Exam 1

STAT 021

Swarthmore College

2019/10/4

Name: Mats Shoraka

Instructions:

There are seven questions on this exam. The points allotted for each question are given at the end of the problem. Please don't write an entire page response for any of the answers. Rather, answer these questions to the best of your ability with succinct, informative statements or observations. You may or may not use the following formulas and definitions.

Formulas and Definitions Linear model: $Y = \beta_0 + \beta_1 x + \epsilon$ or, equivalently, $E[Y] = \beta_0 + \beta_1 x$.

In the model(s) above, if we assume that the mean of ϵ is 0 and the variance of ϵ is some unknown number, σ^2, then the mean of the random variable Y is $\beta_0 + \beta_1 x$ and the variance of Y is σ^2.

Fitted/estimated model: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

In the fitted model above, we solve for the least squares estimates of the parameters using these equations:

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Definition of residuals: $\hat{y}_i - y_i = e_i$

Regression model sums of squares: $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

Residual sums of squares: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Total sums of squares: $\sum_{i=1}^{n}(y_i - \bar{y})^2$

Relationship among the sums of squares terms: $SS_{tot} = SS_{reg} + SS_{res}$

The sums of squares terms are used to calculate the following statistics:

$$\hat{\sigma} = \sqrt{\frac{SS_{res}}{n-2}}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}$$

Problem 1 Suppose that the observational units in a study are patients who entered the emergency room at French Hospital in the previous week. For each of the following, indicate whether it is a categorical variable, a numerical variable, **or** not a variable with regard to these observational units. (10 points) qual

quant a) How long the patient waits to be seen by a medical professional

numerical variable

b) Day of the week on which the patient arrives

Categorical variable

c) Average wait time before the patient is seen by a medical professional

numerical variable

d) Whether or not wait times tend to be longer on weekends than weekdays

not a variable with regard to these observations.

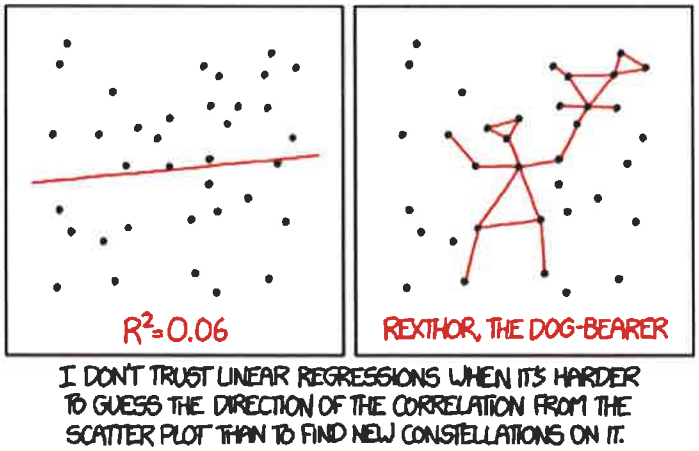
Problem 2 Consider the transactions at the Science Center coffee bar to be the observational units in a statistical study. In a paragraph or less, state a research question that involves two quantitative variables for these observational units. Also clearly identify what roles the two variables would play in the study and why. (10 points)

The objective of this study is to test the correlation between the amount of time each transaction takes (at) and the amount of points (money) spent. The predictor variable is time, and the response variable is numbe of points both of which are quantitative variables. I chose time to be the predictor variable because I have a hypothesis that the more time spent on the transaction predicts the amount of ~~more~~ points the student spent,

Problem 3 Suppose a professor has a paper titled: *Estimation and hypothesis testing in regression in the presence of nonhomogeneous error variances* lying out on her desk.[1] In 1-2 sentences, explain what you think this paper is about? (10 points)

This paper is about one of the assumptions necessary to preform Simple Linear Regression: variance of the errors must be constant. It details how we would first test this assumption through looking at a scatter plot of the predicted $\hat{y}$: ~~response~~ ~~variable~~ ~~vs~~ as the predictor variable withe the residuals as the response variables, and then would take the reader through methods of data transformation in an attempt ~~eff~~ to remedy this.

Problem 4 Based on the data shown in the scatter plot of this comic[2], what can you tell me about the relationship between the SS_{reg} and SS_{res} terms? (10 points)


R²=0.06
REXTHOR, THE DOG-BEARER
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.

