CS555 B1 Data Analysis and Visualization

Scatterplots, Correlation, Linear Regression

Lecture 5

Kia Teymourian

Significance level α, confidence level and p value

Before we run any statistical test, we must first determine an alpha level, also called the "significance level."

Confidence level + alpha = 1 or **Confidence level = 1 - alpha**

The **alpha level** is the **probability of rejecting the null hypothesis** when the null hypothesis is **true**.

Statistically speaking, the p-value is the probability of obtaining a result as extreme as, or more extreme than, the result actually obtained when the null hypothesis is true.

Confidence intervals are ranges based on sample data that provides us with a range of values that the population parameter is likely to be in with a specified level of certainty.

In order to decide whether or not to reject the null hypothesis, we can either compare:

- the p-value to a pre-defined significance level
- the test statistic to a critical value

These two criteria will always lead to the same conclusion.

Scatterplots

A very useful way to graphically display the relationship between two continuous or quantitative factors.

Show the relationship between **two “paired” factors**

The values of one factor are shown on the **horizontal axis (x-axis)** while values for the other factor are shown on the **vertical axis (y-axis)**

Each “pair” of data is shown in the graph 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

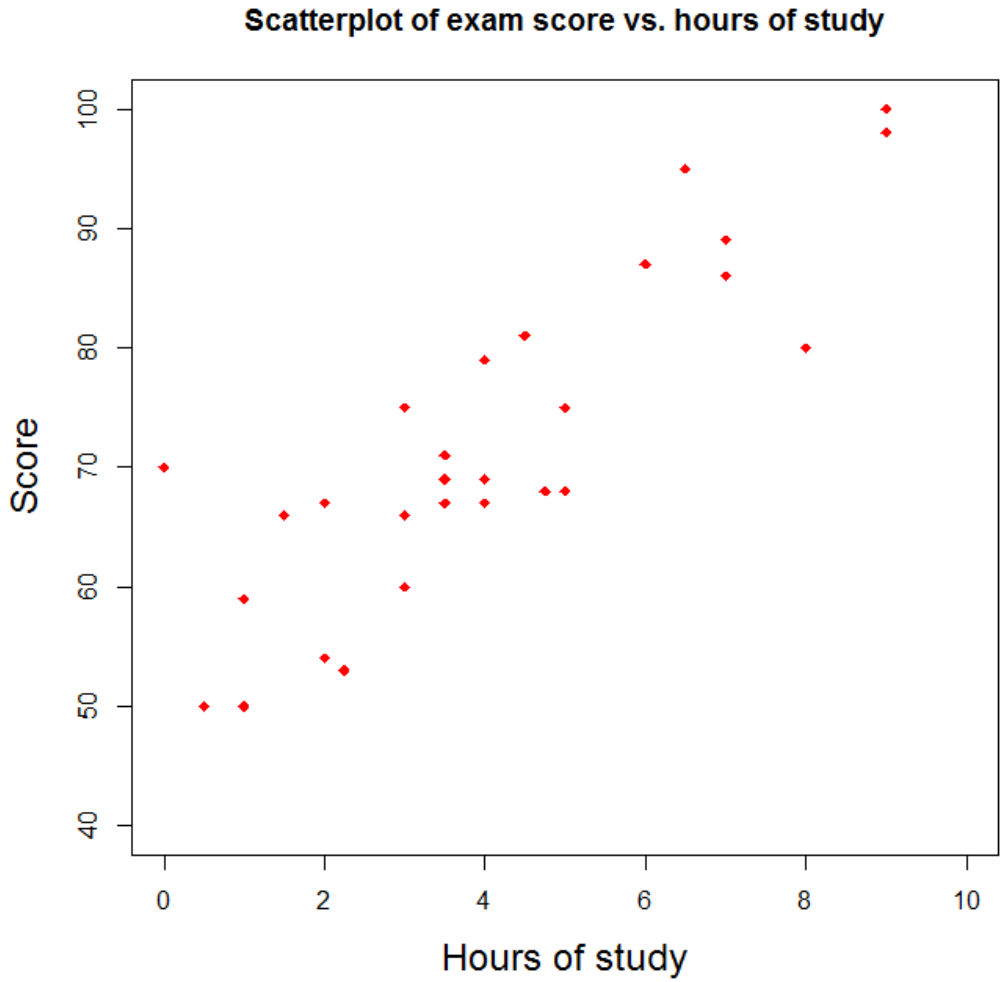
> plot(data$explanatoryvariable, data$responsevariable)
- Use *main, xlab,* and *ylab* to label the picture appropriately
- Use *xlim* and *ylim* to control x and y axises
- Change the type of point using *pch* and/or the color of the point using *col*
- Change the size of the points or the labels using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

