# Linear Regression - Data Analysis and Visualization

Dr. Farshid Alizadeh-Shabdiz

Spring 2021

# Agenda

- Scatterplots
- Correlation
- Simple Linear Regression

# Learning at The High Level

- Data and data collection
- Data segmentation and sampling
- Selecting a Model
- Selecting a fitness function
- Selecting optimization algorithm (not covered in this course)

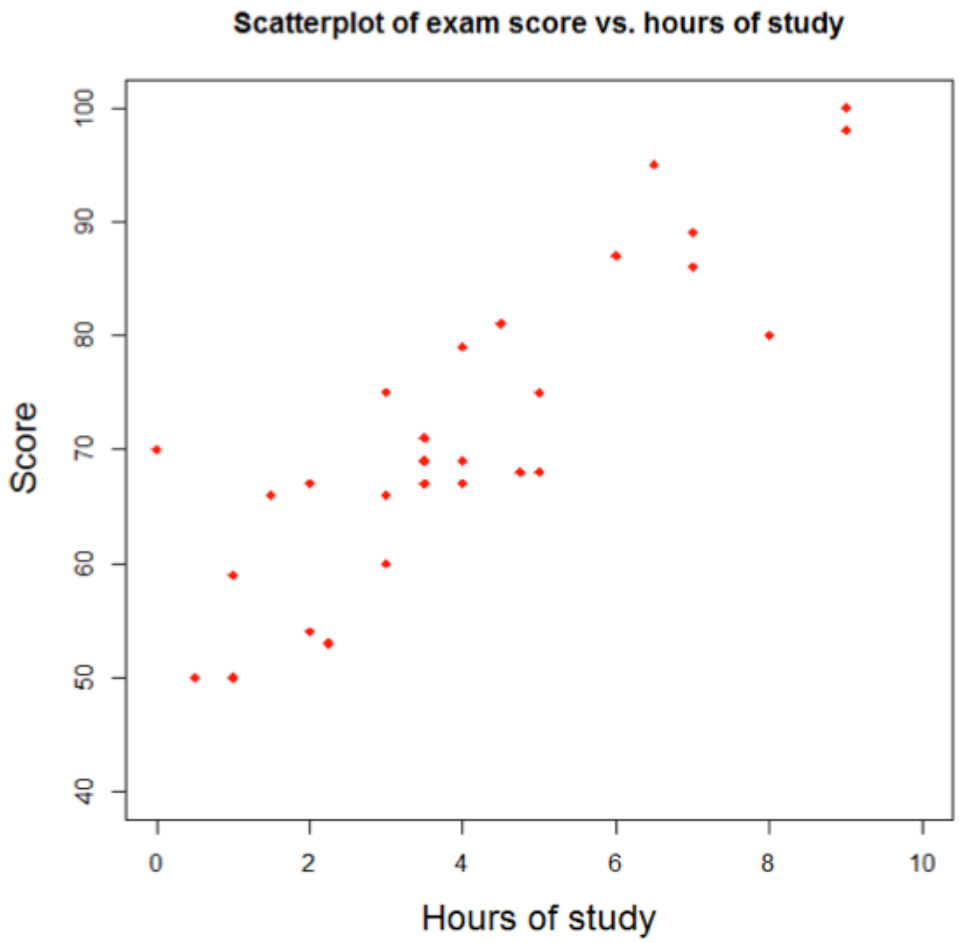
# Scatterplots

# Scatterplots

- Display the relationship between two **continuous or quantitative factors.**

- Shows the relationship between **two paired factors**

- **Each "pair" of data** is shown with one single point.



## An example - Scatterplots

Is there an association between the number of hours spent studying and the performance on the final exam?

**Use plot() function to draw the scatterplot**

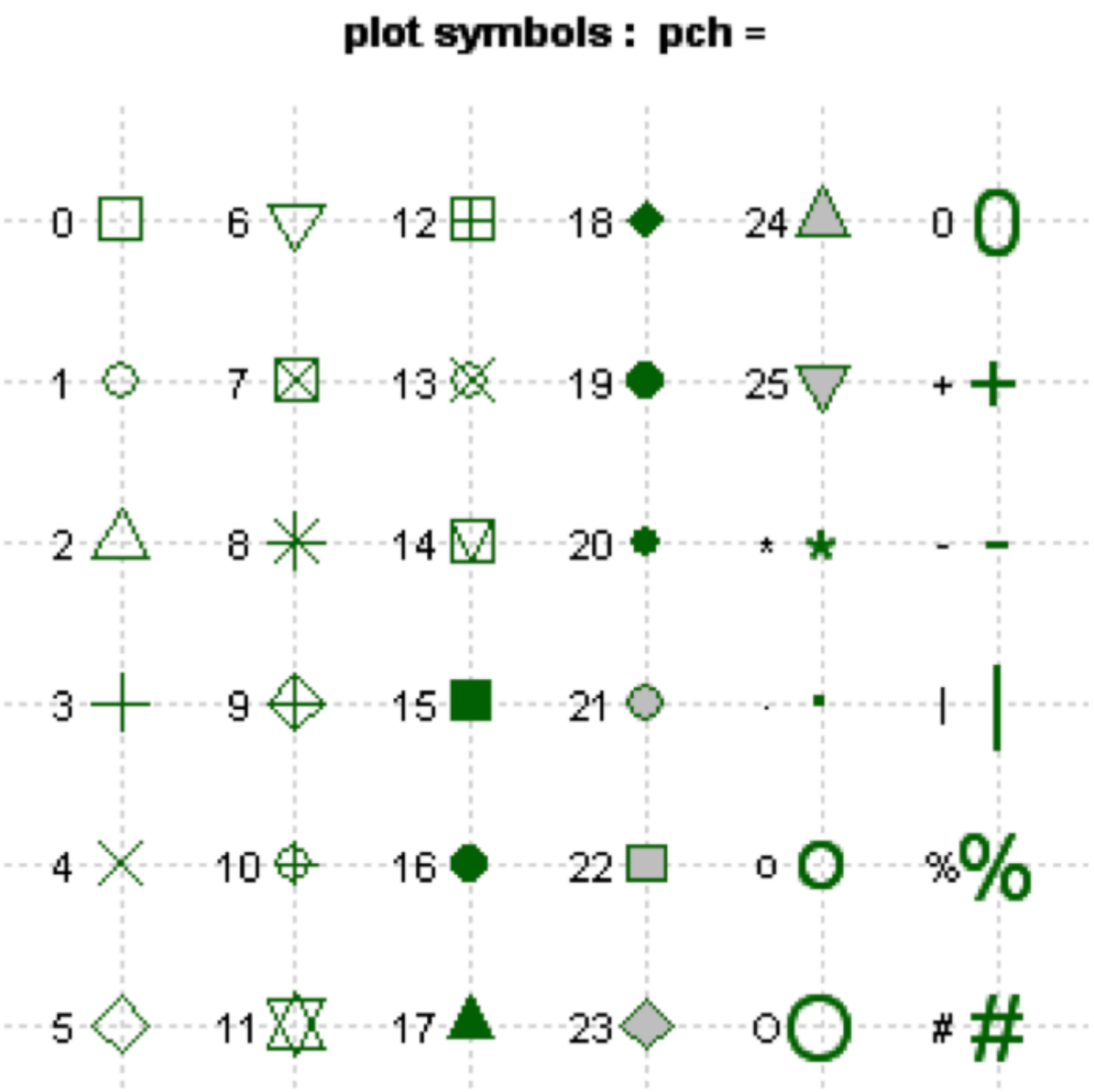
```
> plot(data$explanatoryvariable, data$responsevariable)
```

Many Attributes:

- Use **main, xlab, and ylab** to label the picture appropriately
- Use **xlim and ylim** to control x and y axises
- Change the type of point using **pch** and/or the color of the point using **col**
- Change the size of the points or the labels using cex, cex.axis, cex.main, etc.

