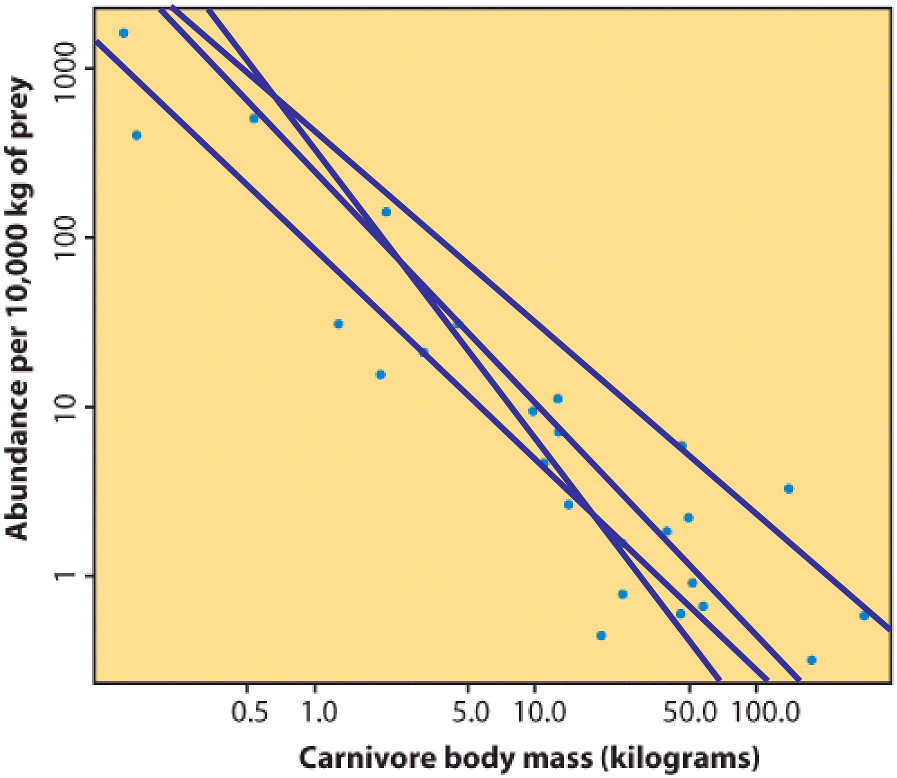
Lecture 7

Simple Linear Regression

Least squares regression. Review of the basics: Sections 2.3-2.5

- The regression line
- Making predictions
- Coefficient of determination R^2
- Transforming relationships
- Residuals
- Outliers and influential points
- Lurking variables
- The question of causation



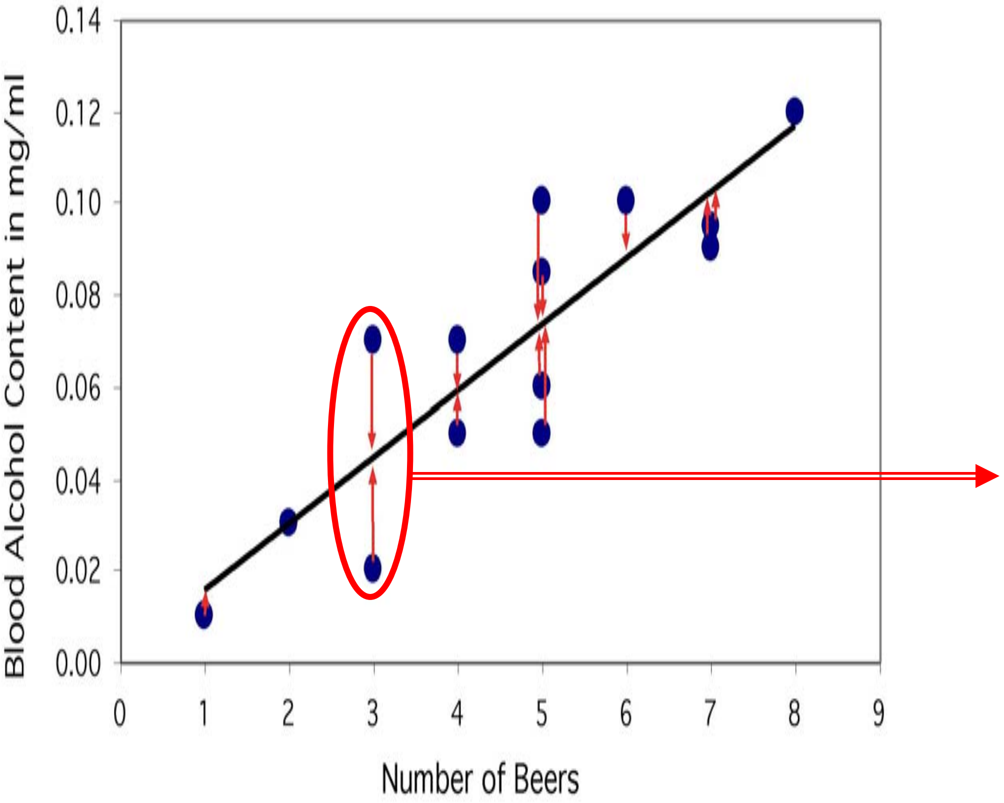
Correlation tells us about *strength* (scatter) and *direction* of the linear relationship between two quantitative variables.

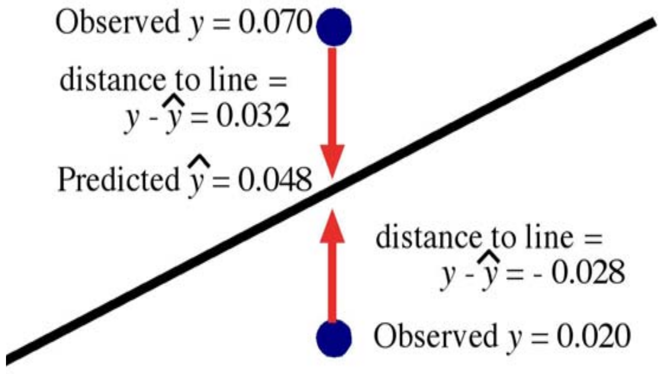
In addition, we would like to have a numerical description of how both variables vary together. For instance, is one variable increasing faster than the other one? And we would like to make predictions based on that numerical description.

But which line best describes our data?

The regression line

The least-squares regression line is the unique line such that the sum of the squared vertical (*y*) distances between the data points and the line is the smallest possible.



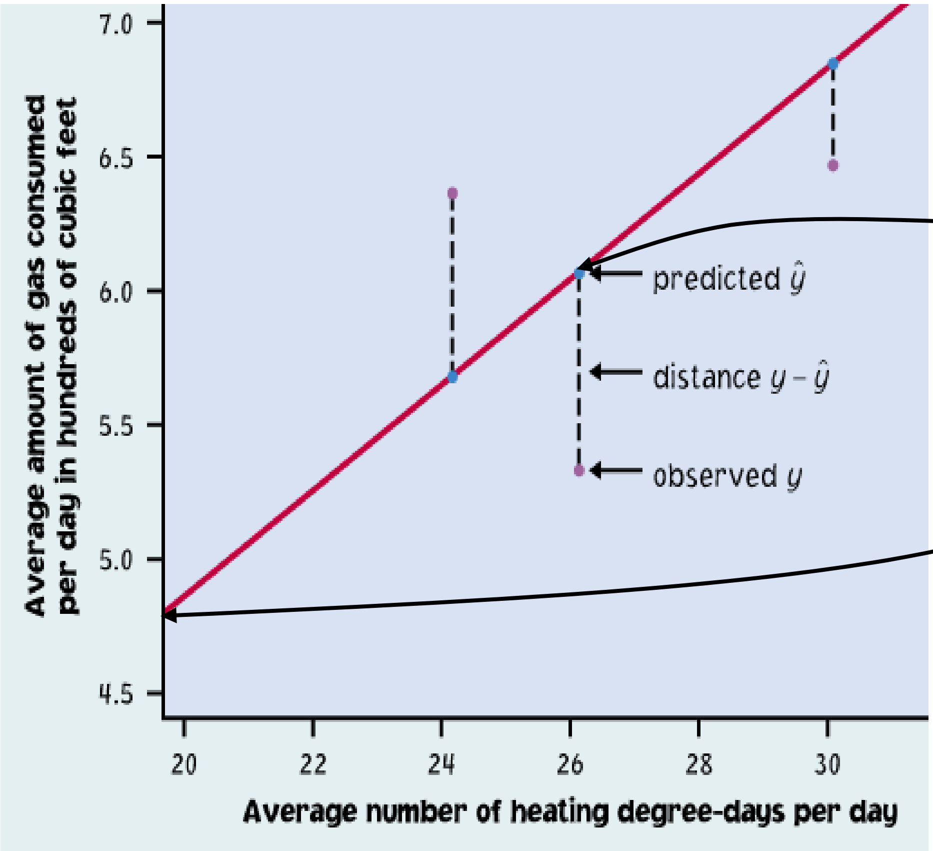


Distances between the points and line are squared so all are positive values. This is done so that distances can be properly added (Pythagoras).

Properties

The least-squares regression line can be shown to have this equation:

$$\hat{y} = (\bar{y} - r\bar{x}\frac{s_y}{s_x}) + r\frac{s_y}{s_x}x, \quad \text{or} \quad \boxed{\hat{y} = a + bx}$$



$\hat{y}$ is the predicted *y* value (y hat)

b is the **slope**

a is the ***y*-intercept**

"a" *is in units of* y
"b" *is in units of* y */ units of* x

How to:

First we calculate the **slope of the line, *b***; from statistics we already know:

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

r is the correlation.
s_y is the standard deviation of the response variable *y.*
s_x is the the standard deviation of the explanatory variable *x.*

Once we know *b,* the slope, we can calculate **a, the *y*-intercept:**

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

where $\bar{x}$ and $\bar{y}$ are the sample means of the *x* and *y* variables

This means that we don't have to calculate a lot of squared distances to find the least-squares regression line for a data set. We can instead rely on the equation.

*But typically, we use a **2-var stats calculator** or stats software.*

BEWARE!!!

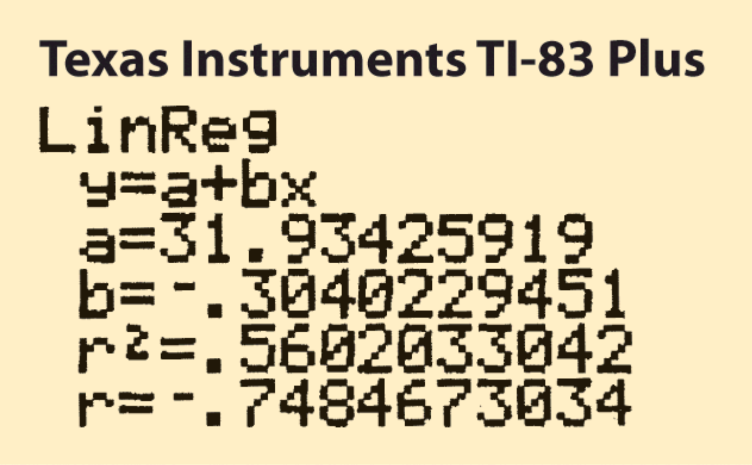
Not all calculators and software use the same convention:

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

Some use instead:

$$\hat{y} = ax + b$$

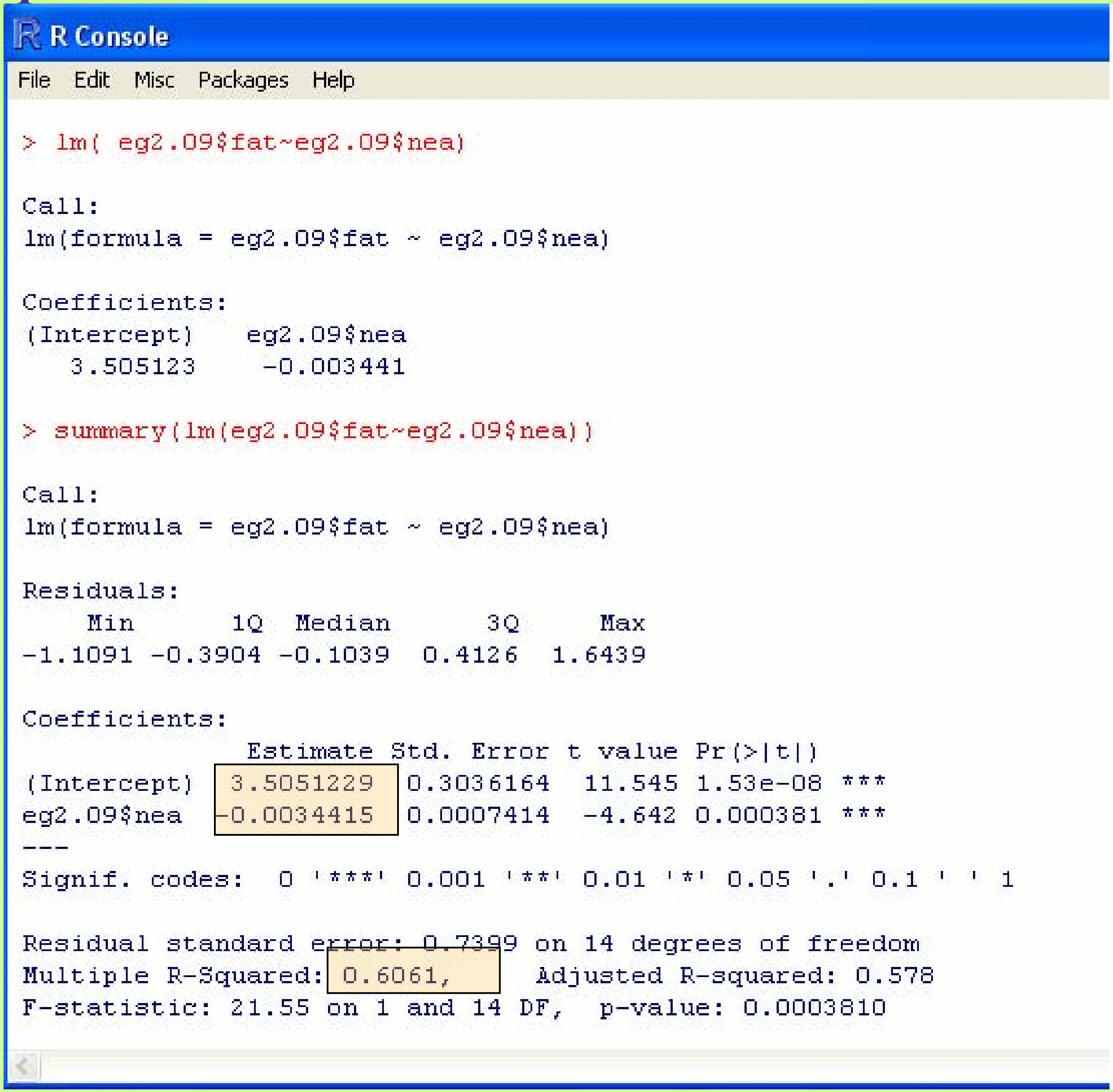
Make sure you know what YOUR calculator gives you for a and b before you answer homework or exam questions.



