4. Relationships: Regression

The Practice of Statistics in the Life Sciences

Third Edition

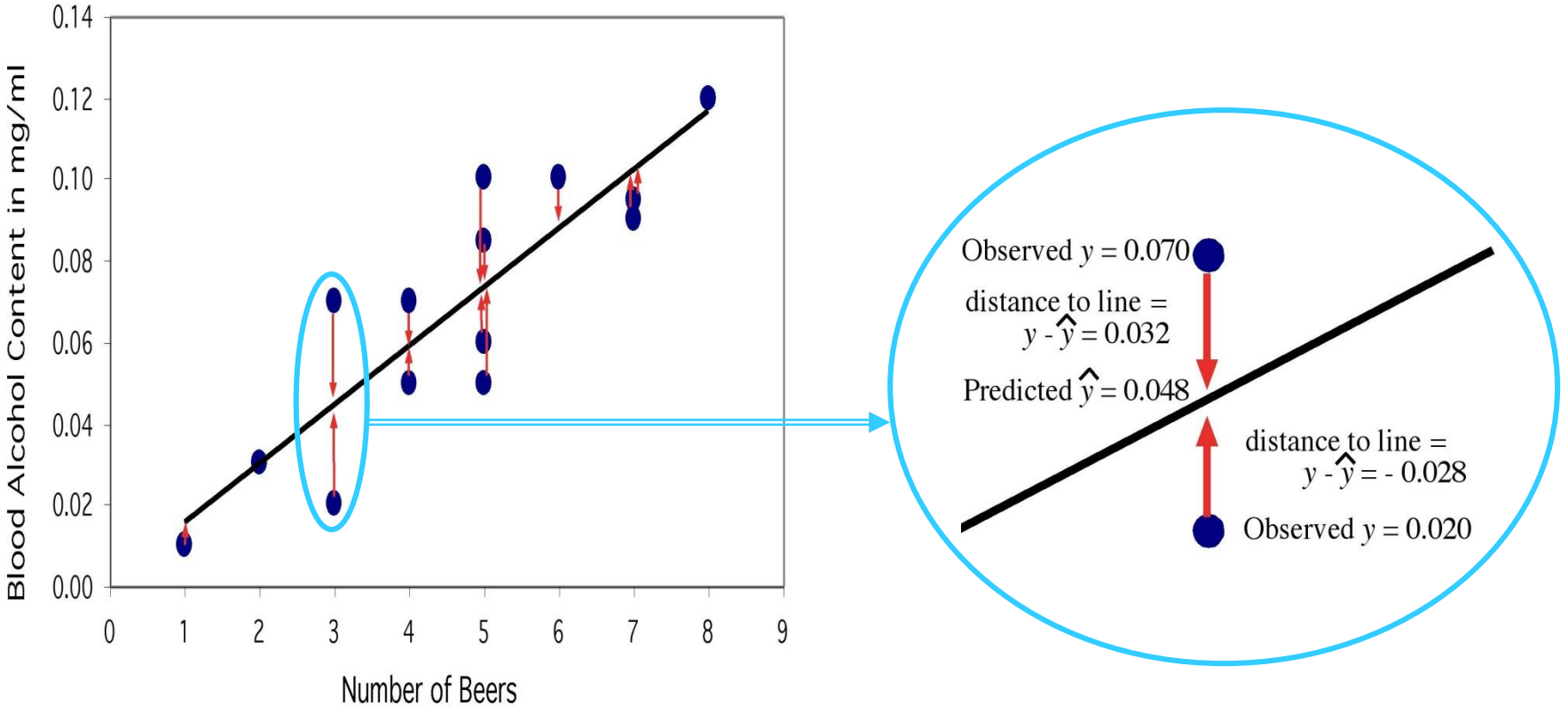
Objectives (PSLS Chapter 4)

Regression

- The least-squares regression line
- Finding the least-squares regression line
- The coefficient of determination, r^2
- Outliers and influential observations
- Making predictions
- Association does not imply causation

The least-squares regression line

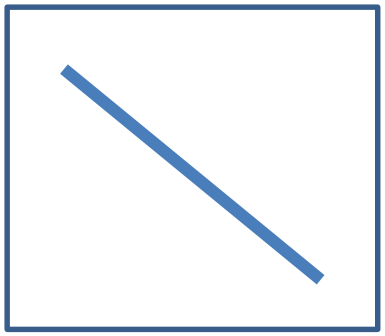
The **least-squares regression line** is the unique line such that the sum of the **vertical distances** between the data points and the line is zero, and the sum of the squared vertical distances is the smallest possible.



