Lecture 4

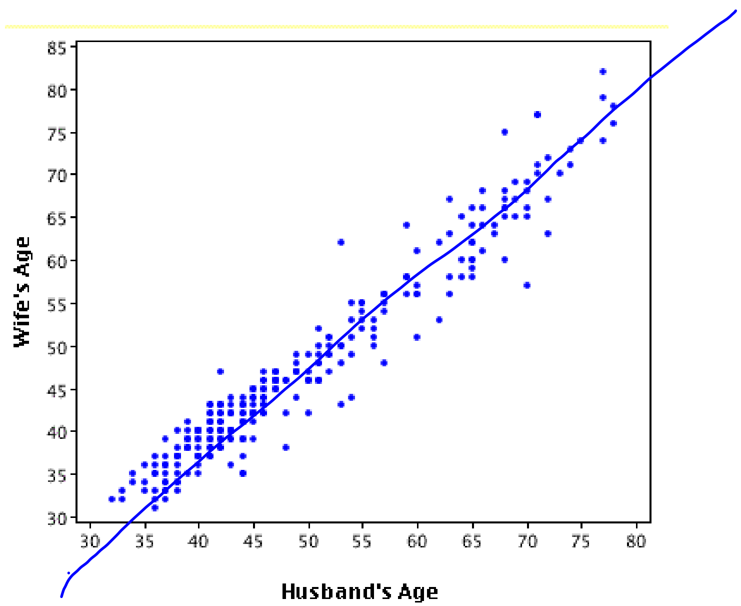
Scatterplots, Association, and Correlation

Previously, we looked at

- Single variables on their own
- One or more categorical variables

In this lecture: We shall look at two quantitative variables.

First tool to do so: a **scatterplot**!



Two variables measured on the same cases are **associated** if knowing the value of one of the variables tells you something about the values of the other variable that you would not know without this information.

Example: You visit a local Starbucks to buy a Mocha Frappuccino. The barista explains that this blended coffee beverage comes in three sizes and asks if you want a Small, a Medium, or a Large.

The prices are \$3.15, \$3.65, and \$4.15, respectively. There is a clear association between the size and the price.

Size $\longrightarrow$ Price

$\hookrightarrow$ strong association

When you examine the relationship, ask yourself the following questions:

- What individuals or cases do the data describe?

- What variables are present? How are they measured?

- Which variables are quantitative and which are categorical?

For the example above: Size is categorical

Price is quantitative

New question might arise:

- Is your purpose simply to explore the nature of the relationship, or do you hope to show that one of the variables can explain variation in the other?

<u>Definition</u>: A **response variable** measures an outcome of a study. An **explanatory variable** explains or causes changes in the response variable.

Example: How does drinking beer affect the level of alcohol in our blood?

The legal limit for driving in most states is 0.08%. Student volunteers at Ohio State University drank different numbers of cans of beer. Thirty minutes later, a police officer measured their blood alcohol content. Here,

Explanatory variable: # of beer

Response variable: % of alcohol in blood

Remark: You will often see explanatory variables called *independent* variables and response variables called *dependent* variables. We prefer to avoid those words.

Scatterplots:

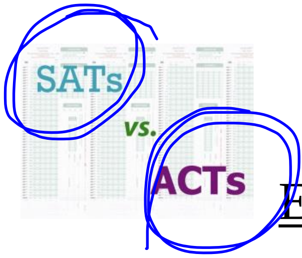
- A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals.

- The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis.

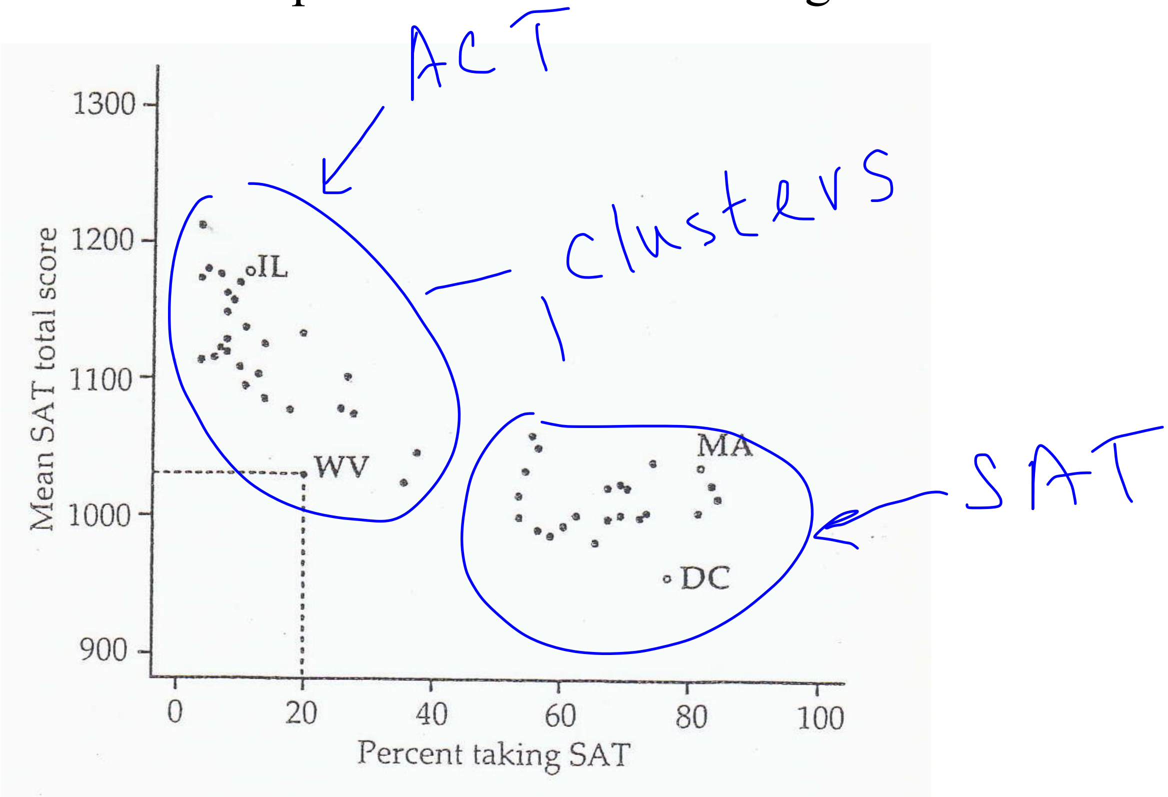
- Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

- Always plot the explanatory variable, if there is one, on the horizontal axis (the x axis) of a scatterplot.

As a reminder, we usually call the explanatory variable $\boldsymbol{x}$ and the response variable $\boldsymbol{y}$. If there is no explanatory response distinction, either variable can go on the horizontal axis.



Example: More than a million high school seniors take the SAT college entrance examination each year. We sometimes see the states "rated" by the average SAT scores of their seniors. Rating states by SAT scores makes little sense, however, because average SAT score is largely explained by what percent of a state's students take the SAT. The scatterplot below allows us to see how the mean SAT score in each state is related to the percent of that state's high school seniors who take the SAT.



WV (20, 1030)

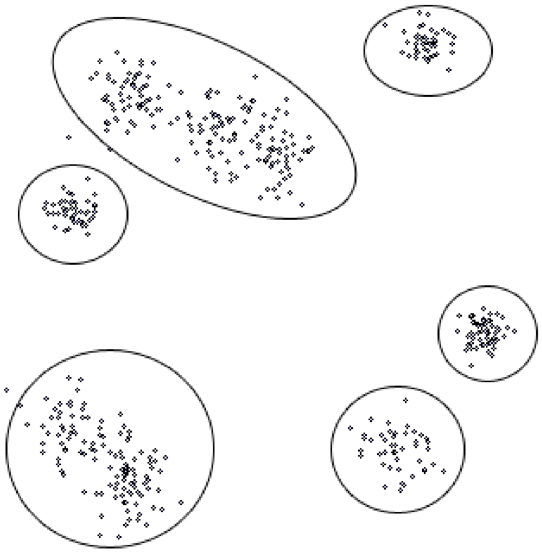
Examining a scatterplot:

- Look for the **overall pattern** and for striking **deviations** from that pattern.

- Describe the overall pattern of a scatterplot by the **form**, **direction**, and **strength** of the relationship.

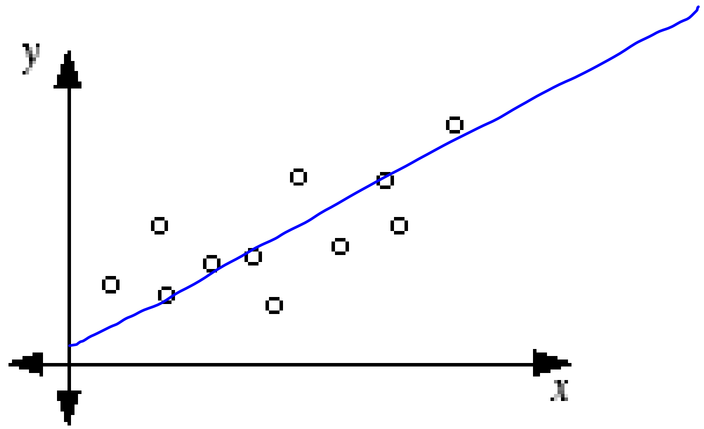
- An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern.

- **Clusters** in a graph suggest that the data describe several distinct kinds of individuals.



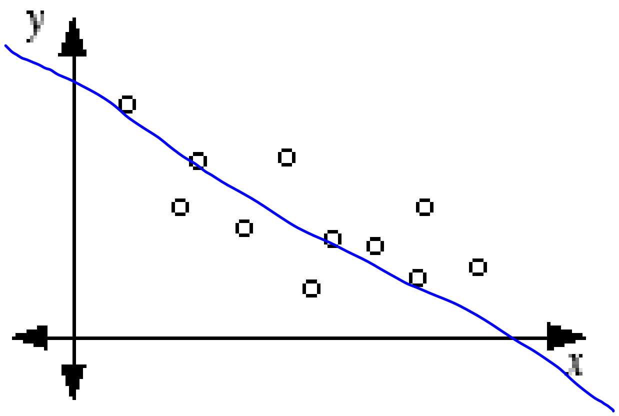
- Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other and below-average values also tend to occur together.

Positively Associated Data



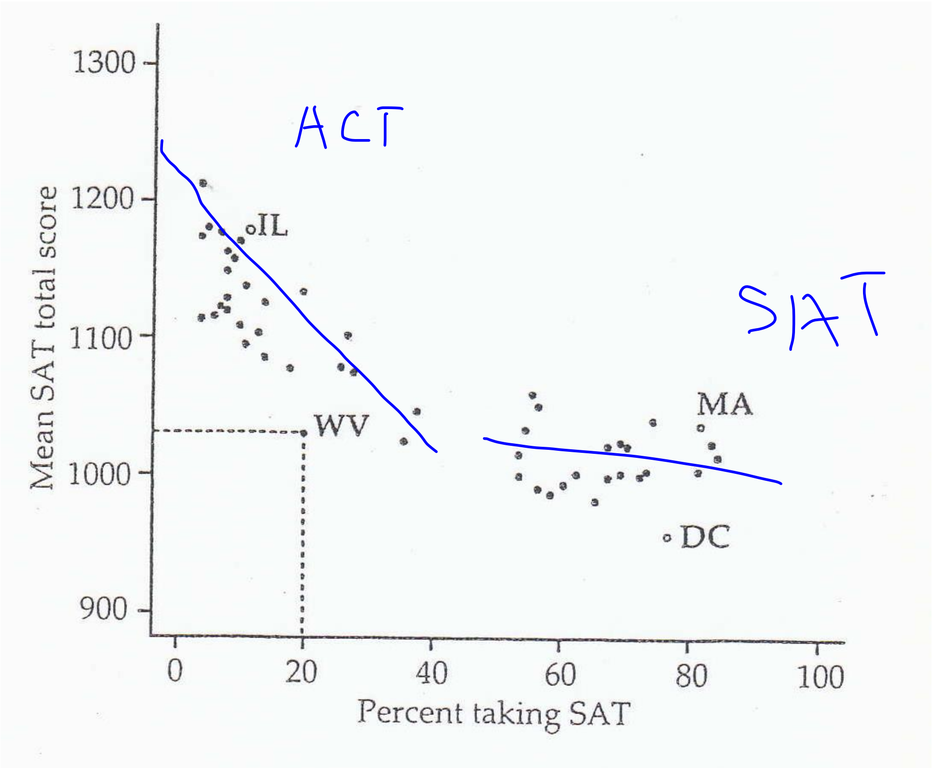
- Two variables are **negatively associated** when above-average values of one accompany below-average values of the other, and vice versa.

Negatively Associated Data



- The **strength** of a relationship in a scatterplot is determined by how closely the points follow a clear form.

