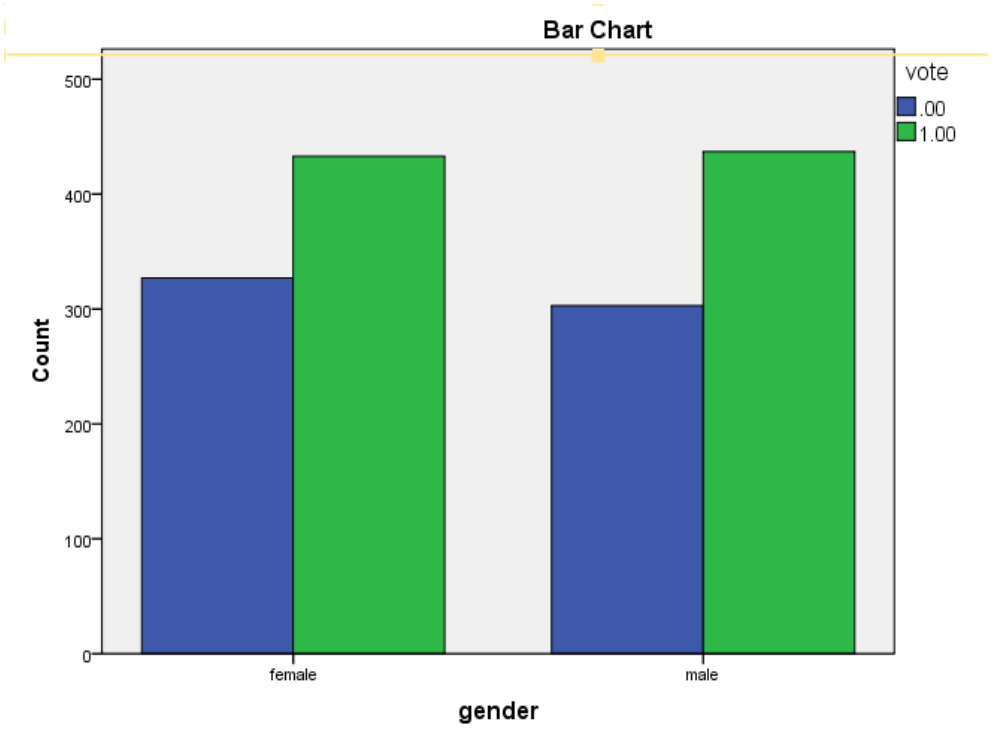
1 Looking at data – Relationships

1.1 Two categorical variables

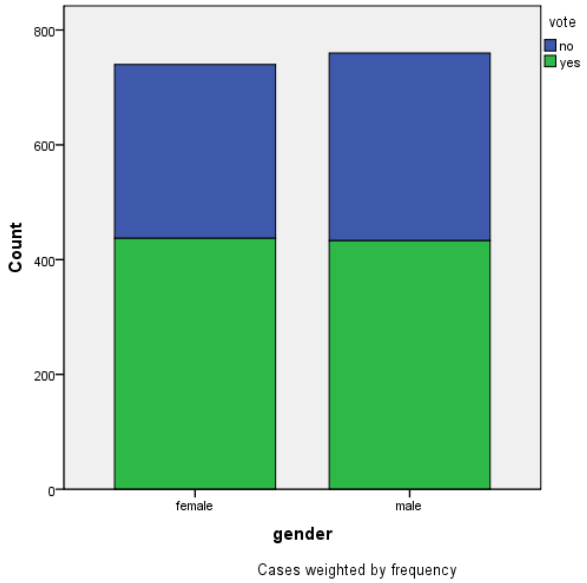
Most of the time graphs are used for the visual comparison of two or more groups in regard to a specific characteristic. We can use bar charts for displaying bivariate categorical data.
For this purpose use a clustered or stacked bar graph. This kind of graph is created by drawing two or more bar charts in one set of horizontal and vertical axes.
The following *two way table* displays the bivariate data for two categorical variables.