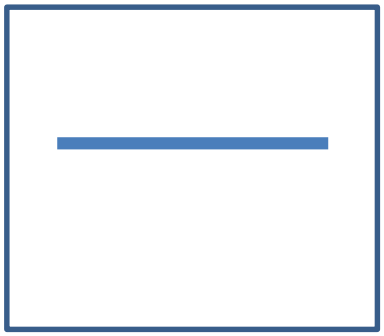
Notation

$\hat{y}$ is the predicted *y* value on the regression line

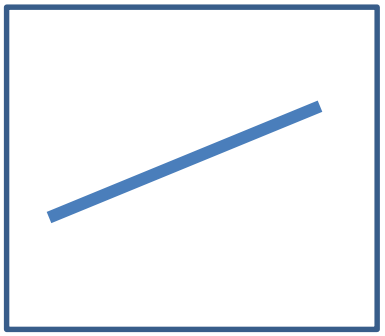
$$\hat{y} = \text{intercept} + \text{slope}\, x \qquad \hat{y} = a + bx$$



slope < 0



slope = 0



slope > 0

Not all calculators/software use this convention. Other notations include:

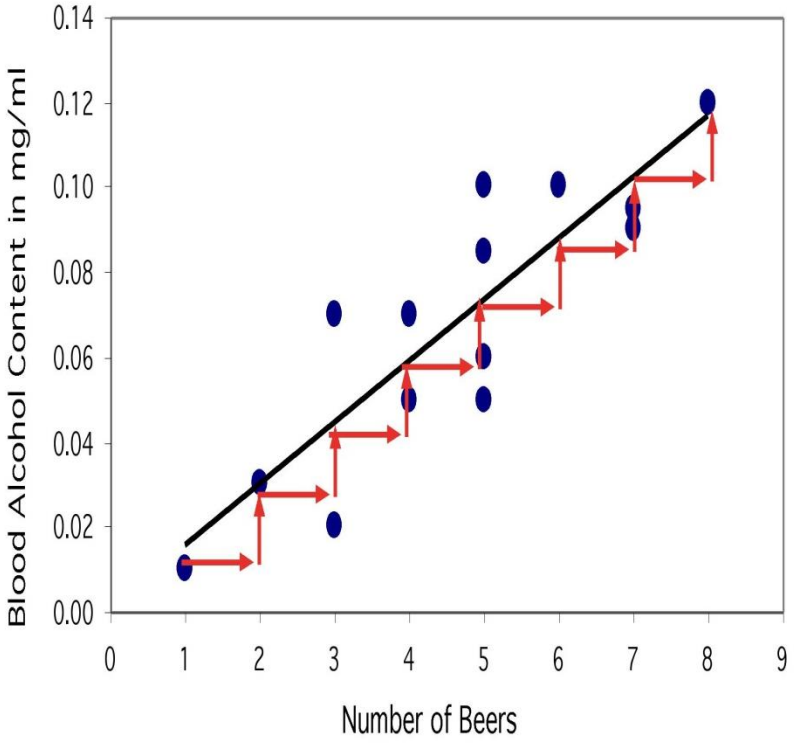
$$\hat{y} = ax + b$$

$$\hat{y} = b_0 + b_1 x$$

$$\hat{y} = \text{variable\_name}\, x + \text{constant}$$

Interpretation

The **slope** of the regression line describes how much we expect *y* to change, on average, for every unit change in *x*.



The **intercept** is a necessary mathematical descriptor of the regression line. It does not describe a specific property of the data.

Finding the least-squares regression line

The **slope of the regression line, *b*,** equals: $b = r\frac{s_y}{s_x}$

r is the correlation coefficient between x and y
s_y is the standard deviation of the response variable y
s_x is the standard deviation of the explanatory variable x

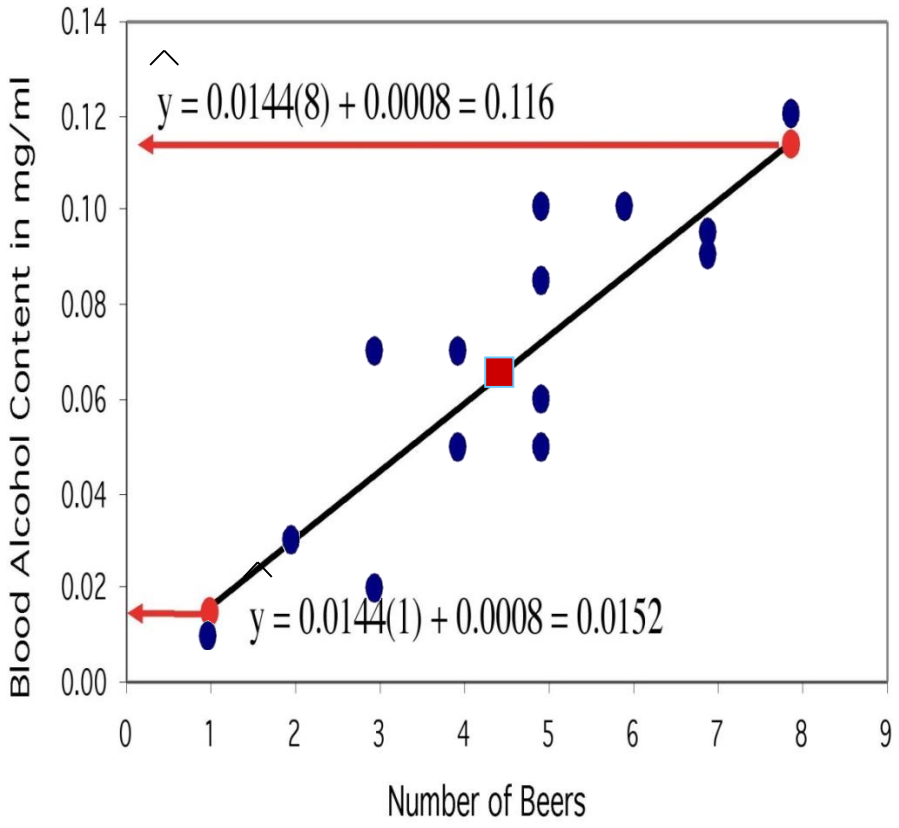
The **intercept, *a*,** equals: $a = \bar{y} - b\bar{x}$

