STA 200 Statistics II

Midterm Exam #1

Summer 2025

Please circle the correct answer. If you circle more than one answer, only the first one will be counted. Due to rounding error in some questions, please always choose the one that is closest to the answer you.

Problem 1.

Which function is used to calculate the mean of a numeric vector in R?

a) mean()
b) average()
c) median()
d) sum()

Answer: a)

Problem 2.

What does the **summary()** function provide for a numeric vector?

a) Minimum and maximum values only

b) 1st and 3rd quartiles only

c) Median and mean only

d) five-number summary (Min, Q1, Median, Q3, Max))

Answer: d)

Problem 3.
Six measurements were made of the magnesium ion concentration (in parts per million, or ppm) in a city's municipal water supply, with the following results.

202 164 157 177 113 213

Which of the following commands defines the R dataset to store the above concentrations?

a) conc <- c(202 164 157 177 113 213)
b) conc <- c(202, 164, 157, 177, 113, 213)
c) conc <- c(202. 164. 157. 177. 113. 213)

d) conc <- c(“202” “164” “157” “177” “113” “213”)

Answer b

Problem 4:

What does the **table()** function do in R?

a) Creates a frequency table for categorical data

b) Displays a structured data frame

c) Calculates summary statistics

d) Performs a t-test

Answer: a) Creates a frequency table for categorical data

Problem 5:
Which function is used to compute the correlation between two numeric vectors?

a) cor()
b) cov()
c) correlation()
d) rel()

Answer: a) cor()

Problem 6.
The following are the midterm exam grades of an introductory stats class.

70.6, 88.2, 94.1, 88.2, 88.2, 76.5, 85.3, 88.2, 85.3, 76.5, 85.3,
94.1, 97.1, 70.6, 91.2, 55.9, 91.2, 83.2, 79.4, 91.2, 94.1, 94.1,
67.5, 82.2, 77.8, 88.2, 62.4, 78.0, 83.3, 71.3, 91.0, 81.1, 75.7

Define an R data set first, then use the R function hist() to create a histogram. The histogram indicates that the distribution of the exam grades is

a) symmetric
b) skewed to the right
c) skewed to the left

Answer: c.

Problem 7

One of the primary foods for beef cattle is corn. The following table presents the average price in dollars for a bushel of corn and a pound of ribeye steak for 10 consecutive months.

Corn Price ($/bu)	Ribeye Price ($/lb)
6.34	13.22
5.93	12.06
6.45	13.72
6.19	12.75
6.56	13.27
5.94	12.17

What is the correlation coefficient between Corn Price and Ribeye Price? [*Hint: define two variables in R first, then use an appropriate R function to find the correlation coefficient.*]

a) 0.967
b) 0.875
c) 0.935
d) -0.875

Answer: C

```
Corn  <- c(6.34, 5.93, 6.45, 6.19, 6.56, 5.94)
Ribeye <- c(13.22, 12.06, 13.72, 12.75, 13.27, 12.17)
cor(
```

Problem 8

One of the primary foods for beef cattle is corn. The following table presents the average price in dollars for a bushel of corn and a pound of ribeye steak for 10 consecutive months.

Corn Price ($/bu)	Ribeye Price ($/lb)
6.34	13.22
5.93	12.06
6.45	13.72
6.19	12.75
6.56	13.27
5.94	12.17

