Statistics 252 – Final Exam – Version A

Instructor: Paul Cartledge

Instructions:

1. *Read <u>all</u> 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 70 marks and has 14 pages (in two parts).*
6. *Show your work for the long answer section 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. *When referring to "log", I am always referring to the natural log.*
10. *Unless instructed otherwise, give a range for the p-value. Also, use the "judgment approach" to help state your conclusion in plain English.*

Name: _____

Signature: _____

Component	Notes	Worth	Mark
Short Answer	3 questions	6	
What test?	2 questions (2 marks each)	4	
Case Study 1	7 parts	30	
Case Study 2	7 parts	30	
Total		70	

ID:_____

Short Answer Problems (6 marks)

Question 1 (2 marks) In testing for a difference in the median responses for two treatments after back-transforming the data, you set up the following set of hypotheses on the original scale. 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: $Median(Y_1) / Median(Y_2) > 0$
H_A: $Median(Y_1) / Median(Y_2) \leq 0$

$$H_0: \frac{Median(Y_1)}{Median(Y_2)} \leq 1 \qquad H_A: \frac{Median(Y_1)}{Median(Y_2)} > 1$$

Question 2 (2 marks) Suppose Optimus Prime determines a simple linear relationship between the cost of his monthly power (in thousands of dollars) and how many kilolitres (kL) of oil the Autobots consume at each of his parties. Checking assumptions, both variables required log-transformations. The estimated regression line is $\hat{\mu}(\ln(cost) \mid \ln(oil)) = 2.908 + 0.458\ln(oil)$. Estimate the change in mean cost of monthly power associated with a change of 15 kL to 75 kL in oil consumption at Optimus's Autobot parties. Give a statement relating this change to the given variables.

$$k^{\hat{\beta}_1} = \left(\frac{75}{15}\right)^{0.458} = 5^{0.458} = 2.090$$

"A multiplicative change from 15 to 75 kL is associated with a multiplicative change of 2.090 in median cost of monthly power."

Question 3 (2 marks) Consider a dataset with 7 groups and 3 models: One-Mean, *J*-Mean, and Seven-Mean. If 10 observations are collected for each group, fill in the shaded areas and find *J*.

ANOVA Table for testing the One-Mean vs. the Seven-Mean model

Source of Variation	Sum of Squares	df	Mean Square	F-Statistic	*p*-value
Between (Extra)	**414**	**6**	**69**	11.5	
Within (Full)	**378**	**63**	6		
Total (Reduced)	792	**69**			

ANOVA Table for testing the One-Mean vs. the *J*-Mean model → ***J* = 6**

Source of Variation	Sum of Squares	df	Mean Square	F-Statistic	*p*-value
Between (Extra)	280	**5**	**56**	**7**	
Within (Full)	**512**	**64**	8		
Total (Reduced)	**792**	**69**			

What test would you use? (4 marks)

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 (2 marks) An organization (Protecting Animals Caringly & Kindly) wants to see if wild dogs have a longer lifespan than those held in captivity. In a time-consuming study, the leaders of P.A.C.K. record the lifespan of 15 randomly selected dogs (10 wild and 5 held in captivity) from each of six different African countries. What test would you use to see if the lifespan of wild and captivity-based dogs are different within each country? What is the distribution of the test statistic under the null hypothesis?

vi) Some Extra-Sum-of-Squares F-test comparing two models for the *I* population means

$F_{df(r) - df(f),\ df(f)} = F_{(6*15 - 12) - (6*15 - 6),\ 6*15 - 12} = F_{6,\ 78}$

Question 5 (2 marks) Guillermo is attempting to market a new kind of special super-spicy salsa. He's not sure where to market his product to maximize his profit, so he observes prices from two random samples (each with 42 observations) from Canada and Mexico, adjusting the latter values to match Canadian currency. What test will denote that the average price in Canada is greater than that in Mexico? What is the distribution of the test statistic under the null hypothesis?

Two Independent Sample *t*-test [OR v) or viii)]

$t_{n1 + n2 - 2} = t_{42 + 42 - 2} = t_{82}$

Case Study 1 – Raiders of the Lost Statistical Analysis (25 marks)

When needed, use the output on pages 12 and 13 to answer the following questions.