$\bar{x}$ and $\bar{y}$ are the respective means of the x and y variables

Plotting the least-square regression line

Use the regression equation to find the value of y for two distinct values of x, and draw the line that goes through those two points.

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



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

They are NOT actual points from the data set (except by pure coincidence).

Least-squares regression is only for linear associations

Don't compute the regression line until you have confirmed that there is a linear relationship between *x* and *y*.

ALWAYS PLOT THE RAW DATA

These data sets all give a linear regression equation of about $\hat{y} = 3 + 0.5x$.

But don't report that until you have plotted the data.

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

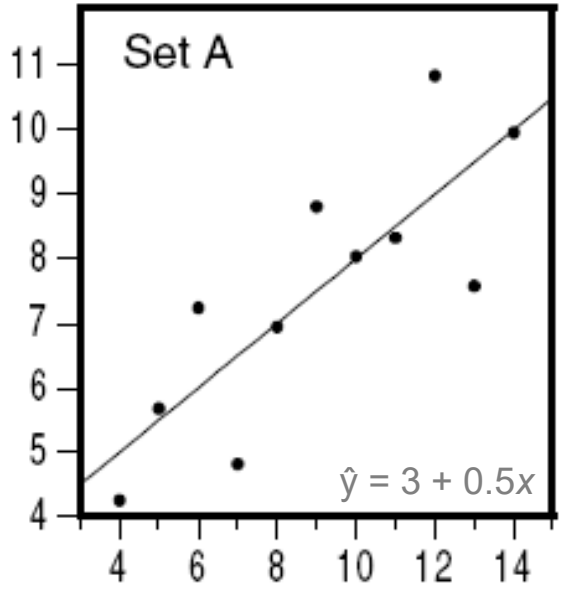
x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

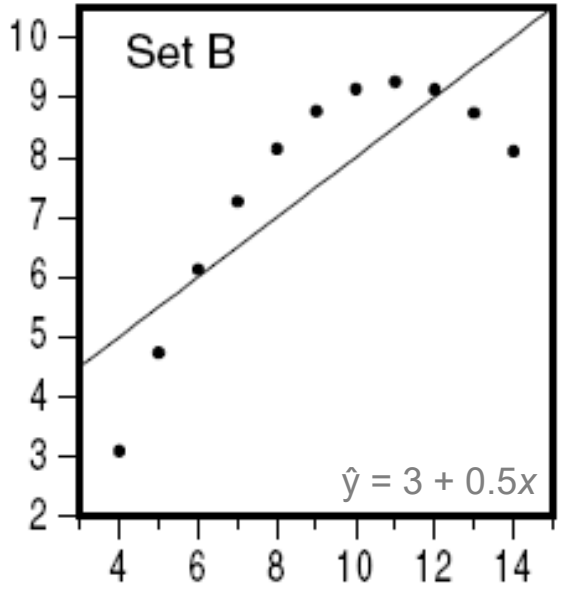
Data Set D

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.

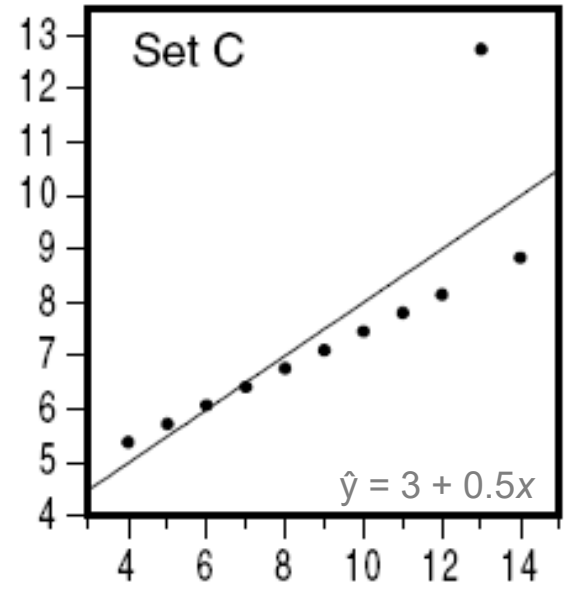
Moderate linear association; regression OK.