For the example above (Interpretation):



ACT : negative moderately strong association

SAT : negative weak association

StatCrunch -> Graphics - > Scatter Plot

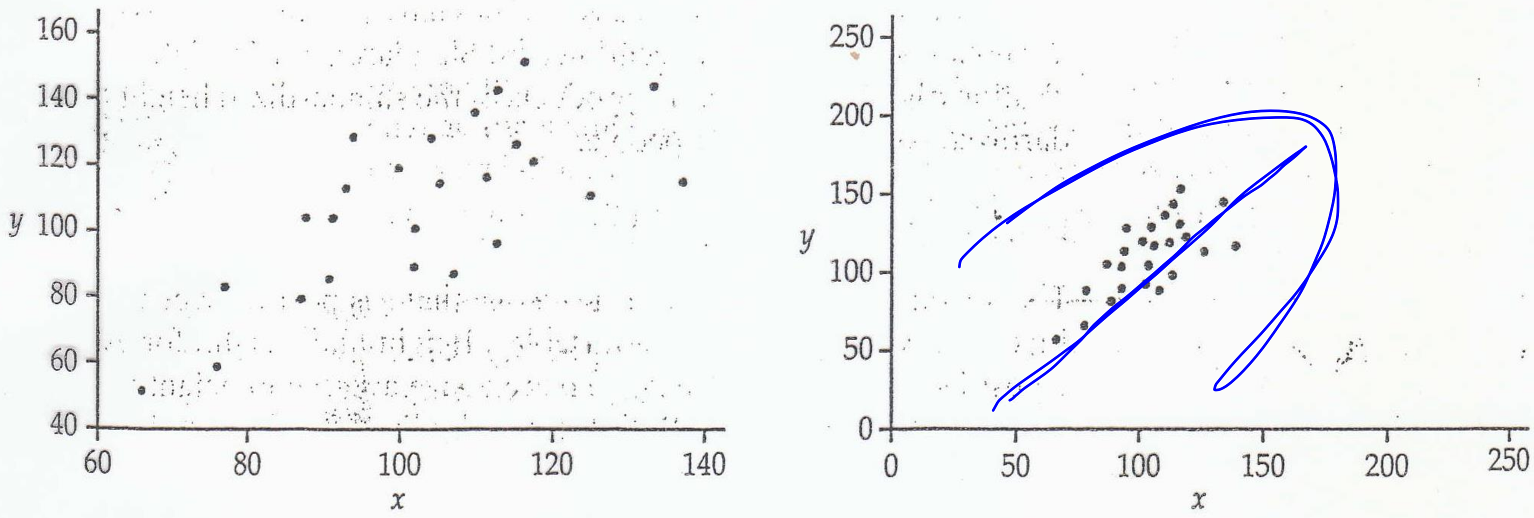
Correlation

We say a linear relationship is

- strong if the points lie close to a straight line, and

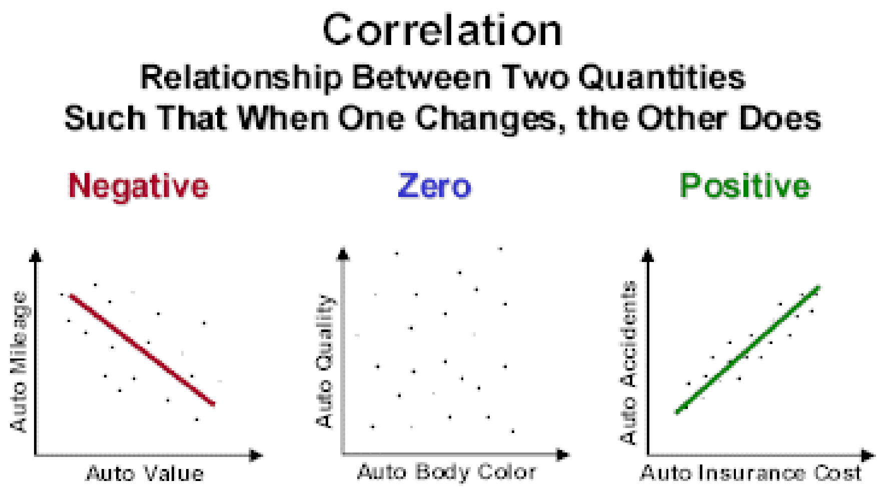
- weak if they are widely scattered about a line.

Sometimes graphs might be misleading:



Two scatterplots of the same data. The linear pattern in the plot on the right appears stronger because of the surrounding space.

We use *correlation* to measure the relationship.



<u>Definition</u>: The **correlation** measures the direction and strength of the linear relationship between two quantitative variables. Correlation is usually written as ***r***.

Suppose that we have data on variables *x* and *y* for *n* individuals. The means and standard deviations of the two variables are $\bar{x}$ and s_x for the *x*-values, and $\bar{y}$ and s_y for the *y*-values. The correlation *r* between *x* and *y* is

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2(y_i-\bar{y})^2}}$$

no units of measurement, just a number

standardized x and y values

Properties of Correlation:

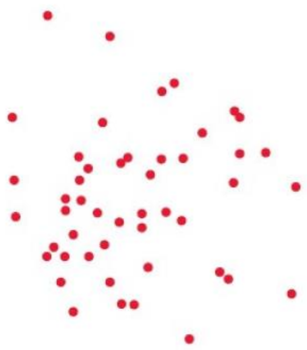
- Correlation does not distinguish between explanatory and response variables.
- Correlation requires that both variables be quantitative.
- Because *r* uses the standardized values of the observations, it does not change when we change the units of measurement of *x*, *y*, or both. The correlation itself has no unit of measurement; it is just a number.
- Positive *r* indicates positive association between the variables, and negative *r* indicates negative association.
- The correlation *r* is always a number between -1 and 1.

  -1 0 1

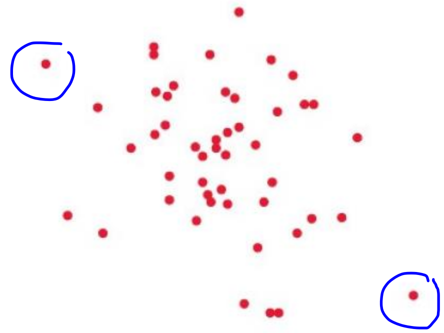
  perfect negative no relationship perfect positive

  - Values of *r* near 0 indicate a very weak linear relationship.
  - The strength of the relationship increases as *r* moves away from 0 toward either -1 or 1.
  - The extreme values *r* = -1 and *r* = 1 occur only when the points in a scatterplot lie exactly along a straight line.
- Correlation measures the strength of only the linear relationship between two variables
- Like the mean and standard deviation, the correlation is not resistant: *r* is strongly affected by a few outlying observations. Use *r* with caution when outliers appear in the scatterplot.

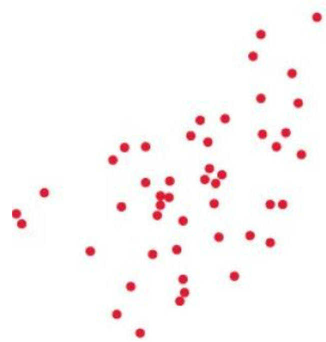
Here is how correlation *r* measures the direction and strength of a linear association:



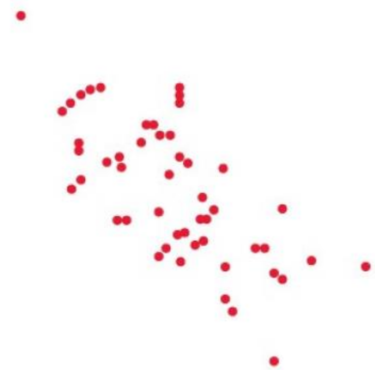
Correlation $r = 0$



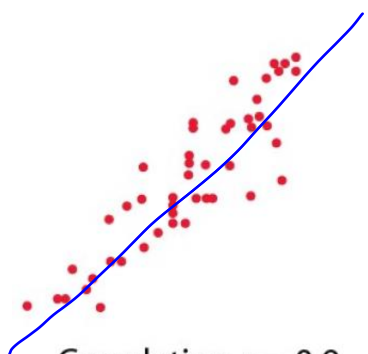
Correlation $r = -0.3$



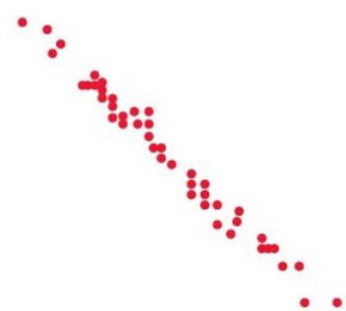
Correlation $r = 0.5$



Correlation $r = -0.7$



Correlation $r = 0.9$



Correlation $r = -0.99$

Correlation does not prove causation!

Examples:

1. There is a high correlation between number of sodas sold in one year and number of divorces, years 1950- 2010. Does that mean that having more sodas makes you more likely to divorce?

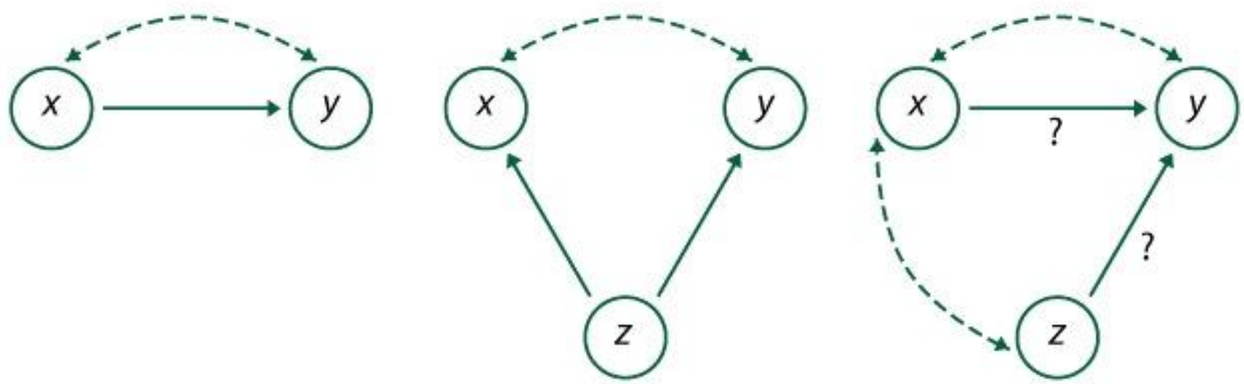
2. There is also a high correlation between number of teachers and number of bars for cities in California. So teaching drives you to drink?

3. What about the high correlation between amount of daily walking and quality of health for men aged over 65?

- In many studies of the relationship between two variables the goal is to establish that changes in the explanatory variable **cause** changes in response variable.

- Even a strong association between two variables, does not necessarily imply a causal link between the variables.

Some explanations for an observed association.

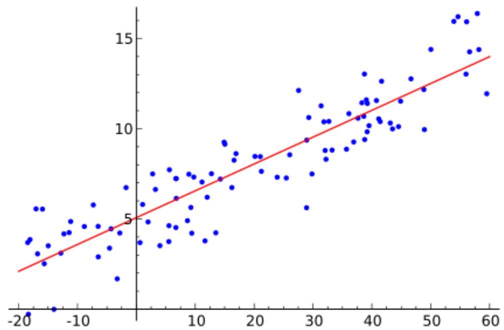


Causation (a) | Common response (b) | Confounding (c)

lurking but included in the study

The dashed double arrow lines show an association. The solid arrows show a cause and effect link. The variable x is explanatory, y is response and z is a **lurking variable**.

Least-Squares Regression



A *regression line* summarizes the relationship between two variables, but only in a specific setting:

- when one of the variables helps explain or predict the other.

Definition:

- A **regression line** is a straight line $\boxed{y = b_0 + b_1 x}$ 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.

- Regression, unlike correlation, requires that we have an explanatory variable and a response variable.

Example: Does fidgeting keep you slim?