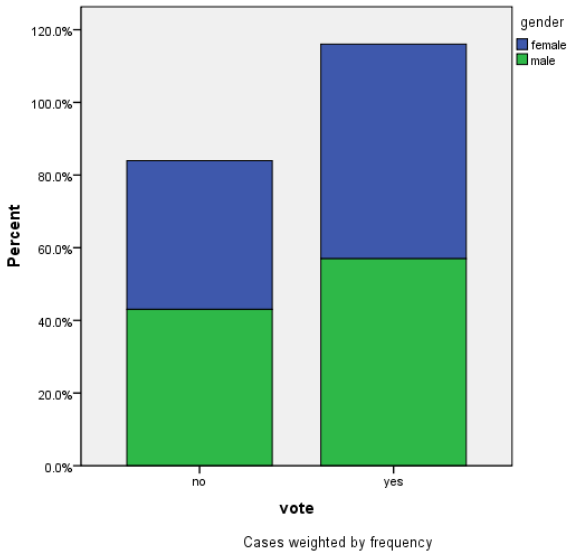
Example 1
The two variables of interest are gender and voted in the 2011 election. A random sample of 1500 eligible voters for the the 2011 federal election resulted in the following summary for variables gender and voting

Gender	Voted		Total
	yes	no	
male	433	327	760
female	437	303	740
Total	870	630	1500

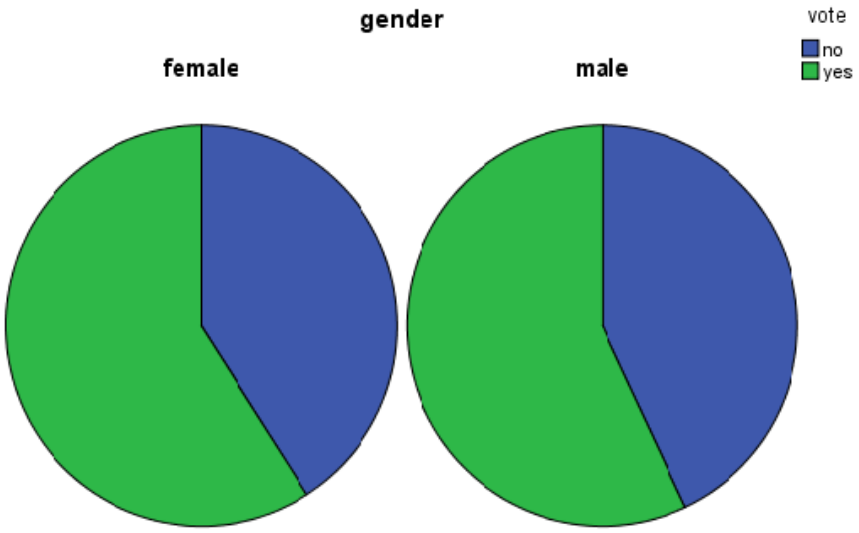


A cluster chart for illustrating the frequencies from the sample.





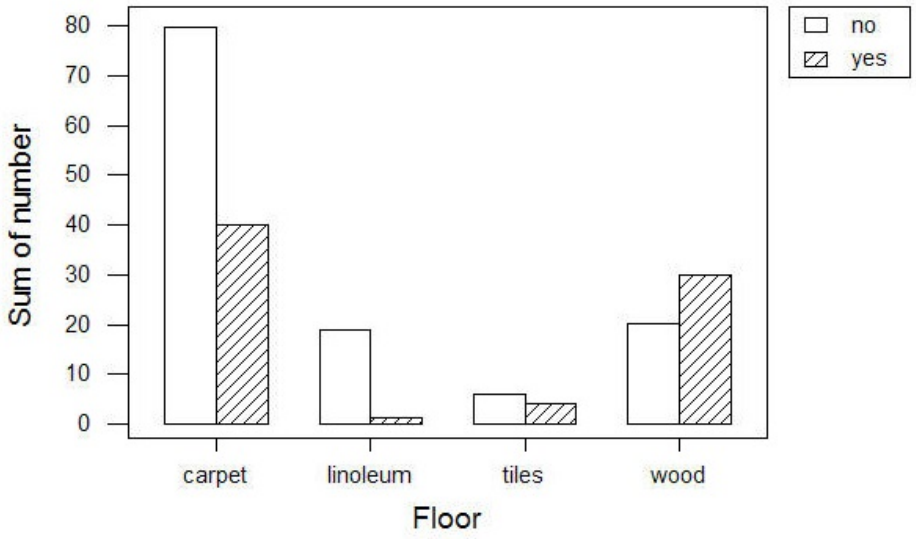
Both graphs show the same data. Explain the difference(s).