Cex – the number indicating the amount by which plotting text and symbols should be scaled relative to the default. 1=default, 1.5 is 50% larger, 0.5 is 50% smaller, etc.

cex.axis magnification of axis annotation relative to cex
cex.lab magnification of x and y labels relative to cex
cex.main magnification of titles relative to cex
cex.sub magnification of subtitles relative to cex

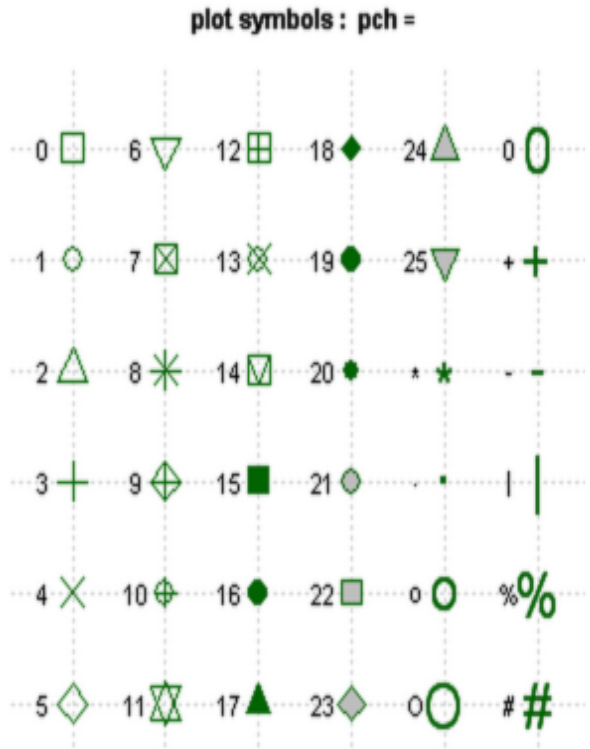
The scatterplot

```
> student <- read.csv("student.csv")
> 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)
```



Plot() options for pch

border color (col=) and fill color (bg=).



Response versus Explanatory Variables

- Generally, we will put the response variable (the outcome of interest) on the y-axis (vertical axis) and the explanatory variable (one that may explain or influence changes on the response variable) on the x-axis (horizontal 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).

- 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 1

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” (similar to tons per hectare), 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. As such, a natural choice is that the response variable is pounds of apples and the explanatory variable is average rain fall.

