STA 302/1001
Summer 2016
Midterm A
5/30/2016
Time Limit: 1h 40 min

Last Name (Print): ____________________

First Name: ____________________

Student Number: ____________________

Check one: STA302 ☐ STA1001 ☐

This exam contains 8 pages (including this cover page) and 3 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.
- SLR stands for Simple Linear Regression; MLE for Maximum Likelihood Estimation; OLS for Ordinary Least Squares
- You may use a scientific calculator, the formulae below, and the t-table on the last page (round DF down).
- Show your work on each problem on this exam, and 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.

Problem	Points	Score
1	10	
2	10	
3	30	
Total:	50	

Some formulae:

$$b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} = \frac{\Sigma X_i Y_i - n\bar{X}\bar{Y}}{\Sigma X_i^2 - n\bar{X}^2} \qquad b_0 = \bar{Y} - b_1\bar{X}$$

$$Var(b_1) = \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2} \qquad Var(b_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}\right)$$

$$Var(\hat{Y}_h) = \sigma^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right) \qquad \sigma^2\{pred\} = Var(Y_h - \hat{Y}_h) = \sigma^2\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right)$$

$$SSTO = \Sigma(Y_i - \bar{Y})^2 \qquad SSE = \Sigma(Y_i - \hat{Y}_i)^2 \qquad SSR = \Sigma(\hat{Y}_i - \bar{Y})^2 = b_1^2\Sigma(X_i - \bar{X})^2$$

$$r = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2\Sigma(Y_i - \bar{Y})^2}} \qquad Cov(b_0, b_1) = -\frac{\sigma^2\bar{X}}{\Sigma(X_i - \bar{X})^2}$$

1. (10 points) **Multiple Choice** Answer the following questions by circling the *best* answer.

   I. Which of the following is a Gauss-Markov assumption for regression errors?

      A. They are Normally distributed
      B. They sum to zero
      C. They must come from a large sample
      D. Their variance is not related to a predictor variable

   II. The p-value is:

      A. The probability of the null hypothesis, given the data
      B. The probability of the data, given the null hypothesis
      C. The probability of the alternative hypothesis, given the data
      D. The probability of the data, given the alternative hypothesis

   III. Which of the following statements is false?

      A. The OLS method yields the same slope and intercept estimates as MLE
      B. OLS estimates for SLR are unbiased BLUE
      C. There are no estimators with lower variance than the OLS estimators
      D. OLS estimators are considered linear estimators

      false, nonlinear estimator with lower variance
      BLUE best linear unbiased estimator

   IV. In R, the command `order(c(1,5,3,2,4))` will return:

      A. `[1] 1 2 3 4 5`
      B. `[1] 5 4 3 2 1`
      C. `[1] 1 4 3 5 2`
      D. `[1] 2 5 3 4 1`

   V. Which of the following lines of R code will cause an error?

      A. `"fac" + "tor"`
      B. `as.numeric("4") - 3`
      C. `c(1,2) + 4`
      D. `c(factor("fac"), "tor")`

Answer the following True or False questions by writing 'T' or 'F' in the blank
Do not write something ambiguous like $\mp$ or $\Im$!

T In R, factors are stored as numbers

T The line $Y = \beta_0 + \beta_1 X$ describes the functional relationship between X and Y if they are linearly related.

F Confidence intervals can be wider than prediction intervals in some circumstances

T The ANOVA F-test is equivalent to the regression slope t-test for SLR

F When the sample size grows to infinity, confidence and prediction intervals will shrink to zero

2. Consider the SLR model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ with fixed (non-random) X_i

   (a) (6 points) Derive the MLEs for the SLR parameters β_0 and β_1. Show all work.

   Solution: In notes.

   (b) (4 points) Suppose instead of the OLS slope estimator we use $b_1 = \frac{\Sigma X_i Y_i}{\Sigma (X_i - \bar{X})^2}$. Is this estimator unbiased? If not, are there any conditions under which it is unbiased?

   Solution:
   $E[b_1] = E[\frac{\Sigma X_i Y_i}{\Sigma (X_i - \bar{X})^2}]$ 1
   $= \frac{\Sigma X_i E[Y_i]}{\Sigma (X_i - \bar{X})^2} = \frac{\Sigma X_i E[\beta_0 + \beta_1 X_i + \epsilon_i]}{\Sigma (X_i - \bar{X})^2} = \frac{\beta_0 \Sigma X_i + \beta_1 \Sigma X_i^2}{\Sigma (X_i - \bar{X})^2}$ 1
   The estimator is not unbiased 1
   Except when $\bar{X} = 0$ 1