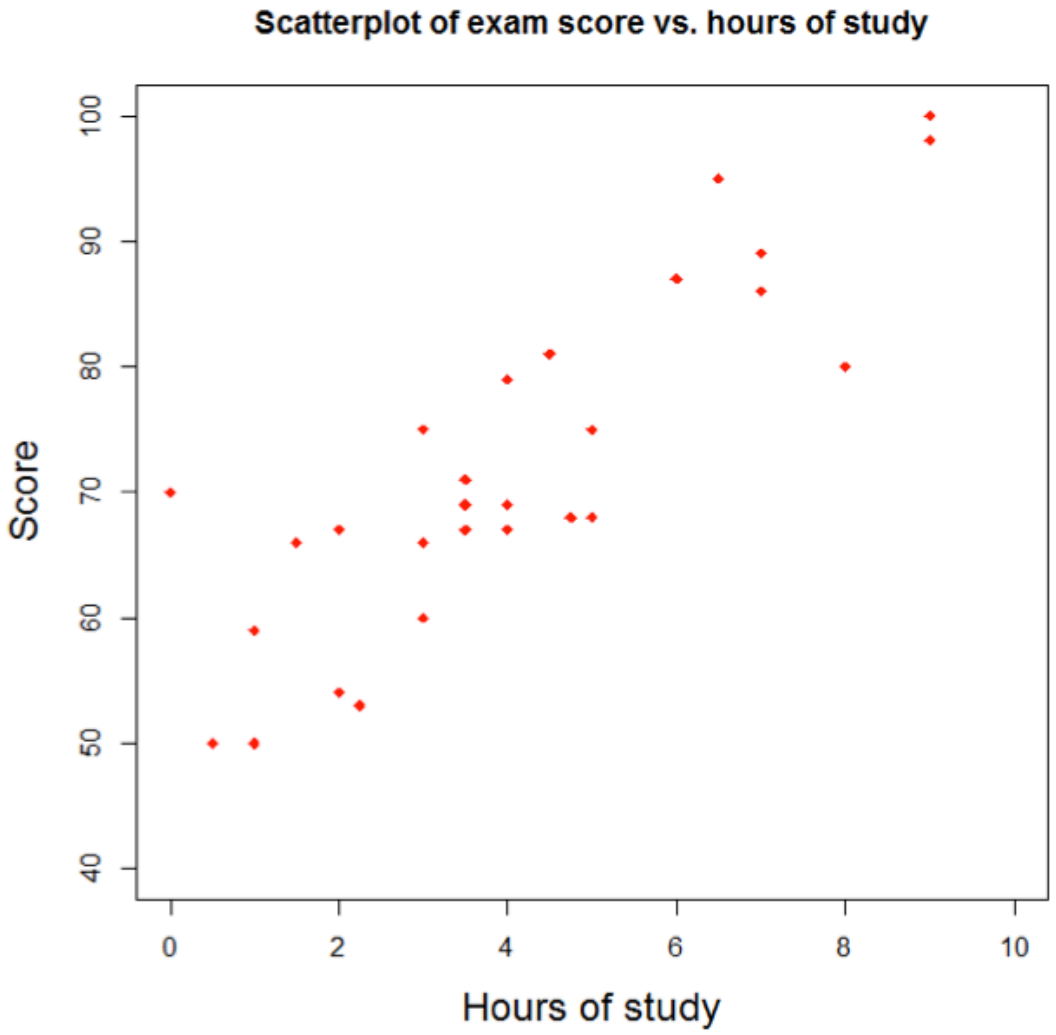
See **?plot()** for documentation

# Plot() options for pch



# Scatterplot Example

```
> attach(student)
> plot(study.hours, score, main="Scatterplot of exam score vs. hours of
    study", xlab="Hours of study", ylab="Score", xlim=c(0,10), ylim=c
    (40,100), pch=18, col="red", cex.lab=1.5)
```



# Response versus Explanatory Variables

- Generally, we put
  - the response variable (the outcome of interest) on the **y-axis** (vertical axis) and
  - the explanatory variable on the **x-axis**.
- **Explanatory variables** are also called **independent variables**.
- **Response variables** are called **dependent variables** (due to the fact that the response variable may depend on the explanatory variable).

## Response versus Explanatory Variables

- If there is **a temporal relationship** (if one variable comes before another) then the **explanatory variable is the generally the one that occurs first.**
- If there is **not a temporal relationship**, then to figure out which may be a better choice for the explanatory variable, you may ask yourself **which factor may depend on the other.**

## Example - Investing in an irrigation system

A farm in upstate New York is evaluating whether or not to invest in an irrigation system. They have data over the last few years on average rain fall (in inches during the growing season) and pounds of apples produced. If the amount of rain fall is associated with the crop yield, they may choose to make the investment.

**Which is the response variable and which is the explanatory variable in this case?**

Rain fall during the growing season occurs temporally **before the crops produce their yield**.

A choice is that the response variable is pounds of apples and the explanatory variable is average rain fall.

# Example - Are SAT math and verbal scores associated?

Are SAT math and verbal scores associated? To find out, a high school collects the math and verbal scores of their students.

**If they were to make a scatterplot of these data, which should go on the xaxis?**

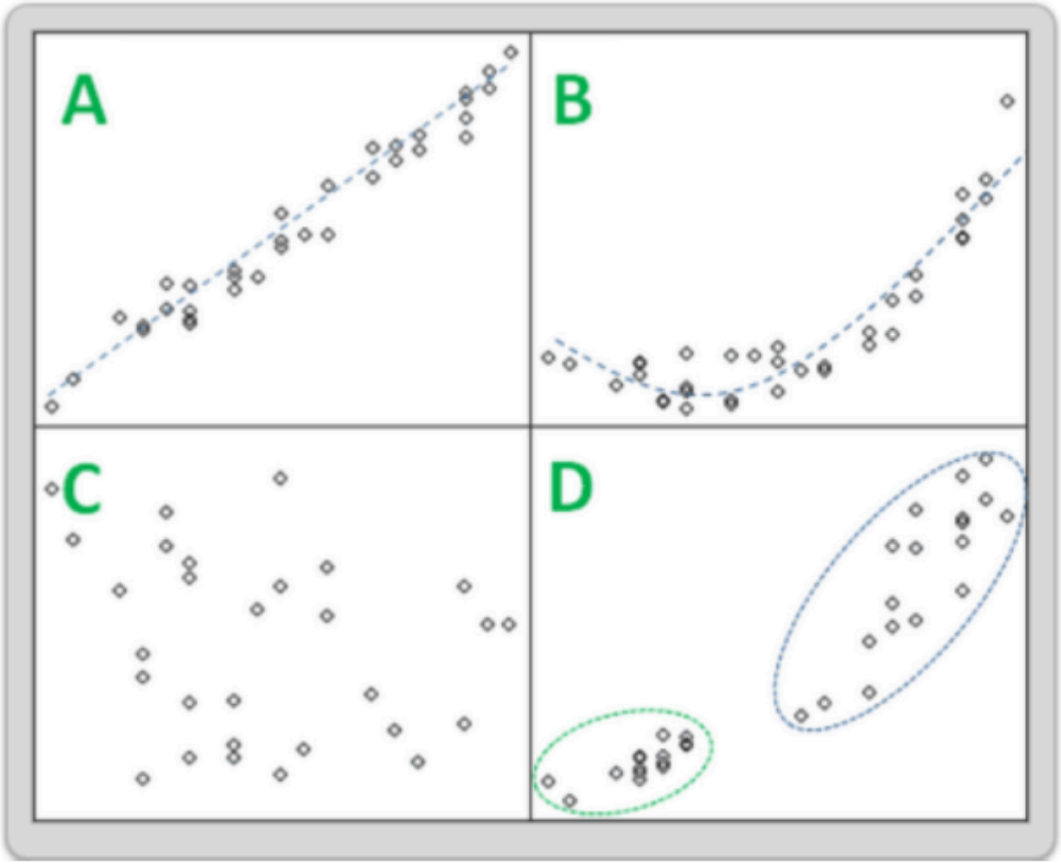
Generally the explanatory variable goes on the x axis of the scatterplot. However, in this case, **there is not a clear choice for the response and explanatory variables.**

There is no reason to think that ones SAT math score depends on ones SAT verbal score, or vice versa. As such, either score could go on the x axis.

# Interpreting Scatterplots - Form

**Relationships between variables:**

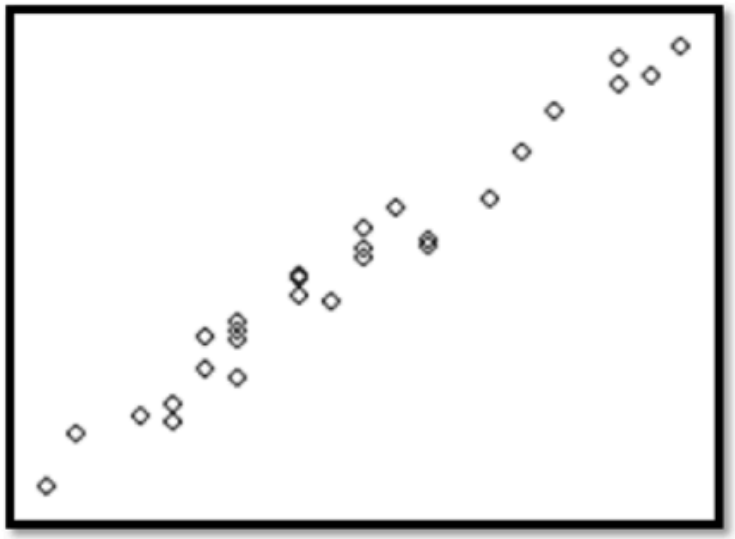
- **Linear** (where the points tend towards a straight line pattern)
- **Curved** (where the points tend toward a U-shape or arced pattern)
- **Random** (where the points don't seem to follow any pattern)
- **Clusters** may also be apparent