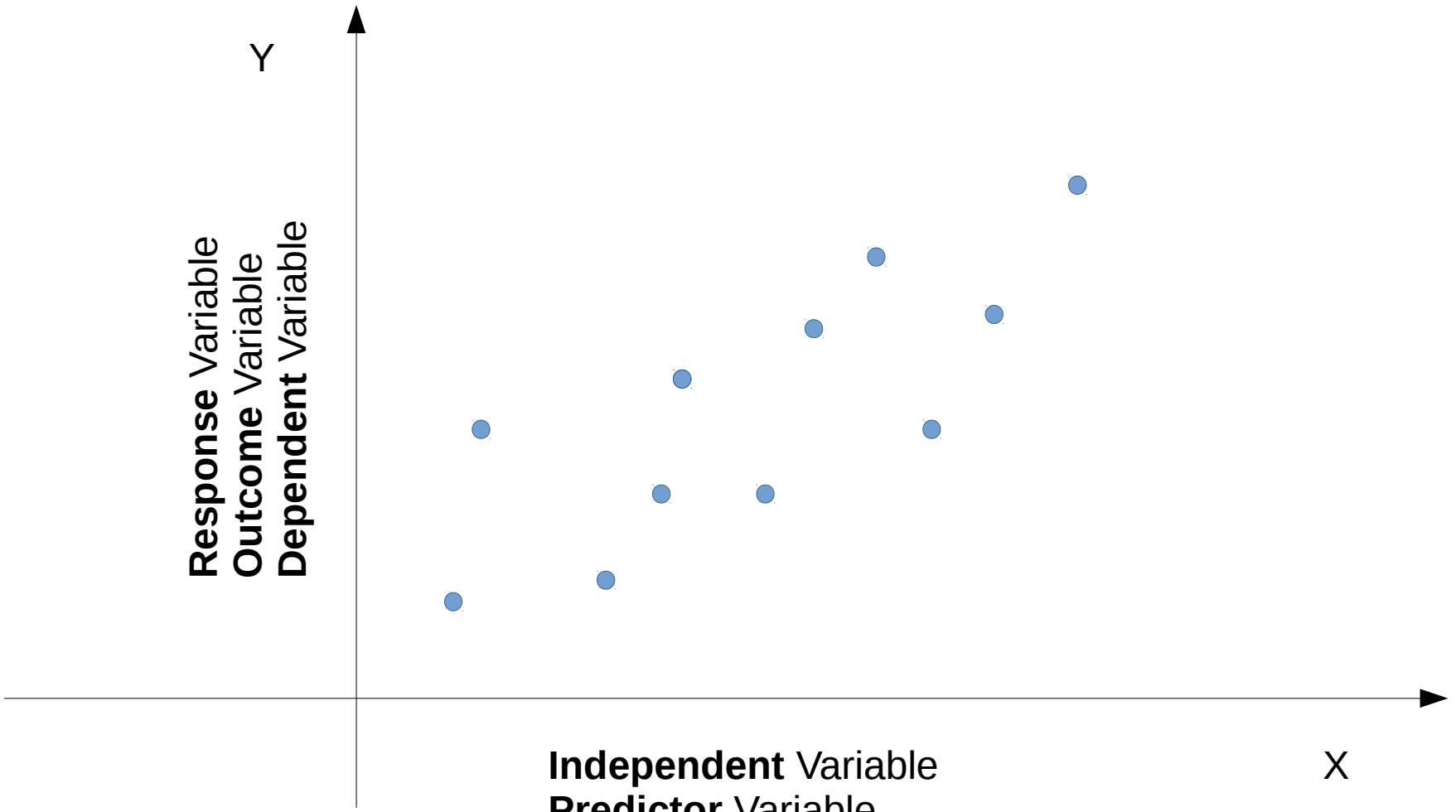
Example 2

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 x-axis?

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 one's SAT math score depends on one's SAT verbal score, or vice versa. As such, either score could go on the x axis.

