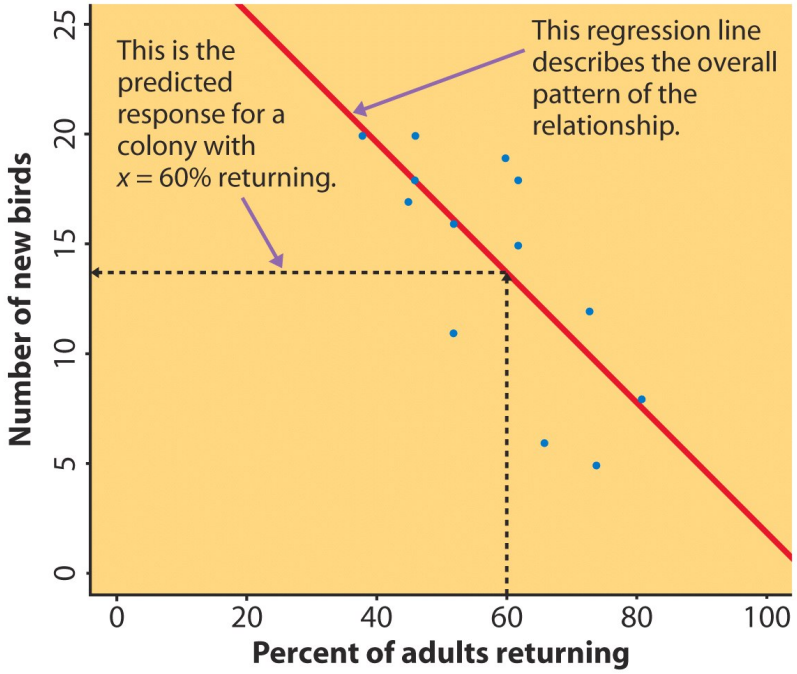
Chapter 5. Regression

A **regression line** is a straight line that describes how a response variable *y* changes as an explanatory variable *x* changes.

We can use a regression line to predict the value of *y* for a given value of *x*.

Example: Predict the number of new adult birds that join the colony based on the percent of adult birds that return to the colony from the previous year.

- **If 60% of adults return, how many new birds are predicted?**



Regression Line

The least-squares regression line (LSRL) which minimize the sum of the squares of the vertical distances of the data points from the line:

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

- ***x*** is the value of the explanatory variable.
- ***"y-hat"*** is the predicted value of the response variable for a given value of *x*.

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

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

- ***b*** is the **slope,** the amount by which *y* changes for each one-unit increase in *x.*
- ***a*** is the **intercept,** the value of *y* when *x = 0*
- ***r*** is their **correlation.**

Prediction via Regression Line

For the returning birds example, the LSRL is

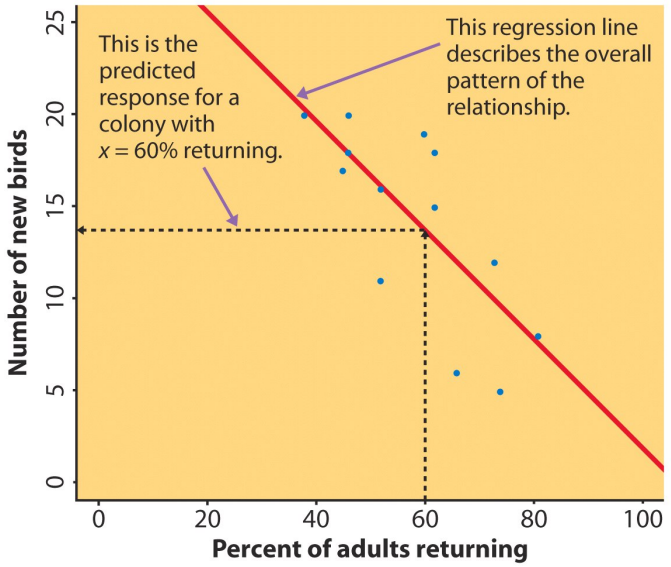
y-hat* = 31.9343 - 0.3040*x

- ✓ ***y-hat*** is the predicted number of new birds for colonies with ***x*** percent of adults returning.

