STAT 302 - Analysis of Experimental and Observational Data
Department of Statistics and Actuarial Science
Simon Fraser University
Midterm Exam - Summer 2023

Time: 2 hours

First name (please write as legibly as possible within the boxes)

S	O	L	U	T	I	O	N	S								

Last name

Student ID number

---

1. **Answer All Questions**
2. **The midterm is 120 minutes in length**
3. **The only aids that you can bring into the midterm are writing utensils, a calculator, and the single side A4 aid sheet**
4. **Provide the answers in the space given. Solutions must be legible and well-presented**

---

1. **(15 points)** A simple linear regression model was fit to data concerned with eruptions of Old Faithful. The response was the duration of the eruption (in minutes) and the explanatory variable was the waiting time since the last eruption (in minutes). R output appears below :

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.29917 -0.37689  0.03508  0.34909  1.19329

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.874016   0.160143  -11.70   <2e-16
waiting      0.075628              34.09

Residual standard error: 0.4965 on 270 degrees of freedom
Multiple R-Squared: 0.8115,     Adjusted R-squared: 0.8108
F-statistic:       on   and 270 degrees of freedom,     p-value:     0
```

(a) **(1 point)** How many total data points in the data set?

272

(b) **(2 points)** What is the least square regression line for predicting the duration of the eruption (in minutes) from waiting time since the last eruption (in minutes) ?

$\hat{y} = -1.874 + 0.076\,x$

(c) **(1 point)** Interpret the slope estimator in the context of this setting.

When waiting time since the last eruption increases by one minute, the average duration of the eruption increases by 0.076 minutes.

(d) **(1 point)** Based on the output what is the size of the typical error when predicting the duration of the eruption (in minutes) from waiting time since the last eruption (in minutes) ?

0.4965

(e) **(1 point)** What is the standard error of the slope estimator?

$$\frac{0.076}{34.09} = 0.002$$

(f) **(2 points)** Test the hypothesis that $\beta_1 = 0$ at 5% significance level. (Use the t-table to calculate the approximate p-value).

$t = 34.09$ $DF = 270$ p-value < 0.001 (1)

$H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$ (1)

$\therefore$ We reject the H_0 at $\alpha = 0.05$.

(g) **(2 points)** Construct a 95% confidence interval for β_1 (Use the t-table to find the approximate t* critical value) .

(1)

$0.076 \pm 1.984 \cdot 0.002$

0.076 ± 0.004

$(0.072, 0.08)$ (1)

(h) **(2 points)** What is the F-statistic value and the numerator degrees of freedom ?

$F = (34.09)^2 = 1162.12$ (1)

Numerator $DF = 1$ (1)

(i) **(1 point)** What percent of the variation is explained by the simple linear regression model with waiting time since the last eruption as a predictor and duration of the eruption as the response ?

81.15 %

(j) **(2 points)** What is SSE in the model ? (Hint : Use Residual Standard Error)

$$\sqrt{\frac{SSE}{270}} = 0.4965 \qquad SSE = 66.55$$

(1) (1)

2. **(8 points - 1 point for each)** Match the statements below with the corresponding terms from this list :

A. Multicollinearity B. Extrapolation C. Residual plots D. Indicator Variables
E. R^2 F. Residual G. Outlier H. Quadratic regression

(a) Used to check the assumptions of the regression model : C

(b) Proportion of the variability in y explained by the regression model : E

(c) Used when a numerical predictor has a curvilinear relationship with the response : H

(d) Predicting outside the observed range of x's : B

(e) Used in a regression model to represent categorical variables : D

(f) Is the observed value of y minus the predicted value of y for the observed x : F

(g) One or more of the predictors is strongly correlated with some combination of the other predictors in the set : A

(h) A point that lies far from the rest : G

3. **(16 points)** A marketing company wants to build a model for estimating sales (in thousand units) based on the advertising budget (in thousand dollars) invested in youtube and facebook. Refer the below R output and answer the questions.

```
Call:                y          x1          x2
lm(formula = sales ~ youtube + facebook, data = marketing)

Residuals:
    Min      1Q  Median      3Q     Max
-10.557  -1.050   0.291   1.405   3.399

