CHAPTER 5:

Regression

Basic Practice of Statistics

7th Edition

Lecture PowerPoint Slides

In chapter 5, we cover …

- Regression lines
- The least-squares regression line
- Examples of technology
- Facts about least-squares regression
- Residuals
- Influential observations
- Cautions about correlation and regression
- Association does not imply causation
- Correlation, prediction, and big data*

Linear Regression

- Objective: To *quantify* the linear relationship between an explanatory variable (x) and response variable (y).

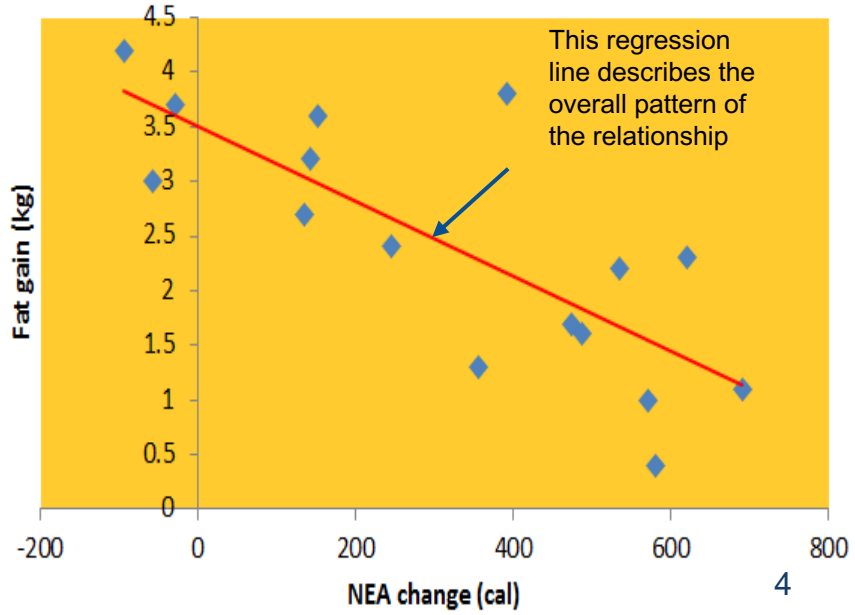
- We can then ***predict*** the average response for all subjects with a given value of the explanatory variable.

Regression line (1 of 2)

A regression line is a straight line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x, when we believe the relationship between y and x is linear.

Example: Predict the gain in fat (in kilograms) based on the change in Non-Exercise Activity (NEA change, in calories).

- **If the NEA change is 400 calories, what is the expected fat gain?**



Regression line (2 of 2)

REVIEW OF STRAIGHT LINES

- Suppose that y is a response variable (plotted on the vertical axis) and x is an explanatory variable (plotted on the horizontal axis). A straight line relating y to x has an equation of the form

$$y = a + bx$$

- In this equation, b is the **slope**—the amount by which y changes when x increases by one unit. The number a is the **intercept**—the value of y when $x = 0$.

- To **plot the line** on the scatterplot, use the equation to find the predicted y for two values of x, one near each end of the range of x in the data. Plot a line each y above its x-value, and draw the line through the two points.

Apply your knowledge

5.1 City mileage, highway mileage. We expect a car's highway gas mileage to be related to its city gas mileage. Data for all 1040 vehicles in the government's *2010 Fuel Economy Guide* give the regression line

$$\text{highway mpg} = 6.554 + (1.016 \times \text{city mpg})$$

for predicting highway mileage from city mileage.

(a) What is the slope of this line? Say in words what the numerical value of the slope tells you.

(b) What is the intercept? Explain why the value of the intercept is not statistically meaningful.

(c) Find the predicted highway mileage for a car that gets 16 miles per gallon in the city. Do the same for a car with city mileage 28 mpg.

(d) Draw a graph of the regression line for city mileages between 10 and 50 mpg. (Be sure to show the scales for the x and y axes.)

Regression Line

The regression line for body fat index (bfi) as a function of triceps thickness (in mm) is given below.

$$\text{bfi} = 14.59 + (0.74 \times \text{triceps thickness})$$

In this equation, bfi is the _____ variable and triceps thickness is the _____ variable.

a) explanatory; response

b) response; explanatory

Regression Line

The regression line for body fat index (bfi) as a function of triceps thickness (in mm) is given below.

$$\text{bfi} = 14.59 + (0.74 \times \text{triceps thickness})$$

What is the average amount by which bfi changes when triceps thickness increases by 1 mm?

a) 0.74

b) 1.48

c) 14.59

d) 15.33

Regression Line

The regression line for body fat index (bfi) as a function of triceps thickness (in mm) is given below.

$$\text{bfi} = 14.59 + (0.74 \times \text{triceps thickness})$$

Predict the bfi for individuals with triceps thickness of 35 mm.

a) $14.59 + 0.74 = 15.33$

b) $14.59 + 35 = 49.59$

c) $14.59 + (0.74 \times 35) = 40.49$

d) $0.74 \times 35 = 25.9$

Regression Line

The regression line for body fat index (bfi) as a function of triceps thickness (in mm) is given below.

$$\text{bfi} = 14.59 + (0.74 \times \text{triceps thickness})$$

What is 14.59 in this equation?

a) the predicted bfi when triceps thickness = 0 mm

b) the bfi intercept

c) Both of the answer options are correct.

Regression Line

Consider the following regression line.

$$MPG = 48.74 - 0.0082 \text{ (weight)}$$

Calculate the predicted value of MPG when the weight is 3000 pounds.

a) 30.5
b) 24.14
c) 28.2
d) 17.5

Regression Line

Consider the following regression line, where y is motor vehicle fatalities per 100,000 residents and x is percent population for ages 18 to 24.

$$y = -10 + 2.7x$$

Which of the following correctly predicts the number of motor vehicle fatalities in a state with 25% population age 18 to 24?

a) $2.7(25) = 67.5$
b) $-10 + 2.7(25) = 57.5$
c) $-10(25) = -250$
d) $-10(25) + 2.7 = -247.3$

Regression Line

Consider the following regression line, where y is motor vehicle fatalities per 100,000 residents and x is percent population for ages 18 to 24.

$$y = -10 + 2.7x$$

For every one unit increase in x, the model predicts a change of _____ in the predicted value of y.

a) -10
b) $2.7(-10)$
c) 2.7
d) $-10 + 2.7x$

Least Squares

- Used to determine the “best” line
- **Least Squares**: use the line that minimizes the sum of the squares of the vertical distances of the data points from the line
- Draw a picture

The least-squares regression line

The least-squares regression line of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

