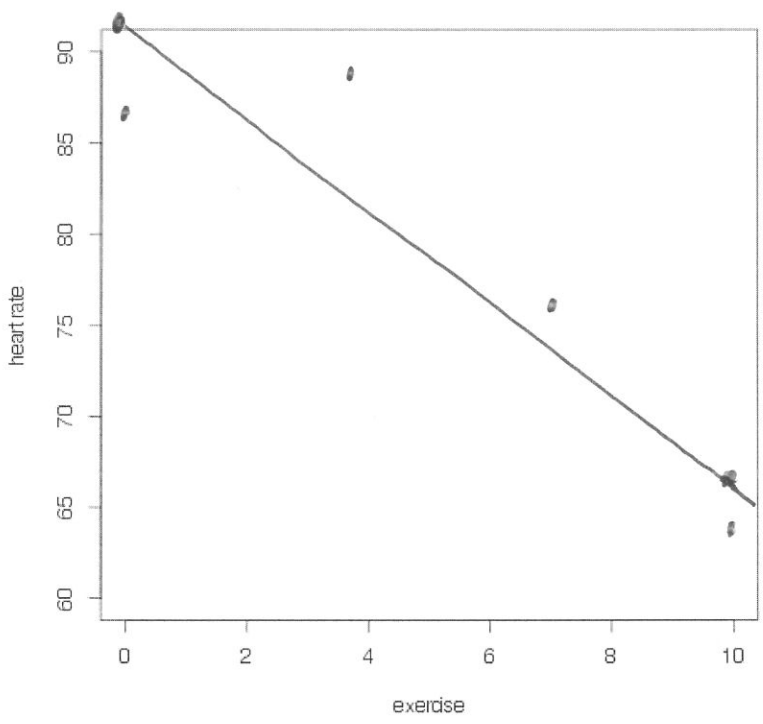
HW 2 Name: _____

Due Wednesday, Sept. 17 Red ID: _____

Use the following to answer questions 1 through 7

We are interested in predicting what a subjects resting heart rate would be, given the number of hours they exercise per week. We measured the resting heart rate and average number of hours exercised per week for 4 subjects, resulting in the data below:

Hours exercise	Resting heart rate
0	86
4	88
7	75
10	64



1. Identify the:
Explanatory variable: Hours exercise

Response variable: Resting heart rate

2. Find the linear model using the formulas for the slope and intercept provided in class. Show all work. You can round to just 1 decimal place during calculations.

$$a=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} \qquad b=\bar{y}-a\bar{x}$$

$$\bar{x}=\frac{0+4+7+10}{4}=5.25 \quad \bar{y}=\frac{86+88+75+64}{4}=78.25$$

$$a=\frac{(0-5.25)(86-78.25)+(4-5.25)(88-78.25)+(7-5.25)(75-78.25)+(10-5.25)(64-78.25)}{(0-5.25)^2+(4-5.25)^2+(7-5.25)^2+(10-5.25)^2}=-2.31$$

$$b=78.25-(-2.31)(5.25)=90.38$$

$$\hat{y}=-2.3(x)+90.4$$

3. Plot the points on the scatterplot provided. Draw the line given by the linear model on this scatterplot.

Find predicted values at any 2 points and plot them

At $x=0$, $\hat{y}=-2.3(0)+90.4=90.4$
At $x=10$, $\hat{y}=-2.3(10)+90.4=67.4$

4. For a subject who exercises 5 hours per week, what is their predicted resting heart rate?

$\hat{y}=-2.3(5)+90.4=78.9$

5. One subject who exercises 3.5 hours a week has an actual resting heart rate of 78. Find the residual for this subject, and state whether we made an over or an under prediction.

$e=y-\hat{y}$ $\hat{y}=-2.3(3.5)+90.4=82.4$

$e=78-82.4=-4.4$ overprediction

6. One subject who exercises 6 hours a week has a residual of 2.4. What is the actual resting heart rate for this subject?

$e=y-\hat{y}$ $\hat{y}=-2.3(6)+90.4=76.6$

$2.4=y-76.6$

$y=79$

7. Find the value of the correlation coefficient for this model.

r = -0.89

From calculator

8. We are interested in predicting the number of days it takes for a soybean plant to reach maturity from sprouting using either the amount of nitrogen found in the soil, or the amount of phosphorous. In a study we recorded the days to reach maturity, as well as the amount of nitrogen and phosphorous found in the soil, resulting in the following linear models:

height = 68 - 1.2(phosphorus) R^2 = 87% height = 64 - 0.8(nitrogen) R^2 = 93%

- Find and interpret the correlation coefficient between height and nitrogen

$r=\pm\sqrt{(R^2)}=\pm\sqrt{(0.93)}=-0.96$ (negative since slope is negative, so negative relationship)

There is a strong, negative, linear relationship between height and nitrogen.

- If you had to predict the number of days it takes for a soybean plant to reach maturity, would you use the amount of phosphorous, or the amount of nitrogen, in the soil, and why?

Nitrogen, higher R^2 so better predictive power

9. We measure the average number of TV's owned per person and average life expectancy for each of the world's nations. There is a high positive correlation between average number of TV's owned and life expectancy.

- Does this mean owning more TV's causes an increase in life expectancy?

No correlation does not imply causation

- What might be responsible for this correlation?

