Name: _____
Score: _____ / _____

Final exam

Part 1: True or False

1

Suppose we want to construct a confidence interval for the true population mean μ using a t-distribution with 9 degrees of freedom: $\bar{x}\pm t_9^*\cdot\text{SE}$. Consider the following two events:

- A: A 95% confidence interval constructed using the above formula contains μ.
- B: A 90% confidence interval constructed using the above formula contains μ.

Then we can conclude that

$$P(A\mid B) = P(A),$$

or in other words, A and B are independent events.

- True
- False

Answer Point Value: 4.0 points
Answer Key: False

2

The binomial distribution can be approximated by a normal distribution more accurately as n, the total number of Bernoulli trials, increases. Similarly, the distribution of a sample proportion approaches the normal distribution as n increases.

- True
- False

Answer Point Value: 2.0 points
Answer Key: True

3

An economist is studying economic development in different parts of the world, measured by Gross Domestic Product ($\texttt{GDP}$). They plan to report their final results **using 95% confidence intervals (or, at the 5% significance level)**. During model selection, they consider two predictors: median household income ($\texttt{HHI}$) and unemployment rate ($\texttt{UR}$) and fit the following multiple linear regression:

$$\widehat{\texttt{GDP}} = \beta_0 + \beta_1\cdot\texttt{HHI} + \beta_2\cdot\texttt{UR}$$

The model output in $\texttt{R}$ produces the following excerpted table:

Predictor	P-value
$\texttt{HHI} $	0.03
$\texttt{UR}$	0.02

Using the backward elimination p-value approach, they should remove $\texttt{HHI}$ from her model and refit her model using only $\texttt{UR}$ as a predictor because it has the largest p-value of all predictors.

- True
- False

Answer Point Value: 4.0 points
Answer Key: False

4

The p-value quantifies the strength of the data as evidence in support of the alternative hypothesis. A smaller p-value implies stronger evidence that the data provide evidence in favor of the alternative hypothesis.

- True
- False

Answer Point Value: 2.0 points
Answer Key: True

5

When comparing the means of two groups, given a fixed dataset, a two-sample t-test and an ANOVA will give us the same result, if we use a pooled standard error when calculating the test statistic.

- True
- False

Answer Point Value: 2.0 points
Answer Key: True

6

The margin of error of a confidence interval would increase if the sample size increases.

- True
- False

Answer Point Value: 2.0 points
Answer Key: False

7

A χ^2-statistic follows a right-skewed distribution and can take on negative values.

- True
- False

Answer Point Value: 2.0 points
Answer Key: False

Part 2: Fill in the blank

8

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation).
NOTE: For scientific notation, a period MUST be used as the decimal point marker.

A geneticist is studying a particular gene that determines hair color in humans, and she knows that there are 3 possible genotypes (combinations of genetic material) an individual can have. She denotes them HH, Hh, and hh. She is preparing for a presentation in which she compares her expected distribution against the observed distribution using a chi-square test. She obtains a χ^2-statistic of 2. In a hurry, she forgot to include the number of observed individuals in each category on her slides to demonstrate the calculation. However, she remembers the following facts:

- The observed number of individuals who have Hh and hh are each equal to their expected numbers.
- The number of expected individuals for HH is 50, and she observed more than she expected.

How many individuals did she observe who had HH? ____ (Note: the answer should be an integer.)

Answer Point Value: 4.0 points
Answer Key: 60

9

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation).
NOTE: For scientific notation, a period MUST be used as the decimal point marker.

[NOTE: this is an extra-credit question.]
A researcher wants to know if high-sugar diets lead to overweight. He randomly assigns guinea pigs into a control group and a treatment group, where the control group receive a regular diet and the treatment group receive a high-sugar diet. Suppose the body weight of guinea pigs follows a normal distribution with a standard deviation of 70 grams. If the researcher wants to detect a weight difference of 20 grams and sets the significance level at $\alpha = 0.05$, then he will need at least ____ guinea pigs **in each group** to achieve a desired power level of 80%. (Please write down an **integer** for the sample size.)

Answer Point Value: 4.0 points
Answer Key: 191|195

10

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation).
NOTE: For scientific notation, a period MUST be used as the decimal point marker.

For the follow-up tests of the results of an ANOVA we use a modified significance level α^* to adjust for multiple comparisons. If we want an overall type 1 error rate of 1%, then the significance level α^* should be ____ for the individual pairwise tests if there are 4 groups. (Please enter a decimal number and round to 4 digits after the decimal point.)

