STA 303/1002
Winter 2017
Midterm
3/2/2016
Time Limit: 1h 40 min

Last Name (Print): _____
First Name (Print): _____
Student Number (Print): _____
Mail (print): _____@mail.utoronto.ca

Check one: STA303 ☐ STA1002 ☐

This exam contains 8 pages (including this cover page) and 4 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, the formulae below, and the t-table on the last page.
- MLE stands for Maximum Likelihood Estimate. AIC stands for Akaike Information Criterion. MLR stands for Multiple Linear Regression.
- Show your work on each problem, and please carry all possible precision through a numerical question. Give your final answer to four (4) decimals, unless they are trailing zeroes.
- You may use a benchmark of $\alpha = 5\%$ for all inference, unless otherwise indicated. Do not write in the table to the right.

Problem	Points	Score
1	10	
2	15	
3	15	
4	10	
Total:	50	

Some formulae:

Linear Regression & Pooled t-test

$$b_1 = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2} = \frac{\sum x_i y_i - n\overline{x}\overline{y}}{\sum x_i^2 - n\overline{x}^2} \qquad b_0 = \overline{y} - b_1\overline{x} \qquad t_{obs} = \frac{\overline{y}_1-\overline{y}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

One-way analysis of variance

$$\text{SSTO} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{..})^2 \qquad \text{SSE or } SS_W = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i.})^2$$
$$SS_{trt} \text{ or } SS_B = \sum_{i=1}^{r} n_i(\bar{Y}_{i.}-\bar{Y}_{..})^2$$

Bernoulli and Binomial distributions

If $Y \sim \text{Bernoulli}(\pi)$
$\text{E}(Y) = \pi$, $\text{Var}(Y) = \pi(1-\pi)$

If $Y \sim \text{Binomial}(m, \pi)$
$\text{E}(Y) = m\pi$, $\text{Var}(Y) = m\pi(1-\pi)$

Logistic Regression with Binomial Response formulae

$$\text{Deviance} = -2ln(\hat{\beta}) = 2\sum_{i=1}^{n}\{y_i\log(y_i) + (m_i-y_i)\log(m_i-y_i) - y_i\log(\hat{y}_i) + (m_i-y_i)\log(m_i-\hat{y}_i)\}$$

Model Fitting Criteria

$$\text{AIC} = -2\log(L) + 2(p+1) \qquad \text{SC} = -2\log(L) + (p+1)\log(N)$$

1. Short answer.

   (a) (3 points) What do ANOVA and ANCOVA stand for? What is the difference between them?

   Solution:
   ANOVA stands for Analysis of Variance while ANCOVA stands for Analysis of Covariance. (1)
   An ANOVA is a regression where all of the covariates are categorical (1). An ANCOVA is a regression with qualitative and continuous covariates, but without interaction terms between the factors and the continuous explanatory variables (i.e., the so called "homogeneous slopes assumption") (1)

   (b) (3 points) For Locally weighted regression, we solve a separate weighted least squares problem at each target point X_0

   $$\hat{\beta} = \arg\min_\beta \sum_{i=1}^{n} w_i(Y_i - X_i\hat{\beta})^2, w_i = K(X_0, X_i) \text{ is a kernel function}$$

   Can locally weighted regression exactly reproduce the learning behaviour of ordinary least-squares regression, given a suitable kernel, for any data? If so, how? If not, why not?

   Solution: Yes.(1) Simply use the kernel K(X_0, X_i)= c for any constant c. Then all examples are weighted equally and the algorithm behaves exactly like ordinary unweighted regression. (2)

   (c) (2 points) The ridge regression uses L_2 penalty with the following loss function.

   $$L = (Y - X\beta)'(Y - X\beta) + \lambda\beta'\beta,\ \lambda > 0$$

   Suppose we accidentally write $L = (Y - X\beta)'(Y - X\beta) + \lambda Y'Y$ instead. Explain why this form of "regularization" has no effect.

   Solution: The optimal solution for β is determined by setting $\partial L/\partial\beta$ to zero; (1) since $\partial Y'Y/\partial\beta = 0$, it has no effect on the solution. (1)

   (d) (2 points) In a logistic model, we estimate the unknown parameter β by maximizing the likelihood function, this is equivalent to finding parameter values that minimize the deviance. Is this statement true or false? Explain.