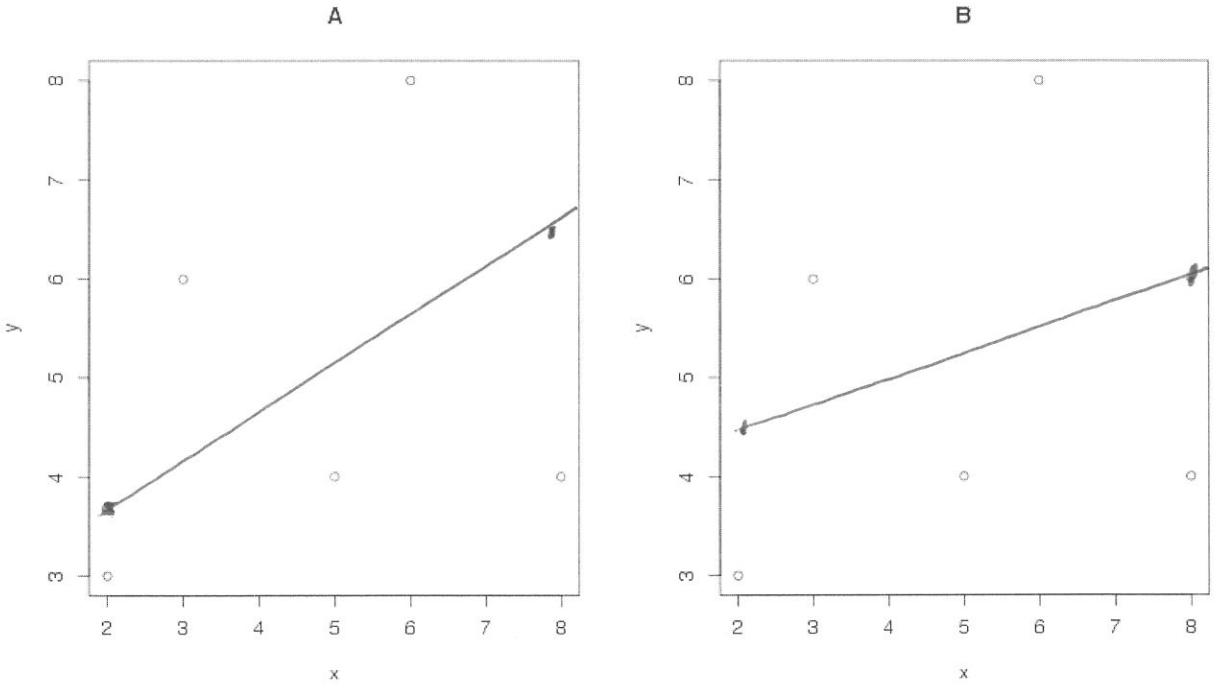
A lurking variable could be income. Nations with higher average income will have citizens who are more likely to be able to afford to buy a TV, and nations with higher income will tend to have better healthcare, education, sanitation services, etc.

10. In linear regression, how do we determine the which linear model best fits the data?

Minimize the sum of squared errors, SSE

11. Consider the following data, which has been plotted in each of the scatterplots below:

x	y
2	3
3	6
5	4
6	8
8	4



We would like to determine which of the following two linear models best fits the data:

model A: $\hat{y}=0.4(x)+3$ model B: $\hat{y}=0.2(x)+4$

- Plot the line given by model A on scatterplot A. Plot the line given by model B on scatterplot B.

model A: At x = 2, $\hat{y}=0.4(2)+3=3.8$ At x = 8, $\hat{y}=0.4(8)+3=6.2$

model B: At x = 2, $\hat{y}=0.2(2)+4=4.4$ At x = 8, $\hat{y}=0.2(8)+4=5.6$

- Find the sum of squared errors for model A, and for model B.

model A:	model B:
$\hat{y}_1=0.4(2)+3=3.8$ $\hat{y}_2=0.4(3)+3=4.2$ $\hat{y}_3=0.4(5)+3=5$ $\hat{y}_4=0.4(6)+3=5.4$ $\hat{y}_5=0.4(8)+3=6.2$ $SSE=\sum(y_i-\hat{y}_i)^2$ $=(3-3.8)^2+(6-4.2)^2+(4-5)^2+(8-5.4)^2+(4-6.2)^2=16.48$	$\hat{y}_1=0.2(2)+4=4.4$ $\hat{y}_2=0.2(3)+4=4.6$ $\hat{y}_3=0.2(5)+4=5$ $\hat{y}_4=0.2(6)+4=5.2$ $\hat{y}_5=0.2(8)+4=5.6$ $SSE=\sum(y_i-\hat{y}_i)^2$ $=(3-4.4)^2+(6-4.6)^2+(4-5)^2+(8-5.2)^2+(4-5.6)^2=15.32$

- Which linear model would you use, and why?

Linear model B, it has a smaller sum of squared errors

12. We are interested in estimating subjects bone density using their age and amount of calcium consumed on a daily basis. Using the data below, the following linear model was obtained:

density = 895 - 3(age) + 210(calcium)

Age	Calcium (g)	Density (mg/cm^2)	Predicted	Residual
38	1.3	1055.7	$\hat{y}=895-3(38)+210(1.3)=1054$	$e=y-\hat{y}=1055.7-1054=1.7$
54	0.9	925.3	$\hat{y}=895-3(54)+210(.9)=922$	$925.3-922=3.3$
62	0.7	853.9	$\hat{y}=895-3(62)+210(.7)=856$	$853.9-856=-2.1$
65	0.7	844.2	$\hat{y}=895-3(65)+210(.7)=847$	$844.2-847=-2.8$
71	1.0	888.3	$\hat{y}=895-3(71)+210(1.0)=892$	$888.3-892=-3.7$
78	0.8	831.7	$\hat{y}=895-3(78)+210(0.8)=829$	$831.7-829=2.7$
84	1.0	853.9	$\hat{y}=895-3(84)+210(1.0)=853$	$853.9-853=0.9$

Plot the residuals vs predicted values using the axis below