Suppose we know that an individual colony has 60% returning. What would we ***predict*** the number of new birds to be for just that colony?

For colonies with 60% returning, we ***predict*** the average number of new birds to be:

31.9343 - (0.3040)(60) = **13.69** birds



Using technology: Minitab

1. Enter these data in the worksheet
2. Click on the Stat tab and

Select Regression → Regression

↓	C1	C2	C3
	Temperature	Growth	
1	29,68	2,63	
2	29,87	2,58	
3	30,16	2,60	
4	30,22	2,48	
5	30,48	2,26	
6	30,65	2,38	
7	30,90	2,26	
8			

Regression Analysis: Growth versus Temperature

The regression equation is
Growth = 12,4 - 0,328 Temperature

intercept
b

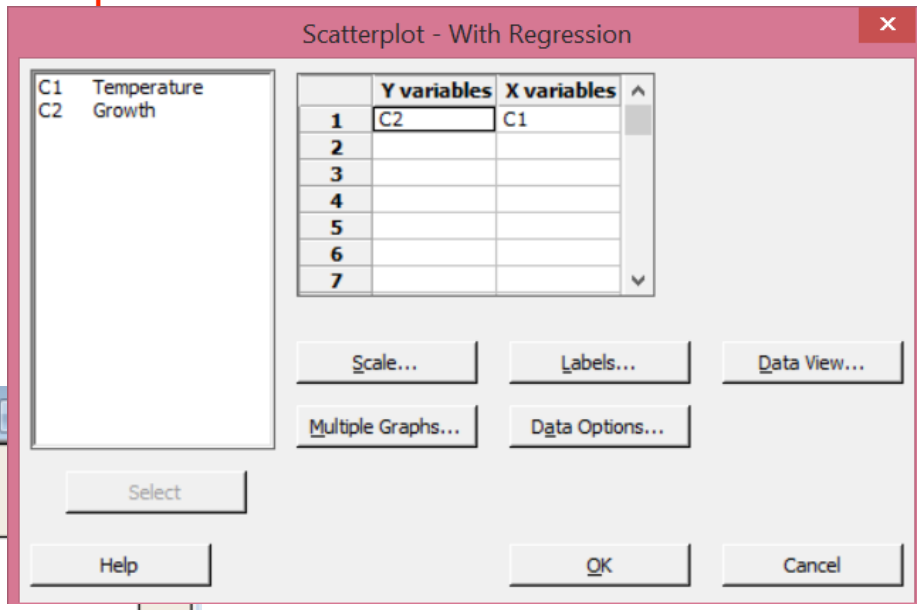
slope a

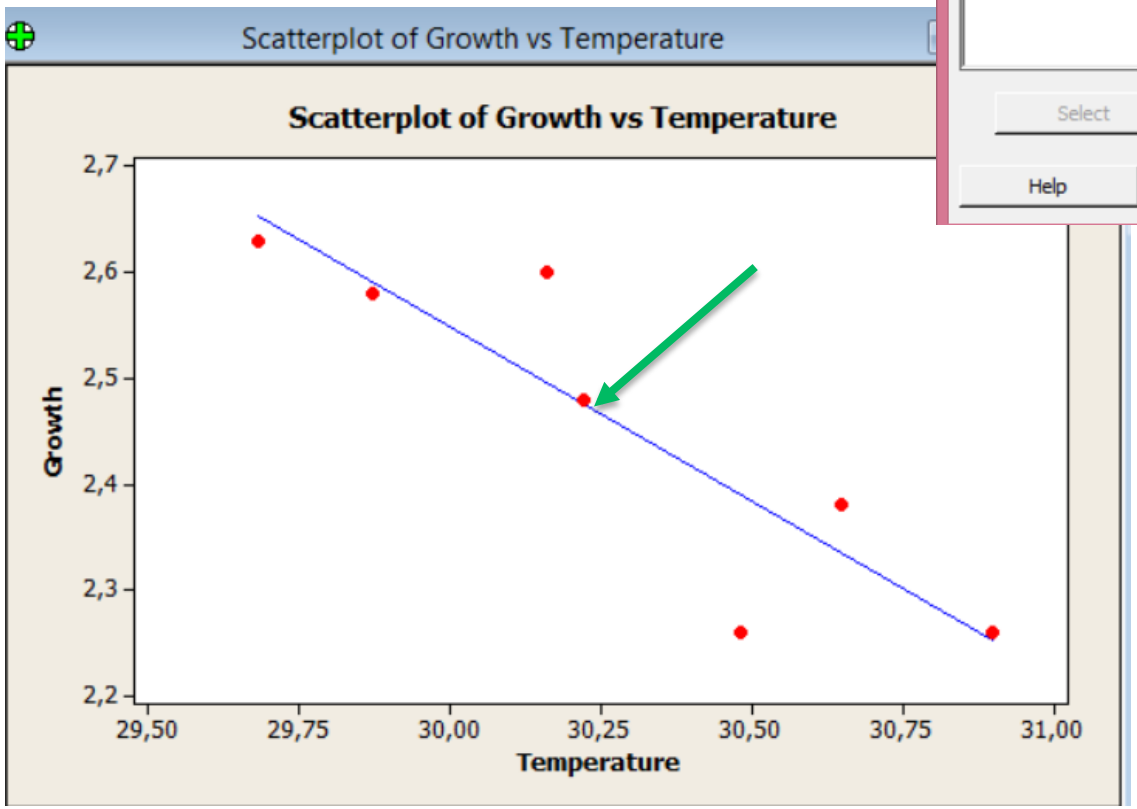