... and a pie-chart illustrating the relationship between gender and election participation.

Example 2
Another example about apartments.

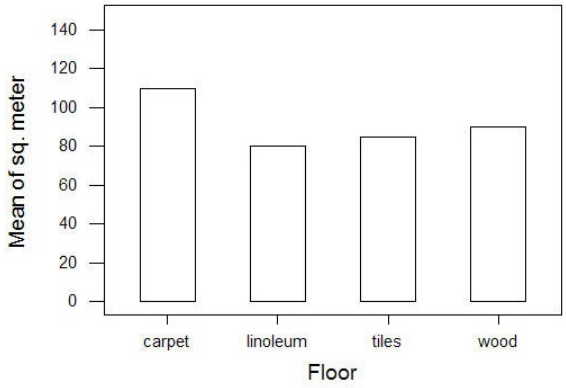
	gym	
Floor	yes	no
wood	30	90
linoleum	1	19
carpet	40	80
tiles	4	6



Remark: Pie charts may also be used. In putting several pie charts beside each other you can obtain a graphical representation of two categorical data variables.

1.2 One categorical, one numerical variable

If summarizing one numerical variable depending on one categorical variable you will calculate mean, standard deviation, median, range, interquartile range, etc. for the numerical variable for all observations that fall into one category. That is the data set will be broken apart according to the categorical variable and then for all those subsets a numerical summary of the numerical variable would be calculated.
For a graphical display bar charts can also be used. They are used to visualize the mean of a numerical variable in different subgroups. In this case the height of the bars are proportional to the mean in the corresponding subgroup of the data.



NEEDS converting data from MiniTab to SPSS do side by side boxplot + rows of histograms

1.3 Two quantitative variables

In this section we will be investigating the relationship and association between two quantitative variables, such as height and weight,
the dose of an injected drug and heart rate,
or the consumption level of some nutrient and weight gain.
The tools used to explore this relationship, are regression and correlation analysis.

These tools can be used to find out if the outcome from one variable depends on the value of the other variable, which would mean a dependency from one variable on the other.

Regression and correlation analysis can be used to describe the nature and strength of the association between two continuous variables.
Sometimes the purpose of a study is to show that one variable can explain the outcome of another variable. In this case we talk about
a **response variable**, which measures an outcome of a study and an **explanatory variable** explains or causes changes in the response variable.

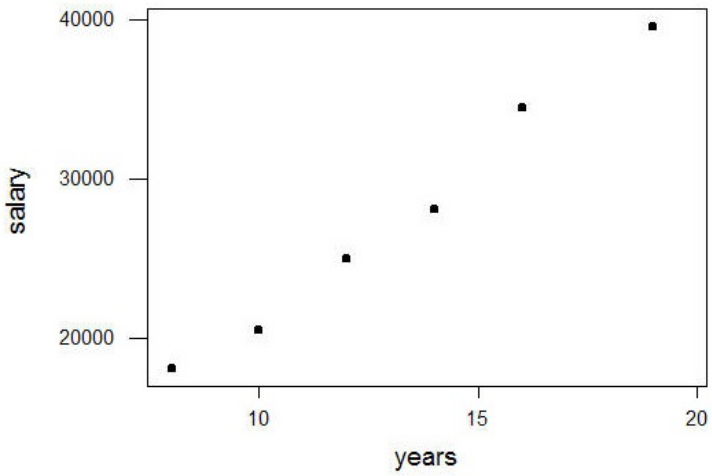
1.3.1 Scatterplot

The first step in the investigation of the relationship between two continuous variables is to create a scatterplot.
A scatterplot allows to visualize the relationship of two variables and to evaluate the **quality** of the association.