   Solution: True. (1), Deviance is defined as

   $$D = -2ln\left(\frac{\text{Likelihood of fitted model}}{\text{Likelihood of saturated model}}\right) = -2ln\{\text{Likelihood of fitted model}\}$$

   So to maximize $L(\beta)$ is the same to miminize -2ln $L(\beta)$. (1)

2. Answer the following three questions.

   (a) (6 points) $X_1, \ldots, X_{n_x} \sim_{IID} N(\mu_x, \sigma^2), Y_1, \ldots, Y_{n_y} \sim_{IID} N(\mu_y, \sigma^2)$. A 95% confidence interval to $\mu_y - \mu_x$ was calculated to be (2, 10) when using a pooled procedure from samples of size $n_x = 25$ and $n_y = 37$. We are interested in the following testing:

   $$H_0 : \mu_y = \mu_x \quad vs \quad H_a : \mu_y \neq \mu_x$$

   What is the test statistic under H_0 and what distribution does it follow? Based on the given information, calculate the observed value for the test statistic.

Table 1: Critical value under t_{df} distribution

	df=24	df=35	df=62	df=61	df=60
$P(T_{df} < t^0) = 0.025$	-2.0639	-2.0281	-1.9990	-1.9996	-2.000

Solution:

Test statistic under H_0 and its distribution:

$$T = \frac{\bar{Y} - \bar{X} - 0}{s_p\sqrt{1/25 + 1/37}} \ (1) \sim t_{60}, \ (1)$$

where

$$s_p^2 = \frac{\sum_{i=1}^{n_x}(X_i - \bar{X})^2 + \sum_{j=1}^{n_y}(Y_j - \bar{Y})^2}{n_x + n_y - 2} = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2} \ (1).$$

Since the $(1 - \alpha)$% CI to $\mu_y - \mu_x$ is

$$\bar{Y} - \bar{X} \pm t_{60,1-\alpha/2} s_p \sqrt{1/25 + 1/37} = (2, 10),$$

this gives $\bar{Y} - \bar{X} = (2 + 10)/2 = 6$ (1), $t_{60,0.975} s_p \sqrt{1/25 + 1/37} = (10 - 2)/2 = 4$. (1)
It follows that the observed test statistic is

$$T = \frac{\bar{Y} - \bar{X}}{s_p\sqrt{1/25 + 1/37}} = \frac{6 - 0}{4/t_{60,0.975}} = \frac{6 - 0}{4/2} = 3. \ (1)$$

(b) (4 points) Show that in one-way ANOVA model, $SS_{tot} = SS_W + SS_B$, that is

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{..})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i.})^2+\sum_{i=1}^{r} n_i(\bar{Y}_{i.}-\bar{Y}_{..})^2 = \sum_{i=1}^{r}(n_i-1)S_i^2+\sum_{i=1}^{r} n_i(\bar{Y}_{i.}-\bar{Y}_{..})^2$$

Solution:

$$\begin{aligned}
\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{..})^2 &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i.}+\bar{Y}_{i.}-\bar{Y}_{..})^2 \ (1)\\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left\{(Y_{ij}-\bar{Y}_{i.})^2+(\bar{Y}_{i.}-\bar{Y}_{..})^2+2(Y_{ij}-\bar{Y}_{i.})(\bar{Y}_{i.}-\bar{Y}_{..})\right\}\\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i.})^2+\sum_{i=1}^{r} n_i(Y_{i.}-\bar{Y}_{..})^2+2\sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{Y}_{i.}-\bar{Y}_{..})(Y_{ij}-\bar{Y}_{i.})\\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i.})^2+\sum_{i=1}^{r} n_i(Y_{i.}-\bar{Y}_{..})^2+2\sum_{i=1}^{r}(\bar{Y}_{i.}-\bar{Y}_{..})\Big(\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i.})\Big) \ (1)\\
&= SS_W+SS_B+2\sum_{i=1}^{r}(\bar{Y}_{i.}-\bar{Y}_{..})\Big(\sum_{j=1}^{n_i}Y_{ij}-n_i\bar{Y}_{i.}\Big) \ (1)\\
&= SS_W+SS_B+2\sum_{i=1}^{r}(\bar{Y}_{i.}-\bar{Y}_{..})*0 \ (1)\\
&= SS_W+SS_B
\end{aligned}$$

(c) (5 points) A horticultural experiment investigates the effects of t = 4 different herbicide formulations on weed growth. Summary statistics are given below. Fill in all 10 blanks

Treatment	Sample size (n_i)	sample mean $\bar{Y}_{i.}$	stad. dev. S_i
Control	5	300	50
A300	5	260	40
A400	5	240	30
Surflan	5	200	40

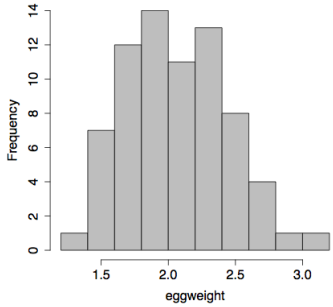
in the ANOVA table below (or write NP if not possible). You may use the fact that $\sum_{i=1}^{4}\sum_{j=1}^{5}(\bar{Y}_{i.}-\bar{Y}_{..})^2$= 26000. Mark is given only for final answer filled in blank.

