Exam 1

STAT 021

Swarthmore College

2019/10/4

Name: Kayonna L Tindle

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:

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

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

7 27

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)

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

Numericae variable:

[illegible]

b) Day of the week on which the patient arrives

categorical variable

[illegible]

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

Numericae variable

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

Not a variable, but rather a research question.

categ = qualitative

= quant.

observ

Problem 2 Consider the transactions at the Science Center coffee bar to be the observational units in a statistical study. In a paragraph of 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)

A potential observation research question with this observationoe unit could be is: the number of transactions the science Center coffee bar, recieves significanty influence by whether it's daytime or nightime. The two variables being compared in this equation is the qualitative predictor variable, time of day, and the quantitative response variable # of transactions. The latter variable would be our response variable as this the variable that we are trying to measure or gauge the potential influence of the first variable.

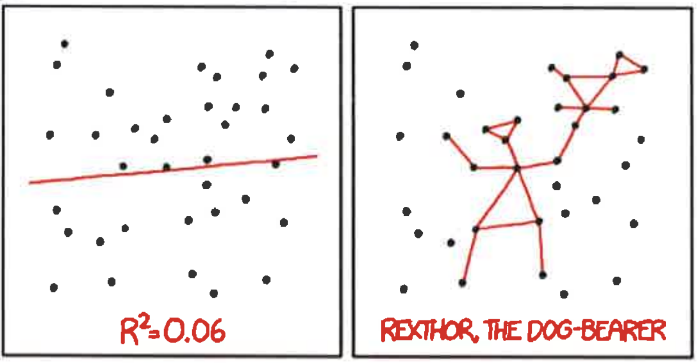
Problem 3 Suppose a professor has a paper titled: *Estimation and hypothesis testing in regression in the presence of nonhomogeneous error variances* lying out on her desk.[1] In 1-2 sentences, explain what you think this paper is about? (10 points)

It would seem that the paper is about
how estimateing and fitting linear regression in
testing for significance and correlation is impacted
by whether or not your error are
independent of eachother and the variance
of your errors is constant. This paper will likely
be able to explain how heteroskedacity can be resolved
potentioecy by transforming your data for your regression
model.

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)

(predicted - mean) ?

(observed - predicted) ?



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.

As observedable in the data from this comic, SS reg & SSres ends up being the sum of SS_{total} or the eutoe sum of squares of the data.

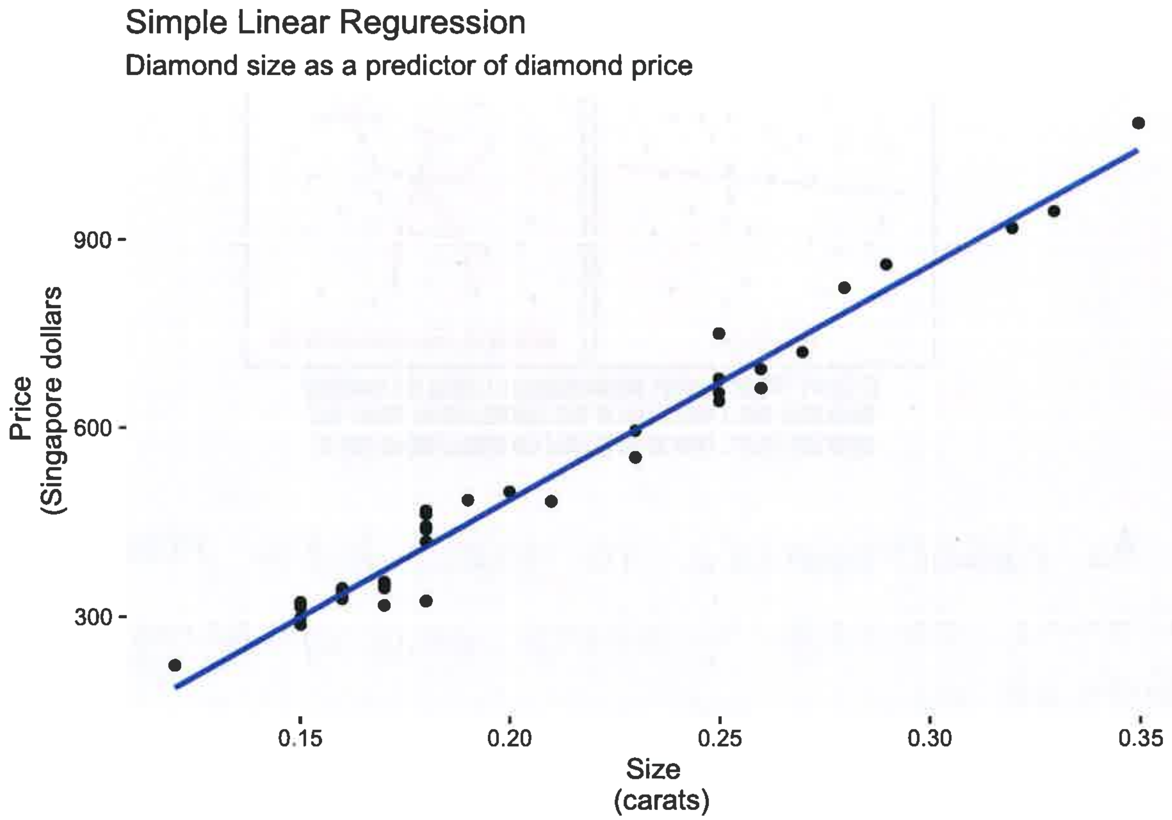
Problem 5 Suppose we have observed a small data set (say $n = 5$) without any significant measurement error (e.g. we are collecting data on vapor pressure and temperature but our instruments to measure each are exact). How do we find the line of best fit? (10 points)