Predictor	Coef	SE Coef	T	P
Constant	12,376	2,255	5,49	0,003
Temperature	-0,32761	0,07448	-4,40	0,007

S = 0,0783747 R-Sq = 79,5% R-Sq(adj) = 75,4%

Using technology: Minitab

Graph tab → Scatterplot → With regression → put Y and X





Facts About Least-Squares Regression

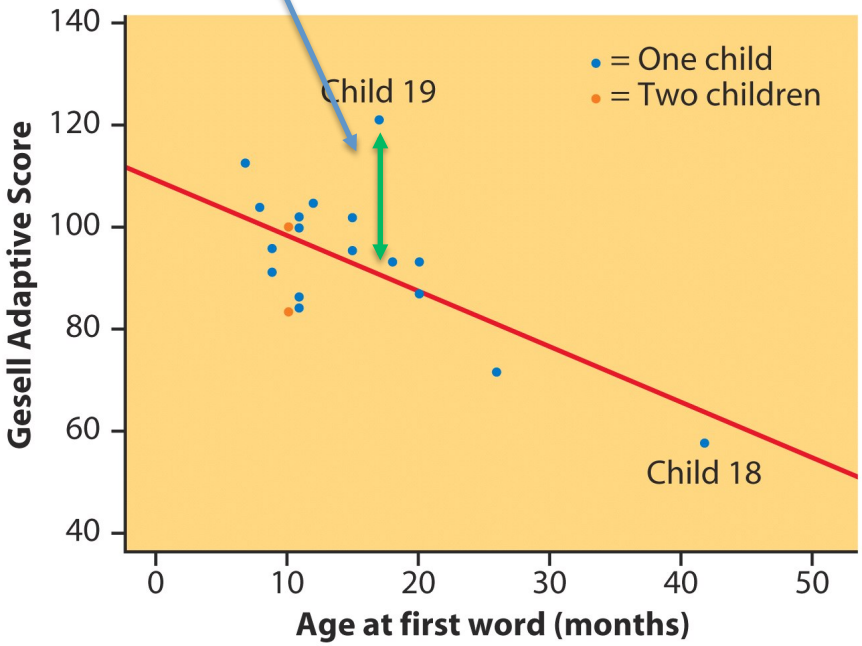
- The distinction between explanatory and response variables is essential.
- The LSRL always passes through (*x*-bar, *y*-bar).
- The square of the correlation, r^2, is the fraction of the variation in the values of *y* that is explained by the least-squares regression of *y* on *x*.
- r^2 is called the **coefficient of determination**.
- r=.7: r^2=.49: regression line explains 49% of the variation in *y*

