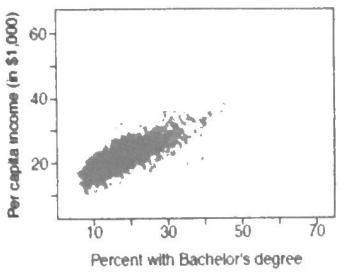
Brock Francom
A02052161

Homework Basic Linear Regression

1. Income and education in US counties: The scatterplot below shows the relationship between per capita income (in thousands of dollars) and percent of population with a bachelor's degree in 3,143 counties in the US in 2010.



   (a) What are the explanatory and response variables?

   Bachelor's degree → income

   (b) Describe the relationship between the two variables.

   as more people get bachelors degrees, income increases, there is a positive relationship

   (c) Can we conclude that having a bachelor's degree increases one's income?

   no, but there is an association

2. **Guess Correlation:**
   Use the applet at http://istics.net/Correlations/ to guess correlations. This is similar to what I will give you on the exam. Play 5 rounds. How many did you get correct?

   16/20 = 80%.

3. Suppose a researcher discovers that during the months with high ice cream sales, there are higher death rates from drowning.

   (a) Does this mean that eating ice cream causes people to drown?

   no.

   (b) Can you think of a possible lurking variable?

   heat.

4. A study was done that showed that children who slept more gained less weight.

   (a) Should we automatically draw the conclusion that more sleep will prevent weight gain?

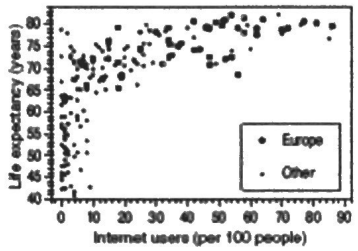
   possibly

   (b) Can you think of a possible lurking variable?

   exercise

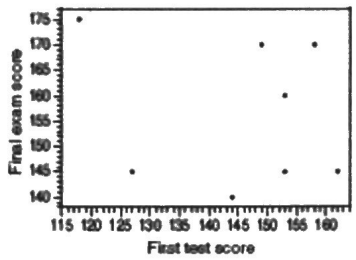
5. If you want to see some more "silly" correlations of variables that do NOT affect each other, but are highly correlated, check here: http://www.tylervigen.com/spurious-correlations

6. Does the scatter plot below suggest that linear regression would be appropriate (can we use the number of internet users to predict life expectancy)?



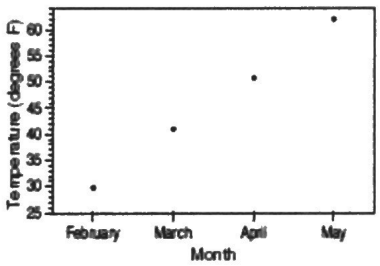
no. not linear.

7. Does the scatter plot below suggest that linear regression would be appropriate for this data set.



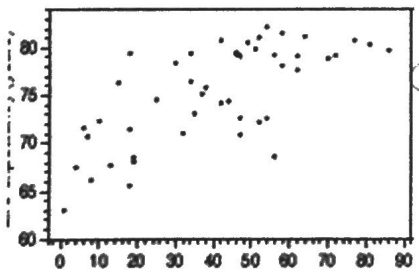
no. doesn't appear linear.

8. Does the scatter plot below suggest that linear regression would be appropriate for this data set.



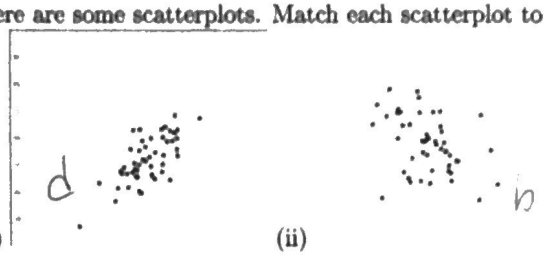
yes.