Texas Instruments TI-83 Plus

```
LinReg
 y=a+bx
 a=31.93425919
 b=⁻.3040229451
 r²=.5602033042
 r=⁻.7484673034
```

Software output



```
> lm( eg2.09$fat~eg2.09$nea)

Call:
lm(formula = eg2.09$fat ~ eg2.09$nea)

Coefficients:
(Intercept)   eg2.09$nea
   3.505123    -0.003441

> summary(lm(eg2.09$fat~eg2.09$nea))

Call:
lm(formula = eg2.09$fat ~ eg2.09$nea)

Residuals:
    Min       1Q   Median       3Q      Max
-1.1091 -0.3904 -0.1039  0.4126  1.6439

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.5051229  0.3036164  11.545 1.53e-08 ***
eg2.09$nea  -0.0034415  0.0007414  -4.642 0.000381 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7399 on 14 degrees of freedom
Multiple R-Squared: 0.6061,     Adjusted R-squared: 0.578
F-statistic: 21.55 on 1 and 14 DF,  p-value: 0.0003810
```

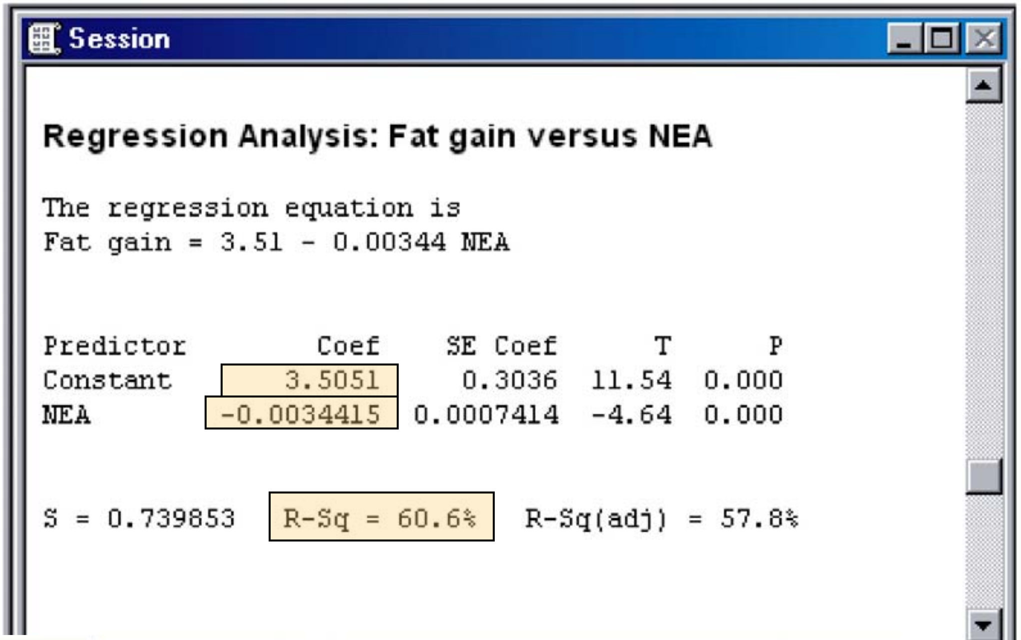
intercept
slope

R^2

Software output (cont)

intercept
slope

R^2



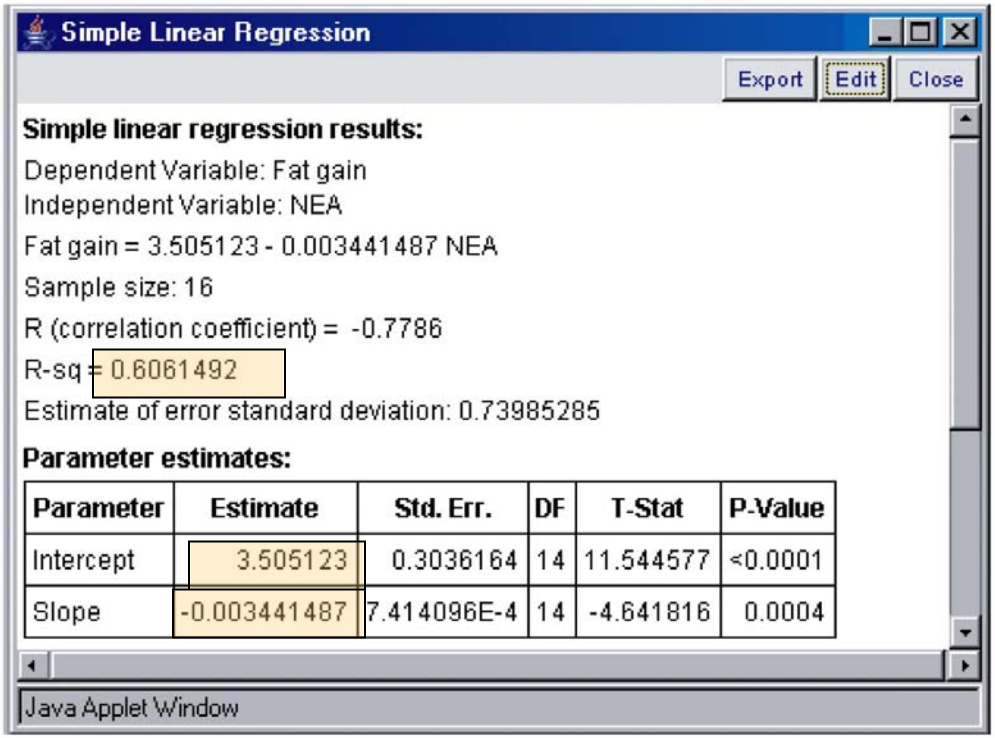
```
Regression Analysis: Fat gain versus NEA

The regression equation is
Fat gain = 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%
```

e Practice of Statistics, Fifth Edition
nan and Company



Simple linear regression results:

Dependent Variable: Fat gain
Independent Variable: NEA
Fat gain = 3.505123 - 0.003441487 NEA
Sample size: 16
R (correlation coefficient) = -0.7786
R-sq = 0.6061492
Estimate of error standard deviation: 0.73985285

Parameter estimates:

Parameter	Estimate	Std. Err.	DF	T-Stat	P-Value
Intercept	3.505123	0.3036164	14	11.544577	<0.0001
Slope	-0.003441487	7.414096E-4	14	-4.641816	0.0004

Figure 2-14a
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W.H. Freeman and Company

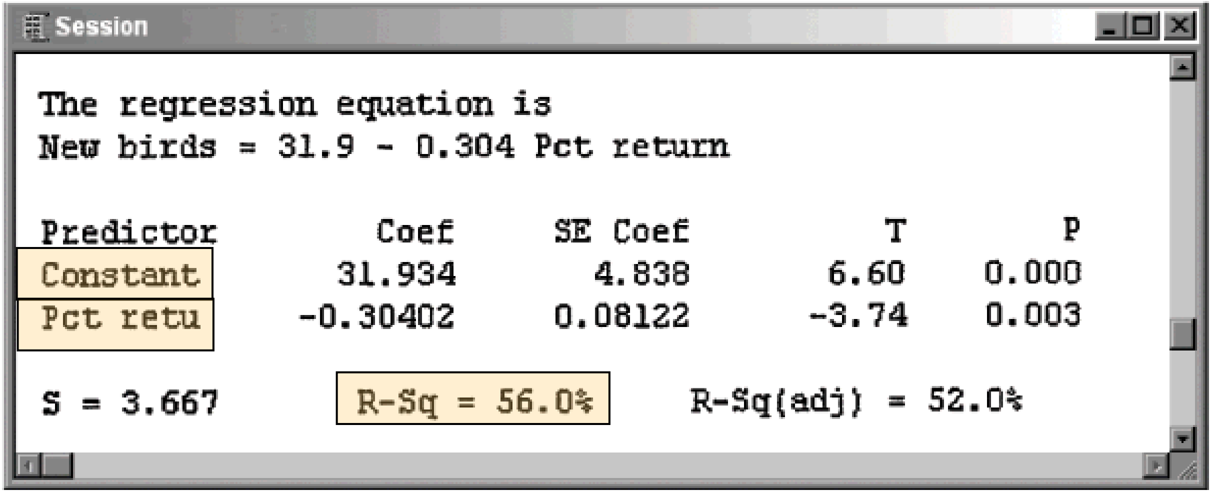
Software output (another example)

intercept

slope

R^2

Minitab

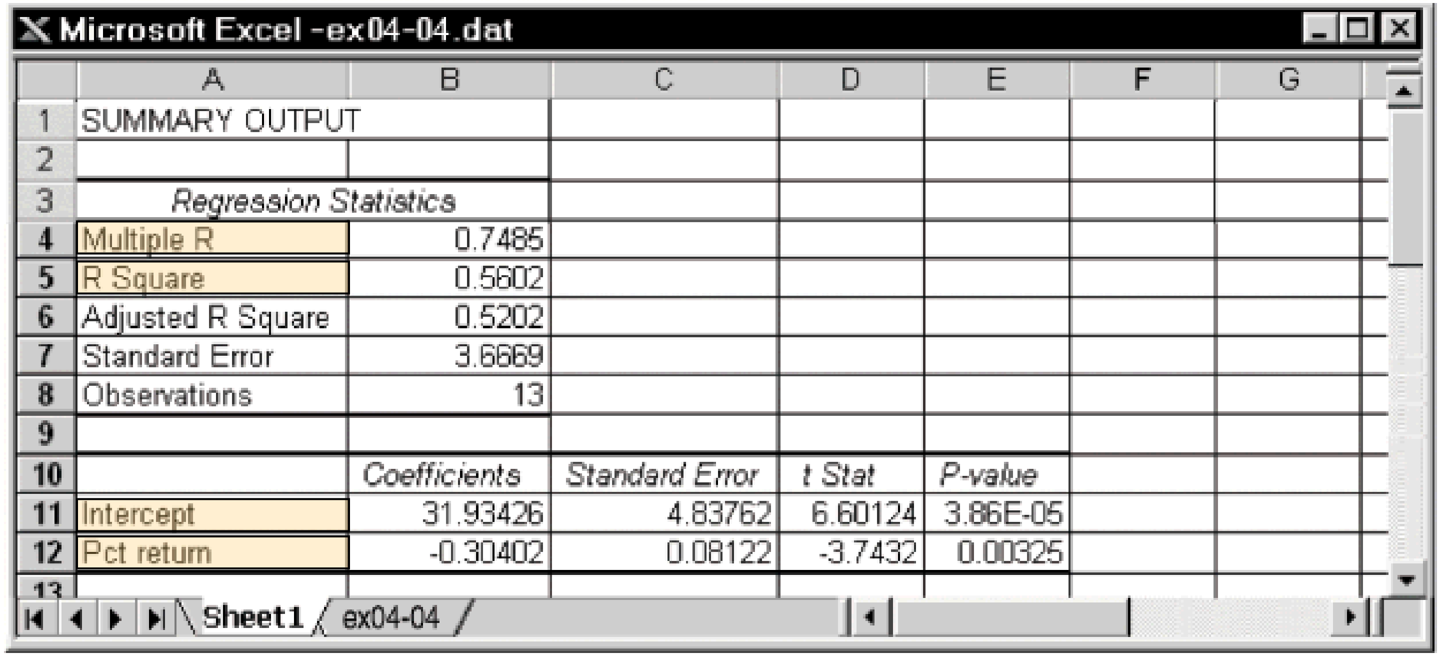


```
The regression equation is
New birds = 31.9 - 0.304 Pct return

Predictor        Coef     SE Coef        T        P
Constant       31.934       4.838     6.60    0.000
Pct retu     -0.30402     0.08122    -3.74    0.003

S = 3.667        R-Sq = 56.0%      R-Sq(adj) = 52.0%
```

Excel



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.7485					
5	R Square	0.5602					
6	Adjusted R Square	0.5202					
7	Standard Error	3.6669					
8	Observations	13					
9							
10		*Coefficients*	*Standard Error*	*t Stat*	*P-value*		
11	Intercept	31.93426	4.83762	6.60124	3.86E-05		
12	Pct return	-0.30402	0.08122	-3.7432	0.00325		
13							

r

R^2

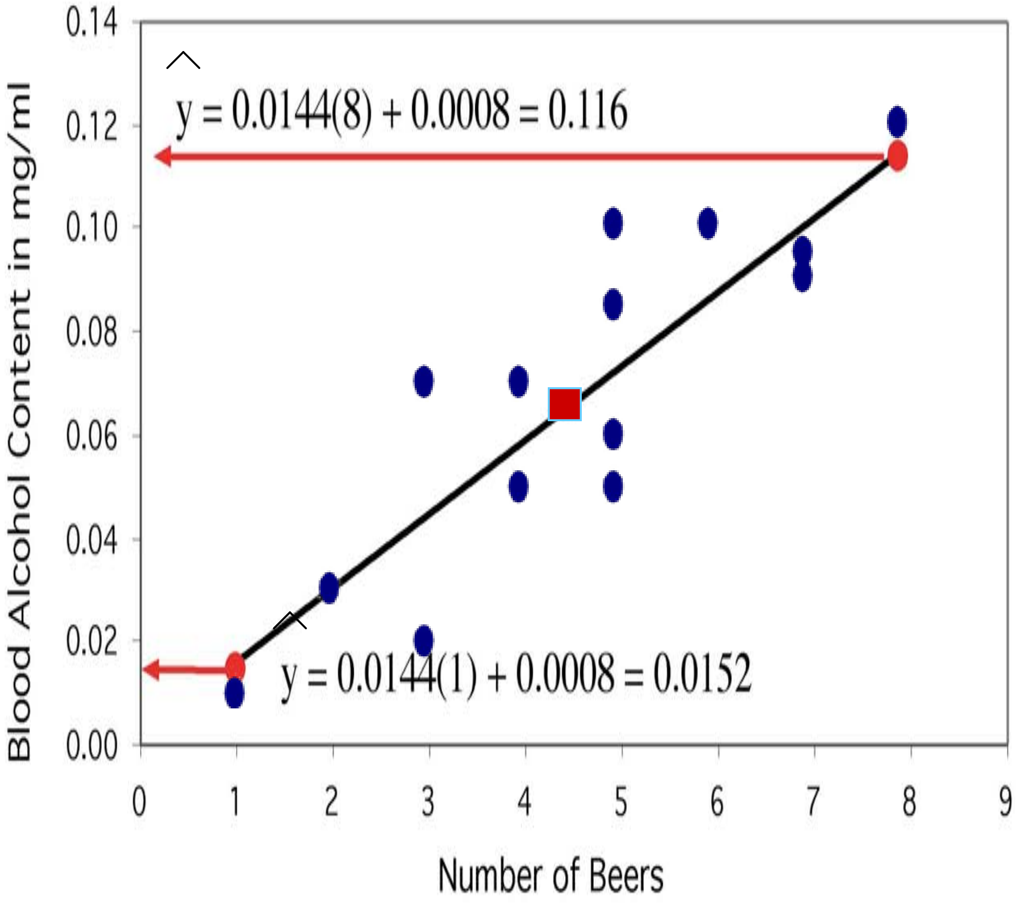
intercept

slope

The equation completely describes the regression line.

To plot the regression line you only need to plug two *x* values into the equation, get *y*, and draw the line that goes through those those points.

Hint: The regression line always passes through the mean of x *and* y.



The points you use for drawing the regression line are derived from the equation.