Residuals: $y - \hat{y}$

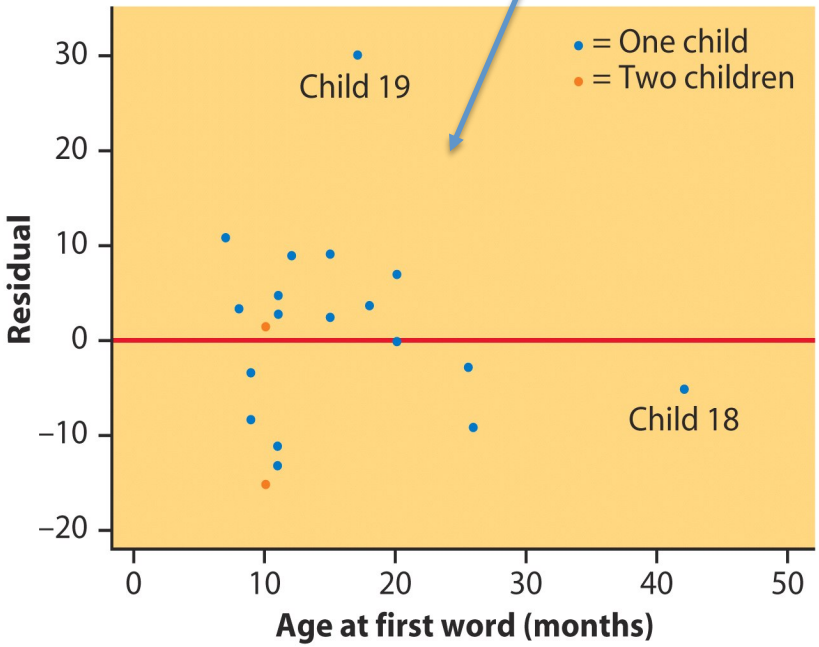
Gesell Adaptive Score and Age at First Word

Let X be the age in months a child speaks his/her first word and let Y be the Gesell adaptive score, a measure of a child's aptitude. How does the child's aptitude change with how long it takes them to speak?