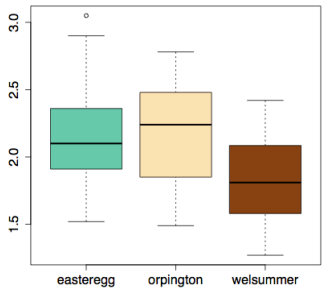
Source	DF	Sum of Squares	Mean Squares	F statistic	P-value
Treatment	A=3	B=26000	C=8667	D=5.25	F=NP
Error	F=16	G=26400	H= 1650		
Total	I=19	J=52400			

(0.5)*10

3. Comparison of egg size across 3 breeds. Tom has 7 egg-laying chickens comprised of 3 breeds on his farm: 2 Easter Eggers (E), 2 Buff Orpingtons (B), and 2 Welsummers (W). He wanted to perform a scientific study to determine whether the size of eggs layed by his chickens over the course of 14 days, and also recorded which breed they came from. Let Y_{ij} be the j-th egg size in i-th breed. The summary statistics table, histogram of egg weights, and the boxplot of egg weights across the 3 chicken breeds are presented blow.

Breed	Mean	Std. Dev.	n
Easter Egger	2.145	0.3543	33
Buff Orpington	2.182	0.3970	19
Welsummer	1.852	0.3277	20





Source	df	Sum Sq	Mean Sq	F-stat	p-value
breed	2	1.3818	0.6969	5.36	0.00714
Error	69	8.894	0.1289		

(a) (2 points) Specify the linear regression model behind the ANOVA model.

Solution:

$$Y_{ij} = \beta_0 + \beta_1 I_B + \beta_2 I_W + \epsilon_{ij}, i = 1, 2, 3, j = 1, \ldots, n_i \quad (1)$$

$$I_B = \begin{cases} 1, & \text{Breed is Buff Orpington} \\ 0, & \text{otherwise} \end{cases}, \quad I_W = \begin{cases} 1, & \text{Breed is Welsummer} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

(b) (3 points) Specify a fixed effect model for the ANOVA table. For the p-value =0.00714, what is the null and alternative hypothesis (answer should use notations you define in the fixed effect model)? What conclusion do you have?

Solution:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad (1)$$

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \quad versus \quad H_a : \text{ at least one is not zero} \quad (1)$$

The observed F-stat is 5.36 and the corresponding p-value is 0.00714 which is less than 0.05. We have strong evidence that the average egg weight varies across these 3 groups. (1)

(c) (5 points) What are the assumptions for the test in part (b) above? Comment on whether these assumptions are reasonable.

Solution:

- Independence: the independence within in groups is definitely violated since many eggs are from the chicken. Independence between groups might also be violated since the feeding could be the same and it affects all the egg size. (2)
- Normality: a bell shape is observed in the histogram, and boxplots appear symmetric, so the normality assumption looks reasonable but there is observation in E group with relative large size but it is not too extreme. (1)
- Constant variance: this looks pretty good from the boxplot and the column of Std. dev. in the summary table. Consider the variance ratio, $S^2_{max}/S^2_{min} = 0.3970^2/0.3277^2 = 1.4677 < 2$, it suggests that we don't have evidence that the constant variance is violated.(2)

(d) (5 points) It is claimed that Buff Orpingtons will lay the largest eggs of these 3 breeds, on average. Conduct an appropriate 1-sided test to the following hypotheses where μ_{E+W} is mean egg size of E and W breeds.

$$H_0 : \mu_B = \mu_{E+W} \quad versus \quad H_a : \mu_B > \mu_{E+W}$$

Make sure you give the test statistic under H_0 and specify the null distribution correctly.