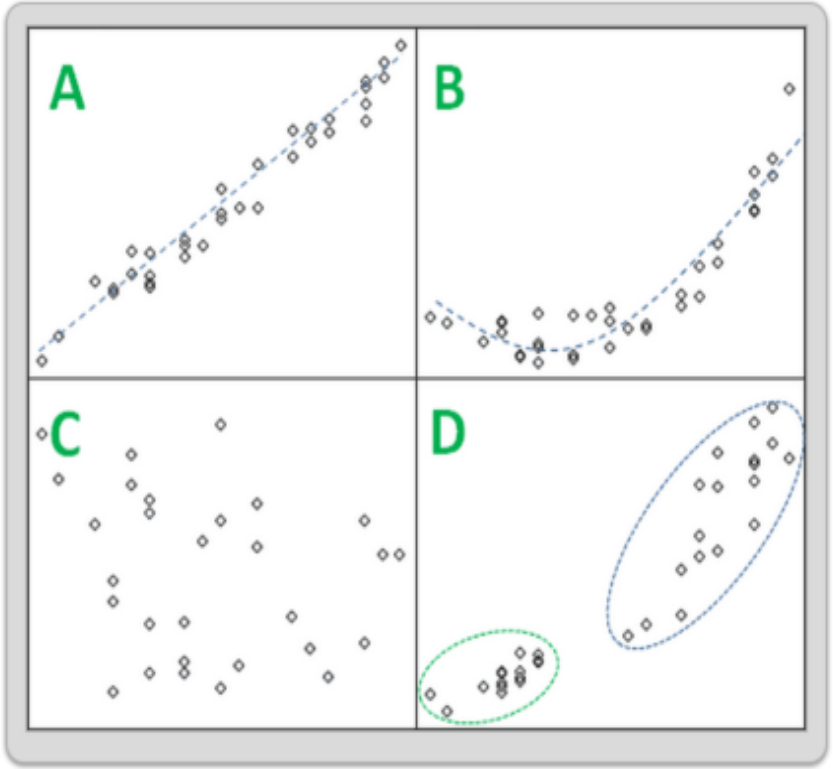
Variables & Observations



Interpreting scatterplots - Form

Relationships between variables:

A) linear (where the points tend towards a straight line pattern)
B) curved (where the points tend toward a U-shape or arced pattern)
C) random (where the points don't seem to follow any pattern)
D) clusters may also be apparent

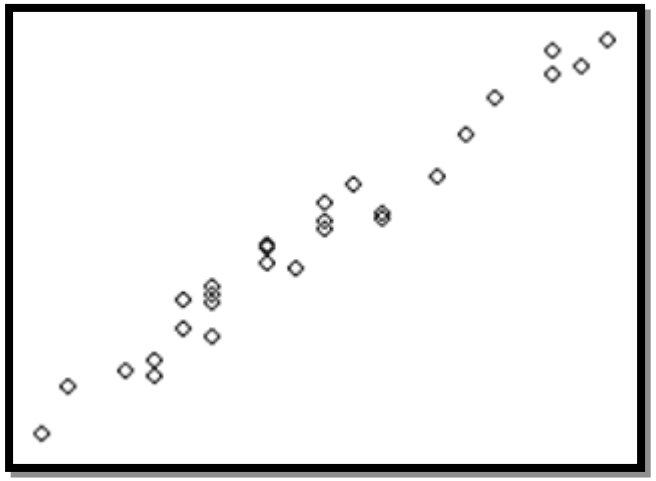


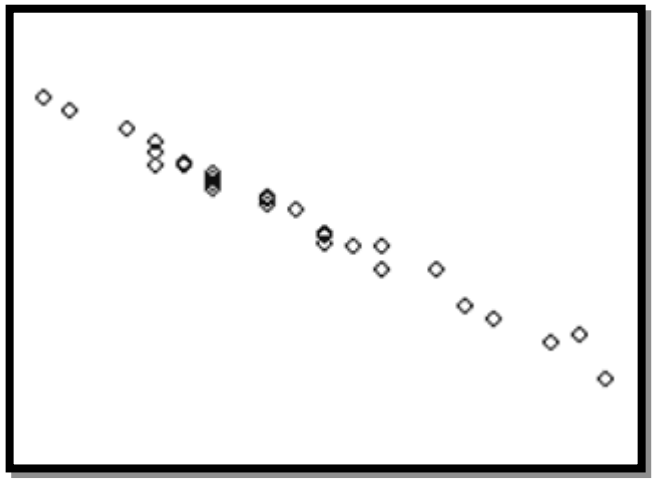
Test Yourself 3.2

Interpreting scatterplots - Direction

The relationship between two factors is:

