STAT 350 Exam 3 Fall 2023

V1

Name: _____	PUID _____

Instructor (circle one): **Anand Dixit** **Timothy Reese** **Halin Shin** **Khurshid Alam**

Class Start Time: ○ 11:30 AM ○ 12:30 PM ○ 1:30 PM ○ 2:30 PM ○ 3:30 PM ○ 4:30 PM ○ Online

As a boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do. Accountable together - we are Purdue.

Instructions:

1. **IMPORTANT** Please write your **name** and **PUID clearly** on every **odd page.**
2. **Write your work in the box. Do not run over into the next question space.**
3. You are expected to uphold the honor code of Purdue University. It is your responsibility to keep your work covered at all times. Anyone caught cheating on the exam will automatically fail the course and will be reported to the Office of the Dean of Students.
4. It is strictly prohibited to smuggle this exam outside. Your exam will be returned to you on Gradescope after it is graded.
5. The only materials that you are allowed during the exam are your **scientific calculator**, **writing utensils, erasers, your crib sheet**, and **your picture ID**. Colored scratch paper will be provided if you need more room for your answers. Please write your name at the top of that paper also.
6. The crib sheet can be a handwritten or type double-sided 8.5in x 11in sheet.
7. Keep your bag closed and cellphone stored away securely at all times during the exam.
8. If you share your calculator or have a cell phone at your desk, you will get a **zero** on the exam.
9. The exam is only 60 minutes long so there will be no breaks (including bathroom breaks) during the exam. If you leave the exam room, you must turn in your exam, and you will not be allowed to come back.
10. You must show **ALL** your work to obtain full credit. An answer without showing any work may result in **zero** credit. If your work is not readable, it will be marked wrong. Remember that work has to be shown for all numbers that are not provided in the problem or no credit will be given for them. All explanations must be in complete English sentences to receive full credit.
11. All numeric answers should have **four decimal places** unless stated otherwise.
12. After you complete the exam, please turn in your exam as well as your table and any scrap paper that you used. Please be prepared to **show your Purdue picture ID**. You will need to **sign a sheet** indicating that you have turned in your exam.

Your exam is not valid without your signature below. This means that it won't be graded.

I attest here that I have read and followed the instructions above honestly while taking this exam and that the work submitted is my own, produced without assistance from books, other people (including other students in this class), notes other than my own crib sheet(s), or other aids. In addition, I agree that if I tell any other student in this class anything about the exam BEFORE they take it, I (and the student that I communicate the information to) will fail the course and be reported to the Office of the Dean of Students for Academic Dishonesty.

Signature of Student: _____

You may use this page as scratch paper.
The following is for your benefit only; we will not use this for grading:

Question Number	Total Possible	Your points
Problem 1 (True/False) (2 points each)	12	
Problem 2 (Multiple Choice) (3 points each)	9	
Problem 3	29	
Problem 4	34	
Problem 5	21	
Total	105	

1. (12 points, 2 points each) True/False Questions. Please indicate the correct answer by filling in the circle. *If you indicate the correct answer by any other way, you may receive 0 points for the question*.

1.1. In **ANOVA** the mean squared error $\mathbf{MS_E}$ measures the within group variability.

Ⓣ or Ⓕ If the homogeneity of variance assumption is valid then $\mathbf{MS_E}$ is also an unbiased estimate of the variance.

1.2. A least squares regression is conducted, and the estimated slope of the mean response, denoted as $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$**,** is found to have a value close to zero in magnitude.

Ⓣ or Ⓕ It follows that the sample Pearson correlation must also be close to zero in magnitude.

1.3. Both **Bonferroni** and **Tukey's** method are statistical techniques used to control the **Family-Wise Error Rate (FWER)** in multiple comparison procedures.

Ⓣ or Ⓕ Tukey's method is generally less conservative than Bonferroni's method in controlling the $\mathbf{FWER}$.

1.4. Suppose you are analyzing a dataset that represents the annual population growth of a city over the past 50 years. You have data from 1970 to 2020, and you've performed a linear regression analysis to model the population growth. The linear regression equation is: $\hat{y} = \mathbf{250,000} + \mathbf{5,000} * \boldsymbol{x}_{\mathbf{year}}$**.** In this equation, $\boldsymbol{x}_{\mathbf{year}}$ denotes the number of years since 1970 with $\boldsymbol{x}_{\mathbf{year}} = \mathbf{0}$ for the year 1970.