Residual for Child #19

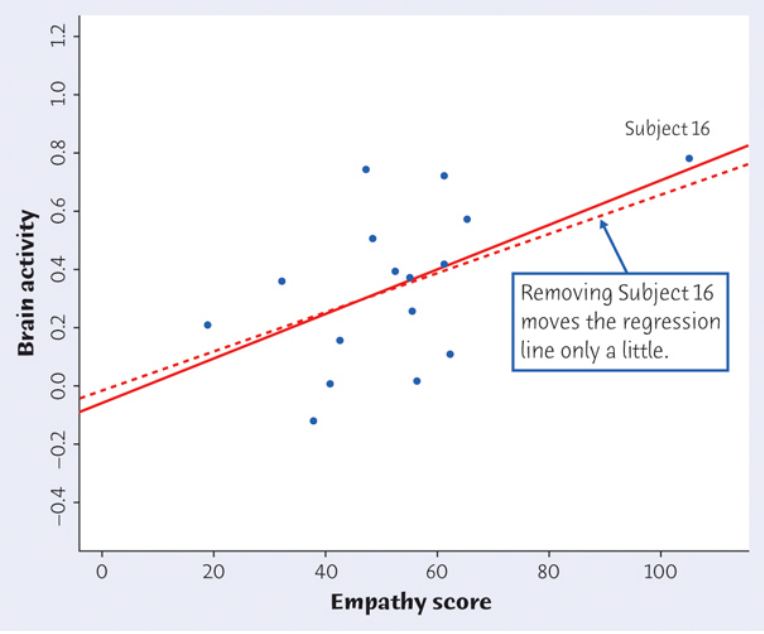


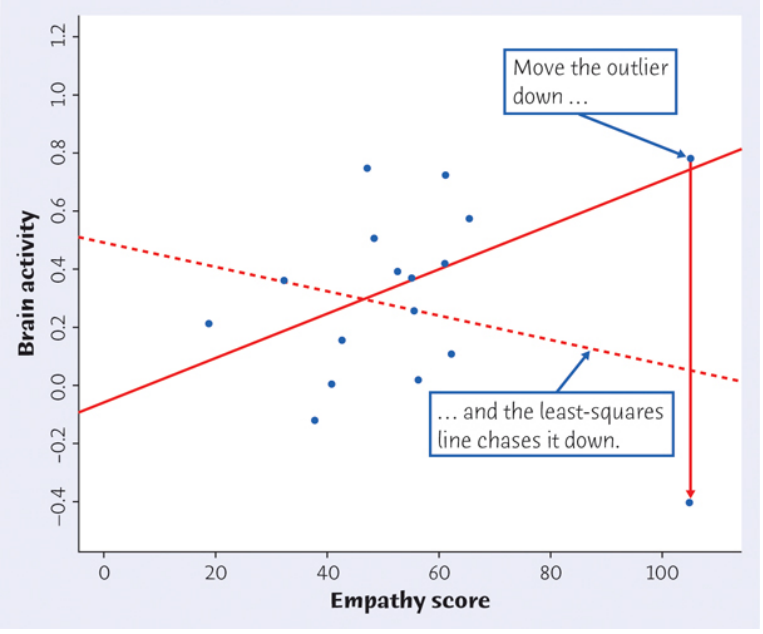
Residual plot



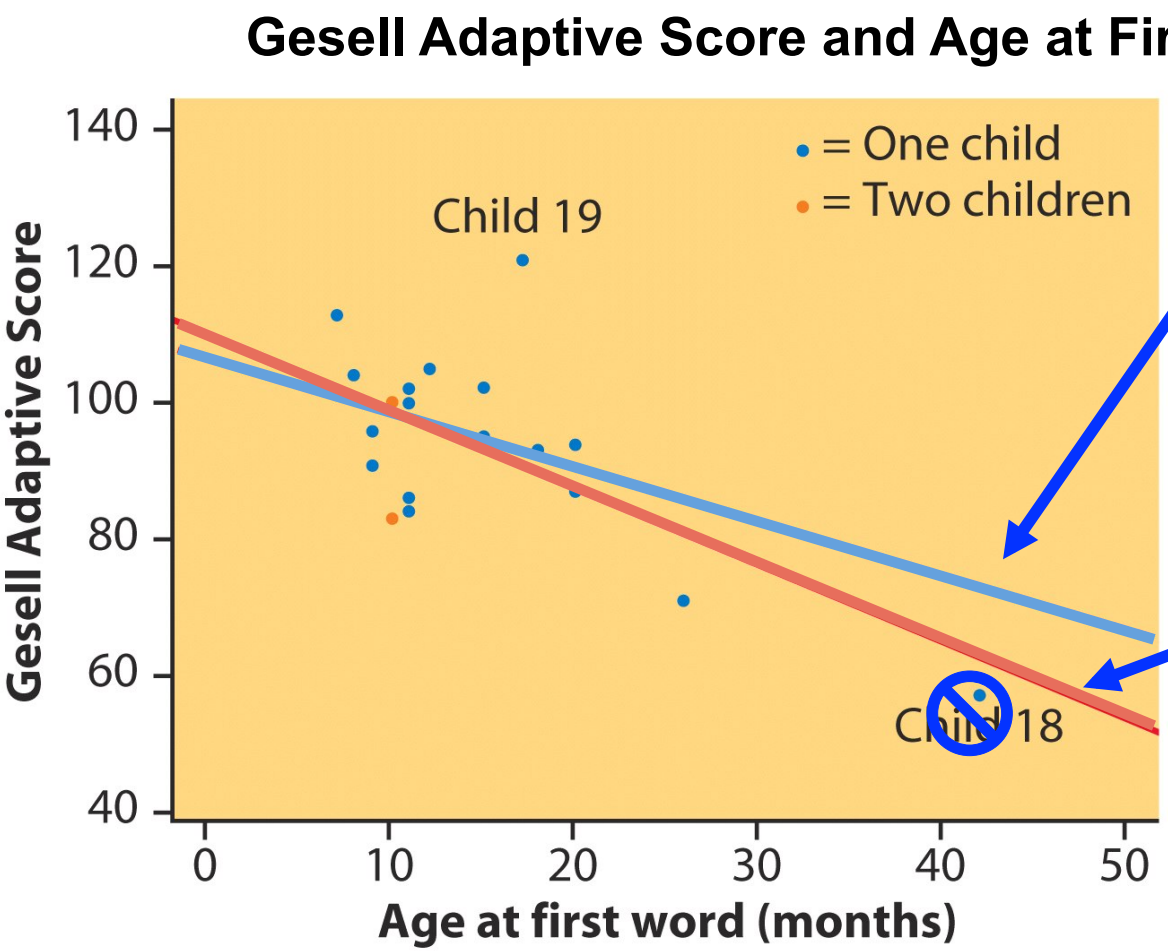
Outliers and Influential Points

- An **outlier** is an observation that lies far away from the other observations.
  - Outliers in the *y* direction have large residuals.
  - Outliers in the *x* direction are often **influential** for the least-squares regression line, meaning that the removal of such points would markedly change the equation of





Outliers and Influential Points