3. In a not-so-recent (1905) experiment, British scientists measured the head size and brain weight of several persons. Some R output from a fitted SLR model follows; you may assume all G-M assumptions are met.

```
> anova(fit)
Analysis of Variance Table

Response: brainWeight
           Df  Sum Sq Mean Sq F value    Pr(>F)
headSize  [A] 2184982     [B]     [C] < 2.2e-16
Residuals 235 1232728     [D]


> summary(fit)
Call:
lm(formula = brainWeight ~ headSize, data = brain)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 325.57342   47.14085     [E] 4.61e-11
headSize      0.26343    0.01291  20.409  < 2e-16
```

R^2 = SSReg / (SST)

```
Residual standard error: 72.43 on 235 degrees of freedom
Multiple R-squared:   [F],Adjusted R-squared:  0.6378
F-statistic: ----- on 1 and 235 DF,  p-value: < 2.2e-16

> summary(brain) # Sample stats
sex              age              headSize        brainWeight
Min.   :1.000    Min.   :1.000    Min.   :2720    Min.   : 955
1st Qu.:1.000    1st Qu.:1.000    1st Qu.:3389    1st Qu.:1207
Median :1.000    Median :2.000    Median :3614    Median :1280
Mean   :1.435    Mean   :1.536    Mean   :3634    Mean   :1283
3rd Qu.:2.000    3rd Qu.:2.000    3rd Qu.:3876    3rd Qu.:1350
Max.   :2.000    Max.   :2.000    Max.   :4747    Max.   :1635
```

(a) (6 points) Some values have been replaced with letters. Fill in those values. You do not need to show any work for this part.

(A) (B) (C)

(D) (E) (F)

Solution: (A) 1 (B) 2184982 (C) 416.5273
(D) 5246.1 or 5245.7 (E) 6.9064 (F) 0.6393

(b) (2 points) What is the sample standard deviation of the predictor variable?

S_{xx}

Solution: $SS_x = SSR/b_1^2 = 2184982/0.26343^2 = 31485993$ (1)
$SD(X) = \sqrt{\frac{SS_x}{n-1}} = \sqrt{\frac{31485993}{236}} = 365.26$ (1)

(c) (2 points) Give a 95% CI for the true regression slope β_1.

Solution: $95\% CI\ for\ \beta_1 : b_1 \pm t_{235,0.975} s\{b_1\}$ (1)
$0.26343 \pm (1.97)(0.01291)$
0.2634 ± 0.0254 (1)

(d) (2 points) Give a 99% CI for the true regression intercept β_0.

Solution: $99\% CI\ for\ \beta_0 : b_0 \pm t_{235,0.995} s\{b_0\}$ (1)
$325.5734 \pm (2.60)(47.14085)$
325.5734 ± 122.5662 (1)

(e) (1 point) What is the expected head size for subjects who have a brain weight of 1300?

Solution: You cannot get an unbiased answer from the output given. (1)

(f) (1 point) What is the expected brain weight for subjects who have a head size of 1300?

Solution: You cannot safely make this prediction as it is out of range. (1)

(g) (5 points) What is the expected brain weight for subjects who have a head size of 3100? Give an appropriate 95% Interval estimate for this prediction.

Solution: $\hat{Y}_h = 325.57342 + 0.26343(3100) = 1142.204$ 1
$SS_x = 31485993$
$s^2\{\hat{Y}_h\} = MSE\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{SS_x}\right)$ 1
$= 5246.1\left(\frac{1}{237} + \frac{(3100-3634)^2}{31485993}\right) = 5246.1(0.013276) = 69.647$ 1
$95\% CI\ for\ E[\hat{Y}_h] : \hat{Y}_h \pm t_{235,0.975} s\{\hat{Y}_h\}$ 1
$1142.204 \pm 1.97\sqrt{69.647}$
1142.204 ± 16.44 1

(h) (4 points) Test the hypothesis that the intercept is equal to 200 (vs. the alternative that it is not 200) at the 5% level. State the hypotheses formally, give the test statistic, df and p-value range, and your conclusion in a plain English sentence.

Solution: $H_0 : \beta_0 = 200\ vs\ H_a : \beta_0 \neq 200$
$t^* = \frac{b_0 - 200}{s\{b_0\}} = \frac{325.5734 - 200}{47.14085} = 2.6638$ on 235 df 1
One-sided p-value $\epsilon(0.001, 0.005)$
Two-sided p-value $\epsilon(0.002, 0.01)$ 1
$\therefore$ We can reject the claim that the intercept is 200 at this significance level. 1

Two separate models were fit using subsets of the data for males and females. Some R output follows:

```
> summary(fitM) # Men
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 430.30269   77.19134   5.574 1.34e-07
headSize      0.23736    0.02025  11.723  < 2e-16

Residual standard error: 74.54 on 138 degrees of freedom
Multiple R-squared:  0.5101,Adjusted R-squared:  0.5063
F-statistic: 137.4 on 1 and 138 DF,  p-value: < 2.2e-16

> summary(fitW) # Women
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 286.08702   75.95198   3.767 0.000278
headSize      0.27280    0.02212  12.330  < 2e-16

Residual standard error: 65.92 on 101 degrees of freedom
Multiple R-squared:  0.6008,Adjusted R-squared:  0.5969
F-statistic:   152 on 1 and 101 DF,  p-value: < 2.2e-16
```

(i) (2 points) For which sex do we have a stronger indication of a relationship? How do you know this?