EQUATION OF THE LEAST-SQUARES REGRESSION LINE

- We have data on an explanatory variable x and a response variable y for n individuals. From the data, calculate the means $\bar{x}$ and $\bar{y}$ and the standard deviations s_x and s_y of the two variables and their correlation r. The least-squares regression line is the line

$$\hat{y} = a + bx$$

- with slope

$$b = r\frac{s_y}{s_x}$$

- and intercept

$$a = \bar{y} - b\bar{x}$$

Least Squares Regression Line

- Regression equation: $\hat{y} = a + bx$
  - ***x*** is the value of the explanatory variable
  - ***"y-hat"*** is the average value of the response variable *(predicted response for a value of x)*
  - note that ***a*** and ***b*** are just the intercept and slope of a straight line
  - note that ***r*** and ***b*** are not the same thing, but their signs will agree

Prediction via regression line

- For the non-exercise activity example, the least-squares regression line is:

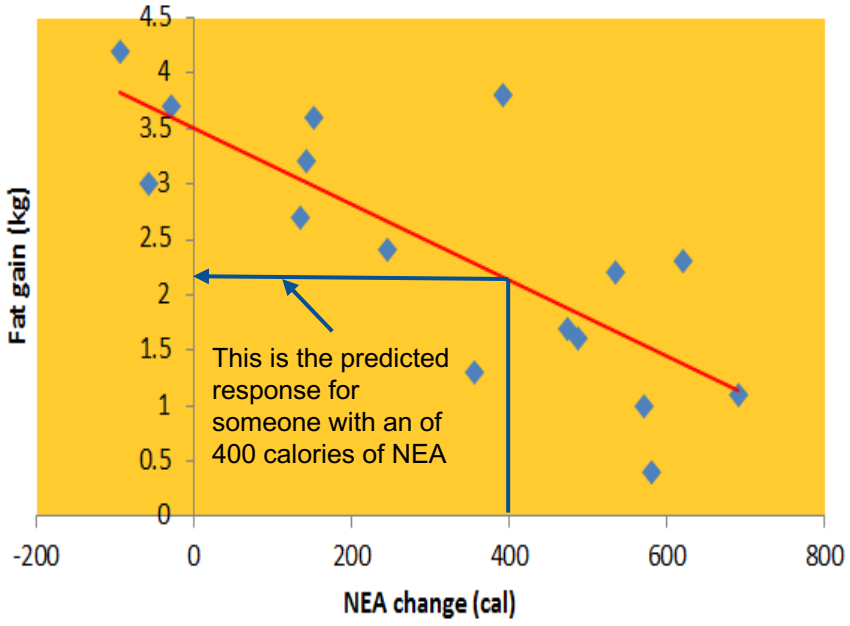
$$\hat{y} = 3.5051 - 0.0034x$$

- $\hat{y}$ is the predicted fat gain (in kg) with x calories of Non-Exercise Activity

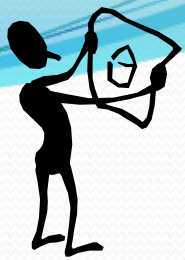
- Suppose we know someone has an increase of 400 calories of NEA. What would we predict for fat gain?

- For someone with 400 calories of NEA, we would predict fat gain of:

$$3.5051 - 0.0034(400) = 2.1451 \text{ kg}$$



Regression Calculation Case Study



Per Capita Gross Domestic Product and Average Life Expectancy for Countries in Western Europe

Regression Calculation Case Study

Country	Per Capita GDP (*x*)	Life Expectancy (*y*)
Austria	21.4	77.48
Belgium	23.2	77.53
Finland	20.0	77.32
France	22.7	78.63
Germany	20.8	77.17
Ireland	18.6	76.39
Italy	21.5	78.51
Netherlands	22.0	78.15
Switzerland	23.8	78.99
United Kingdom	21.2	77.37

Regression Calculation Case Study

Linear regression equation:

$$\bar{x} = 21.52 \qquad \bar{y} = 77.754 \qquad r = 0.809$$

$$s_x = 1.532 \qquad s_y = 0.795$$

$$b = r\frac{s_y}{s_x} = (0.809)\left(\frac{0.795}{1.532}\right) = 0.420$$

$$a = \bar{y} - b\bar{x} = 77.754 - (0.420)(21.52) = 68.716$$

$$\hat{y} = 68.716 + 0.420x$$

Apply your knowledge

student's name	Choo	Kang	Lee	Park	Ryu
Time studied(hours)	2	10	6	8	5
Score	60	95	78	88	72

1) Draw scatterplot.

2) Caculate r.

3) Find b and a. And explain in words what the slope of the regression line tell us.

4) Draw the regression line on the scatterplot of 1)

5) Another person in their group studies for 7 hours. What is his predicted score?

Coral reefs. Exercises 4.2 and 4.10 discuss a study in which scientists examined data on mean sea surface temperatures (in degrees Celsius) and mean coral growth (in millimeters per year) over a several-year period at locations in the Red Sea. Here are the data:[2]  CORAL

Sea surface temperature	29.68	29.87	30.16	30.22	30.48	30.65	30.90
Growth	2.63	2.58	2.60	2.48	2.26	2.38	2.26

1) Draw scatterplot.

2) Caculate r.

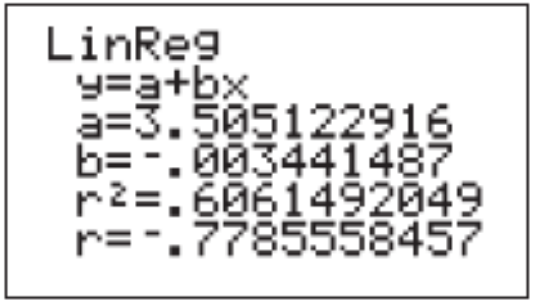
3) Find b and a. And explain in words what the slope of the regression line tell us.

4) Draw the regression line on the scatterplot of 1)

5) Another sea surface temperature is 30 degree in Celsius. What is its predicted growth?

Examples of technology

Texas Instruments Graphing Calculator



Minitab

Minitab

Regression Analysis: fat versus nea

The regression equation is
fat = 3.51 - 0.00344 nea

Predictor	Coef	SE Coef	T	P
Constant	3.5051	0.3036	11.54	0.000
nea	-0.0034415	0.0007414	-4.64	0.000

S = 0.739853 R-Sq = 60.6% R-Sq (adj) = 57.8%

JMP

JMP

Linear Fit

Fat = 3.5051229 - 0.0034415*NEA

Summary of Fit

RSquare	0.606149
RSquare Adj	0.578017
Root Mean Square Error	0.739853
Mean of Response	2.3875
Observations (or Sum Wgts)	16

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.5051229	0.303616	11.54	<.0001*
NEA	-0.003441	0.000741	-4.64	0.0004*

Microsoft Excel

Excel

	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*Regression statistics*					
4	Multiple R	-0.778555846				
5	R Square	0.606149205				
6	Adjusted R Square	0.578017005				
7	Standard Error	0.739852874				
8	Observations	16				
9						
10		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
11	Intercept	3.505122916	0.303616403	11.54458	1.53E-08	
12	nea	-0.003441487	0.00074141	-4.64182	0.000381	
13						

CrunchIt!