Given $SSreg = \sum_{i=1}^{\wedge}(\hat{y}_i-\bar{y})^2$ and $SSres = \sum_{i=1}^{\wedge}(y_i-\hat{y}_i)^2$, along with the graphs above, we can see SSreg is close to 0 wile SSresidual is close to SStot. ~~Since~~ because

$$.06 = R^2 = 1 - \underbrace{\frac{SSresidual}{SStot}}_{\text{close to } 1} = \underbrace{\frac{SSreg}{SStot}}_{\substack{\text{close to } 0 \\ \text{but } SStot \neq 0}}.$$

In general if ~~R~~ SSresidual is close to SStot, SSreg is Close to 0 and if SSresidual is ~~much~~ much smaller ~~or larger~~ than SStot SSreg is close to SStot.

$\hat{B}_1 = \dfrac{\sum_{i=1}^{n=5}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n=5}(x_i-\bar{x})}$ $\hat{B}_0 = \bar{y} - \hat{B}_1\bar{x}$

Problem 5 Suppose we have observed a small data set (say $n = 5$) without any significant measurement error (e.g. we are collecting data on vapor pressure and temperature but our instruments to measure each are exact). How do we find the line of best fit? (10 points)

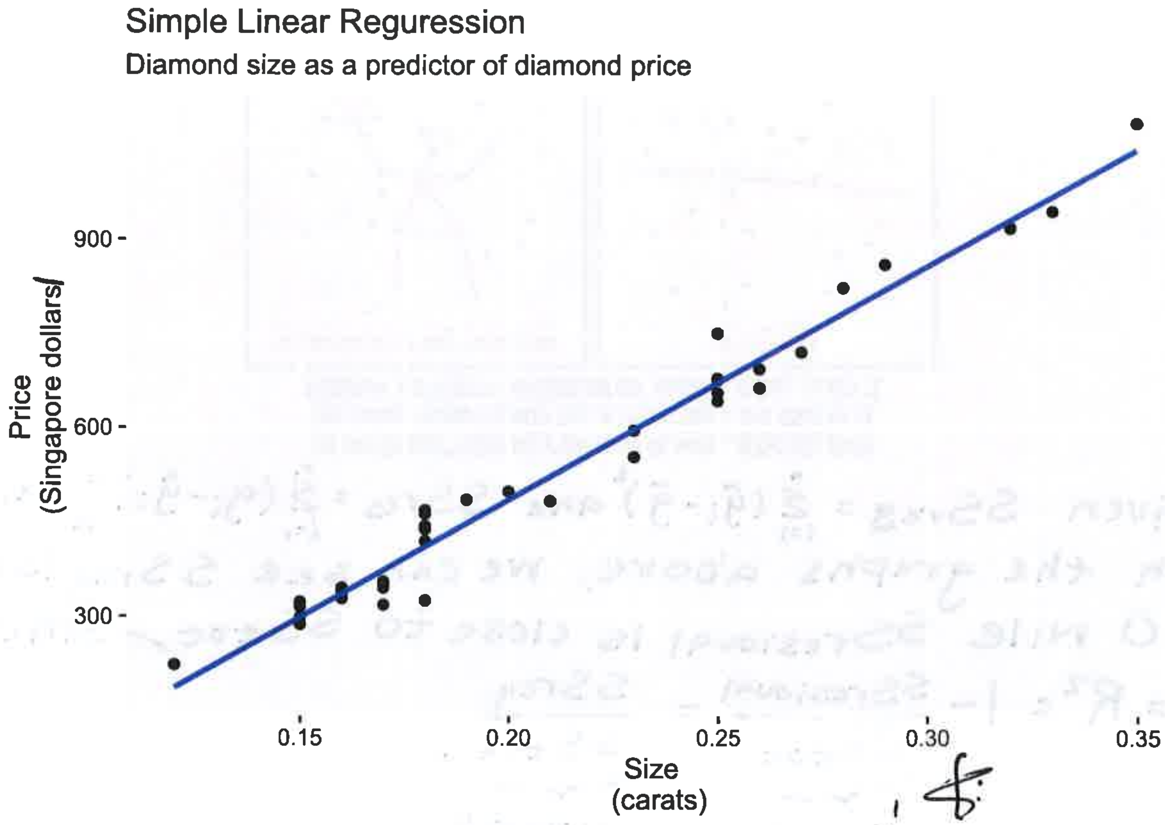
There are many ways to find the line of best fit. We could minimize the error produced by $\text{diff}(\hat{B}_0+\hat{B}_1x_i, y_i)$ to find the best $\hat{B}_0$ and $\hat{B}_1$. We could also plug our values into R and use the lm call. More concretly we could plug ~~each~~ our data into the equations for B_1, B_0 found on the front of this test, after calculating $\bar{y}$ and $\bar{x}$ $\left(\frac{1}{n}\sum_{i}^{n} x_i, \frac{1}{n}\sum_{i}^{n} y_i\right)$. $-n=5$

use linear algebra approximation

Problem 6 Recall the diamond data that we discussed in class. For this data, we have a simple random sample of Singaporean diamonds and 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. (25 points)

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

Analyse the following three plots based on this regression model to answer the next two questions.