After removing child 18

$r^2 = 11\%$

From all of the data

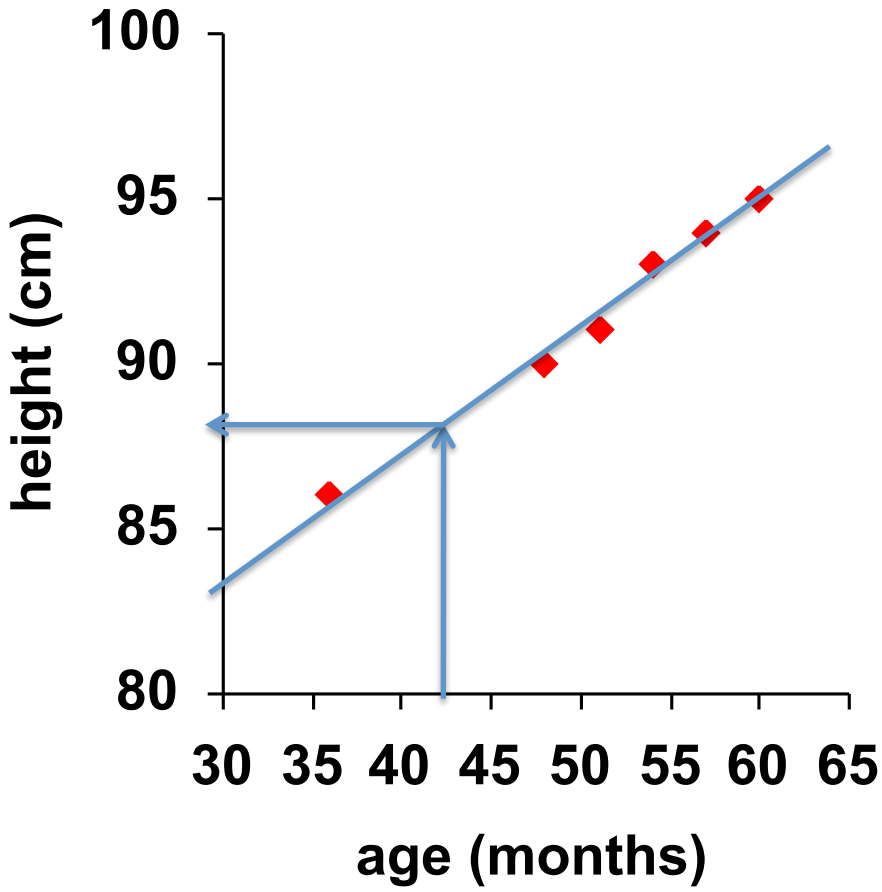
$r^2 = 41\%$

Cautions About Correlation and Regression

- Both describe linear relationships.
- Both are affected by outliers.
- Always plot the data before interpreting.
- Beware of ***extrapolation.***
  - The use of a regression line for prediction outside of the range of values of *x* used to obtain the line. Such predictions are often not accurate.
- 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.
- Correlation does not imply causation!