CRUNCHIT! CRUNCHIT!

Export

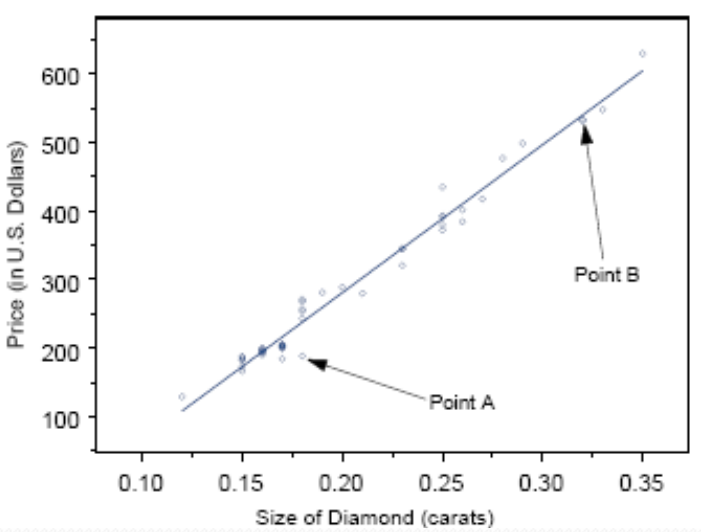
Fitted Equation:
Fat = 3.505 - 0.003441 * NEA

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.505	0.3036	11.54	<0.0001
NEA	-0.003441	0.0007414	-4.642	0.0003810

r-Squared: 0.6061 Adjusted r-Squared: 0.5780 sigma: 0.7399

Least-Squares Regression Line (1 of 2)

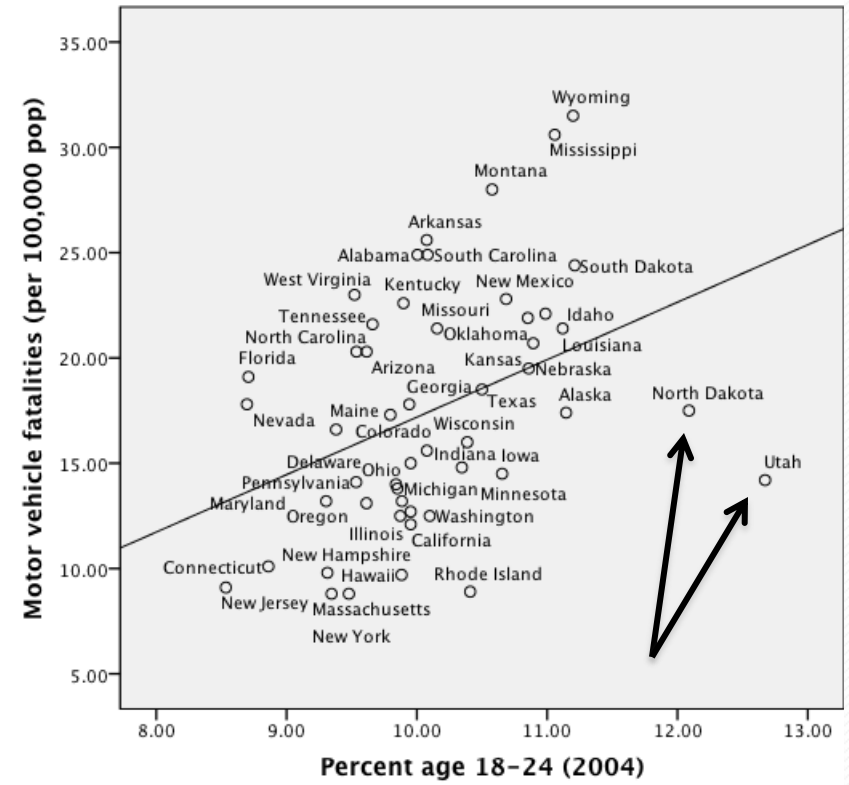
Look at the following least-squares regression line. Compare the squared errors from the points A and B.



a) Point A's squared errors would be greater than Point B's.
b) Point A's squared errors would be less than Point B's.
c) Point A's squared errors would be equal to Point B's.
d) The answer cannot be determined from the information given.

Least-Squares Regression Line (2 of 2)

What will happen to the slope of the least-squares regression line if we remove Utah and North Dakota?



a) increase (more positive)
b) decrease (more negative)
c) remain the same
d) The answer cannot be determined from the information given.

Facts about least-squares regression

- The distinction between explanatory variables and response variables is essential in regression.
- There is a close connection between correlation and the slope of the least-squares line. The slope is

$$b = r\frac{s_y}{s_x}$$

- The slope b and correlation r always have the same sign.
- Along the regression line, a change of 1 standard deviation in x corresponds to a change of r standard deviations in y.
- The least-squares regression line always passes through $(\bar{x}, \bar{y})$ on the graph of y against x.

Facts about Least Squares

The least-squares regression line always passes through which point?

a) $(0, 0)$
b) $(\overline{x}, 0)$
c) $(0, \overline{y})$
d) $(\overline{x}, \overline{y})$

Facts about Least Squares

If y is the response variable and x is the explanatory variable, what does the regression line allow you to do that correlation does not?

a) calculate the exact value of x, given a value for y

b) calculate the predicted value of x, given a value for y

c) calculate the exact value of y, given a value for x

d) calculate the predicted value of y, given a value for x

The square of the correlation

- The square of the correlation, r^2, is the fraction of the variation in the values of y that is explained by the least-squares regression of *y* on *x*.

$$r^2 = \frac{\text{variation in } \hat{y} \text{ along the regression line as } x \text{ varies}}{\text{total variation in observed values of } y}$$

- *Caution:* You can find a regression line for any relationship between two quantitative variables, but the usefulness of the line for prediction depends on the strength of the linear relationship.
- r^2 near 0 means a *weak* linear relationship.
- r^2 near 1 means a *strong* linear relationship.
- Note that, with r^2, we lose information about the *direction* of the association.
- r=1: R^2=1: regression line explains all (100%) of the variation in *y*
- r=.7: R^2=.49:regression line explains almost half (50%) of the variation in *y*

Facts about Least Squares

Which of the following best indicates the usefulness of a regression line for prediction?

a) twice the correlation coefficient: $2r$

b) square of the correlation coefficient: r^2

c) square root of the correlation coefficient: $\sqrt{r}$

Facts about Least Squares

The squared correlation (r^2) of bfi and triceps thickness is 0.72. The squared correlation (r^2) of bfi and neck thickness is 0.18. Which regression line is more useful for prediction of bfi?

a) bfi vs. triceps thickness
b) bfi vs. neck thickness
c) Both answer options are equally good.
d) The answer cannot be determined from the information given.

