[redacted] Application Assessment Unit 3

Introduction: overview of investigation.

This investigation is exploring the potential correlation between the time a person spends reading during the day and their assumed vocabulary level. The point of this investigation is to demonstrate the impact leisure reading has on a persons fluency within a single language. The variety of data used consists of highschool aged students 15-17 in order to keep the potential learning gap between teenagers to a minimum.

Hypothesis: critical assumptions relevant to investigation.

The greater amount of time someone spends reading leisurely in a day, the higher level of fluency within their vocabulary will be noticeable.

Response Variable: The vocabulary test result (in words known) (from Preply)

Explanatory Variable: The amount of time spent leisurely reading within a day (measured in minutes on average)

Controlled variables;

- The surveyed being within an age range (15-17)
- All subjects utilised same vocabulary test
- Scale of measurement used (words known out of total dictionary)

Critical Assumption;

- Everyone one is honest with their answers
- Everyone interpreted the parameters of the test in a similar way
- All participants have a similar level of raw vocabulary knowledge (reading time excluded)

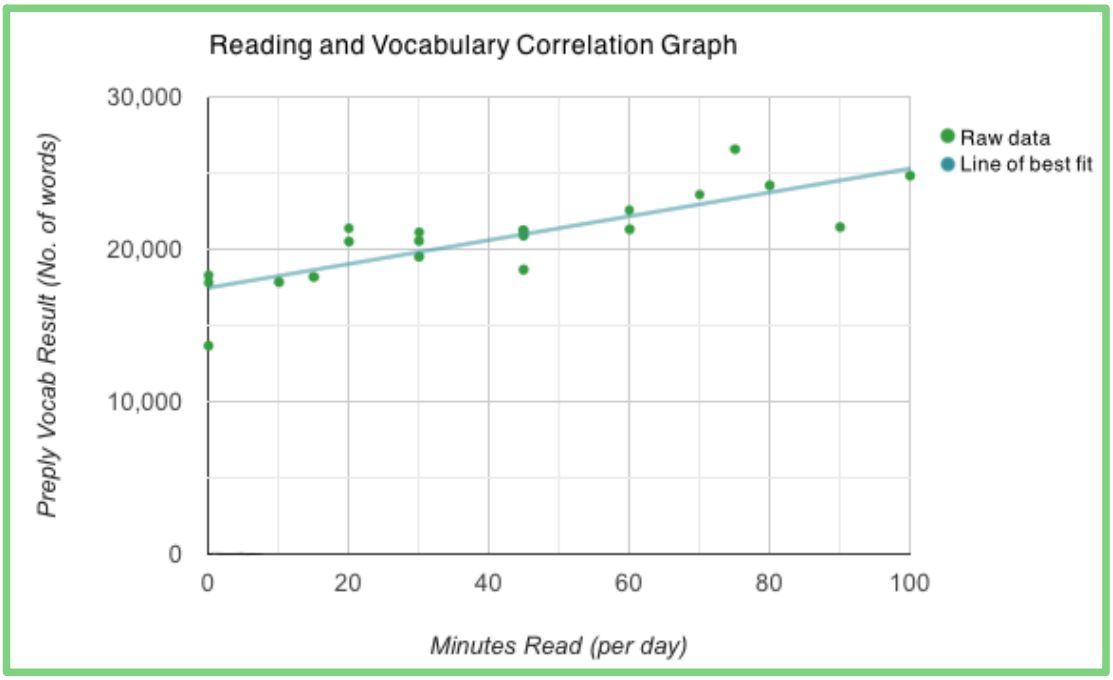
Discussion of data collection method. Include a copy of raw data.

My data was collected via myself asking a randomly selected 20 students within years 11 and 12 (therefore ageing them at 15-17) how often they read for leisure per day on average (in minutes) and asking them to complete a Preply Vocabulary test I linked to each participant. The test in question is a two section, result level based test, meaning depending on how you perform in the first section dictates the difficulty of the second section. This is measured as an estimated count of words you can fluently use within day to day life. As I obtained this data myself it is primary data, with the knowledge I used an external test, I would still consider it primary as I created the questions and have obtained all data including outliers which wouldn't necessarily happen if you obtained secondary data, by collecting primary data I know it is accurate and trustworthy on my end, however my bias could have affected the way I asked questions, I have a smaller sample size than I would of liked and it took a considerable amount of my time. My raw data is demonstrated in the table below.

Minutes read per day	100	90	75	20	0	45	45	20	15	60	30	10	80	60	70	30	30	45	0	0
Vocabulary test result	24961	21762	26926	21690	13856	18607	21184	20516	18128	21550	20600	17967	24128	22505	23460	21040	19500	21050	17960	18105

The results from CAS (all maths can be seen in appendix)
m = 78.3948 (78.4)
b = 17540.96 (17541.0)
Full Equation(y-hat=mx+b): y-hat = 78.4x + 17541
r = 0.8240
r^2 = 0.6790

Analyse: consider appropriate graphs, calculations must be shown.
The scattplot below is the visually represented correlation between the table and the line is the linear line of best fit demonstrated above.