Earlier in the term, we (namely me) made fun of a certain director's talent. This time, however, we'll be nicer and completely ignore him. Let's suppose a study was done to investigate the association of box office gross in the United States with the year the film came out, the film's rating at IMDB.com, and the appearance of certain actors in the films under study. The study consists of 21 random and independent films directed by Steven Spielberg. "Actor appearance" is categorized into 3 levels: Harrison Ford, Tom Hanks, and Other. Box office gross (*BoxOffice*, measured in millions of US$), the year the film came out (*Year*), and the film's rating at IMDB.com (*Rating*, measured on a scale of 0 to 10, with 10 being the highest) are modeled as continuous (numerical) variables.

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

$$\mu(BoxOffice \mid Year, Rating, Actor) = \beta_0 + \beta_1 Year + \beta_2 Rating + \beta_3 Ford + \beta_4 Hanks + \beta_5 Year \times Rating + \beta_6 Year \times Ford + \beta_7 Year \times Hanks + \beta_8 Rating \times Ford + \beta_9 Rating \times Hanks + \beta_{10} Year \times Rating \times Ford + \beta_{11} Year \times Rating \times Hanks$$

a) (3 marks) What is the effect of the film's rating on mean box office gross, after accounting for year and actor appearance?

$\mu(BoxOffice \mid Year, Rating + 1, Actor) - \mu(BoxOffice \mid Year, Rating, Actor)$

$= [\ \beta_0 + \beta_1 Year + \beta_2(Rating + 1) + \beta_3 Ford + \beta_4 Hanks + \beta_5[Year \times (Rating + 1)] + \beta_6[Year \times Ford] + \beta_7[Year \times Hanks] + \beta_8[(Rating + 1) \times Ford] + \beta_9[(Rating + 1) \times Hanks] + \beta_{10}[Year \times (Rating + 1) \times Ford] + \beta_{11}[Year \times (Rating + 1) \times Hanks]\ ]$

$- [\beta_0 + \beta_1 Year + \beta_2 Rating + \beta_3 Ford + \beta_4 Hanks + \beta_5[Year \times Rating] + \beta_6[Year \times Ford] + \beta_7[Year \times Hanks] + \beta_8[Rating \times Ford] + \beta_9[Rating \times Hanks] + \beta_{10}[Year \times Rating \times Ford] + \beta_{11}[Year \times Rating \times Hanks]\ ]$

$= \beta_2 + \beta_5 Year + \beta_8 Ford + \beta_9 Hanks + \beta_{10} Year \times Ford + \beta_{11} Year \times Hanks$

b) (4 marks) What is the effect of actor appearance on mean box office gross, after accounting for year and rating, for each listed pair of levels below? (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 actor appearance on mean box office gross
Ford	Hanks	$(\beta_3 - \beta_4) + (\beta_6 - \beta_7)Year + (\beta_8 - \beta_9)Rating + (\beta_{10} - \beta_{11})Year \times Rating$
Ford	Other	$\beta_3 + \beta_6 Year + \beta_8 Rating + \beta_{10} Year \times Rating$
Hanks	Other	$\beta_4 + \beta_7 Year + \beta_9 Rating + \beta_{11} Year \times Rating$

c) (3 marks) Using the original model, state the null and alternative hypothesis to test whether the year a film comes out depends on actor appearance, after accounting for the film's rating. What is the distribution of the test statistic under the null hypothesis?

H_0: $\beta_6 = \beta_7 = \beta_{10} = \beta_{11} = 0$ H_A: at least one $\beta_i \neq 0$ $i = 6, 7, 10, 11$

$F_0 \sim F_{\#\text{ of parameters in question},\ n-(p+1)} = F_{4,\ 21-(11+1)} = F_{4,\ 9}$

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

d) (3 marks) Calculate a 95% confidence interval for the mean difference in box office gross between films starring Harrison Ford and Other films.