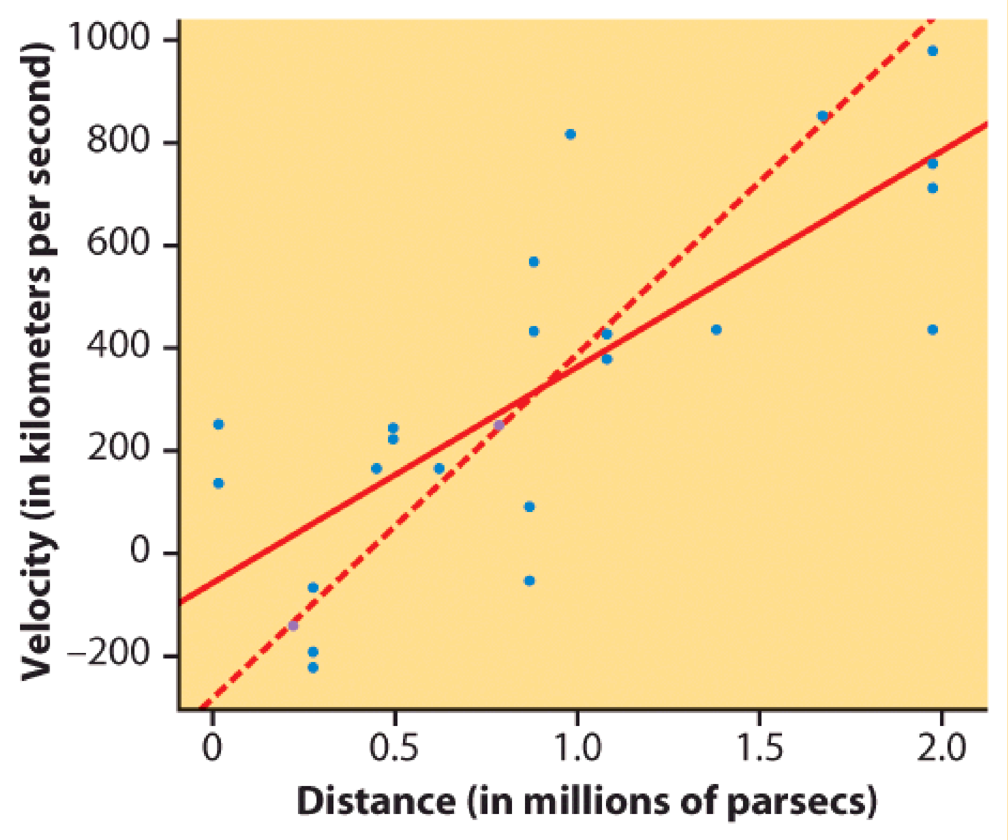
They are NOT points from your sample data (except by pure coincidence).

The distinction between explanatory and response variables is crucial in regression. If you exchange *y* for *x* in calculating the regression line, you will get the wrong line.

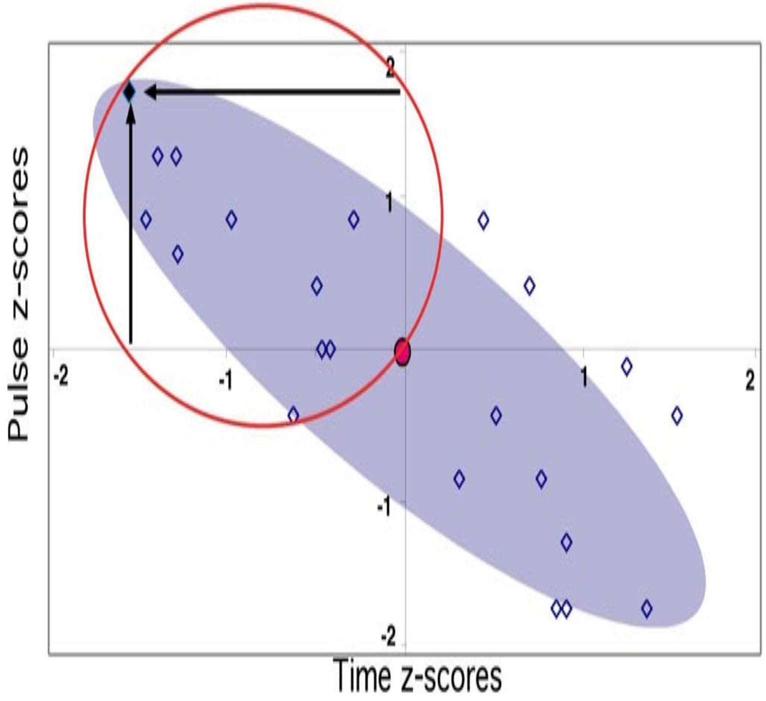
Regression examines the distance of all points from the line **in the *y* direction only.**

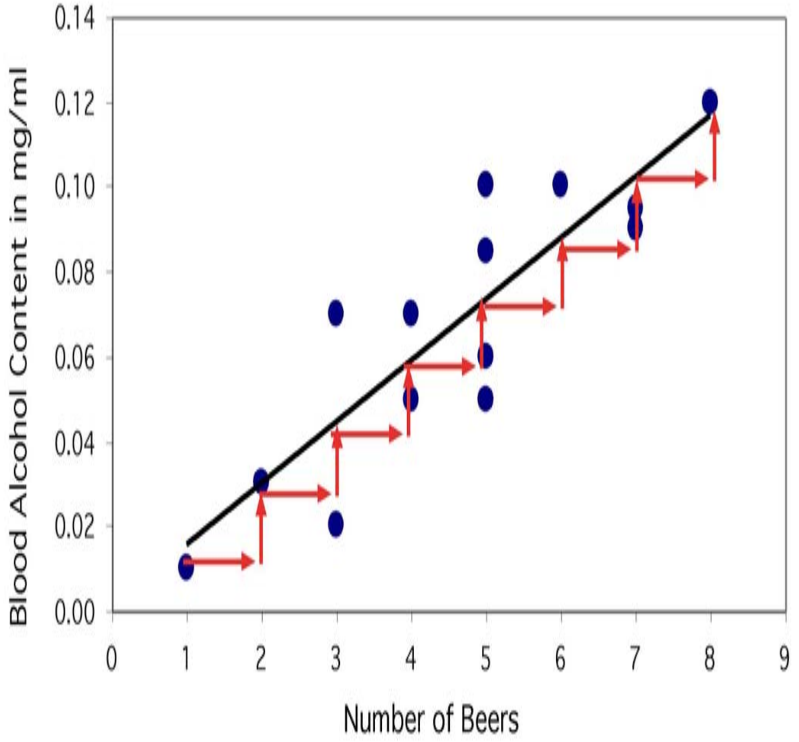
Hubble telescope data about galaxies moving away from earth:

These two lines are the two regression lines calculated either correctly (*x* = distance, *y* = velocity, solid line) or incorrectly (*x* = velocity, *y* = distance, dotted line).



Correlation versus regression



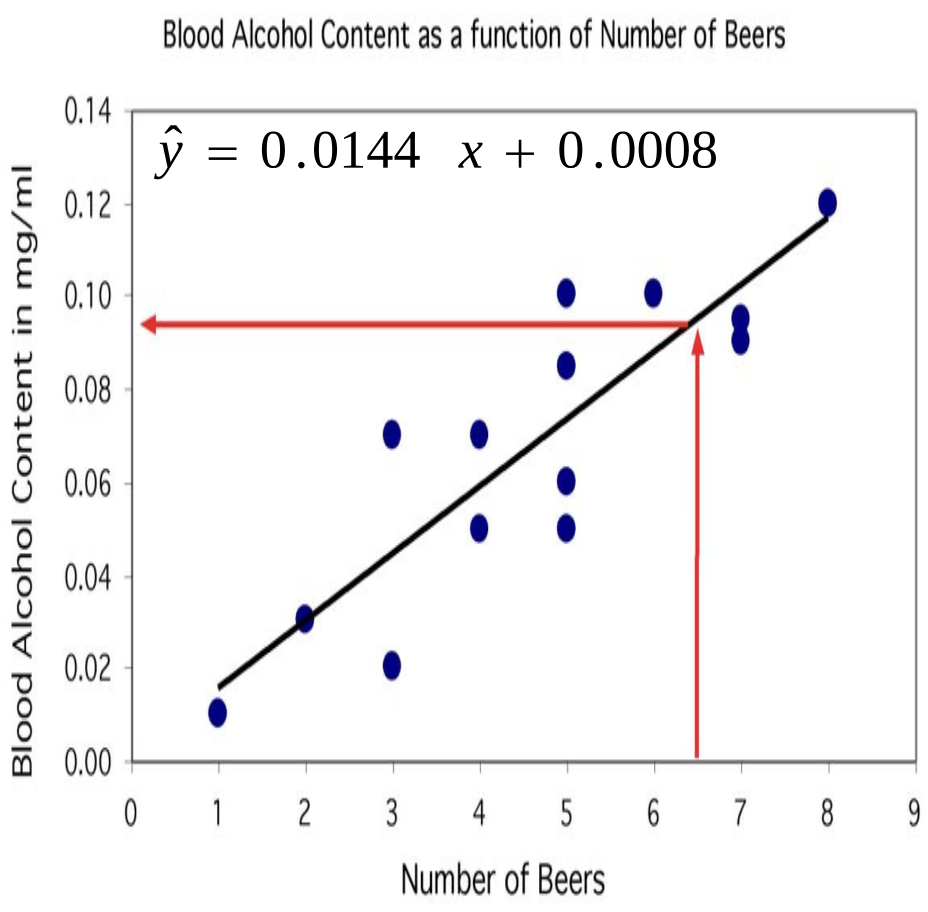


The **correlation** is a measure of spread (scatter) in both the *x* and *y* directions in the linear relationship.

In **regression** we examine the variation in the response variable (*y*) given change in the explanatory variable (*x*).

Making predictions: interpolation

The equation of the least-squares regression allows to predict *y* for any *x* within the range studied. This is called **interpolating**.



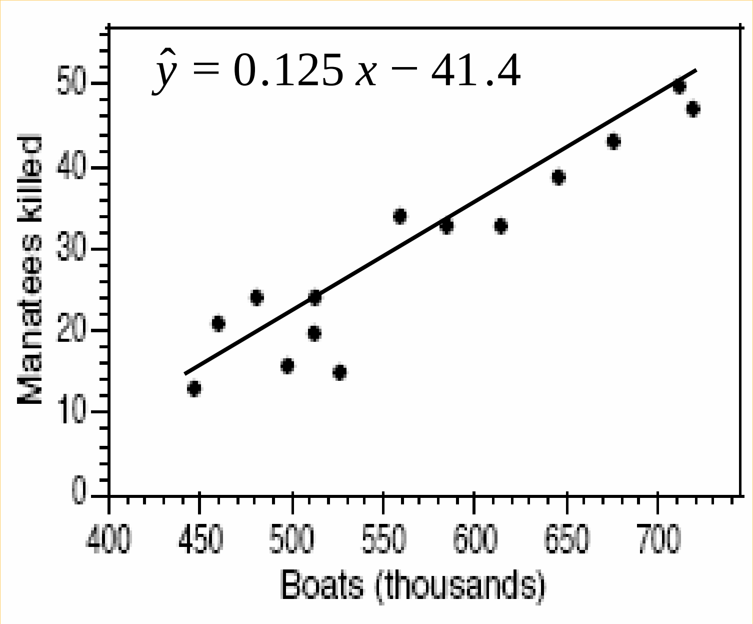
Nobody in the study drank 6.5 beers, but by finding the value of $\hat{y}$ from the regression line for *x* = 6.5 we would expect a blood alcohol content of 0.094 mg/ml.

$$\hat{y} = 0.0144*6.5 + 0.0008$$

$$\hat{y} = 0.936 + 0.0008 = 0.0944 \text{mg/ml}$$

(in 1000's)

Year	Powerboats	Dead Manatees
1977	447	13
1978	460	21
1979	481	24
1980	498	16
1981	513	24
1982	512	20
1983	526	15
1984	559	34
1985	585	33
1986	614	33
1987	645	39
1988	675	43
1989	711	50
1990	719	47



There is a positive linear relationship between the number of powerboats registered and the number of manatee deaths.

The least squares regression line has the equation: $\hat{y} = 0.125\,x - 41.4$

Thus if we were to limit the number of powerboat registrations to 500,000, what could we expect for the number of manatee deaths?

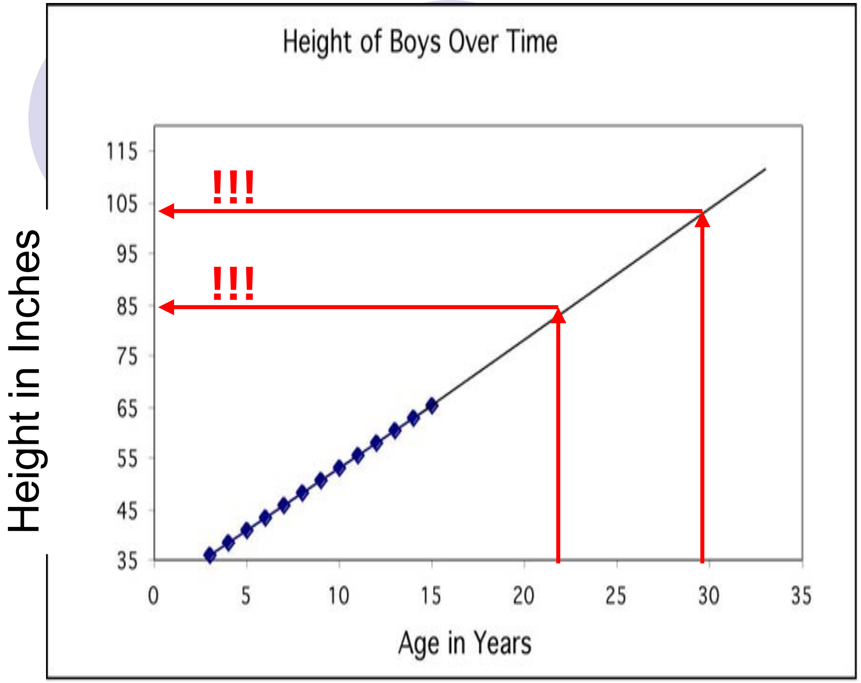
$$\hat{y} = 0.125(500) - 41.4 \Rightarrow \hat{y} = 62.5 - 41.4 = 21.1$$

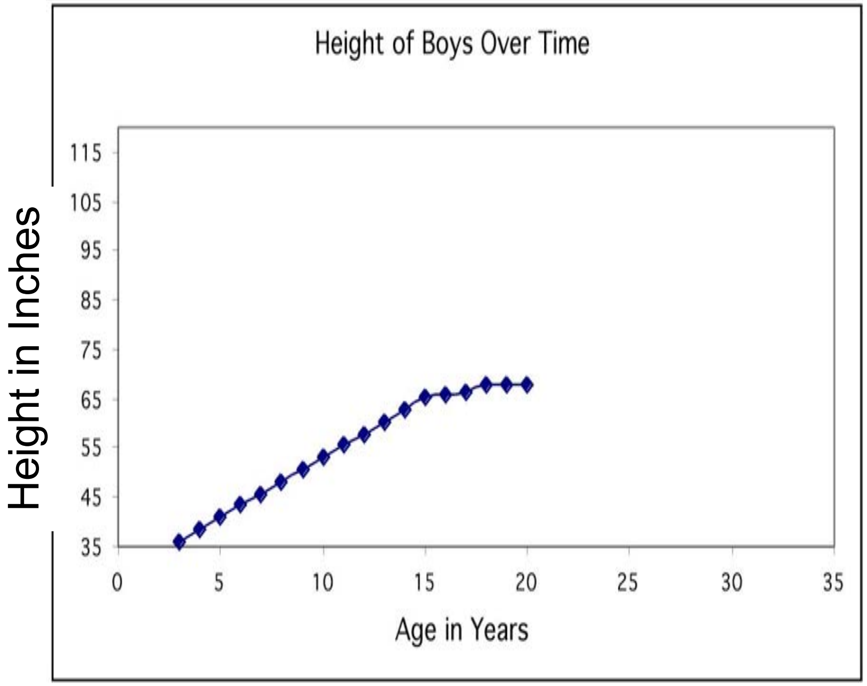
Roughly 21 manatees.