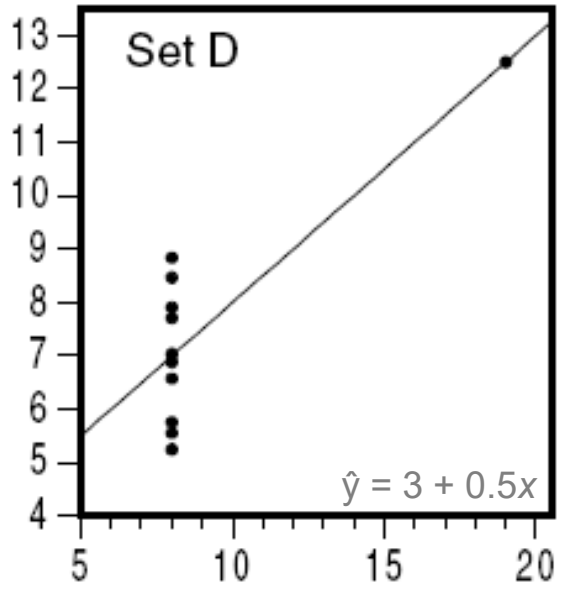




Obvious nonlinear relationship; regression inappropriate.

One extreme outlier, requiring further examination.



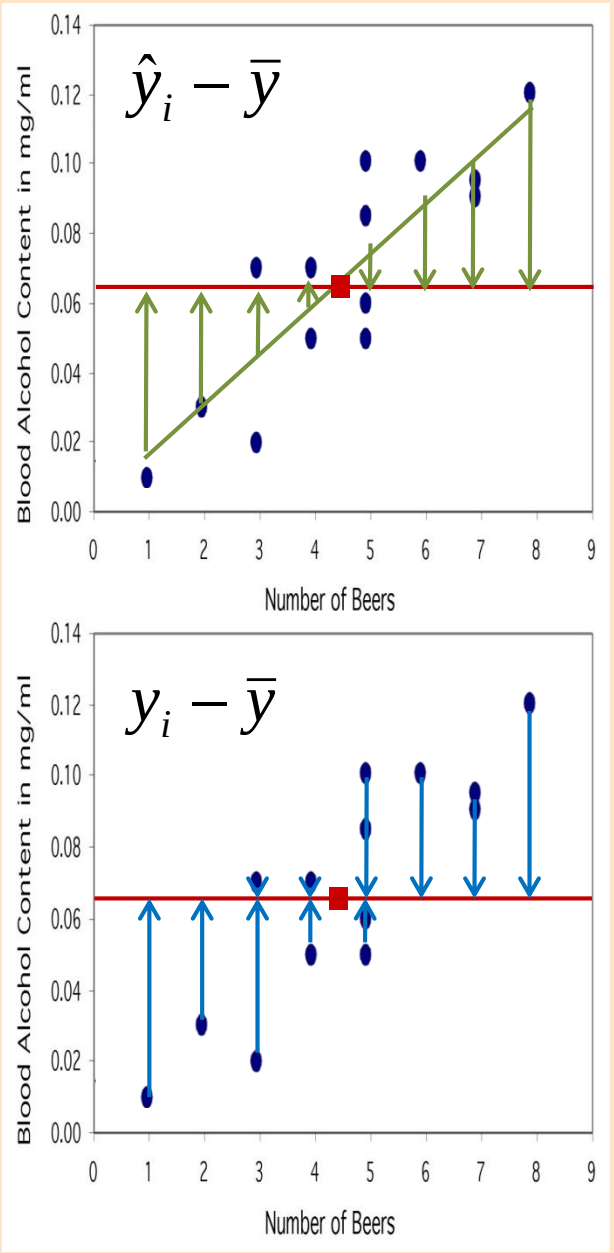


Only two values for *x;* a redesign is due here…

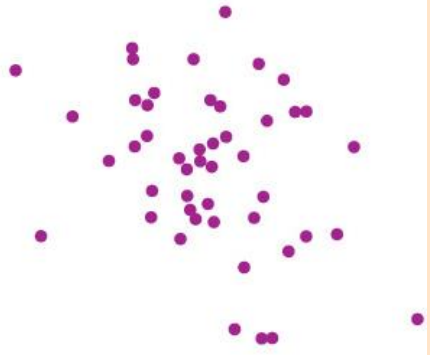
The coefficient of determination, r^2

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

r^2 represents **the fraction of the variance in *y* that can be explained by the regression model.**