Answer Point Value: 3.0 points
Answer Key: 0.0015|0.0018

11

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation).
NOTE: For scientific notation, a period MUST be used as the decimal point marker.

In the Rock-Paper-Scissors (RPS) game, a player can take three actions in one round of the game: Rock, Paper, and Scissors. In a study about strategies of the RPS game, 72 volunteers **each** played 300 rounds of the game (that is, $72 \times 300 = 21600$ total rounds were played) and their adopted actions in each round were recorded. Below is the table for the average relative frequencies of their adopted actions:

	Rock	Paper	Scissors
Average frequency	0.35	0.33	0.32

The researchers suspect that the true probabilities of actions adopted in a round of the RPS game are not equal. Assuming that all volunteers' actions are independent from each other and their actions are also independent across the rounds, if we use a chi-square test to test for their suspicion, then the χ^2 statistic is ____ (round to 2 digits after the decimal point), and we should compare this statistic with a χ^2 distribution with df = ____ to compute the p-value.

(Note: this question is based on a published study conducted by Wang et. al. .)

Answer Point Value: 4.0 points
Answer Key: 30.22|30.26, 2

12

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation).
NOTE: For scientific notation, a period MUST be used as the decimal point marker.

Suppose we want to compare the average vitamin C concentration levels (in mg/100mL) of orange juice for two brands, A and B. Below is a summary table of vitamin C concentration levels from a sample of 40 bottles of orange juice from these two brands.

	Brand A	Brand B
sample mean ($\bar x$)	35.2	37.1
standard deviation (s)	2.9	3.3
sample size (n)	20	20

If we use a two sample t-test to compare the mean vitamin C levels of these two brands of orange juice, the t statistic will be ____, and the p-value will be ____. (Please round to 3 digits after the decimal point for both blanks).

Answer Point Value: 4.0 points
Answer Key: 1.930|1.938, 0.066|0.070

13

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation).
NOTE: For scientific notation, a period MUST be used as the decimal point marker.

Duke's COVID-19 surveillance testing protocol involves "pool testing". Samples from 5 individuals are combined into a single mixture (the "pooled sample") and sent to the lab for the COVID-19 nucleic acid test. The pooled sample will test positive if **at least one** individual among the five is COVID positive.
Suppose that the probability for one individual at Duke to test positive is 15%, and assume that all the individual samples are independent of one another. Then the probability that a pooled sample will test positive is ____.
(Please write down a decimal number, and round to **3 digits** after the decimal point.)

Answer Point Value: 3.0 points
Answer Key: 0.545|0.565

Part 3: Multiple choices

14

After completing STA 101, you and a lab partner decide to revisit the $\texttt{ames}$ data set from some earlier labs, which contains data about house sales in Ames, Iowa. You decide to construct a simple linear regression that explains the sale prices of houses ($\texttt{price}$) using the square footage of each house ($\texttt{area}$):

$$\widehat{\texttt{price}} = \beta_0 + \beta_1\cdot\texttt{area}.$$

You find in $\texttt{R}$ that the estimate for β_1 is 111.7. Your partner makes the following statements:

- **Statement 1**: "Our model says that each additional square foot in a house increases the expected price by $111.7."
- **Statement 2**: "If our hypotheses are $H_0: \beta_1=0$ and $H_A: \beta_1\neq 0$ and set $\alpha=0.05$, then the probability we fail to reject H_0 when H_A is actually true is 5%."

Which of these two statements is(are) TRUE?

- A. Only Statement 1.
- B. Only Statement 2.
- C. Both statements.
- D. Neither statement.

Answer Point Value: 4.0 points
Answer Key: A

15

Which of the following statements are **incorrect** about the power of a hypothesis test? (Select all incorrect statements)

- [ ] A. Increasing the sample size doesn't help increase the power.
- [ ] B.
When we perform hypothesis testing, it is less important to ensure a high power level than to ensure a low significance level.
- [ ] C.
To increase power, we need to consider a smaller effect size.
- [ ] D.
For a particular hypothesis test, if the Type 2 error rate is 0.1, then the power is 90%.
- [ ] E.
To increase power, we can increase the significance level α, thus allowing a higher Type 1 error rate.