Example:
Does the number of years invested in schooling pay off in the job market?
Apparently so – the better educated you are, the more money you will earn. The data in the following table give the median annual income of full-time workers age 25 or older by the number of years of schooling completed.

x=Years of Schooling	y=Salary (dollars)
8	18,000
10	20,500
12	25,000
14	28,100
16	34,500
19	39,700

Start of with creating a scatterplot for X and Y.



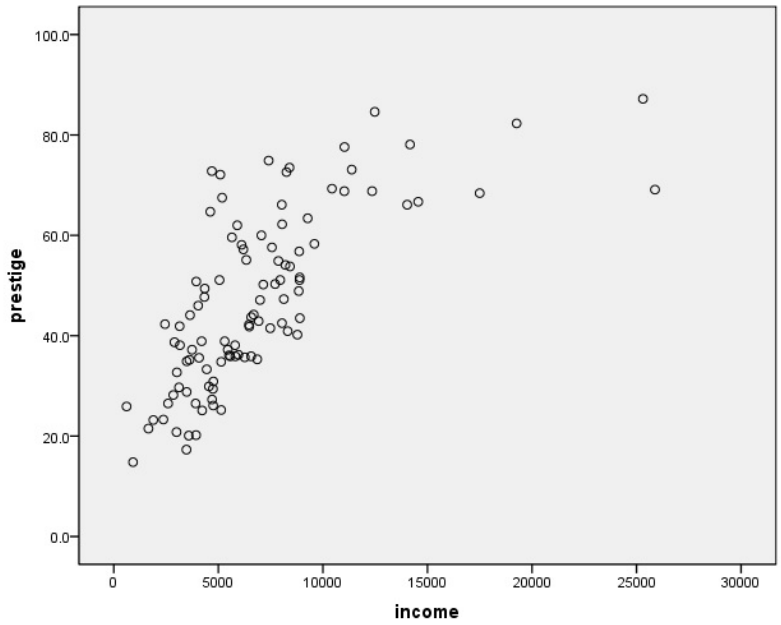
Evaluate the relationship for

- being linear (yes/no)
- strength
- direction
- presence of outliers (yes/no)

This scatterplot shows a strong (strength), positive (direction), linear (linear?) association between years and salary.

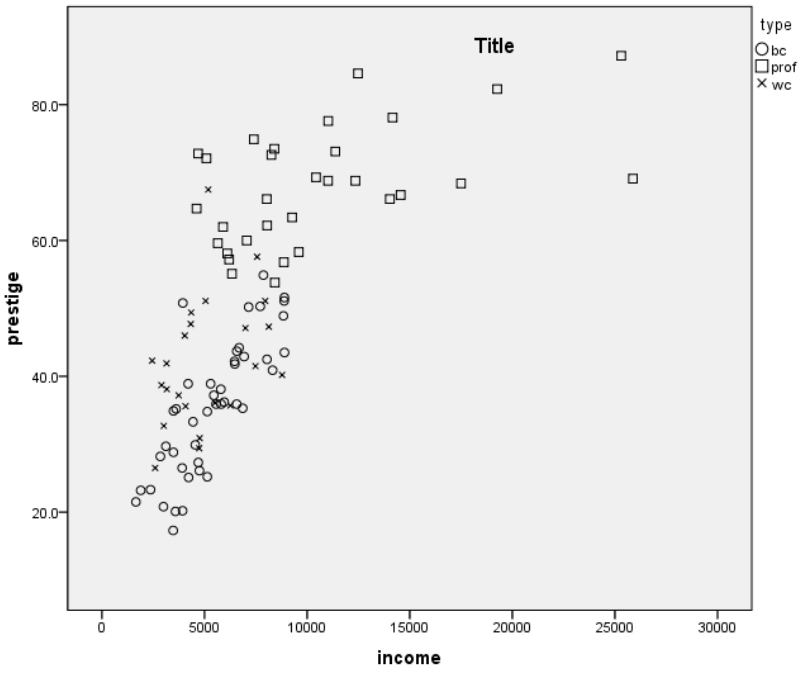
Example:
The prestige data set from 1964 includes 98 occupations. Each occupation has been assigned a prestige value based on survey results, average income according to Stats Canada, mean time of education, percentage of women in this occupation, and each occupation has been categorized as either blue collar, white collar, or professional.