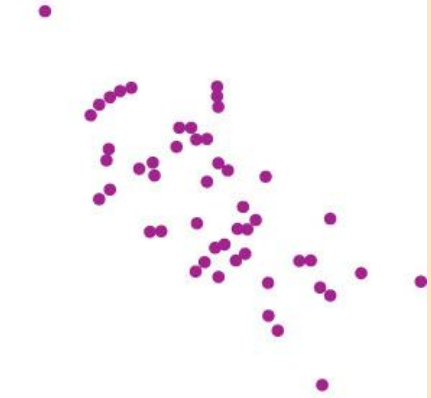
$r = 0.87$, so $r^2 = 0.76$
This model explains 76% of individual variations in BAC



Correlation $r = -0.3$

$r = -0.3$, $r^2 = 0.09$, or 9%

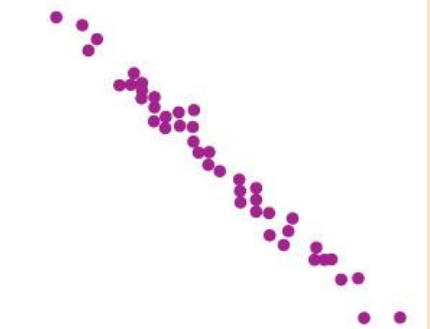
The regression model explains not even 10% of the variations in y.



Correlation $r = -0.7$

$r = -0.7$, $r^2 = 0.49$, or 49%

The regression model explains nearly half of the variations in y.



Correlation $r = -0.99$

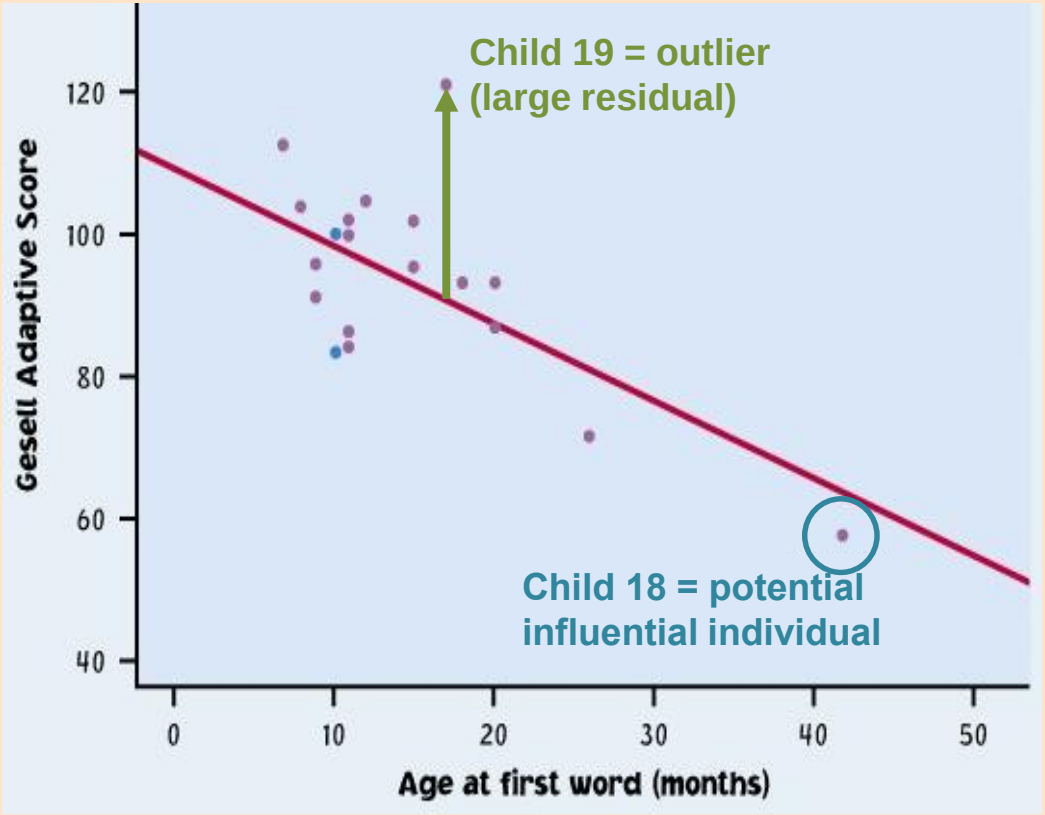
$r = -0.99$, $r^2 = 0.9801$, or ~98%

The regression model explains almost all of the variations in y.

Outliers and influential points

Outlier: An observation that lies outside the overall pattern.

"Influential individual": An observation that markedly changes the regression if removed. This is often an isolated point.