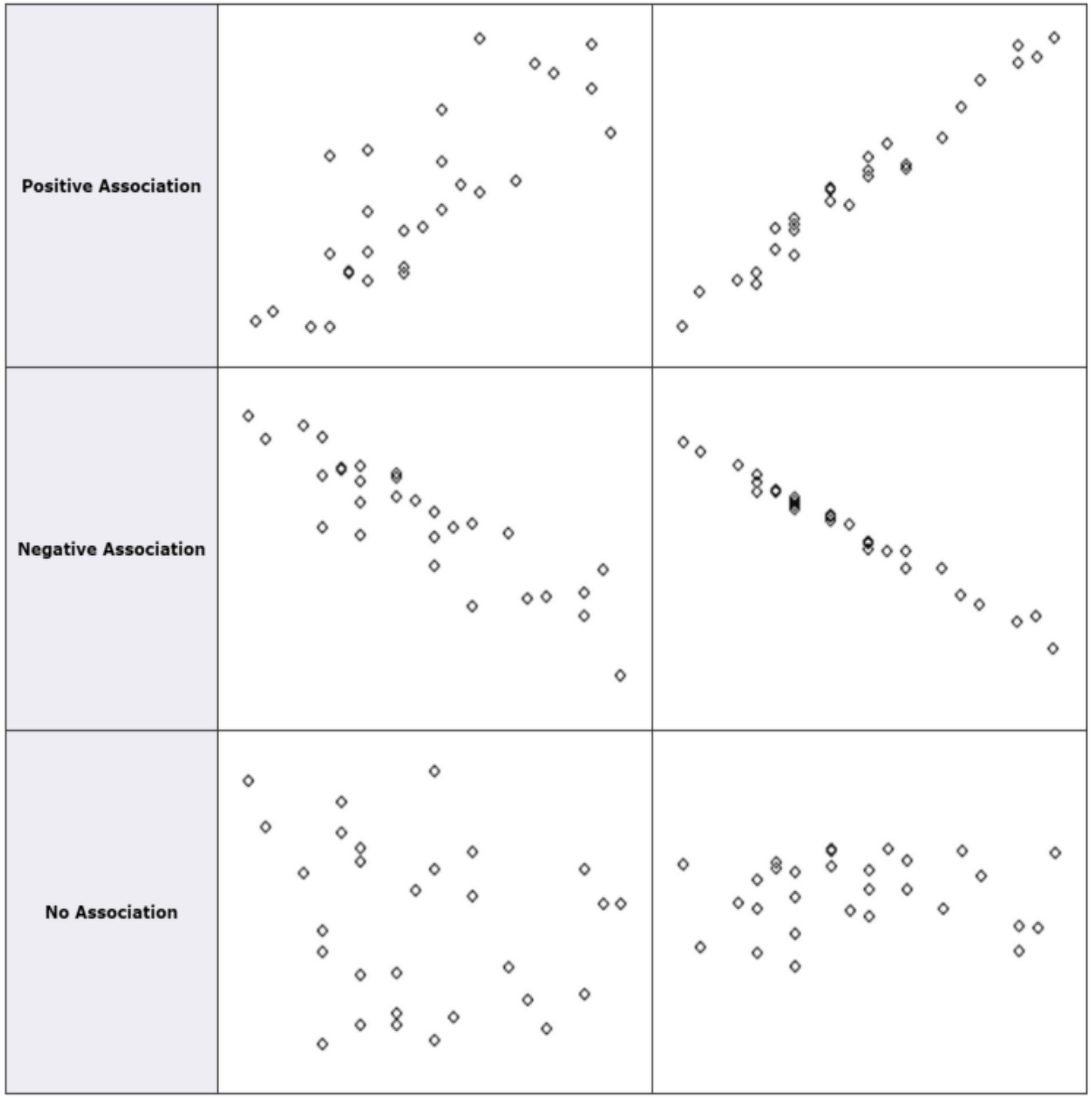
- "**positively associated**" when as one factor increases in value the other factor also tends to increase in value
- "**negatively associated**" when as one factor increases in value the other factor tends to decrease in value
- The scatterplot of two positively associated variables tends to look like points hovering around a line with a positive slope
- The scatterplot of two negatively associated variables tends to look like points hovering around a line with a negative slope





Interpreting scatterplots - Direction

ictors that appear to not be associated.

