Statistics 252 – Final Exam – Version A

Instructor: Paul Cartledge

Instructions:

1. *Read all the instructions CAREFULLY.*
2. *This is a closed book exam.*
3. *You may only use the formula sheets, the output provided, and a non-programmable calculator.*
4. *You have 3 hours to complete the exam.*
5. *The exam is out of a total of 80 marks and has 14 pages (in two parts).*
6. *Show your work in all sections to receive full credit.*
7. *Use the reverse side of pages for scrap work.*
8. *Make sure your name and signature are on the front and that your ID number is on the top of page two.*
9. *All biological examples approved by a quasi-medical professional.*
10. *When referring to "log", I am always referring to the natural log.*
11. *Unless instructed otherwise, give a range for the p-value. Also, use the "judgment approach" to help state your conclusion in plain English.*
12. *When asked for a "confidence interval", state the estimate, the standard error, and the critical value. Then, calculate and interpret the interval.*
13. *When asked to "carry out a full analysis in detail", set up the hypotheses, calculate the test statistic, state the distribution of the test statistic (i.e. t_9 or $F_{3,10}$), find the p-value (or its range), and state your conclusion in plain English.*

Name: **SOLUTION**

Signature: ___________________________________

Component	Notes	Worth	Mark
Short Answer	3 questions	8	
What test?	2 questions	6	
Case Study 1	7 parts	31	
Case Study 2	7 parts	35	
Total		80	

ID:____________________________

Short Answer Problems (8 marks)

Question 1 (2 marks) In testing for five equal means, you set up the following set of hypotheses. Is there something wrong? If so, correct the mistake by simply re-writing the hypotheses. If not, say why there is nothing wrong.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = 0$

H_A: at least one $\mu_i \neq 0 \quad (i = 1, \ldots, 5)$

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_A: at least one μ is different from the others

Question 2 (3 marks) Suppose Harry Potter determines a simple linear relationship between his grade average (in %) each year at Hogwarts School of Witchcraft and Wizardry and how many hours he spends studying per week. Checking assumptions, the response variable required a log-transformation. The estimated regression line is $\hat{\mu}(\ln(grade) \mid study) = 4.587 - 0.0300 study$. Estimate the change in grade associated with a change of 3 to 5 hours of studying per week. Give a statement relating this change to the given variables. Also, predict the grade for first value.

$$e^{k\hat{\beta}_1} = e^{(5-3)(-0.03)} = e^{-0.06} = 0.942$$

"An additive change from 3 to 5 hours is associated with a multiplicative change of 0.942 in median grade average."

$\hat{\mu}(\ln(grade) \mid study = 3) = 4.587 - 0.0300(3) = 4.497$

$e^{4.497} = 89.747$

Question 3 (3 marks) Suppose an ANOVA table for simple linear regression gave a test statistic of 8.423. If the standard deviation of the errors is 2.52 and there are 50 observations, what is the coefficient of determination? [Hint: $R^2 = SS(Extra)/SS(Total)$]

Source	**SS**	**df**	**MS**	***F***
Regression	**53.489**	1	**53.489**	8.423
Residual	**304.819**	48	6.3504	
Total	**358.309**	49		

$s_e = 2.52 \rightarrow MSE = s_e^2 = 2.52^2 = 6.3504$

$$R^2 = \frac{SS(Extra)}{SS(Total)} = \frac{53.489}{358.309} = 0.149 \rightarrow 14.9\%$$

<u>What test would you use? (6 marks)</u>

In each scenario, identify the appropriate procedure needed to answer the question. Be as descriptive as possible.
Choose from the following:

i) One-Sample t-test for a single population mean,
ii) Paired t-test for the difference between two population means,
iii) Two Independent Sample t-test for the difference between two population means,
iv) One-Factor ANOVA F-test for any difference among *I* population means,
v) A t-test for a linear combination of means,
vi) Some Extra-Sum-of-Squares F-test comparing two models for the *I* population means,
vii) An ANOVA F-test for any regression model effects,
viii) A t-test for a single regression coefficient,
ix) Some Extra-Sum-of-Squares F-test comparing two regression models (testing a subset of coefficients),
x) An F-test for any factor effects (main OR interaction) in Two-Factor ANOVA,
xi) The F-test for additivity in a Two-Factor ANOVA.

