I gather, young man, that you wish to be a Member of Parliament. The first lesson that you must learn is, when I call for statistics about the rate of infant mortality, what I want is proof that fewer babies died when I was Prime Minister than when anyone else was Prime Minister. That is a political statistic.

–Winston Churchill

Review of scatterplots and correlation.

Review of scatterplots and correlation.

For two observed quantities in the same data set, a scatterplot "plots them against each other"

Review of scatterplots and correlation.

For two observed quantities in the same data set, a scatterplot "plots them against each other" To measure correlation, we take the value r, which depends on the normalized variables (deviations from the mean)

Review of scatterplots and correlation.

For two observed quantities in the same data set, a scatterplot "plots them against each other" To measure correlation, we take the value r, which depends on the normalized variables (deviations from the mean) (and thus doesn't care about the units of measurement, for example).

Example 1. *We take the data from exercise 4.27. This gives, for 19 wealthy countries, the average number of liters of alcohol per year from wine, and the yearly rate of death from heart disease (per 100,000 people).*

Example 1. *We take the data from exercise 4.27. This gives, for 19 wealthy countries, the average number of liters of alcohol per year from wine, and the yearly rate of death from heart disease (per 100,000 people).*

INDIVIDUALS: *countries*

Example 1. *We take the data from exercise 4.27. This gives, for 19 wealthy countries, the average number of liters of alcohol per year from wine, and the yearly rate of death from heart disease (per 100,000 people).*

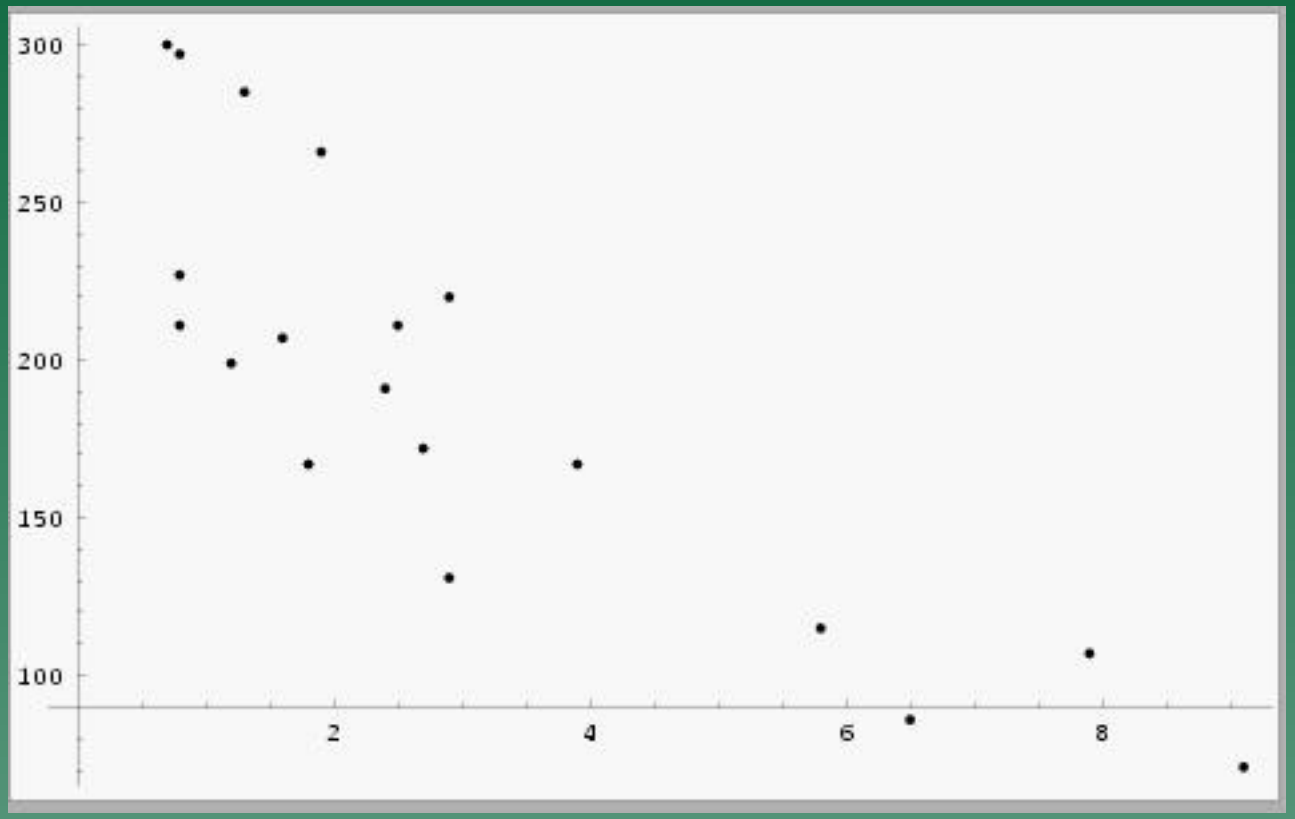
INDIVIDUALS: *countries*
EXPLANATORY VARIABLE:

Example 1. *We take the data from exercise 4.27. This gives, for 19 wealthy countries, the average number of liters of alcohol per year from wine, and the yearly rate of death from heart disease (per 100,000 people).*

INDIVIDUALS: *countries*
EXPLANATORY VARIABLE: *liters of alcohol per year from wine*

Example 1. *We take the data from exercise 4.27. This gives, for 19 wealthy countries, the average number of liters of alcohol per year from wine, and the yearly rate of death from heart disease (per 100,000 people).*

INDIVIDUALS: *countries*
EXPLANATORY VARIABLE: *liters of alcohol per year from wine*
RESPONSE VARIABLE: *deaths per 100,000 from heart attack per year.*


300
250
200
150
100
2
4
6
8

This seems to show a strong correlation. Does it look positive or negative?

