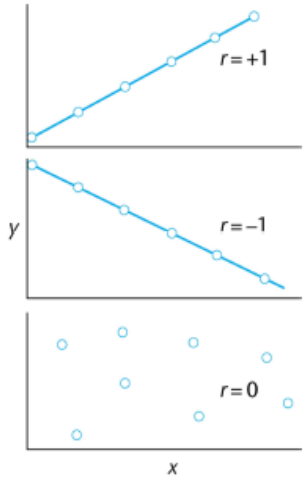
Review of Correlation and Simple Linear Regression

Correlation

A correlation coefficient is a number between -1 and 1 which measures the degree to which two variables are linearly related. If there is perfect linear relationship with positive slope between the two variables, we have a correlation coefficient of 1; if there is positive correlation, whenever one variable has a high (low) value, so does the other.

If there is a perfect linear relationship with negative slope between the two variables, we have a correlation coefficient of -1; if there is negative correlation, whenever one variable has a high (low) value, the other has a low (high) value.

A correlation coefficient of 0 means that there is no linear relationship between the variables.



It can be shown that the correlation coefficient satisfies

$$-1 \leq r \leq 1.$$

and $|r| \approx 1$ then the relation is close to linear.

Pearson's Product Moment Correlation Coefficient

Pearson's product moment correlation coefficient, usually denoted by ***r***, is one example of a correlation coefficient. It is a measure of the linear association between two variables that have been measured on interval or ratio scales, such as the relationship between height in inches and weight in pounds.

However, it can be misleadingly small when there is a relationship between the variables but it is a non-linear one.

Example 5.3.1

Standard aqueous solutions of fluorescein are examined in a fluorescence spectrometer, and yield the following fluorescence intensities (in arbitrary units):

Fluorescence intensities:	2.1	5.0	9.0	12.6	17.3	21.0	24.7
Concentration, pg ml^{-1}	0	2	4	6	8	10	12

Determine the correlation coefficient, *r*.

In practice, such calculations will almost certainly be performed on a calculator or computer, alongside other calculations covered below, but it is important

We can determine the Pearson Correlation coefficient in R using the `cor()` command. To get a more complete statistical analysis, with formal tests, we can use the command `cor.test()`

The interpretation of the output from the `cor.test()` procedure is very similar to procedures we have already encountered. The null hypothesis is that the correlation coefficient is equal to zero. This is equivalent to saying that there is no linear relationship between variables.

```
> Conc=c(0,2,4,6,8,10,12)
> Fluo=c(2.1,5.0,9.0,12.6,17.3,21.0,24.7)
>
> cor(Fluo,Conc)
[1] 0.9988796
>
> cor.test(Fluo,Conc)

        Pearson's product-moment correlation

data:  Fluo and Conc
t = 47.1967, df = 5, p-value = 8.066e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9920730 0.9998421
sample estimates:
      cor
0.9988796
```

Remark upon the following outputs:

- The correlation coefficient: **0.9988796** (very strong positive linear relationship)
- The 95% confidence interval for the correlation coefficient estimate: **(0.9920730,0.9998421)**
- p-value: **8.066e-08** (i.e. Reject the Null Hypothesis)

There are procedures, based on ***Pearson's coefficient***, for making inferences about the population correlation coefficient. However, these make the implicit assumption that the two variables are jointly normally distributed.

When this assumption is not justified, a non-parametric measure such as the Spearman Rank Correlation Coefficient might be more appropriate.

(Let us assume for a moment that both ***Fluo*** and ***Conc*** are not normally distributed)

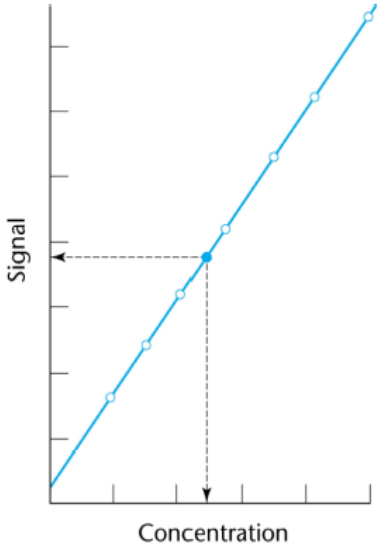
The specification is the same as for Pearson's test, with the additional argument "method=spearman").
The interpretation is very similar, but there are no confidence intervals for the estimates.

```
> cor.test(Conc,Fluo,method="spearman")

        Spearman's rank correlation rho

data:  Conc and Fluo
S = 0, p-value = 0.0003968
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
  1
```

Regression Line

A regression line is a line drawn through the points on a scatterplot to summarise the relationship between the variables being studied. When it slopes down (from top left to bottom right), this indicates a negative or inverse relationship between the variables; when it slopes up (from bottom right to top left), a positive or direct relationship is indicated.

The regression line often represents the regression equation on a scatterplot.



Simple Linear Regression

Simple linear regression aims to find a linear relationship between a response variable and a possible predictor variable by the method of least squares.

Ordinary Least Squares

The method of least squares is a criterion for fitting a specified model to observed data. For example, it is the most commonly used method of defining a straight line through a set of points on a scatterplot.

