Test 1

STAT 021

Swarthmore College

Do not flip this page until instructed to do so.

Test organization: There are 9 questions in total on this test and they are organized into two subsections: the first **5** questions are select all that apply multiple choice questions and the last **4** questions are free response. If you need additional paper you may come to the front of the class and pick some up. There are a total of 30 points possible on this test.

Instructions: The first part of this test are multiple choice questions that do not require any additional explanation or work. No extra work will be considered in the grading of these questions but *you can get partial credit* for many of these questions. The last part of this test involves short answer questions. For these questions, you must show all your work and/or provide enough justification and explain your reasoning in order to get full credit or be considered for partial credit. You do not need a calculator to evaluate any expressions. For any calculation problems, simply writing out the formula to find the answer will suffice.

First and Last Name: Xinxin Li

Swarthmore Username: xli5

Take a deep breath.

You have prepared for this test and with a clear and well-rested mind, you are ready to show me what you have learned this semester. As with the other tests, the purpose of this test is to measure your understanding of the material we have covered. This is nothing more than a metric for me to evaluate your preparedness to think statistically at this particular moment in time and in this particular setting. This is not a perfect measure of your knowledge and does not predict your future statistical skills.

Multiple choice problems (2 points each)

1

The average prevalence of lung cancer across the entire state is 31.6 cases per 100,000 individuals. A doctor wants to determine how closely the prevalence of lung cancer for her patients from a rural area of that state aligns with the state wide average. She calculates a 90% confidence interval of $[37.2, 40.1]$ for the average prevalence (per 100,000 individuals) of lung cancer based on a random sample of her patients and patients from nearby doctors over the last three years. Which of the following statements are valid interpretations of this confidence interval? Circle all that apply.

(a) This confidence interval indicates that the prevalence of lung cancer in this area is likely much higher than the state-wide average.

(b) If another doctor were to conduct the same study on a new random sample of patients from the region, there is a 90% chance that he would calculate the same confidence interval.

(c) If this study were to repeated with new random samples, the resulting confidence intervals would contain the true prevalence of lung cancer in this area 90% of the time.

(d) There is a 10% chance that the interval $[37.2, 40.1]$ does not capture the true prevalence of lung cancer (per 100,000 individuals) for this area.

2

A candidate running for mayor of a town in Minnesota suspects that there is evidence of racial bias in his town and state judicial system. He investigates the incarceration rates of white and non-white offenders finding a p-value of 0.003 when testing the null hypothesis that the incarceration rates of white offenders is the same as that of non-white offenders vs the alternative that the incarceration rates of white offenders is smaller than that of non-white offenders. Which of the following statements are valid interpretations of this p-value? Circle all that apply.

(a) There is a large difference between the incarceration rates of white offenders and non-white offenders.

(b) The difference between the incarceration rates of white offenders and non-white offenders is likely not due to random chance.

(c) The probability of a randomly selected white offender being incarcerated is 0.003 times smaller than the probability of a non-white offender being incarcerated.

(d) For randomly selected offenders in this state, if the incarceration rates are equal, then the probability we observe a difference in incarceration rates as large or larger than observed in this data set is 0.003.

x/15

3

Which of the following statements are true? Circle all that apply.

(a) For the same value of the predictor, the 95% prediction interval for a new observation is always wider than the 95% confidence interval for the mean response.

(b) The variability due to error (SSE) is always smaller than the variation explained by the model (SS_{Mod}).

(c) If (absolute value of) the correlation between a predictor X_1 and Y is greater than the (absolute value of) the correlation between another predictor, X_2, and Y, then the coefficient of determination for regressing Y on X_1 is greater than the coefficient of determination for regressing Y on X_2.

4

Which of the following conditions for inference in regression does a fitted-values vs. residual plot aid in assessing? Circle all that apply.

(a) Linearity

(b) Constant variance

(c) Independence

(d) Normality

5

In which of the following situations do we need to asses the normality and randomness conditions for inference in regression? Circle all that apply.

(a) A scientist studying a random sample of sparrows at Kent Island wants to determine how the average weight (in grams) of the birds changes for each additional mm in wing length.