Extrapolation

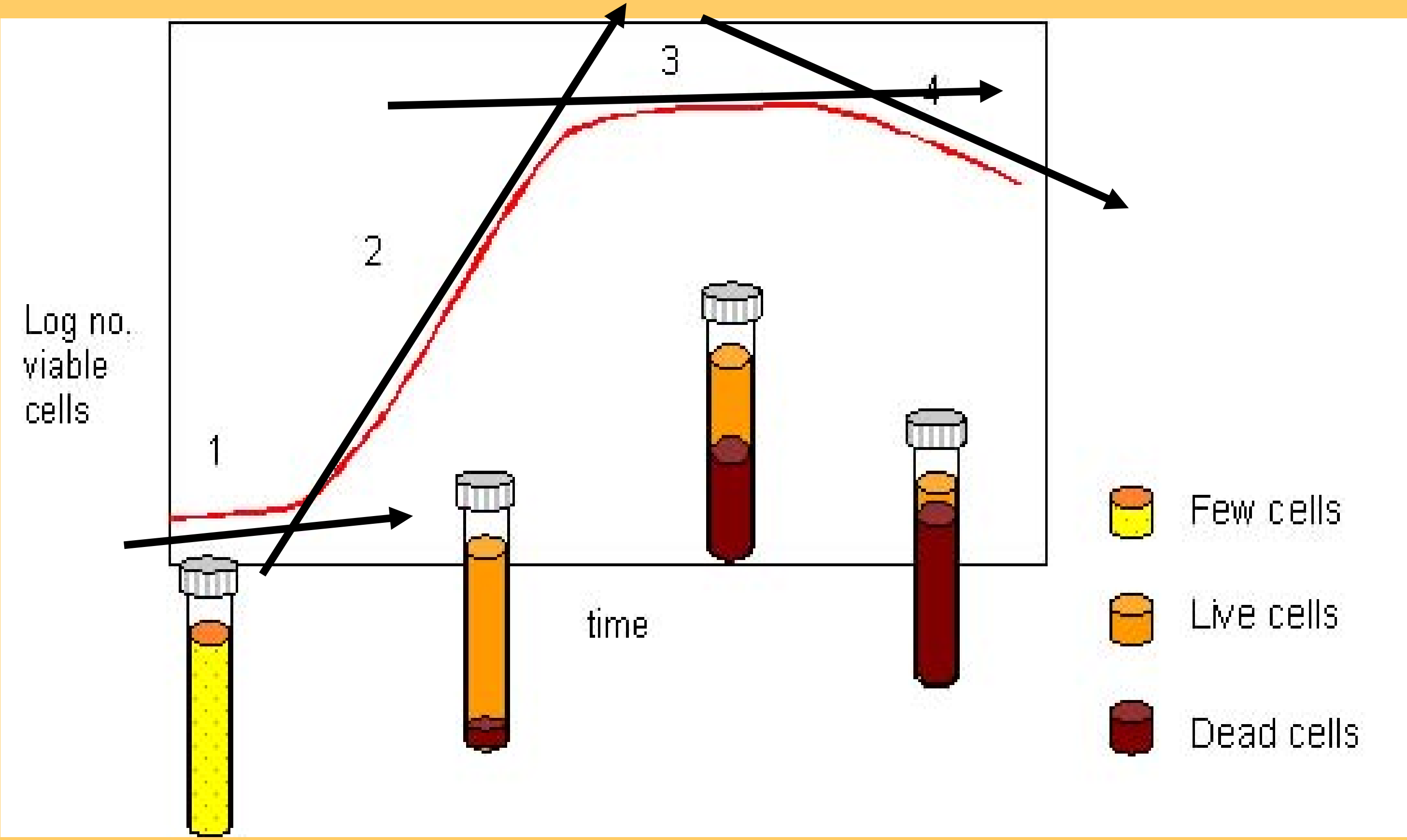
Extrapolation is the use of a regression line for predictions outside the range of *x* values used to obtain the line.

This can be a very stupid thing to do, as seen here.





Example: Bacterial growth rate over time in closed cultures



If you only observed bacterial growth in test-tube during a small subset of the time shown here, you could get almost any regression line imaginable.

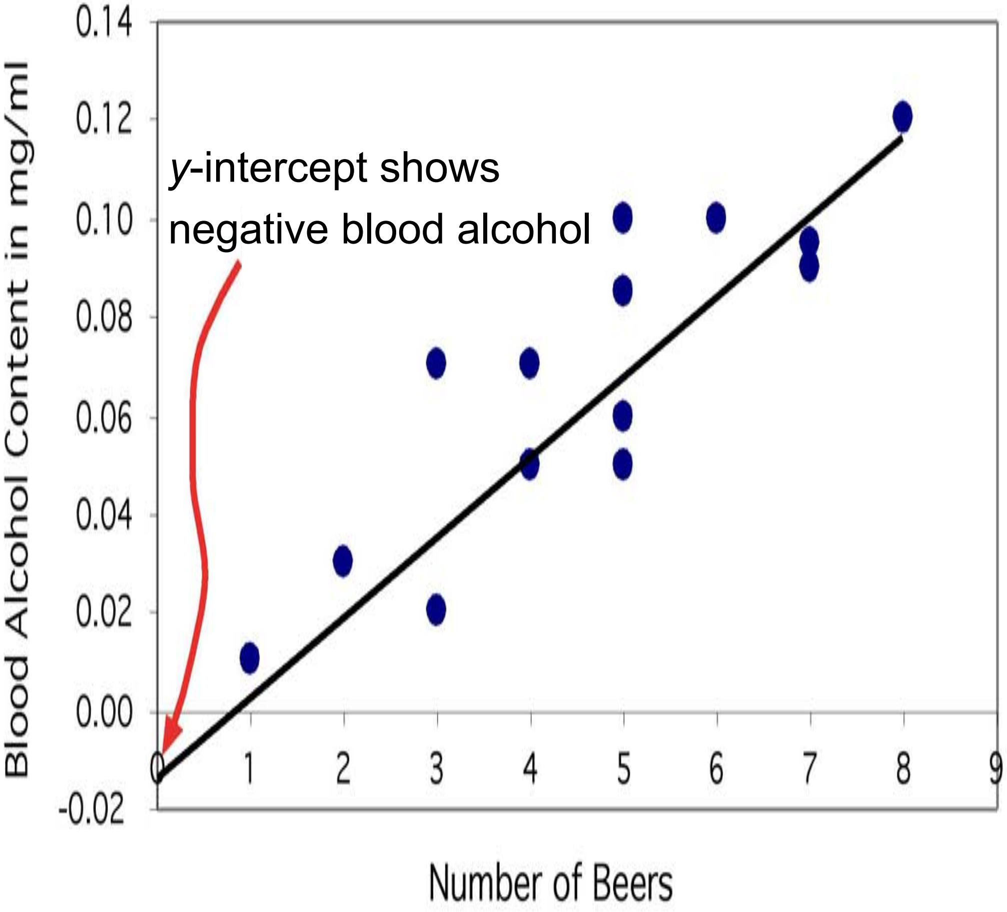
Extrapolation = big mistake.

The *y* intercept

Sometimes the *y*-intercept is not biologically possible. Here we have negative blood alcohol content, which makes no sense…

But the negative value is appropriate for the equation of the regression line.

There is a lot of scatter in the data, and the line is just an estimate.

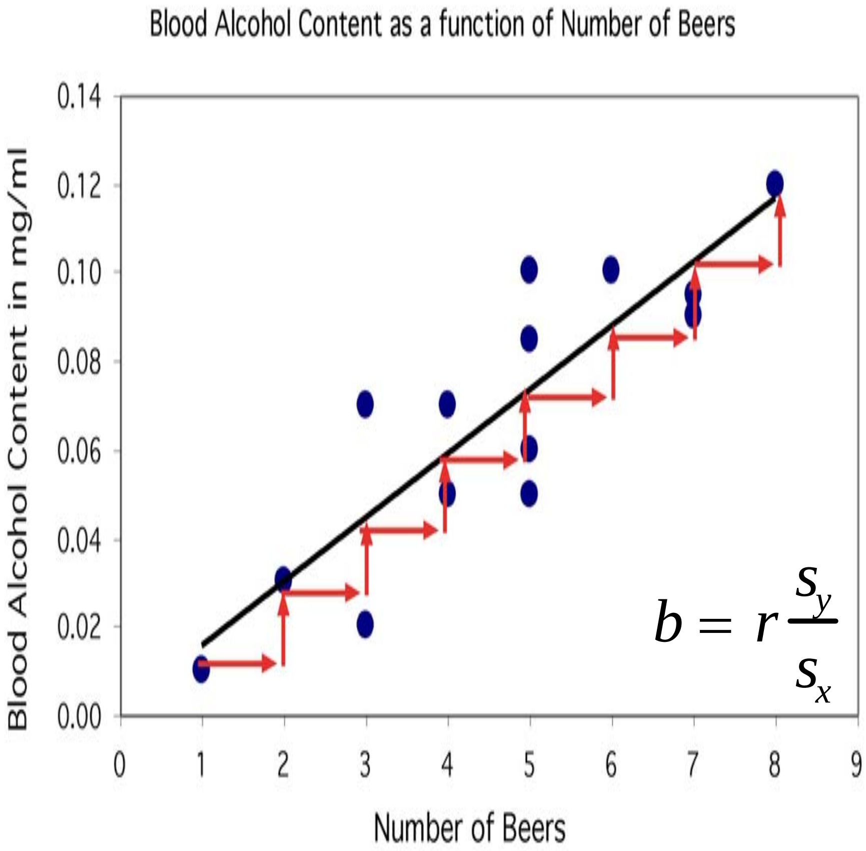


y-intercept shows negative blood alcohol

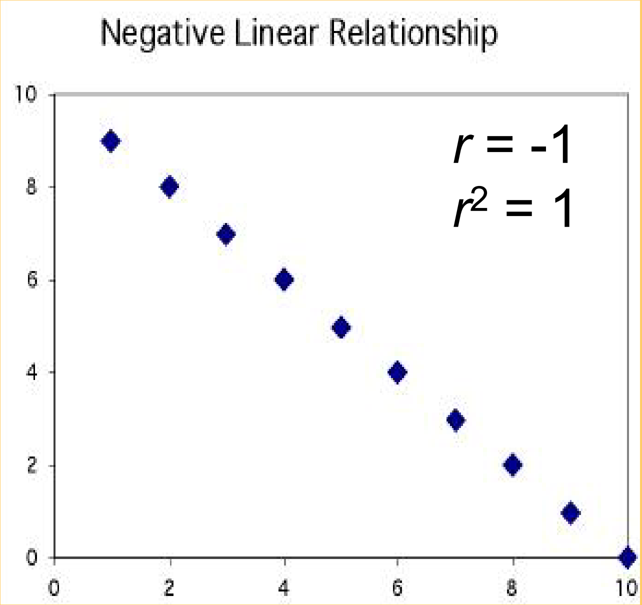
Coefficient of determination, r^2

***r^2*, the coefficient of determination,** is the square of the correlation coefficient.

r^2 represents **the percentage of the variance in *y*** (vertical scatter from the regression line) **that can be explained by the linear relationship with *x*.**

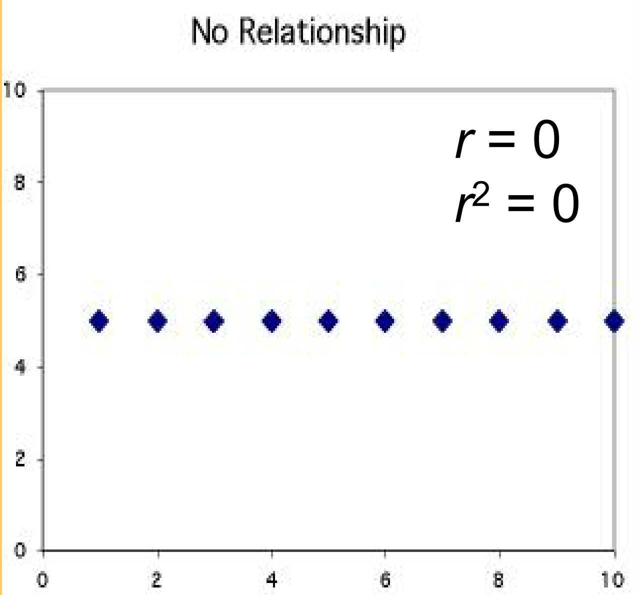


$$b = r\frac{s_y}{s_x}$$



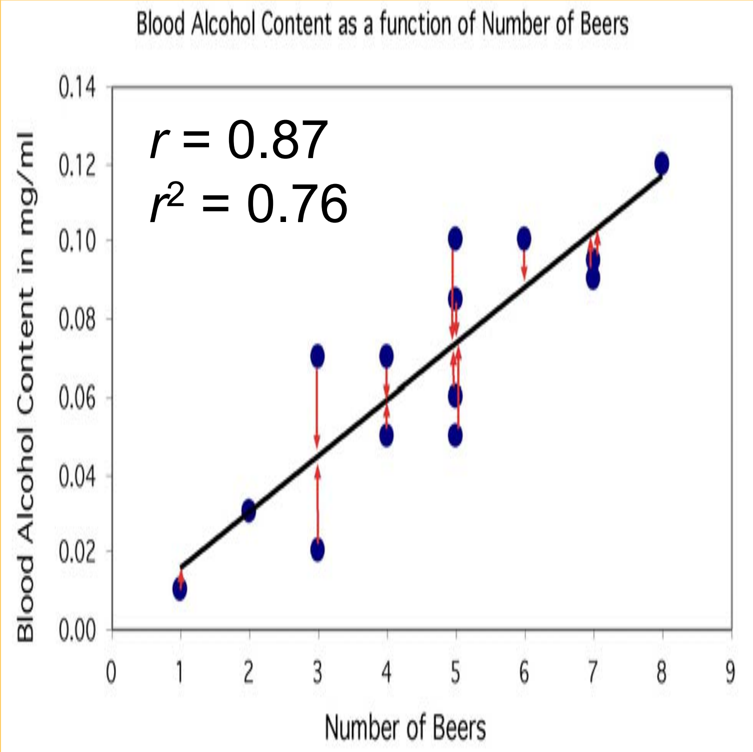
Changes in *x* explain 100% of the variations in *y*.

Y can be entirely predicted for any given value of *x*.

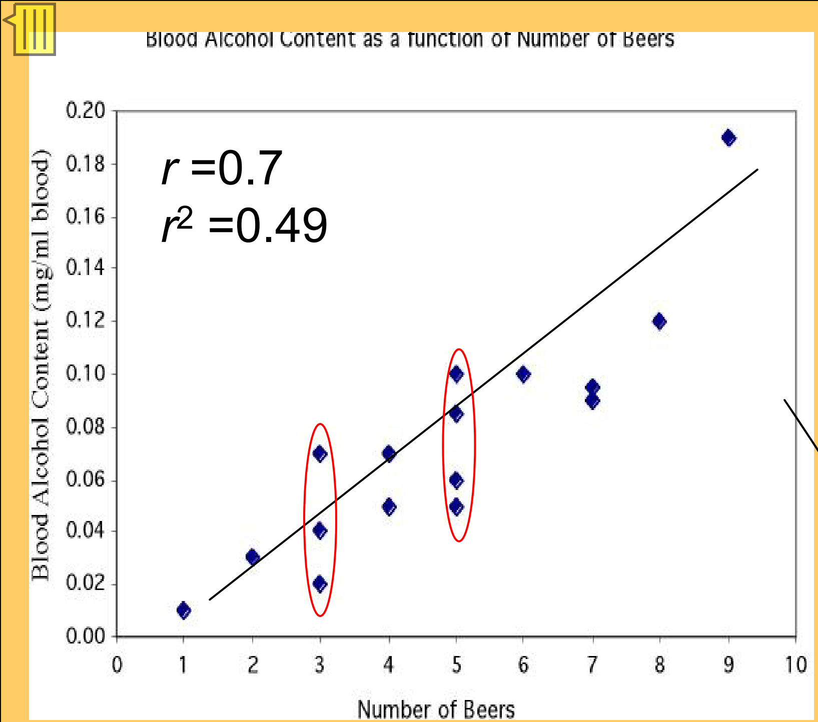


Changes in *x* explain 0% of the variations in y.

The value(s) *y* takes is (are) entirely independent of what value *x* takes.