Residuals (1 of 2)

A residual is the difference between an observed value of the response variable and the value predicted by the regression line. That is, a residual is the prediction error that remains after we have chosen the regression line:

$$\begin{aligned} \text{residual} &= \text{observed } y - \text{predicted } y \\ &= y - \hat{y} \end{aligned}$$

Residuals represent “leftover" variation in the response after fitting the regression line.

Residuals (2 of 2)

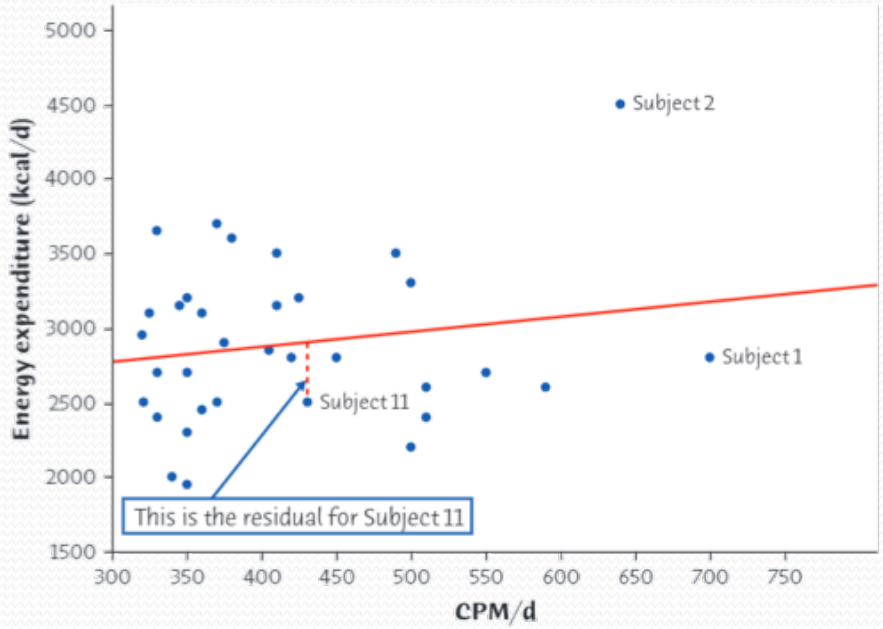
The residuals from the least-squares line have a special property: **the mean of the least-squares residuals is always zero**.

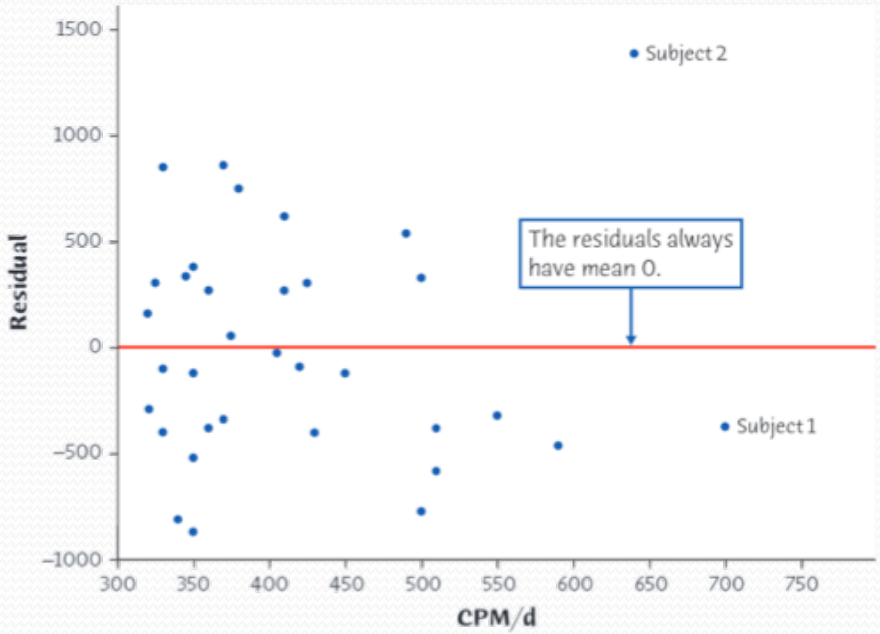
RESIDUAL PLOTS

- A residual plot is a scatterplot of the regression residuals against the explanatory variable.
- Residual plots help us assess how well a regression line fits the data.
- Look for a "random" scatter around zero.

Residual plot

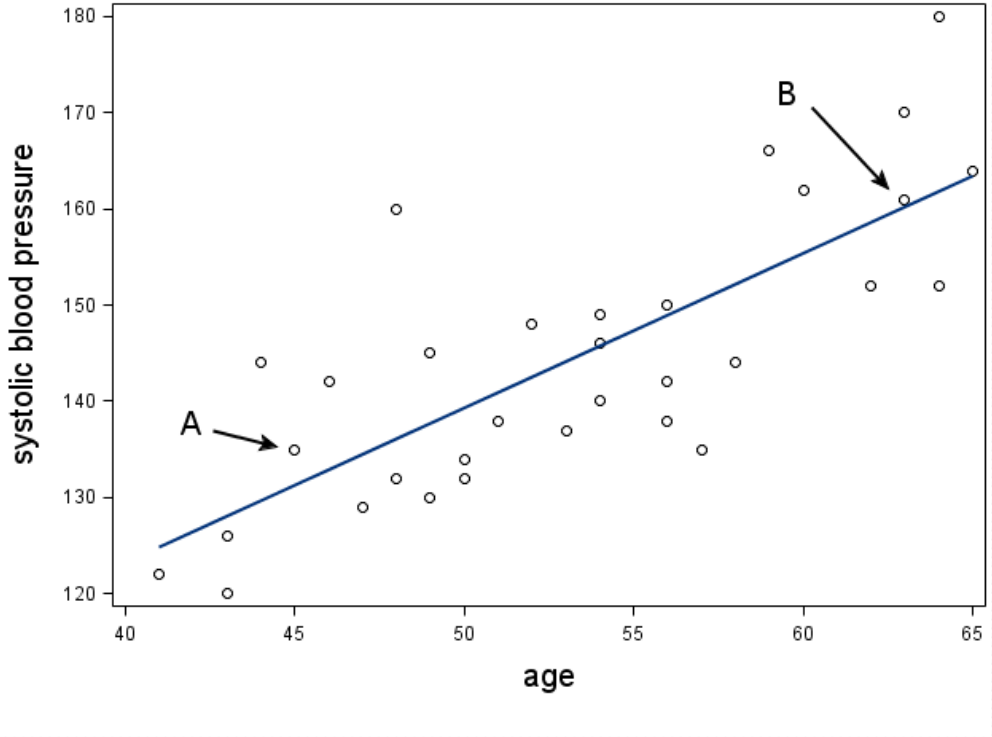
PHYSICAL ACTIVITY AND WEIGHT LOSS





Residuals (1 of 4)

Which point, A or B, has a larger residual?



a) point A
b) point B
c) The answer cannot be determined from the information given.