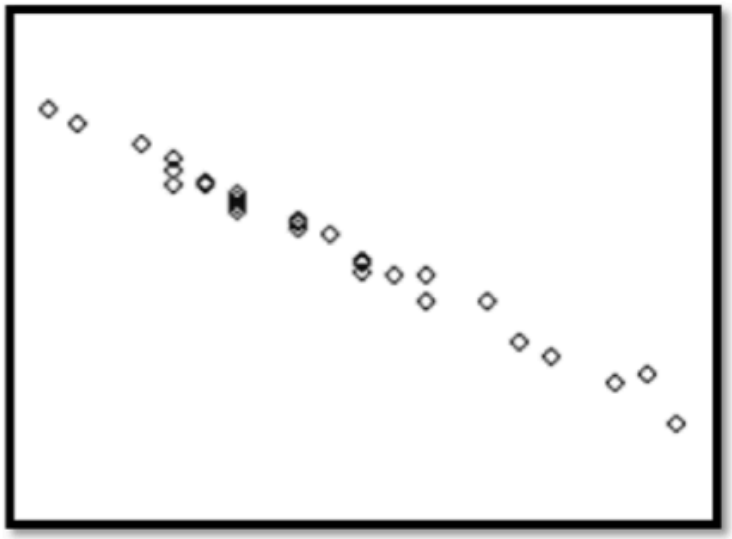
# Interpreting scatterplots - Direction

**The relationship between two factors is:**

- **"Positively Associated"** when as one factor increases in value the other factor also tends to increase in value
- **"Negatively Associated"** when as one factor increases in value the other factor tends to decrease in value

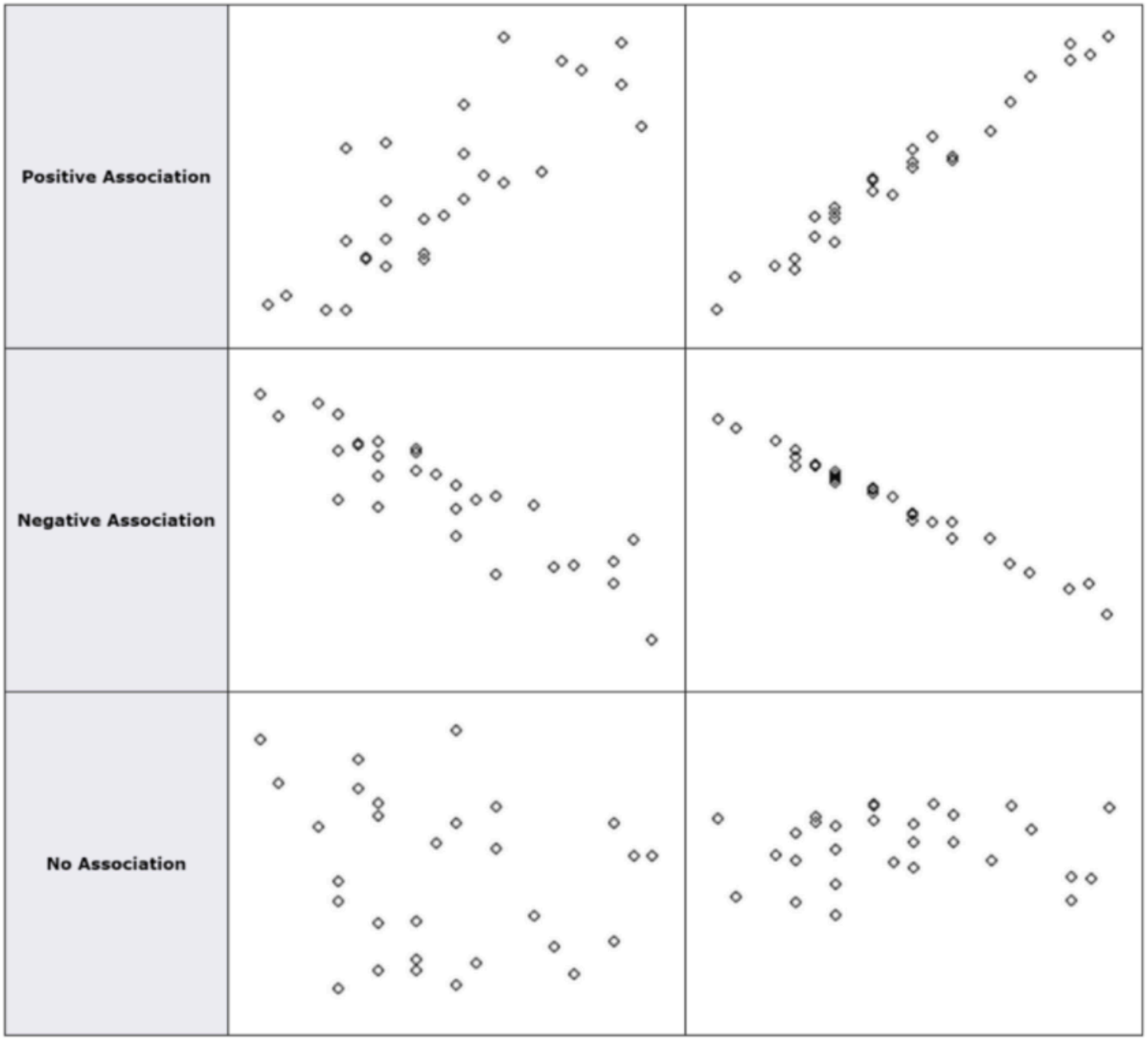


Positive Association



Negative Association

# Interpreting scatterplots - Direction

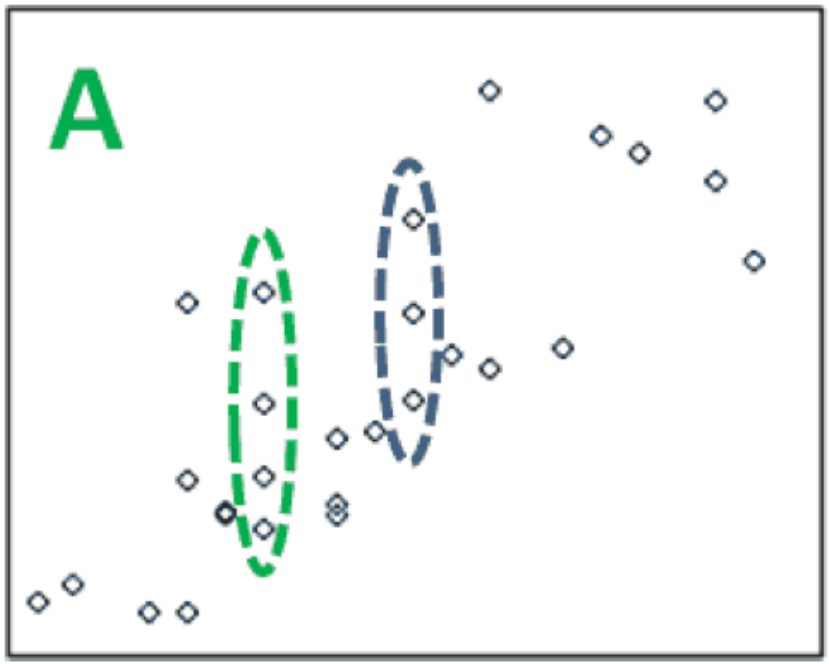


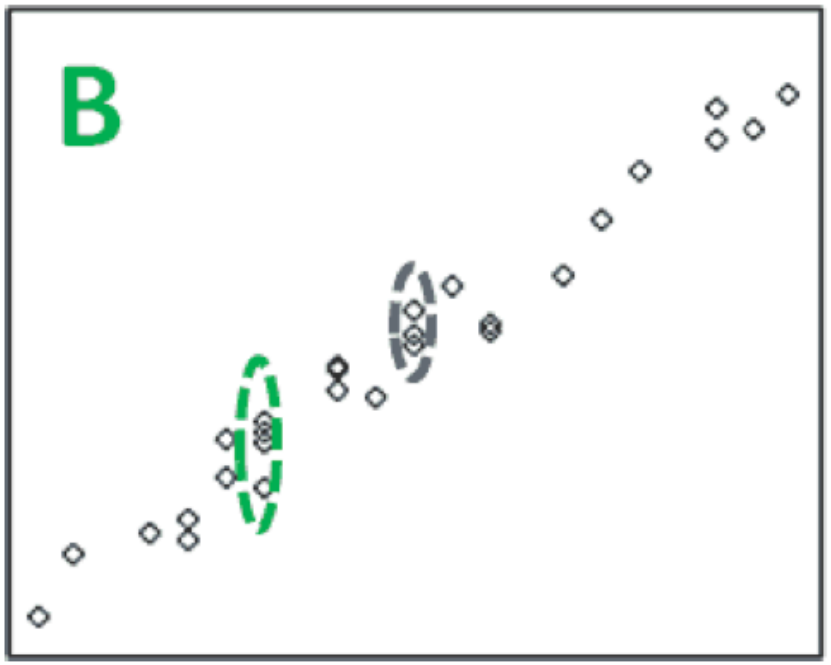
# Interpreting scatterplots - Strength of the Relationship

**The strength of the association** between the factors describes how closely the points appear to follow a clear form or pattern

Scatterplots A and B show the age in months vs length in centimeters of **baby koala bears at two different zoos** (A and B, respectively).

