1. True/False Questions. Please indicate the correct answer by filling in the circle. *<u>If you indicate the correct answer by any other way, you may receive 0 points for the question</u>*.

1.1. In a statistical analysis, an ANOVA hypothesis test is conducted, in which all assumptions were valid. The test obtained significance concluding at least one of the 5 population means was different from the rest. To further investigate which means are significantly different, a follow-up Tukey multiple comparison procedure is to be conducted.

Ⓣ or Ⓕ The total number of Tukey Multiple comparisons would be 5.

1.2. A simple linear regression is conducted in which the **coefficient of determination R^2** is found to be **0.85**.

Ⓣ or Ⓕ Using the R^2 as the only reference we can conclude due to the large R^2 that a linear relationship best describes the relationship between the explanatory variable and the response variable.

1.3. In an ANOVA hypothesis test in which all assumptions were valid,

Ⓣ or Ⓕ if the within-group variation is much smaller than the between-group variation, the ***p*-value** will be large.

1.4. In an ANOVA testing procedure with three groups labeled **A, B** and **C,** we test the following hypothesis:

Ⓣ or Ⓕ: $H_0: \mu_A = \mu_B = \mu_C$

$H_a: \mu_A \neq \mu_B$ **or** $\mu_A \neq \mu_C$ **or** $\mu_B \neq \mu_C$ **or** $\mu_A \neq \mu_B \neq \mu_C$

1.5. In simple linear regression we must check the assumptions of linearity, homogeneity of variance, and normality of the residuals using diagnostic plots.

Ⓣ or Ⓕ The residual plot is a useful diagnostic plot to check the assumption of normality of the residuals.

1.6. In simple linear regression, a prediction interval is wider than a confidence interval for the mean response.

Ⓣ or Ⓕ This is because the prediction interval has to account for two sources of variation, whereas the confidence interval for the mean response only needs to account for one source of variation.

2. 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. Only one option should be selected in each multiple-choice problem.*

2.1 In a statistical analysis, an ANOVA hypothesis test is conducted, in which all assumptions were valid. The test revealed a significant difference among population means. To further investigate which means are significantly different, a multiple comparison procedure was performed with a family-wise error rate of $\alpha = 0.01$. A **99% confidence interval** of **(-2.17, 6.47)** was obtained from a Tukey multiple comparison procedure between **levels 1** and **2**, what conclusion can you draw about the hypothesis test?

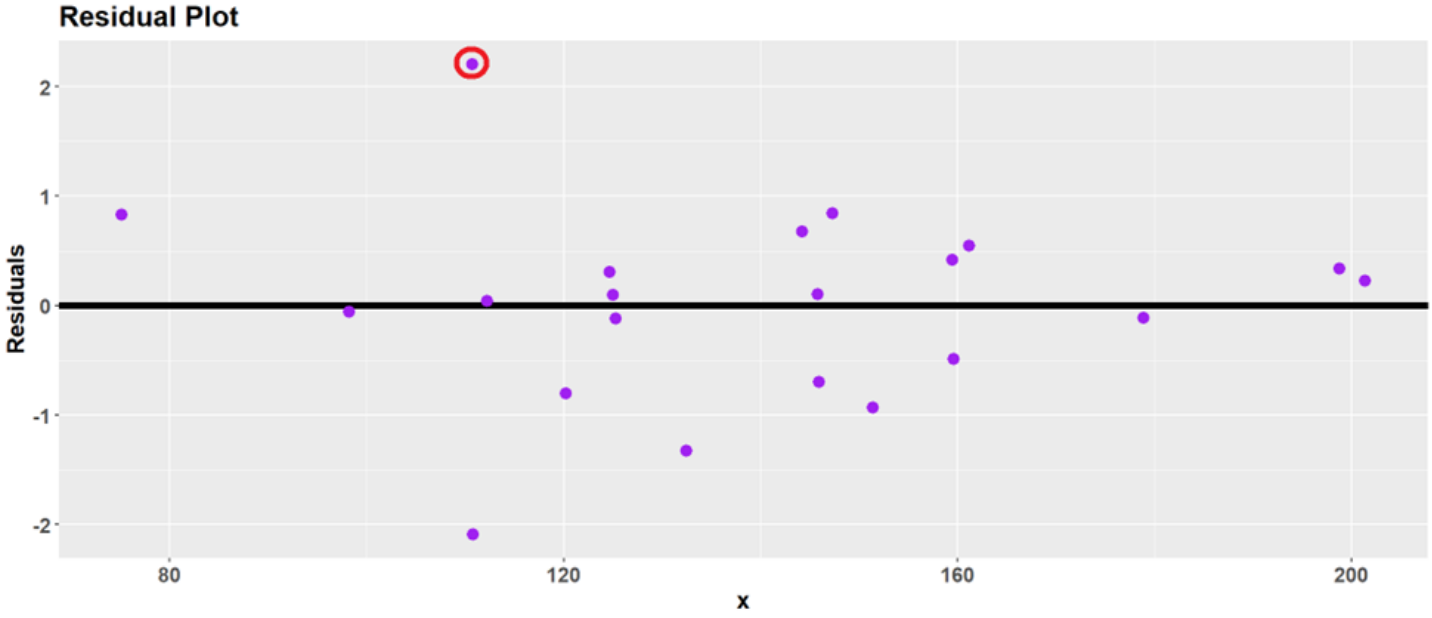
Ⓐ Fail to reject H_0 at $\alpha = 0.01$; there is no evidence that the true means of levels 1 and 2 are different.