$\hat{\beta}_2 \pm t_{n-(p+1)} S.E.(\hat{\beta}_2)$ $t_{n-(p+1),\alpha/2} = t_{18,\ 0.025} = 2.101$

80.107 ± (2.101)(73.320)

80.107 ± 154.045 → (-73.938, 234.152)

e) (4 marks) Calculate a 90% prediction interval for the box office gross of a film starring Tom Hanks that came out in 2002 and has a rating of 7.7.

$\hat{Y} = 4716.707 + 63.805(7.7) - 2.544(2002) + 18.460(0) - 12.339(1) = 102.579$

$\hat{Y} \pm t_{n-(p+1)} \hat{\sigma}$ $t_{n-(p+1),\alpha/2} = t_{16,\ 0.05} = 1.746$

102.579 ± (1.746)(107.007) $\hat{\sigma} = \sqrt{11450.39} = 107.007$

102.579 ± 186.833

(-89.255, 289.412) (Yes, it strangely appears that a number of people may have asked for refunds to *Catch Me If You Can*.)

f) (5 marks) Carry out a test to determine if there is significant evidence that the mean box office gross has a negative association with the year the film comes out, after accounting for the film's rating. State the null and alternative hypothesis in terms of the regression coefficients, the test statistic and all of its components (see its formula), the distribution of the test statistic under the null hypothesis, and the *exact* p-value of the test. Conclude in plain English.

H_0: $\beta_1 \geq 0$ H_A: $\beta_1 < 0$

$$t_0 = \frac{\hat{\beta}_1}{S.E.(\hat{\beta}_1)} = \frac{-2.918}{2.294} = -1.272$$ (output gives same exact result of -1.272)

$t_0 \sim t_{n-(p+1)} = t_{18}$ p-value = 0.220/2 = 0.110

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

→ Mean box office gross may not have a negative association with the year the film comes out, after accounting for the film's rating.

g) (8 marks) Carry out a test to determine if there is significant evidence that mean box office gross depends on the film's rating and the year it came out, after accounting for actor appearance. State the null and alternative hypothesis in terms of the regression coefficients, the sum-of-squares residuals for the models under the null and alternative hypotheses, and the distribution of the test statistic under the null hypothesis. Calculate the test statistic and the *p*-value of the test. Conclude in plain English.

H_0: $\beta_1 = \beta_2 = 0$ H_A: at least one $\beta_i \neq 0$ $i = 1, 2$

Reduced model (under null hypothesis):

$$\mu(BoxOffice \mid Year, Rating, Actor) = \beta_0 + \beta_3 Ford + \beta_4 Hanks$$

$SSR(r) = 241\,914.031$
$df(r) = 18$

Full model (under alternative hypothesis):

$$\mu(BoxOffice \mid Year, Rating, Actor) = \beta_0 + \beta_1 Year + \beta_2 Rating + \beta_3 Ford + \beta_4 Hanks$$

$SSR(f) = 183\,206.278$
$df(f) = 16$

$$F_0 = \frac{(SSR(r) - SSR(f)) / (df(r) - df(f))}{SSR(f) / df(f)}$$
$$= \frac{(241914.031 - 183206.278) / (18 - 16)}{183206.278 / 16}$$
$$= \frac{58707.753 / 2}{183206.278 / 16}$$
$$= 2.564$$

$F_0 \sim F_{16}^2$

$1.51 < F_0 = 2.564 < 2.67$

$0.25 > p\text{-value} > 0.1$ → $p\text{-value} = P\left(F_{df(f)}^{df(r)-df(f)} > F_0\right) = P\left(F_{16}^2 > 2.564\right) \in (0.1, 0.25)$

→ weak evidence against H_0 → Do not reject H_0.

→ Mean box office gross may not depend on film rating or year, after accounting for actor appearance.

Case Study 2 – Feel the Rhythm, Feel the Rhyme, Time to Analyze the Bobsled Time! (30 marks)

When needed, use additional output on page 14 to answer the following questions.

With Vancouver 2010 just over two months away, a scientist (Hercules) decides to observe the bobsled times at the 2006 Torino Olympics while simultaneously identifying the measurements by two factors: whether they belong to one of three country pairings (Canada/U.S., Italy/Russia, or Switzerland/Germany) as well as the specific sport (Two Woman, Two Man, or Four Man). Recording the results from 4 random and independent subjects per combination, the table below summarizes their "finishing times" (the time it takes to get to the finish line) in seconds.