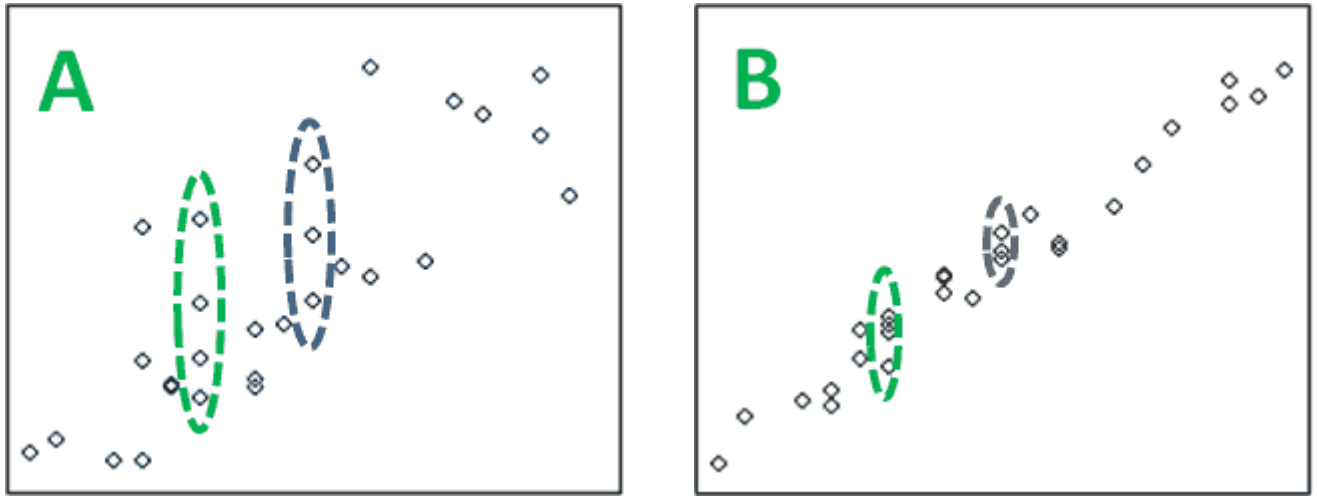


Interpreting scatterplots - Strength of the Relationship

- 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). Let's also assume that the scale of the graphs is the same.

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 (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:

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

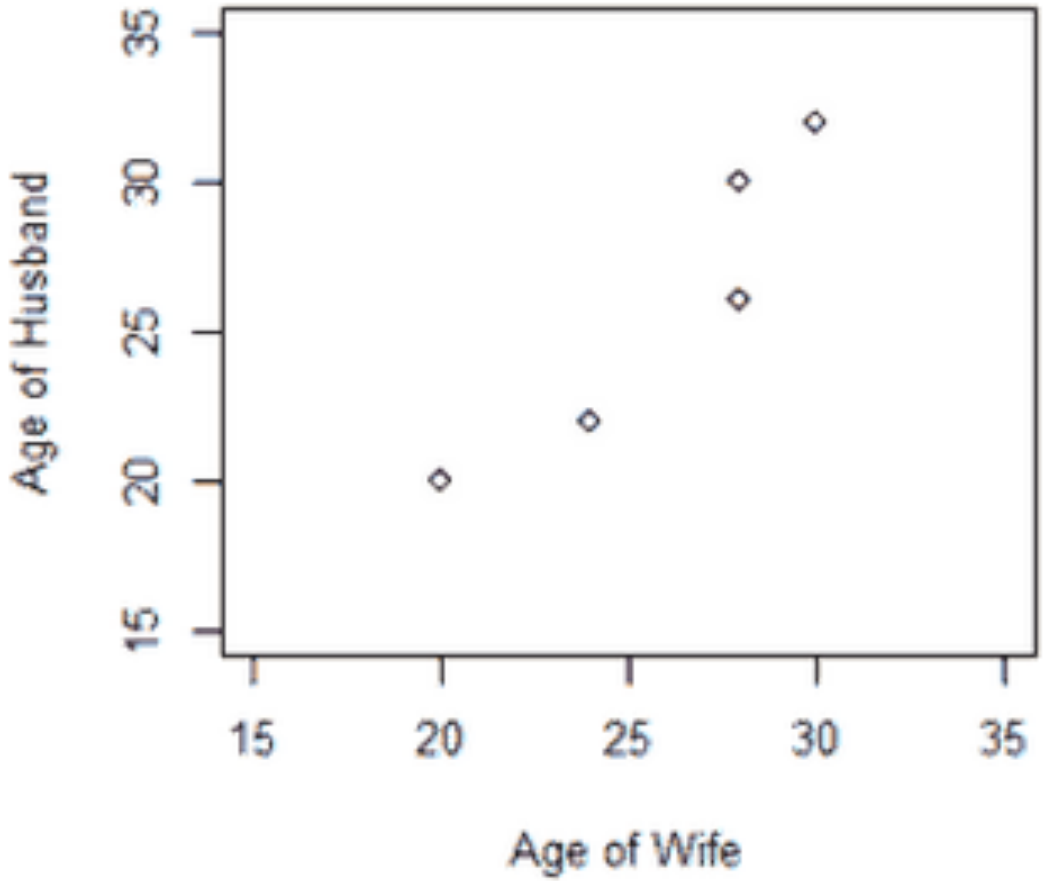
n is the number of data points or pairs
x_i is the i^{th} data point for variable x
x bar is the sample mean from variable x
s_x is the sample standard deviation of variable x
y_i is the i^{th} data point for variable y
y bar is the sample mean from variable y
s_y is the sample standard deviation of variable y

