i^a (\bar{Y}_{i.} - \bar{Y}_{..})^2, \quad SSE = SST - SS_{trmt} = b\sum_i^a (Y_{ij} - \bar{Y}_{i.})^2$$

Deviance for Bernoulli, Binomial and Poisson distribution

$$D_{Bern} = -2\{\ell(\hat{\beta}) - \ell_S(\tilde{\beta})\} = 2\sum_i^n \left\{Y_i \log \frac{Y_i}{\hat{\pi}} + (1 - Y_i)\log \frac{1 - Y_i}{1 - \hat{\pi}}\right\}$$

$$D_{Bino} = -2\{\ell(\hat{\beta}) - \ell_S(\tilde{\beta})\} = 2\sum_i^g \left\{Y_i \log \frac{Y_i}{\hat{Y}_i} + (n_i - Y_i)\log \frac{n_i - Y_i}{n_i - \hat{Y}_i}\right\}, \hat{Y}_i = n_i\hat{\pi}_i$$

$$D_{Pois} = -2\{\ell(\hat{\beta}) - \ell_S(\tilde{\beta})\} = 2\sum_i^n \left\{Y_i \log \frac{Y_i}{\hat{\lambda}_i} - (Y_i - \hat{\lambda}_i)\right\}, \hat{\lambda}_i = \exp(X_i\hat{\beta})$$

Log-linear model for 2-way/3-way Contingency Table

indep. model $\log(E(Y_{ij})) = \mu + \alpha_i + \beta_j, \pi_{ij} = \pi_i\pi_j, \hat{\lambda}_{ij} = n_{i.}n_{.j}/n$

Model	$\log E\{Y_{ijk}\} =$	$\pi_{ijk} =$	$\lambda_{ijk} =$
Mut. Indep	$\mu + \alpha_i + \beta_j + \gamma_k$	$\pi_i\pi_j\pi_k$	$n_{i++}n_{+j+}n_{++k}/n^2$
Joint Indep.	$\mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij}$	$\pi_{ij}\pi_k$	$n_{ij+}n_{++k}/n$
Cond. Indep.	$\mu + \alpha_i + \beta_j + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}$	$\pi_{ik}\pi_{jk}/\pi_k$	$n_{i+k}n_{+jk}/n_{++k}$
Unif. Assoc.	$\mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}$	$\pi_{ij}\pi_{ik}\pi_{jk}$	Iterative
Saturated	$\mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}$	π_{ijk}	n_{ijk}

Two-Way ANOVA table
(which could be reduced to two-way additive model and one-way ANOVA table)

Source	d.f.	SS	MS	F
Factor A	a-1	SS_A	MSA	MSA/MSE
Factor B	b-1	SS_B	MSB	MSB/MSE
AB	(a-1)(b-1)	SS_{AB}	MSAB	MSAB/MSE
Error	ab(n-1)	SSE	MSE	
Total	abn-1	SST		

Expected Mean Squares for Balanced Two-factor ANOVA models

Mean Square	d.f.	Fixed ANOVA model (A and B fixed)	Random ANOVA model (A and B random)	Mixed ANOVA model (A fixed, B random)
MSA	a-1	$\sigma^2 + nb\sum_i^a \alpha_i^2/(a-1)$	$\sigma^2 + nb\sigma_\alpha^2 + n\sigma_{\alpha\beta}^2$	$\sigma^2 + nb\frac{\sum_i^a \alpha_i^2}{a-1} + n\sigma_{\alpha\beta}^2$
MSB	b-1	$\sigma^2 + na\sum_j^b \beta_j^2/(b-1)$	$\sigma^2 + na\sigma_\beta^2 + n\sigma_{\alpha\beta}^2$	$\sigma^2 + na\sigma_\beta^2 + n\sigma_{\alpha\beta}^2$
MSAB	(a-1)(b-1)	$\sigma^2 + n\frac{\sum_i^a\sum_j^b(\alpha\beta)_{ij}^2}{(a-1)(b-1)}$	$\sigma^2 + n\sigma_{\alpha\beta}^2$	$\sigma^2 + n\sigma_{\alpha\beta}^2$
MSE	ab(n-1)	σ^2	σ^2	σ^2