(b) Someone interested in selling their house creates a SLR model using the list price of local houses as a predictor of the final sale price of each house once sold. They want to predict the actual sale price of their home for a list price of \$189,000.

(c) A nutritionist wants to determine if an increase in sugar content (in grams) of breakfast cereal corresponds to a positive (non-zero) change in calories per serving while also obtaining an upper and lower bound on the size of the increase.

(d) A used car salesperson wants to determine if an increase in the number of miles on a car has a statistically significant impact on the sale price of the vehicle.

4. for a increase in payment from own account of 1, the APR is $\hat{\beta}_1$, on average, expected to be changed by $\hat{\beta}_1$. We can also use confidence interval for mean response and prediction interval for a specific predictor value to make inference.

Short answer questions

6 (4 points)

We've discussed the construction of statistical models as following four general steps:

1. Choose which relationship to model.
2. Fit the data to your chosen model.
3. Assess the fit of the model.
4. Use the estimated model to answer statistical questions.

3. Use ANOVA F test to check the overall fitness of $H_0: \beta_1 = 0$; $H_A: \beta_1 \neq 0$, where β_1 is the true slope for the SLR model. If p-value < 0.001, then we reject null hypothesis and there is statistically significant evidence to say that there is relationship between y and x.

Credit card companies are allowed to change their annual percent interest rate (APR) depending on different factors than involve your personal spending habits and decisions made by the Federal Reserve. Suppose you are interested in exploring a linear relationship between payments made on your credit card account and the APR at the time of your payment. Describe what you would do at each of the four modeling steps above in no more than 1-3 sentences per step. Make sure to provide some justification/explanation for each activity.

1. Since I'm interested in exploring a linear relationship between payments made on my own credit card account and APR at the time of payment, I would choose to use simple linear regression model. The predictor variable, which is the independent variable, would be the payments made on my own credit card account. The response variable is the APR at the time of payment.

2. use formula or R to calculate the model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

where $\hat{y}_i$ is the predicted APR at the time of payment for a x_i, and x_i is the payments made on credit card account.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad ; \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$\bar{x}$ is the average of x_i

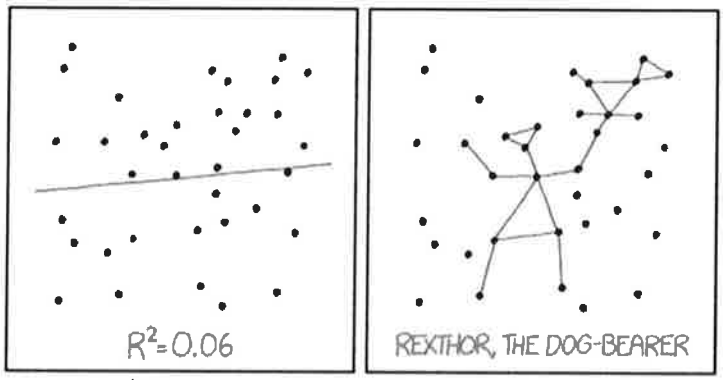
$\bar{y}$ is the average of y_i

1. Use residual vs. fitted value plot to check linearity and constant variance; use normal quantile plot to check normality; If the points on both plot fit closely with the predicted line, then there is no issue. Check independence of error and randomness of data collecting, if the data are from SRS, then it's likely that there is no issue, and mean for error should be zero.

x175

7 (4 points)

Based on the data shown in the scatter plot of this comic[1], what can you tell me about the relationship between the SS_{Mod} and SSE terms?



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

response

SS_{mod} is for SLR model. It checks how the observed value is far from the average of observed response value.

response

SSE ~~is~~ checks the difference between the ~~the~~ observed response value and the predicted response value based on the ~~use~~ model chosen.