Ⓑ Fail to reject H_0 at $\alpha = 0.01$; there is evidence that the true means of levels 1 and 2 are different.

Ⓒ Reject H_0 at $\alpha = 0.01$; there is no evidence that the true means of levels 1 and 2 are different.

Ⓓ Reject H_0 at $\alpha = 0.01$; there is evidence that the true means of levels 1 and 2 are different.

2.2 In the following residual plot, which statement **best describes** the **point circled in red**? Recall that a point is said to be influential if it has a significant impact on the regression line and the overall fit of the model.



Ⓐ The point is a potential outlier for the explanatory variable (x) and is highly influential.

Ⓑ The point is a potential outlier for the explanatory variable (*x*) but is not influential.

Ⓒ The point is a potential outlier for the response variable (*y*) but is not influential.

Ⓓ The point is a potential outlier the response variable (y) and is highly influential.

Ⓔ The point is not an outlier.

2.3 The table below presents the results of an **ANOVA** analysis conducted to determine if there are significant differences among the group means. However, the table is not complete.

You are required to determine the **estimate of the population standard deviation** and identify the **correct number** of **groups** from the options provided below.

Source	Degrees of Freedom	Sum of Squares	Mean Square	F value	$\Pr(>F)$
Factor A (Between Groups)	$df_A =$	$SS_A = 10162$	$MS_A = 2540.6$		
Error (Within Groups)	$df_E = 452$	$SS_E = 2872$	$MS_E =$		
Total	$df_T = 456$	$SS_T = 13034$	$MS_T = 28.5833$		

Ⓐ 4 groups, **s** = 6.3540

Ⓑ 5 groups, **s** = 2.5207

Ⓒ 5 groups, **s** = 6.3540

Ⓓ 4 groups, **s** = 2.5207

2.4 A **98% confidence interval** for $\boldsymbol{\beta_1}$ is given as: **(1.97, 7.65)**. What would be the conclusion of the appropriate hypothesis test with H_a: $\boldsymbol{\beta_1 \neq 0}$?

Ⓐ Fail to reject the null hypothesis at $\boldsymbol{\alpha = 0.02}$ and there is an association between x and y.

Ⓑ Fail to reject the null hypothesis at $\boldsymbol{\alpha = 0.02}$ and there is no association between x and y.

Ⓒ Reject the null hypothesis at α = 0.02 and there is an association between x and y.

Ⓓ Reject the null hypothesis at α = 0.02 and there is no association between x and y.

2.5 In the event that the linear regression model fails to accurately represent the underlying relationship between the explanatory and response variables, what behavior is expected to be observed in the residual plot if we proceed to fit the linear model?

Ⓐ The residual plot will show a linear pattern.

Ⓑ The residual plot will conform to a normal distribution.

Ⓒ The residual plot will manifest a non-random pattern.

Ⓓ The residual plot will exhibit a mean of zero.

Ⓔ The residual plot cannot be obtained as the relationship is not linear.