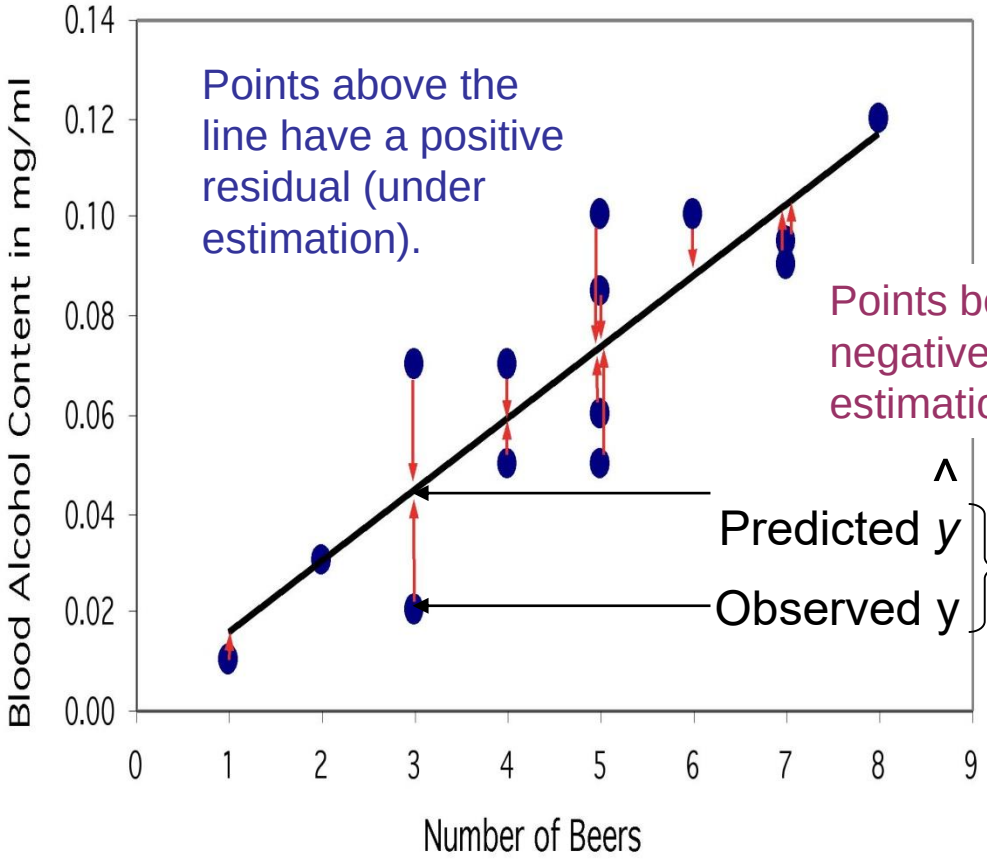
Child 19 is an outlier of the relationship (it is unusually far from the regression line, vertically).

Child 18 is isolated from the rest of the points, and might be an influential point.