This seems to show a strong correlation. Does it look positive or negative?

Do you think that there is some causality?

This seems to show a strong correlation. Does it look positive or negative?

Do you think that there is some causality?

The value of r is $-.843$ - how does that fit with what we see informally?

The least squares regression line

The least squares regression line

When we introduced the r-value, we talked about it as related to some line

The least squares regression line

When we introduced the r-value, we talked about it as related to some line (since for example if $r = \pm 1$, the data lies along some line). We will now study that line.

The least squares regression line

When we introduced the r-value, we talked about it as related to some line (since for example if $r = \pm 1$, the data lies along some line). We will now study that line.

Definition 2. *The* least square regression line *is decribed as follows:*

- *Put the explanatory variable x along the horizontal axis and the response variable y along the vertical axis.*

The least squares regression line

When we introduced the r-value, we talked about it as related to some line (since for example if $r = \pm 1$, the data lies along some line). We will now study that line.

Definition 2. *The* least square regression line *is decribed as follows:*

- *Put the explanatory variable x along the horizontal axis and the response variable y along the vertical axis. The least squares regression line is the line that minimizes the sum of the squares of the vertical distances from*

the points to the line.

the points to the line.

- *Numerically, take*

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

Take

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

Then the least squares regression line is

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

the points to the line.

- *Numerically, take*

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

Take

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

Then the least squares regression line is

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

In these formulas, $\overline{x}$ is the mean for the variable x, s_x is the standard deviation for the variable x.

In these formulas, $\overline{x}$ is the mean for the variable x, s_x is the standard deviation for the variable x.

We use $\hat{y}$ instead of y to remind ourselves that the line will only give a prediction for the response variable, not its actual value.

In these formulas, $\overline{x}$ is the mean for the variable x, s_x is the standard deviation for the variable x.

We use $\hat{y}$ instead of y to remind ourselves that the line will only give a prediction for the response variable, not its actual value.

You can derive the equation of the regression line using calculus, as is done sometimes in Math 242.

In these formulas, $\overline{x}$ is the mean for the variable x, s_x is the standard deviation for the variable x.

We use $\hat{y}$ instead of y to remind ourselves that the line will only give a prediction for the response variable, not its actual value.

You can derive the equation of the regression line using calculus, as is done sometimes in Math 242.

There would be many criteria by which we might find a "best line" - why might the least squares criterion be better in many situations?

In these formulas, $\overline{x}$ is the mean for the variable x, s_x is the standard deviation for the variable x.

We use $\hat{y}$ instead of y to remind ourselves that the line will only give a prediction for the response variable, not its actual value.

You can derive the equation of the regression line using calculus, as is done sometimes in Math 242.

There would be many criteria by which we might find a "best line" - why might the least squares criterion be better in many situations?

Example 3. *We calculate with our wine consumption /life expectancy data. With*

$$\overline{x} = 3.026, s_x = 2.51, \overline{y} = 191.053, s_y = 68.396, r = -.843.$$

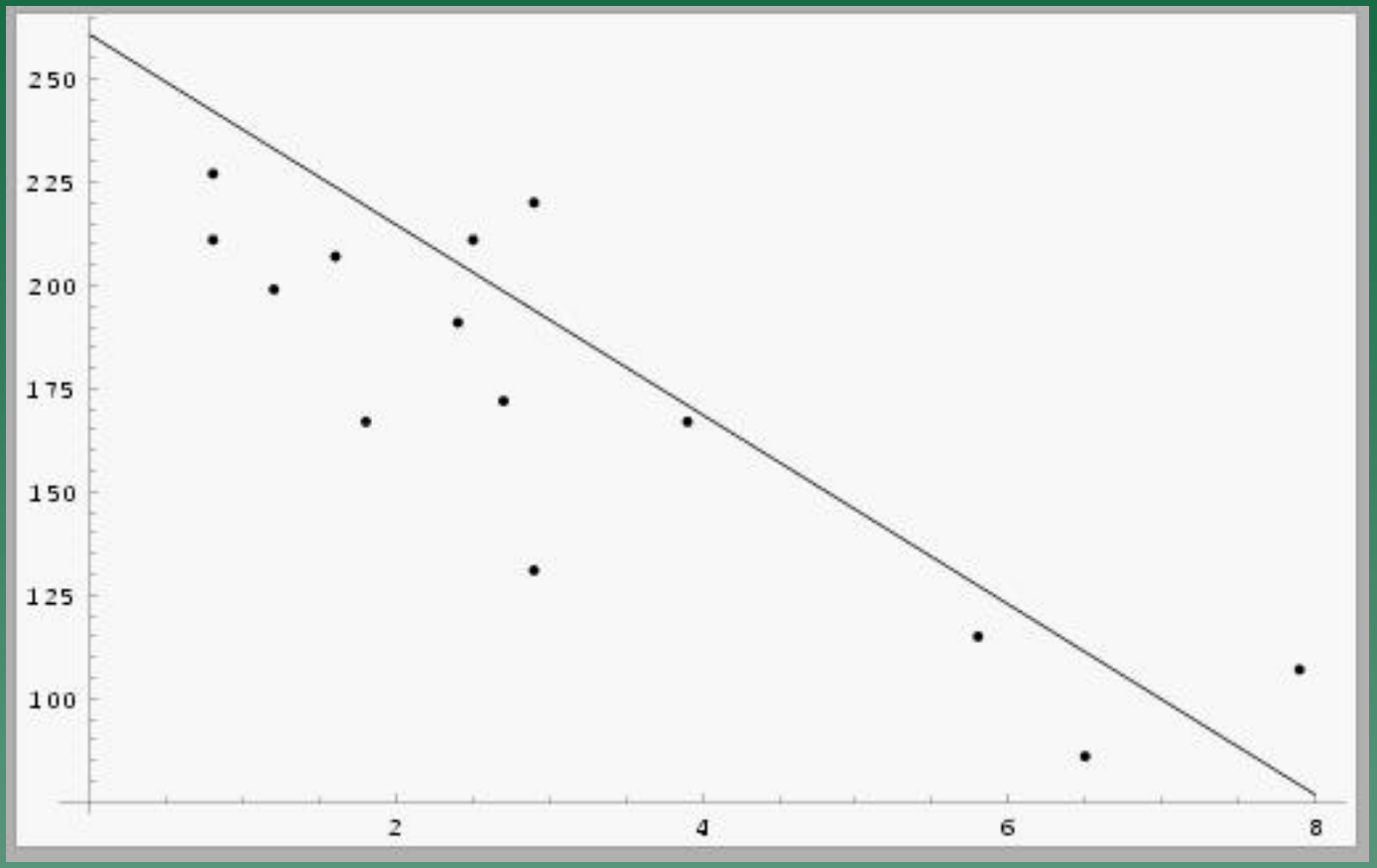
What is the equation of the regression line?