8 (6 points)

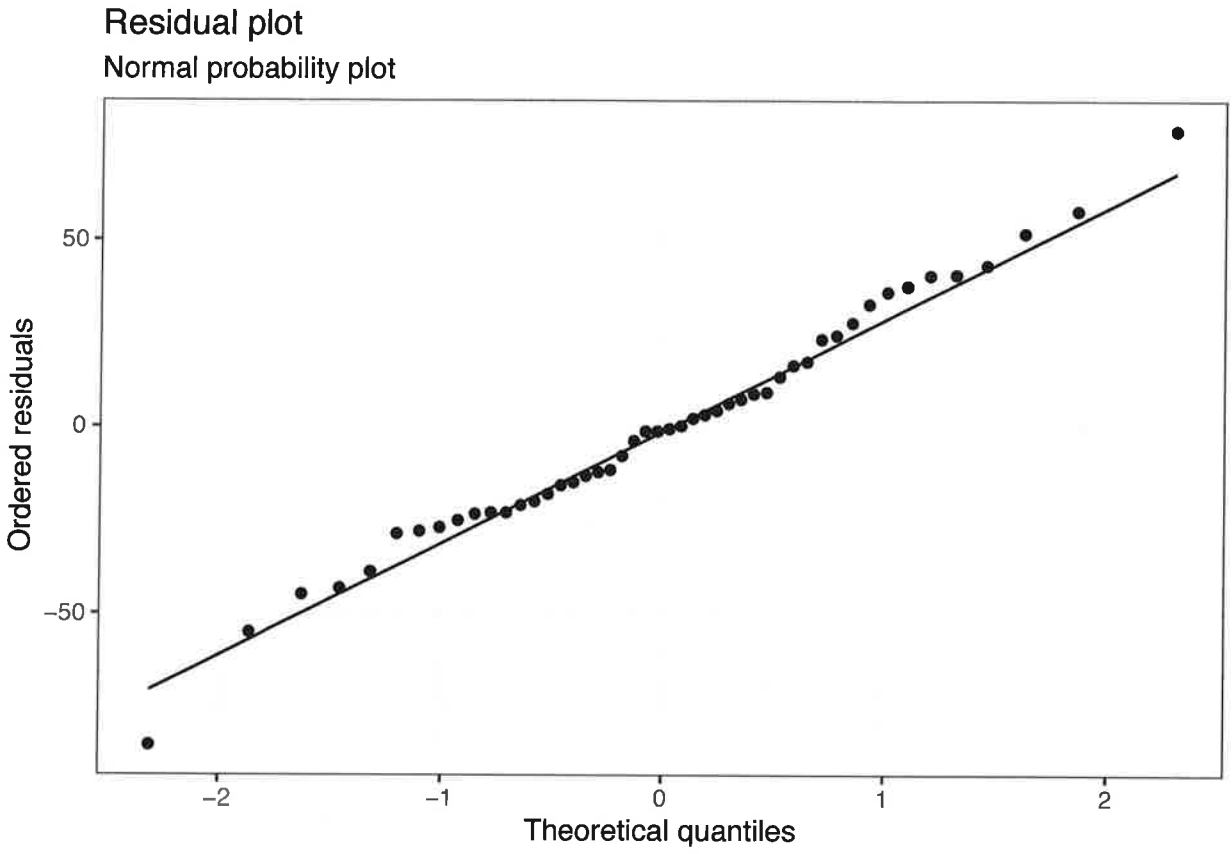
The data below represents a simple random sample of Singaporean diamonds. Suppose we are interested in how the size of the diamond (in carats) can predict (or explain) what the cost of the diamond will be. Below is the R code for fitting this simple linear regression model.