Residuals (2 of 4)

Residual equals _____.

a) $\hat{y} - y$

b) $y - \hat{y}$

c) $\hat{x} - x$

d) $x - \hat{x}$

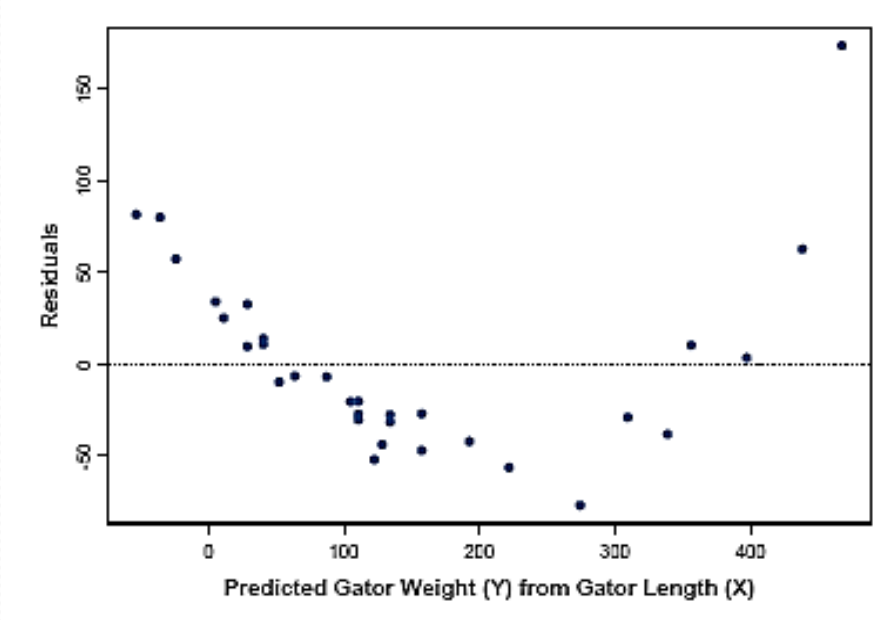
Residuals (3 of 4)

The residual plot is used to:

a) examine the relationship between two variables.
b) identify the mean and spread of the residuals.
c) check for independence of observations.
d) examine how well a regression line fits the data.

Residuals (4 of 4)

What does this residual plot indicate?



a) The relationship between *X* and *Y* can be modeled with a straight line.

b) The relationship between *X* and *Y* cannot be modeled with a straight line.

c) Neither of the answer options is correct.

Apply your knowledge

student's name	Choo	Kang	Lee	Park	Ryu
Time studied(hours)	2	10	6	8	5
Score	60	95	78	88	72

1) Find the residuals.

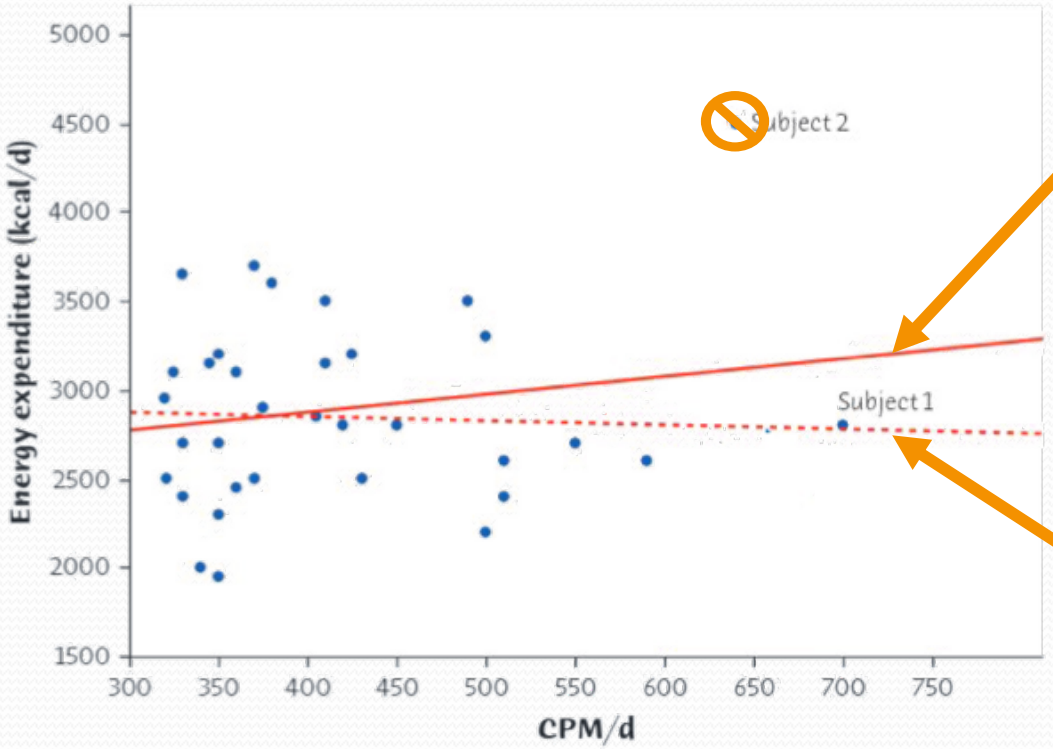
2) Draw the residual plot.

Influential observations

- An observation is influential for a statistical calculation if removing it would markedly change the result of the calculation.
- The result of a statistical calculation may be of little practical use if it depends strongly on a few influential observations.
- Points that are outliers, in either the x or the y direction of a scatterplot, are often influential for the correlation. Points that are outliers in the x direction are often influential for the least-squares regression line.