~~$\begin{bmatrix} x_1 y_1 \\ x_2 y_2 \\ x_3 y_3 \\ x_4 y_4 \end{bmatrix}$~~

4 ~~$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$~~

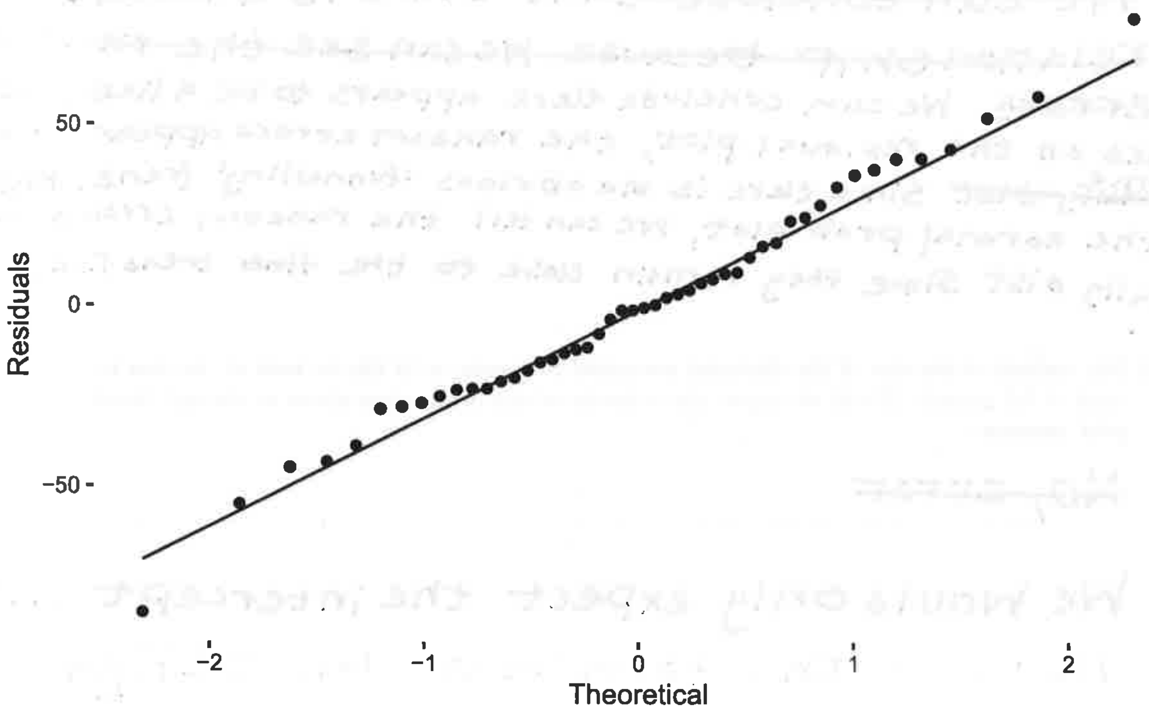
~~$[x_1\ x_2\ x_3\ x_4]$~~

4 ~~$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$~~

Norm Prob

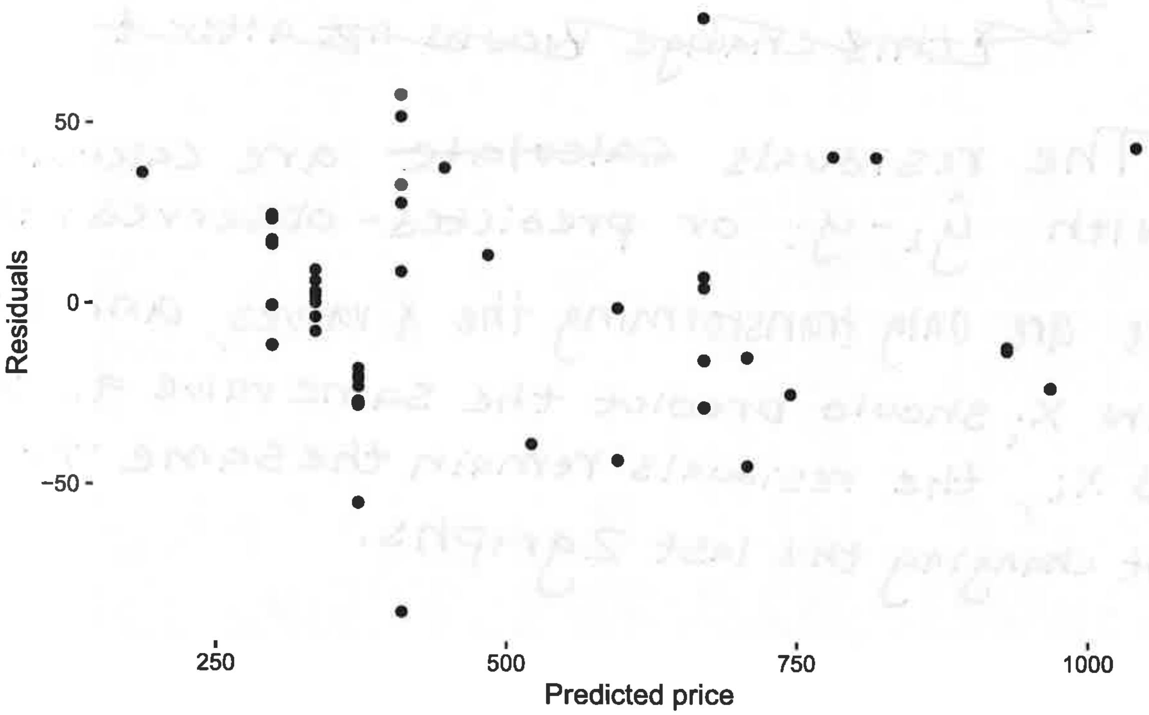
~~Residual plot~~

Fitting diamond prices as a linear function of size



Residual plot

Fitting diamond prices as a linear function of size



random errors themselves shouldn't be constant right! else they aren't random

variance?

a) Based on these plots, what conclusions can we make about the presence of a linear relationship, if the random errors are constant, and if the random errors are Normally distributed?

~~We can conclude that there is a linear relationship because we can see the random errors~~ We can conclude there appears to be a linear rel. Based on the residual plot, the random errors appear ~~to be~~ to have const variance ~~Normality dist~~ since there is no obvious 'funnelling' trend. ~~Based~~ Based on the normal prob plot, we can tell the random errors are normally dist since they remain close to the ~~time~~ trend line.

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 answers.

~~No, error~~

We would only expect the intercept and slope value of the formula for the SLR to change. So, overall no.

~~this change would not alter t~~

The residuals ~~calculate~~ are calculated with $\hat{y}_i - y_i$ or predicted - observed. Because we are only transforming the x values, and each new x_i should predict the same value as the old x_i, the residuals remain the same thus not changing the last 2 graphs.