Group	Country Pair	Sport	n	Sample Mean	Sample S.D.
1	Canada/U.S.	Two Woman	4	231.84	1.40
2	Italy/Russia	Two Woman	4	232.01	0.80
3	Swiss/German	Two Woman	4	231.55	1.22
4	Canada/U.S.	Two Man	4	225.05	1.15
5	Italy/Russia	Two Man	4	225.40	1.07
6	Swiss/German	Two Man	4	224.18	0.50
7	Canada/U.S.	Four Man	4	221.81	1.16
8	Italy/Russia	Four Man	4	221.98	1.03
9	Swiss/German	Four Man	4	221.02	0.58

The following table is the ANOVA output.

ANOVA

Time

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	656.125	**8**	**82.016**	**77.306**	
Within Groups	28.645	**27**	**1.061**		
Total	684.770	**35**			

a) (4 marks) Is there any significant evidence of a difference in average finishing times among the nine different groups? State the sum-of-squares residuals for the model under the alternative hypothesis, the test statistic, the distribution of the test statistic under the null hypothesis, and the range of the *p*-value (you do not have to answer the question).

SSR for the model under H_A: 28.645

Test statistic: 77.306

Distribution: $F_{8, 27}$

p-value range: (0, 0.001)

There is a treatment contrast that might be of interest in the experiment for estimating the main effects of the two factors on mean heart rate. Let μ_1, μ_2, μ_3, μ_4, μ_5, μ_6, μ_7, μ_8, and μ_9 correspond to the population mean responses for groups 1, 2, 3, 4, 5, 6, 7, 8, and 9, respectively.

b) (5 marks) Does the pairing of Canada/U.S. take more time on average at all bobsled sports compared to the pairing of Switzerland/Germany?

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

$$\gamma = \left(\frac{1}{3}\right)\mu_1 + 0\mu_2 + \left(-\frac{1}{3}\right)\mu_3 + \left(\frac{1}{3}\right)\mu_4 + 0\mu_5$$

$$+ \left(-\frac{1}{3}\right)\mu_6 + \left(\frac{1}{3}\right)\mu_7 + 0\mu_8 + \left(-\frac{1}{3}\right)\mu_9$$

ii. (2 marks) Determine the test statistic and the *exact* *p*-value.

$t_0 = 1.540$

p-value = 0.135/2 = 0.0675

iii. (1 mark) Make a decision using the *p*-value and answer the question at the top of this page.

There is moderate to suggestive evidence against H_0. Thus, do not reject H_0.

The Canada/U.S. pairing may not take more time on average at all bobsled sports compared to the pairing of Switzerland/Germany.

Another approach to test the effects of the two factors, as well as their interaction, is to model the data as a Two-Way ANOVA with the two factors:

Factor A – Sport (Two Woman, Two Man, Four Man)
Factor B – Country Pair (Canada/U.S., Italy/Russia, Switzerland/Germany)

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

Tests of Between-Subjects Effects

Dependent Variable: Time

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	656.125[a]	8	82.016	**77.306**	.000
Intercept	1840264.077	1	1840264.077	1734583.002	.000
A	650.436	**2**	325.218	**306.542**	.000
B	5.006	**2**	2.503	**2.359**	.114
A * B	.684	**4**	.171	**.161**	.956
Error	28.645	**27**	**1.061**		
Total	1840948.848	36			
Corrected Total	684.770	35			

a R Squared = .958 (Adjusted R Squared = .946)

i. (5 marks) Is there any significant evidence that Sport depends on the presence of Country Pair? State the null and alternative hypothesis, the test statistic, the distribution of the test statistic under the null hypothesis, and the *exact* p-value of the test. Conclude in plain English.

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

$F_0 = 0.171/1.061 = 0.161 \sim F_{4, 27}$

p-value = 0.956

p-value > 0.1 → (extremely) weak evidence against H_0 → do not reject H_0
→ Sport may not depend on the presence of Country Pair.

ii. (3 marks) Does it appear that either Sport or Country Pair have any effect on mean time? Simply refer to the appropriate test statistic, the distribution of the test statistic and p-value in the table above.