Ⓣ or Ⓕ Given that all assumptions of the linear regression model are valid, and the coefficient of determination $\boldsymbol{R}^{\mathbf{2}} = \mathbf{0.99}$, we can safely use the linear regression model to accurately predict the city's population for the year 2030 to be 550,000.

1.5. A regression analysis between weight **($\boldsymbol{y}$ kg)** and height **($\boldsymbol{x}$ cm)** resulted in the following least-squares regression line: $\hat{y} = -\mathbf{5} + \mathbf{0.4}\boldsymbol{x}$.

Ⓣ or Ⓕ In this context, the estimated value of the slope ($\boldsymbol{b}_{\mathbf{1}} = \mathbf{0.4}$) indicates that if the height is increased by **1 cm**, the weight will exactly increase by **0.4 kg**.

1.6. Consider a random variable ***X*** that follows an F distribution with numerator and denominator degrees of freedom equal to 5 and 15, respectively.

Ⓣ or Ⓕ In this context, it is theoretically possible for the random variable ***X*** to take negative values.

2. (9 points, 3 points each) Multiple Choice Questions. Please indicate the correct answer by filling in the circle. *If you indicate the correct answer by any other way, you may receive 0 points for the question*. For each question, there is only one correct option given.

2.1. In the context of a researcher conducting an ANOVA analysis to compare the population means of 4 populations, and all necessary ANOVA assumptions have been met, and the ANOVA procedure has resulted in statistical significance, how many total pairwise comparisons should the researcher conduct as a follow-up?

Ⓐ **2**

Ⓑ **4**

Ⓒ **6**

Ⓓ **10**

Ⓔ **24**

2.2. A researcher ran regression analysis between two numerical variables, and performed a hypothesis test with the following null and alternative hypothesis:

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

During the analysis, the researcher discovers that there is a positive association between these two variables and all the data points lie on the regression line. **Choose the statement that must be false in regard to this analysis.**

Ⓐ $\mathbf{MS_E = 0}$

Ⓑ The coefficient of determination is 1.

Ⓒ $|\hat{\beta}_1| = 0$

Ⓓ The Pearson correlation coefficient r is 1.

Ⓔ The measured residuals error terms are all zero.

2.3. A large value of the coefficient of determination in a least-squares regression analysis indicates:

Ⓐ The model can produce accurate and reliable predictions for any value of the explanatory variable.

Ⓑ The model's explained variation is smaller than the unexplained (error) variation.

Ⓒ The relationship between the explanatory variable and the response is linear.

Ⓓ The majority of the variation in the response has been explained by the regression line.

Ⓔ The total variation is equal to the unexplained (error) variation.

3. (29 points) A college dietitian has developed three diet plans, namely Diet 1, Diet 2, and Diet 3, to assist college students in losing weight. The weight of participants who followed these diet programs was measured in pounds both before starting the diet and after completing three months on the respective diet plan. The dietitian is investigating whether there are any differences in the effectiveness of these three diets in terms of the average weight loss achieved. Weight loss is defined as the difference between the weight measured after three months and the weight measured before starting the diet. The summary information for each diet is provided below.

	Diet 1	Diet 2	Diet 3
n	10	10	10
$\bar{x}$	10.24	9.72	11.48
s	0.80	1.21	0.86

(a) (2 points) Check the **assumption of constant variance** using the provided summary information. Show your work and state clearly whether the assumption was satisfied or not. Apart from constant variance, you may assume that all other assumptions have been satisfied for the rest of the analysis.

$$\frac{s_{\max}}{s_{\min}} = \frac{1.21}{0.8} = 1.5125 < 2$$

Since, the ratio of the largest variance to the minimum variance is smaller than 2 **the assumption is valid** and we may proceed with the analysis.

(b) (7 points) Complete the **ANOVA** table below.

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Test Statistic
Factor	$k - 1 = 3 - 1 = 2$	$SS_T - SS_E = 42.05 - 25.74 = 16.31$	$\frac{SS_A}{df_A} = \frac{16.31}{2} = 8.155$	$F_{TS} = \frac{MS_A}{MS_E} = \frac{8.155}{0.9533} = 8.5545$ or 8.5542
Error	$n - k = 3 * 10 - 3 = 27$	25.74	$\frac{SS_E}{df_E} = \frac{25.74}{27} = 0.9533$	
Total	$n - 1 = 3 * 10 - 1 = 29$	42.05		

(c) (2 points) What is the estimated value of the assumed common variance among the three diet plans in the analysis?