Outliers and influential points

PHYSICAL ACTIVITY AND WEIGHT LOSS



From all of the data

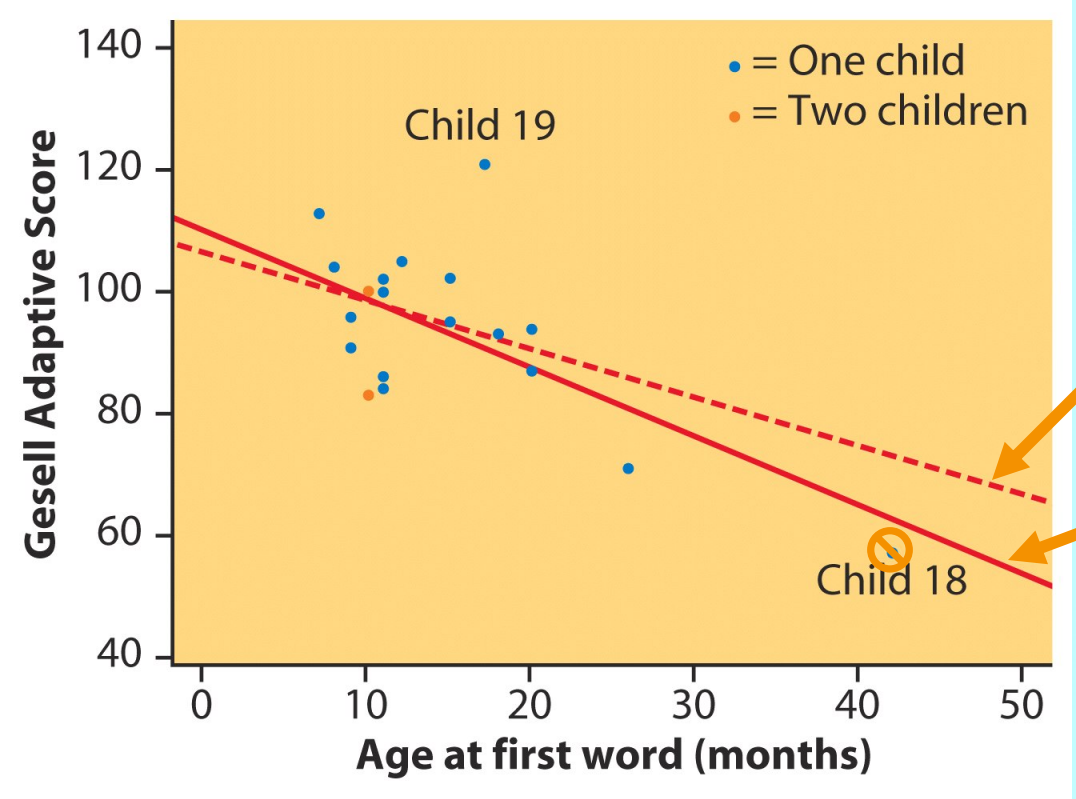
$r = 0.181$

After removing subject 2

$r = -0.043$

Outliers:
Case Study

Gesell Adaptive Score and Age at First Word



After removing child 18

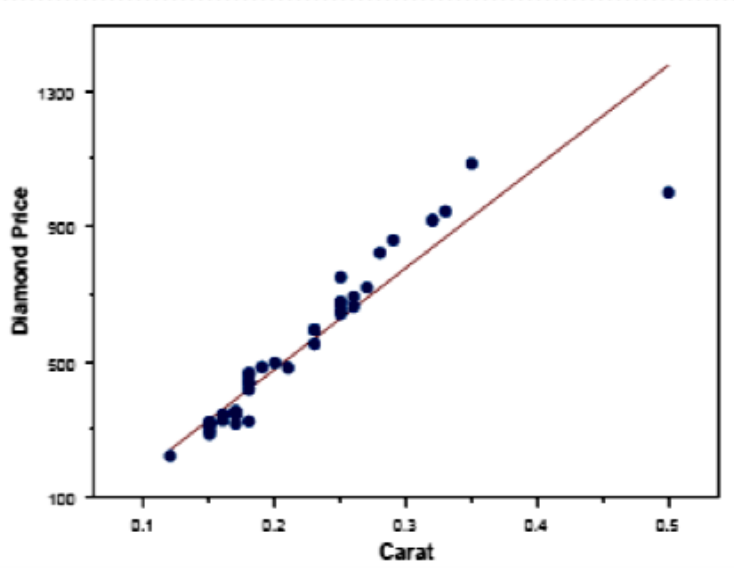
$r^2 = 11\%$

From all the data

$r^2 = 41\%$

Influential Observation (1 of 3)

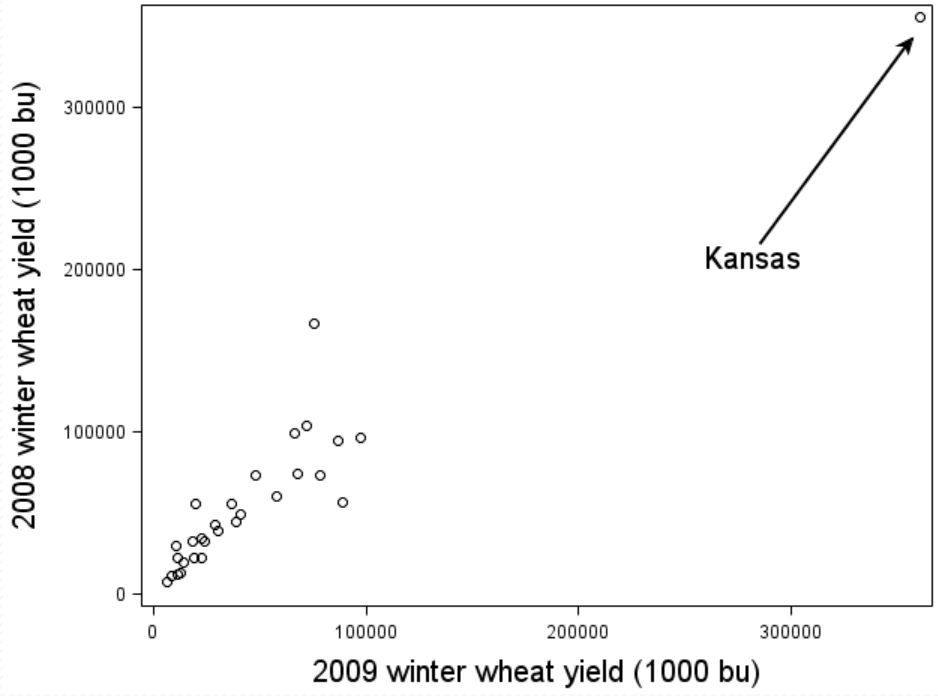
The graph indicates the presence of:



a) a non-influential outlier.
b) an influential outlier.
c) a lurking variable.
d) an outlier in the *y* direction only.

