

# 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:** \_\_\_\_\_

**Swarthmore Username:** \_\_\_\_\_

**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 *not* 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 be 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 *not* 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.

3

Which of the following statements are *false*? 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 *not* 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 *not* need to assess 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.

## Short answer questions

### 6 (4 points)

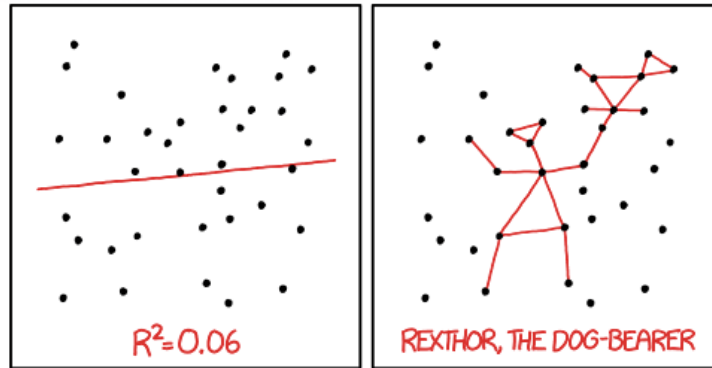
We've discussed the construction of statistical models as following four general steps:

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.

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.

7 (4 points)

Based on the data shown in the scatter plot of this comic<sup>[1]</sup>, 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.

## 8 (6 points)

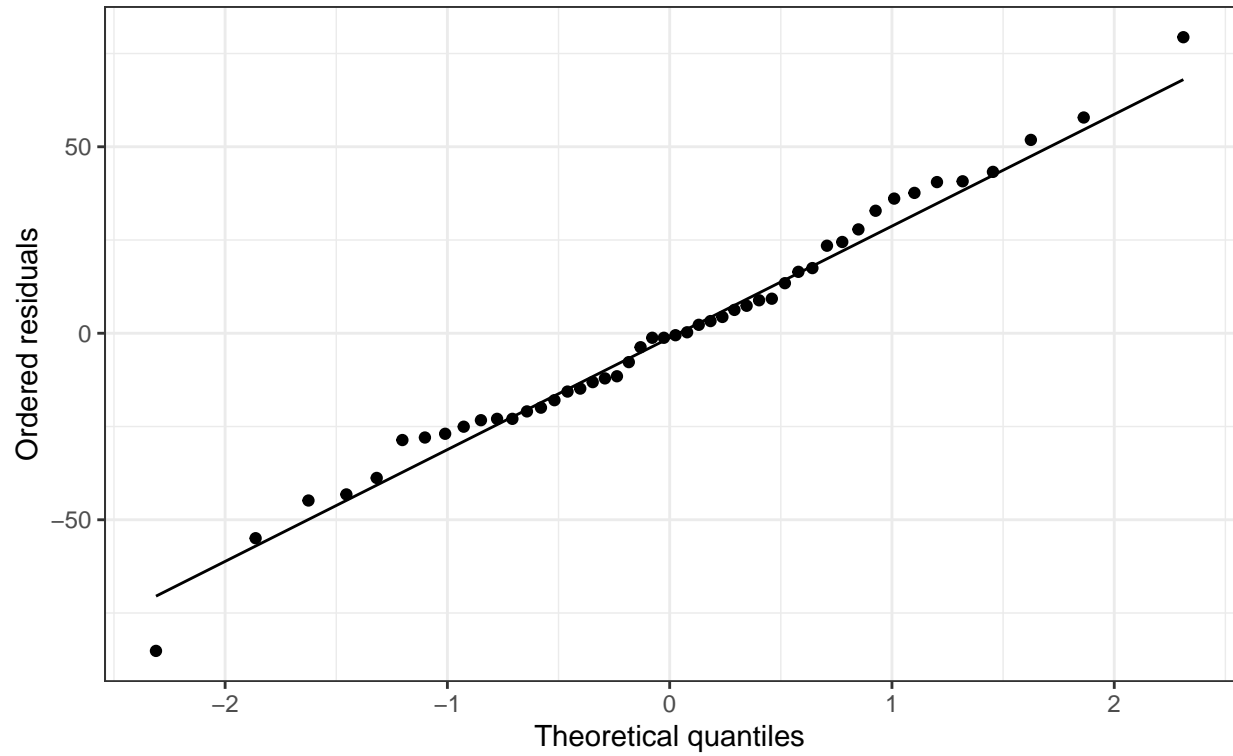
The data below represents a simple random sample of Singaporean diamonds. Suppose we are interested in how the size of the diamond (in carats) can predict (or explain) what the cost of the diamond will be. Below is the R code for fitting this simple linear regression model.