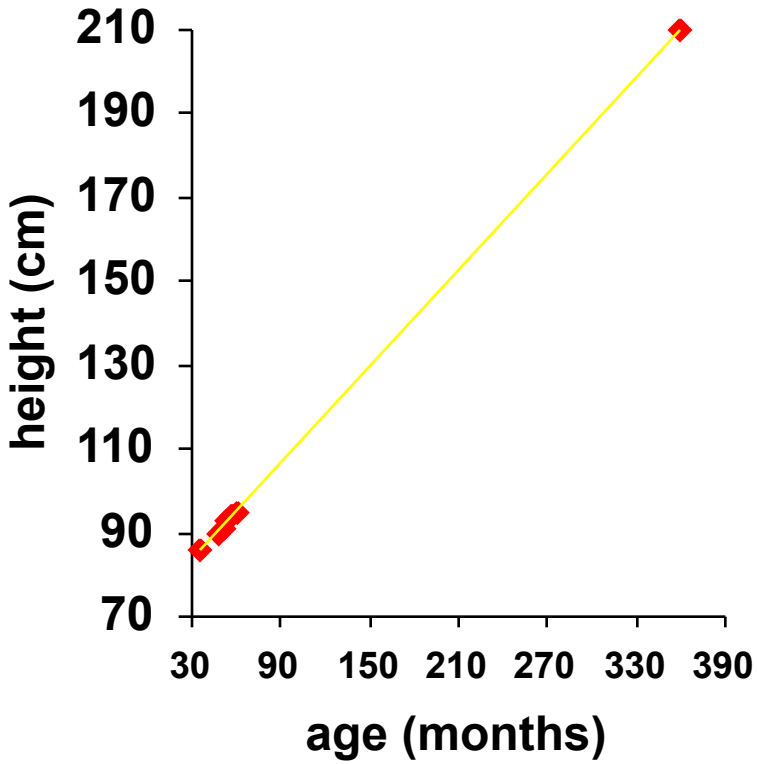
Caution: Beware of Extrapolation

- 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

- Regression line:
  y*-hat = 71.95 + .383 *x
- Height at age 42 months?
  ***y*-hat = 88**
- Height at age 30 years?
  ***y*-hat = 209.8**
  - She is predicted to be 209,8 cm at age 30!



Caution: Beware of Lurking Variables

Does having more cars make you live longer?

A serious study once found that people with two cars live longer than people who own only one car.
Could we lengthen our lives by buying more cars?

No. The study used number of cars as a quick indicator of affluence. Well-off people tend to have more cars. They also tend to live longer, probably because they are better educated, take better care of themselves, and get better medical care. The cars have nothing to do with it.

There is no cause–and–effect tie between number of cars and length of life. A **lurking variable**—such as personal affluence in Example —that influences both *x* and *y* can create a high correlation even though there is no direct connection between *x* and *y*.

To earn more, get married?

Data show that men who are married, and also divorced or widowed men, earn quite a bit more than men the same age who have never been married. This does not mean that a man can raise his income by getting married, because men who have never been married are different from married men in many ways other than marital status.

Suggest several lurking variables that might help explain the association between marital status and income.

Age is probably the most important lurking variable: married men would generally be older than single men, so they would have been in the workforce longer, and therefore had more time to advance in their careers

Association 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

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 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.*
    - The connection happens in *repeated trials.*
    - The connection happens under *varying conditions.*
  - Higher doses are associated with stronger responses.
  - Alleged cause *precedes* the effect in time.
  - Alleged cause is *plausible* (reasonable explanation).