Influential Observation (2 of 3)

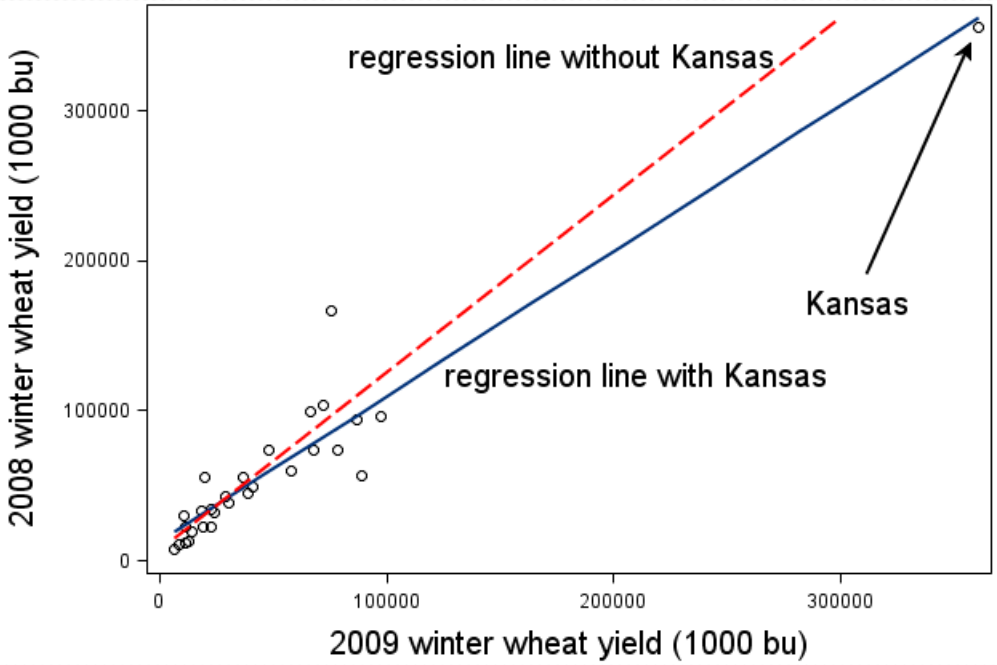
If Kansas was removed from the plot, the correlation would _____.



a) increase
b) decrease
c) The answer cannot be determined from the information given.

Influential Observation (3 of 3)

Kansas is an influential outlier.



a) true
b) false
c) The answer cannot be determined from the information given.

Cautions about correlation and regression

- Correlation and regression lines describe only *linear* relationships.
- Correlation and least-squares regression lines are not *resistant*.
- Beware *ecological correlation*, or correlation based on *averages* rather than individuals.
- A correlation based on averages rather than on individuals is called an ecological correlation.

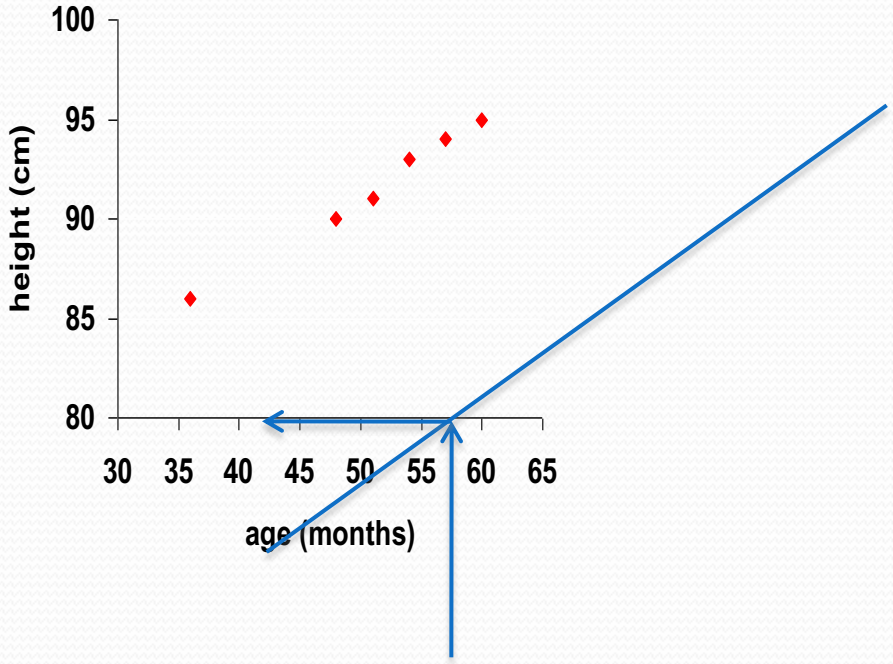
Cautions about correlation and regression

- Beware of *extrapolation*—predicting outside of the range of *x*.
- Beware of *lurking variables*—these have an important effect on the relationship among the variables in a study, but are not included in the study.
- A lurking variable is a variable that is not among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables.
- Correlation does not imply causation!
- An association between an explanatory variable *x* and a response variable *y*, even if it is very strong, is not by itself good evidence that changes in x actually cause changes in *y*.

Caution: beware of extrapolation (1 of 2)

Sarah's height was plotted against her age.

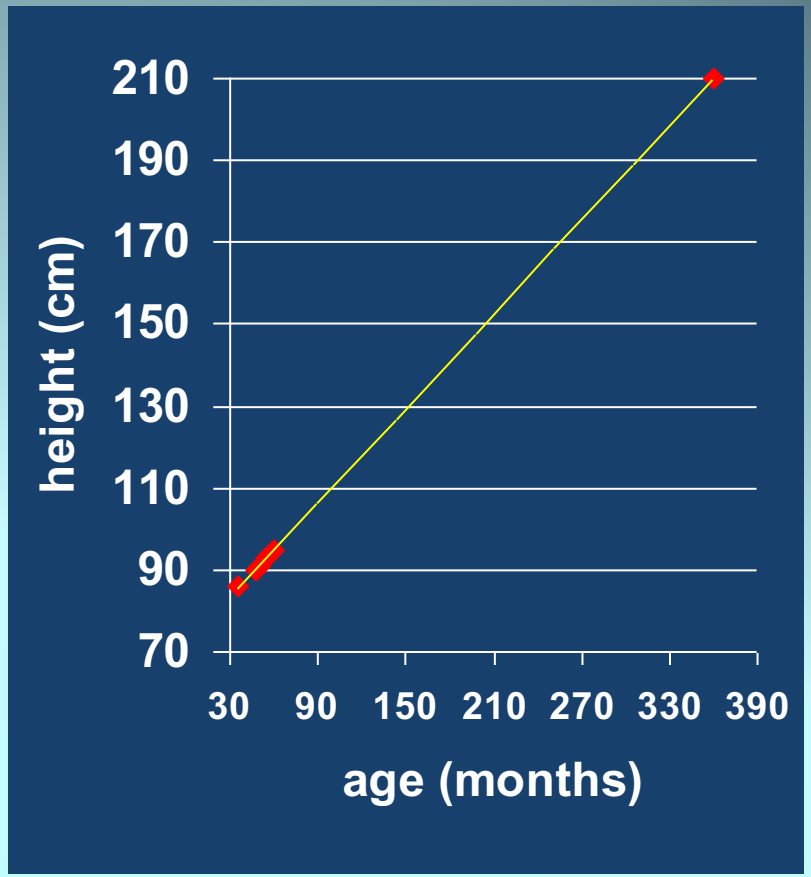
- Can you predict her height at age 42 months?
- Can you predict her height at age 30 years (360 months)?