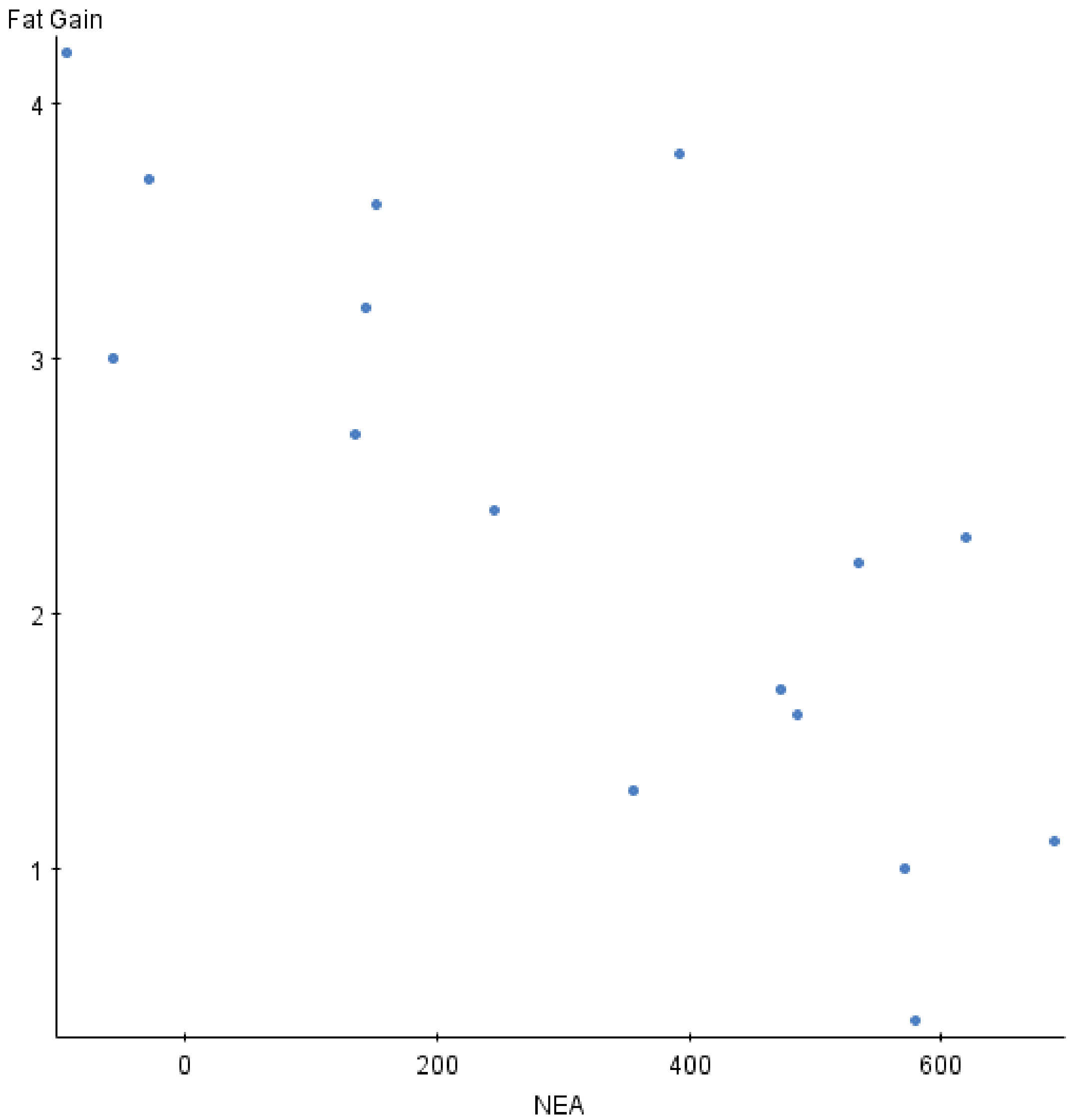
Some people don't gain weight even when they overeat. Perhaps fidgeting and other «nonexercise activity» (NEA) explains why – the body might spontaneously increase nonexercise activity when fed more. Researchers deliberately overfed 16 healthy young adults for 8 weeks. They measured fat gain (in kilograms) and, as an explanatory variable, increase in energy use (in calories) from activity other than deliberate exercise – fidgeting, daily living, and the like. Here are the data:

x

y

NEA increase (cal)	-94	-57	-29	135	143	151	245	355
Fat gain (kg)	4.2	3.0	3.7	2.7	3.2	3.6	2.4	1.3
NEA increase (cal)	392	473	486	535	571	580	620	690
Fat gain (kg)	3.8	1.7	1.6	2.2	1.0	0.4	2.3	1.1

Figure below is a scatterplot of these data. The plot shows a moderately strong negative linear association with no outliers. The correlation is $\boxed{r = -0.7786.}$



People with larger increase in NEA gain less fat

What does it mean «fitting a line to data»?

- It means drawing a line that comes as close as possible to the points representing our data.

Definition: 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

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

- b_1 is the **slope**, the amount by which y changes when x increases by one unit.
- b_0 is the **intercept**, the value of y when $x = 0$.

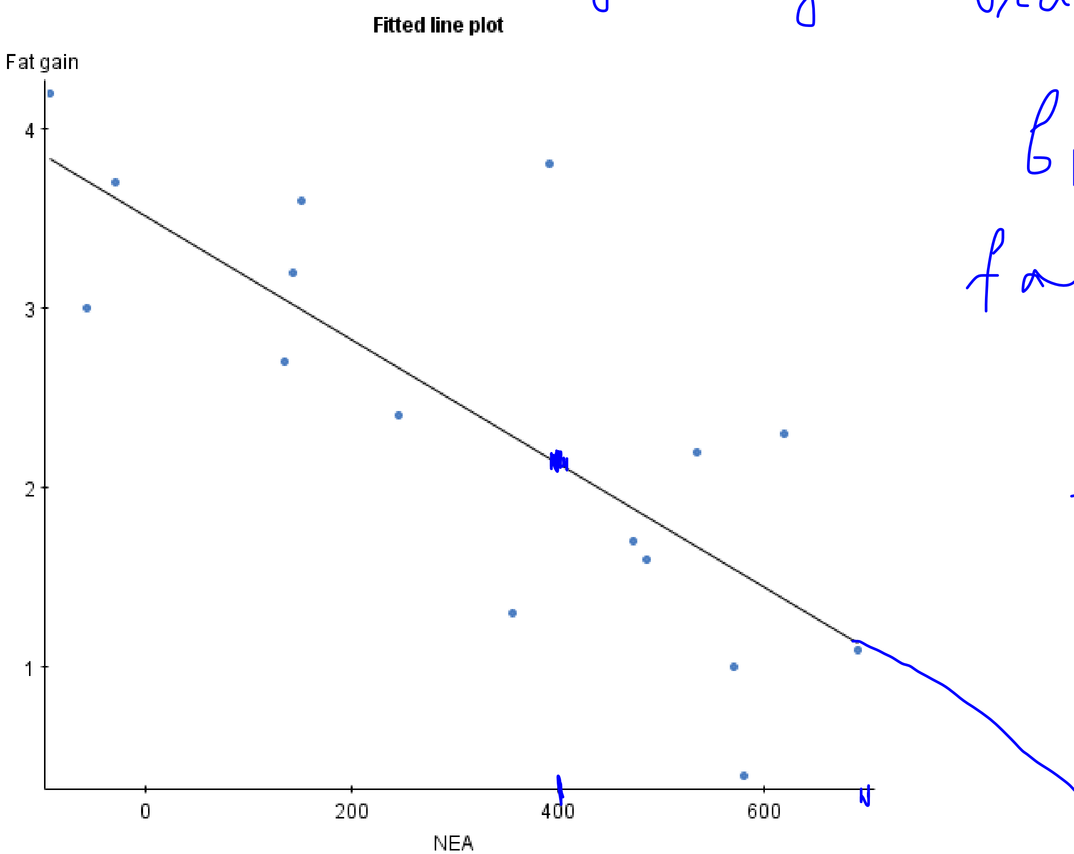
Example: Regression line for fat gain.

In figure below we have drawn the regression line with the equation

$y = b_0 + b_1 X$

$$\text{fat gain} = 3.505 - 0.00344 \times \text{NEA increase}$$

kg kg kg/cal cal



$b_1 = -0.00344$ kg/cal

fat gain goes down by 0.00344 kg for each added cal of NEA.

So b_1 is the rate of change in y when X changes

$b_0 = 3.505$ kg $\longrightarrow$ fat gain if NEA does not change

We can use a regression line to **predict** the response *y* for a specific value of the explanatory variable *x*.

Example: Say, we want to predict the fat gain for an individual whose NEA increases by 400 calories when she overeats. $x_0 = 400$

$$\text{fat gain} = 3.505 - 0.00344 \times \text{NEA increase}$$

$$\text{predicted fat gain} = 3.505 - 0.00344 \cdot 400$$
$$= 2.13 \text{ kg}$$

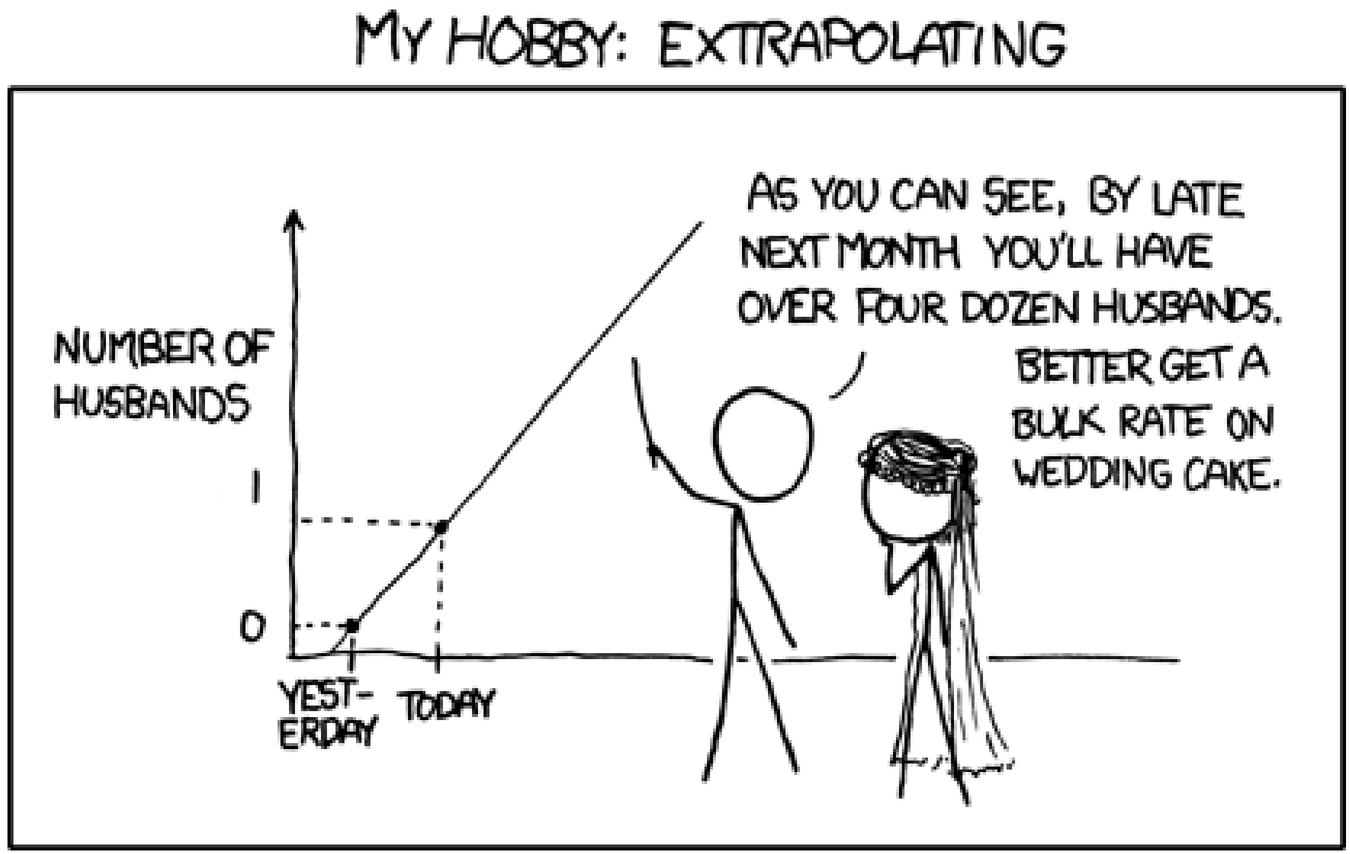
Is this prediction reasonable? Can we predict the fat gain for someone whose nonexercise activity increases by 1500 calories when she overeats?

$x_0 = 1500$

$$\text{predicted fat gain} = 3.505 - 0.00344 \cdot 1500$$
$$= -1.66 \text{ kg}$$

$\Rightarrow$ loses fat!

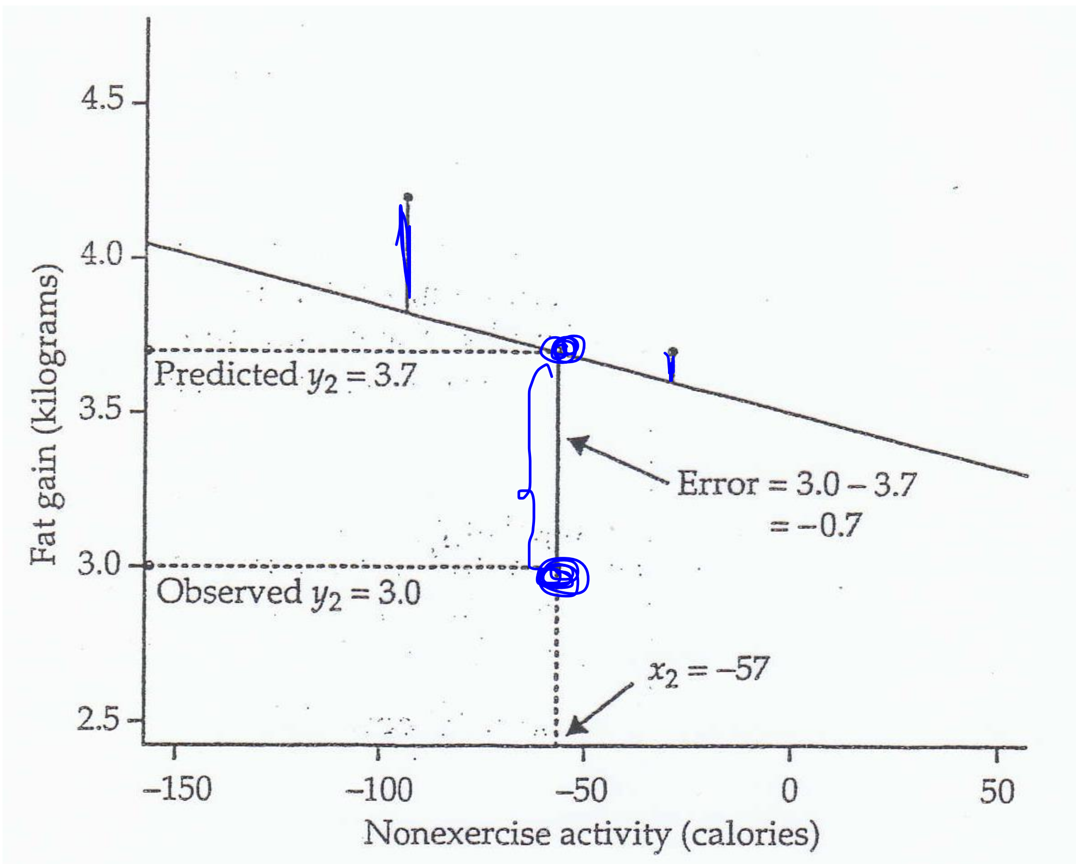
Definition: **Extrapolation** is the use of a regression line for prediction far outside the range of values of the explanatory variable *x* used to obtain the line. Such predictions are often not accurate.