Using the command lm(Ribeye ~ Corn) produces the following output.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.7478     2.7632  -0.633  0.56140
Corn          2.3437     0.4429   5.292  0.00612 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2601 on 4 degrees of freedom
Multiple R-squared:  0.875,  Adjusted R-squared:  0.8438
F-statistic: 28.01 on 1 and 4 DF,  p-value: 0.006119
```

Which of the following is the correct regression line?

a) Ribeye = -1.7478 + 2.3437 Corn
b) Ribeye = 2.3437 -1.7478 Corn
c) Corn = -1.7478 + 2.3437 Ribeye
d) Corn = 2.3437 -1.7478 Ribeye

Answer: a)

Problem 9.
Which of the following is **NOT** an assumption of simple linear regression?

A) The errors are normally distributed.
B) The relationship between *X* and *Y* is quadratic.
C) The errors have constant variance (homoscedasticity).
D) The errors are independent.

Answer: B) The relationship between *X* and *Y* is quadratic.

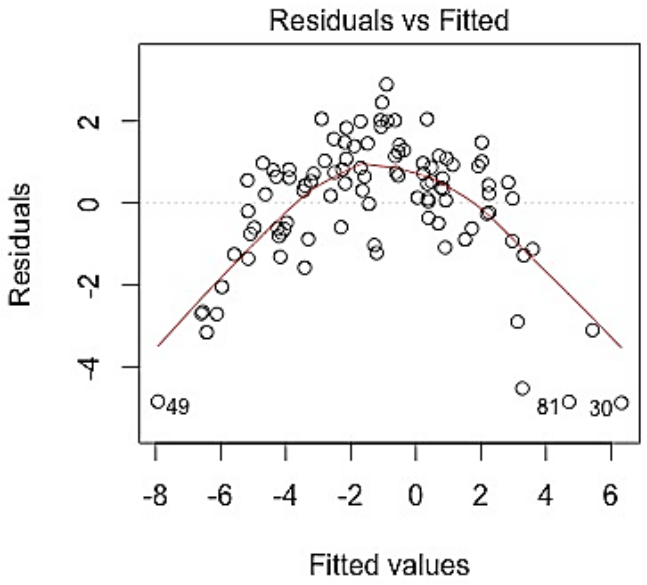
Problem 10.

If the estimated regression equation is $\hat{Y} = 5 + 3X$, and *X* increases by 2 units, what is the expected change in *Y*?

A) 3
B) 6
C) 11
D) 5

Answer: B) 6

Problem 11.
The following residual plot (residuals vs. fitted values) suggests:

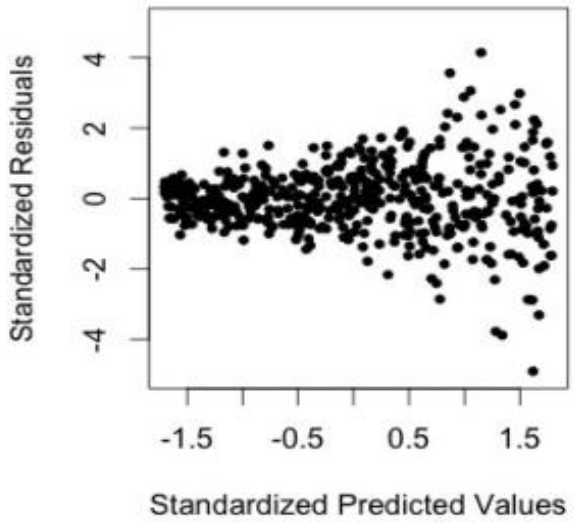


A) The model is mis-specified (nonlinear relationship).
B) The homoscedasticity assumption is satisfied.
C) There is multicollinearity.
D) The errors are autocorrelated.

Answer: A) The model is mis-specified (nonlinear relationship).

Problem 12

What problem does the following residual plot indicate?



A) Heteroscedasticity (non-constant variance).
B) Nonlinearity.
C) Autocorrelation.
D) Multicollinearity.

Answer: A) Heteroscedasticity (non-constant variance).

Problem 13.

The following screenshot is a least squares regression model.

```
Call:
lm(formula = Sales ~ Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
 -3385  -2097    258   1726   3034