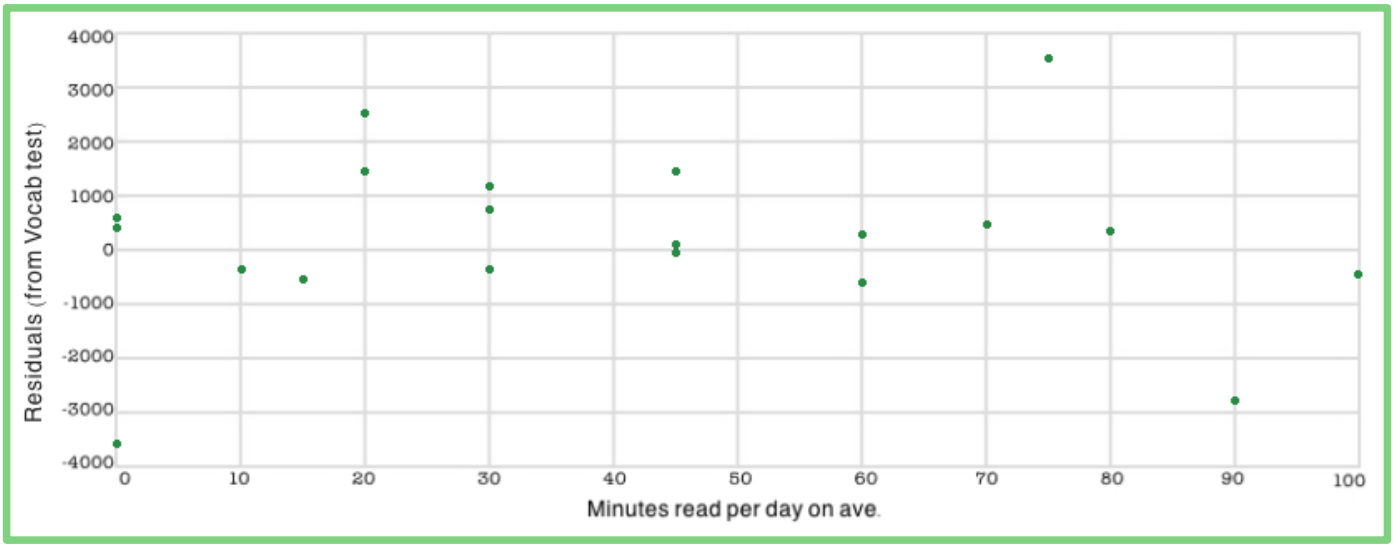


As seen from the line of best fit on the graph it travels in a positive direction meaning there is a positive correlation and as you read more minutes per day, your vocabulary abilities grow as well.
Y-intercept interpretation: If you read for 0 leisure minutes per day, your vocabulary would still be 17541 words as you learn words in alternative places.
Gradient interpretation: there is an m value of 78.4, this means that for every one minute increase in leisure reading per day your vocabulary increases 78.4 words.
The correlation coefficient or r value is 0.824 this demonstrates a strong positive linear correlation within the data, additionally the r^2 value or coefficient of determination is 0.6790, this means that 67.9% of the variation within a person's vocabulary can be explained by the variation in the amount of time they read on average per day (in minutes.) Despite this strong correlation there may be lurking variables that contribute to a difference in vocabulary level such as a greater access to education, differing levels of work ethic or learning difficulties within the participants.

A residual plot is used to determine the validity of a linear graph to measure the correlation. My residuals have been calculated and graphed below. The residuals (rounded) added up are -4 and raw 0. This demonstrates all data points have been accounted for within the plot.

Minutes read per day	100	90	75	20	0	45	45	20	15	60	30	10	80	60	70	30	30	45	0	0
Vocabulary test result	24961	21762	26926	21690	13856	18607	21184	20516	18128	21550	20600	17967	24128	22505	23460	21040	19500	21050	17960	18105
Y=78.4x + 17541	25381	24597	23421	19109	17541	21069	21069	19109	18717	22245	19893	18325	23813	22245	23029	19893	19893	21069	17541	17541
Residuals (rounded)	-420	-2835	3505	2581	-3685	-2462	115	1407	-589	-695	707	-357	315	260	431	1147	-393	-19	419	564

Residual Plot



The formula used to calculate this plot was Residual = Yi - Y-hati where i is an individual value. As seen within the residual plot there is no identifiable pattern within the data, and this randomness points towards a least squares regression line suitable data set.