How do we get the regression line?

NEA increase (cal) Fat gain (kg)	-94 4.2	-57 3.0	-29 3.7	135 2.7	143 3.2	151 3.6	245 2.4	355 1.3
NEA increase (cal) Fat gain (kg)	392 3.8	473 1.7	486 1.6	535 2.2	571 1.0	580 0.4	620 2.3	690 1.1

$$\text{fat gain} = 3.505 - 0.00344 \times \text{NEA increase}$$



$x_2 = -57, \quad y_2 = 3.0$

predicted $y_2 =$

$= 3.505 - 0.00344 \cdot (-57)$

$= 3.7$

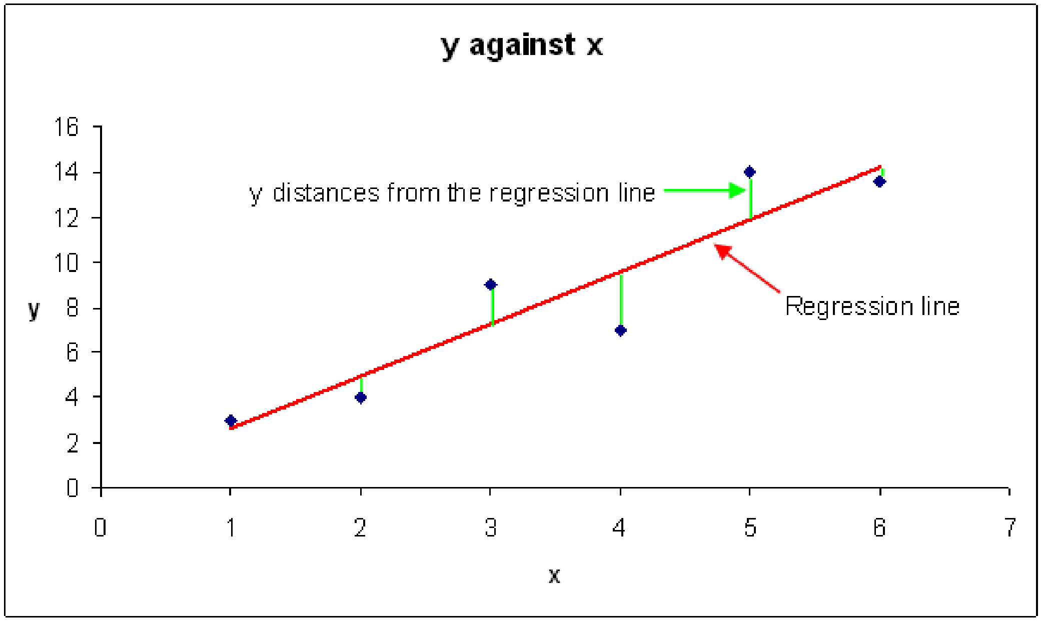
$3.0 - 3.7 = -0.7$

→ error

error = observed − predicted

Goal: minimize errors

Definition: 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.



$(x_1, y_1) \ldots (x_n, y_n)$

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

$\hat{y}_i = b_0 + b_1 x_i$

$\hat{y}_i$ is predicted value

y_i is observed value

error $= y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$

Idea: Find b_0, b_1 that minimize

$$\sum (\text{errors})^2 = \sum (y_i - (b_0 + b_1 x_i))^2$$

Equation of the Least-Squares Regression Line:

- We have data on an explanatory variable x and a response variable y for n individuals.

- The means and standard deviations of the sample data are $\bar{x}$ and s_x for x and $\bar{y}$ and s_y for y, and the correlation between x and y is r.

- The equation of the least-squares regression line of y on x is

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

with slope

$$\boxed{b_1 = r\frac{s_y}{s_x}}$$

and intercept

$$\boxed{b_0 = \bar{y} - b_1\bar{x}}$$

Example: Let's check the calculations for our example.

Using software we get

Summary statistics:

	Column	n	Mean	Std. Dev.
x	NEA	16	324.75	257.65674
y	Fat gain	16	2.3875	1.1389322

$\bar{x} = 324.75$ $\quad s_x = 257.67$

$\bar{y} = 2.3875$ $\quad s_y = 1.14$

$r = -0.7786$

$$b_1 = r\frac{s_y}{s_x} = -0.7786 \cdot \frac{1.14}{257.67} = 0.00344$$

$$b_0 = \bar{y} - b_1\bar{x} = 2.3875 + 0.00344 \cdot 324.75$$

$$= 3.505$$

$$\boxed{\hat{y} = 3.505 - 0.00344x}$$

StatCrunch -> Stat -> Regression -> Simple Linear

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 $= r$
R-sq = 0.6061492
Estimate of error standard deviation: 0.73985285 $= S_e$

Parameter estimates:

	Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Value
b_0	Intercept	3.505123	0.3036164	≠ 0	14	11.544577	<0.0001
b_1	Slope	-0.003441487	7.414096E-4	≠ 0	14	-4.641816	0.0004

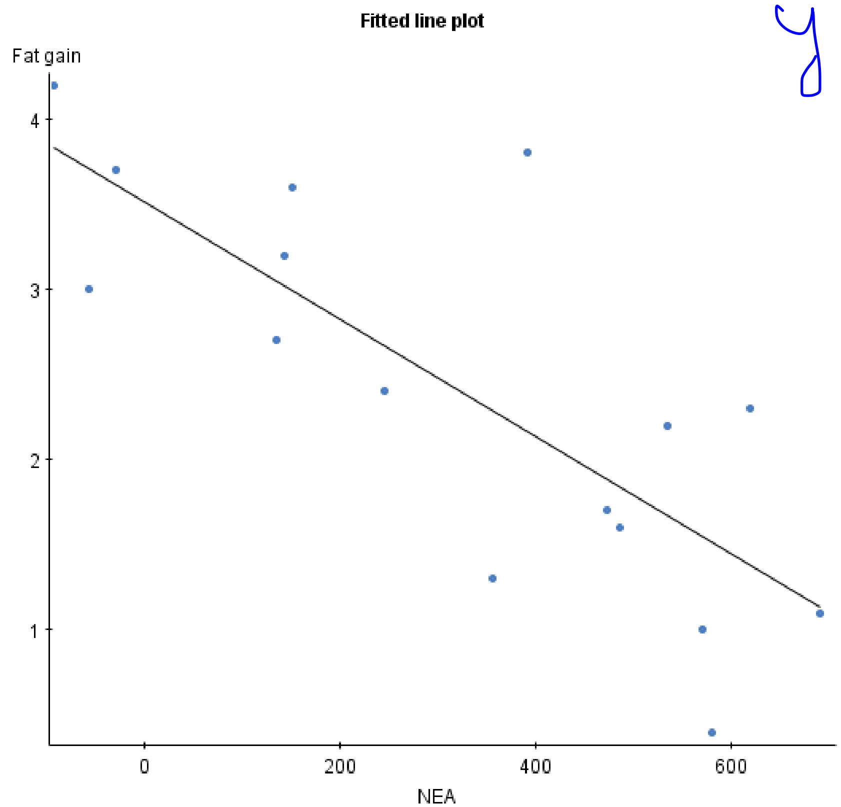
Predicted values:

X value	Pred. Y	s.e.(Pred. y)	95% C.I. for mean	95% P.I. for new
400	2.128528	0.1931943	(1.7141676, 2.5428886)	(0.48849356, 3.7685626)

Coefficient of determination (R^2):

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*.
In the above example: R-sq = 0.6061492 = appr. 61%

$$y - \bar{y} = \underbrace{y - \hat{y}}_{\underbrace{39\%}_{0}} + \underbrace{\hat{y} - \bar{y}}_{\underbrace{61\%}_{100}}$$

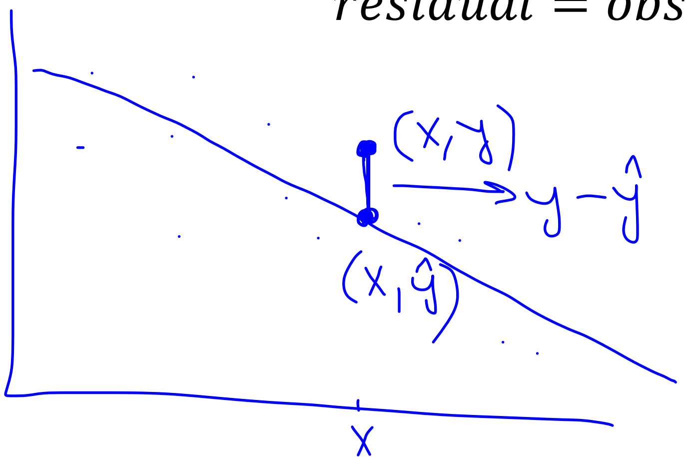


i.e. 61% of the variation in fat gain is explained by the regression. Other 39% is the vertical scatter in the observed responses remaining after the line has fixed the predicted responses.

Residuals

<u>Definition</u>: A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

$$residual = observed\ y - predicted\ y = y - \hat{y}$$



Note: If residual > 0, then $y - \hat{y} > 0$, $y > \hat{y}$

So the model underestimates

If residual < 0, then $y < \hat{y}$

So the model overestimates

For our example: fat gain = 3.505 – 0.00344 × NEA increase

NEA increase (cal) Fat gain (kg)	-94 4.2	-57 3.0	-29 3.7	135 2.7	143 3.2	151 3.6	245 2.4	355 1.3
NEA increase (cal) Fat gain (kg)	392 3.8	473 1.7	486 1.6	535 2.2	571 1.0	580 0.4	620 2.3	690 1.1

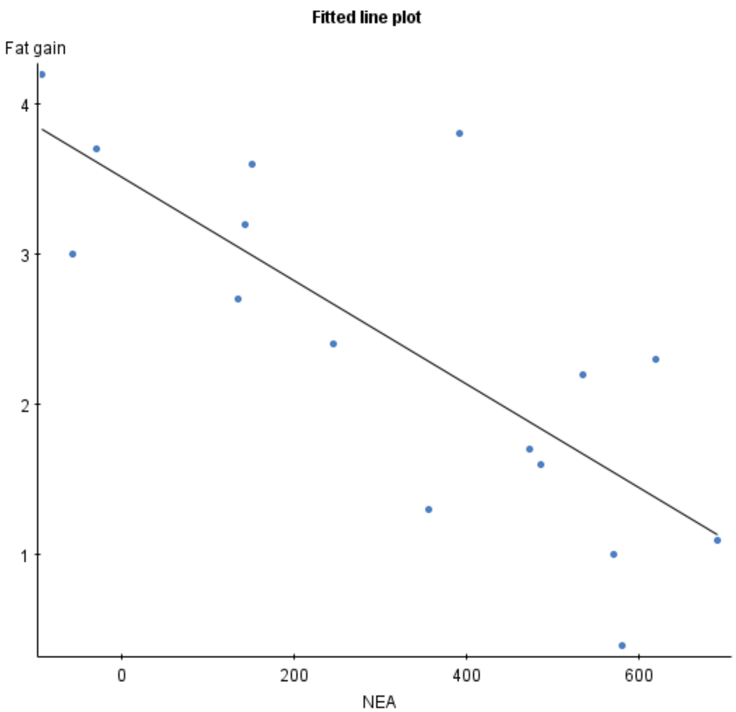
$(x_4, y_4) = (135, \underline{2.7})$

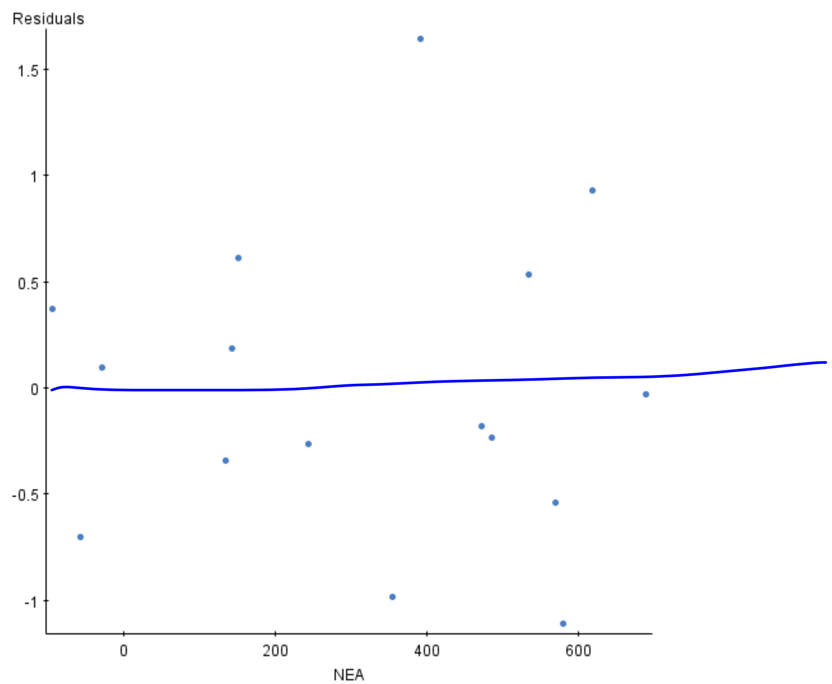
$\hat{y} = 3.505 - 0.00344 \cdot 135 = 3.04$