Answer Point Value: 4.0 points
Answer Key: A,B,C

16

The "*trimmed mean*" is a statistical measure of central tendency, much like the mean and median. It involves the calculation of the mean after discarding given parts of the sample at the high and low ends. For example, the 10% trimmed mean is calculated after discarding the bottom and top 5% of the sample data. Based on this information, how does the robustness of the trimmed mean to outliers and extreme skew compare to that of the (regular) mean?

- A.
Trimmed mean is more robust than regular mean.
- B.
Trimmed mean is as robust as regular mean.
- C.
Cannot tell from the information given.
- D.
Trimmed mean is less robust than regular mean.

Answer Point Value: 4.0 points
Answer Key: A

17

Consider a simple linear regression model that predicts a response variable Y with an explanatory variable X. About the R^2 of this linear model, which of the following statements are correct? (Select ALL correct ones.)

- A. If R^2 is high, then Y and X are highly and positively correlated.
- B. If $R^2 = 0.6$, then the linear model can predict the value of Y correctly 60% of the time.
- C. If $R^2 = 0.6$, then 60% of the variability of Y can be explained by the model.
- D. If $R^2 = 0.6$, then 60% of the variability of X can be explained by the model.
- E. If a leverage point is added to the data, and a least squares regression line is re-fitted, then the value of R^2 may still be the same as before.

Answer Point Value: 4.0 points
Answer Key: C,E

18

A random sample of 200 runners who completed a 5-mile run yielded an average finishing time of 45 minutes. A 95% confidence interval for the 5-mile run time calculated based on this sample is 40 to 50 minutes. Which of the following statements are FALSE? (Select ALL incorrect ones.)

- A. The margin of error for this confidence interval is 10 minutes.
- B. If we calculate a 90% confidence interval for the average finishing time of this 5-mile run based on the same sample, and let the confidence interval be (L, U). Then we have $40 < L < U < 50$.
- C. We are 95% confident that the true average finishing time for all runners who completed this 5-mile run is between 40 to 50 minutes.
- D. If we collect 100 random samples of 200 runners who completed this 5-mile run and calculate the average finishing time in each sample, then approximately 95 of those averages will be between 40 to 50 minutes.

Answer Point Value: 4.0 points
Answer Key: A,D

19

Match the appropriate methods of analysis to the types of data and applications.

1. z-test for single proportions	A. Lindsay wants to compare the quality of burritos of two local restaurants. She invites 100 volunteers to taste the burritos offered by the two places. Each volunteer tries the signature burrito of the first restaurant and then tries the signature burrito of the second restaurant; after that, each volunteer will give a score (from 0 to 100) for the quality of each burrito that he has tasted. Lindsay then analyzes the scores from those 100 volunteers to compare the quality of burritos offered by these two places.
2. z-test for comparing two proportions.	B. Luna wants to know if urban and rural residents have different political ideologies. She surveys 200 representative residents from urban and from rural areas each and asks about their views on political issues to classify them into 3 categories of political ideology: conservative, moderate, liberal. She then uses the data to test for whether or not there is association between the residential area and political ideology among American people.
3. t-test for paired data	C. Leon wants to know if a die he has is fair. He throws the die 2000 times and counts the frequency that each number (1, 2, 3, 4, 5, or 6) shows on top. He then analyzes the data to test if the chance that each number shows on top is equal.
4. two sample t-test	D. We want to know if the slope of a simple linear regression line is 0.

5. chi-square test for goodness-of-fit

6. one sample t-test

7. chi-square test for independence

8. ANOVA

E. Lisa wants to know if a coin she has is fair. She tosses the coin 500 times and finds that for 220 times the coin has heads up. She wishes to use such data to test whether or not the chance that this coin has heads up is 50%.

F. Lucas wants to know if the poverty rate is different between the White and non-White communities in his town. He surveyed 200 White and non-White residents and asked about their annual household income in order to calculate the proportion of residents under the poverty line for each racial community. Lucas then compares the sample proportions of residents in poverty between the two communities.

G. None of the Above

Answer Point Value: 6.0 points
Answer Key: 1:E, 2:F, 3:A, 4:G, 5:C, 6:D, 7:B, 8:G

20

In a study on whether or not meditation can help treat insomnia, researchers *randomly* divided 400 people into two equal-sized groups. One group meditated daily for 30 minutes, the other group attended a 2-hour information session on insomnia. At the beginning of the study, the average difference between the number of minutes slept between the two groups was about 0. After the study, the average difference was about 24 minutes, and the meditation group had a higher average number of minutes slept.