9. Guess the correlation in the scatter plot.



.46

10. Here are some scatterplots. Match each scatterplot to the correlation.



(i)

(ii)

(iii)

(iv)

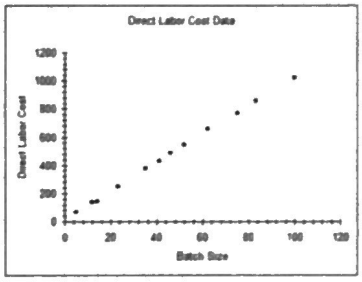
~~(a) $r = -.82$~~

~~(b) $r = -.37$~~

(c) $r = .16$

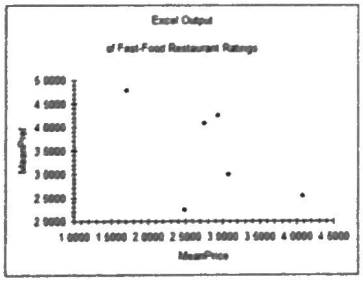
~~(d) $r = .7$~~

11. Guess the correlation in the scatter plot.



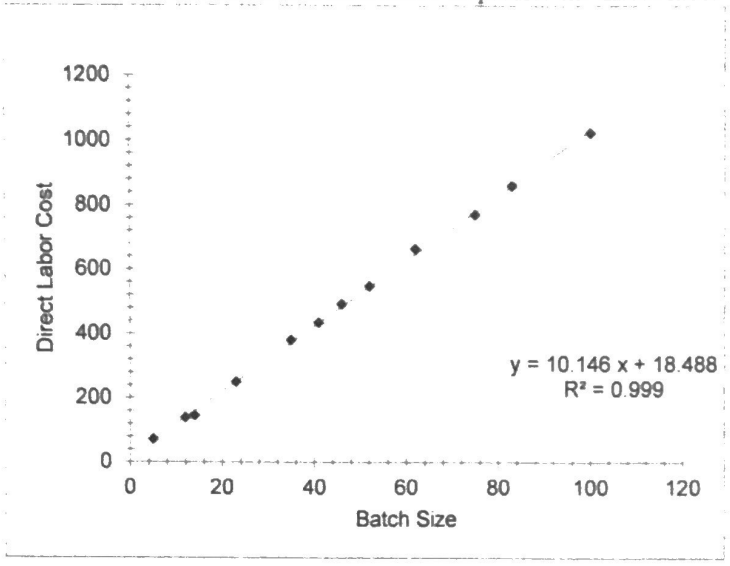
.99

12. Guess the correlation in the scatter plot.



-.52

13. We want to use the Batch Size as a factor to predict the direct labor cost.



(a) Looking at the scatter plot, do you think that linear regression would be appropriate?

yes.

(b) What is the least squares line, or the linear regression equation?

$y = 10.146x + 18.488$

(c) What is $\hat{\beta}_0$? Interpret $\hat{\beta}_0$. Does it have a practical interpretation in this case?

18.488

at a batch size of 0, production labor [illegible] year [illegible] have to pay workers even if not doing anything

(d) What is $\hat{\beta}_1$? Interpret $\hat{\beta}_1$.

10.146, with each additional batch, you [illegible] an average of $10.146 more for labor.

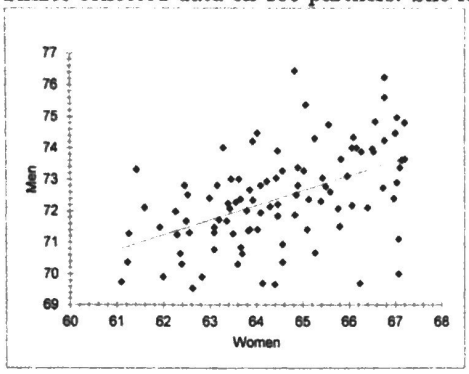
(e) Predict Labor Cost when batch size is 23.

$\hat{y}(23) = 257.85$

(f) The values for batch size in the data set range from 5 to 100. Should we use this data set and our linear regression equation, $\hat{y} = 18.488 + 10.146x$, to predict the labor cost for a batch size of 1000? Why or why not.

no, it probably wouldn't be accurate.