3. A research group at Purdue is studying the effect of use of portable electronic devices on sleep duration among adolescents ages 10 to 17. A random sample of 50 adolescents was selected from Indiana to report their daily use of portable electronic devices. The participants measured their duration of sleep (DS). The records of use of portable electronic devices were used to calculate a screen use score (SUS) ranging from 0 to 100. A simple linear regression is performed with DS (Y) vs SUS (X). The R output is provided **at the end of the exam**.

List the assumptions for linear regression in the following table. Determine whether the assumptions have been satisfied or not using the information and graphs provided. Please explain your choice. Be sure to identify which information/graphs you use for each assumption. Please mention **ALL** graphs that are relevant to check each assumption. You may state the figure numbers in your response.

Assumption	Information/ Graph Used	Satisfied (Yes/No)	Explanation of whether the assumption is satisfied or not
SRS/Independence of observations	assumed or from the situation in the question	Yes	No explanation necessary
Linear relationship	Scatter plot (1) Residual plot (2)	no	From both plots, the points are not randomly spaced above and below the line. They are above the line at low and high values and below the line in the middle.
Constant variance (or standard deviation) of residuals	Scatter plot (1) Residual plot (2)	Yes	From both the scatter plot and the residual plot, the standard deviation is approximately the same for all values.
residuals have a normal distribution	Histogram (3) Normal probability plot (4)	Yes	Histogram: the thick and thin lines are close (slightly right skewed) with a sample size of 50 Normal probability plot: the dots are close enough to the line with a sample size of 50.

4. To study the effectiveness of learning platforms in students, researchers used three separate (**not simultaneously**) learning platforms. Namely: **Brightspace (BS), DataCamp (DC) and TopHat (TH)**. They measured the grades of the students when using these platforms and want to identify whether there is any measurable difference between the performance on a standardized test. The summary information for each learning platform is given below. Assume that all of the assumptions are valid. **The R output is at the end of the exam.**

Diet Plan	BS	DC	TH
n	10	10	10
$\bar{x}$	83.1	74.2	86.0
s	2.04	2.77	3.43

a. Using the R output and your knowledge, complete the ANOVA table below. Work is required for all values **NOT** in the output. You do not need to include the values for the greyed-out cells.

Source	Degrees of Freedom	Sum of Squares	Mean Square	F	p-value
Factor	$\mathrm{df}_A = k - 1 = 3 - 1 = 2$	$SS_A = 756.2$	$\mathrm{MS_A} = \frac{SS_A}{\mathrm{df}_A} = \frac{756.2}{2} = 378.1$	$\frac{378.1}{7.87} = 48.04$	1.28e-09
Error	$\mathrm{df_E} = n - k = 3 \times 10 - 3 = 27$	$SS_E = 212.5$	$\mathrm{MS_E} = \frac{SS_E}{\mathrm{df}_E} = \frac{212.5}{27} = 7.87$		
Total	$\mathrm{df_T} = n - 1 = 3 \times 10 - 1 = 20$	$\mathrm{SS_T} = SS_A + SS_E = 756.2 + 212.5 = 968.7$			

b. What is the estimated value of the variance?

c. The researcher wanted to know if the true mean score on a standard exam is dependent on the type of learning platform used. Please perform the hypothesis test at a 5% significance level. Provide all four steps as discussed in class. Be sure to include all degrees of freedom.

Step 1:

Let μ_{BS} = the population or true mean or average test score for Brightspace (BS)
μ_{DC} = the population or true mean or average test score for DataCamp (DC)
μ_{TH}= the population or true mean or average test score for TopHat (TH)

Step 2:

H_0: $\mu_{BS} = \mu_{DC} = \mu_{TH}$
H_a: **At least two $\mu_i's$ are different. OR**
H_a: $\mu_i \neq \mu_j$ for at least one $i \neq j$

Step 3:

$F_{ts} = 47.861$
$df_1 = dfa = 2$
$df_2 = dfe = 29$
p = 1.28e-9

Step 4:

Reject H_0 because 1.28e-9 ≤ 0.05

The data shows support (p = 1.29e-9) the claim that at least two of the population mean test scores for the different learning platforms are different.

d. From the results of the Tukey multiple-comparison method provided in the R output, please indicate whether each of the differences is significant or not (Yes or No). Then draw the graphical representation of the results. No work is required for this part.

