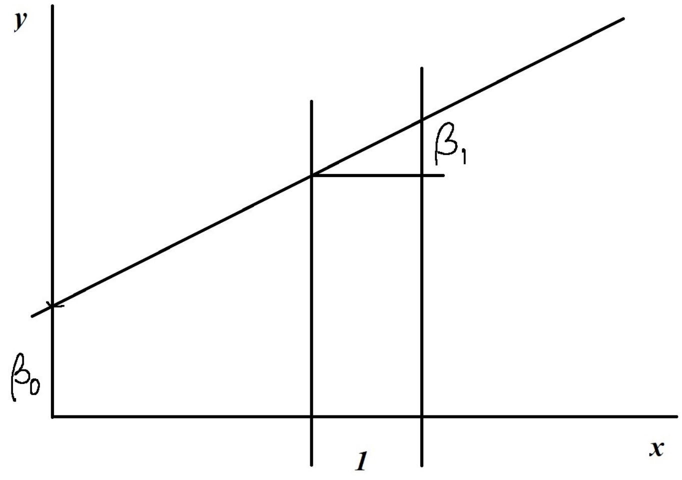
# Contents

# 1 Simple Linear Regression - Descriptive only

Simple Linear Regression is used to assess, measure, and describe the association between two quantitative variables.
We often perceive one of the two variables as the response variable,$y$, and the other as the predictor variable, $x$.
E.g. temperature, $x$, affects the yield of corn, $y$
narcissististy, $x$, affects the number of self portraits posted on instagram, $y$
temperature, $x$, affects how many people are on a ski hill, $y$
years of schooling, $x$, affects salary, $y$
In this course we only consider linear relationships between the two variables of interest, i.e. the mean of $y$ can be described as a general pattern as a line in $x$.
mean($y$) $= \beta_0 + \beta_1 x$ with

$\beta_0 =$ intercept

$\beta_1 =$ slope



**Example 1**
**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 | $y$=Median Income (in 1000 dollars) |
|---|---|
| 8 | 18 |
| 10 | 20.5 |
| 12 | 25 |
| 14 | 28.1 |
| 16 | 34.5 |
| 19 | 39.7 |

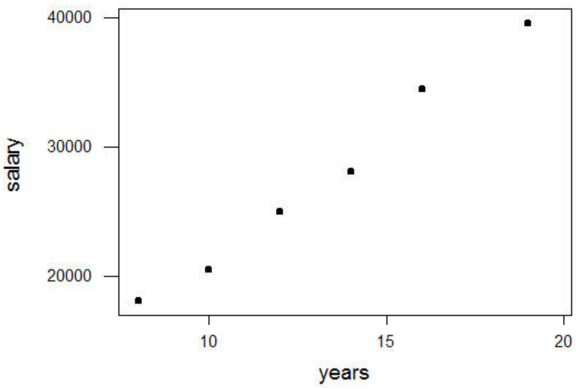
Task: describe and measure the association between Years of Schooling and Median Income.

## 1.1 Describing the Relationship with a Scatterplot

In a first step obtain a scatterplot to check visually if a linear relationship seems to be a reasonable description for the data.
Comment on

- **direction**: positive (the larger $x$ the larger $y$) or negative (the larger $x$ the smaller $y$)
- **strength**: weak/moderate/strong. Judgement depends on how closely the data follow a curve, or how far the measurements fall away from the curve
- **shape**: linear versus non-linear versus no association

**Example 2**
Scatterplot for $x$ and $y$.



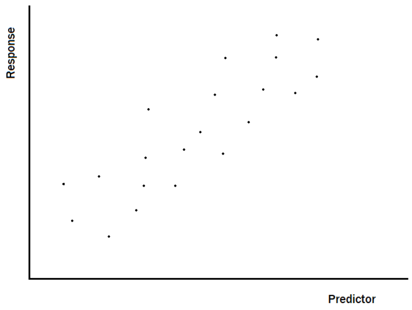
The scatterplot shows a strong, positive, linear association between years of schooling and the median salary. Since there is a linear pattern the tools we will be learning here are appropriate to be applied to these data.
We will introduce the correlation coefficient to measure the strength of the linear association, and will see how to find the line of best fit to summarize the relationship between the variables.