**The association between age in months and length in centimeters is more strongly** associated when looking at the data from **Zoo B** than the data from **Zoo A**.



# Correlation

## Correlation

**Correlation (denoted as $r$) or the correlation coefficient is a measure of the strength and direction of a linear relationship between two quantitative variables in a sample.**

The correlation (or the sample correlation coefficient) between two variables x and y can be computed using:

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

- $(x_i - \bar{x})/s_x$ gives the number of standard deviations from the mean that the $i$*th* observation of $x$ is.
- The sample correlation is an average of the product of the standardized $x$ and $y$ data points.

## Properties of Correlation

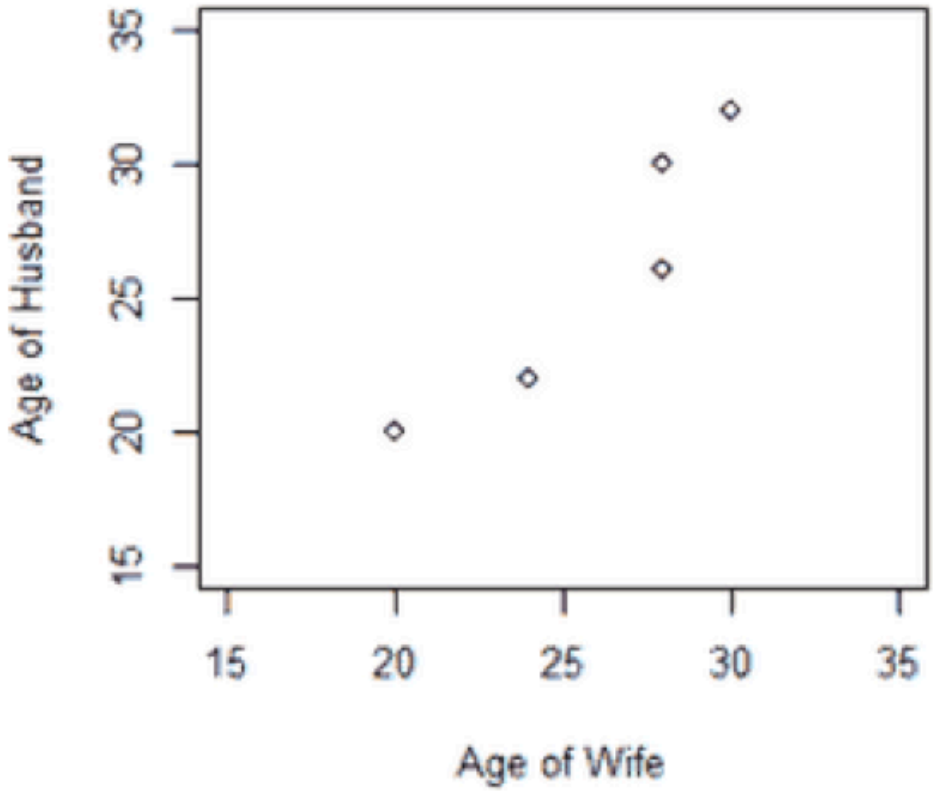
1. The **correlation** takes on **values between -1 and +1.**
2. The correlation **between variables x and y** is **the same as the correlation** between variables **y and x**.
3. Correlations can be computed between paired values of two **quantitative variables**.
4. The correlation coefficient **does not have units** and it is **independent of unit** of measure of variables x and y.
5. **Correlation measures the strength** of a linear relationship only. Correlation should **not be used to describe a curved** relationship - even if the association is strong.
6. **Outliers affect correlation.** Correlation in the presence of outliers should be interpreted with caution.

# An example - Correlation

We have data about age of 5 couples.

**Calculate the sample correlation between the ages of husbands and wives.**

| Couple | Age of Wife | Age of Husband |
|---|---|---|
| 1 | 20 | 20 |
| 2 | 30 | 32 |
| 3 | 24 | 22 |
| 4 | 28 | 26 |
| 5 | 28 | 30 |
| Sample mean | 26 | 26 |
| Sample standard deviation | 4.0 | 5.1 |