$residual = 2.7 - 3.04 = -0.34 < 0$

our model overestimates

Residual Plots





(a) (b)

(a) Scatterplot of fat gain versus increase in NEA

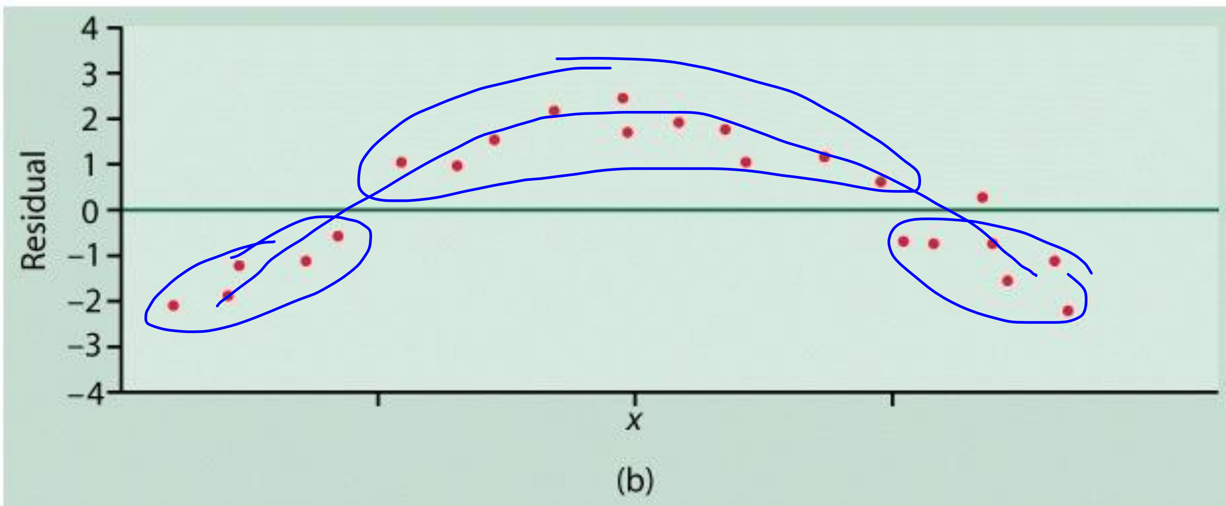
(b) Residual plot for this regression.

Because the residuals show how far the data fall from our regression line, examining the residuals helps assess how well the line describes the data.

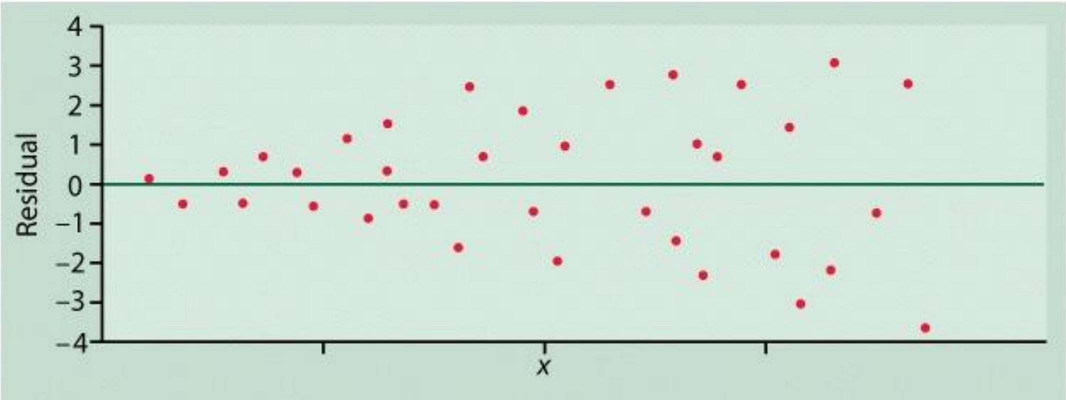
Definition: A **residual plot** is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess the model assumptions.

- If the regression line catches the overall pattern of the data, there should be no pattern in the residuals.

- On the other hand, curvature would suggest using higher order models or transformations.



- Also look for trends in dispersion, e.g. an increasing dispersion as the fitted values increase, in which case a transformation of the response may help (e.g. log or square root).

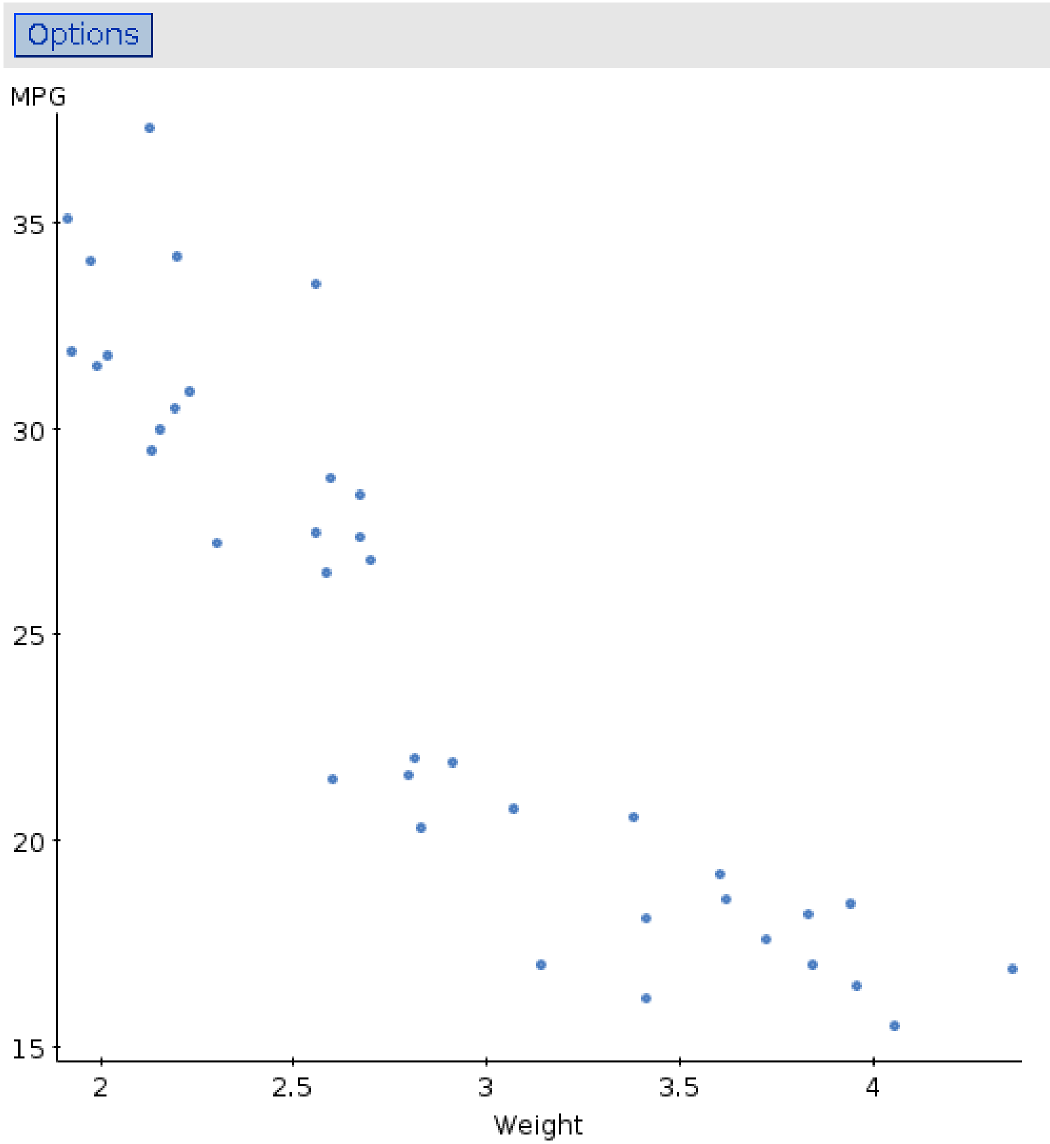


(c)

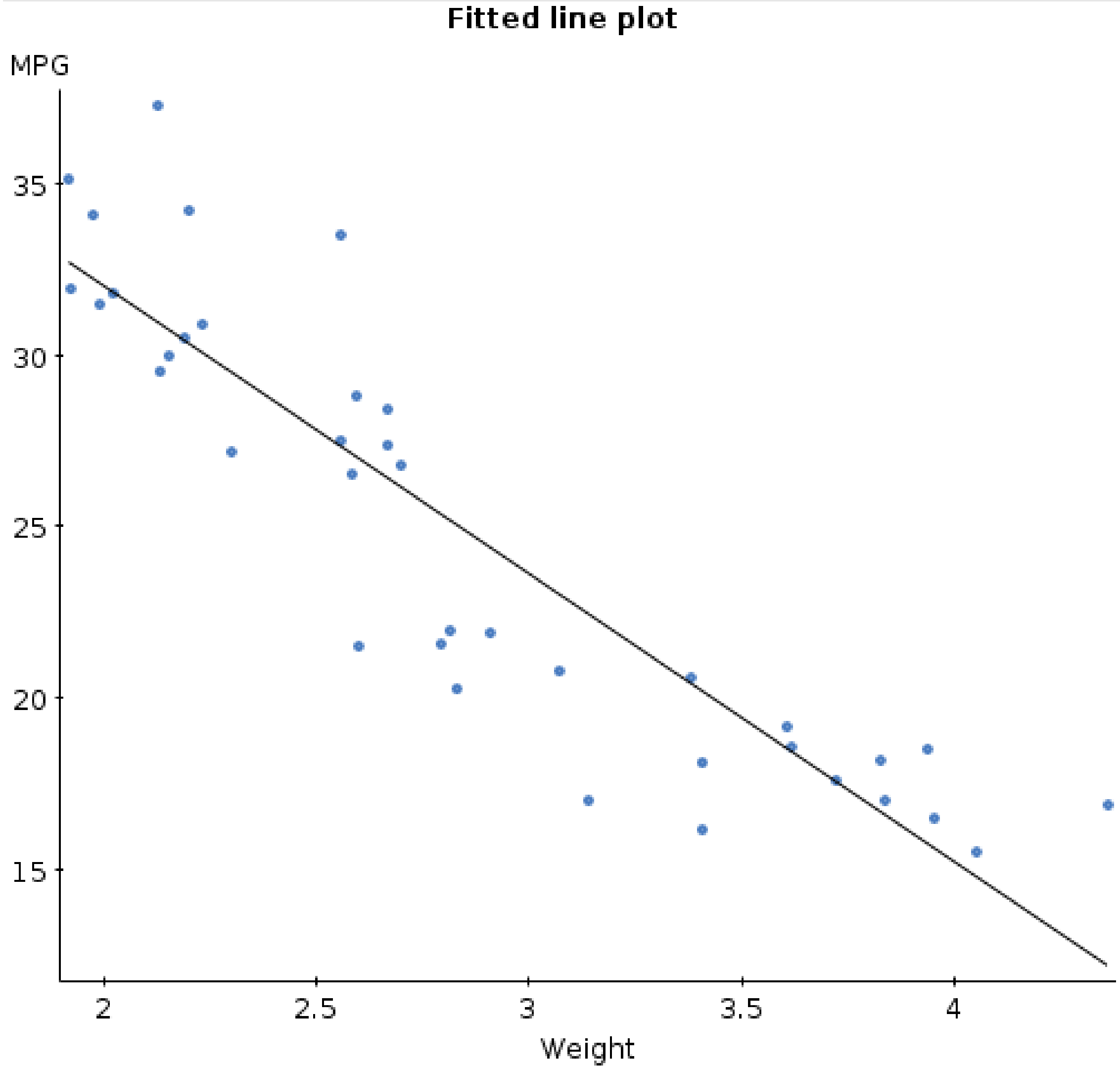
- No regression analysis is complete without a display of the residuals to check that the linear model is reasonable.

- Residuals often reveal things that might not be clear from a plot of the original data.