$s^2 = MS_E = 0.9533$

(d) (3 points) Choose the correct p-value associated with the **ANOVA** table in part (b) from the options below. Assume that F-ts, df-factor and df-error are the correct test statistic, degrees of freedom for factor, and degrees of freedom for error, respectively.

Ⓐ pf(F-ts, df1 = df-factor, df2 = df-error, lower.tail = TRUE) = 0.9986727

Ⓑ pf(F-ts, df1 = df-factor, df2 = df-error, lower.tail = FALSE) = 0.001327266

Ⓒ 2*pf(F-ts, df1 = df-factor, df2 = df-error, lower.tail = FALSE) = 0.002654533

Ⓓ pf(F-ts/2, df1 = df-factor, df2 = df-error, lower.tail = TRUE) = 0.9756361

(e) (6 points) At **5% level of significance**, is there evidence that the average weight lost due to at-least one Diet is different? Provide a formal decision and conclusion in context.

Decision: The ***p*-value** $= 0.001327266 \leq 0.05 = \alpha$ therefore we have evidence to reject the null hypothesis H_0.

The data **does** give **strong** support (***p*-value** = 0.001327266) to the claim that that the true average weight lost due to at least one diet is different from the other two diets.

(f) (4 points) Given the results in part (e), will it be meaningful to conduct a pairwise comparison? Provide an explanation for your answer.

Yes, it will be meaningful. This is because we have identified that at least one diet results in an average weight loss that is different from the rest. The next step is to examine which diets are different.

The college dietician did conduct a pairwise comparison while maintaining the overall type I error at 5% using Tukey's HSD method. The R output for it is given below:

	diff	lwr	upr	p adj
Diet 2-Diet 1	-0.5165791	-1.5992860	0.5661278	0.473240087
Diet 3-Diet 1	1.2403638	0.1576569	2.3230707	0.022312262
Diet 3-Diet 2	1.7569429	0.6742360	2.8396498	0.001173322

(g) (5 points) Using the above output and the table of sample statistics draw a graphical representation of the Tukey's HSD results and write one to two complete English sentences stating which Diet program is the best to reduce weight. Please explain your answer.

(2 points graph)

$\bar{x}_{\text{Diet 2}}$	$\bar{x}_{\text{Diet 1}}$	$\bar{x}_{\text{Diet 3}}$
9.72	10.24	11.48

Since, the pairwise Tukey comparison resulted in no significant difference between diet 1 and diet 2 they are statistically indistinguishable in the population. Diet 3 is different from both diet 1 and diet 2 and resulted in the largest reduction in weight so there is evidence to suggest that **diet 3** is the better program to reduce weight.

4. (33 points) A scientist is studying the relationship between annual rainfall (***x***) measured in centimeters and shoreline erosion (***y***) which is also measured in centimeters. For each of ten annual rainfall levels, a randomly selected shoreline erosion value was measured. The study reported the following data. **You may assume all model assumptions hold.**

x	**30**	**25**	**90**	**60**	**50**	**35**	**75**	**110**	**45**	**80**
y	**0.3**	**0.2**	**5.0**	**3.0**	**2.0**	**0.5**	**4.0**	**6.0**	**1.5**	**4.0**

You are also given the following summary information:

$\sum x_i = 600$	$\sum x_i^2 = 43,100$	$\sum y_i = 26.5$	$\sum y_i^2 = 108.63$	$\sum x_i * y_i = 2,109$
$\sqrt{\text{MSE} * \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right)} = 0.139718$ where $x^* = 100\ cm$			$\sqrt{\text{MSE} * \left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right)} = 0.2819949$ where $x^* = 100\ cm$	

(a) (10 points) Fit the linear regression line for the association between shoreline erosion and rainfall and interpret the value of b_1.

i) Determine the slope (b_1) of the least-squares regression line.

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i * y_i - \frac{1}{n} * (\sum x_i) * (\sum y_i)}{\sum x_i^2 - \frac{1}{n} * (\sum x_i)^2} = \frac{2,109 - \frac{1}{10} * 600 * 26.5}{43,100 - \frac{1}{10} * 600^2} = \frac{519}{7100} = 0.0731$$

ii) Determine the intercept (b_0) of the least squares regression line.

$$b_0 = \bar{y} - b_1 * \bar{x} = \frac{26.5}{10} - 0.0731 * \frac{600}{10} = -1.736$$

iii) Write out the equation of the regression line.

$$\hat{y} = -1.736 + 0.0731 * x_{\text{rainfall}}$$

(b) (5 points) Complete the ANOVA table. (You don't need to show work.)