We can find the line of best fit by calculating the slope of the points (if done by hand, just calculating the slope of two representative points) and using the slope ($\hat{\beta}_1$) and the data points of the data to calculate the y-intercept ($\hat{\beta}_0$) for the equation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x$ wherein x is fixed.

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)
```

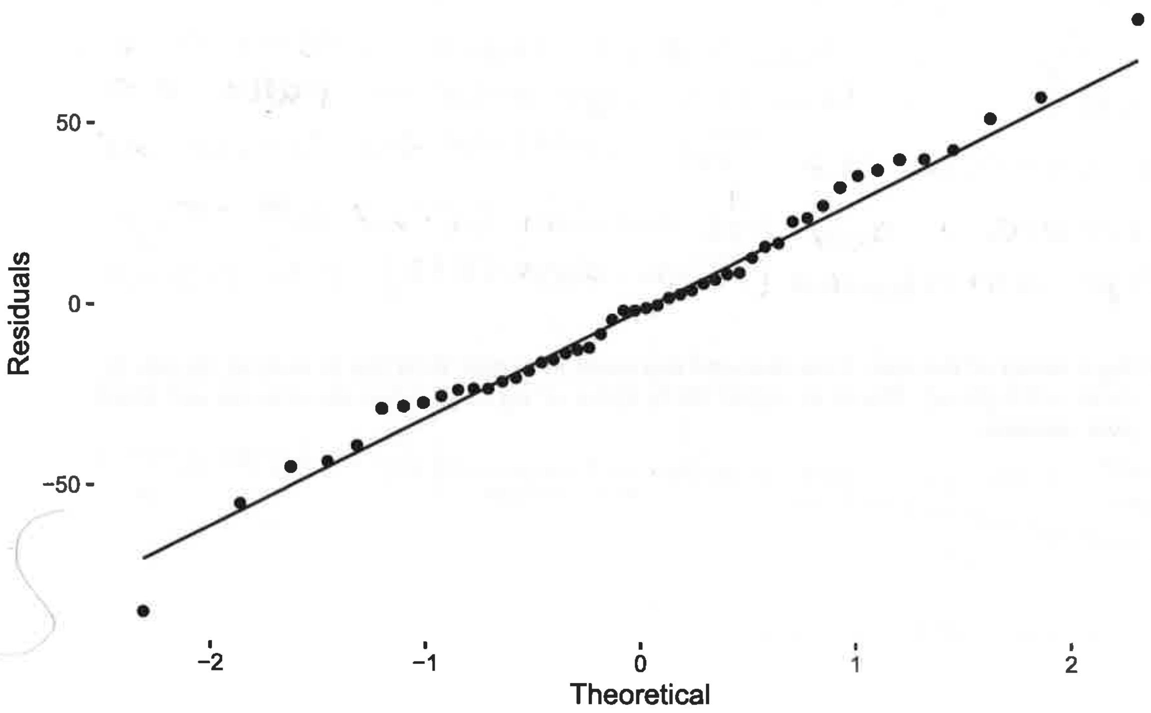
Analyse the following three plots based on this regression model to answer the next two questions.



Normal Probability Plot

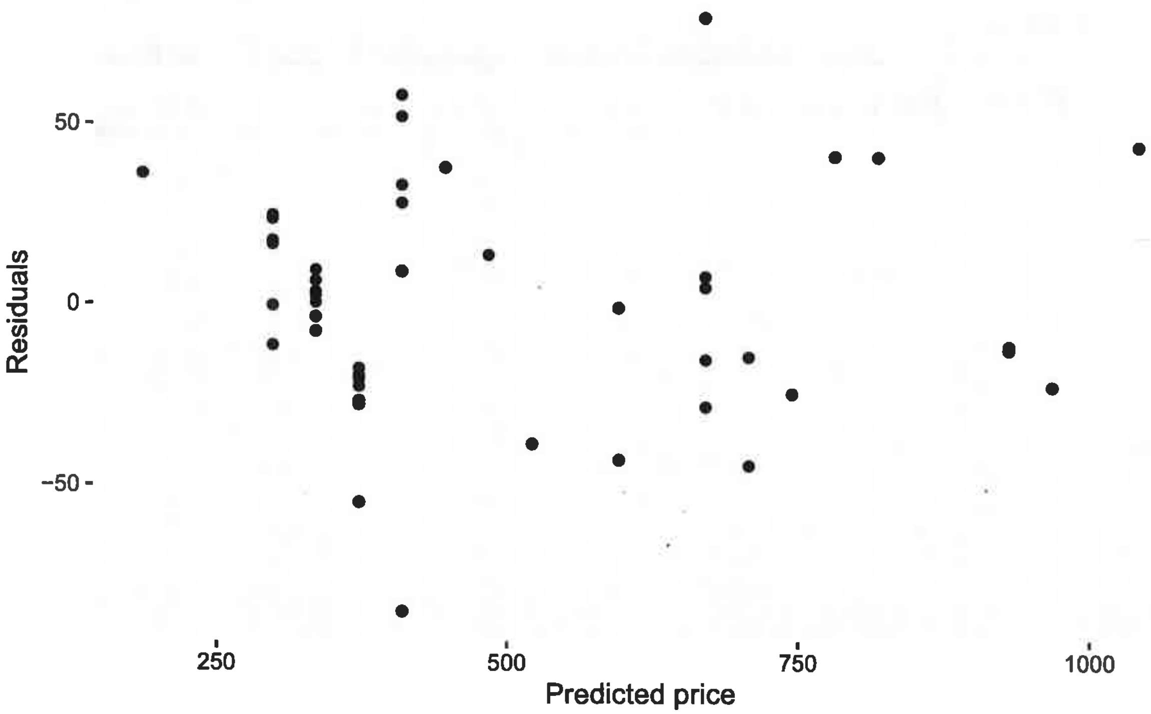
~~Residual~~ plot

Fitting diamond prices as a linear function of size



Residual plot

Fitting diamond prices as a linear function of size



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?

The simple linear regression plot and the normol probability plot shows that there seems to be a linear relationship between our predictor value x and our response variable y. The spread of the random errors appear to be constant and the random errors are only somewhat normolly distributed (from observing the residual plot).

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.

~~If we standardize the data, we would~~ not expect

We should not expect the linear regression plot to change from this linear transformation. Also if we standardized the data we will not need to be concerned about our other data plot behavior changing as well.

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?

The estimate. for std : 80.82

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

We can expected our runner to burn by a $\hat{\beta}_1$ positive slope coefficint

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 could cook at the t-statistic and p-value to see if his value are significantlydiffrient $p < 0.05$.

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.)

Our p-value and our R^2 Value can help us determine if there is a relationship while the R^2 tells how well good the predict fit model represents the data.

[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]