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

An example - Correlation

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 (continued)

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	$\frac{28-26}{4.0}=\frac{2}{4}$	$\frac{30-26}{5.1}=\frac{4}{5.1}$	$\left(\frac{2}{4}\right)\left(\frac{4}{5.1}\right)=\frac{8}{20.4}$

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

standardized age of the wife multiplied by the standardized age of the husband.

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**. You cannot use correlation to compute the correlation between gender and SAT scores

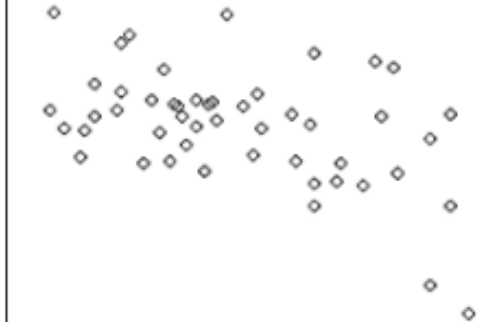
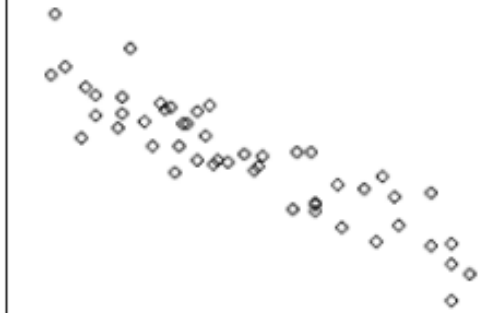
4. The correlation coefficient **does not have units** and it is **independent** of **unit of measure of variables x and y**.

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.