Sources	DF	Sum of Squares	Mean Squares	F Value	Pr(>F)
Model	1	37.938	37.938	654.8276	< 0.0001
Error	8	$SS_E = SS_T - SS_R = 38.405 - 37.938 = 0.467$	0.058		
Total	$n - 1 = 9$	$SS_T = \sum y_i^2 - \frac{1}{n} * \left(\sum y_i\right)^2 = 108.63 - \frac{1}{10} * (26.5)^2 = 38.405$			

(c) (5 points) Is there a significant linear association between rainfall and shoreline erosion at a $\alpha = 0.01$ level of significance? Provide a formal decision and conclusion in context. Regardless of the results of the hypothesis test, determine the coefficient of determination and assess what this tells us about the linear regression in context.

Decision: The p-value $< 0.0001 \leq 0.01 = \alpha$ therefore we have evidence to reject the null hypothesis H_0.

The data **does** give **strong** support (p-value< 0.0001) to the claim that that there is a linear association between rainfall and shoreline erosion in the population.

$$R^2 = \frac{SS_R}{SS_T} = \frac{37.938}{38.405} = 0.9878$$

R^2 tells us that 98.78% of the variation in the shoreline erosion has been explained by the regression line.

(d) (3 points) What proportion of the total variation of the shoreline erosion is **NOT** explained by the rainfall?

$$1 - R^2 = 1 - 0.9878 = 0.0122$$

Therefore, **1.22%** of the total variation of the shoreline erosion is not explained by the linear relationship with rainfall.

(e) (5 points) Construct a **99% confidence interval** for $\boldsymbol{\beta_1}$. Select an appropriate critical value for the calculation.

`> qt(0.01/2, 9, lower.tail= FALSE)` `[1] 3.249836`	`> qt(0.01/2, 8, lower.tail= FALSE)` `[1] 3.355387`
`> qt(0.01, 9, lower.tail= FALSE)` `[1] 2.821438`	`> qt(0.01, 8, lower.tail= FALSE)` `[1] 2.896459`

$$t_{\alpha/2,8} = 3.355387$$
$$b_1 = 0.0731$$
$$\text{MSE} = 0.058$$
$$b_1 \pm t_{\alpha/2,8} * \sqrt{\frac{\text{MSE}}{S_{xx}}}$$
$$0.0731 \pm 3.355387 * \sqrt{\frac{0.058}{7100}} = (\mathbf{0.0635, 0.0827})$$

(f) (5 points) Using the same critical value as **(e)** construct a **99% prediction interval** of shoreline erosion when the **amount of rainfall is 100 cm**.