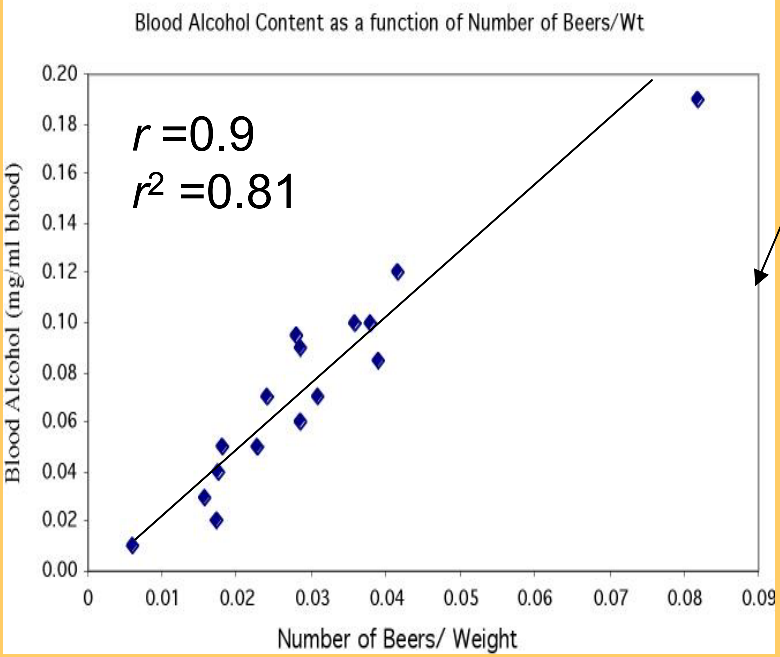


Here the change in *x* only explains 76% of the change in *y*. The rest of the change in *y* (the vertical scatter, shown as red arrows) must be explained by something other than *x*.



There is quite some variation in BAC for the same number of beers drunk. A person's blood volume is a factor in the equation that was overlooked here.

We changed number of beers to number of beers/weight of person in lb.



- In the first plot, number of beers only explains 49% of the variation in blood alcohol content.
- But number of beers / weight explains 81% of the variation in blood alcohol content.
- Additional factors contribute to variations in BAC among individuals (like maybe some genetic ability to process alcohol).

Grade performance

If class attendance explains 16% of the variation in grades, what is the correlation between percent of classes attended and grade?

1. We need to make an assumption: attendance and grades are **positively** correlated. So r will be positive too.

2. $r^2 = 0.16$, so $r = +\sqrt{0.16} = +\ 0.4$

A weak correlation.

Transforming relationships

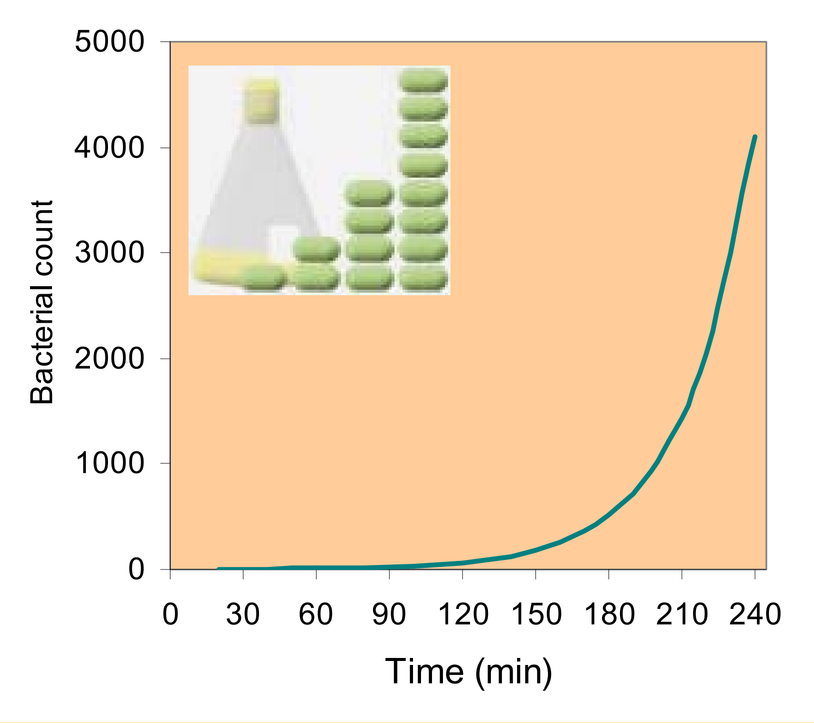
A scatterplot might show a clear relationship between two quantitative variables, but issues of influential points or non linearity prevent us from using correlation and regression tools.

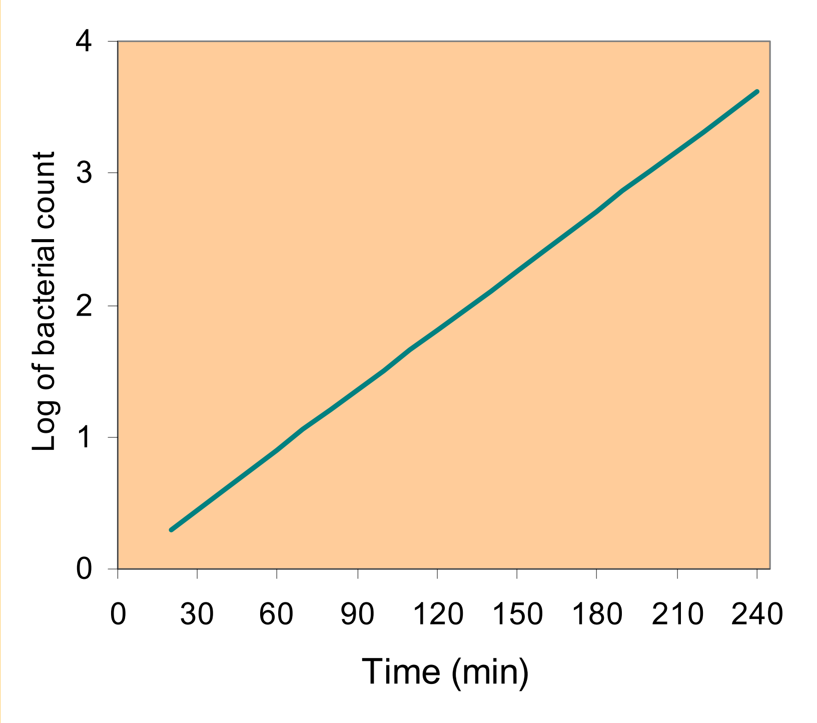
Transforming the data – changing the scale in which one or both of the variables are expressed – can make the shape of the relationship linear in some cases.

Example: Patterns of growth are often exponential, at least in their initial phase. Changing the response variable *y* into log(*y*) or ln(*y*) will transform the pattern from an upward-curved exponential to a straight line.

Exponential bacterial growth

In ideal environments, bacteria multiply through binary fission. The number of bacteria can double every 20 minutes in that way.





1 - 2 - 4 - 8 - 16 - 32 - 64 - …

Exponential growth 2^n, not suitable for regression.

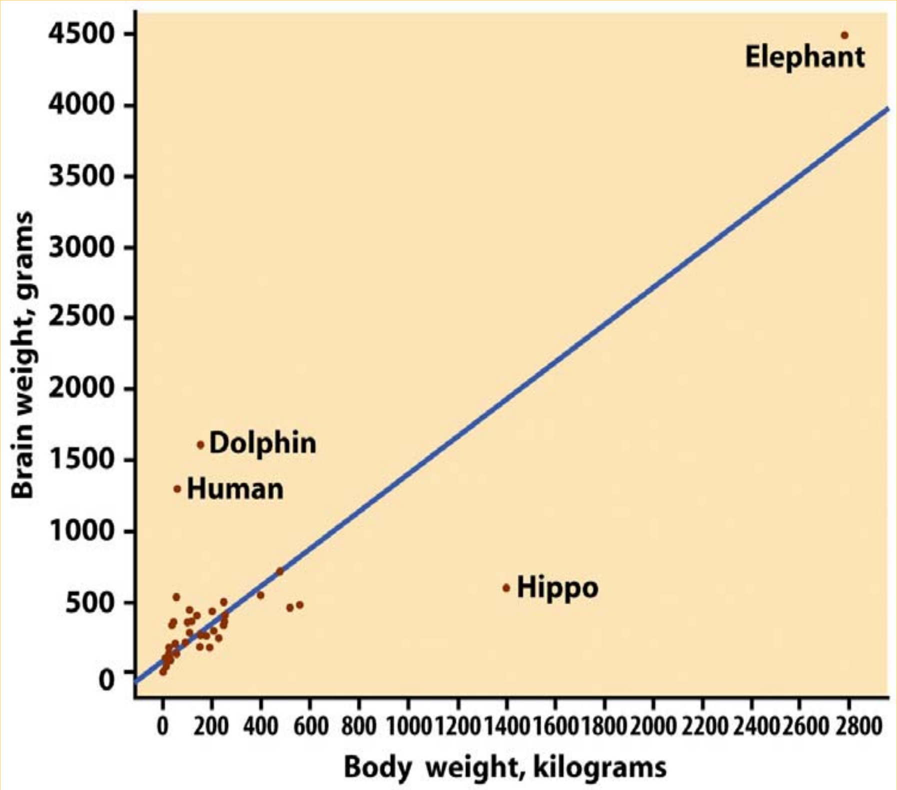
$\log(2^n) = n*\log(2) \approx 0.3n$

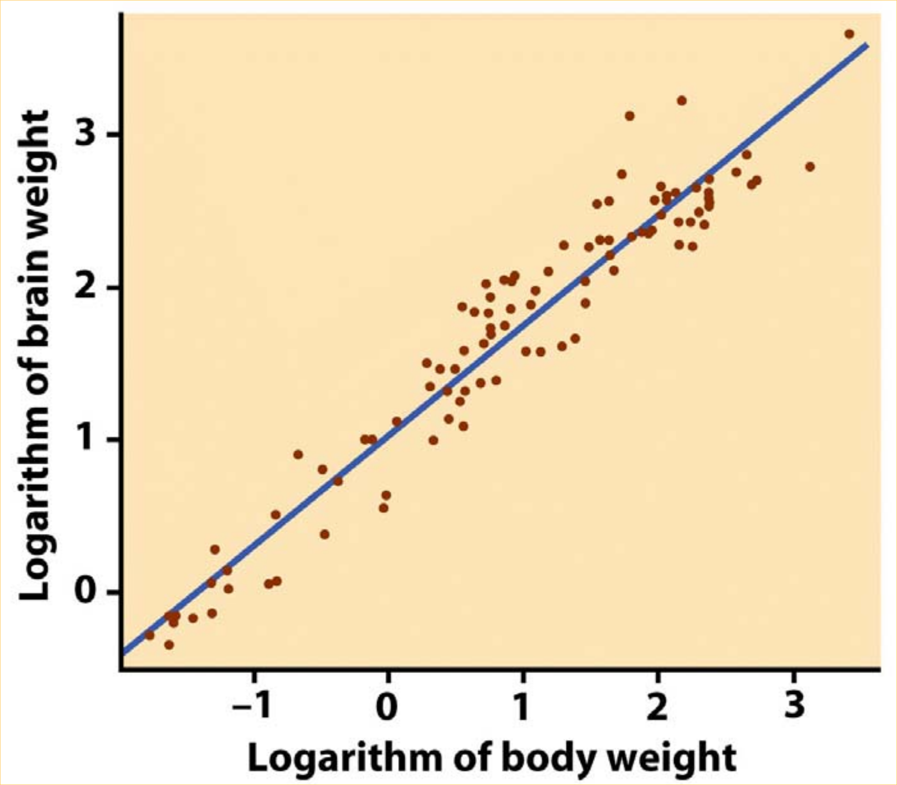
Taking the log changes the growth pattern into a straight line.

Body weight and brain weight in 96 mammal species

r = 0.86, but this is misleading.

The elephant is an influential point. Most mammals are very small in comparison. Without this point, *r* = 0.50 only.



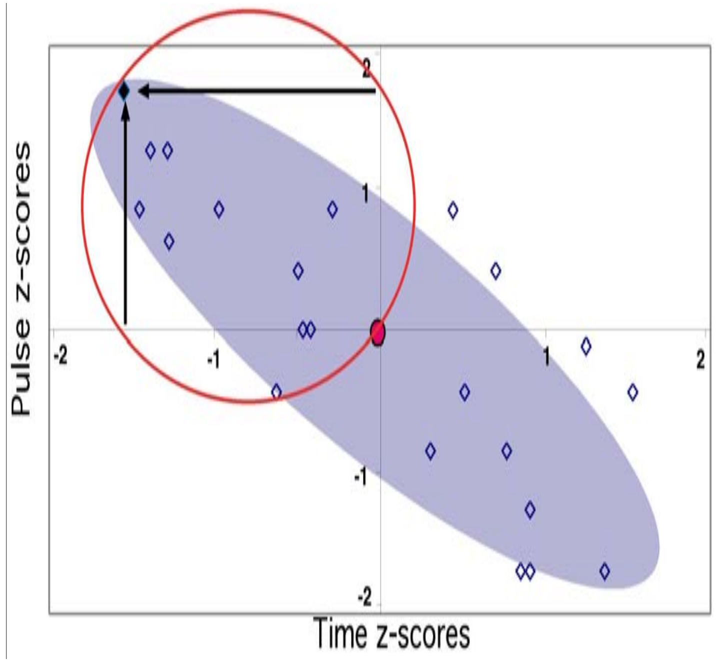


Now we plot the log of brain weight against the log of body weight.

The pattern is linear, with *r* = 0.96.
The vertical scatter is homogenous
→ good for predictions of brain weight from body weight (in the log scale).

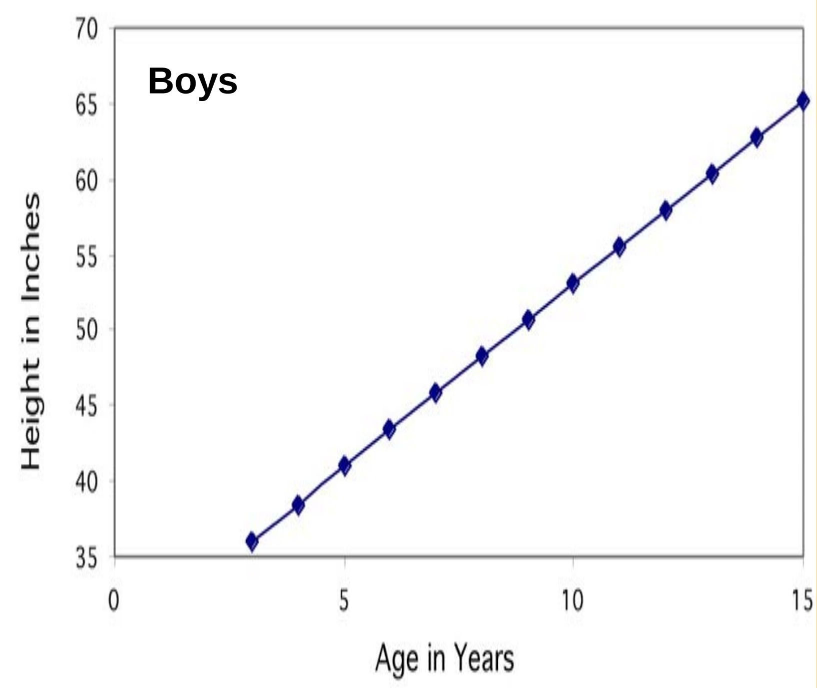
Caution about regression 2.4, 2.5 Correlation/regression using averages

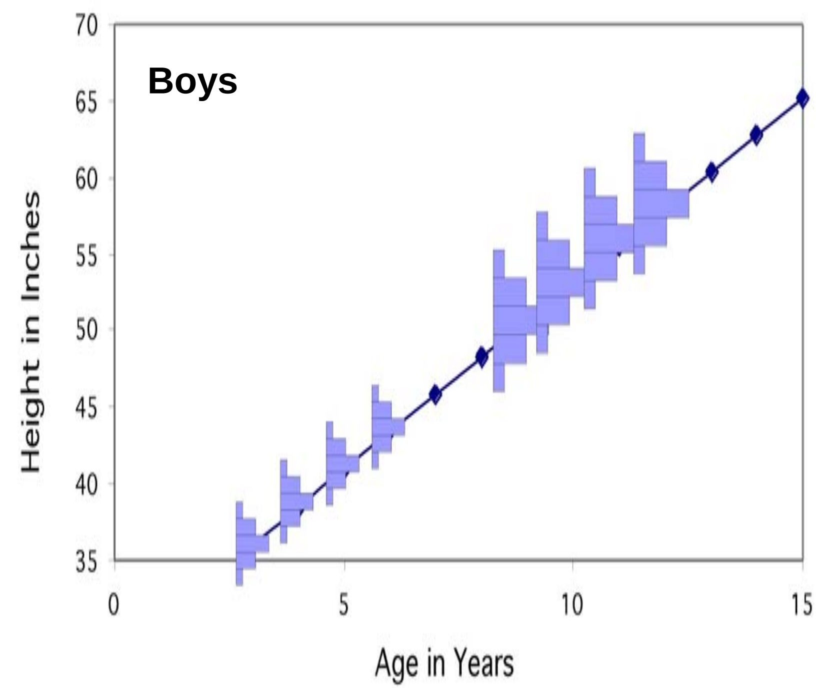
- Many regression or correlation studies use average data.
- While this is sometimes appropriate, you should know that correlations based on averages are usually quite higher than when made on the raw data.