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1383.4714  1255.2404   1.102    0.296
Spend         10.6222     0.1625  65.378 1.71e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2313 on 10 degrees of freedom
Multiple R-squared:  0.9977, Adjusted R-squared:  0.9974
F-statistic:  4274 on 1 and 10 DF,  p-value: 1.707e-14
```

Based on the information in the output. What is the correlation coefficient between spending and sales?

a) 0.9977
b) 0.9977^2
c) $\sqrt{0.9977}$
d) $-\sqrt{0.9977}$

Answer C

Problem 14.

Which is the best plot to assess the normality of residuals of a regression model?

a) Residual plot
b) Q-Q plot
c) Leverage plot (Cook's distance)
d) Scatter plot

Answer: b

Problem 15.

Cook's distance is used to:
A) Detect multicollinearity.
B) Measure the influence of individual data points on regression estimates.
C) Test for heteroscedasticity.
D) Assess model overfitting.

Answer: B) Measure the influence of individual data points on regression estimates.

Problem 16.

In a simple linear regression model, $Y=\beta_0+\beta_1 X+\epsilon$, what does β_1 represent?

A) The predicted value of Y when $X=0$.
B) The change in Y for a one-unit increase in X.
C) The error term.
D) The variance of X.

Answer: B) The change in Y for a one-unit increase in X.

Problem 17.

In R, which function is used to fit a simple linear regression model of `y` on `x`?
A) `lm(y ~ x)`
B) `fit(y, x)`
C) `regress(y, x)`

Answer: A) `lm(y ~ x)`

Problem 18.

After fitting a model `model <- lm(y ~ x)`, how do you extract the estimated coefficients?

A) `summary(model)`

B) coef(model)
C) model$coefficients
D) All of the above

Answer: D) All of the above

Problem 19.

Which function computes the 95th percentile of a t-distribution with 10 degrees of freedom?
A) qt(0.95, df = 10)
B) pt(0.95, df = 10)
C) dt(0.95, df = 10)
D) rt(0.95, df = 10)

Answer: A) qt(0.95, df = 10)

Problem 20.
What does **qnorm(0.975)** return?
A) The p-value for a z-score of 1.96
B) The z-score corresponding to the 97.5th percentile
C) The probability that a standard normal variable is less than 0.975
D) The 97.5th percentile of a t-distribution

Answer: **B) The z-score corresponding to the 97.5th percentile**

Problem 21.

How do you find the critical t-value for a 90% confidence interval with 25 degrees of freedom?
A) qt(0.95, df = 25)
B) pt(0.90, df = 25)
C) dt(0.90, df = 25)
D) rt(0.90, df = 25)

Answer: A) qt(0.95, df = 25) *(Because a 90% CI leaves 5% in each tail.)*

Problem 22.

Which function is commonly used to read a CSV file from a local drive in R?

A) read.table()
B) read.csv()
C) load()
D) read_excel()

Answer: B) read.csv()

Problem 23.

The URL of a comma-separated value (CSV) format data is at: https://pengdsci.github.io/STA200/dataset/diabetes-dataset.csv. Which of the following ways correctly imports the data to R?

A) load.table("https://pengdsci.github.io/STA200/dataset/diabetes-dataset.csv")

B) read.csv("https://pengdsci.github.io/STA200/dataset/diabetes-dataset.csv")

C) read.text("https://pengdsci.github.io/STA200/dataset/diabetes-dataset.csv")

D) import.cdv("https://pengdsci.github.io/STA200/dataset/diabetes-dataset.csv")

Answer: B

Problem 24.

Consider the following R code for a linear regression using artificial data.

```
# Create toy data
advertising <- c(1, 2, 3, 4, 5)  # Thousands of dollars
sales <- c(2, 4, 5, 4, 6)        # Thousands of units sold

# Fit model
model <- lm(sales ~ advertising)
model.summary <- summary(model)
```

Which of the given R commands extracts the following coefficients matrix **only**

```
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept)     1.8  0.9380832 1.918806 0.1508144
advertising     0.8  0.2828427 2.828427 0.0662756
```

a) coef(model)
b) model$coef
c) model.summary$coef
d) summary(model)

Answer: C

Problem 25.

The following R code builds a least squares regression model.