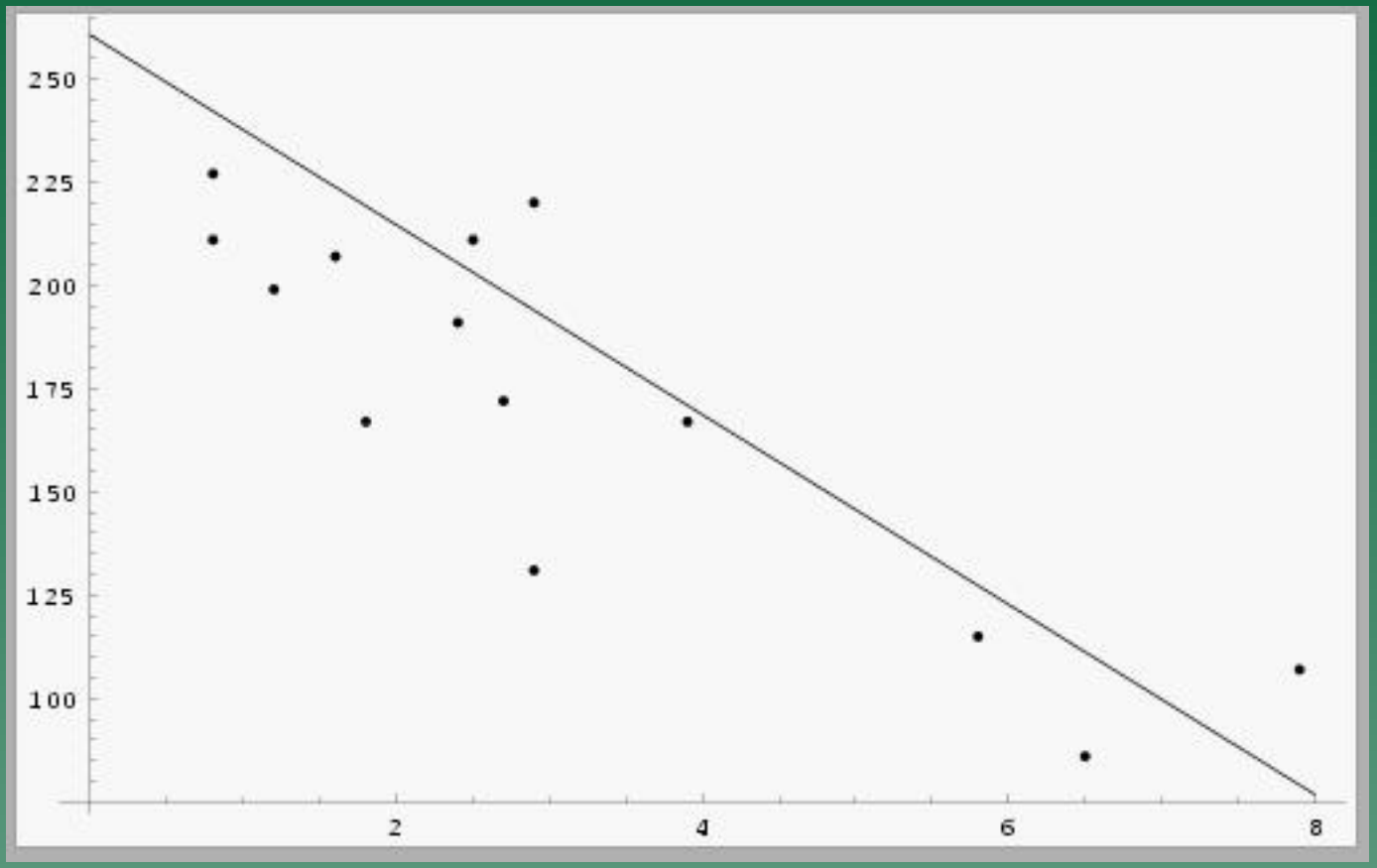
Example 3. *We calculate with our wine consumption /life expectancy data. With*

$$\overline{x} = 3.026, s_x = 2.51, \overline{y} = 191.053, s_y = 68.396, r = -.843.$$

What is the equation of the regression line?

Let's plot the regression line over our scatterplot:


250
225
200
175
150
125
100
2
4
6
8



What do you notice?