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

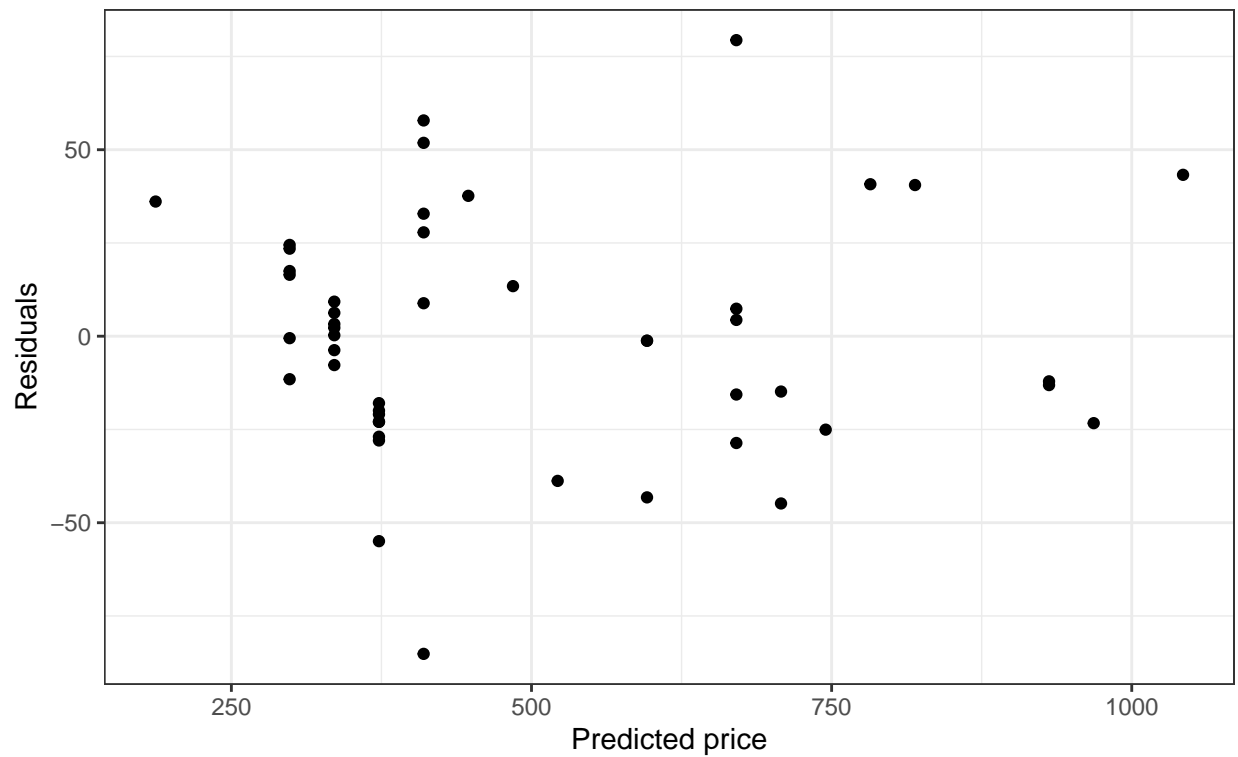
##
## Call:
## lm(formula = price ~ size, data = diamond_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -85.159 -21.448  -0.869  18.972  79.370
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -259.63      17.32  -14.99  <2e-16 ***
## 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
```

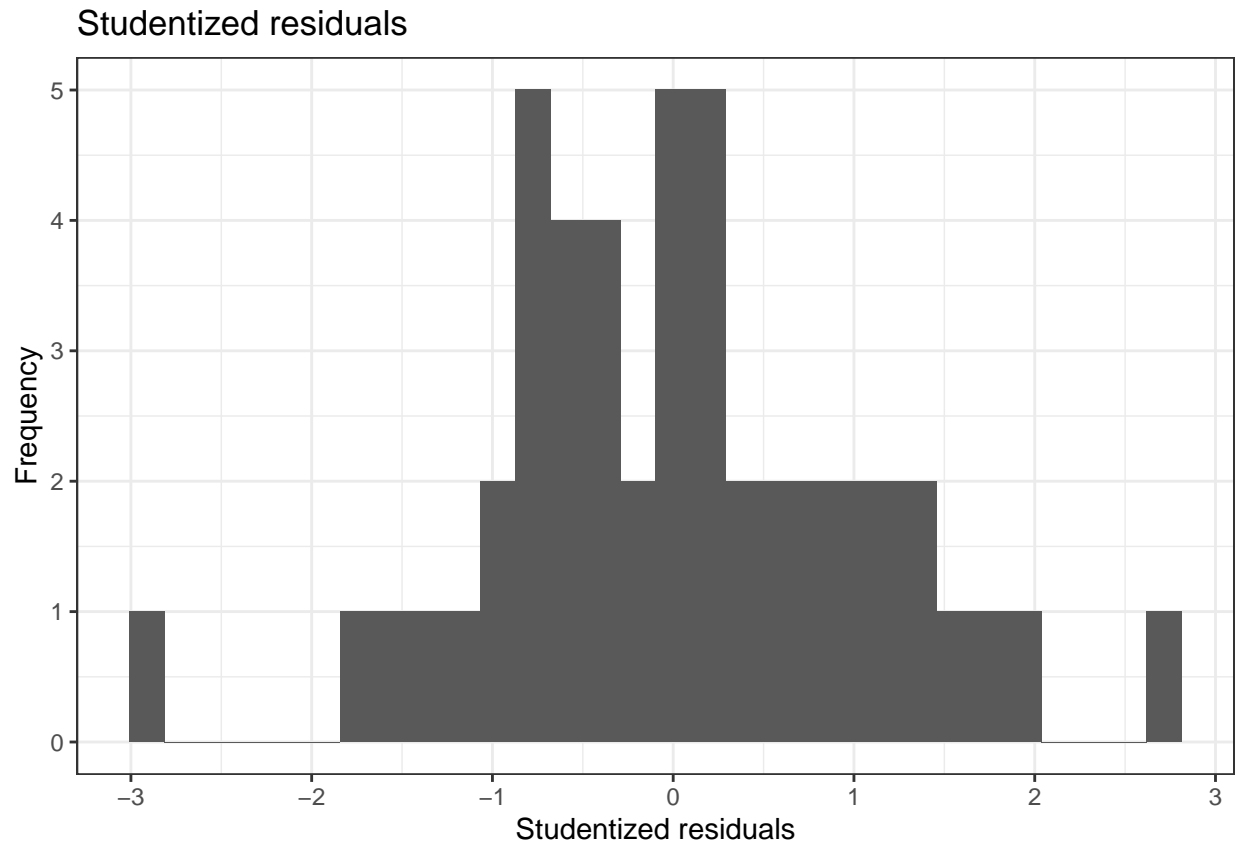
Analyse the following plots based on this regression model to answer this two-part question.

Residual plot  
Normal probability plot



Residual plot  
Fitting diamond prices as a linear function of size





- a) Based on these plots, what conclusions can we make about the conditions necessary for inference with a SLR model?
- b) Say instead of the size of the diamond measured in carats, we'd like to look at the size in grams ( $1 \text{ carat} = 0.2 \text{ grams}$ ). Would we expect the behavior of any of the plots above to change? Briefly explain your answer.



## 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.<sup>[2]</sup>

```
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?

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

## 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]