> **Solution:** Females (1), because t-statistic is higher (and n is lower) (1)

(j) (2 points) For which sex does head size explain a higher proportion of variation in brain weight? How do you know this?

> **Solution:** Females (1), because R^2 is higher. (1)

(k) (3 points) Which of the two models do you prefer *for making a confidence interval for the* $E[Y_h]$ at the average head size for each sex? How did you arrive at this conclusion?

> **Solution:** Males (1), because of a lower MSE/n. (2)

Critical values of the t distribution. Upper tail area is the column heading.

DF	0.25	0.2	0.15	0.1	0.05	0.025	0.01	0.005	0.001	5e-04	1e-04
1	1.00	1.38	1.96	3.08	6.31	12.71	31.82	63.66	318.31	636.62	3183.10
2	0.82	1.06	1.39	1.89	2.92	4.30	6.96	9.92	22.33	31.60	70.70
3	0.76	0.98	1.25	1.64	2.35	3.18	4.54	5.84	10.21	12.92	22.20
4	0.74	0.94	1.19	1.53	2.13	2.78	3.75	4.60	7.17	8.61	13.03
5	0.73	0.92	1.16	1.48	2.02	2.57	3.36	4.03	5.89	6.87	9.68
6	0.72	0.91	1.13	1.44	1.94	2.45	3.14	3.71	5.21	5.96	8.02
7	0.71	0.90	1.12	1.41	1.89	2.36	3.00	3.50	4.79	5.41	7.06
8	0.71	0.89	1.11	1.40	1.86	2.31	2.90	3.36	4.50	5.04	6.44
9	0.70	0.88	1.10	1.38	1.83	2.26	2.82	3.25	4.30	4.78	6.01
10	0.70	0.88	1.09	1.37	1.81	2.23	2.76	3.17	4.14	4.59	5.69
11	0.70	0.88	1.09	1.36	1.80	2.20	2.72	3.11	4.02	4.44	5.45
12	0.70	0.87	1.08	1.36	1.78	2.18	2.68	3.05	3.93	4.32	5.26
13	0.69	0.87	1.08	1.35	1.77	2.16	2.65	3.01	3.85	4.22	5.11
14	0.69	0.87	1.08	1.35	1.76	2.14	2.62	2.98	3.79	4.14	4.99
16	0.69	0.86	1.07	1.34	1.75	2.12	2.58	2.92	3.69	4.01	4.79
18	0.69	0.86	1.07	1.33	1.73	2.10	2.55	2.88	3.61	3.92	4.65
20	0.69	0.86	1.06	1.33	1.72	2.09	2.53	2.85	3.55	3.85	4.54
24	0.68	0.86	1.06	1.32	1.71	2.06	2.49	2.80	3.47	3.75	4.38
28	0.68	0.85	1.06	1.31	1.70	2.05	2.47	2.76	3.41	3.67	4.28
32	0.68	0.85	1.05	1.31	1.69	2.04	2.45	2.74	3.37	3.62	4.20
36	0.68	0.85	1.05	1.31	1.69	2.03	2.43	2.72	3.33	3.58	4.14
40	0.68	0.85	1.05	1.30	1.68	2.02	2.42	2.70	3.31	3.55	4.09
50	0.68	0.85	1.05	1.30	1.68	2.01	2.40	2.68	3.26	3.50	4.01
60	0.68	0.85	1.05	1.30	1.67	2.00	2.39	2.66	3.23	3.46	3.96
70	0.68	0.85	1.04	1.29	1.67	1.99	2.38	2.65	3.21	3.44	3.93
80	0.68	0.85	1.04	1.29	1.66	1.99	2.37	2.64	3.20	3.42	3.90
100	0.68	0.85	1.04	1.29	1.66	1.98	2.36	2.63	3.17	3.39	3.86
150	0.68	0.84	1.04	1.29	1.66	1.98	2.35	2.61	3.15	3.36	3.81
200	0.68	0.84	1.04	1.29	1.65	1.97	2.35	2.60	3.13	3.34	3.79
500	0.67	0.84	1.04	1.28	1.65	1.96	2.33	2.59	3.11	3.31	3.75
1000	0.67	0.84	1.04	1.28	1.65	1.96	2.33	2.58	3.10	3.30	3.73
Inf	0.67	0.84	1.04	1.28	1.64	1.96	2.33	2.58	3.09	3.29	3.72