Example 4. *Both your calculators and statistical programs such as Excel can do regression analysis.*

Example 4. *Both your calculators and statistical programs such as Excel can do regression analysis. We use Excel to analyze Data Set A from Table 5.2 of your book.*

Example 4. *Both your calculators and statistical programs such as Excel can do regression analysis. We use Excel to analyze Data Set A from Table 5.2 of your book. Steps: enter data;*

Example 4. *Both your calculators and statistical programs such as Excel can do regression analysis. We use Excel to analyze Data Set A from Table 5.2 of your book. Steps: enter data; highlight data;*

Example 4. *Both your calculators and statistical programs such as Excel can do regression analysis. We use Excel to analyze Data Set A from Table 5.2 of your book. Steps: enter data; highlight data; "Insert" a chart;*

Example 4. *Both your calculators and statistical programs such as Excel can do regression analysis. We use Excel to analyze Data Set A from Table 5.2 of your book. Steps: enter data; highlight data; "Insert" a chart; under Chart menu, add a trendline (with options to give equation).*

Using the regression line for prediction

Using the regression line for prediction

Prediction using the regression line is simple -

Using the regression line for prediction

Prediction using the regression line is simple - just plug in.

Using the regression line for prediction

Prediction using the regression line is simple - just plug in.

Example 5. *Suppose we know that in Elbonia the average person drinks 5 liters of wine per year. What is the predicted rate of heart attack?*

Using the regression line for prediction

Prediction using the regression line is simple - just plug in.

Example 5. *Suppose we know that in Elbonia the average person drinks 5 liters of wine per year. What is the predicted rate of heart attack?*

Suppose in the Grand Duchy of Fenwick, we know the average person drinks 1.8 liters of wine each year.

Using the regression line for prediction

Prediction using the regression line is simple - just plug in.

Example 5. *Suppose we know that in Elbonia the average person drinks 5 liters of wine per year. What is the predicted rate of heart attack?*

Suppose in the Grand Duchy of Fenwick, we know the average person drinks 1.8 liters of wine each year. What is our predicted value of y?

Using the regression line for prediction

Prediction using the regression line is simple - just plug in.

Example 5. *Suppose we know that in Elbonia the average person drinks 5 liters of wine per year. What is the predicted rate of heart attack?*

Suppose in the Grand Duchy of Fenwick, we know the average person drinks 1.8 liters of wine each year. What is our predicted value of y?

Example 6. *Using our linear regression, what would we predict from Data Set A of Table 5.2 for a y-value if we observe an x-value of 10?*

Example 6. *Using our linear regression, what would we predict from Data Set A of Table 5.2 for a y-value if we observe an x-value of 10? of 15?*

Example 6. *Using our linear regression, what would we predict from Data Set A of Table 5.2 for a y-value if we observe an x-value of 10? of 15?*

Example 6. *Using our linear regression, what would we predict from Data Set A of Table 5.2 for a y-value if we observe an x-value of 10? of 15?*

It is somewhat funny to "predict" an associated y-value for an x-value which has already been observed;

Example 6. *Using our linear regression, what would we predict from Data Set A of Table 5.2 for a y-value if we observe an x-value of 10? of 15?*

It is somewhat funny to "predict" an associated y-value for an x-value which has already been observed; what does it mean to do so?

Fact 7. *Facts about regression line:*

- *It matters which variable you take as "explanatory" and which as "response."*

Example 6. *Using our linear regression, what would we predict from Data Set A of Table 5.2 for a y-value if we observe an x-value of 10? of 15?*

It is somewhat funny to "predict" an associated y-value for an x-value which has already been observed; what does it mean to do so?

Fact 7. *Facts about regression line:*

- *It matters which variable you take as "explanatory" and which as "response."*

- *Slope is related to* r.

- *Slope is related to r.More precisely, a change of one standard deviation in x predicts a change of r standard deviations in y.*

- *The line goes through $(\overline{x}, \overline{y})$.*

- *Slope is related to r.More precisely, a change of one standard deviation in x predicts a change of r standard deviations in y.*

- *The line goes through $(\overline{x}, \overline{y})$.*

- *r gives the strength of the relationship.*

- *Slope is related to r.More precisely, a change of one standard deviation in x predicts a change of r standard deviations in y.*