Question 4 (3 marks) An animated biologist in need of money is trying to see if he can get frogs to sing. Randomly sampling 20 frogs in a nearby swamp, 20 frogs in a swamp 2 km away, and 20 frogs in a swamp 20 km away, he measures each frog's pitch level. In testing to see if the closest frogs have a different average pitch level than the frogs further away, what test would you use? What is the distribution of the test statistic?

v) A t-test for a linear combination of means

$t_{N-I} = t_{3\times 20-3} = t_{57}$

Question 5 (3 marks) A local irate samurai is known to businesses as an expert in "fixing" aggravating photocopiers. For future potential customers, he decides to prove his worth by randomly sampling 60 photocopiers and measuring their performances, before and after he has "fixed" them. What test would you use to assist this samurai? What is the distribution of the test statistic?

Paired *t*-test

$t_{n-1} = t_{60-1} = t_{59}$

Case Study 1 – Music written by that guy… (25 marks)

When needed, use the output on pages 11 and 12 to answer the following questions.

In class, we determined that everyone's movie knowledge could be more awesome, especially when it comes to film music composer, John Williams. Suppose an observational study was done to investigate 58 random films where John Williams composed the music, where the response variable is the film's rating at IMDB.com (*Rating*, measured on a scale of 0 to 10, with 10 being the highest). The film's rating, the film's length (*Runtime*, measured in minutes), and budget (measured in millions of US$, but requiring log transformation) are modeled as continuous (numerical) variables. "Director credit" is categorized into 3 levels: Steven Spielberg, George Lucas, and Other. Each film was also categorized (*Nom*) by whether or not it received an Academy Award nomination for Best Music Score.

To fit an MLR model, the categorical variable *Dir* uses the first two levels listed to correspond to indicator variables. Use the following "original model" to answer questions.

$$\mu(Rating \mid Runtime, ln(Budget), Dir, Nom) = \beta_0 + \beta_1 Runtime + \beta_2 ln(Budget) + \beta_3 Dir1 + \beta_4 Dir2 + \beta_5 Nom + \beta_6 Runtime \times ln(Budget) + \beta_7 Runtime \times Dir1 + \beta_8 Runtime \times Dir2 + \beta_9 Runtime \times Dir1 \times Nom + \beta_{10} Runtime \times Dir2 \times Nom$$

a) (3 marks) What is the effect of runtime on mean rating, after accounting for director credit, Academy Award nomination, and *ln*(budget)?

$\mu(Rating \mid Runtime + 1, \ln(Budget), Dir, Nom) - \mu(Rating \mid Runtime, \ln(Budget), Dir, Nom)$

$= [\ \beta_0 + \beta_1(Runtime + 1) + \beta_2\ln(Budget) + \beta_3 Dir1 + \beta_4 Dir2 + \beta_5 Nom + \beta_6[(Runtime + 1)\times\ln(Budget)] + \beta_7[(Runtime + 1)\times Dir1] + \beta_8[(Runtime + 1)\times Dir2] + \beta_9[(Runtime + 1)\times Dir1\times Nom] + \beta_{10}[(Runtime + 1)\times Dir2\times Nom]\ ]$

$-\ [\beta_0 + \beta_1 Runtime + \beta_2\ln(Budget) + \beta_3 Dir1 + \beta_4 Dir2 + \beta_5 Nom + \beta_6[Runtime\times\ln(Budget)] + \beta_7[Runtime\times Dir1] + \beta_8[Runtime\times Dir2] + \beta_9[Runtime\times Dir1\times Nom] + \beta_{10}[Runtime\times Dir2\times Nom]\ ]$

$= \beta_1 + \beta_6\ln(Budget) + \beta_7 Dir1 + \beta_8 Dir2 + \beta_9 Dir1\times Nom + \beta_{10} Dir2\times Nom$

b) (4 marks) What is the effect of director credit on mean rating for each listed pair of levels below, after accounting for runtime, Academy Award nomination, and *ln*(budget)? (Hint: If you need more room, please direct me to where you did your work…perhaps the back of page 3?)

Level 1	Level 2	Effect of director credit on mean rating
Steven Spielberg	George Lucas	$(\beta_3 - \beta_4) + (\beta_7 - \beta_8)Runtime + (\beta_9 - \beta_{10})Runtime\times Nom$
Steven Spielberg	Other	$\beta_3 + \beta_7 Runtime + \beta_9 Runtime\times Nom$

c) (3 marks) Using the original model, state the null and alternative hypothesis to test whether Academy Award nomination has any effect on mean rating, after accounting for runtime, director credit, and *ln*(budget). What is the distribution of the test statistic under the null hypothesis?