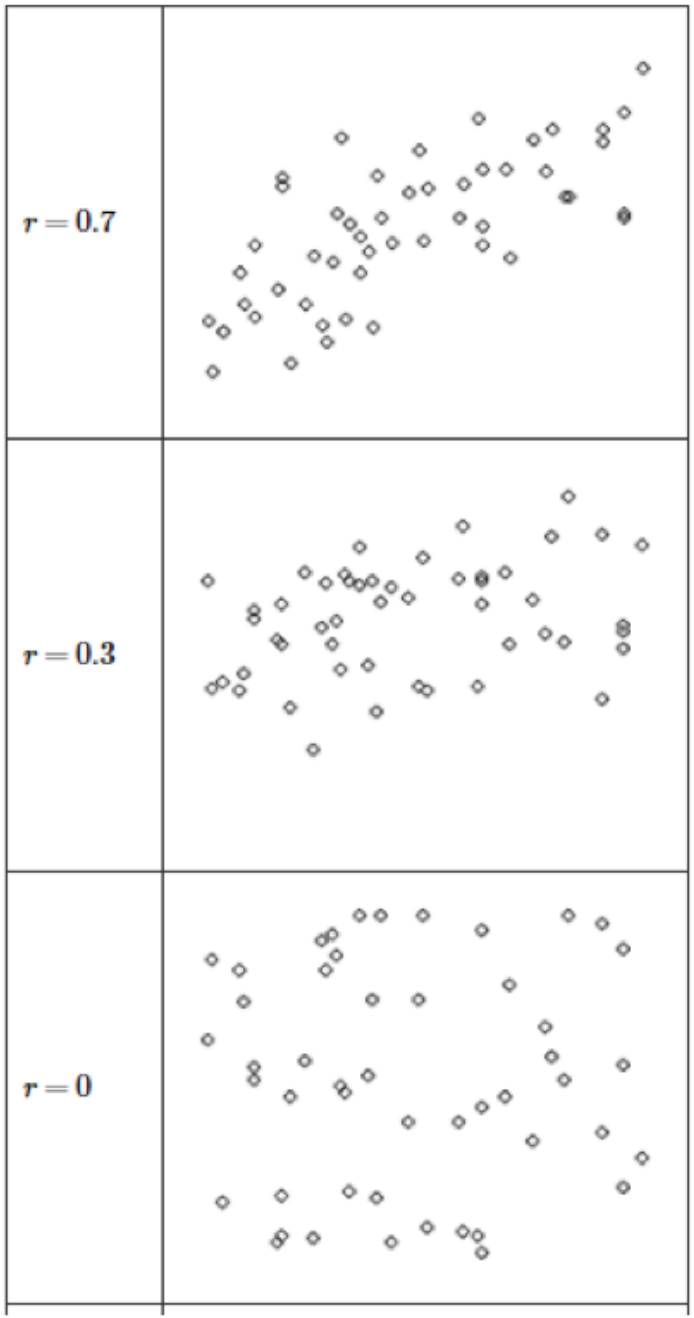


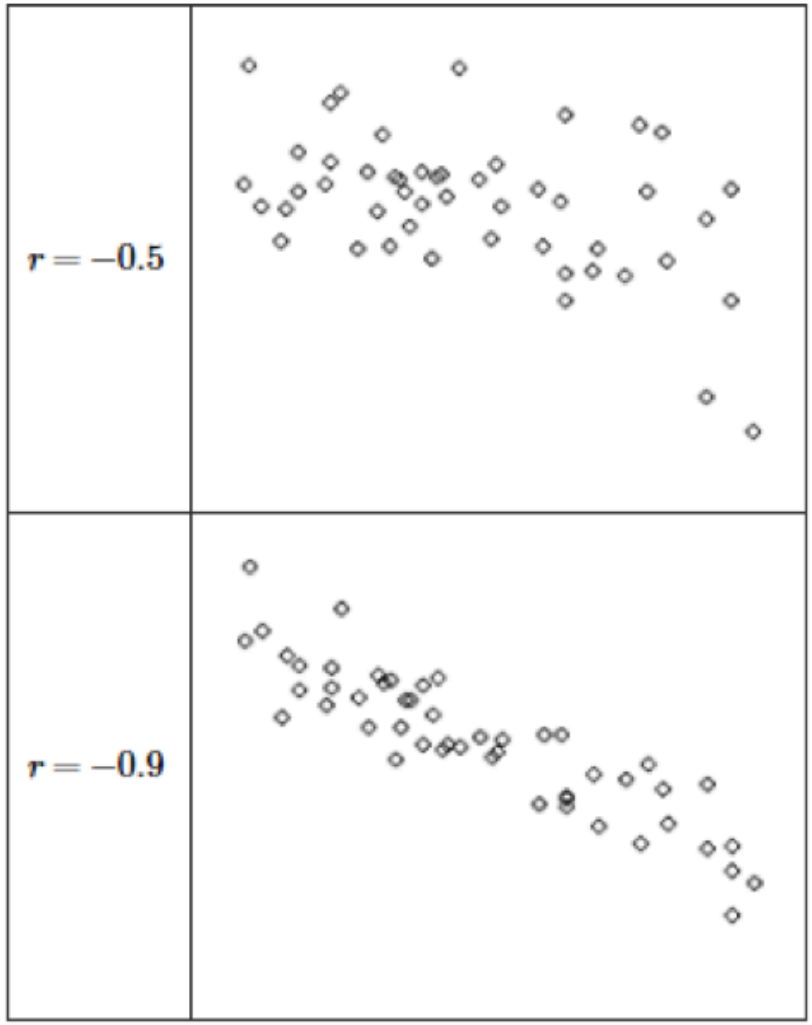
# An Example - Correlation

| Couple | Age of Wife | Age of Husband | Standardized Age of Wife $\left(\frac{x_i-\bar{x}}{s_x}\right)$ | Standardized Age of Husband $\left(\frac{y_i-\bar{y}}{s_y}\right)$ | $\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$ |
|---|---|---|---|---|---|
| 1 | 20 | 20 | $\frac{x_i-\bar{x}}{s_x}=\frac{20-26}{4.0}=\frac{-6}{4}$ | $\frac{y_i-\bar{y}}{s_y}=\frac{20-26}{5.1}=\frac{-6}{5.1}$ | $\left(\frac{-6}{4}\right)\left(\frac{-6}{5.1}\right)=\frac{36}{20.4}$ |
| 2 | 30 | 32 | $\frac{30-26}{4.0}=\frac{4}{4}=1$ | $\frac{32-26}{5.1}=\frac{6}{5.1}$ | $\left(\frac{4}{4}\right)\left(\frac{6}{5.1}\right)=\frac{24}{20.4}$ |
| 3 | 24 | 22 | $\frac{24-26}{4.0}=\frac{-2}{4}$ | $\frac{22-26}{5.1}=\frac{-4}{5.1}$ | $\left(\frac{-2}{4}\right)\left(\frac{-4}{5.1}\right)=\frac{8}{20.4}$ |
| 4 | 28 | 26 | $\frac{28-26}{4.0}=\frac{2}{4}$ | $\frac{26-26}{5.1}=\frac{0}{5.1}$ | $\left(\frac{2}{4}\right)\left(\frac{0}{5.1}\right)=\frac{0}{20.4}$ |
| 5 | 28 | 30 | $\frac{28-26}{4.0}=\frac{2}{4}$ | $\frac{30-26}{5.1}=\frac{4}{5.1}$ | $\left(\frac{2}{4}\right)\left(\frac{4}{5.1}\right)=\frac{8}{20.4}$ |

$$
\begin{aligned}
r &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) \\
&= \frac{1}{5-1}\left(\frac{36}{20.4}+\frac{24}{20.4}+\frac{8}{20.4}+\frac{0}{20.4}+\frac{8}{20.4}\right) \\
&= \frac{1}{4}\left(\frac{76}{20.4}\right) \\
&\approx 0.93
\end{aligned}
$$

# Correlation coefficients



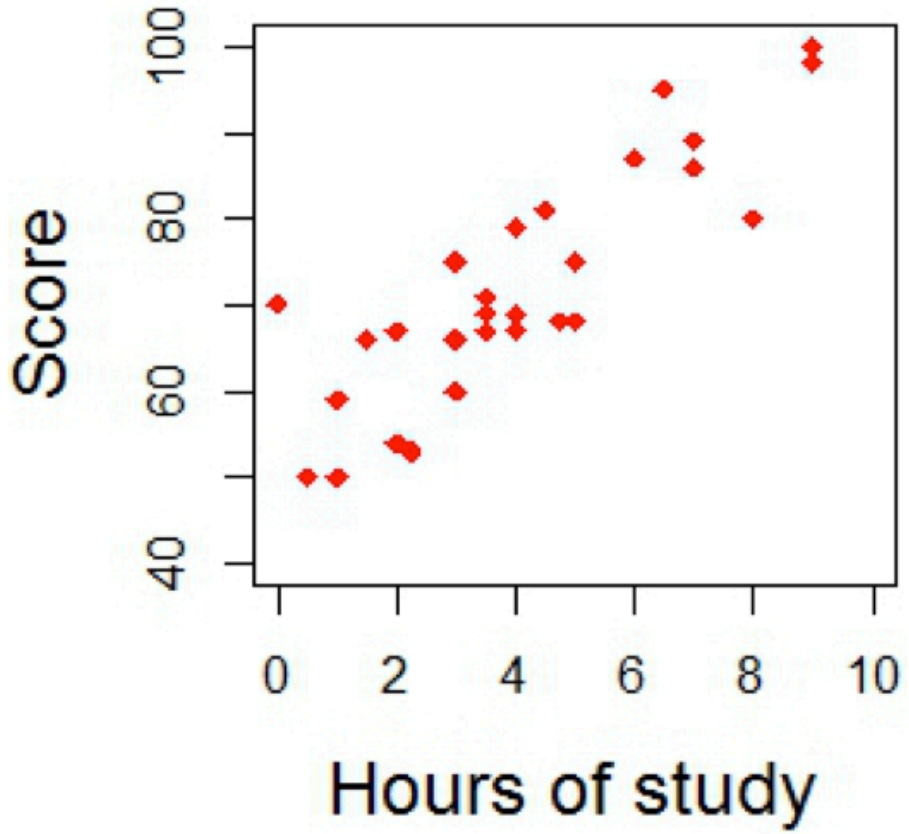



# R function cor()

Use **cor()** to calculate sample correlation coefficient

```
> cor(data$explanatoryvariable, data$responsevariable)

# calculate sample correlation
> cor(study.hours, score)
> cor(score, study.hours)
0.8835101
```



## Inference about population correlation coefficient

- Use sample data to make conclusions about **the population correlation coefficient.**
- **The Sample correlation, $r$, is a point estimate for the population correlation coefficient, $\rho$.**
- Formal tests of hypotheses concerning $\rho$ seek to determine **whether there is a linear association between the variables in the population.**

**$H_0 : \rho = 0$ (there is no linear association)**

**$H_1 : \rho \neq 0$ (there is a linear association)**

## Inference about population correlation coefficient

**$H_0 : \rho = 0$ (there is no linear association)**

**$H_1 : \rho \neq 0$ (there is a linear association)**

We use the test statistic:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

which follows a **t-distribution with $n-2$ degrees of freedom under $H_0$.**

- The decision rule for a two-sided level $\alpha$ test is:
  Reject $H_0 : \rho = 0$ if $t \geq t_{n-2,\frac{\alpha}{2}}$ or if $t \leq -t_{n-2,\frac{\alpha}{2}}$
  Otherwise do not reject $H_0 : \rho = 0$
- $t_{n-2,\frac{\alpha}{2}}$ is the value from the t-distribution table with $n-2$ degrees of freedom and associated with a right hand tail probability of $\alpha/2$.

## An Example - Inference

`Is there a linear relationship between hours of study and exam score?`

**Using the data we collected on the 31 students**, we can test the hypothesis that $H_0 : \rho = 0$ (no linear association) versus $H_1 : \rho \neq 0$ (linear association).

**1. Set up the hypotheses and select the alpha level**

**$H_0 : \rho = 0$ (there is no linear association)**

**$H_1 : \rho \neq 0$ (there is a linear association)**

$\alpha = 0.05$

**2. Select the appropriate test statistic**

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

# An Example - Inference

## 3. State the decision rule

Decision Rule: Reject $H_0$ if $p - value < \alpha$.

Otherwise do not reject $H_0$.

OR

Determine the appropriate value from the t-distribution table with $n - 2 = 31 - 2 = 29$ degrees of freedom and associated with a right hand tail probability of $\frac{\alpha}{2} = 0.025$

Using the table, $t_{n-2,\frac{\alpha}{2}} = t_{29,0.025} = 2.045$

Using R

```
> qt(0.975, df=29) = 2.04523
```

Decision Rule: Reject $H_0$ if $t \geq 2.045$ or if $t \leq -2.045$ $(|t| \geq 2.045)$.