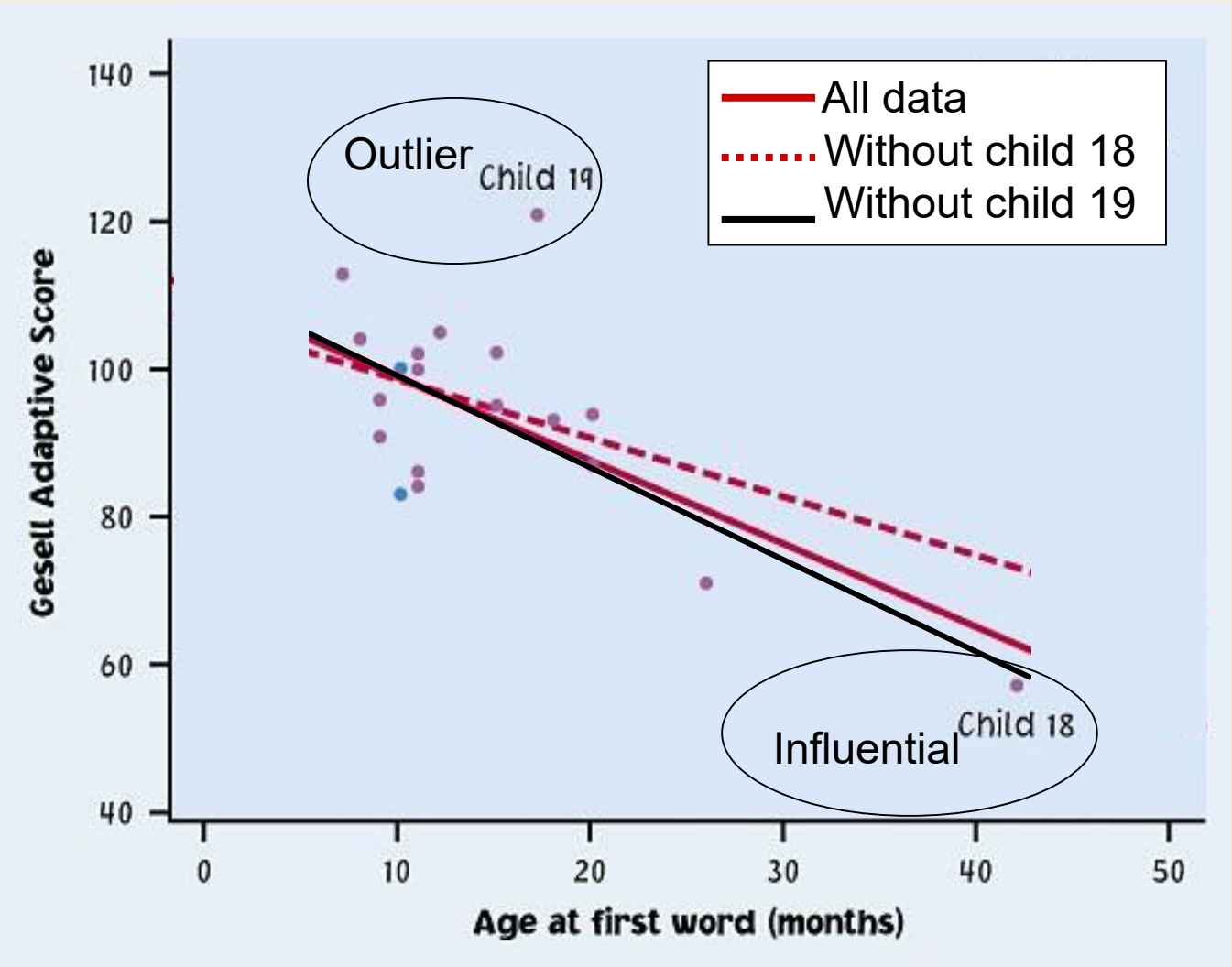
Residuals

The vertical distances from each point to the least-squares regression line are called **residuals.** The sum of all the residuals is by definition 0.

Outliers have unusually large residuals (in absolute value).



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



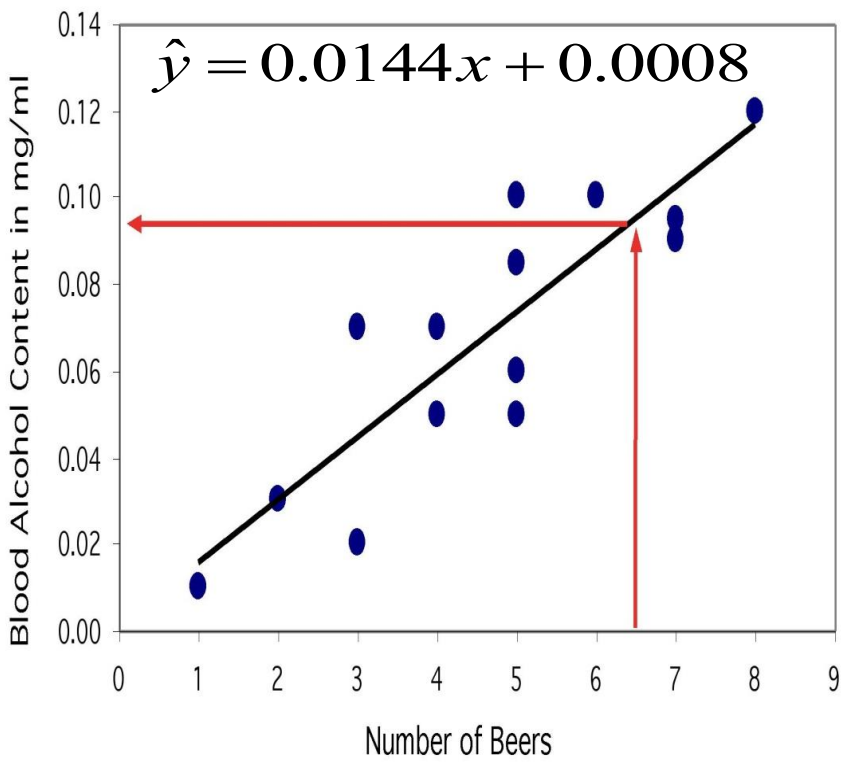
Child 18 changes the regression line substantially when it is removed. So, Child 18 is indeed an influential point.

Child 19 is an outlier of the relationship, but it is not influential (regression line changed very little by its removal).

Making predictions

Use the equation of the least-squares regression to **predict** *y* for any value of *x* **within the range studied**.