H_0: $\beta_5 = \beta_9 = \beta_{10} = 0$ $\quad\quad$ H_A: at least one $\beta_i \neq 0$ $\quad i = 5, 9, 10$

$F_0 \sim F_{\text{\# of parameters in question}, n-(p+1)} = F_{3, 58-(10+1)} = F_{3, 47}$

Note: For parts d) – g), remove all interaction terms from the "original model".

d) (4 marks) Calculate a 90% <u>confidence interval</u> for the effect of changing budget from 50 to 100 million $US on mean rating.

$\ln(k) = \ln\left(\frac{100}{50}\right) = \ln(2) = 0.693$ $\quad\quad$ $t_{n-(p+1),\alpha/2} = t_{56,\,0.05} \approx t_{50,\,0.05} = 1.676$

$\hat{\beta}_1 \pm t_{n-(p+1)} \times S.E.(\hat{\beta}_1)$

0.068 ± (1.676)(0.104)

0.068 ± 0.174

(-0.106, 0.242)

→ (0.693(-0.106), 0.693(0.242)) → (-0.0737, 0.1680)

With 90% confidence, the effect of changing budget from 50 to 100 million $US on mean rating is between -0.0737 and 0.1680.

e) (5 marks) <u>Carry out a full analysis in detail</u> to determine if *ln*(budget) has a negative association with mean rating, after accounting for runtime. You must provide the *exact* p-value for this analysis.

H_0: $\beta_2 \geq 0$ $\quad\quad$ H_A: $\beta_2 < 0$

$$t_0 = \frac{\hat{\beta}_2}{S.E.(\hat{\beta}_2)} = \frac{-0.121}{0.111} = -1.090$$ (output gives result of -1.088)

$t_0 \sim t_{n-(p+1)} = t_{55}$ $\quad\quad$ p-value = 0.281/2 = 0.1405

→ p-value > 0.1 → Weak evidence against H_0. → Do not reject H_0.

→ Mean rating may not have a negative association with *ln*(budget), after accounting for runtime.

f) (4 marks) Calculate a 98% prediction interval for the rating of a film directed by Oliver Stone that has a runtime of 150 minutes, a budget of \$US 30 million, and had an award nomination.

$\hat{Y} = 5.116 - 0.153*\ln(30) + 0.012(150) + 0.646(1) + 1.089(0) + 0.942(0) = 7.042$

$\hat{Y} \pm t_{n-(p+1),\alpha/2} \times \hat{\sigma}$ $\qquad$ $t_{n-(p+1),\alpha/2} = t_{52,\,0.01} \approx t_{50,\,0.01} = 2.403$

$7.042 \pm (2.403)(0.677)$ $\qquad$ $\hat{\sigma} = \sqrt{0.458} = 0.677$

7.042 ± 1.626

$(5.415, 8.668)$

g) (8 marks) Carry out a full analysis in detail to determine if mean rating depends on award nomination or director credit, after accounting for *ln*(budget) and runtime. You must also clearly label the sum-of-squares residuals for the models under the null and alternative hypotheses.

H_0: $\beta_3 = \beta_4 = \beta_5 = 0$ $\qquad$ H_A: at least one $\beta_i \neq 0$ $\quad i = 3, 4, 5$

Reduced model (under null hypothesis):

$\mu(Rating \mid Runtime, ln(Budget), Dir, Nom) = \beta_0 + \beta_1 Runtime + \beta_2 ln(Budget)$

$SSR(r) = 40.499$ $\qquad$ $df(r) = 55$

Full model (under alternative hypothesis):

$\mu(Rating \mid \ldots) = \beta_0 + \beta_1 Runtime + \beta_2 ln(Budget) + \beta_3 Dir1 + \beta_4 Dir2 + \beta_5 Nom$

$SSR(f) = 23.827$ $\qquad$ $df(f) = 52$

$$F_0 = \frac{(SSR(r) - SSR(f))/(df(r) - df(f))}{SSR(f)/df(f)} = \frac{(40.499 - 23.827)/(55-52)}{23.827/52}$$

$$= \frac{16.672/3}{23.827/52} = 12.128$$

$F_0 \sim F_{52}^{3} \approx F_{50}^{3}$

$6.34 < F_0 = 12.128$

$0.001 >$ p-value $\qquad$ → $p\text{-value} = P\left(F_{df(f)}^{df(r)-df(f)} > F_0\right) = P\left(F_{50}^{3} > 12.128\right) \in (0, 0.001)$

→ (Strong to) convincing evidence against H_0. → Reject H_0.

→ Mean rating does depend on award nomination or director credit, after accounting for *ln*(budget) and runtime.