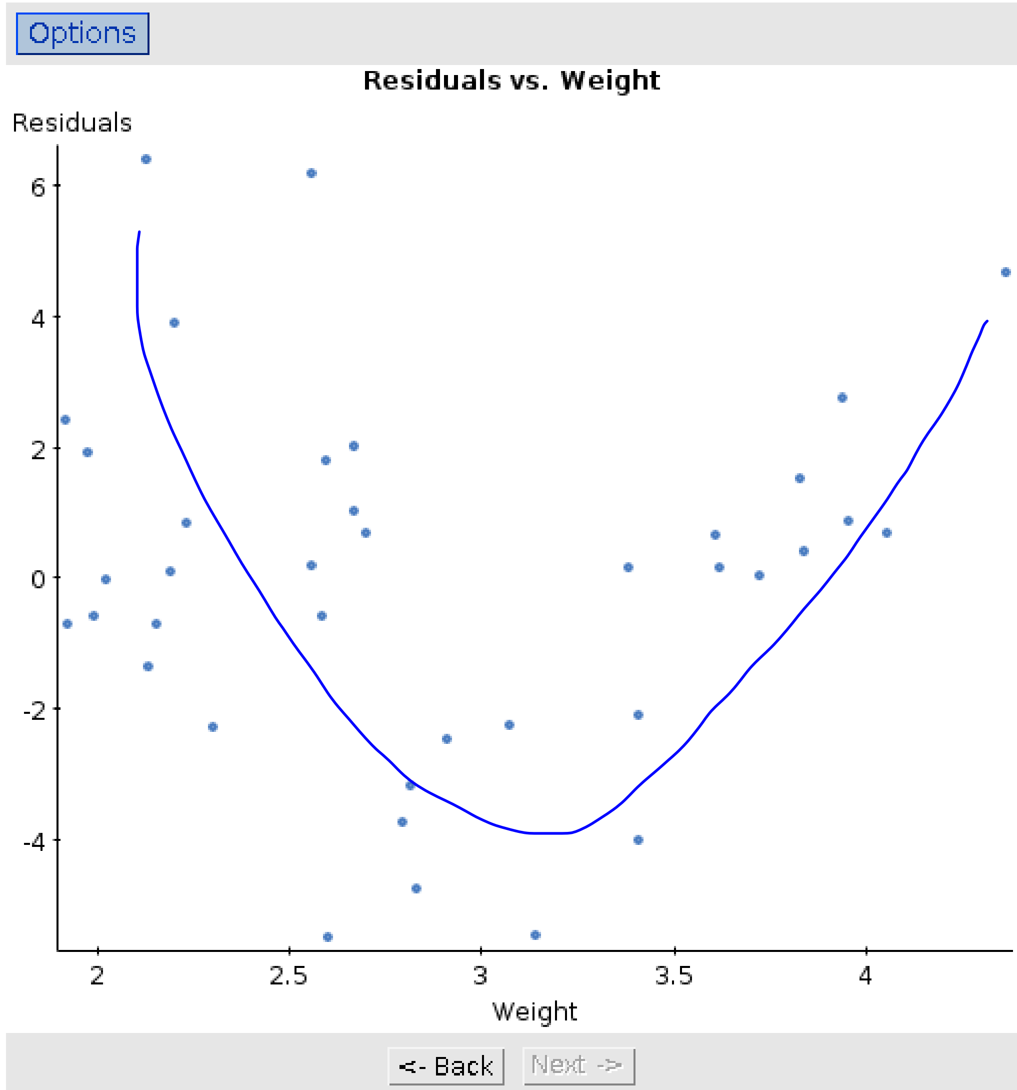
<u>Example</u>:



Options

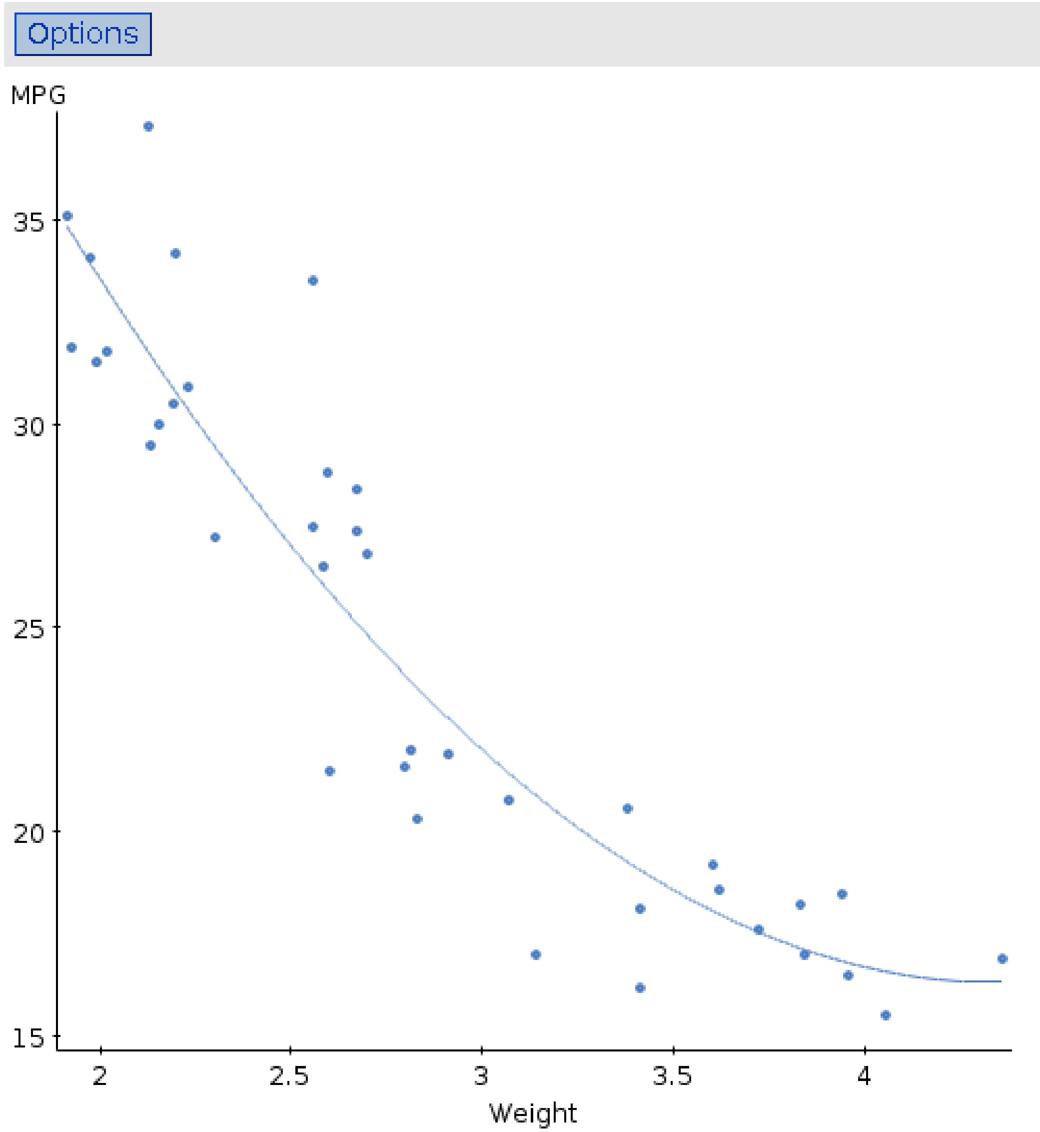


<- Back Next ->



- The residual plot doesn't look completely random, but a bit curved.

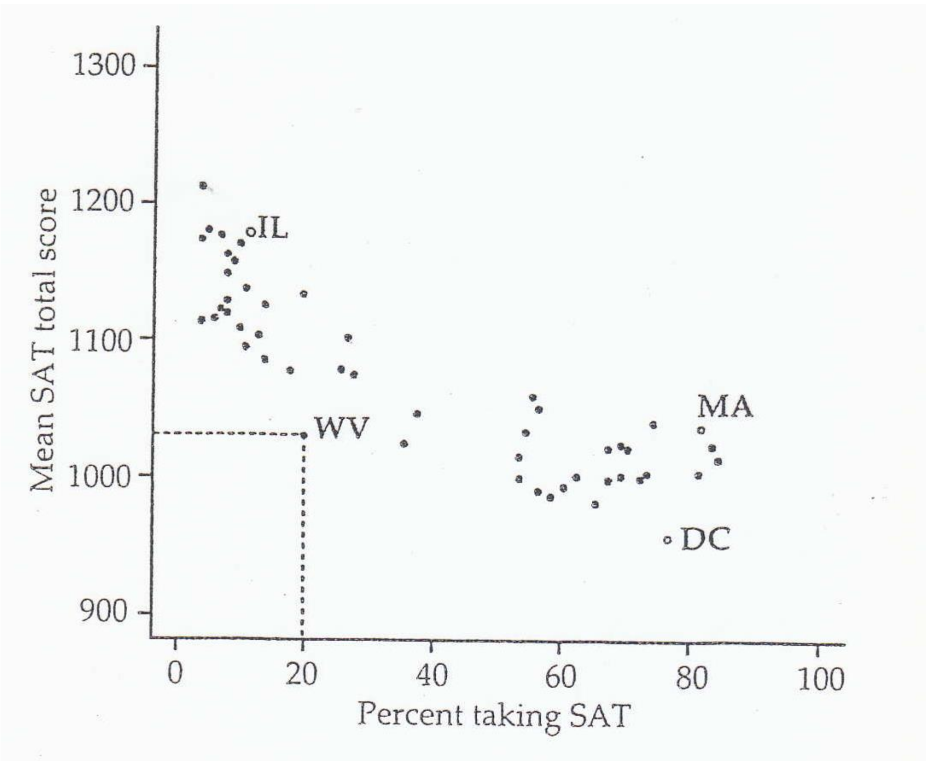
- Curve does seem to go through points better:



- Sometimes residuals reveal violations of the regression conditions that require our attention.

- An examination of residuals often leads us to discover groups of observations that are different from the rest.

- When we discover that there is more than one group in a regression, we may decide to analyze the groups separately, using a different model for each group.



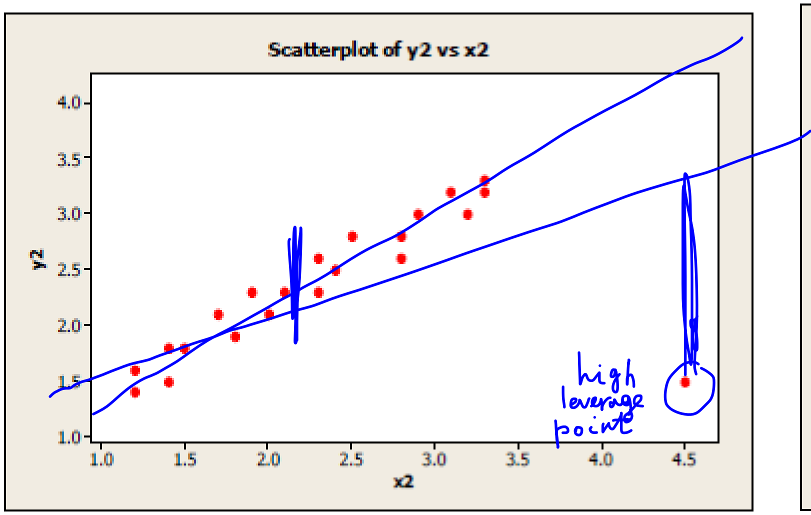
Outliers, Leverage, and Influential Observations

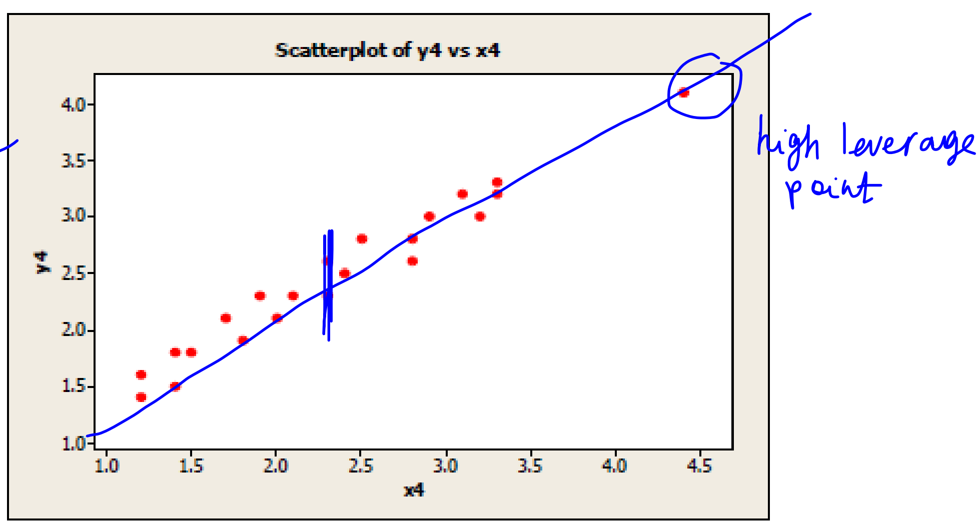
- Outliers: Any point that stands away from the others can be called an **outlier** and deserves your special attention.

- Outlying points can strongly influence a regression. Even a single point far from the body of the data can dominate the analysis.

- High Leverage Point: A data point that has an *x*-value far from the mean of the *x*-values is called a **high leverage point**.