$F_0 = 82.016/1.061 = 77.306$

$F_0 \sim F_{8, 27}$

p-value ≈ 0.000

Note: With a p-value of approximately 0.956, we will remove the interaction term and fit the additive model.

d) (5 marks) The Two-Way ANOVA table is given below (Additive Fit)

Tests of Between-Subjects Effects

Dependent Variable: Time

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	655.442[a]	4	163.860	173.198	.000
Intercept	1840264.077	1	1840264.077	1945126.620	.000
A	650.436	2	325.218	343.750	.000
B	5.006	2	2.503	2.645	.087
Error	29.329	31	.946		
Total	1840948.848	36			
Corrected Total	684.770	35			

a R Squared = .957 (Adjusted R Squared = .952)

i. (3 marks) What are the sum-of-squares residuals for the full and reduced models for one-way ANOVA for only Sport?

$$SSR(f) = SSR(\text{3-mean: only Sport}) = SSE + SSB = 29.329 + 5.006 = 34.335$$

$$SSR(r) = SSR(\text{One-Mean}) = 684.770$$

ii. (2 marks) Is there any significant evidence that Country Pair has an effect on mean time, after accounting for Sport? Simply refer to the appropriate test statistic and *p*-value in the table above.

$F_0 = 2.503/0.946 = 2.645$

p-value = 0.087

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 Country Pair be represented by two indicator variables (*Pair1* and *Pair2*) that will indicate the Canada/U.S. and Switzerland/Germany pairs, respectively; the Italy/Russia pair is the "default". Let Sport be represented by two indicator variables (*W2* and *M2*) that will indicate the Two Woman and Two Man sports, respectively; the Four Man sport is the "default".

The corresponding regression model is:

$$\mu(Time \mid Country\ Pair, Sport) = \beta_0 + \beta_1 Pair1 + \beta_2 Pair2 + \beta_3 W2 + \beta_4 M2 + \beta_5 W2 \times Pair1 + \beta_6 W2 \times Pair2 + \beta_7 M2 \times Pair1 + \beta_8 M2 \times Pair2$$

e) (4 marks) In terms of the coefficients, what is the effect of Sport on mean time for each pair of sport levels?

Fill in the chart:

Level 1	Level 2	Effect of sport on mean time	Estimate
Two Woman	Two Man	$\beta_3 - \beta_4 + (\beta_5 - \beta_7)Pair1 + (\beta_6 - \beta_8)Pair2$	6.790
Two Woman	Four Man	$\beta_3 + \beta_5 Pair1 + \beta_6 Pair2$	10.028
Two Man	Four Man	$\beta_4 + \beta_7 Pair1 + \beta_8 Pair2$	3.238

f) (2 marks): Estimate each defined effect in the rightmost column for the **Canada/U.S.** pair. Show your work below. No marks will be given if no work is shown.

$\hat{\beta}_3 - \hat{\beta}_4 + \left(\hat{\beta}_5 - \hat{\beta}_7\right)Pair1 + \left(\hat{\beta}_6 - \hat{\beta}_8\right)Pair2$

= 10.033 – 3.418 + (-0.005 – (-0.180))(1) + (0.495 – (-0.260))(0) = 6.790

$\hat{\beta}_3 + \hat{\beta}_5 Pair1 + \hat{\beta}_6 Pair2$ = 10.033 + (-0.005)(1) + (0.495)(0) = 10.028

$\hat{\beta}_4 + \hat{\beta}_7 Pair1 + \hat{\beta}_8 Pair2$ = 3.418 + (-0.180)(1) + (-0.260)(0) = 3.238

g) (2 marks): Consider a different MLR model that uses factors such as gender and how many people are in the bobsled.

$$\mu(Time \mid Gender, How\ Many) = \beta_0 + \beta_1 Male + \beta_2 Two + \beta_3 Male \times Two$$

Is there something wrong with this model? Why or why not?

If both indicators equal zero, this implies a group consisting of four women on a bobsled team, but such a group does not exist in the original data on page 7. Thus, this model is wrong because it refers to a default group that does not exist in the provided data structure.