To test whether an average difference of 24 minutes could be attributed to pure chance, a statistics student decided to conduct a randomization test. She wrote the number of minutes slept by each subject in the study on an index card. She shuffled the cards together very well, and then dealt them into two equal-sized groups.

Select ALL statements that are TRUE:

- [ ] A. If she repeats the card-shuffling many times, the differences between the two stacks of cards will be centered around 24 approximately.
- [ ] B. If meditation does have an effect, the average the average difference between the two stacks of cards will be more than 24 minutes.
- [ ] C. This study is an observational study, so no causal conclusions can be made.
- [ ] D. If the statistics student can show that the difference of 24 minutes can NOT be attributed to only chance, then she can conclude that meditation does help treat insomnia.
- [ ] E. If she repeats the card-shuffling many times, the differences between the two stacks of cards will be centered around 0 approximately.

Answer Point Value: 6.0 points
Answer Key: D,E

Part 4: ANOVA

21

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation).
NOTE: For scientific notation, a period MUST be used as the decimal point marker.

The "iris" dataset contains measurements for 50 flowers from each of 3 species of iris (150 observations in total). Suppose we want to see if the average sepal lengths of the 3 species are different, and we conduct an ANOVA for such comparison. Below is a partial ANOVA table:
Response variable: Sepal.Length

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species	a	63.212	31.606		
Residuals	b		0.265		
(Total)		102.168			

Given that there are 3 species and 150 total observations, the degrees of freedom in the cell of (a) should be ____, and the degrees of freedom in the cell of (b) should be ____.

Answer Point Value: 4.0 points
Answer Key: 2, 147

22

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation).
NOTE: For scientific notation, a period MUST be used as the decimal point marker.

The "iris" dataset contains measurements for 50 flowers from each of 3 species of iris (150 observations in total). Suppose we want to see if the average sepal lengths of the 3 species are different, and we conduct an ANOVA for such comparison. Below is a partial ANOVA table:

Response variable: Sepal.Length

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species		63.212	31.606		
Residuals		(a)	0.265		
(Total)		102.168			

The sum of squares error - the sum of squares in the cell of (a) - should be ____. (Please keep 3 digits after the decimal point.)

Answer Point Value: 3.0 points
Answer Key: 38.950|38.965

23

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation).
NOTE: For scientific notation, a period MUST be used as the decimal point marker.

The "iris" dataset contains measurements for 50 flowers from each of 3 species of iris (150 observations in total). Suppose we want to see if the average sepal lengths of the 3 species are different, and we conduct an ANOVA for such comparison. Below is a partial ANOVA table:

Response variable: Sepal.Length

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species		63.212	31.606		
Residuals			0.265		
(Total)		102.168			

The F value should be ____. (Please round to 3 digits after the decimal point.)

Answer Point Value: 3.0 points
Answer Key: 119.266|119.269

The "iris" dataset contains measurements for 50 flowers from each of 3 species of iris (150 observations in total). Suppose we want to see if the average sepal lengths of the 3 species are different, and we conduct an ANOVA for such comparison. Below is a partial ANOVA table:

Response variable: Sepal.Length

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species		63.212	31.606		
Residuals			0.265		
(Total)		102.168			

Given the degrees of freedom and F statistic value that you have calculated in previous questions, the p-value is

- A. Below 0.0001.
- B. Between 0.0001 and 0.01.
- C. Between 0.01 and 0.05.
- D. Between 0.05 and 0.95.
- E. Between 0.95 and 0.9999.
- F. Above 0.9999.

Answer Point Value: 3.0 points
Answer Key: A

25

A winery in Sonoma Valley, CA is reviewing data on its most recent yield, taking into account various quality factors such as alcohol content, hue, and alkalinity levels. You have been hired as an analyst to run statistical tests on their data for quality control purposes.
The winery wants to ensure consistent magnesium levels across its three most popular varieties, labelled as A, B, and C. Let μ_A, μ_B,μ_C denote the average magnesium level of the three varieties, and you are asked to conduct an ANOVA to test for difference between the magnesium levels. Which of the following statements are correct? (Select ALL that are correct.)