Regression Equation

A regression equation allows us to express the relationship between two (or more) variables algebraically. It indicates the nature of the relationship between two (or more) variables. In particular, it indicates the extent to which you can predict some variables by knowing others, or the extent to which some are associated with others.

Recall that we class one variable as the response or dependent variable, usually denoted Y, and the other as the predictor, or dependent variable, usually denoted X.
X is said to “cause” changes in Y.

If a linear relationship , the relationship between X and Y is formulated as follows:

$$Y = \beta_o + \beta_1 X + \varepsilon$$

- is the predicted value for the dependent variable
- β_0 is the intercept coefficient
- β_1 is the slope coefficient
- X is the independent variable
- ε is the residual term (i.e random error)

Both β_0 and β_1 are almost always unknown (population) values. However these are the key terms in the model. From a sample of data estimates for the slope and estimate coefficient are derived.

A fitted line to model the data as a linear regression mode (i.e. a regression equation) is usually written as

$$= b_o + b_1 X$$

where

- is the predicted value for the dependent variable
- b_0 is the intercept estimate
- b_1 is the slope estimate (or regression coefficient)
- X is the independent variable

Simple linear regression is from a family of models known as Linear Models. The ***R*** command used to implement such models is `lm()` .

The regression model is specified in the following form: `lm(Y ~ X)`

The operator "~" (the tilde sign) is taken to mean "is explained by" or "is predicted by".

For our previous example, the simple linear model can be implemented as follows:

```
> lm(Fluo~Conc)

Call:
lm(formula = Fluo ~ Conc)

Coefficients:
(Intercept)         Conc
      1.518        1.930
```

Using the coefficients given in the computer output, the regression equation is therefore

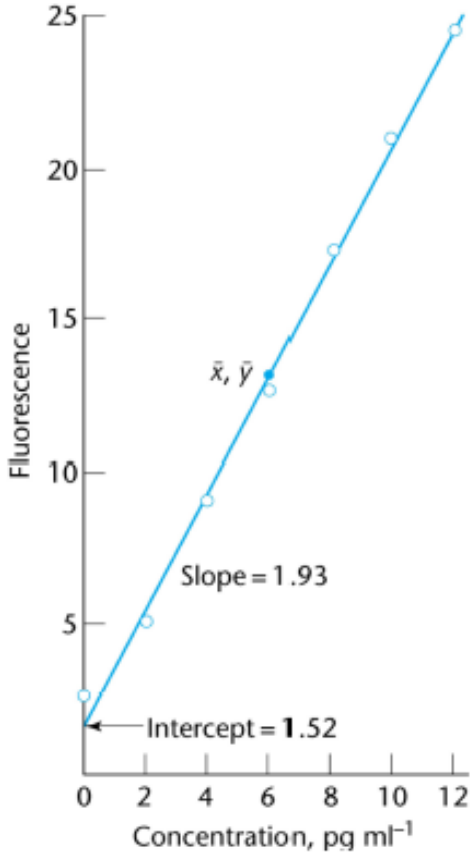
= 1.52 + 1.93X

Where

= Fluoresence

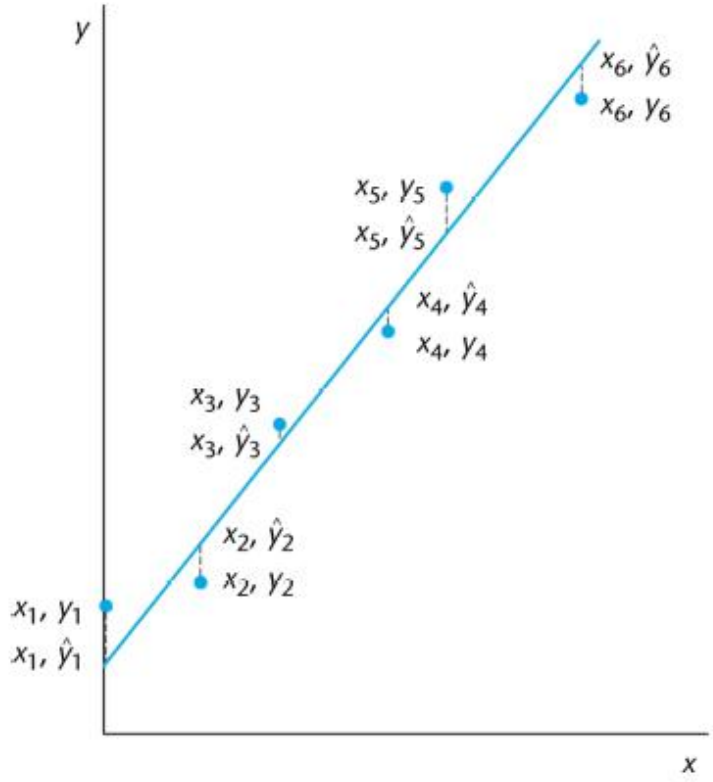
X = Concentration

The equation will specify the average magnitude of the expected change in Y given a change in X. The regression equation is often represented on a scatterplot by a regression line.



Residual

Residual (or error) represents unexplained (or residual) variation after fitting a regression model. It is the difference (or left over) between the observed value of the variable and the value suggested by the regression model.



For the example used in this class, the residuals are very small.