Is the prestige of an occupation associated with its average income?

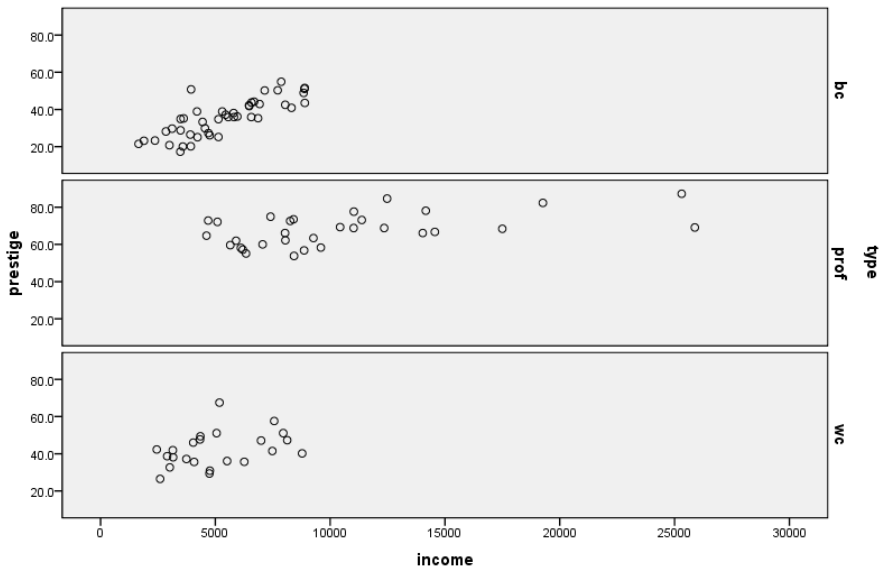


The graph shows a moderately strong, non-linear (maybe logarithmic), positive association between income and prestige.

Interpret the following scatterplot, in which we will add different symbols for the different types of occupation.



Maybe the following graph shows what happens a bit better.



The last two graphs permitted us to include a third variable with the scatterplot. A very powerful tool you will explore in the lab.

1.3.2 Correlation

If the scatterplot shows a reasonable linear relationship (straight line) between the two variables of interest **calculate Pearson's correlation coefficient** to "measure" the **strength** of the **linear** relationship.

Notation:
Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ denote a sample of n measurements of (x, y) pairs.
Assume $n = 201$ and x=height, y=weight, these pairs represent the measurement of hwight and weight from 201 people.

The quantity

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

is called the *covariance* of x and y.