Table 2: Critical value under t_{df} distribution

	df=18	df=51	df=72	df=71	df=70	df=69
$P(T_{df} < t^0) = 0.05$	-1.734	-1.675	-1.666	-1.667	-1.667	-1.667

Solution:

$$\hat{\mu}_{E+W} = \frac{33 * 2.145 + 20 * 1.852}{33 + 20} = 2.0344 \quad (1)$$

$$T|_{H_0} = \frac{(2.182 - 2.0344) - 0}{\sqrt{MSE(1/19 + 1/53)}} = \frac{2.182 - 2.0344}{\sqrt{0.1289(1/19 + 1/53)}} = 1.5375 \quad (2)$$

$T|_{H_0} \sim t_{69}$ and $t_{69,0.975} = 1.667$ (1)

1.5375<1.667, we do not have enough evidence to that Buff Orpingtons will lay the largest eggs of these 3 breeds, on average. (1)

4. Below is some output from the Donner Party example from lecture and the problem set 2.

```
     AGE            SEX           STATUS
Min.   :15.0   FEMALE:15   DIED    :25
1st Qu.:24.0   MALE  :30   SURVIVED:20
Median :28.0
Mean   :31.8
3rd Qu.:40.0
Max.   :65.0


Call:
glm(formula = STATUS ~ SEX + AGE, family = "binomial", data = donner)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.23041    1.38686   2.329   0.0198
SEXMALE     -1.59729    0.75547  -2.114   0.0345
AGE         -0.07820    0.03728  -2.097   0.0359

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 51.256  on 42  degrees of freedom
AIC: 57.256

Number of Fisher Scoring iterations: 4
```

(a) (4 points) OR (odds-ratio) of A compared to B is defined as the ratio of odds of A divided by odds of B. From the output, find the OR of survival for female compared to male and its associated 95% confidence interval. $P(Z > 1.96) = 0.025, Z \sim N(0, 1)$.

Solution:
The odds-ratio of survival for male compared to female is $OR_M = \exp(-1.59729) = 0.2024444$. And the confidence interval for OR_M is

$$\exp(-1.59729 \pm 1.96 * 0.75547) = (0.0461, 0.8900) \quad (1)$$

So the odds-ratio of survival for female compared to male OR_F is

$$OR_F = \frac{1}{OR_M} = \frac{1}{\exp(-1.59729)} = 4.9396 \quad (1)$$

The 95% CI for OR_F is

$$(\frac{1}{0.8900}, \frac{1}{0.0461}) = (1.1236, 21.6920) \quad (2)$$

(b) (1 point) What is the fitted male model and what is the fitted female model?

Solution: The fitted male model: (0.5)

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 2.070681 - 0.07820 * age, \quad \hat{\pi} = \hat{P}(\text{survivorship} = 1 | age, sex = 1)$$

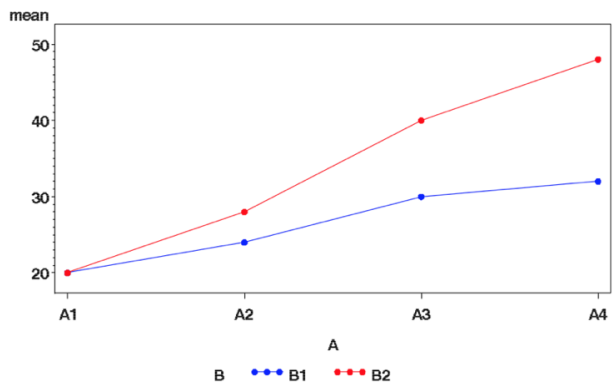
The fitted female model: (0.5)

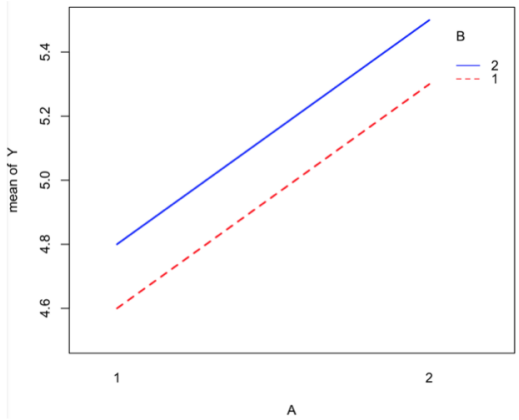
$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 3.23041 - 0.07820 * age, \quad \hat{\pi} = \hat{P}(\text{survivorship} = 1 | age, sex = 0)$$

(c) (2 points) If age is increased by 10 years, the odds of survival decreases by 54.25%. This is true for either a male of a female. Is this statement true? Justify your answer.

Solution:
True. (1)
Odds(age+10,sex) = Odds(age,sex)∗exp(−0.07820) = Odds(age,sex)∗0.4575 (1)

(d) (3 points) Give comments on the following two interaction plots.





Solution:

- Left plot: both main effect and interaction effect (reinforcement effect) exist. The main effect of A is bigger for B=B2 than for B=B1. (1.5)
- Right plot, there is no interaction effect since lines are parallel. As A changes from level 1 to 2, the marginal mean increases. Similar, the marginal mean of B moves as the level of B goes up. We conclude that we have both main effects but no interaction effect in this case. (1.5)