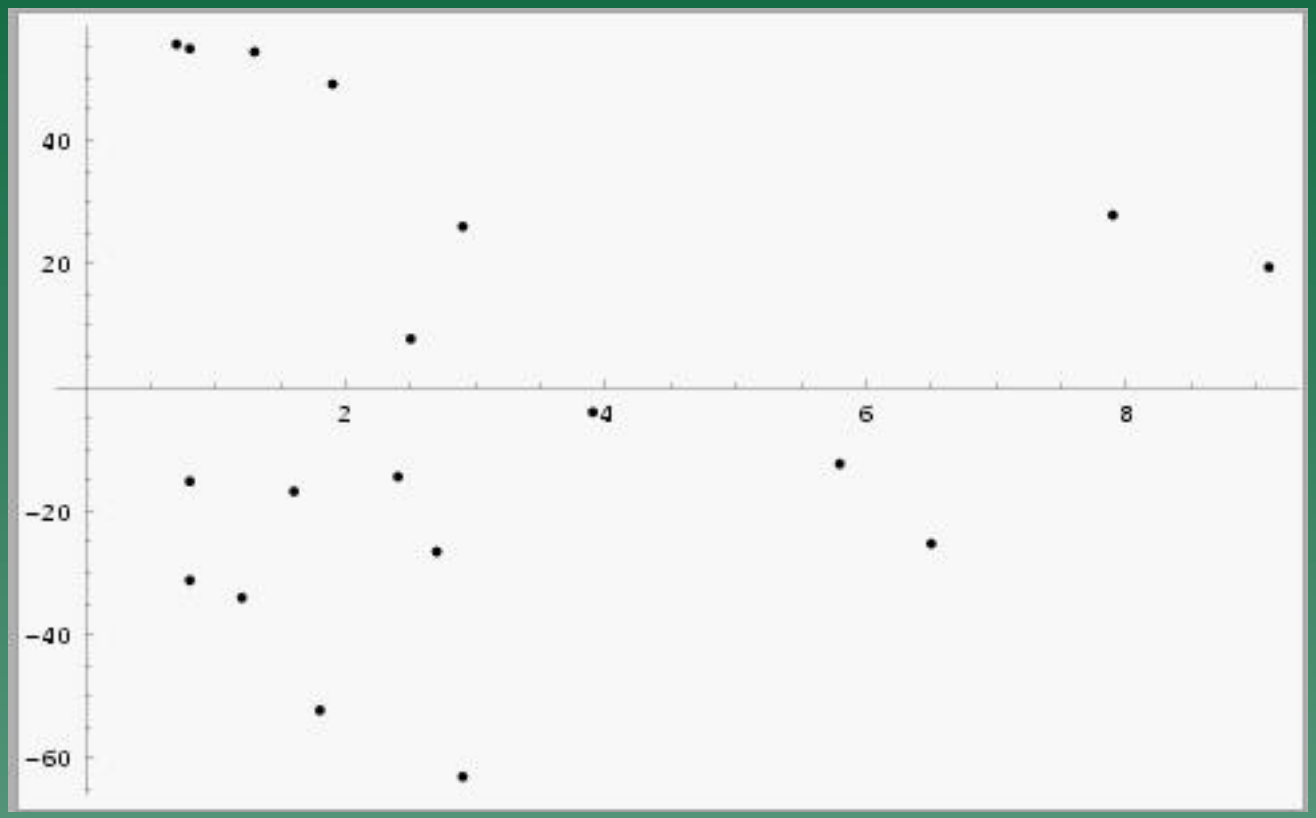
- *The line goes through $(\overline{x}, \overline{y})$.*

- *r gives the strength of the relationship.*

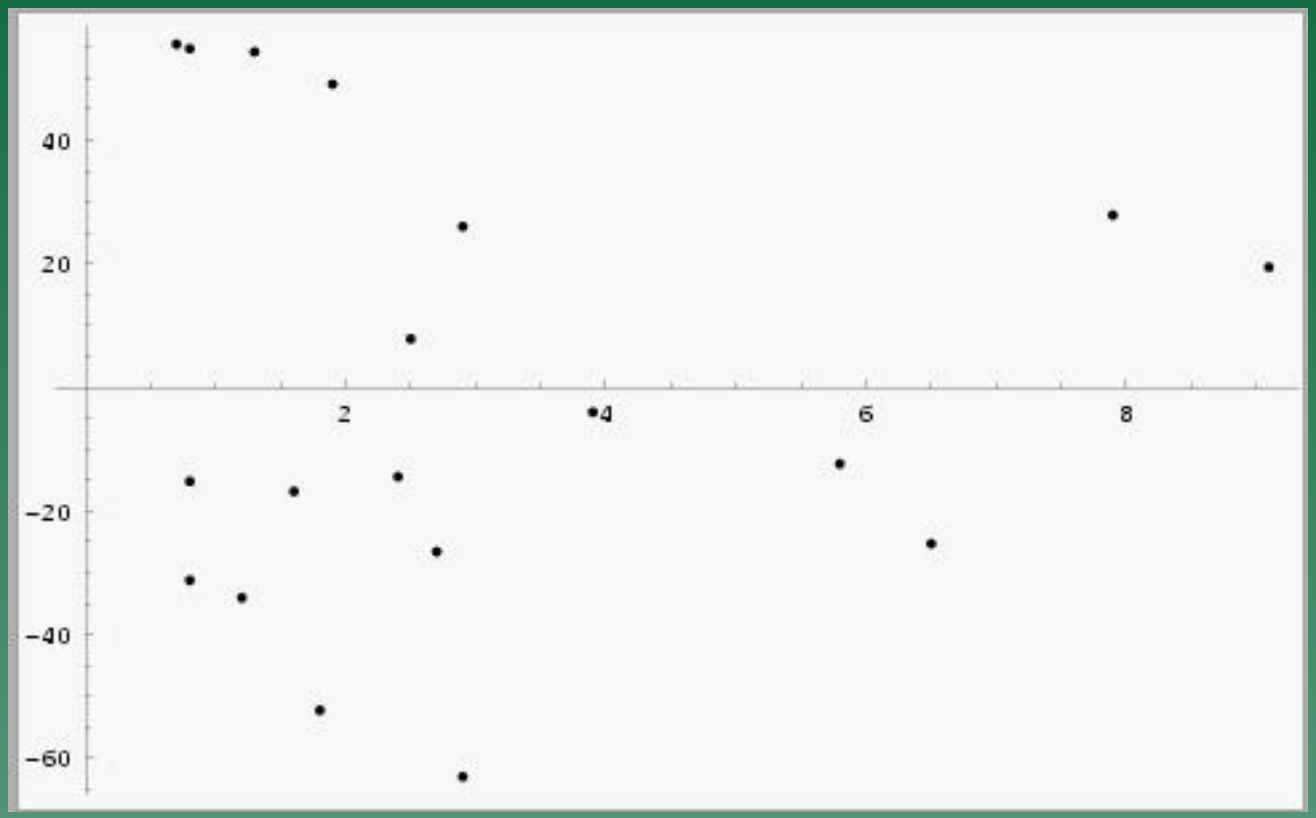
- *r^2 tells the fraction of the variation of y that is explained by the variation of x. (In our wine example, $r^2 = .71$.)*

Residual plot

We can make a residual plot by measuring how far each y value is from the predicted $\hat{y}$ value (for each x).