14. A study looked at the heights of women and the heights of the men that they date. Shalee collected data on 100 partners. She found a correlation of $r = .502$.



(a) What do you think would happen to the correlation if all of the men magically grew 7 inches taller, but the women stayed the same?

It would stay the same.

(b) What do you think would happen to the correlation if we measured the heights in centimeters instead of inches?

It wouldn't change.

(c) What do you think the correlation would be if suddenly every woman would only date a man that was exactly 4 inches taller than her?

It would increase and probly be almost 1

15. A college newspaper interviews a psychologist about student ratings of the teaching of faculty members. The psychologist says,

"The evidence indicates that the correlation between the research productivity and teaching rating of faculty members is close to zero."

The paper reported this as

"Professor McDaniel said that good researchers tend to be poor teachers, and vice versa."

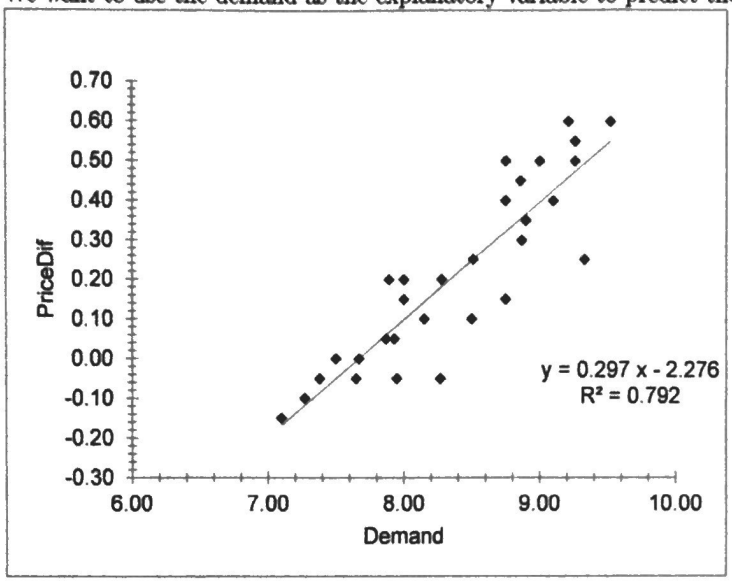
(a) Explain why the paper's report is wrong. Write a statement in plain language (don't use the word "correlation") to explain the psychologist's meaning.

Their findings are interpreted wrong. A value of 0 means there is no evidence that there is a relationship between the 2 things.

(b) The paper interpreted the statement as if the correlation was _____.

i. negative
ii. zero
iii. positive

16. We want to use the demand as the explanatory variable to predict the response variable PriceDif.



(a) What is the least squares line, or the linear regression equation?

$y = .297x - 2.276$

(b) What is $\hat{\beta}_0$? Interpret $\hat{\beta}_0$. Does it have a practical interpretation in this case?

-2.276, if demand = 0, price diff = -2.276

yeah, you would need to drop the price for low demand (sale)

(c) What is $\hat{\beta}_1$? Interpret $\hat{\beta}_1$. Does it have a practical interpretation in this case?

.297, for each unit of demand, the price diff increases by .297.

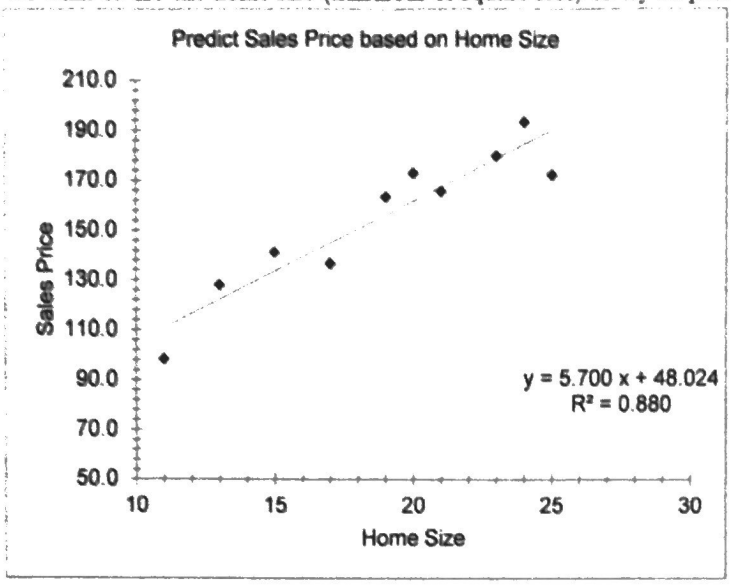
(d) Predict the ~~PriceDif~~ when demand is 11.