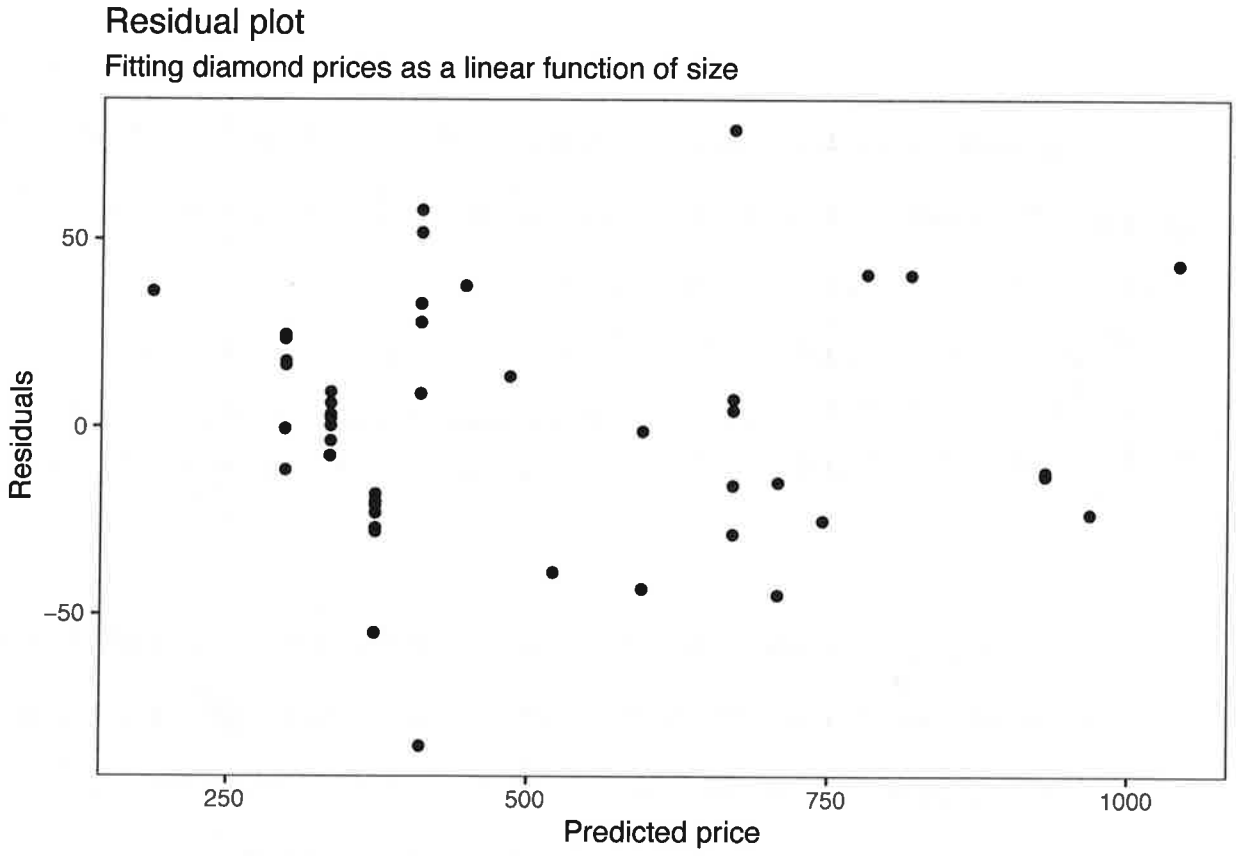
```
diamond_mod <- lm(price~size, data=diamond_dat)
diamond_mod %>% summary
```