This gives a subjective impression of how well the regression line fits.



This gives a subjective impression of how well the regression line fits. If most y values are close to zero, the

fit is good.

Correlation and the regression line

Correlation and the regression line

We take data from our dataset involving life expectancies (male, female and both), people per TV and people per doctor. We examine the variables "people per television" as a possible explanatory variable and "life expectancy" as a possible response variable.

Correlation and the regression line

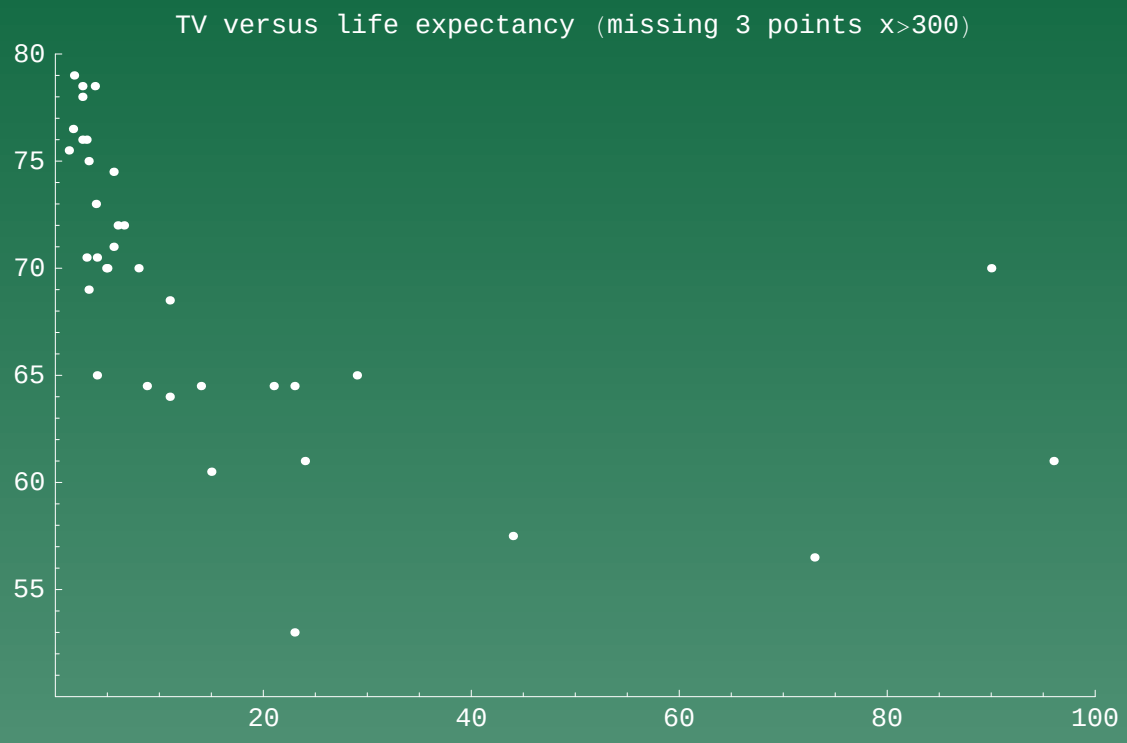
We take data from our dataset involving life expectancies (male, female and both), people per TV and people per doctor. We examine the variables "people per television" as a possible explanatory variable and "life expectancy" as a possible response variable.

We'd like to use this data for two purposes.

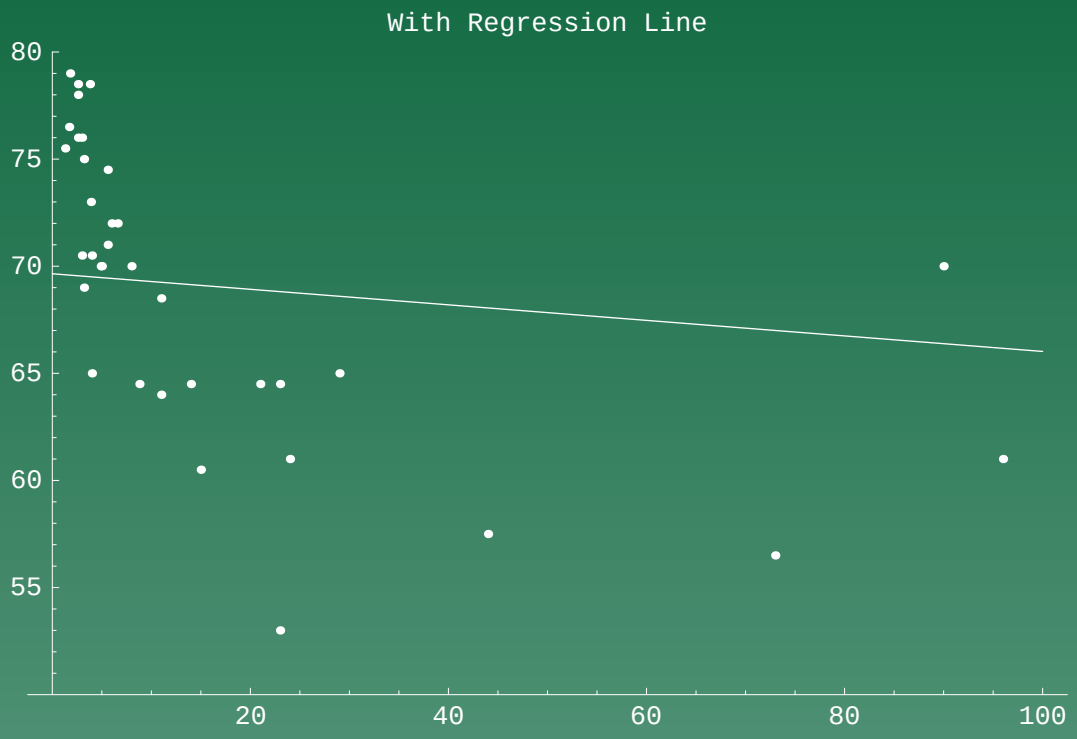
- Correlation does *not* imply causation. These two variable are reasonably strongly correlated, but obviously abundance of TVs does not cause a long lifespan.