Problem 7 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. We are interested in answering the question: is the speed of the run (in mph) related to the number of calories burned. Below is the R code and output for fitting such a linear model to this data.[3] (25 points)

```
run_dat <- read_table2("~/Google Drive Swat/Swat docs/Stat 21/Data/Four-Mile-Run-data.txt")
summary(lm(calories~aveSpeed, run_dat))

##
## 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 calories

Residual standard error

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

80.82 calories

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. How can we determine if our runner's rate of burning calories is different from this average for all people in the age group?

We can run a t-test on B_1 with

$H_0: B_1 = 100$ $H_a: B_1 \neq 100$.

If we get a p-val $< .05$ we can say at the ~~α=.05~~ 95% confidence level that this is abnormal (ie reject H_0). We could also calculate a CI around $\hat{B}_1 = 80.82$ at a 95% level, and if 100 falls outside this range we would deem it abnormal.

d) What numbers in the R output above can help us determine if this model is a good fit for the data? Explain briefly. (There are at least two.)

F-statistic, p-value
aveSpeed p-value

} Test the hypotesis tes $H_0: B_1 = 0$ $H_a: B_1 \neq 0$ thereby conveying to us whetler X has any correlation with Y.

Multiple R^2 → Tells us the percent of Y that can be predicted from X.

[1] Michael L. Deaton, Mation R. Reynolds Jr. & Raymond H. Myers (1983) Estimation and hypothesis testing in regression in the presence of nonhomogeneous error variances, Communications in Statistics - Simulation and Computation, 12:1, 45-66, DOI: 10.1080/03610918308812299

[2] https://xkcd.com/1725

[3] 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]