```
# Explicitly define height (cm) and weight (kg) pairs
height <- c(150, 155, 160, 165, 170, 175, 180, 185, 190, 195)
weight <- c(45, 52, 54, 60, 63, 68, 72, 77, 80, 85)
# Fit the linear model
M01 <- lm(weight ~ height)
M01.summary <- summary(M01)
```

Which of the following commands extracts R^2 (coefficient of determination)

a) M01$r.squared
b) M01.summary$r.squared
c) R.squared(M01)
d) R.squared(M01.summary)

Answer B

Problem 26.

In a study of reaction times, the time to respond to a visual stimulus (x) and the time to respond to an auditory stimulus (y) were recorded for each of 6 subjects. Times were measured in thousandths of a second. The results are presented in the following table.

The following computer output describes the fit of a linear model to these data. Assume that the assumptions of the linear model are satisfied.

The regression equation is
Auditory = 181.501124 + 0.408378 Visual

Predictor	Coef	SE Coef	T-statistic	P-value
Constant (b)	181.501124	21.861296	8.302395	0.00115
Visual (m)	0.408378	0.113544	3.596642	0.022827

What is the slope of the least-squares regression line?

A). 181.501124 B) 0.113544 C) 8.302395 D) 0.408378

Answer D.

Problem 27
Which of the following is true of R^2 ?

A). R^2 is also called the standard error of regression.
B). A low R^2 indicates that the Ordinary Least Squares line fits the data well.
C). R^2 usually decreases with an increase in the number of independent variables in a regression.
D. R^2 shows what percentage of the total variation in the dependent variable, Y, is explained by the explanatory variables

Answer D

Problem 28.
A regression analysis between sales (in $1000) and price (in dollars) resulted in the following equation:

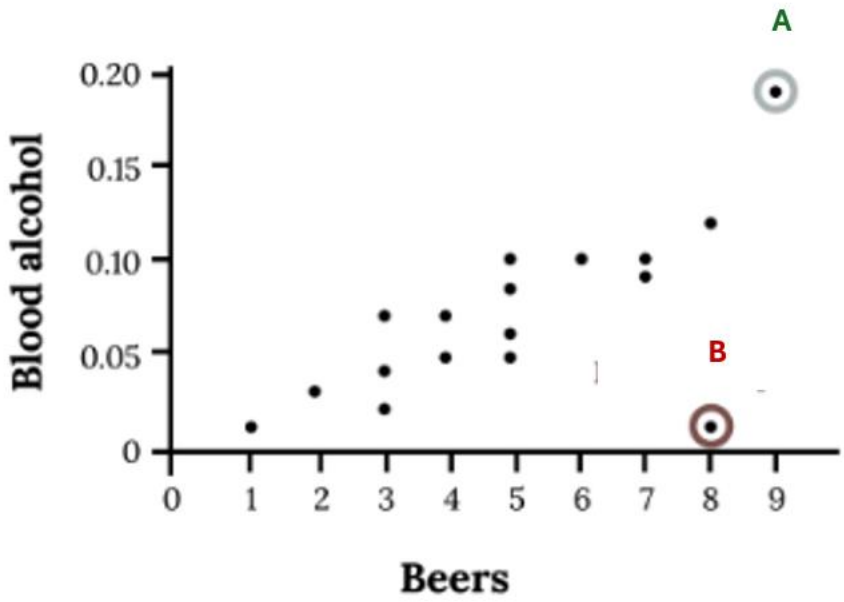
$$\hat{y} = 50{,}000 - 8x$$

The above equation implies that an
A). increase of $1 in price is associated with a decrease of $8000 in sales
B). increase of $8 in price is associated with an increase of $8,000 in sales
C). increase of $1 in price is associated with a decrease of $42,000 in sales
D). decrease of $1 in price is associated with a decrease of $8000 in sales

Answer A

Problem 29

Which of the following statements is correct based on the following plot?



a) A is an outlier and influential.
b) B is an outlier and influential.
c) A is not an outlier and not influential
d) B is not an outlier but influential

Answer: B

Summary of Weekly Exam #1

The class boundary is: 70,80,90,100

cut.data.freq	Freq	midpts	rel.freq	cum.freq	rel.cum.freq
[70,80]	2	75.00	0.12	2	0.12
(80,90]	8	85.00	0.47	10	0.59
(90,1e+02]	7	95.00	0.41	17	1.00