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.50532    0.35339    9.92   <2e-16 ***
youtube      0.04575    0.00139   32.91   <2e-16 ***
facebook     0.18799    0.00804   23.38   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.02 on 197 degrees of freedom
Multiple R-squared:  0.897,  Adjusted R-squared:  0.896
F-statistic:  860 on 2 and 197 DF,  p-value: <2e-16
```

(a) **(1 point)** What is the least square regression equation ?

$\hat{y} = 3.505 + 0.046 x_1 + 0.188 x_2$

(b) **(2 points)** Conduct the ANOVA F-Test to check the effectiveness of the multiple linear regression model ? (Write the correct hypotheses, Test statistic value, P-value and the conclusion)

$H_0: \beta_1 = \beta_2 = 0$ Vs $H_a:$ At least one $\beta_i \neq 0$ (1)

$F = 860$

P-value ~ 0

$\therefore$ We reject H_0 at $\alpha = 0.05$. $\therefore$ at least one of these x's are important. (1)

(c) **(4 points)** Conduct separate hypothesis testing for Youtube and Facebook coefficients in the model (Write the correct hypotheses, Test statistic value, P-value and the conclusion)

Youtube - x_1

$H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

$t = 32.91$ (1)

p-value ~ 0

$\therefore$ We reject H_0 at $\alpha = 0.05$.

(1)

Facebook - x_2

$H_0: \beta_2 = 0$ vs $H_a: \beta_2 \neq 0$

$t = 23.38$ (1)

p-value ~ 0

$\therefore$ We reject H_0 at $\alpha = 0.05$.

(1)

(d) **(2 points)** In the context of this setting interpret the coefficient estimates of Youtube and Facebook.

β_1 : As the Youtube advertising budget increases by thousand dollars, after accounting for the facebook advertising budget, the average number of sales increase by 45.75 units. (1)

β_2 : As the Facebook advertising budget increases by thousand dollars, after accounting for the Youtube advertising budget, the average number of sales increase by 187.99 units. (1)

(e) **(4 points)** Calculate the 95% confidence intervals for Youtube and Facebook coefficients. (Use the t-table to find the approximate t* critical value)

Youtube : $0.046 \pm 1.984 \cdot 0.001$ (1)

0.046 ± 0.001

$(0.045, 0.047)$ (1)

Facebook : $0.188 \pm 1.984 \cdot 0.008$ (1)

0.188 ± 0.016

$(0.172, 0.204)$ (1)

(f) **(1 point)** How much variability in sales can be explained by the model ?

89.7 %

(g) **(2 points)** Find SSE and SSTotal in the model.

$\sqrt{\frac{SSE}{197}} = 2.02$

$SSE = 803.84.$ (1)

$1 - \frac{SSE}{SST} = 0.897$

$\frac{SSE}{SST} = 0.103$

$SST = 7804.27$ (1)

4. **(11 points)** Children tend to get bigger as they get older, but we might be interested in how growth rates compare. Refer the below R output and answer the questions.

```
Call:
lm(formula = Weight ~ Age + Sex + Sex * Age, data = Kids198)

Residuals:
    Min      1Q  Median      3Q     Max
-46.884 -12.055  -2.782  10.185  58.581

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -33.69254   10.00727  -3.367 0.000917 ***
Age           0.90871    0.06106  14.882  < 2e-16 ***
Sex          31.85057   13.24269   2.405 0.017106 *
Age:Sex      -0.28122    0.08164  -3.445 0.000700 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\hat{\beta}_0$ $\hat{\beta}_1$ $\hat{\beta}_2$ $\hat{\beta}_3$

```
Residual standard error: 19.19 on 194 degrees of freedom
Multiple R-squared:  0.6683, Adjusted R-squared:  0.6631
F-statistic: 130.3 on 3 and 194 DF,  p-value: < 2.2e-16
```

$y = \beta_0 + \beta_1 \cdot Age + \beta_2 \cdot Sex + \beta_3 \cdot Age \cdot Sex + \varepsilon$

(a) **(1 point)** What is the full regression model ?

$\widehat{Weight} = -33.693 + 0.909\ Age + 31.851\ Sex - 0.281\ Age \cdot Sex$

(b) **(2 points)** If Sex = 0 means Boys and Sex = 1 means Girls, write the two separate regression models for Boys and Girls.

Boys : $\widehat{Weight} = -33.693 + 0.909 \cdot Age$ (1)

Girls : $\widehat{Weight} = -1.842 + 0.628\ Age$ (1)

(c) **(2 points)** Do a hypothesis test to compare the intercepts of the models. What is your conclusion at 5% significance level ?

$H_0: \beta_2 = 0$ Vs $H_a: \beta_2 \neq 0$ (1)

$t = 2.405$

P-value = 0.017

∴ We reject H_0 at $\alpha = 0.05$, ∴ intercepts are different. (1)

(d) **(2 points)** Do a hypothesis test to compare the slopes of the models. What is your conclusion at 5% significance level ?

$H_0: \beta_3 = 0$ Vs $H_a: \beta_3 \neq 0$ (1)

$t = -3.445$

P-value = 0.0007

∴ We reject H_0 at $\alpha = 0.05$. ∴ slopes are different. (1)

(e) **(2 points)** Calculate the 95% confidence interval for the difference between the intercepts(Use the t-table to find the approximate t* critical value).

$\hat{\beta}_2 \pm t_{194} \cdot S.E(\hat{\beta}_2)$ (1)

$31.851 \pm 1.984 \cdot 13.243$

31.851 ± 26.274

$(5.577, 58.125)$ (1)

(f) **(2 points)** Calculate the 95% confidence interval for the difference between the slopes (Use the t-table to find the approximate t* critical value).

$\hat{\beta}_3 \pm t_{194} \cdot S.E(\hat{\beta}_3)$ (1)

$-0.281 \pm 1.984 \cdot 0.082$

-0.281 ± 0.163

$(-0.444, -0.118)$ (1)