$i - j$	Is this significant? (Y/N)
DC – BS	Yes
TH – BS	No
TH – DC	Yes

$\bar{x}_{DC}$	$\bar{x}_{BS}$	$\bar{x}_{TH}$
74.2	**83.1**	**86.1**

e. Write one to two complete English sentences stating which learning platform(s) are the best (have the highest population means). Please explain your answer.

Learning Platforms Brightspace (BS) and TopHat (TH) have higher scores on the standardized test than data camp but they are indistinguishable from each other.

5. A group of student developers created an app that detects grammar errors in English writing that is uploaded by the users. They designed the app so that its algorithm improves itself by learning from the historical inputs. To see if this feature works, the developers record the error detection rate (Y, error_rate) on fixed test data every time the app analyzes 10,000 sentences (X, number of sentences in units of 10,000). The developers make sure that the algorithm does not remember the test data after each recording. You may assume that all assumptions are valid for this data. **The R output is at the end of the exam.**

a. Using the computer output, write the equation of the least squares regression line. Identify all variables used in the equation.

$$\widehat{error\ detection\ rate} = 65.275946 + 0.055010 \times sentences$$

OR

$$\hat{y} = 65.275946 + 0.055010\ x$$

$\hat{y}$ denotes the error detection rate
x denotes number of sentences (in units of 10,000)

b. Does it make sense to interpret the y-intercept by itself in this situation? Please explain your answer. This question does not involve the calculated value of the y-intercept.

No, the algorithm cannot if there are no sentences in the training set, therefore, this number cannot be obtained.

c. What proportion of the error detection rate is explained by its linear relationship with the number of previously analyzed sentences? Please use at least **4 decimal places**.in your answer and all work.

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE} = \frac{1613}{1613 + 1243} = \frac{1613}{2856} = 0.5648$$

d. What does your answer **from part c) ONLY** (the previous question) say about the linearity of the relationship between the number of sentences and the error detection rate? Please explain your answer.

Answer: Nothing.
Reason: The value of R^2 states how close the values are to the best fit line. Not if the real relationship is linear. This needs to be determined by the scatterplot which was not provided.

e. From the data provided, interpret a 95% confidence interval for the slope. Please clearly indicate the interval and which output is used. Remember that context is required. Please include at least 3 decimal places in your answer.

output: 2
We are 95% confident that the population slope between error detection rate (y) and number of sentences (x) is covered by the interval (0.0391484, 0.07087106).

f. From the data provided, interpret a 95% interval for the population average error detection rate of the same algorithm after it has analyzed a total of **341,000** sentences (coded as **341**). Please clearly indicate the interval and which output is used.

Note: There was a typo in the exam, 341,000 was 314,000 and 341 was 314. We did not count points off if you used one of these two numbers.

output: 4
We are 95% confident that the population average error detection rate when the number of sentences is 341,000 (341), is covered by the interval (80.03079, 85.06721).