Otherwise, do not reject $H_0$

## An Example - Inference

### 4. Compute the test statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.8835\sqrt{\frac{31-2}{1-0.8835^2}} \approx 10.16$$

### 5. Conclusion

- Reject $H_0$ since $10.16 \geq 2.045$.
- We have significant evidence at the $\alpha = 0.05$ level that $\rho \neq 0$.
- There is evidence of a significant linear association between study time and exam score.
- The sample correlation coefficient is 0.8835 indicating a strong positive association between study time and exam score. The positive correlation between these factors indicates that as study time increases, exam scores increase.

## R function cor.test()

Use cor.test() to perform testing

```
> cor.test(data$explanatoryvariable, data$responsevariable,
    alternative=[alternative], method=[method], conf.level=[confidence
        level])
```

**[alternative]** = **"two.sided"**, "less" (corresponds to negative association), or greater (corresponds to positive association)

**[method]** = **"pearson"**, "kendall", or "spearman"

**"pearson"**, the test statistic is based on Pearson's product moment correlation coefficient and follows a t distribution with df =n-2 (samples independent normal distributions).

If method is "kendall" or "spearman", Kendall's tau or Spearman's rho statistic is used to estimate a rank-based measure of association. These might be used if the data is not normal.

# Simple Linear Regression

# Regression

- Regression Function

$$f(x) = E(y \mid x = x_i)$$

- Regression Function minimizes mean squared error (MSE)

# Regression – How to calculate *f*

- The conditional probability cannot be calculated
- Relax the definition and let calculate the conditional probability for a small region
- Nearest Neighbor or local averaging (which also provide smooth solution)

$$f(x) = E(y \mid x \in [x_i - \Delta, x + \Delta])$$

# Linear Regression Function – Supervised Learning

- Parametric – assume a model for *f(x)*
  - *For example – linear model*

$$f(\mathbf{X}_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

- With p number of parameters, there will be (p+1) parameters to completely define the model
- Using the training data , estimate the unknown parameters of the model, e.g. $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.

# Fitness Function – Measuring Quality of Fitness

- Predicting Y by using the model f(x) with regard to mean squared prediction error
- So finding f(x) which minimizes the following function

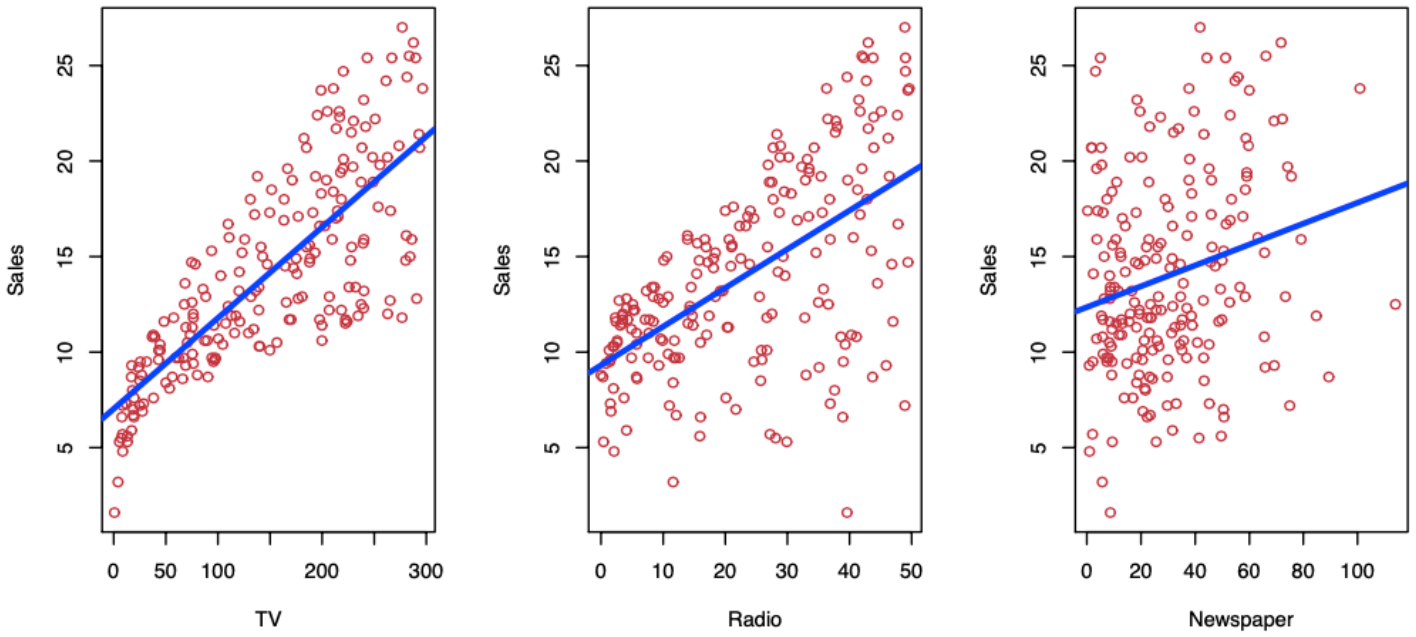
$$E[(Y - f(X))^2 | X = x]$$

- The most common measure of accuracy is Mean Squared Error (MSE)

# What is Statistical Learning

- Predicting sales as money spent on advertisement.

$$\mathrm{Y(sales\ \$)} = f(Xtv + XRadio + XNewspaper)$$



# Linear Regression Questions

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contributes to sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

# Linear Regression – Closed form

- Linear model

$$y = f(x) = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_m X_m$$

Closed form answer – Normal Equation

$$A = (X^T X)^{-1} X^T Y$$

- Fitness function: Mean Squared Error (MSE)

## Simple Linear Regression (SLR)

If a linear relationship exists, **we can model** the nature of the relationship between the variables using **simple linear regression (SLR)**.

Using SLR, we assert a **straight line on the scatterplot** that represents the best fitting line to the data that captures the pattern of the relationship.

- An explanatory variable.
- A response variable

## Simple Linear Regression (SLR)

We build a data model that allows us to:

- **quantify the relationship between the response variable and the explanatory variable**
- **predict the response of a new observation with a given value for x** or what the average response is for observations with a specific value for the explanatory variable