The correlation is a measure of spread (scatter) in a linear relationship. Using averages greatly reduces the scatter.

Therefore *r* and r^2 are typically greatly increased when averages are used.





Each dot represents an average. The variation among boys per age class is not shown.

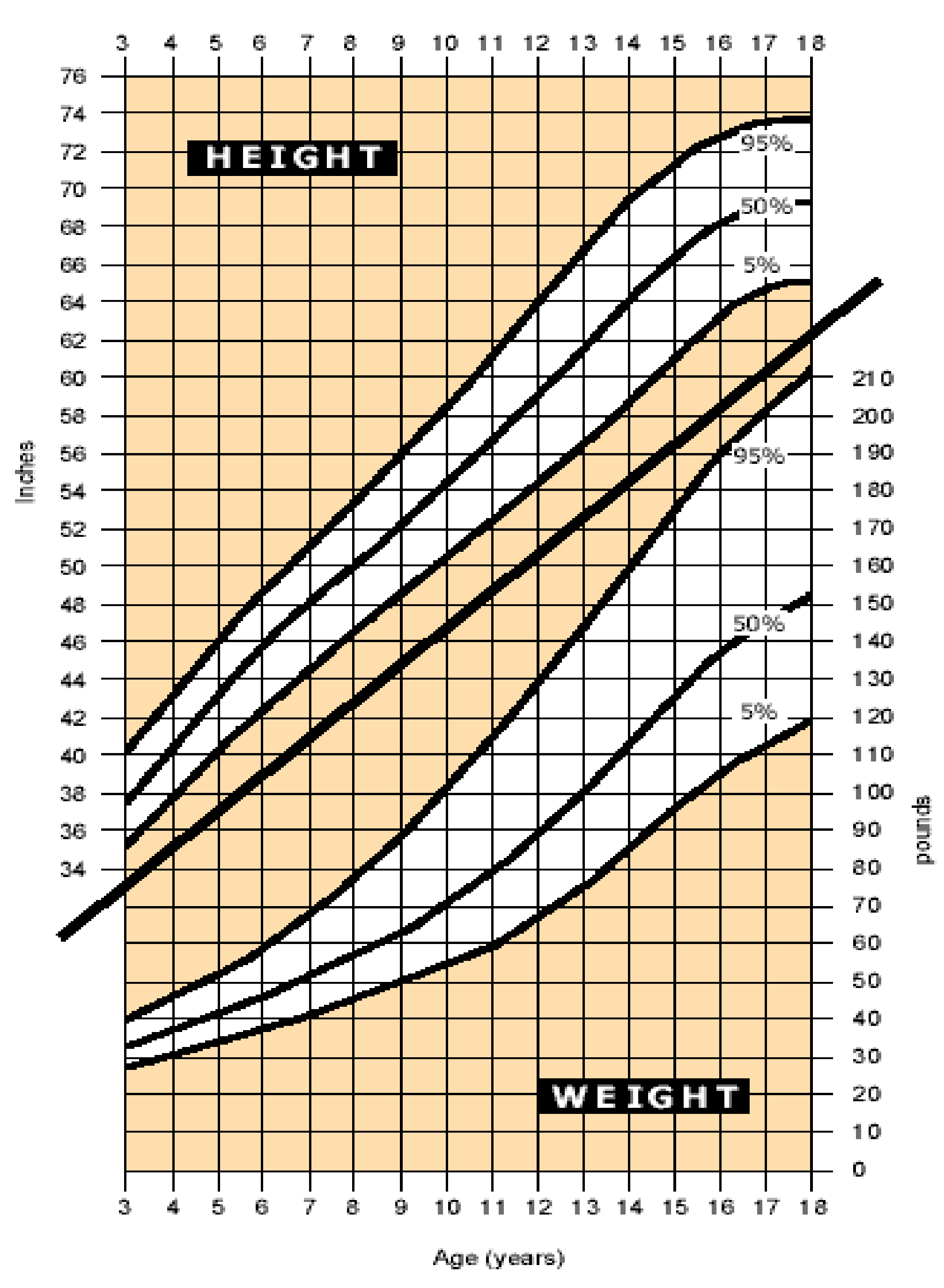
These histograms illustrate that each mean represents a distribution of boys of a particular age.

Should parents be worried if their son does not match the point for his age?

If the raw values were used in the correlation instead of the mean there would be a lot of spread in the *y*-direction, and thus the correlation would be smaller.

That's why typically growth charts show a range of values (here from 5th to 95th percentiles).

This is a more comprehensive way of displaying the same information.

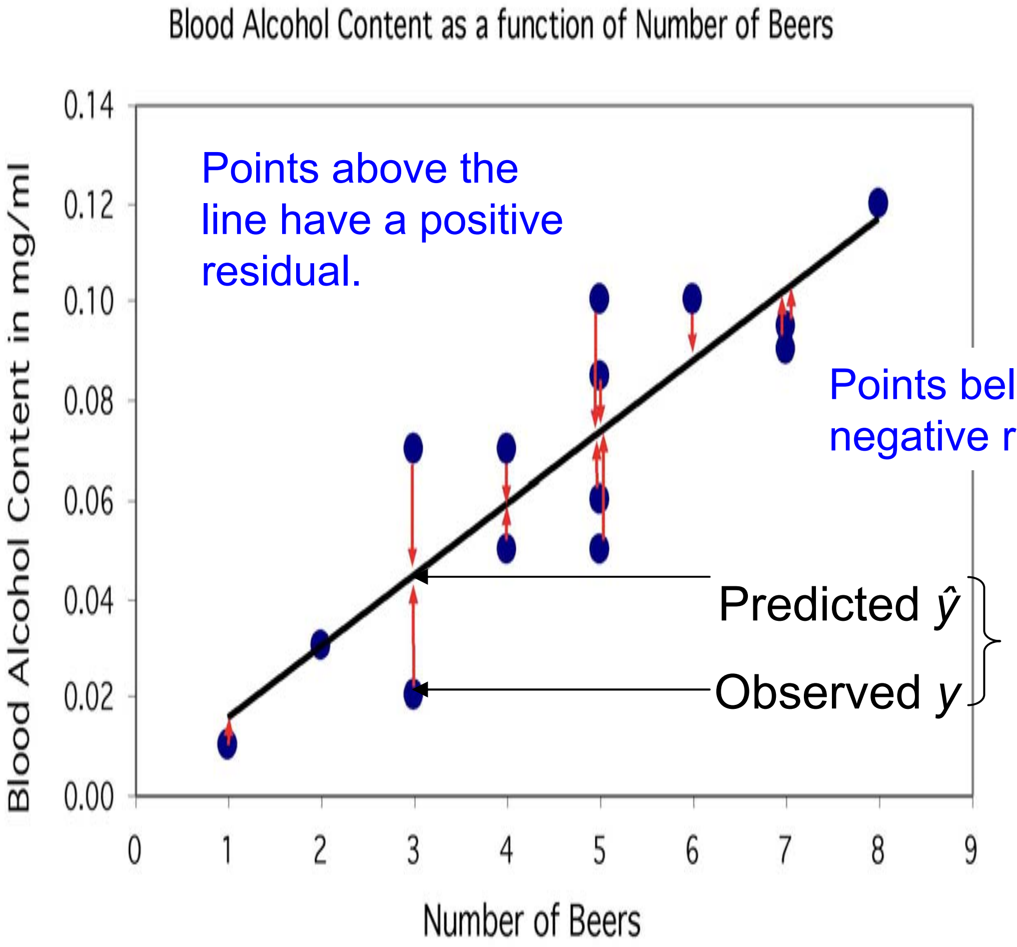


Residuals

The distances from each point to the least-squares regression line give us potentially useful information about the contribution of individual data points to the overall pattern of scatter.

These distances are called **“residuals.”**

The sum of these residuals is always 0.



Points above the line have a positive residual.

Points below the line have a negative residual.

Predicted $\hat{y}$

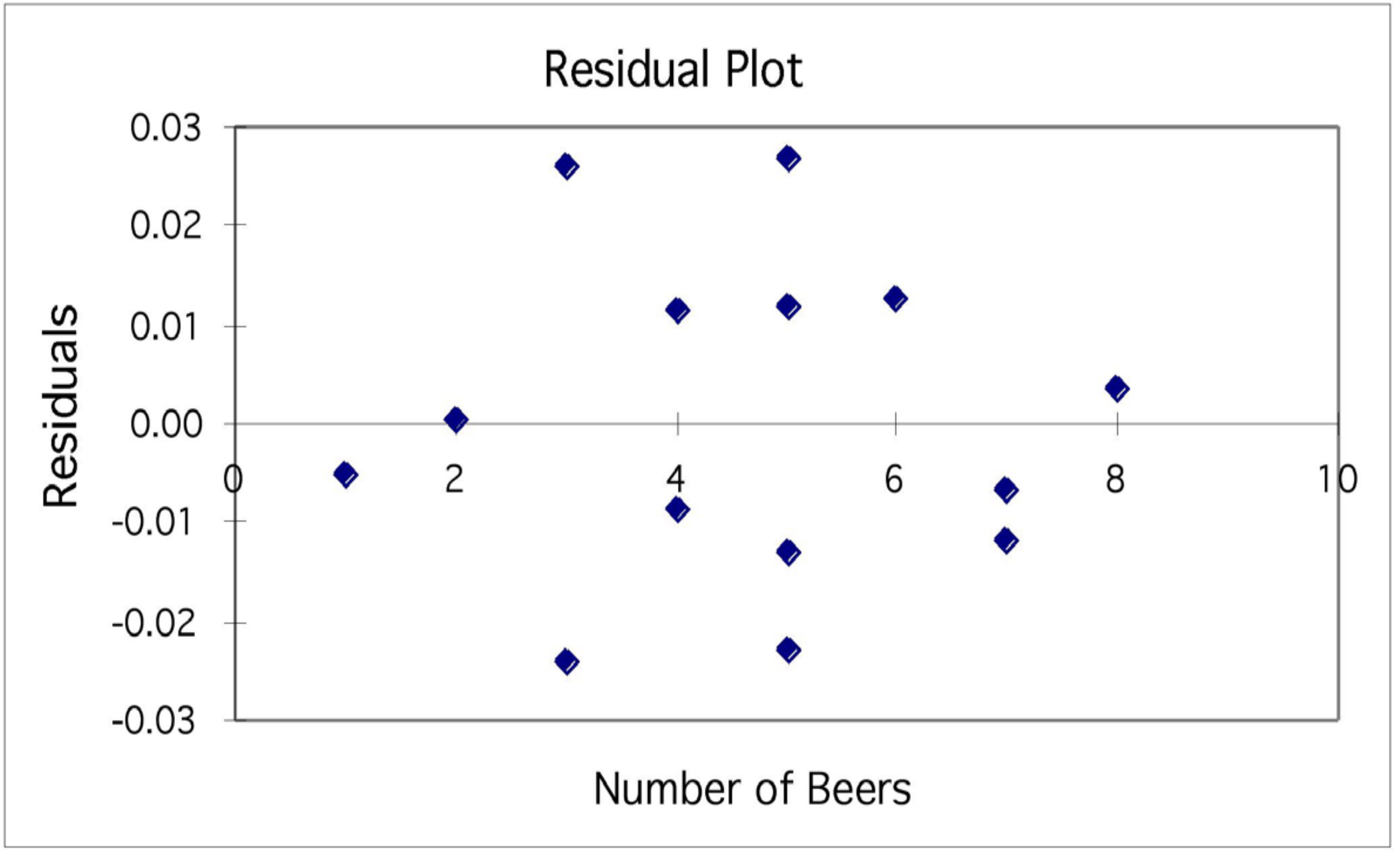
Observed y

$$\text{dist. } (y - \hat{y}) = \text{residual}$$

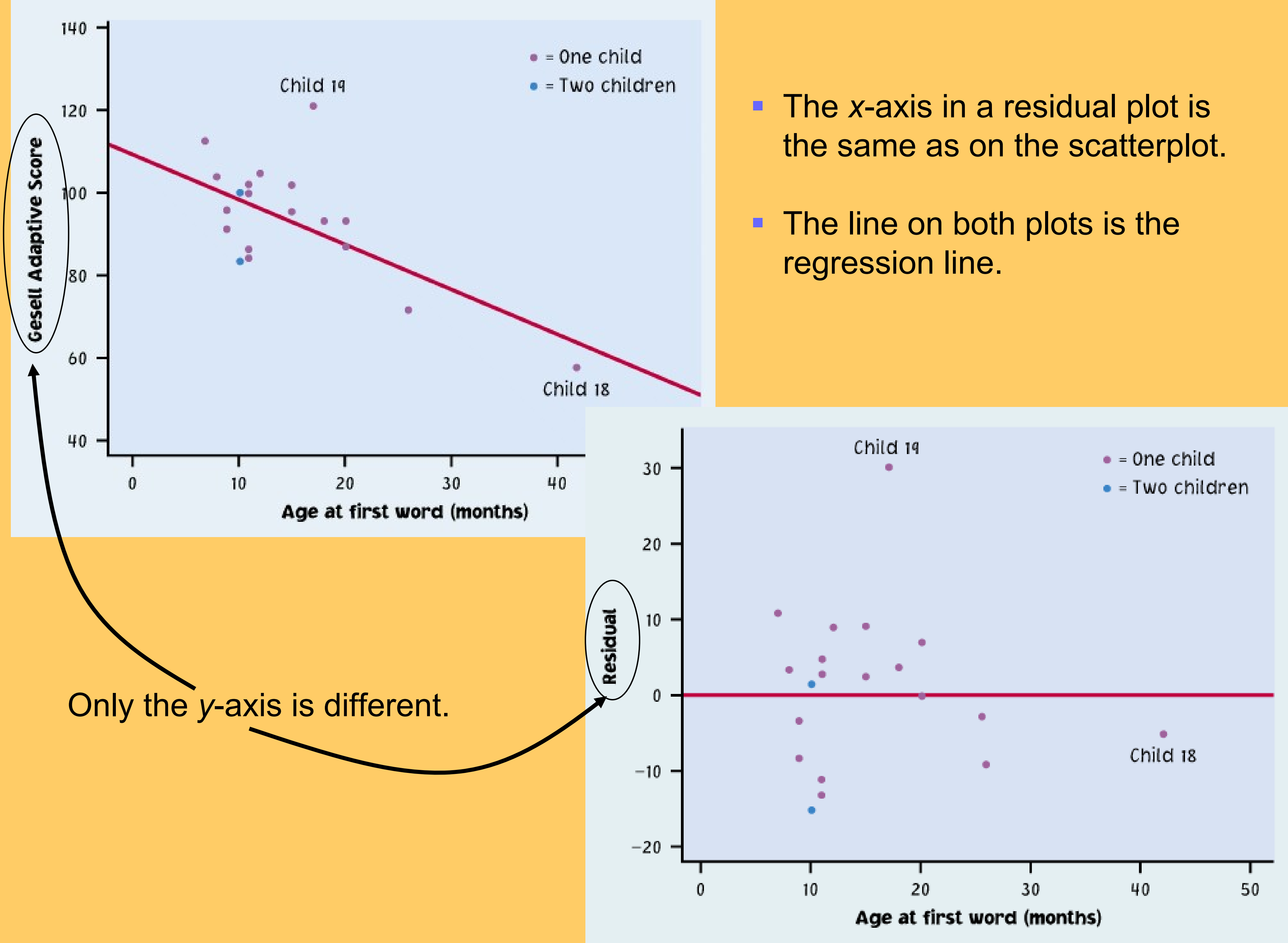
Residual plots

Residuals are the distances between *y*-observed and *y*-predicted. We plot them in a **residual plot.**

