STAT 308 – In Class Exam 1

Name: Solutions

- There are a total of 50 points possible on this exam.
- The score of this in class exam will be combined with your take home exam for your total Exam 1 score.
- For questions where you are asked to perform calculations, you may leave your answers incomplete. For example $\frac{900-644}{7}$ would be an acceptable answer.
- You are allowed one two-sided "cheat sheet" for use on this exam.

Good luck!

This page intentionally left blank.

1. [2 pts each] **TRUE OR FALSE** For each of the following statements, determine if they are true or false (please write out either true or false, so your answer is clear to me). If the statement is false, please correct the statement to make it true.

a. I obtain a 95% confidence interval for a population mean μ of $(0.95, 1.2)$. If I were to test $H_0 : \mu = 1$ vs. $H_a : \mu \neq 1$ at $\alpha = 0.05$, I would reject H_0.

False, I would fail to reject H_0.

b. r^2 is the percent of variation in the dependent variable Y explained by its relationship with the independent variable X.

False, it is the percent variation in Y explained by its linear relationship with X.

c. If I remove an influential observation in the dataset, it will always change the estimated equation of the least squares regression line.

True

d. A random scatter of points across all observed values of X in the residual plot provides evidence that the assumption of normally distributed residuals is not violated.

(a) This means the assumption of homoscedasticity is violated.
(b) Points following the 45° line on a QQ-plot means normality assumption is not violated

e. The p-value from the ANOVA table is the p-value for testing for a significant linear relationship between the explanatory variable X and the response variable Y.

True

2. [3 pts each] Consider the following incomplete ANOVA table for a simple linear regression model.

	df	Sum Sq	Mean Sq	F value	Pr(>F)
Model		490.89			
Error					
Total	34	667.23			

a. What are the model and error degrees of freedom?

$df_{MODEL} = 1$

$df_{ERROR} = 33 = 34 - 1$

b. What are the sum of squared errors?

$SSE = SST - SSM$
$= 667.23 - 490.89$

c. What is the estimate of the regression variance?

$s^2 = \frac{SSE}{df_{ERROR}} = \frac{667.23 - 490.89}{33} = MSE$

d. What is the test statistic used to test for a significant linear relationship between X and Y?

$F = \frac{MSM}{MSE} = \frac{490.89}{\left(\frac{667.23 - 490.89}{33}\right)}$

e. What is the distribution of the test statistic under the null hypothesis in (2d)?

F with 1 & 33 degrees of freedom

3. The amount of time it takes to wait in minutes for the next eruption of the Old Faithful geyser in Yellowstone National Park is considered to be related linearly to the length in minutes of the most recent eruption. We wish to formally test this hypothesis, and the relavent output for this problem can be found in the back of this exam.

a. [2 pts] What are the response and explanatory variables for this problem?

X = Time (in minutes) of most recent eruption

Y = Wait time (in minutes) until next eruption

b. [2 pts] Report the least squares regression line.

$\hat{Y} = 33.47 + 10.73$

c. [3 pts] Report and interpret a 95% confidence interval for the slope of the least squares regression line.

(10.11, 11.35)

We are 95% confident that when the time of the most recent eruption increases by 1 minute, the expected wait time until the next eruption increases between 10.11 and 11.35 minutes.

d. [2 pts] Report the estimate of the regression standard devation.

5.914

e. [2 pts] Interpret the r^2 value in the context of the problem.

81.15% of the variation in wait time until the next eruption can be explained by its linear relationship with time of most recent eruption.

f. [3 pts] Report and interpret a 90% confidence interval for the mean of Y when $x = 5$.

(86.14, 88.11)

We are 90% confident that when the most recent eruption is 5 minutes, the true mean wait time for the next eruption is between 86.14 and 88.11 minutes.

g. [3 pts] Report and interpret a 90% prediction interval for the mean of Y when $x = \bar{x}$ (Note, leave the sample mean in your final answer as $\bar{x}$).

(61.12, 80.68)

We are 90% confident that for a randomly selected eruption that lasted $\bar{x}$ minutes, the wait time for the next eruption is between 61.12 and 80.68 minutes.

h. [4 pts] Perform a formal hypothesis test for a significant linear relationship between duration of the last eruption and time to wait until the next eruption. Be sure to properly state all necessary information to perform this hypothesis test.

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

$t = 34.09$ p-value $< 2.2e^{-16}$

We reject H_0 and conclude there is significant evidence of a linear relationship between time of most recent eruption and wait time until the next eruption.

i. [2 pts] Determine if the assumption of homoskedasticity is violated. Be sure to properly identify which parts of the R output you used to answer this question.

Residual plot on left:
There appears to be a random scatter of points around 0, so the assumption appears to not be violated.
(Multiple possible responses)

j. [2 pts] Determine if the assumption of normally distributed residuals is violated. Be sure to properly identify which parts of the R output you used to answer this question.

QQ plot on right.
The points appear to fall closely to the 45° line, so the assumption appears to not be violated.

Linear Model Output from the Geyser Dataset

Table of the summary output

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.4744	1.1549	28.98	<2e-16
X	10.7296	0.3148	34.09	<2e-16

Residual standard error: 5.914 on 270 degrees of freedom

Multiple R-squared: 0.8115

95% Confidence Intervals for the least squares regression estimates

	2.5 %	97.5 %
(Intercept)	31.20	35.75
X	10.11	11.35

90% confidence and prediction intervals for $x = 5$ and $x = \bar{x}$

(86.14, 88.11), (77.31,96.93), (70.31,71.49), (61.12,80.68).

Plots for Checking Assumptions