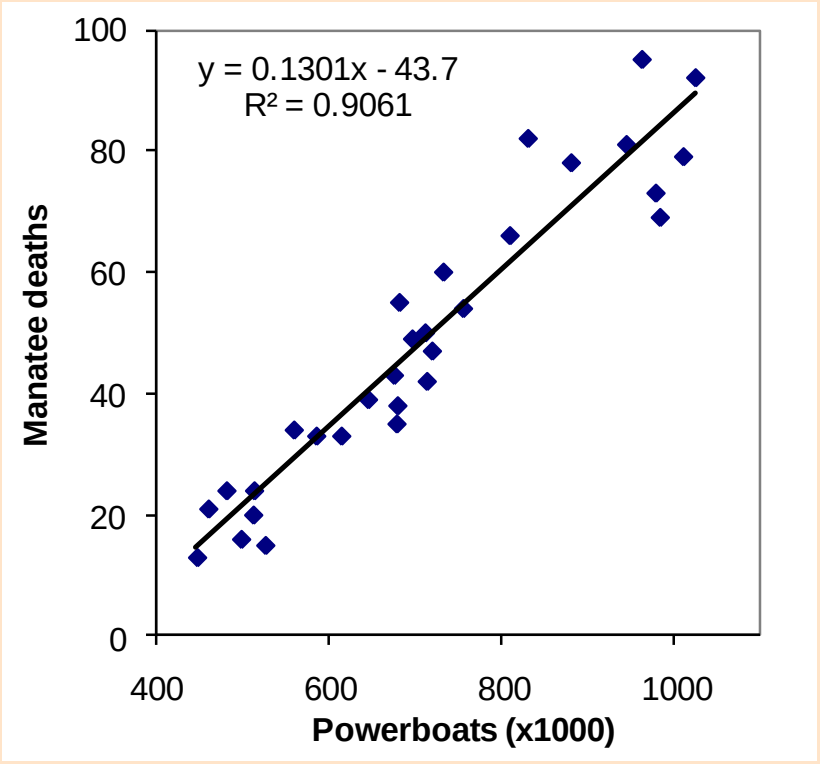
Predication outside the range is extrapolation. ***Avoid extrapolation***.



What would we expect for the BAC after drinking 6.5 beers?

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

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



Thousands powerboats	Manatee deaths
447	13
460	21
481	24
498	16
513	24
512	20
526	15
559	34
585	33
614	33
645	39
675	43
711	50
719	47
681	55
679	38
678	35
696	49
713	42
732	60
755	54
809	66
830	82
880	78
944	81
962	95
978	73
983	69
1010	79
1024	92

The least-squares regression line is:

$$\hat{y} = 0.1301x - 43.7$$

If Florida were to limit the number of powerboat registrations to 500,000, what could we expect for the number of manatee deaths in a year?

$$\hat{y} = 0.1301(500) - 43.7 \Rightarrow \hat{y} = 65.05 - 43.7 = 21.35$$

➔ Roughly 21 manatee deaths.

Could we use this regression line to predict the number of manatee deaths for a year with 200,000 powerboat registrations?

Association does not imply causation

Association, however strong, does NOT imply causation.

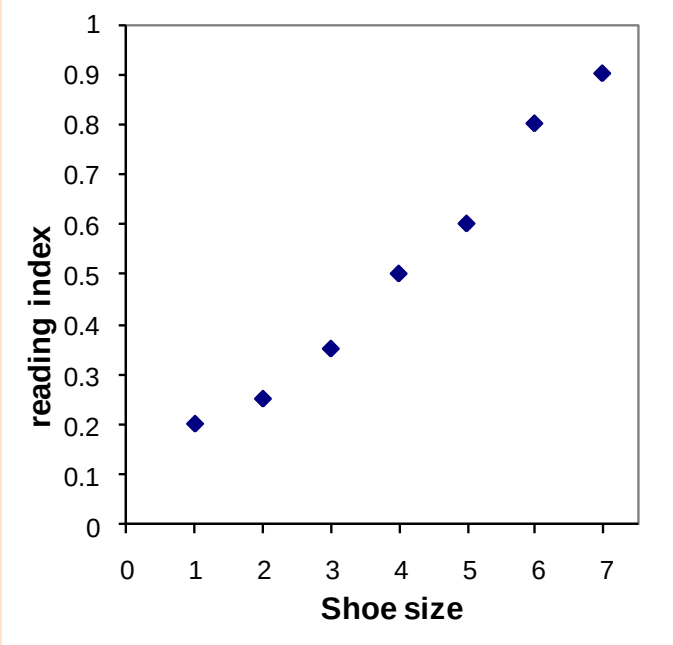
The observed association could have an external cause.

- A **lurking variable** is a variable that is not among the explanatory or response variables in a study, and yet may influence the relationship between the variables studied.

- We say that two variables are **confounded** when their effects on a response variable cannot be distinguished from each other.

In each example, what is most likely the lurking variable? Notice that some cases are more obvious than others.

Strong positive association between the shoe size and reading skills in young children.



Strong positive association between the number firefighters at a fire site and the amount of damage a fire does.

Negative association between moderate amounts of wine-drinking and death rates from heart disease in developed nations.

Establishing causation

Establishing causation from an observed association can be done if:

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

Lung cancer is clearly associated with smoking.

What if a genetic mutation (lurking variable) caused people to both get lung cancer and become addicted to smoking?

It took years of research and accumulated indirect evidence to reach the conclusion that smoking causes lung cancer.