Case Study 2 – Airline food: SURELY it's good for you? (25 marks)

When needed, use additional output on page 13 to answer the following questions.

Dr. Rumack can't tell if people are getting sick via airline food. Randomly sampling months of the year, he observes how many passengers are getting sick for each airline, also differentiating by another factor: what type of food was ordered by at least 50% of the passengers. The table below summarizes the results.

Group	Airline	Food	n	Sample Mean	Sample S.D.
1	Oceanic	Steak	10	30.18	2.72
2	Oceanic	Fish	10	33.23	2.58
3	Qantas	Steak	10	30.44	4.47
4	Qantas	Fish	10	31.37	3.68
5	Northwest	Steak	10	27.54	3.53
6	Northwest	Lasagna	10	22.68	4.11
7	Trans American	Steak	10	33.44	3.74
8	Trans American	Fish	10	48.17	4.98

The following table is the ANOVA output.

ANOVA

NumPass

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3778.180	7	539.740	37.340	
Within Groups	1040.736	72	14.455		
Total	4818.916	79			

a) (4 marks) Is there any evidence of a difference in the average number of sick passengers among the eight different groups? State the sum-of-squares residuals for the models under the alternative hypothesis, the test statistic, and the distribution of the test statistic (you do not have to find the *p*-value or answer the question). Also, estimate the common standard deviation.

SSR for the model under H_A: 1040.736

Test Statistic: $F_0 = 37.340$

Distribution of test statistic: $F_{7,\,72}$

Common Standard Deviation: $\boxed{\sqrt{14.455} = 3.802}$

There is a treatment contrast that might be of interest for estimating the main effects of the two factors on the mean number of sick passengers. Let μ_1, μ_2, μ_3, μ_4, μ_5, μ_6, μ_7 and μ_8 correspond to the population mean responses for groups 1, 2, 3, 4, 5, 6, 7 and 8, respectively.

b) (6 marks) Does the average number of sick passengers decrease if the type of food ordered by at least 50% of the passengers is steak instead of fish?

i. (2 marks) First, define the treatment contrast (i.e. fill in the blanks with the appropriate contrast coefficients) that will define the effect described in the above question.

$$\gamma = \left(\frac{1}{4}\right)\mu_1 + \left(-\frac{1}{3}\right)\mu_2 + \left(\frac{1}{4}\right)\mu_3 + \left(-\frac{1}{3}\right)\mu_4$$

$$+ \left(\frac{1}{4}\right)\mu_5 + 0\mu_6 + \left(\frac{1}{4}\right)\mu_7 + \left(-\frac{1}{3}\right)\mu_8$$

ii. (2 marks) Next, define the null and alternative hypotheses for this contrast.

$H_0: \gamma \geq 0$ $H_A: \gamma < 0$

iii. (2 marks) Then, determine the test statistic and the *exact p*-value.

$t_0 = -7.827$

p-value $\approx 0.000/2 = 0.000$

Dr. Rumack decided to be safe and eat lasagna on a Northwest flight, but his lasagna was a bit fishy, so he got sick. Enlisting the help of research assistants Shirley and Leslie, the lasagna was determined to have large concentrations of fish, so group 6 was then re-categorized as Fish. NOTE: This re-categorization does NOT affect how you answered parts a) and b).

Nevertheless, the new data structure allowed another approach to test the effects of the two factors, as well as their interaction: Two-Way ANOVA with two factors.

Factor A – Airline (Northwest, Oceanic, Qantas, Trans American)
Factor B – Food (Fish, Steak)

c) (8 marks) Use the following incomplete Two-Way ANOVA table.

Tests of Between-Subjects Effects

Dependent Variable: NumPass

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3778.180[a]	**7**	**539.740**	**37.340**	.000
A	2525.281	**3**	841.760	**58.234**	.000
B	239.339	**1**	239.339	**16.558**	.000
A * B	1013.559	**3**	**337.853**	**23.373**	.000
Error	1040.736	**72**	**14.455**		
Corrected Total	4818.916	**79**			

a R Squared = .784 (Adjusted R Squared = .763)

i. (6 marks) Carry out a full analysis in detail to determine if Airline or Food has any effect on the mean number of sick passengers.