- We'll calculate correlation and regression lines removing outliers to demonstrate that correlation and the regression line are *not* resistant.

We will take the opportunity to calculate the regression lines, and try to understand the numbers which go into them, along the way.

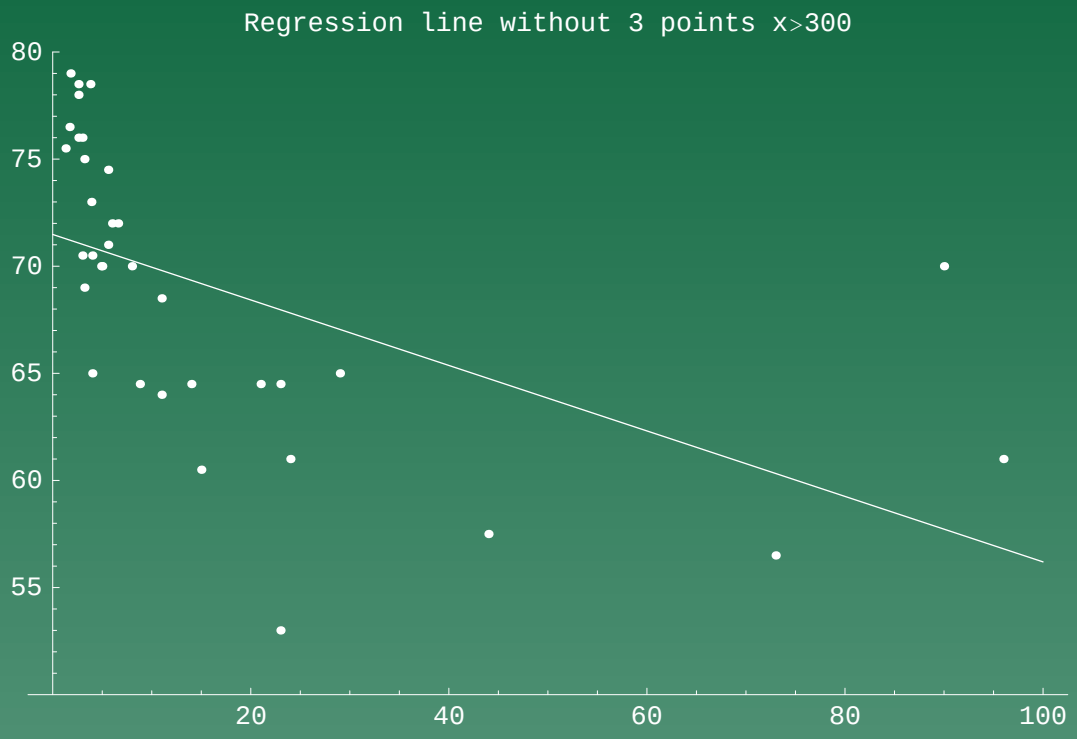


$$r = -.605847, r^2 = .36705$$



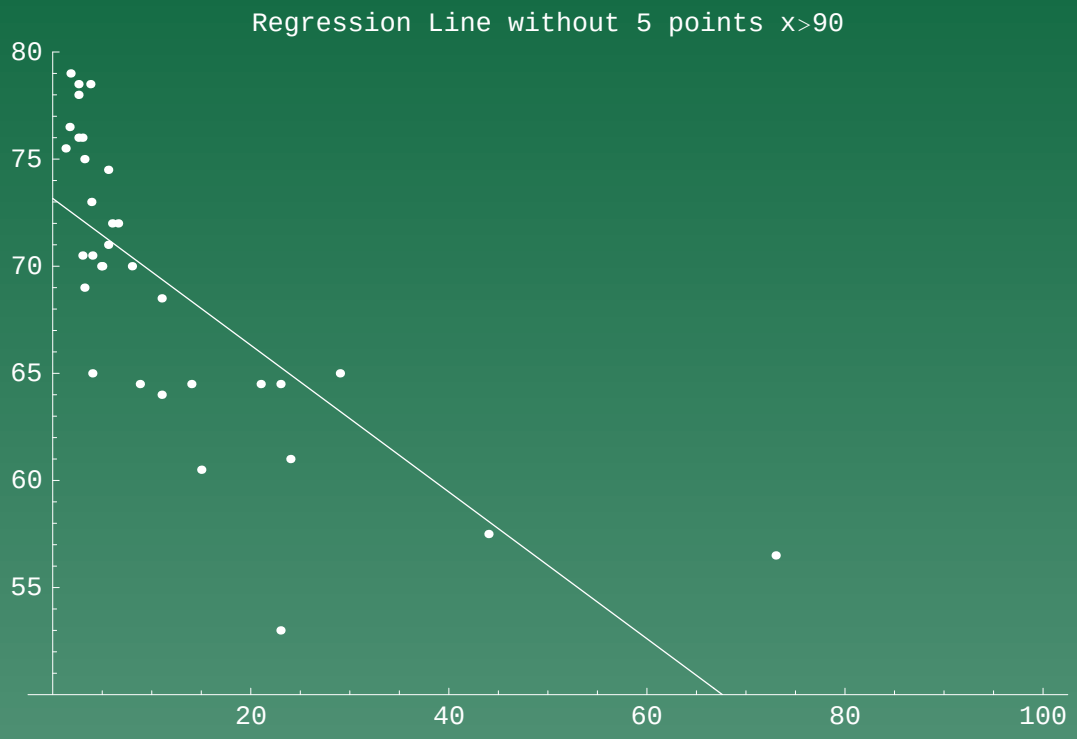
$$r = -.605847, r^2 = .36705$$

$$\hat{y} = 69.64 - 0.036x$$



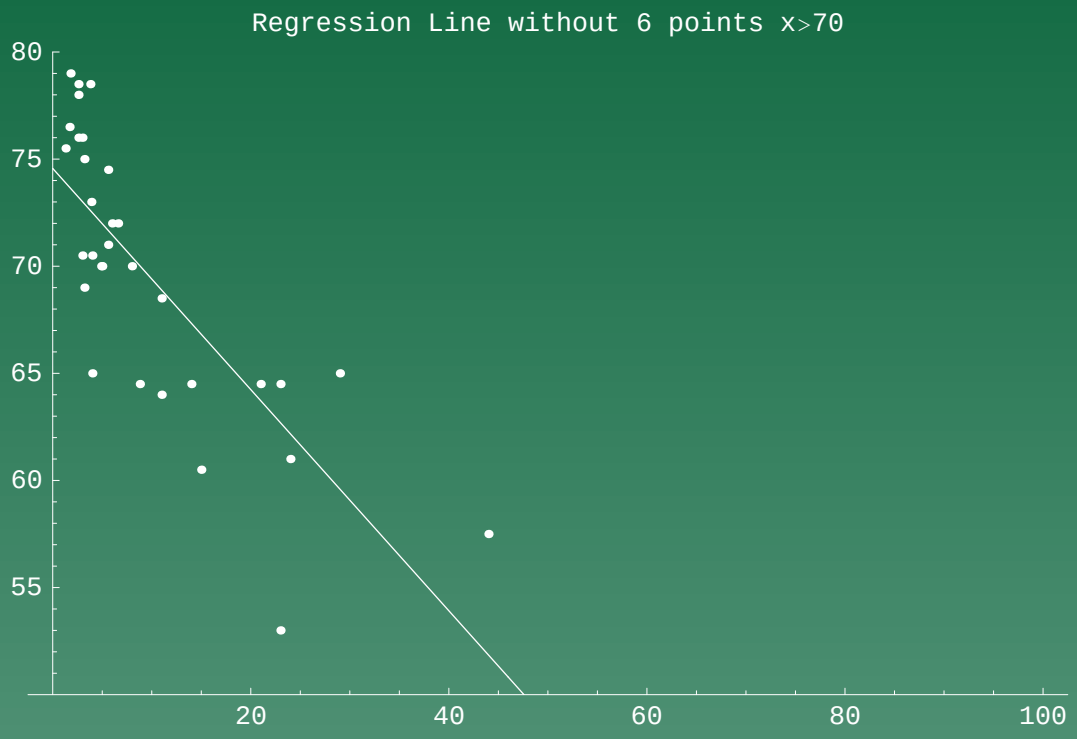
$$r = -.541398, r^2 = .293112$$

$$\hat{y} = 71.48 - 0.153x$$



$$r = -.740354, r^2 = .548123$$

$$\hat{y} = 73.16 - 0.343x$$



$$r = -.784158, r^2 = .614904$$

$$\hat{y} = 74.56 - 0.516x$$

Final cautions about regression

Final cautions about regression

Always plot data you analyze with regression

Final cautions about regression

Always plot data you analyze with regression The equation doesn't tell the whole story.

Final cautions about regression

Always plot data you analyze with regression The equation doesn't tell the whole story. Take for example Data Set D from Table 5.2.

Final cautions about regression

Always plot data you analyze with regression The equation doesn't tell the whole story. Take for example Data Set D from Table 5.2.

Beware the effects of outliers

Final cautions about regression

Always plot data you analyze with regression The equation doesn't tell the whole story. Take for example Data Set D from Table 5.2.

Beware the effects of outliers As we have illustrated with the TV vs. Life Span regressions.

Final cautions about regression

Always plot data you analyze with regression The equation doesn't tell the whole story. Take for example Data Set D from Table 5.2.

Beware the effects of outliers As we have illustrated with the TV vs. Life Span regressions. If a statistic can change a lot with the addition or subtraction of one data point, how heavily can we rely on it?

Avoid extrapolating too much

Avoid extrapolating too much For example if people in Vinland drank 50 liters of wine per year, would we really expect them to be healthier than the Elbonians?

Avoid extrapolating too much For example if people in Vinland drank 50 liters of wine per year, would we really expect them to be healthier than the Elbonians?

Correlation does not imply causation

Avoid extrapolating too much For example if people in Vinland drank 50 liters of wine per year, would we really expect them to be healthier than the Elbonians?

Correlation does not imply causation Or else we would be shipping TV's to sub-Saharan Africa to increase their lifespans.

Avoid extrapolating too much For example if people in Vinland drank 50 liters of wine per year, would we really expect them to be healthier than the Elbonians?

Correlation does not imply causation Or else we would be shipping TV's to sub-Saharan Africa to increase their lifespans.