$$\hat{y}^* = -1.736 + 0.0731 * 100 = 5.574$$

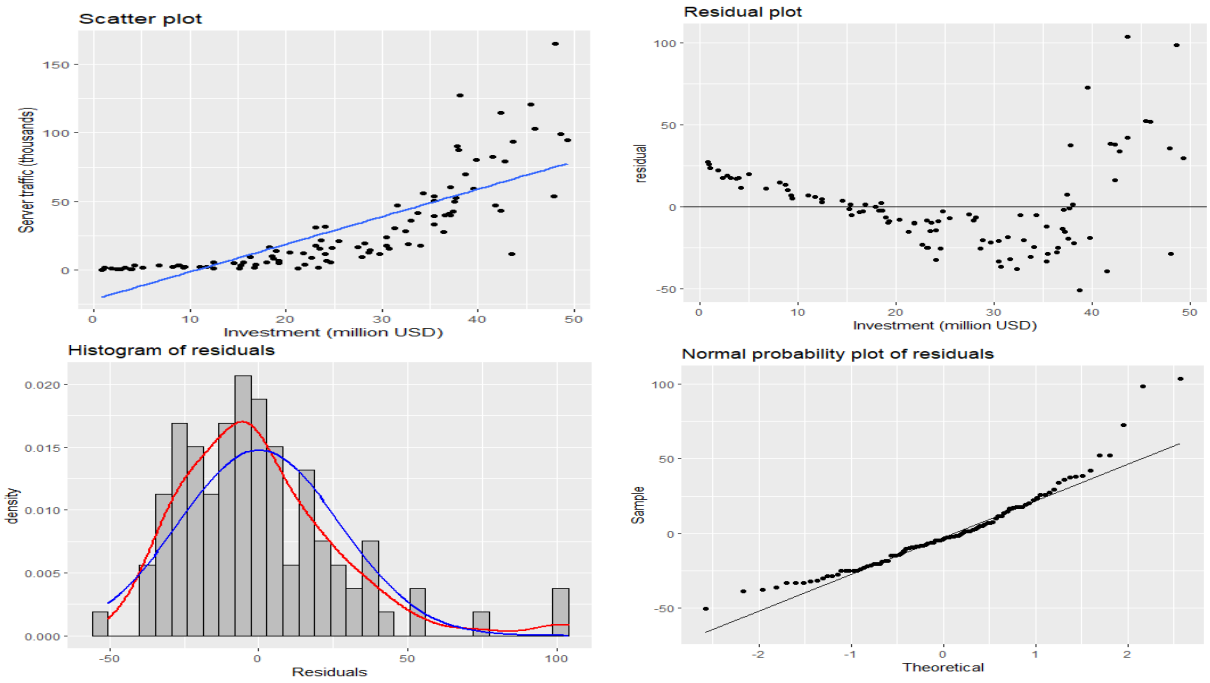
Standard Error: $\sqrt{\text{MSE} * \left(1 + \frac{1}{\text{n}} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right)} = 0.2819949$

$$\hat{y}^* \pm t_{\alpha/2,8} \sqrt{\text{MSE} * \left(1 + \frac{1}{\text{n}} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right)} = 5.574 \pm 3.355387 * 0.2819949 = (\mathbf{4.6278, 6.5202})$$

There was a mistake in the standard error.

The real SE is 0.2780997 so the interval should be: $(\mathbf{4.6409, 6.5071})$

5. (21 points) A university is studying the relationship between the amount of investment in server facilities to the average server traffic. The research team obtained the following set of graphics from their data of 100 pairs, collected from an SRS of other universities with servers of varying sizes.



(a) (4 points) List **ALL** assumptions which should be satisfied to conduct inference in a simple linear regression analysis.

Assumptions:

1) **SRS**: The response **(server traffic)** is obtained as a simple random sample for each fixed x **(investment)**. It is okay if they mention that the pairs are SRS.
2) **Linearity**: Assume that the relationship between server traffic and investment dollars is linear.
3) **Homogeneity/Constant Variance**: Assume that the residual errors have common variance.
4) **Normality**: Assume that the residual errors are normally distribued.

(b) (13 points) Evaluate whether each assumption in **(a)** is satisfied. For each assumption, where appropriate, clearly state **ALL** graphs that can be used to assess the validity of the assumption and the features used. Feel free to organize your answer using bullet points or a table.

1) **SRS**: Is assumed true as stated in the problem description. **(Valid)**
2) **Linearity**: Curvature clearly visible. **(Invalid)**
   Graphs:
   i. Scatter Plot: Curvature of points
   ii. Residual Plot: Curvature of points
3) **Homogeneity/Constant Variance**: Variability is low for small values (below 20) and high for large values (above 20). **(Invalid)**
   Graphs:
   i. Scatter Plot: Concentration of Points
   ii. Residual Plot: Concentration of Points
4) **Normality**: The residuals are positively skewed and have outliers. **(Invalid)**
   Graphs:
   i. Histogram of Residuals: Gaps in histogram indicate outliers and long right tail. Blue normal curve and red kernel curve do not align.
   ii. Normal Probability Plot of Residuals: Concave up and points do not follow the line.

(c) (4 points) Below is the R output from running a linear regression analysis with this dataset. What can you conclude based on this and your answers of **(a)** and **(b)**?

```
Call:
lm(formula = y ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-50.611 -19.427  -3.735  13.730 103.656

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -26.1706     5.5543  -4.712 8.11e-06 ***
x             2.2408     0.1933  11.589  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.15 on 98 degrees of freedom
Multiple R-squared:  0.5782,    Adjusted R-squared:  0.5739
F-statistic: 134.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

The R output does not matter as our assumptions have been strongly violated and we cannot conclude anything other than that the relationship between server traffic and investment is non-linear.

This page is intentionally left blank for scratch paper.

This page is intentionally left blank for scratch paper.