Caution: beware of extrapolation (2 of 2)

- Regression line:
  $\hat{y}$ **= 71.95 + .383** ***x***
- Predicted height at age 42 months? $\hat{y}$ **= 88**
- Predicted height at age 30 years? $\hat{y}$ **= 209.8**

She is predicted to be 6’ 10½” at age 30!



Caution: beware of lurking variables

Example: Meditation and Aging
(*Noetic Sciences Review*, Summer 1993, p. 28)

- Explanatory variable: observed meditation practice (yes/no)
- Response variable: level of age-related enzyme

General concern for one's well-being may also be affecting the response (and the decision to try meditation).

Correlation does not imply causation

- Even very strong correlations may not correspond to a real causal relationship (changes in *x* actually causing changes in *y*).
- Correlation may be explained by a lurking variable

Social Relationships and Health

House, J., Landis, K., and Umberson, D. "Social Relationships and Health," *Science*, Vol. 241 (1988), pp. 540-545.

Does lack of social relationships cause people to become ill?

(There was a strong correlation.)

- **Or**, are unhealthy people less likely to establish and maintain social relationships? *(reversed relationship)*
- **Or**, is there some other factor that predisposes people both to have lower social activity and to become ill?

Evidence of causation

A properly conducted **experiment** may establish causation.

Other considerations when we cannot do an experiment:

- The association is *strong*.
- The association is *consistent*.
- *Higher* doses are associated with *stronger* responses.
- Alleged cause *precedes* the effect *in time*.
- Alleged cause is *plausible* (reasonable explanation).

Correlation, prediction, and big data

- Massive databases, or "big data," that are collected by Google, Facebook, credit card companies, and others contain petabytes (10^{15} bytes of data) and continue to grow in size.
- Proponents for big data often make the following claims for its value:
  - There is no need to worry about causation, because correlations are all we need to know for making accurate predictions.
  - Scientific and statistical theory is unnecessary because, with enough data, the numbers speak for themselves.
- True, to a point, but we have to be aware of pitfalls: if you have no idea what is behind a correlation, you have no idea what might cause prediction to fail; bias (systematic departures from what is true about a particular group because data are not representative of the group) is not eliminated with big data; and even the perception of the infallibility of big data itself has been subject to "confirmation bias," with its practitioners quick to report successes and (perhaps) not so quick to report failures.

Causation (1 of 3)

A 1999 observational study reported in the prestigious journal *Nature* found a positive correlation between ambient lighting in a baby's room and the degree of myopia (nearsightedness) later in life. What should you do with this information?

a) Accept it without question.
b) Accept it tentatively.
c) Throw away your baby's nightlight to prevent myopia.
d) Wait until an experiment is done.

Causation (2 of 3)

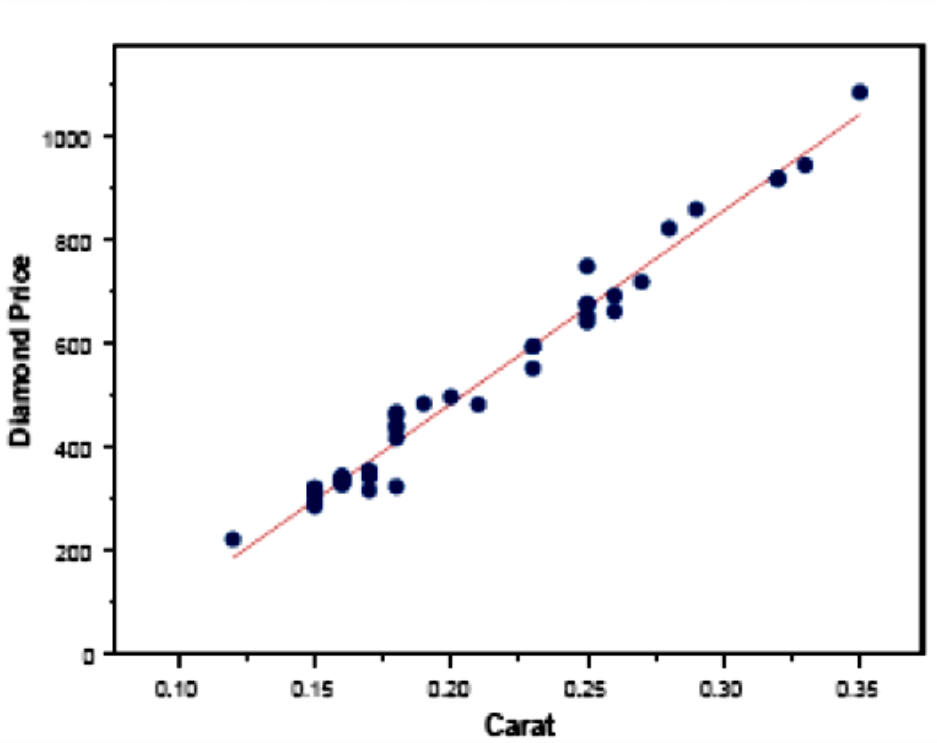
The myopia study referenced on the previous slide was later criticized. A second study showed no relationship. One of the following was suggested as a lurking variable in the first study. Which is it?

(Think carefully about which variable might have influenced the relationship. Who would have tended to both use nightlights and pass on myopia to their children?)

a) socioeconomic status
b) parental myopia
c) gender
d) nationality

Causation (3 of 3)

The graph shows the least-squares regression line for predicting diamond price from diamond size, for diamonds that are 0.35 carat or less. Using this relationship to predict the price of a diamond that is 1 carat is an example of _____.



a) prediction
b) an influential observation
c) a lurking variable
d) extrapolation

Big Data (1 of 2)

We don't worry about correlation versus causation in big data, because correlation is all we need for _____.

a) unbiased estimation
b) making predictions
c) significance
d) extrapolation

Big Data (2 of 2)

Despite the very large data sets associated with big data, _____ due to recording errors and non-random sampling does not go away.

a) correlation
b) bias
c) significance
d) regression