Example:

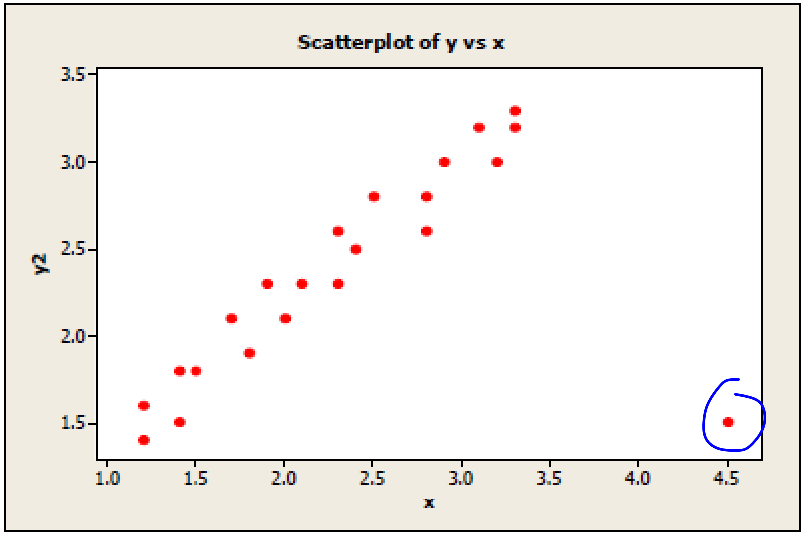


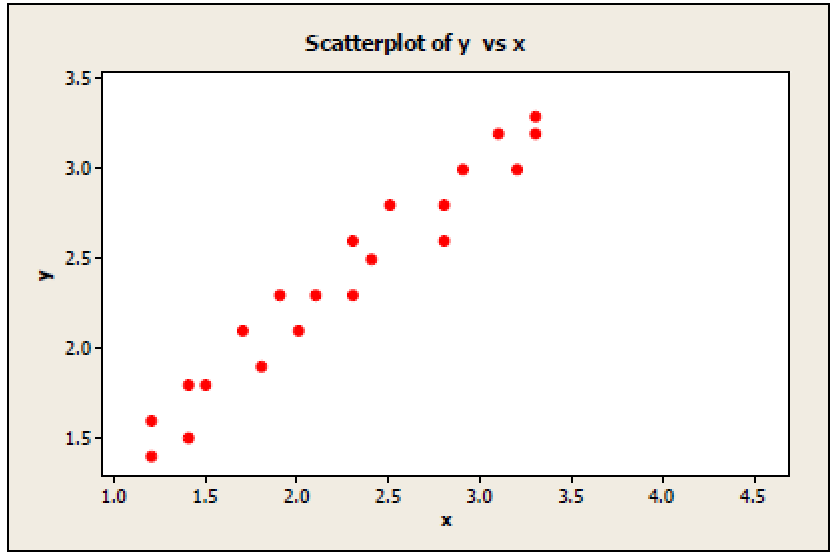


Are they influential?

<u>Influential observations</u>: A data point is influential if omitting it from the analysis gives a very different model.

<u>Example</u>:





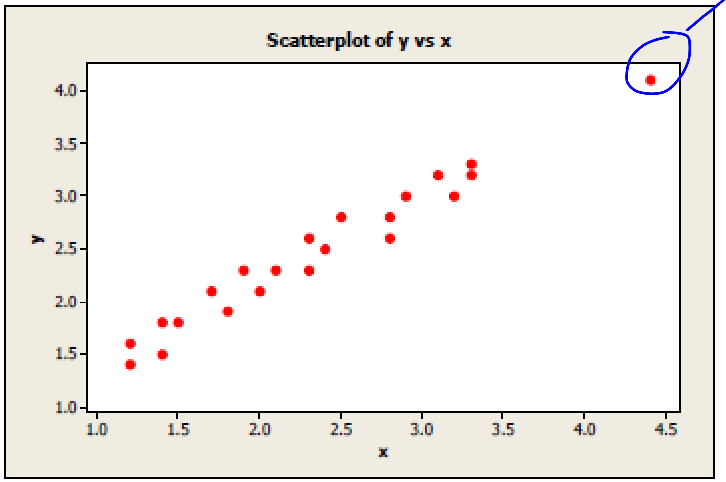

```
y = 1.38 + 0.414 x,
R-Sq = 33.2%
```

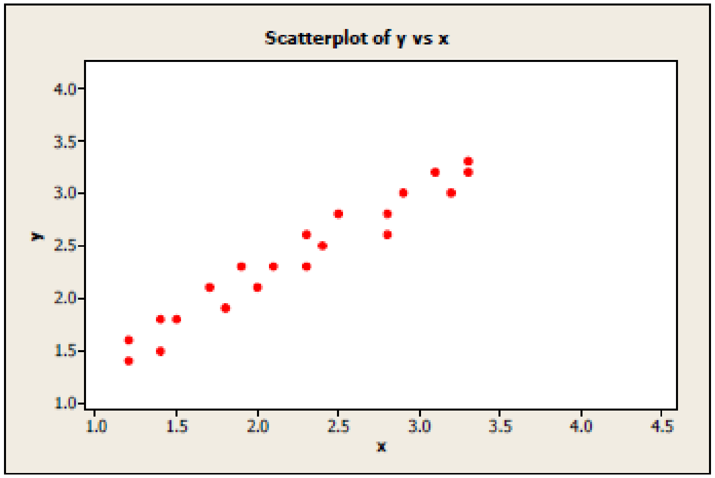
```
y = 0.567 + 0.811 x
R-Sq = 94.8%
```

<u>Note</u>: R^2 is much larger for the second plot.

Example: (A high leverage point that is not influential)

makes the relationship a bit stronger

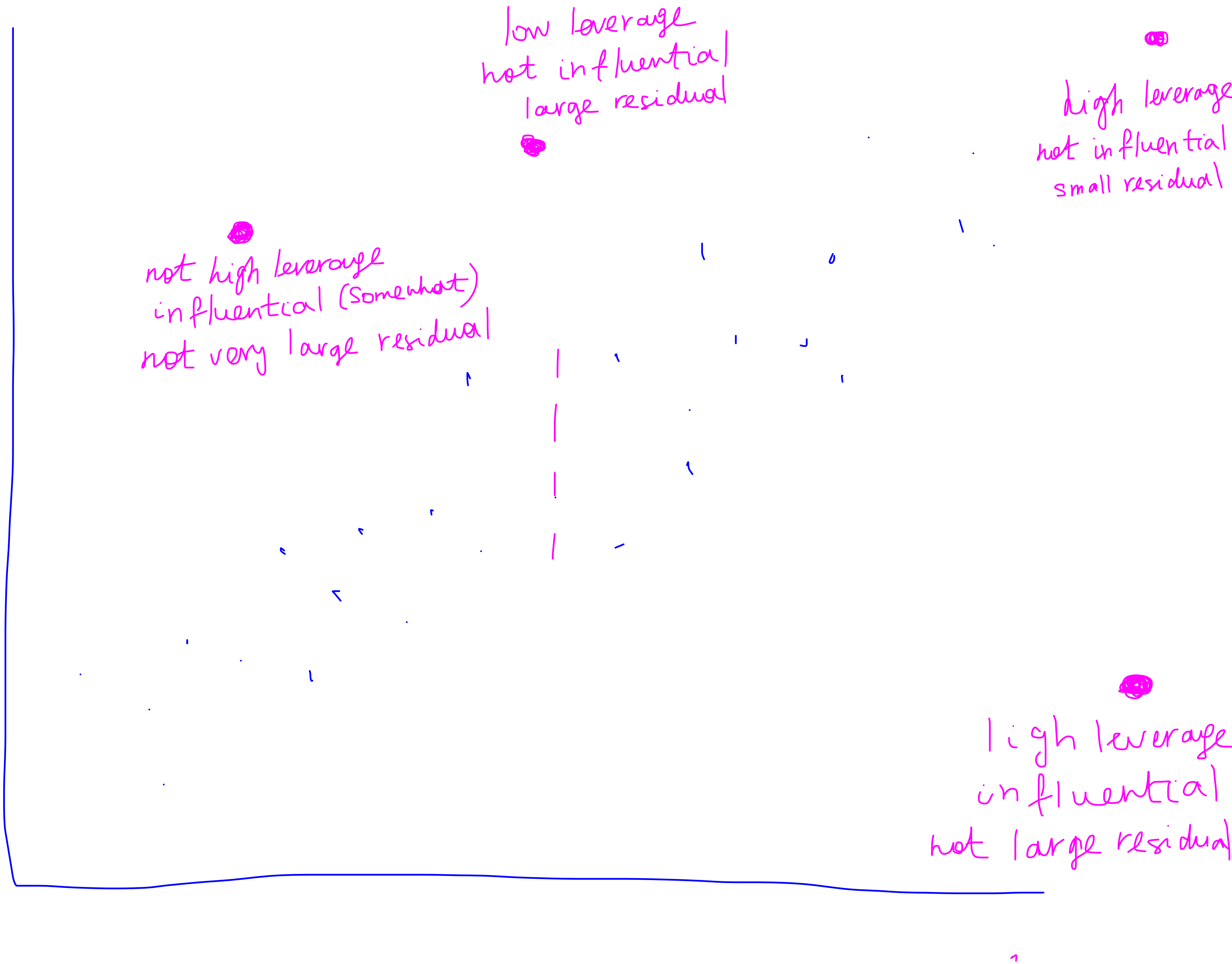




y = 0.577 + 0.806 x
R-Sq = 96.3%

y = 0.567 + 0.811 x
R-Sq = 94.8%

Note: R^2 is a bit less.


low leverage
not influential
large residual
high leverage
not influential
small residual
not high leverage
influential (somewhat)
not very large residual
high leverage
influential
not large residual

Example: People with diabetes must manage their blood sugar levels carefully. They measure their fasting plasma glucose (FPG) several times a day with a glucose meter. Another measurement, made at regular medical checkups, is called HbA. This is roughly the percent of red blood cells that have a glucose molecule attached. It measures average exposure to glucose over a period of several months.

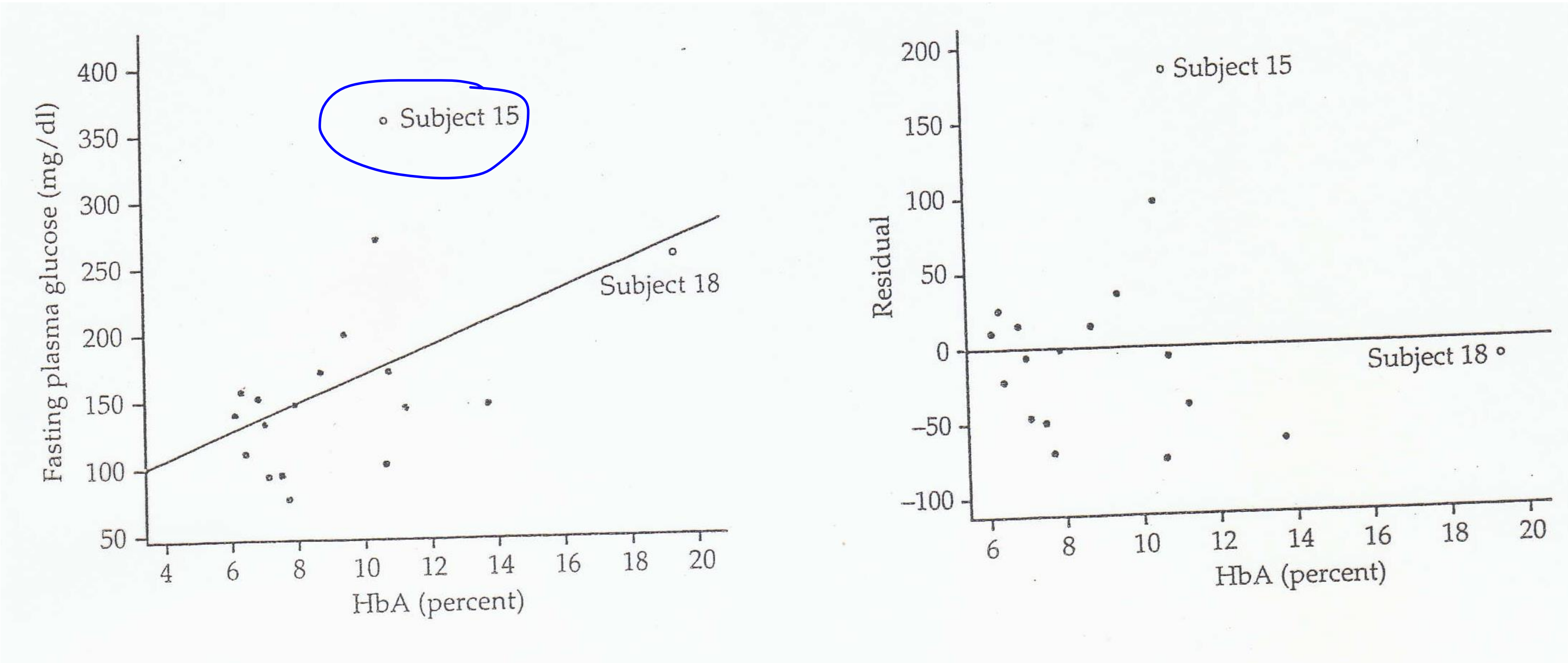
Table below gives data on both HbA and FPG for 18 diabetics five month after they had completed a diabetes education class.