- [ ] A.
The alternative hypothesis should be: $\mu_A \neq \mu_B \neq \mu_C$.
- [ ] B.
The alternative hypothesis should be: at least one of $\mu_A, \mu_B, \text{ and } \mu_C$ is different from others.
- [ ] C.
When performing ANOVA, we are testing if the variability of magnesium levels between different varieties is significantly larger than the variability of magnesium levels within the same varieties.
- [ ] D.
We conduct follow-up tests if we reject the null hypothesis in ANOVA. Then in each follow-up test, we can do a two-sample t-test to compare the mean magnesium levels between each pair of two varieties.
- [ ] E. Performing multiple follow-up tests after ANOVA may blow up the Type 2 error rate.

Answer Point Value: 4.0 points
Answer Key: B,C,D

Following an international standard on nutrition label requirements, the winery (that you are working as an analyst for) has set ranges for designating their wine's level (low, moderate, or high) of flavonoids, a class of phytonutrients. The winery plans to market and label their "high flavonoid" varieties as antioxidant-rich, but wants to ensure that their proanthocyanin levels do not vary within flavonoid categories. You run an ANOVA on a dataset with **130** sampled wines, with the null hypothesis that proanthocyanin levels are the same across the three (3) flavonoid categories, and the following table is the partial output you get:

	Df	Sum Sq	Mean Sq	F value	P-value
flavonoid category		9.427	4.713	21.28	\(1.07 \times 10^{-8}\)
Residuals		28.130	0.221		

Choose ALL statements that are correct:

A.
- [ ] The between-group degrees of freedom is 2, and the within-group degrees of freedom is 129. So we should compare the F statistic with an F distribution with \(df_1 = 2\) and \(df_2 = 129\).

B.
- [ ] The between-group degrees of freedom is 2, and the within-group degrees of freedom is 127. So we should compare the F statistic with an F distribution with \(df_1 = 2\) and \(df_2 = 127\).

C.
- [ ] With significance level \(\alpha = 0.05\), there is **no** convincing evidence to show that there is a difference in the mean proanthocyanin levels across the three flavonoid levels. Our conclusion would change if we adopted a significance level of \(\alpha = 0.01\).

D.
- [ ] With significance level \(\alpha = 0.05\), we would reject the null hypothesis if the F value exceeds 3.0675 (this cutoff value is rounded to 4 digits after the decimal point).

Answer Point Value: 4.0 points
Answer Key: B,D

Part 5: Linear regression - candy matchups

After being hired by Mars, the leading producer of chocolate and other popular confections, you are asked to investigate the candy preferences of American consumers. You randomly sample 9,500 people and ask them to vote on randomly generated matchups between all of the different popular candy types in the United States. In determining potential factors that boost the popularity of Mars' candies, you isolate several variables:

- **Chocolate** ("\(\texttt{chocolate}\)"): Does it contain chocolate (TRUE/FALSE)
- **Bar** ("\(\texttt{bar}\)"): Is it a candy bar? (TRUE/FALSE)
- **Sugar percentile** ("\(\texttt{sugarpercent}\)"): The percentile of sugar content it falls compared to the rest of the candies
- **Price percentile** ("\(\texttt{pricepercent}\)"): The percentile of unit price compared to the rest of the candies.