Let s_x and s_y be the sample standard deviations for the x and the y values, respectively. The Pearson's sample correlation coefficient r is then given by

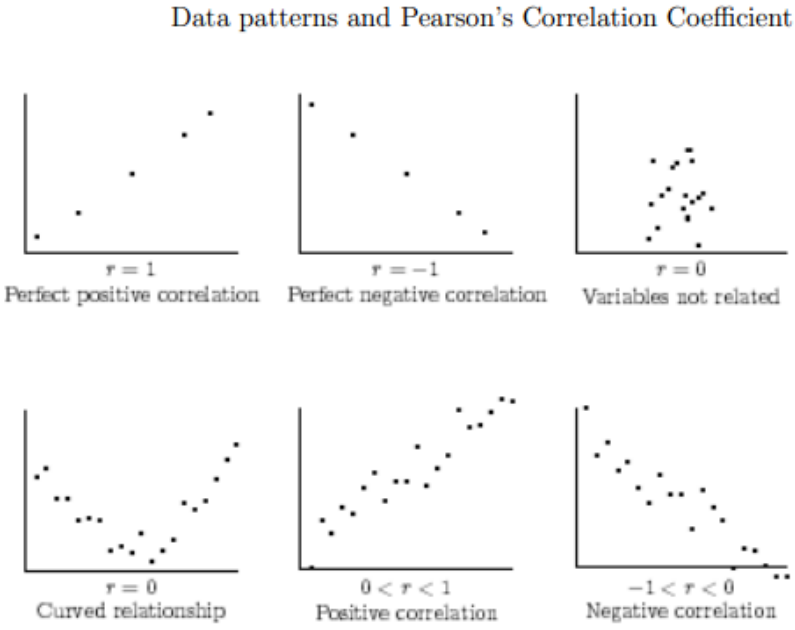
$$r = \frac{s_{xy}}{s_x \cdot s_y}.$$

The easier to compute formula for the covariance is

$$s_{xy} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{(n-1)}$$

Pearson's correlation coefficient (named after Karl Pearson, 1857-1936) is a number between -1 and 1, that measures the strength of a **linear** relationship between two continuous variables.

The **absolute value** of the coefficient measures how closely the variables are related. The closer it is to 1 the closer the relationship. A correlation coefficient over 0.8 indicates a strong correlation between the variables.



The **sign** of the correlation coefficient tells you of the trend in the relationship. A positive (negative) coefficient means that one variable increases (decreases), when the other increases.

For the mathematical inclined reader:
The following formula might help under stand what is measured by r: First standardize all measurements

$$z_{xi} = \frac{x_i - \overline{x}}{s_x}, \quad z_{yi} = \frac{y_i - \overline{y}}{s_y}$$

z_{x_i} measures how many standard deviations the measurement x_i falls away from the mean. For example: Assume $z_{x_1} = 2$ means that the weight of the first person falls 2 standard deviations above the average weight in the sample, the person seems to be quite heavy in comparison to others in the sample.
Assume $z_{x_2} = -0.3$ means that the weight of the second person falls 0.3 standard deviations below the average weight in the sample, the person weighs a bit less than average in comparison to others in the sample.
Then

$$r = \frac{1}{n-1} \sum_i z_{x_i} z_{y_i}$$

From this you can see that r will be positive if larger than average x values are paired with larger than average y values, and less than average x values are paired with less than average y values. In these cases $z_{x_i} z_{y_i}$ will be positive.

But if larger than average x values are paired with less than average y values $z_{x_i} z_{y_i}$ will be negative and r will be negative.

Continue Example: Calculate Pearson's correlation coefficient for years and salary. First find $\bar{x} = 13.17$, $s_x = 4.02$ and $\bar{y} = 27633$, $s_y = 8290$.

x_i=Years of Schooling	y_i=Salary (dollars)	$x_i \cdot y_i$
8	18,000	144000
10	20,500	205000
12	25,000	300000
14	28,100	393400
16	34,500	552000
19	39,700	754300

So that$\sum x_i = 79$, $\sum y_i = 165800$, $\sum x_i y_i = 2348700$. This leads to

$$s_{xy} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{(n-1)} = \frac{2348700 - \frac{(79)(165800)}{6}}{5} = 33133.33$$

So that

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{33133.33}{4.02 \cdot 8290} = 0.9942.$$

The Pearson correlation coefficient of Years of schooling and salary $r = 0.9942$.

A correlation of 0.9942 is very high and shows a strong, positive, linear association between years of schooling and the salary.