Two measures of glucose level in diabetics

Subject	HbA (%)	FPG (mg/ml)	Subject	HbA (%)	FPG (mg/ml)	Subject	HbA (%)	FPG (mg/ml)
1	6.1	141	7	7.5	96	13	10.6	103
2	6.3	158	8	7.7	78	14	10.7	172
3	6.4	112	9	7.9	148	15	10.7	359
4	6.8	153	10	8.7	172	16	11.2	145
5	7.0	134	11	9.4	200	17	13.7	147
6	7.1	95	12	10.4	271	18	19.3	255

Because both FPG and HbA measure blood glucose, we expect a positive association.

The scatterplot in figure below shows a surprisingly weak relationship, with correlation $r = 0.4819$.

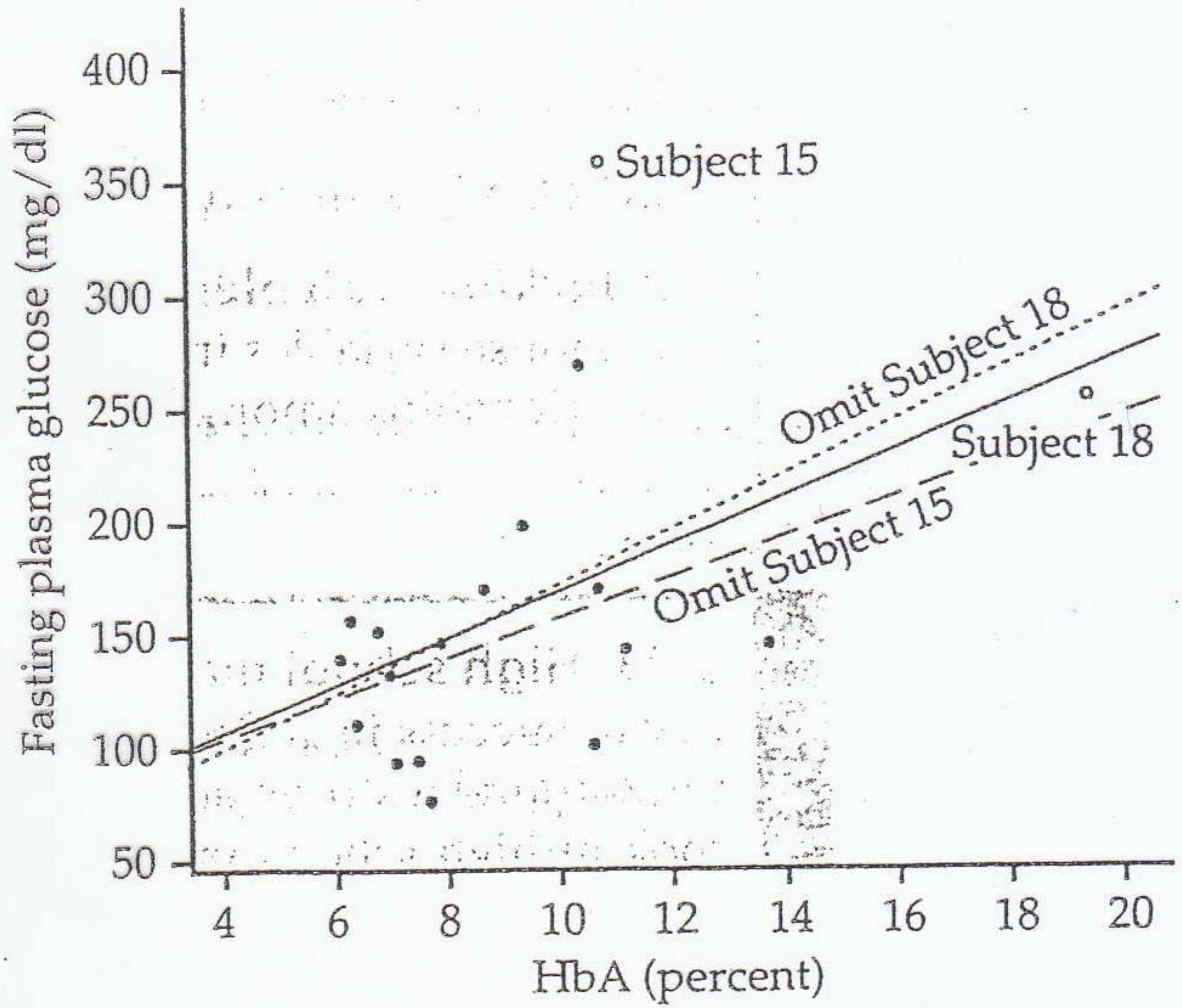


The line on the plot is the least-squares regression line for predicting FPG from HbA. Its equation is

$$\hat{y} = 66.4 + 10.41x$$

If we remove Subject 15, $r = 0.5684$.

If we remove Subject 18, $r = 0.3837$.


Fasting plasma glucose (mg/dl)
400
350
300
250
200
150
100
50
Subject 15
Omit Subject 18
Subject 18
Omit Subject 15
4
6
8
10
12
14
16
18
20
HbA (percent)

Doing regression:

- Start with a scatterplot
- If it does not look like a straight line relationship, stop (see Chapter 10).
- Otherwise, calculate correlation, intercept, and slope of regression line.
- Check whether regression is OK by looking at plot of residuals.
- If not OK, do not use regression.
- Aim: want regression for which line is OK, confirmed by looking at scatterplot and residual plot(s). Otherwise, cannot say anything useful.

Re-expressing data (transformations) – Get it Straight!

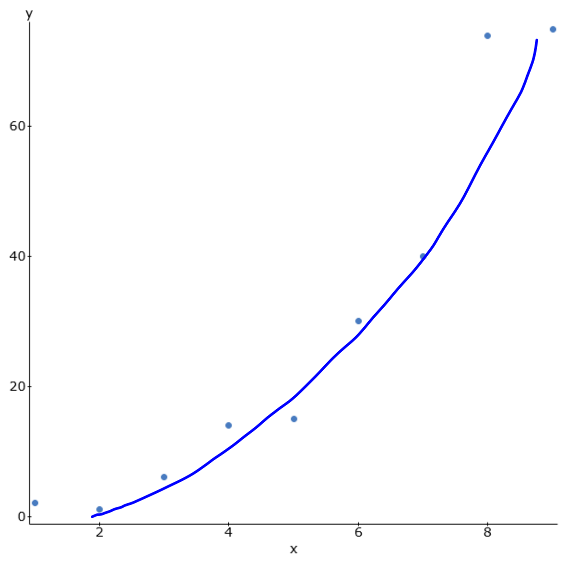
Take a simple function (a transformation) of the data to achieve:

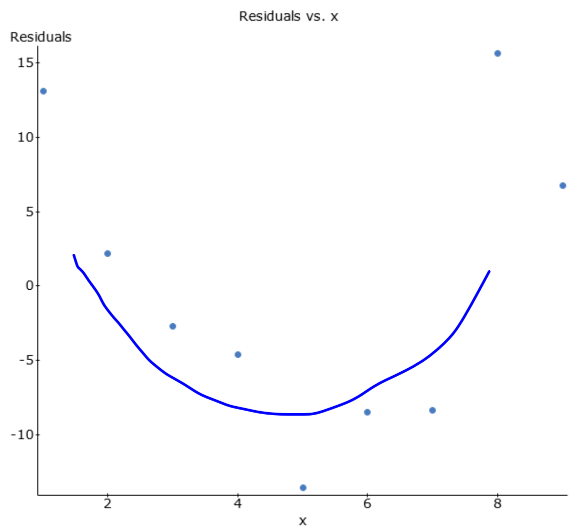
- make the distribution more symmetric

- make a scatterplot more linear

<u>Example</u>:

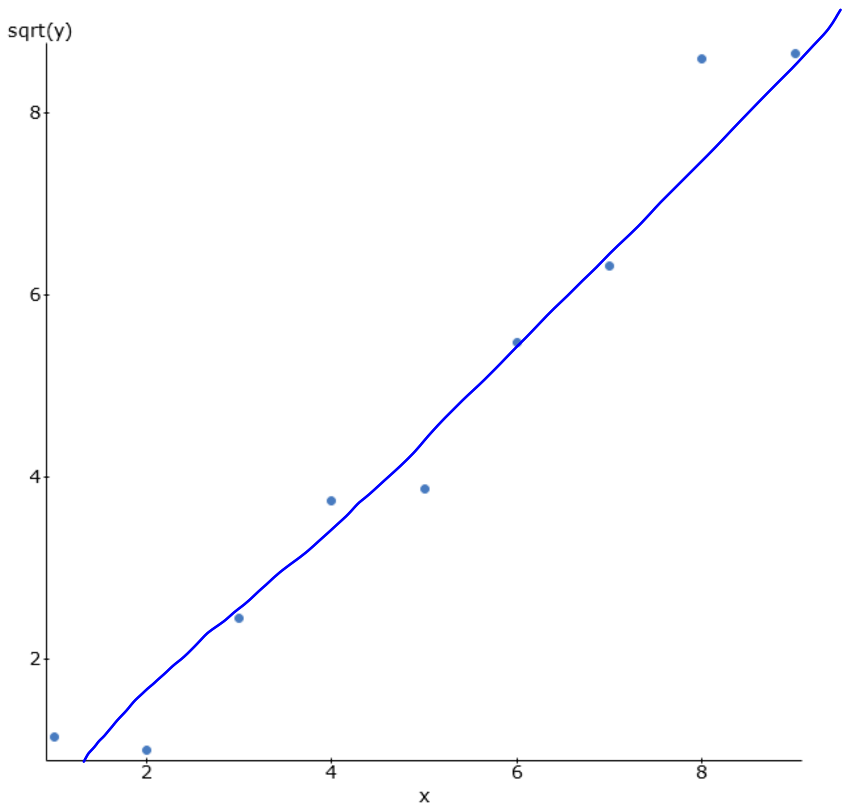
x	1	2	3	4	5	6	7	8	9
y	2	1	6	14	15	30	40	74	75

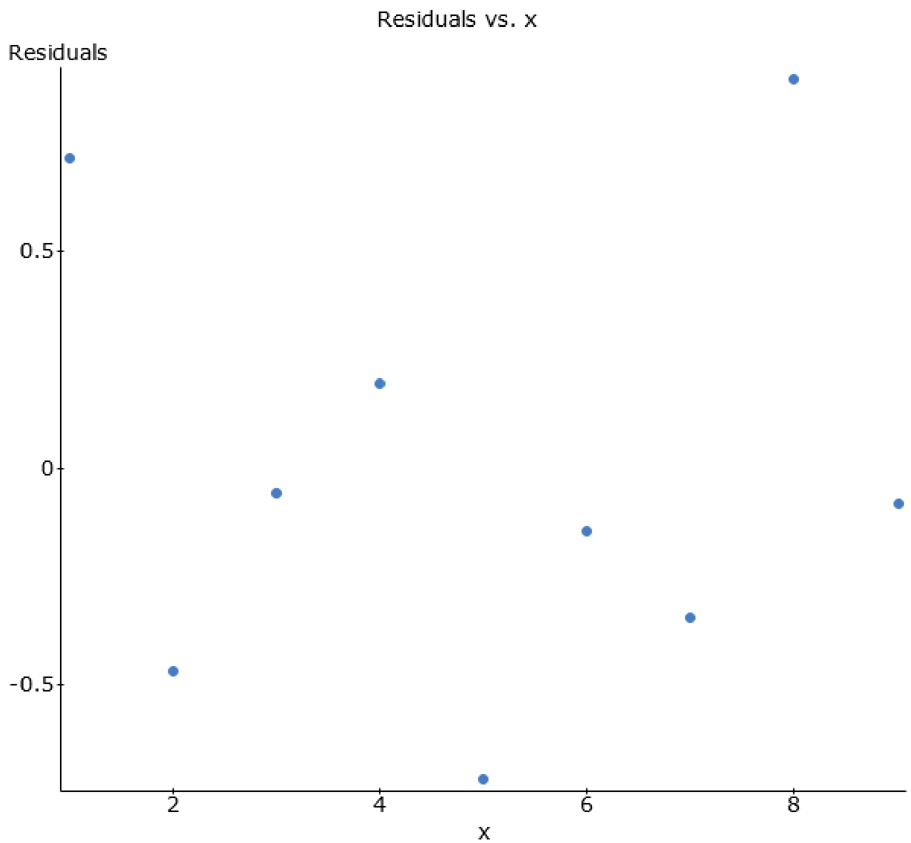




Take $\sqrt{y}$:

x	1	2	3	4	5	6	7	8	9
$\sqrt{y}$	1.14	1	2.45	3.74	3.87	5.48	6.32	8.6	8.66





Example:

Variable: potassium

0 002233333334444444444
0 555556666667789999999
1 00000001111111222233444
1 667799
2 034
2 68
3 23

Variable: log(potassium)

2 7
3 022224444
3 66666777788889
4 0001112244
4 55555666666677777777788889999
5 11112234
5 56688