H_0: $\mu(Y \mid A, B) = \beta_0$
H_A: $\mu(Y \mid A, B) = \beta_0 + A + B + AB$

$F_0 = 539.740/14.455 = 37.340 \sim F_{7, 72}$

p-value ≈ 0.000

p-value < 0.01 → (Strong to) convincing evidence against H_0. → Reject H_0.
→ Either Airline, Food, or their interaction has an effect on the mean response.

ii. (2 marks) Calculate the sum-of-squares residuals for the following models:

1. Two-way ANOVA for Airline and the interaction of Airline and Food.
2. One-way ANOVA for only Food.

$$\boxed{1.\ SSR(Two-Way: \beta_0 + A + AF) = 1040.736 + 239.339 = 1280.075}$$

$$\boxed{2.\ SSR(One-Way: Food) = 1040.736 + 1013.559 + 2525.281 = 4579.576}$$

Delirious from the sickness, Dr. Rumack demanded the removal of the interaction term.

d) (7 marks) The Two-Way ANOVA table is given below (Additive Fit)

Tests of Between-Subjects Effects

Dependent Variable: NumPass

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2764.620[a]	4	**691.155**	**25.233**	.000
A	2525.281	3	**841.760**	**30.732**	.000
B	239.339	1	**239.339**	**8.738**	.004
Error	2054.296	75	27.391		
Corrected Total	4818.916	79			

a R Squared = .574 (Adjusted R Squared = .551)

i. (4 marks) Determine if Airline has an effect on the mean number of sick passengers, after accounting for Food. Only provide the null and alternative hypotheses, the test statistic, and the *p*-value.

H_0: $\mu(Y \mid A, B) = \beta_0 + B$
H_A: $\mu(Y \mid A, B) = \beta_0 + A + B$

$F_0 = 841.760/27.391 = 30.732$

p-value ≈ 0.000

ii. (3 marks) Determine if Food has an effect on the mean number of sick passengers, after accounting for Airline. Only provide the test statistic, its distribution, and the *p*-value.

$F_0 = 239.339/27.391 = 8.738 \sim F_{1,\,75}$

p-value $= 0.004$

Yet another approach to test the effects of the two factors is to model the data as a multiple linear regression model using indicator variables:

Let Airline be represented by three indicator variables (Q, NW, and TA) that will indicate Qantas, Northwest, and Trans American, respectively; Oceanic is the "default". Let Food be represented by one indicator variable (*Steak*); Fish is the "default".

The corresponding regression model is:

$$\mu(NumPass \mid Airline, Food) = \beta_0 + \beta_1 Q + \beta_2 NW + \beta_3 TA + \beta_4 Steak + \beta_5 Q \times Steak + \beta_6 NW \times Steak + \beta_7 TA \times Steak$$

e) (4 marks) In terms of the coefficients, what is the effect of Airline on the mean number of sick passengers, after accounting for Food for each pair of levels? Show your work.

Fill in the table:

Level 1	Level 2	Effect of Airline on mean response
Qantas	Trans American	$\beta_1 - \beta_3 + (\beta_5 - \beta_7)Steak$
Oceanic	Northwest	$-\beta_2 - \beta_6 Steak$

$\mu(NumPass \mid Airline = \text{Q}, Food) - \mu(NumPass \mid Airline = \text{TA}, Food)$

$= [\beta_0 + \beta_1(1) + \beta_2(0) + \beta_3(0) + \beta_4 Steak + \beta_5(1)\times Steak + \beta_6(0)\times Steak + \beta_7(0)\times Steak]$

$- [\beta_0 + \beta_1(0) + \beta_2(0) + \beta_3(1) + \beta_4 Steak + \beta_5(0)\times Steak + \beta_6(0)\times Steak + \beta_7(1)\times Steak]$

$= \beta_1 - \beta_3 + (\beta_5 - \beta_7)Steak$ (Similar calculations give the second pair of levels.)

f) (4 marks): Determine if the average number of sick passengers for flights from the *airline* in the first column, where the type of food ordered by at least 50% of the passengers is the *food* in the first column, is **lower** than the average number of sick passengers for flights from the *airline* in the second column, where the type of food ordered by at least 50% of the passengers is the *food* in the second column. Only provide the appropriate test statistic and *exact* *p*-value.

		Test statistic	*p*-value
Qantas & fish	Oceanic & fish	-1.092	.278/2 = 0.139
Oceanic & steak	Oceanic & fish	-1.794	.077/2 = 0.0385

g) (2 marks): Rewrite the regression model above with Trans American and Fish as the respective default levels.

If 'O' = 'Oceanic', then one possible solution is

$$\mu(NumPass \mid Airline, Food) = \beta_0 + \beta_1 Q + \beta_2 NW + \beta_3 O + \beta_4 Steak + \beta_5 Q \times Steak + \beta_6 NW \times Steak + \beta_7 O \times Steak$$