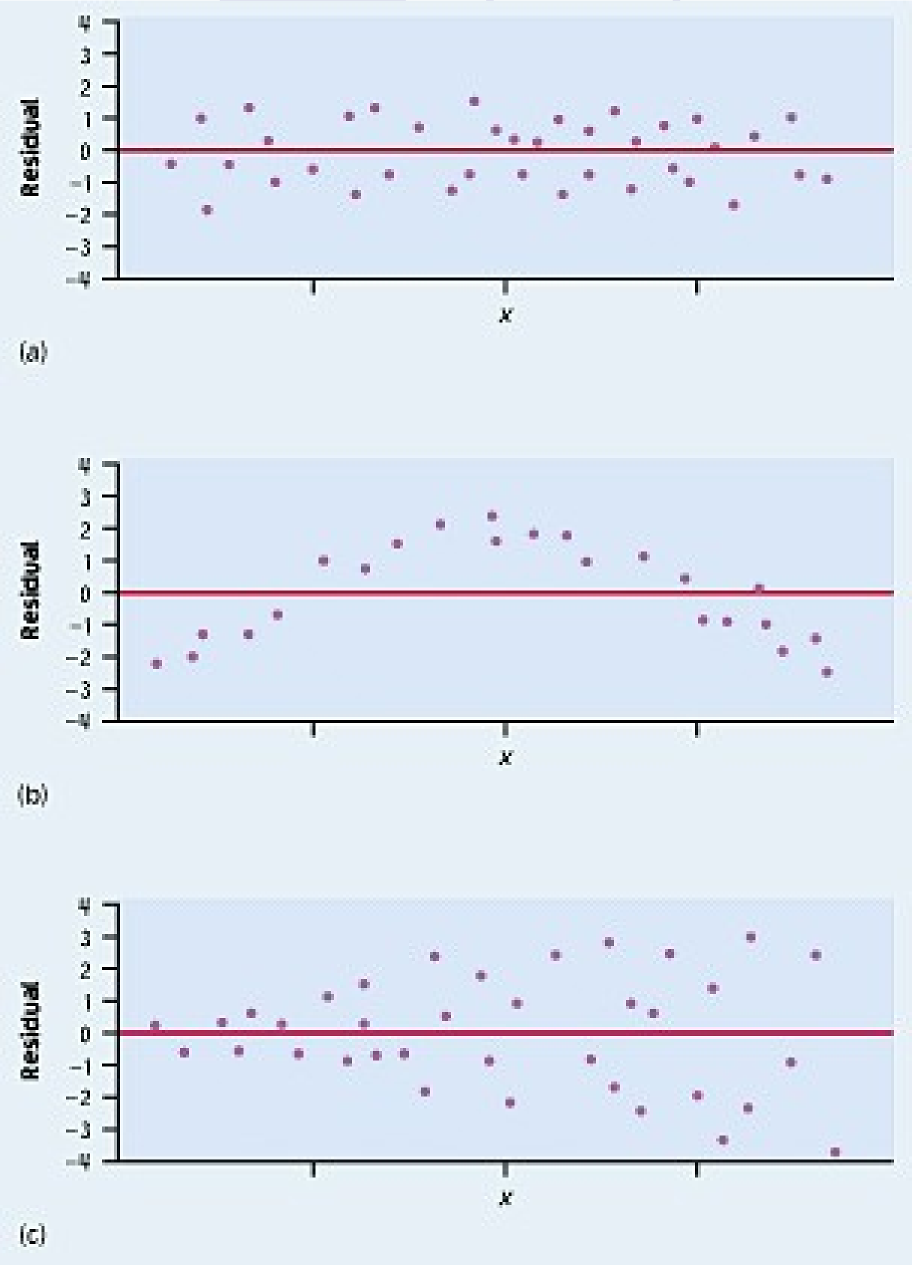
If residuals are scattered randomly around 0, chances are your data will fit a linear model, were normally distributed, and you didn't have outliers.



- The *x*-axis in a residual plot is the same as on the scatterplot.
- The line on both plots is the regression line.



Only the *y*-axis is different.



Residuals are randomly scattered—good!

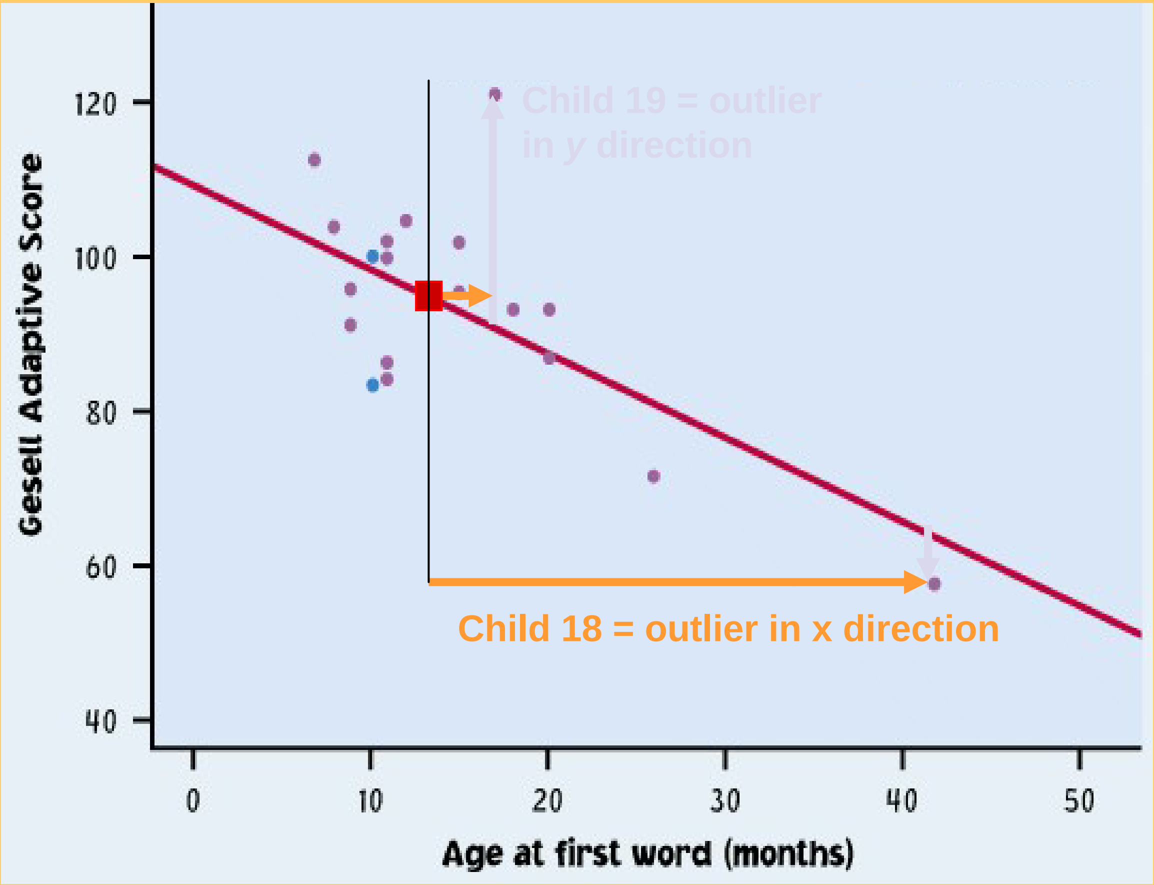
Curved pattern—means the relationship you are looking at is not linear.

A change in variability across plot is a warning sign. You need to find out why it is, and remember that predictions made in areas of larger variability will not be as good.

Outliers and influential points

Outlier: observation that lies outside the overall pattern of observations.

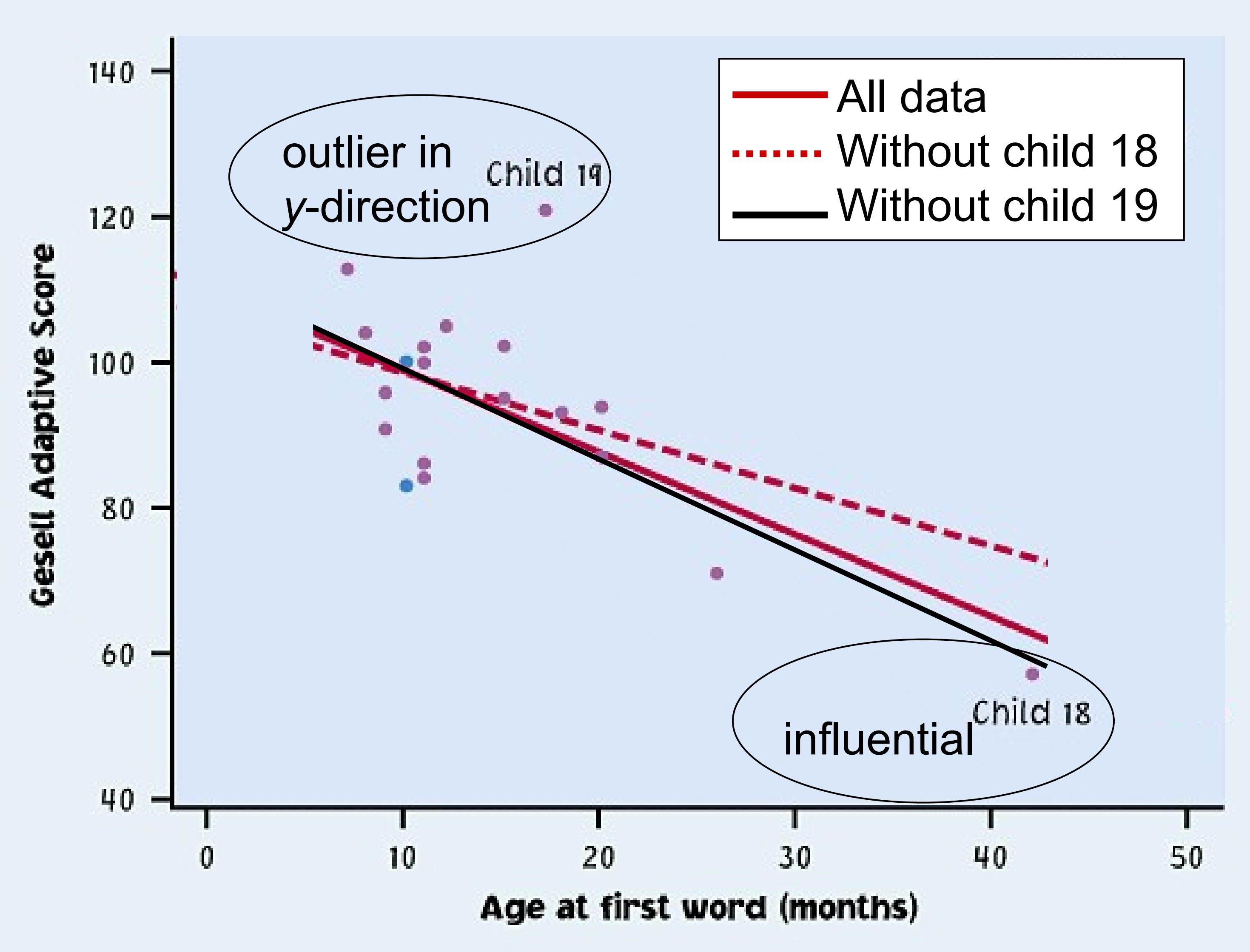
“Influential individual”: observation that markedly changes the regression if removed. This is often an outlier on the *x*-axis.



Child 19 is an outlier of the relationship.

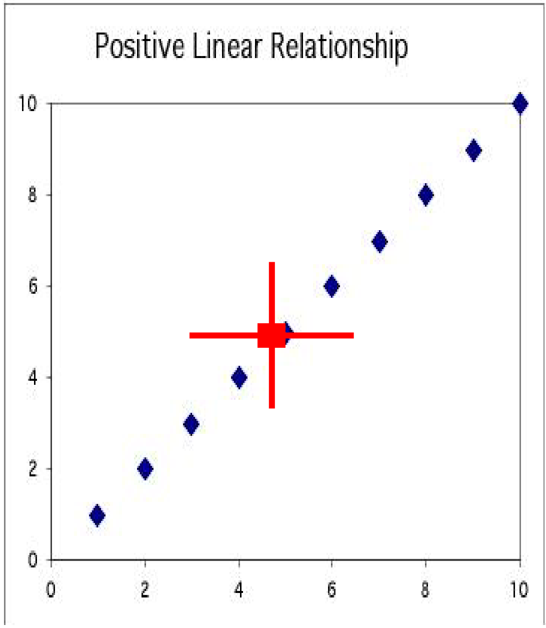
Child 18 is only an outlier in the *x* direction and thus might be an influential point.

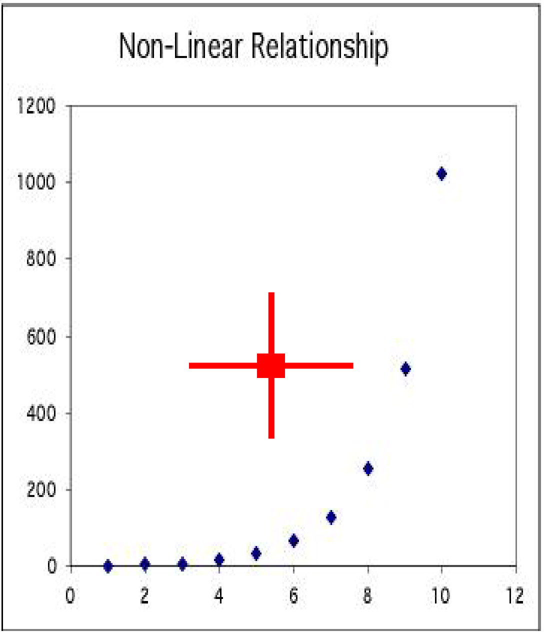
Are these points influential?



Always plot your data

A correlation coefficient and a regression line can be calculated for any relationship between two quantitative variables. However, outliers greatly influence the results and running a linear regression on a nonlinear association is not only meaningless but misleading.





So make sure to always plot your data before you run a correlation or regression analysis.

Always plot your data!

The correlations all give $r \approx 0.816$, and the regression lines are all approximately $\hat{y} = 3 + 0.5x$. For all four sets, we would predict $\hat{y} = 8$ when $x = 10$.