Trends and Hypotheis:

As discussed in the analysis of the scatter plot and residual plot there is a positive trend between leisure reading and the size of a persons vocabulary. This is demonstrated through the least squares regression line and the r value of 0.82 and r^2 value of 67.9%.

The original hypothesis was in relation to whether reading for leisure aided in expanding a persons fluent vocabulary. The above data is representation of how reading impacts somebody's vocabulary in a positive way and this reinforces my hypothesis due to the strong positive correlation and high coefficient of determination with valid least squares regression demonstrated.

Conclusion and Evaluation:

The results of the experiment conclude that reading does impact your vocabulary ignorant of major lurking variables. This was dependant on the validity of data collection and the suitability of my data in regards to a least squares regression line.

Assumptions I made and Limitaions I experienced:

- the amount of people included within my data (as collecting it myself is time consuming is a limitation on accuracy.
- I've assumed that every student had equal access and opportunity to learn as child and grasp the English language
- I've assumed that everybody participating is a native English speaker
- The reliability (trustworthiness) of the participants is a limitation that could be further eradicated by increasing sample size.
- The sample size is within one age, to further reliability it's possible to run the same experiment with different age groups to evaluate whether this stays true throughout a persons whole life.

Every time the experiment is completed the results will differ to an extent. As a generalisation people will have an increase in vocabulary if they read more regularly. However to further improve this experiment the following things could be done;

- An increased sample size, a minimum of 10% of the total group you're generalising. Eg. School of 2000, minimum 200 kids.
- An increased variety of people, by having a small sample size, diversity is naturally reduced so increasing sample size will increase all types of diversity not acting as controlled variables.
- Considering this within a larger scale, or a variety of ages to explore the continuity of the hypothesis.
- To increase validity you could personalise a vocabulary test and monitor the participants reading habits to obtain more accurate data than word of mouth.

Apendix:

Unit 3 Assignment Maths

Finding linear line of best fit

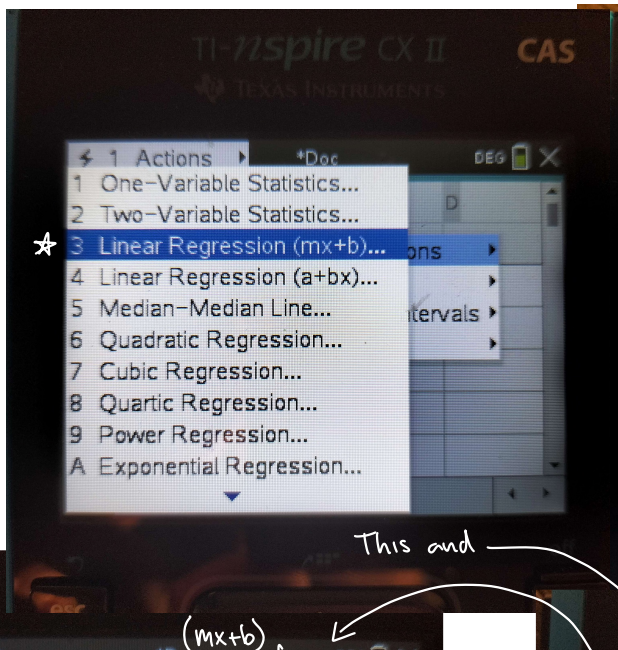
$\hat{y} = 78.39x + 17540.96$

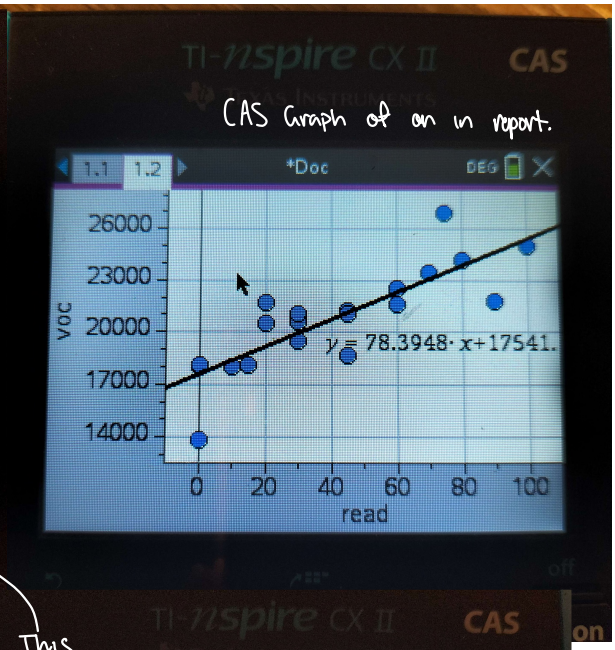
$(\hat{y} = 78.4x + 17541.0)$

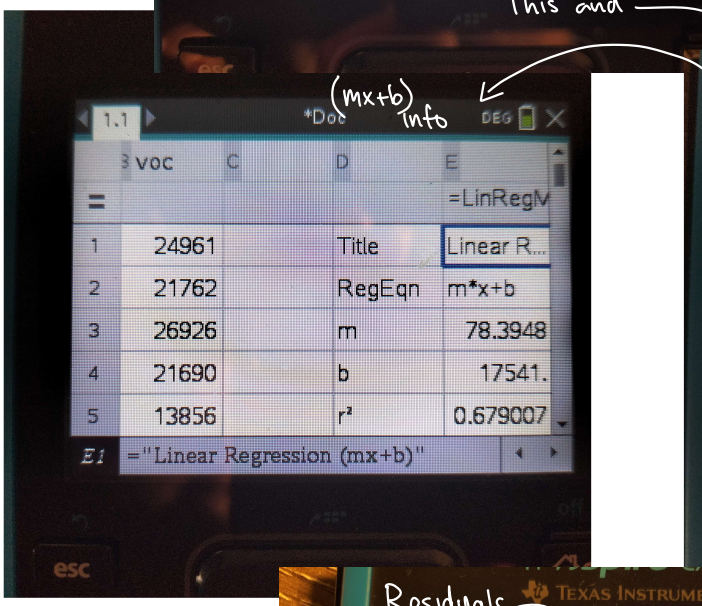
$78.4 \cdot 100 + 17541.0 = 25381$

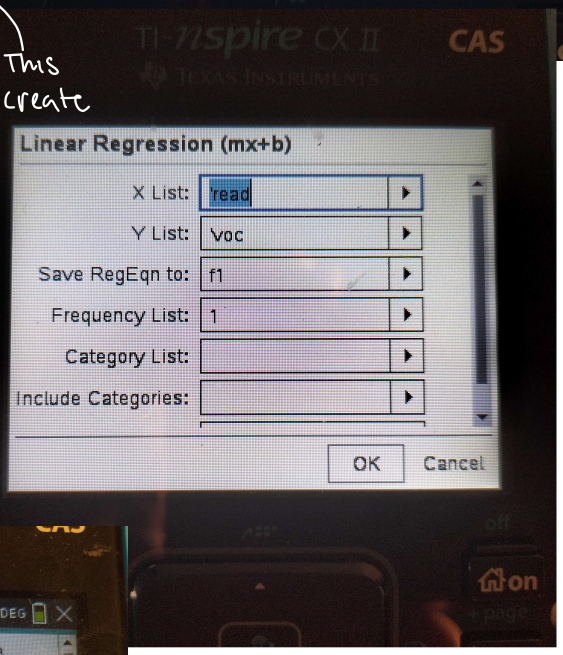
Two points;

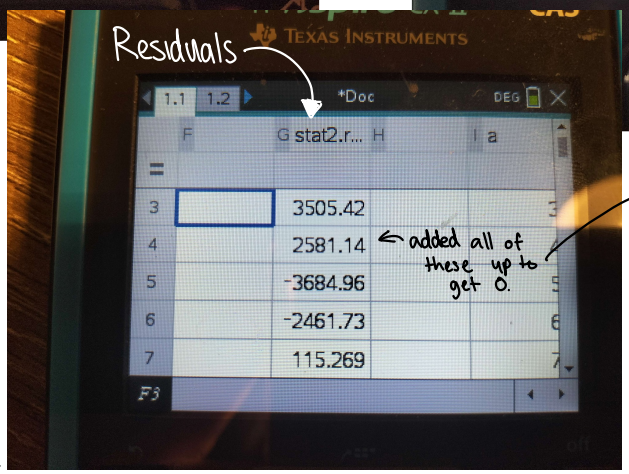
0 / 17541 & 100/25381











Example of calculating residual.

Data piece 100-24961

$\hat{y} = 78.4(100) + 17541 = 25380$

$24961 - 25380 = -419$

Bigger Versions of the tables.

Minutes read per day	100	90	75	20	0	45	45	20	15	60	30	10	80
Vocabulary test result	24961	21762	26926	21690	13856	18607	21184	20516	18128	21550	20600	17967	24128

60	70	30	30	45	0	0
22505	23460	21040	19500	21050	17960	18105

Minutes read per day	100	90	75	20	0	45	45	20	15	60	30	10	80	60
Vocabulary test result	24961	21762	26926	21690	13856	18607	21184	20516	18128	21550	20600	17967	24128	22505
Y=78.4x + 17541	25381	24597	23421	19109	17541	21069	21069	19109	18717	22245	19893	18325	23813	22245
Residuals (rounded)	-420	-2835	3505	2581	-3685	-2462	115	1407	-589	-695	707	-357	315	260

70	30	30	45	0	0
23460	21040	19500	21050	17960	18105
23029	19893	19893	21069	17541	17541
431	1147	-393	-19	419	564