$y(11) = .991$

(e) Predict the PriceDif when demand is 60.

This would be extrapolation.

17. We want to use the home size (hundreds of square feet) to try to predict sales price (thousands of dollars).



(a) What is the least squares line, or the linear regression equation?

$\hat{y} = 5.7x + 48.024$

(b) What is $\hat{\beta}_0$? Interpret $\hat{\beta}_0$. Does it have a practical interpretation in this case?

48.024 if house size = 0, sale price = 48.024 thousand dollars.
no, houses of 0 sq ft don't sell.

(c) What is $\hat{\beta}_1$? Interpret $\hat{\beta}_1$.

5.7 for each hundred sq ft. The price increases.
48.024 thousand dollars.

(d) Predict the sales price for a 2700 square foot house.

$\hat{y}(27) = 201.924$

(e) Predict the sales price for a 4500 square foot home.

$\hat{y}(45) = ?$ extrapolation!

18. Outliers and Influential Observations

- Use the applet https://www.geogebra.org/m/cDFC774m
- Click the "Fit Least Squares Line" checkbox.
- Click the "correlation" checkbox. Pay attention to how the correlation changes as you move points around.

(a) Move your points so they are close to a straight line.

i. What is the equation of the line? $y =$

ii. Move a point so that it is an outlier but still on the line.
(Something like this)

iii. What is the equation of the line now? about same

iv. Did the equation change? no

v. Was this outlier an influential point?
no

(b) Move that point so that it is an outlier but is far away from the line.
(Something like this)

i. What is the equation of the line now? $y =$

ii. Did the equation of the line change? yes

iii. Was this outlier an influential point?

yes.

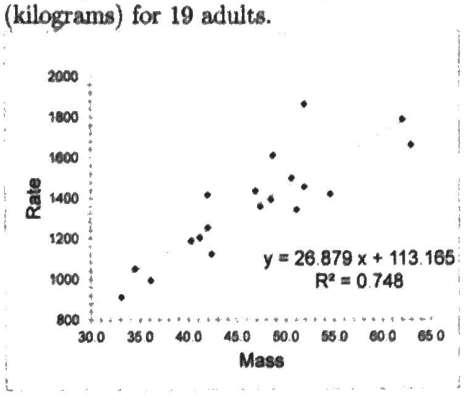
(c) Drag the outlier point around the plot & watch the changes in the equation of the line as you do so.

✓

19. **Read:** Is it be common practice to ignore outliers when they throw the least squares line off considerably?