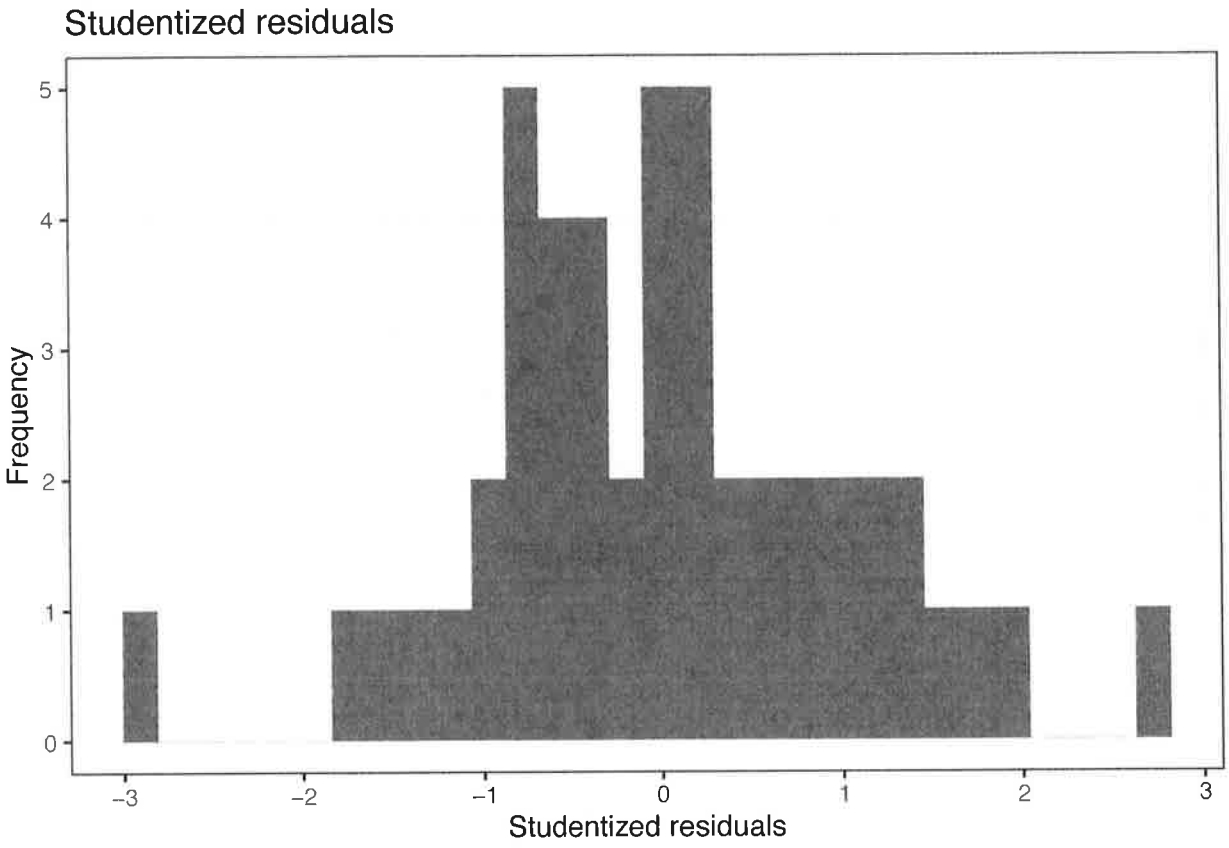
```
##
## Call:
## lm(formula = price ~ size, data = diamond_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.159 -21.448  -0.869  18.972  79.370
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -259.63      17.32  -14.99   <2e-16 ***
## size         3721.02      81.79   45.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.84 on 46 degrees of freedom
## Multiple R-squared:  0.9783, Adjusted R-squared:  0.9778
## F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

Analyse the following plots based on this regression model to answer this two-part question.

x/15


Residual plot
Normal probability plot
Ordered residuals
50
0
−50
−2
−1
0
1
2
Theoretical quantiles


Residual plot
Fitting diamond prices as a linear function of size
Residuals
50
0
−50
250
500
750
1000
Predicted price



a) Based on these plots, what conclusions can we make about the conditions necessary for inference with a SLR model? No obvious outliers from standardized residuals.

The points in normal probability plot lie closely ~~a~~ to the predicted line that ~~a normal data~~ data with normal ~~distributed~~ error should lie on, so the normality condition is ~~not~~ met.

The points in ~~residual~~ residual vs. fitted value plot scatter randomly above and below the zero residual line, and there's no obvious curvature, so linearity and constant variance is met. We can assume the data are from SRS, so independence is met.

b) Say instead of the size of the diamond measured in carats, we'd like to look at the size in grams (1 carat = 0.2 grams). Would we expect the behavior of any of the plots above to change? Briefly explain your answer.

No. All x and y values will be scaled ~~by~~ up by 5. ~~But~~ Since the observed response value and predictor value are still mapped in 1 to 1, and the observed response value, which is the prize, changes at the same scale as x since they follow SLR model.

Their spread and distribution does not change. Since y and x are changed by the same scale, the ~~et~~ distribution of error also does not change

x/7 5

9 (6 points)

The data that appear in the data set "Four-Mile-Run-data.txt" were collected by a GPS watch worn by the runner of a four-mile course. Using heart rate measurements after each run, an analysis of the runner's post-exercise heart rate recovery provides an indication of cardiovascular fitness. Suppose we are interested in determining if the speed of the run (in mph) is related to the number of calories burned. Below is the R code and output for fitting such a linear model to this data.[2]

```
run_dat <- read_table2("Four-Mile-Run-data.txt")
run_reg <- lm(calories~aveSpeed, run_dat)
run_reg %>% summary
```