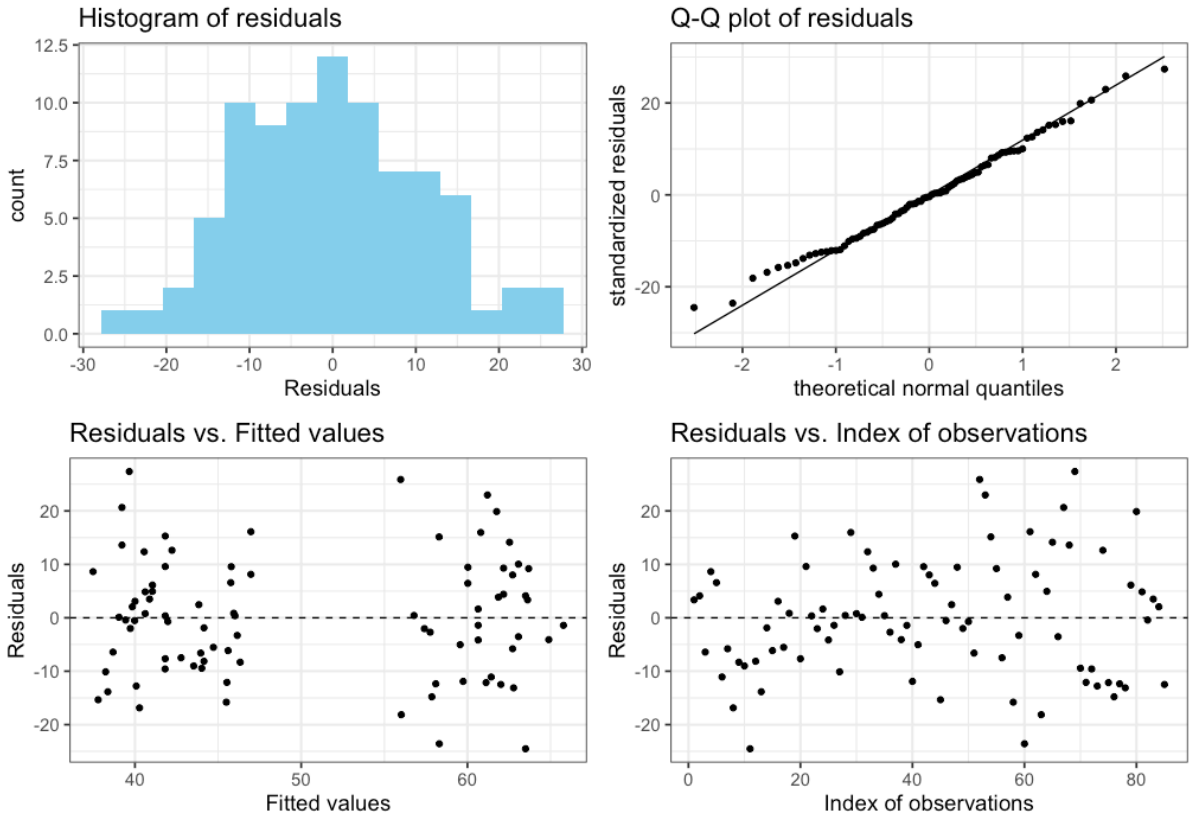
You fit a multiple linear regression model for predicting the percentage of each candy's “won” matchups based on these given variables (a higher "win percentage" means higher popularity level of a candy). Some of the model output is attached below.

Attachments

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    38.932      2.840  13.707  < 2e-16 ***
chocolateTRUE  17.584      3.222   5.458 5.26e-07 ***
barTRUE         2.940      3.743   0.786   0.4344
sugarpercent    9.255      4.646   1.992   0.0498 *
pricepercent   -3.041      5.571  -0.546   0.5867
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The attached diagnostic plots have been produced out of the modeling results. Based on these plots, which of the conditions for multiple linear regression are clearly violated?

Attachments



- A.
  Constant variance of residuals
- B.
  Normality
- C. Linearity
- D.
  Independence of residuals
- E.
  None of the conditions are violated

Answer Point Value: 3.0 points
Answer Key: E

28

For which of the following variables, the 95% confidence interval for its slope would include 0? Select ALL that apply.

- [ ] A.
sugarpercent
- [ ] B.
pricepercent
- [ ] C.
bar
- [ ] D.
chocolate

Answer Point Value: 4.0 points
Answer Key: B,C

29

The R^2 of this model is about 0.437. Which of the following statements are correct? (Select ALL that are correct)

- [ ] A.
About 43.7% of the variability of the matchup win percentage can **not** be explained by this model.
- [ ] B.
About 43.7% of the variability of the matchup win percentage can be explained by this model.
- [ ] C.
The adjusted R^2 of this model should be smaller than 0.437.
- [ ] D. If we include more explanatory variables in a linear model to predict win percentage, then the R^2 of this new model would be ≥ 0.437.
- [ ] E.
The correlation between win percentage (the response variable) and each of the explanatory variables is $\sqrt{0.437} \approx 0.661$.

Answer Point Value: 4.0 points
Answer Key: B,C,D

Based on the model ouput, which of the following statements is **correct** about the estimated slope for "chocolate"?

A.
When a candy contains chocolate, we expect it to win approximately 17.58% more matches against other candy types compared to those without chocolate on average, all else held constant.

B.
When a candy does **not** contain chocolate, we expect it to win approximately 17.58% more matches against other candy types compared to chocolate candy on average, all else held constant.

C.
The average percent of matches won by chocolate candy is 17.58 points higher than non-chocolate candy.

D.
The average percent of matches won by chocolate candy is 17.58 points lower than non-chocolate candy.

Answer Point Value: 4.0 points
Answer Key: A