- I think it depends on your situation.
- If you only care about the majority of the data, if you want to be as accurate as possible that 95% of the time and you don't mind if you are inaccurate at predicting it 5% of the time, and you aren't necessarily trying to figure out what factors influence your variable, then sure you could throw it out.
  - I had a situation like that this week when it was for my own use and I really just wanted to have a good fit for most of the data.
  - I had a student once who had worked in a mine and they needed to be able to predict the conditions pretty accurately. But the problem was that every 50-100 days they had a random weird day that would be an outlier. But they decided to throw out the outliers because they would rather be accurate 350 days, than let those 5 days in the year throw them off. They just accepted that they would be wildly off on those 5 weird days. (And they weren't trying to explain how the conditions worked, they just wanted to predict them.)
  - And if you are **sure** it was a typo, throw it out.
- But if you are doing a study and you want to show how height, exercise, genetics, etc.... affect weight, then you don't really want to throw out outliers because they give you important information.
  - But maybe you can make note of the outliers and report them and then refit the line without the outlier if the line really doesn't fit the rest of the data.
  - Because you are right, you don't really want a line that doesn't go through the data.
  - But it is dangerous from a moral and scientific point of view if you don't tell people about the outliers.
  - As long as you point them out you can say something like: "This regression line fits pretty well for most of the data. 1% of the time a value will come along that doesn't fit this trend, but hey, it's a crazy world, no system is perfect"
  - But make sure you report and maybe even try to explain the outliers (that aren't typos). Sometimes those give us important information about the world.
- Finally, we only consider the most basic tests and models in this class. There are hundreds of different statistical procedures out there. And some of them are designed to deal better with outliers than our basic linear regression.

20. A study looked at the resting metabolic rates (calories burned per 24 hours) and the lean body mass (kilograms) for 19 adults.



(a) Which variable is the response variable?

Rate

(b) What is the equation for the least square regression line?

$\hat{y} = 26.879x + 113.165$

(c) What is $\hat{\beta}_0$?

i. What do we use the symbol for?

y-intercept

ii. What is the value for $\hat{\beta}_0$?

113.165

iii. What is the interpretation? Is it a practical/logical interpretation?

for a mass of 0, rate = 113.165 / no, no one has a mass of 0.

(d) What is $\hat{\beta}_1$?

i. What do we use the symbol for?

Slope

ii. What is the value for $\hat{\beta}_1$?

26.879

iii. What is the interpretation? Is it a practical/logical interpretation?

for every grass unit, ~~the~~ rate increases by 26.879.

(e) Predict the metabolic rate for a person that weighs 55 kilograms..

$\hat{y}(55) = 1591.51$

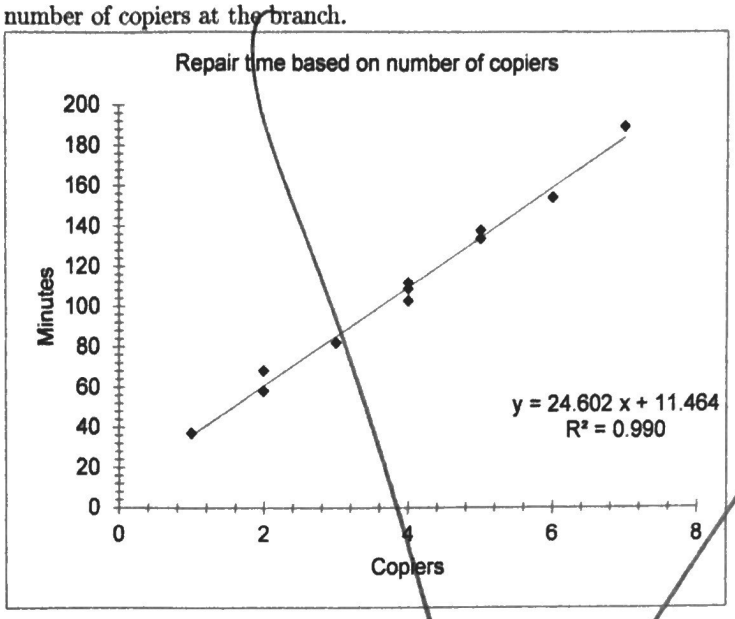
(f) Predict the metabolic rate for a person that weighs 3 kilograms.

Extrapolation

(g) What is the correlation? Interpret.
(Hint: Remember that the symbol for correlation is r. Should r be positive or negative for this data set?)

$r = \sqrt{.748} = .864$

21. **Optional:** A large company wants to predict the service time in minutes for each branch based on the number of copiers at the branch.



(a) What is the least squares line, or the linear regression equation?

(b) What is the intercept? Does it have a practical interpretation in this case?

(c) What is the slope? Interpret it.

(d) Find the predicted service time in minutes if the company has 8 copiers.