## Simple Linear Regression (SLR)

The equation for the simple linear regression line is given by

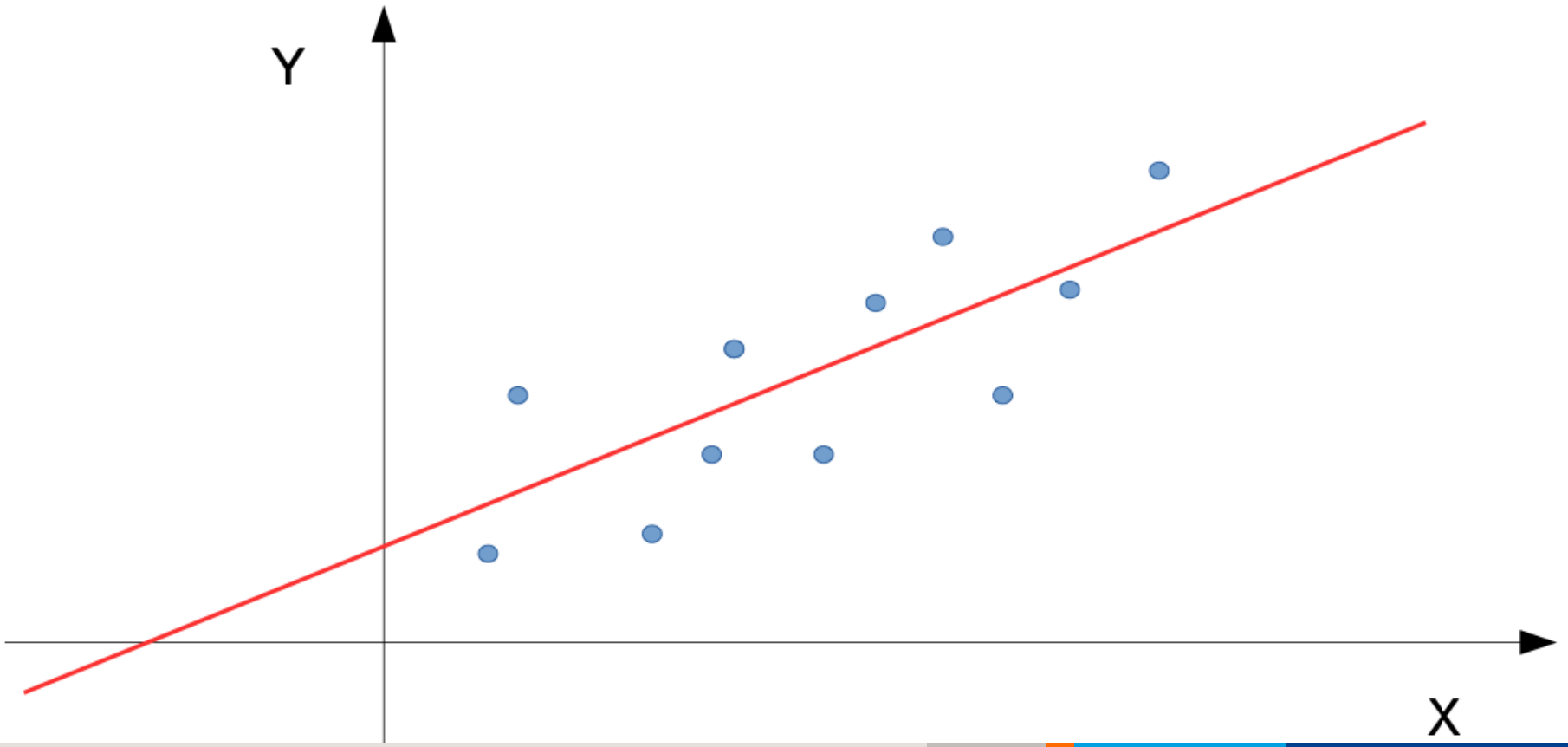
$$y = \beta_0 + \beta_1 x$$

- $y$ is the response or dependent variable
- $x$ is the explanatory or independent variable
- **$\beta_0$ is the intercept (the value of $y$ when $x = 0$)**
- **$\beta_1$ is the slope (the expected change in $y$ for each one-unit change in $x$)**

## Simple Linear Regression (SLR)

The equation for the simple linear regression line is given by

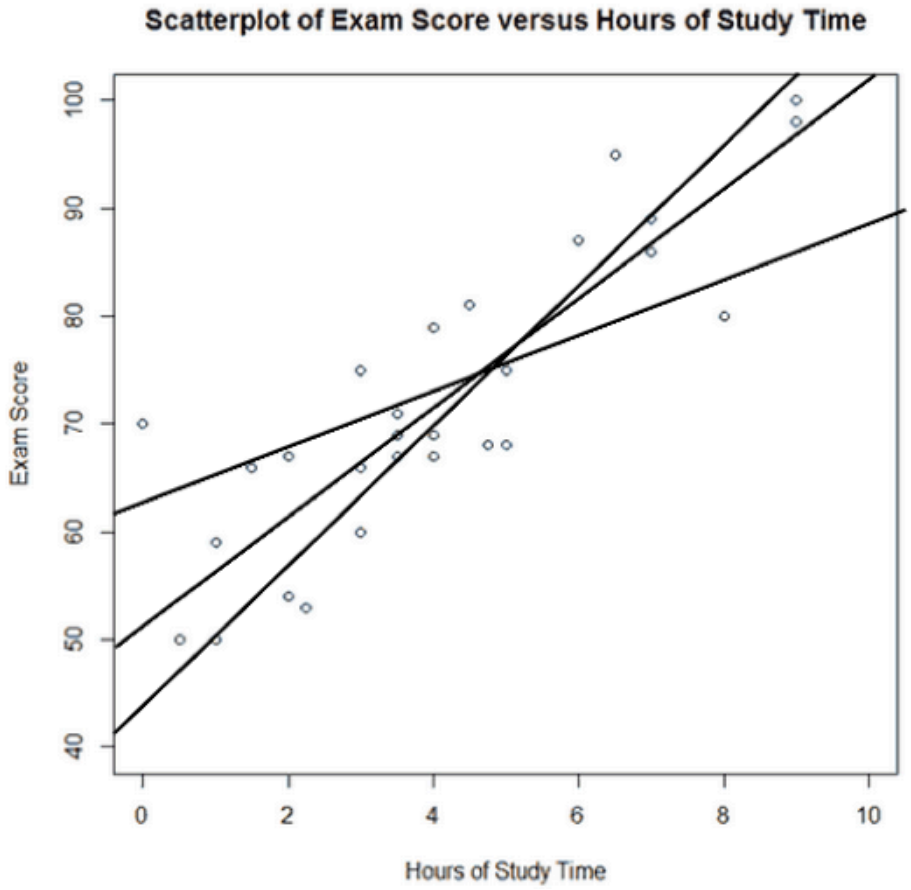
$$y = \beta_0 + \beta_1 x$$



# How to find the regression line that best fits the data

There are several ways to find parameters of the line. The most common way is to **minimize the sum of the squares** of the distances between the points and the regression line.

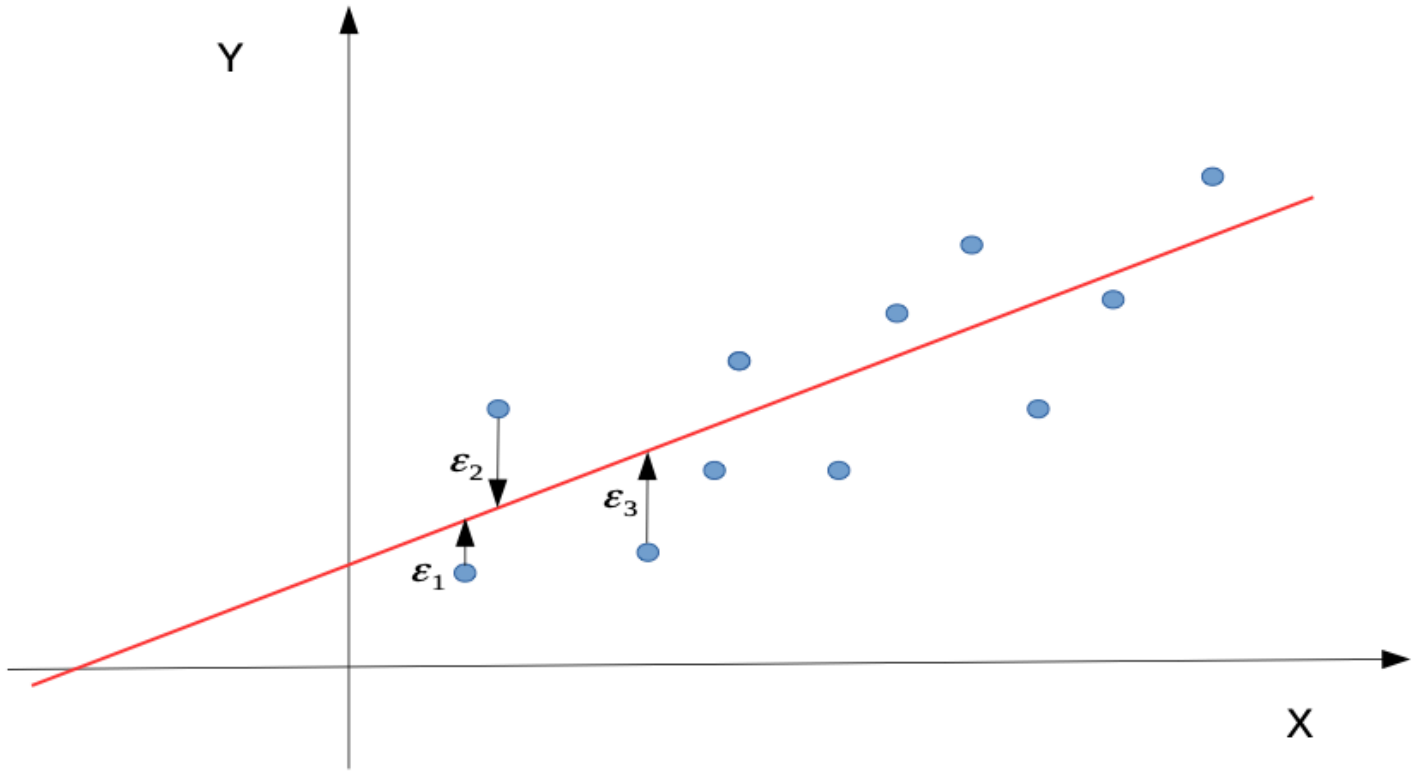
This approach is called the **least-squares method**. We want to minimize the vertical distance between each of the points and the regression line.



## How to find the regression line that best fits the data

There are several ways to find parameters of the line. The most common way is to **minimize the sum of the squares** of the distances between the points and the regression line.

This approach is called the **least-squares method**. We want to minimize the vertical distance between each of the points and the regression line.