## 1.2 Pearson's Correlation Coefficient

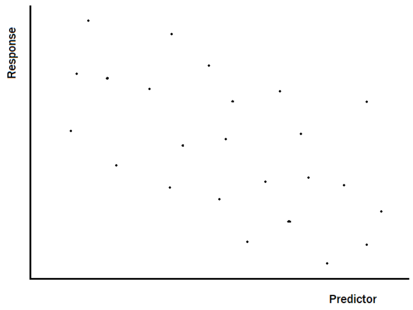
Pearson's correlation coefficient, $r$, is a number that measures the strength of the linear relationship between two numerical variables.

Properties

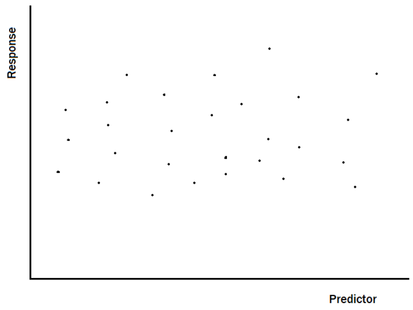
1. $-1 \leq r \leq 1$
2. independent from units used
3. $abs(r) = 1$ indicates that all values fall precisely on a line
4. $r > (<)0$ indicates a positive (negative) relationship.
5. $r \approx 0$ indicates no linear relationship



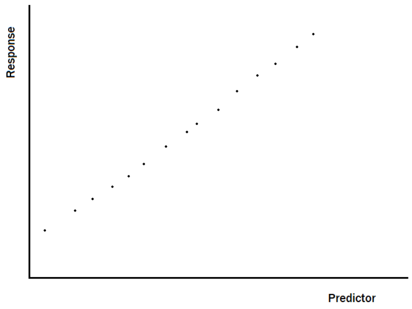
$r \approx 0.8$



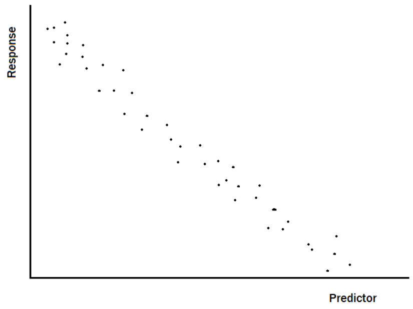
$r \approx -0.5$



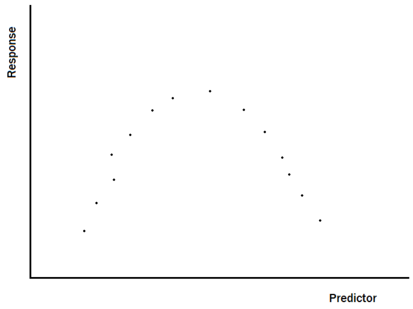
$r \approx 0$



$r \approx 1$



$r \approx -0.9$



$r \approx 0(\text{nonlinear})$

Calculating Pearson's correlation coefficient, $r$

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

with

$$SS_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}, \ SS_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}, \text{ and } \ SS_{yy} = \sum x_i^2 - \frac{(\sum y_i)^2}{n}$$

**Continue example:**

First calculate the three sum of squares and then plug those into the formula for $r$.
To calculate the sum of squares first calculate the required totals from the data:

| | $x_i$ | $y_i$ | $x_i^2$ | $y_i^2$ | $x_i y_i$ |
|---|---|---|---|---|---|
| | 8 | 18 | 64 | 324 | 144 |
| | 10 | 20.5 | 100 | 420.25 | 205 |
| | 12 | 25 | 144 | 625 | 300 |
| | 14 | 28.1 | 196 | 789.61 | 393.4 |
| | 16 | 34.5 | 256 | 1190.25 | 552 |
| | 19 | 39.7 | 361 | 1576.09 | 754.3 |
| totals | 79 | 165.8 | 1121 | 4925.2 | 2348.7 |
| | $\sum x_i$ | $\sum y_i$ | $\sum x_i^2$ | $\sum y_i^2$ | $\sum x_i y_i$ |

Using these totals calculate:

$$\begin{aligned} SS_{xy} &= 2348.7 - \frac{(79)(165.8)}{6} = 165.667 \\ SS_{xx} &= 1121 - \frac{(79)^2}{6} = 80.833 \\ SS_{yy} &= 4925.2 - \frac{(165.8)^2}{6} = 343.593 \end{aligned}$$

and now $r$:

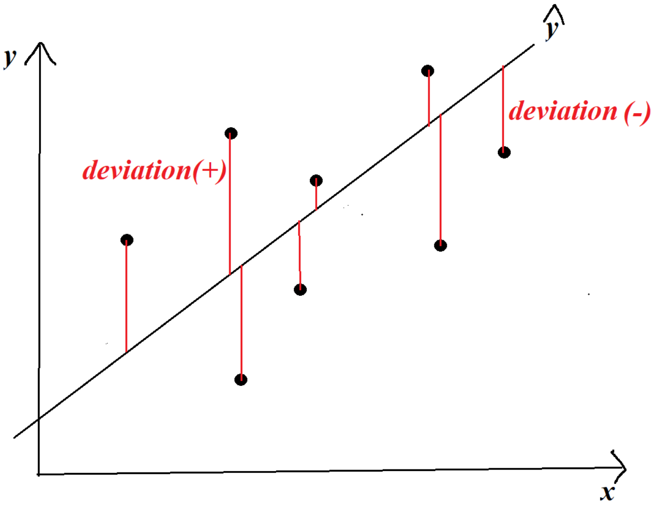
$$r = \frac{165.667}{\sqrt{80.833(343.593)}} = 0.994$$

$r = 0.994$ indicates a very strong positive linear relationship between the years of schooling and the mean income, which is what we should have expected given the scatterplot.
This number indicates that there is a line that is a good representation of the association between Years of Schooling and Median Salary. In a next step we will find that line!

## 1.3 Least Squares Regression Line

How shall we choose the "best" line to represent the relationship?
The "Least Squares Regression Line" is the line that minimizes the squared vertical distances of the measurements to the line.

deviations = difference between observed measurements and the line = error of prediction



The line that results in the least squared deviations (for the sample data) is called the least-squares-regression-line.

Mathematically this means making the sum of squares for error, $SSE$, as small as possible.

$$SSE = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

Asking mathematicians to solve this problem they provide the following formulas for the least squares intercept,$\hat{\beta}_0$, and slope, $\hat{\beta}_1$:

- slope: $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$ and
- intercept: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, where $\bar{y}$ is the average of the $y$ values, and $\bar{x}$ is the average of the $x$ values.

**Continue example:**
Most of the work has already been done!

For years and schooling and median salary this leads to:

$$\begin{aligned}\hat{\beta}_1 &= \frac{165.667}{80.833} &= 2.049\\ \hat{\beta}_0 &= \frac{165.8}{6} - 2.049\frac{79}{6} &= 0.648\end{aligned}$$

The least squares regression line summarizing our data is

$$\hat{y} = 0.648 + 2.049x$$

**Interpretation(important!):**
The intercept ($\hat{\beta}_0$) estimates that the median income is \$648 for a person with zero years of education.
Using this interpretation would mean that we apply the result outside the range of observed values of $x$. This is called extrapolation, which is highly problematic and usually not admissible.

The slope ($\hat{\beta}_1$) estimates that on average the income increases by \$2049 for every extra year of schooling.

## 1.4 Coefficient of Determination

The coefficient of determination measures the usefulness of the line for predicting and estimating future values.

Coefficient of determination, $r^2$.
$r^2$ is always a number between 0 and 1.
Widely use can be explained because of the interpretation:
An alternative formula for $r^2$ is

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}}$$

This formula can help to understand the interpretation of $r^2$.
$SS_{yy}$ measures the spread in the $y$ values, and $SSE$ measures the spread in $y$ we can not explain through the linear relationship with $x$. So that $SS_{yy} - SSE$ is the spread in $y$ explained through the linear relationship with $x$. By dividing this by $SS_{yy}$, we find that $r^2$ gives the proportion in the spread of $y$, that is explained through the linear relationship with $x$.

**Continue Example:**
The coefficient of determination in our example is

$$r^2 = 0.994^2 = 0.988$$

98.8% in the variation of the mean income is explained through years of schooling.
A high coefficient of determination indicates that the model is suitable for estimation and prediction.

## 1.5 Where do we go from here?

To use methods from inferential statistics to be able to generalize the findings from the sample to the underlying population the *Simple Linear Regression Model* will be introduced.
The model describes the population, and correlation coefficient, intercept, and slope will be interpreted as estimates of parameters of the model. Confidence intervals and tests regarding the slope are of particular interest.
What would be the relevance of a test of the slope being different from 0?