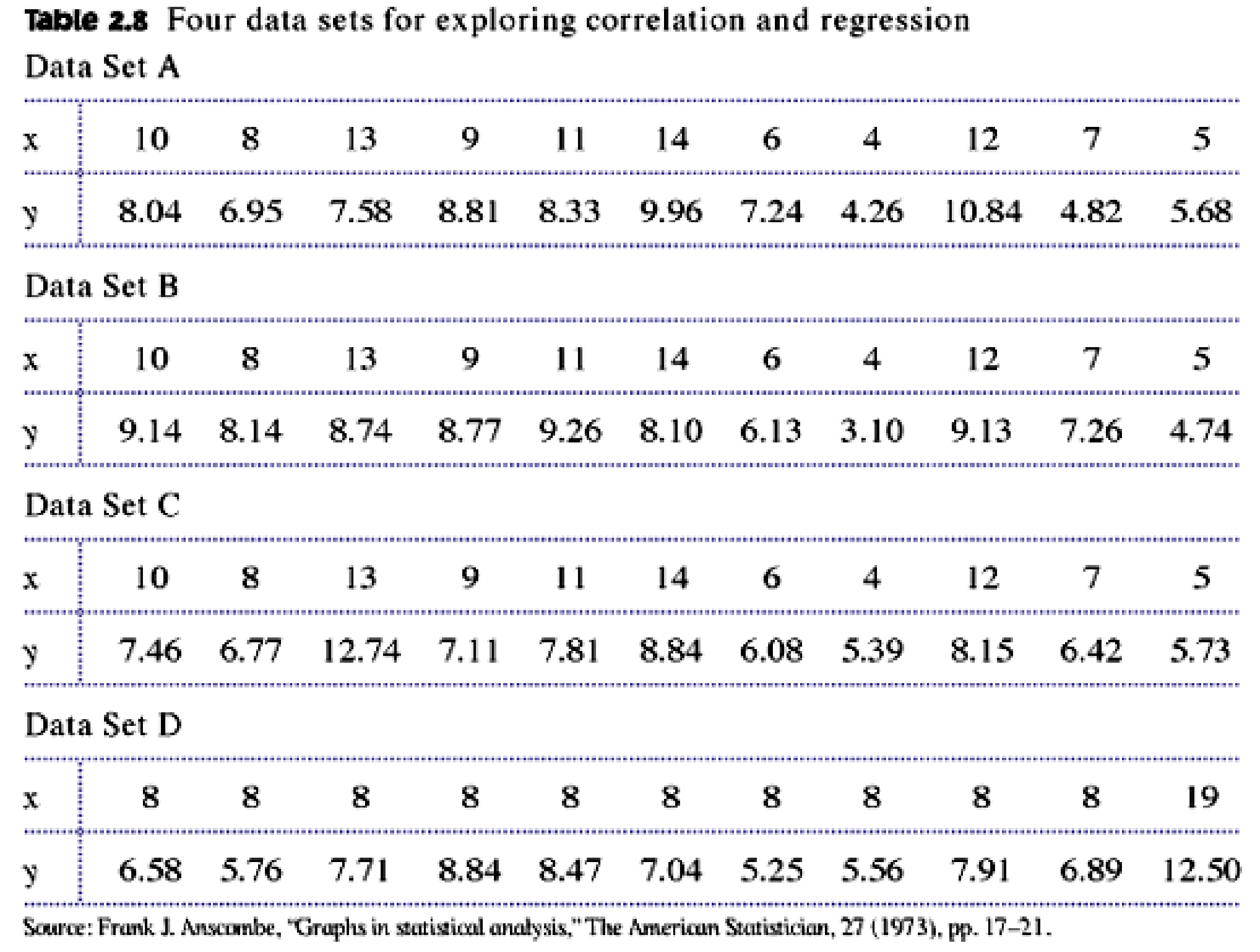


Table 2.8 Four data sets for exploring correlation and regression

Data Set A

x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68

Data Set B

x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74

Data Set C

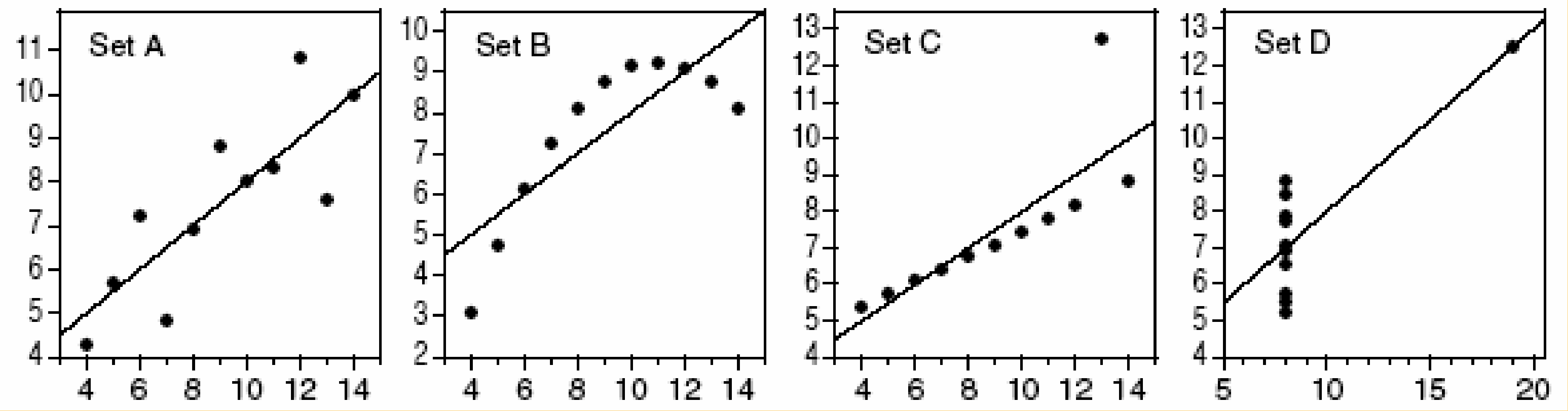
x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

Data Set D

x	8	8	8	8	8	8	8	8	8	8	19
y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

Source: Frank J. Anscombe, "Graphs in statistical analysis," The American Statistician, 27 (1973), pp. 17–21.

However, making the scatterplots shows us that the correlation/ regression analysis is not appropriate for all data sets.



Moderate linear association; regression OK.

Obvious nonlinear relationship; regression not OK.

One point deviates from the highly linear pattern; this outlier must be examined closely before proceeding.

Just one very influential point; all other points have the same *x* value; a redesign is due here.

Lurking variables

A **lurking variable** is a variable not included in the study design that does have an effect on the variables studied.

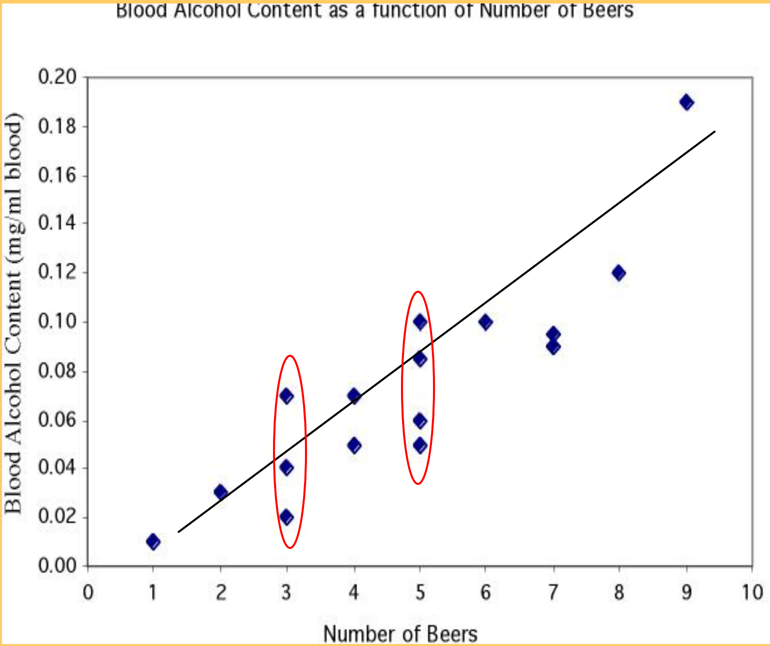
Lurking variables can *falsely*
relationship.

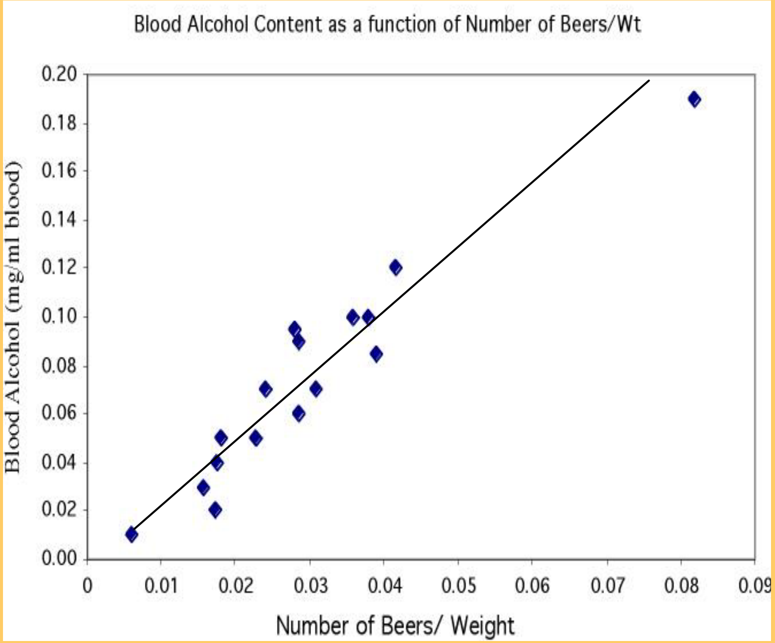
- Strong positive associati
number of firefighters at a f he
amount of damage a fire d

What is the lurking variable in t es?

Ho ould you answer if you didn't know anything about the topic?

○ Negative association between





There is quite some variation in BAC for the same number of beers drunk. A person's blood volume is a factor in the equation that we have overlooked.

Now we change number of beers to number of beers/weight of person in lb.

The scatter is much smaller now. **One's weight was indeed influencing the response variable "blood alcohol content."**

Vocabulary: lurking vs. confounding

- 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.

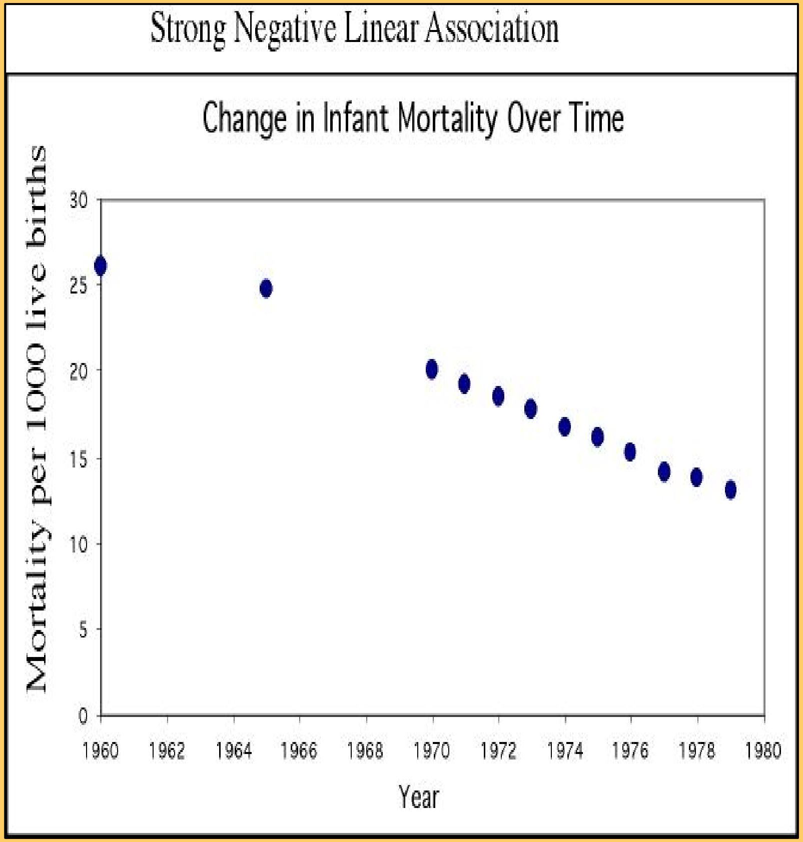
- Two variables are **confounded** when their effects on a response variable cannot be distinguished from each other. The confounded variables may be either explanatory variables or lurking variables.

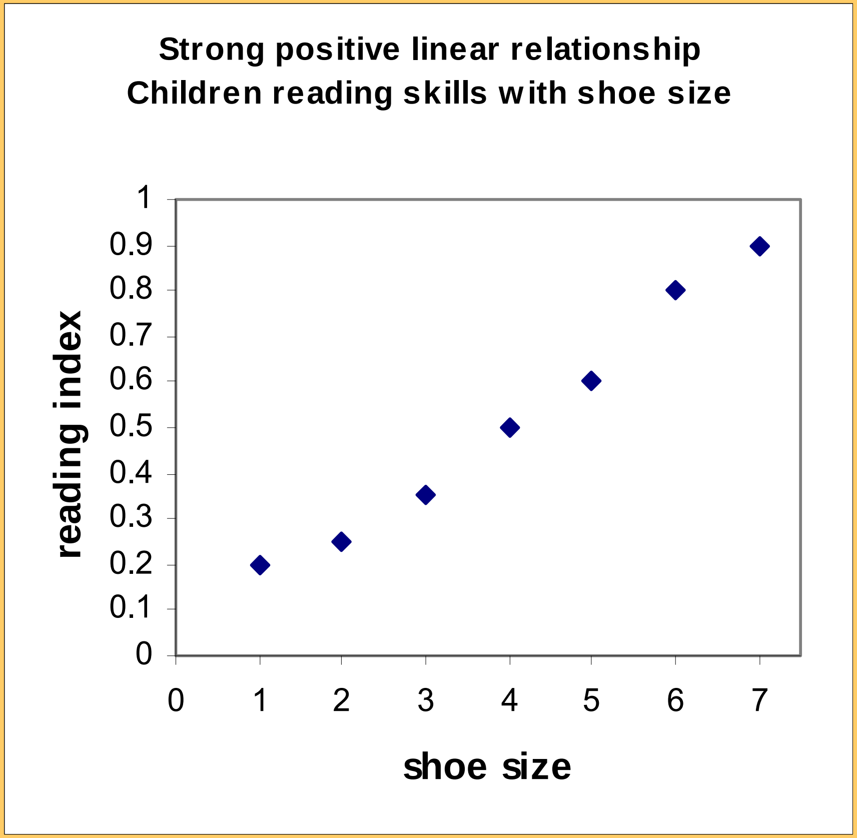
But you often see them used interchangeably…

Association and causation

Association, however strong, does NOT imply causation.

Only careful experimentation can show causation.





Not all examples are so obvious…

Establishing causation

It appears that lung cancer is associated with smoking.

How do we know that both of these variables are not being affected by an unobserved third (lurking) variable?

For instance, what if there is a genetic predisposition that causes people to both get lung cancer *and* become addicted to smoking, but the smoking itself doesn't CAUSE lung cancer?

We can evaluate the association using the following criteria:

1) The association is strong.
2) The association is consistent.
3) Higher doses are associated with stronger responses.
4) Alleged cause precedes the effect.
5) The alleged cause is plausible.

Caution before rushing into a correlation or a regression analysis

- Do not use a regression on inappropriate data.
  - ✓ Pattern in the residuals
  - ✓ Presence of large outliers
  - ✓ Clumped data falsely appearing linear

  } *Use residual plots for help.*

- Beware of lurking variables.
- Avoid extrapolating *(going beyond interpolation).*
- Recognize when the correlation/regression is performed on averages.
- A relationship, however strong, does not itself imply causation.