Continue Prestige example:
The correlation between income and prestige is
0.805 for blue collar occupations,
0.378 for white collar occupations, and
0.528 for professional occupations.

Give an interpretation and explanation for the observed numbers.

1.3.3 Linear Regression

Sometimes the two variables, x and y, are related in a particular way. it may be that the value of y depends on the value of x; that is the value of x somehow explains the value of y.
In the example for example the time of schooling can viewed as somehow explaining the annual salary of an individual.

Terms and Definition: If we want to use a variable x to draw conclusions concerning a variable y:

y is called dependent or response variable.

x is called independent, predictor, or explanatory variable.

If one of the two variables can be classified as the dependent and the other as the independent variable **and** the relationship between two variables follows a linear pattern it is possible to describe the relationship by a straight line and by an equation of the form:

$$y = b_0 + b_1\ x$$

b_0 is called the **intercept** and b_1 the **slope** of the equation.
The slope is the amount by which y increases when x increases by 1 unit, and the intercept is the value of y when $x = 0$.

Fitting a straight line

Given data points (x_i, y_i) a and b_1 shall now be chosen in that way that the corresponding linear line will have the "best fit" for the given data.

The criteria for "best fit" used in regression analysis is the sum of the squared differences between the data points and the line itself, that is the y deviations.
For data points $(x_i, y_i),\ 1 \leq i \leq n$ this can be written as

$$\min_{b_0,b_1} \sum_{i}^{n} (y_i - (b_0 + b_1 x_i))^2$$

In words: minimize the sum of squared differences by choosingthe parameters b_0 and b_1.

The resulting line is called the *least square line* or *sample regression line.*

After the problem is stated it can be solved mathematically and the results are formulas, how to calculate the *best* (in the least-squares sense) parameters.

$$b_1 = r\left(\frac{s_y}{s_x}\right) = \frac{s_{xy}}{s_x^2} \quad \text{and} \quad b_0 = \bar{y} - b_1\bar{x}$$

Write the equation of the least squares line as

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

$\hat{y}$ gives an estimate for y for a given value of x.

Continue Example:
Since the salary and the years of schooling show such a strong linear relationship and the salary can be viewed as depending on the years of schooling, do a linear regression analysis with the salary as the response variable and the years of schooling as the predictor variable.
Calculate

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{33133.33}{4.02^2} = 2050.28 \quad \text{and} \quad b_0 = \bar{y} - b_1\bar{x} = 27633 - 2050.28 \cdot 13.17 = 630.81$$

Our result is the least squares line

$$\hat{y} = b_0 + b_1 x = 630.81 + 2050.28\ x$$

The slope equals \$2050.28, that is for every year of schooling the average salary increases by this amount.

Estimation: To **estimate** the average salary after 18 years of schooling we calculate $\hat{y}$ with $x = 18$

$$\hat{y} = 630.81 + 2050.28 \cdot 18 = 37535.85\$$$

Don't use the regression line for values outside the range of the observed values. This is a model that only has been proved valid for the given range.

Continue Prestige Example:
Only consider the blue collar occupations.y = prestige, and x=income, SPSS reports $n = 44$, $\bar{x} = 5374.14$, s_x =2004.33, $\bar{y}$ =32.52 $s_y = 10.02$, and we already found $r = 0.805$.
From this we can now find:

$$b_1 = r\left(\frac{s_y}{s_x}\right) = 08.805\left(\frac{10.02}{2004.33}\right) = 0.004$$

and

$$b_0 = \bar{y} - b_1\bar{x} = 32.52 - 0.004(5374.14) = 13.91$$

Therefore

$$\hat{y} = 13.91 + 0.004(\text{income})$$

For an occupation with average income of \$3000 this equation results in an estimated prestige of $13.91 + 0.004(3000) = 25.91$, which is below average for blue collar occupations (How do I know?).

Properties of the regression or least squares line

1. The least squares line passes through the balance point $(\bar{x}, \bar{y})$ of the data set.
2. The regression line of y on x should not be used to predict x, since it is not the line that minimizes the sum of squared x deviations.