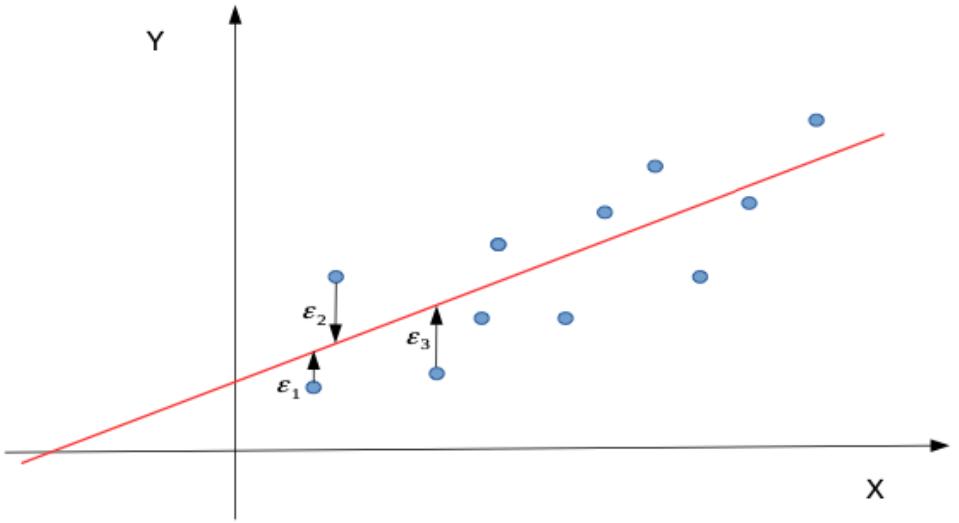
# How to find the regression line that best fits the data

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} \left( y_i - (\beta_0 + \beta_1 x_i) \right)^2$$

Loss function, l2 norm



## How to find the regression line that best fits the data

$$y_i = \beta_0 + \beta_1 x_i + \epsilon$$

$$\text{Best Fit Slope: } \hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\text{Best Fit Y-Intercept: } \hat{\beta}_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

## Equation for the least-squares regression line

The equation for the simple linear regression line is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

In the least-squares regression, the estimates of $\beta_0$ and $\beta_1$ are:

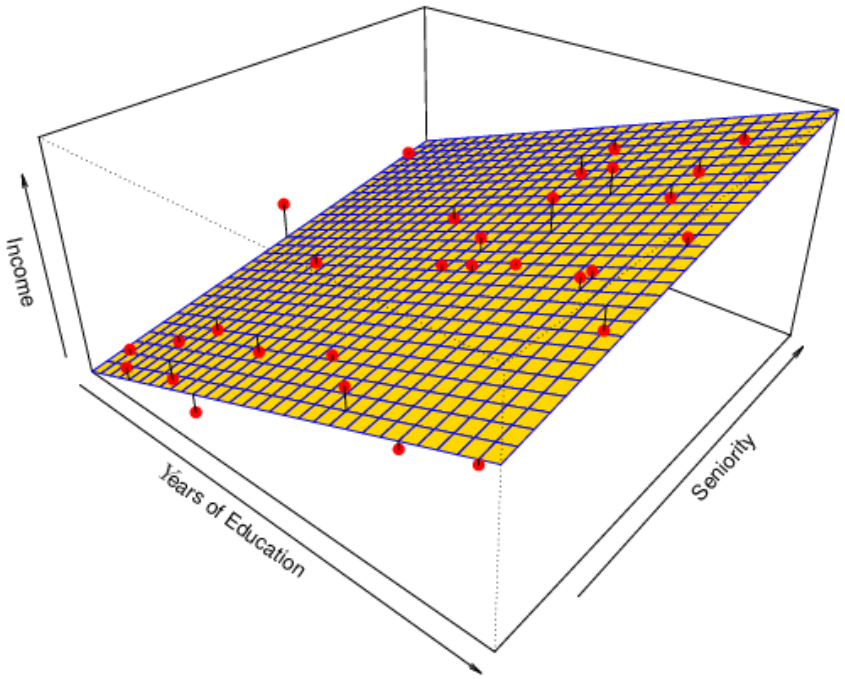
$$\hat{\beta}_1 = r\frac{s_y}{s_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$r$ is correlation coefficient

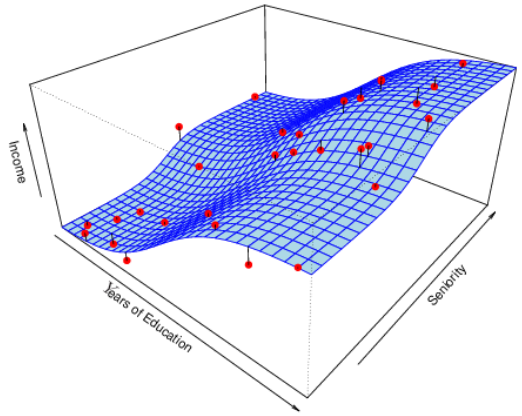
$S_x$ the sample standard deviation of $x$, $S_y$ is SD of $y$

$\bar{x}$ sample mean of $x$, and $\bar{y}$ sample mean of $y$




- Linear Regression Model fit to Income vs Education & Seniority

$$f = \beta_0 + \beta_1 \times Education + \beta_2 \times Seniority$$



# R Function - Simple linear regression

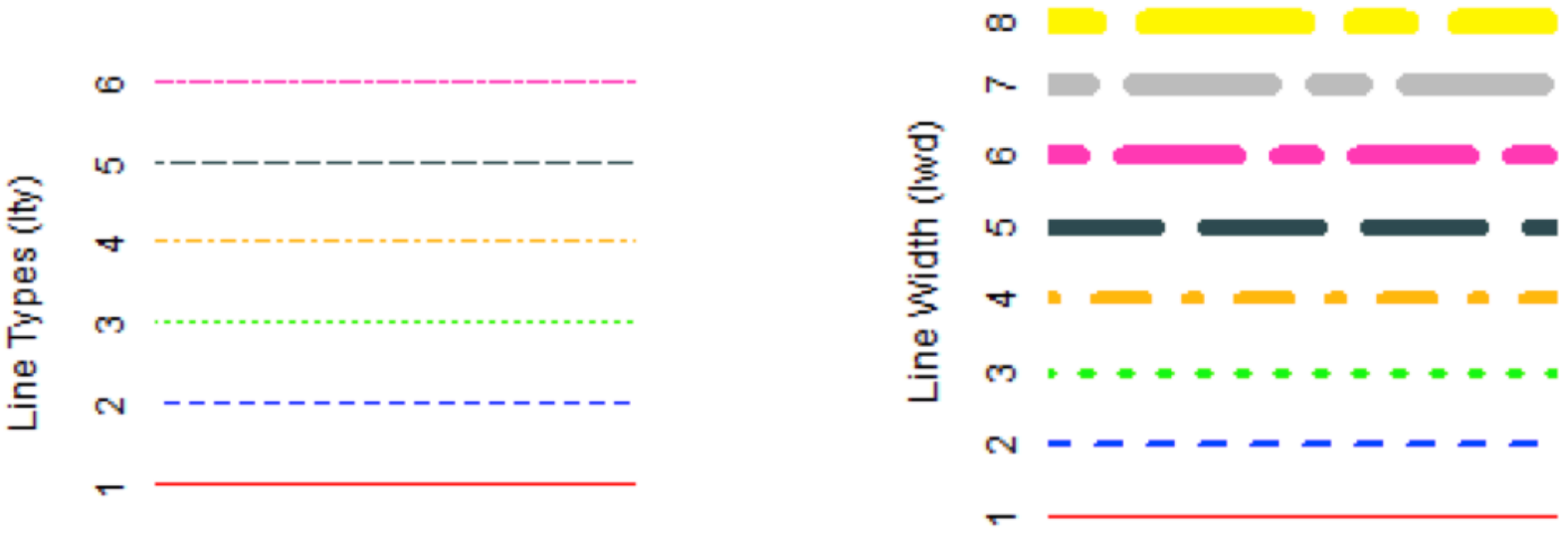
```
> lm(data$responsevariable ~ data$explanatory)

> abline(a=intercept, b=slope)

# create a model object
> m <- lm(score ~ study.hours)

# Add regression line to the scatterplot
> abline(m, lty=3, col="blue")
```

http://www.cookbook-r.com/Graphs/Shapes_and_line_types



# An example - the least-squares regression line

National Unemployment Male Vs. Female Reference: Statistical Abstract of the United States

```
> unemployment <- read.csv("national_unemployment_rate.csv")
> attach(unemployment)

> xbar <- mean(male.unemployment.rate)
> sx <- sd(male.unemployment.rate)
> ybar <- mean(female.unemployment.rate)
> sy <- sd(female.unemployment.rate)
> r <- cor(male.unemployment.rate, female.unemployment.rate)
> beta1 <- r*sy/sx
> beta0 <- ybar - beta1*xbar
```

$$\hat{\beta}_1 = 0.69$$

$$\hat{\beta}_0 = 1.43$$

$$\hat{y} = 1.43 + 0.69x$$

When $x = \bar{x} = 5.95$, $\hat{y} = \bar{y} = 5.57$