Problem 3

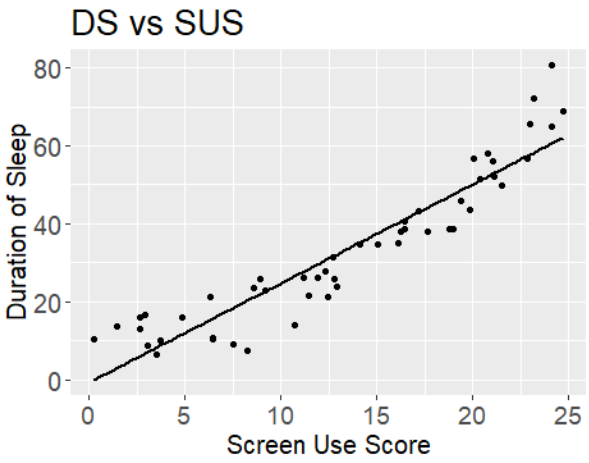


Figure 1

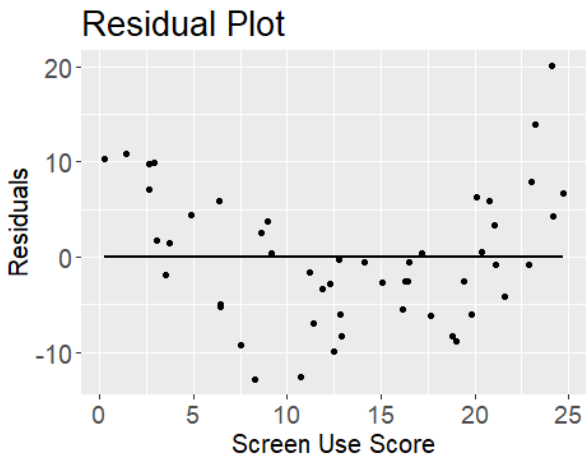
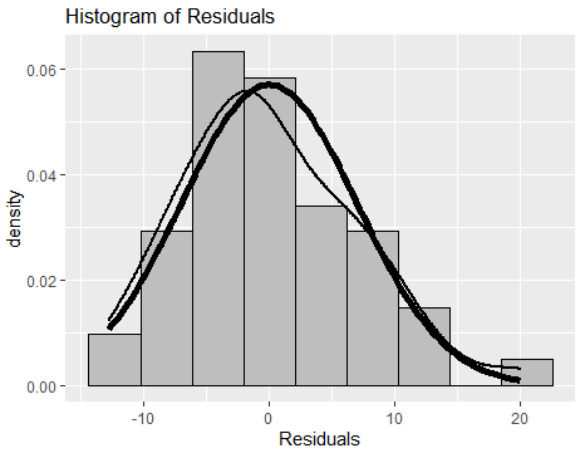
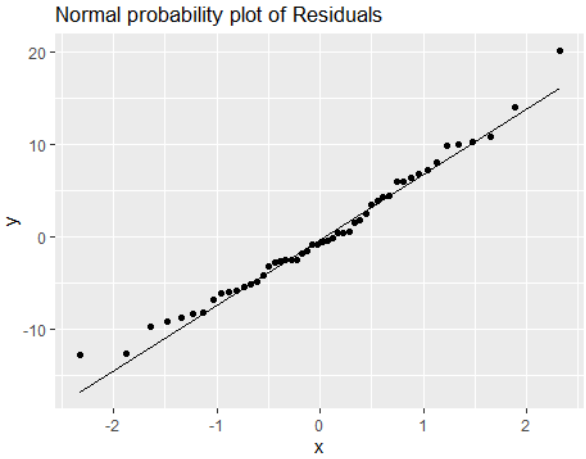


Figure 2





Problem 4

Output 1

```
> fit=aov(grades ~ platform, data= LMS)
> summary(fit)
            Df Sum Sq Mean Sq F value  Pr(>F)
platform    ??  756.2   ??     ????? 1.28e-09
Residuals   ??  212.5   ??
```

Output 2

```
> TukeyHSD(fit,'platform',conf.level=0.95)
DC-BS -8.889576 -12.0006270 -5.778525 0.0000004
TH-BS  2.914268  -0.1967832  6.025319 0.0696062
TH-DC 11.803844   8.6927928 14.914895 0.0000000
```

Problem 5

Output 1

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 65.275946   1.843331  35.412  < 2e-16 ***
sentences    0.055010   0.007835   7.021  2.3e-08

            Df Sum Sq Mean Sq F value  Pr(>F)
sentences   ??   1613  1612.9   ????? 2.3e-08
Residuals   38   1243    32.7
```

Output 2

```
> confint(fit, level=0.95)
                 2.5 %      97.5 %
(Intercept) 61.5443174 69.00757479
sentences    0.0391484  0.07087106
```

Output 3

```
> confint(fit, level=0.99)
                  0.5 %      99.5 %
(Intercept) 60.27764763 70.27424453
sentences    0.03376441  0.07625506
```

Output 4

```
> newdata = data.frame(sentences = 341)
> predict(fit, newdata, interval = "confidence",level = 0.95)
     fit      lwr      upr
1 82.549 80.03079 85.06721
```

Output 5

```
> newdata = data.frame(sentences = 341)
> predict(fit, newdata, interval = "confidence",level = 0.99)
     fit      lwr      upr
1 82.549 79.17601 85.92199
```

Output 6

```
> newdata = data.frame(sentences = 341)
> predict(fit, newdata, interval = "predict",level = 0.95)
     fit     lwr      upr
1 82.549 70.6985 94.39951
```

Output 7

```
> newdata = data.frame(sentences = 341)
> predict(fit, newdata, interval = "predict",level = 0.99)
     fit      lwr      upr
1 82.549 66.67594 98.42206
```