Assessing the fit of a line
Once the least squares line has been obtained, it is natural to examine how effectively the line summarized the relationship between x and y.
The first question that has to be answered is, if the line is an appropriate way to summarize the relationship. In order to answer this question, we will calculate the coefficient of determination r^2.

Definition: The coefficient of determination for he regression of y on x is
r^2 the square of Pearson's Correlation Coefficient (easy!).
It gives the proportion of variation in y that can be attributed to a linear relationship between x and y.
Is r^2 greater than 0.8, the model has an excellent fit and can be used to calculate reliable predictions of the dependent variable by using the independent variable.

In the example, the variable Years of Schooling explains $r^2 = 98.8\%$ of the variation in the variable Salary. Which is very high. The plot showed that the data points are almost on a straight line.

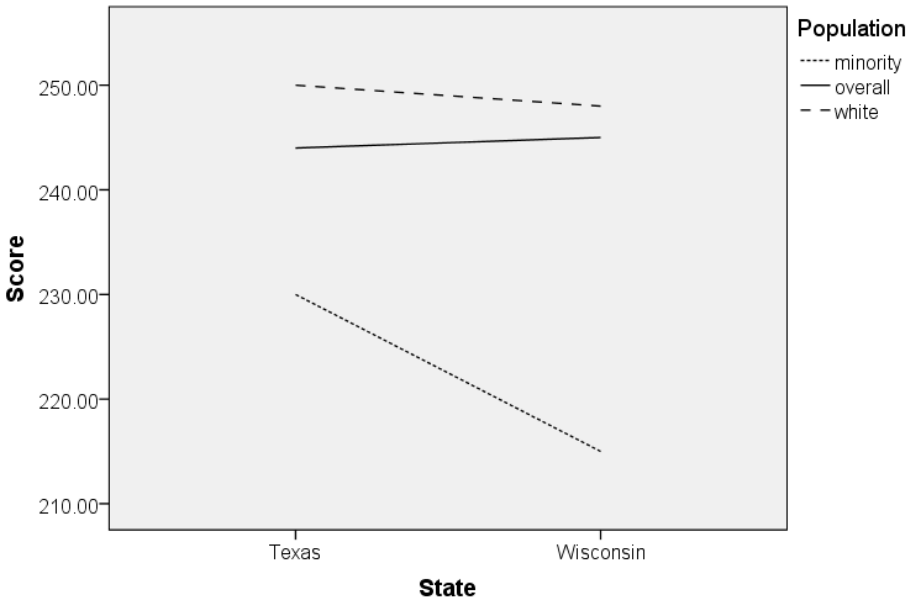
Continue Prestige Example:
$r^2 = 0.805^2 = 0.648$. For blue collar occupations about 65% of the variation in prestige can be explained through the linear relationship with income. Meaning that income is a very important factor but by far not the only contributing variable in the prestige of an occupation.

1.4 Simpson's Paradox

Example:
In 2009 Texas was blamed to underfund its schools and therefore Texas' students underperformed at nationwide tests. They were asked to look at what Wisconsin was doing the second ranked school system according to this test.
Here are the scores for both states depending on ethnicity of the students (not quite the true data, but they capture the gist in what happened)

	State	
Ethnicity	Texas	Wisconsin
white	250	248
minority	230	215
overall	244	245

In calculating the overall average. it had to be taken into account that in Texas the minority population is about 30% and in Wisconsin only 10%.



If you could choose, which school system would you choose if you were white?
If you could choose, which school system would you choose if you were from a minority?
Should learn Texas from Wisconsin or vice versa?

Observe that overall Wisconsin students seem to outperform Texas' students, but in both subpopulations Texas reports the higher averaged score. The reason for this is that Texas has many more minority students, who score lower on average than white students.

This is an example for Simpson's Paradox. The trend in the overall population is different from the trend in the subpopulations.