```
##
## Call:
## lm(formula = calories ~ aveSpeed, data = run_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.542 -18.918   2.212  16.376  56.130
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -208.21     161.63  -1.288  0.21495
## aveSpeed       80.82      22.51   3.590  0.00225 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.84 on 17 degrees of freedom
## Multiple R-squared:  0.4313, Adjusted R-squared:  0.3978
## F-statistic: 12.89 on 1 and 17 DF,  p-value: 0.002255
```

a) What is the estimate for the standard deviation of the number of calories burned based on this linear model?

30.84

standard deviation = ~~22.51~~

this is the ~~Std. error for aveSpeed~~.

RSE.

b) On average, how many more (or fewer) calories can our runner expect to burn for each mph increase in average running speed?

On average, 80.82 more calories can runner expect to burn for each mph increase in average running speed.

Because the slope of SLR model is 80.82

c) Suppose, on average, for any person within the same age group as our runner, every mph increase in running speed corresponds to 100 additional calories burnt. Describe a procedure to determine if our runner's rate of burning calories is different from this average for all people in the age group. Make sure you define any symbols you use. You do not need to actually determine the answer, just describe the statistical procedure you would use.

Calculate a 95% confidence interval (CI) for the slope of the SLR model using formula $\hat{\beta}_1 \pm t^*_{(n-2),\alpha/2} \times SE(\hat{\beta}_1)$. $\hat{\beta}_1$ is the estimated slope for SLR. $SE(\hat{\beta}_1)$ is the standard error for $\hat{\beta}_1$. t^* is the t-value for $n-2$ degrees of freedom, n is the number of individuals we use for the SLR model, and α is 5 since we are finding 95% CI.

If 100 lies with CI, then there is no statistically significant difference, otherwise, there is.

d) List two numbers in the R output above that can help us determine if this model is a good fit for the data. Explain each briefly.

$P(>|t|) = 0.21495$. So p-value for t-test of slope is larger than $\alpha = 0.05$. If we use $H_0: \beta_1 = 0$, $H_a: \beta_1 \neq 0$, then since p-value $> \alpha$, we fail to reject H_0, and there is no statistically significant evidence to show that there is correlation between calories and speed.

p-value for F test is $0.002255 > 0.001$. So again, we fail to reject $H_0: \beta_1 = 0$. So this model may not be a good fit since the linear correlation is too weak to be considered significant.

x175

References

1 https://xkcd.com/1725

2 Paul J. Laumakis & Kevin McCormack (2014) Analyzing Exercise Training Effect and Its Impact on Cardiorespiratory and Cardiovascular Fitness, Journal of Statistics Education, 22:2, , DOI: 10.1080/10691898.2014.11889702]

Xinxin Li

Correction:

1. Correct answer: a, c
   d is incorrect because we cannot say "10% chance" under this case.

3. Correct answer: c, d
   b is incorrect because predicting the average of the response for a specific predictor value is not inference but estimation.

7. For this specific case, r^2 is only 0.06 for SLR.

$$r^2 = \frac{SSModel}{SSTotal} = 1 - \frac{SSE}{SSTotal}$$

Since r^2 is small, we can say that SSModel is less than SSE, so the percentage of variability of the response that can be explained by the SLR is low. So SLR may not be the most suitable model for this situation.

9. a Standard deviation = RSE = 30.84
   This means that based on the linear model, the number of calories burned deviates from the average value by 30.84 calories on average.

9. d T-test for slope:

$H_0: \beta_1 = 0$ β_1 is the true slope for the SLR model

$H_a: \beta_1 \neq 0$

$P(>|t|) = 0.00225$, so p-value for t-test for slope is less than alpha (0.05). So we reject H_0, and there is a statistically significant evidence that there is correlation between the calories burnt and the speed.

ANOVA F test:

$H_0: \beta_1 = 0$ β_1 is the true slope for SLR.

$H_a: \beta_1 \neq 0$

p-value = 0.002255 < alpha = 0.05. So we reject H_0 that $\beta_1 = 0$, and there is a statistically significant evidence that there is correlation between calories and speed, and this model is effective in predicting calories from speed.