Correlation coefficients

$r = 0.7$	
$r = 0.3$	
$r = 0$	

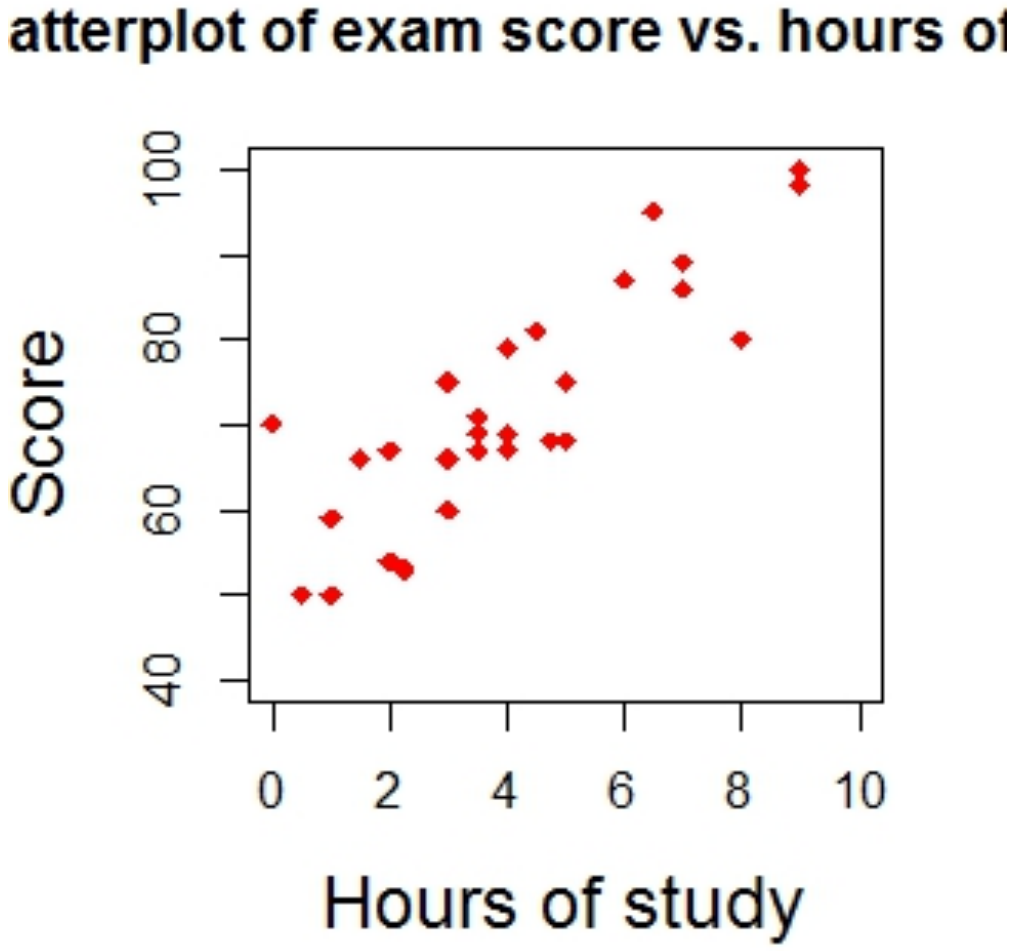
$r = -0.5$	
$r = -0.9$	

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 – whether or not there is a linear association

We often want to use the sample data to **make conclusions about the correlation between the same parameters in the population.**

The **sample correlation, r**, is a point estimate for the **population correlation coefficient, ρ.**

Formal tests of hypotheses concerning ρ seek to determine whether there is a linear association between the variables in the population.

They want to address whether
ρ=0 (H_0: there is no linear association)
or ρ≠0 (H_1: there is a linear association).

We use the test statistic: $$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Inference – whether or not there is a linear association

They want to address whether ρ=0 (**H_0: there is no linear association**) or ρ≠0 (H_1: 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 α test is:

Reject H_0: ρ=0 if $t \geq t_{n-2,\frac{\alpha}{2}}$ or $t < -t_{n-2,\frac{\alpha}{2}}$

Otherwise do not reject H_0 : ρ=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 α/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: ρ=0 (no linear association) versus H_1:ρ≠0 (linear association).

1. Set up the hypotheses and select the alpha level
   H_0:ρ=0 (there is no linear association)
   H_1:ρ≠0 (there is a linear association)
   α=0.05

2. Select the appropriate test statistic

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

An example – Inference (continued)

3. State the decision rule
Decision Rule: **Reject H_0 if p<α**. 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 **α/2=0.025**
Using the table, t=2.045
> qt(0.975, df=29) = 2.04523
Decision Rule: **Reject H_0 if t≥2.045** or if t≤−2.045 (|t|≥2.045).
Otherwise, do not reject H_0

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≥2.045. We have significant evidence at the **α=0.05** level that ρ≠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”

If method is "**pearson**", the test statistic is based on Pearson's product moment correlation coefficient and follows a t distribution with length(x)-2 degrees of freedom if the samples follow independent normal distributions. If there are at least 4 complete pairs of observation, an asymptotic confidence interval is given based on Fisher's Z transform.

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 tests may be used if the data do **not necessarily come from a normal distribution.**

Simple Linear Regression

Simple linear regression

If a linear relationship exists, we can describe 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.

In SLR, we **must select** one variable to be the **response variable** and the other to be the **explanatory variable**.

Not only does SLR allow us to quantify the relationship between the response variable and the explanatory variable, it also allows us a tool for **predicting 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 (based on the pattern of the data).

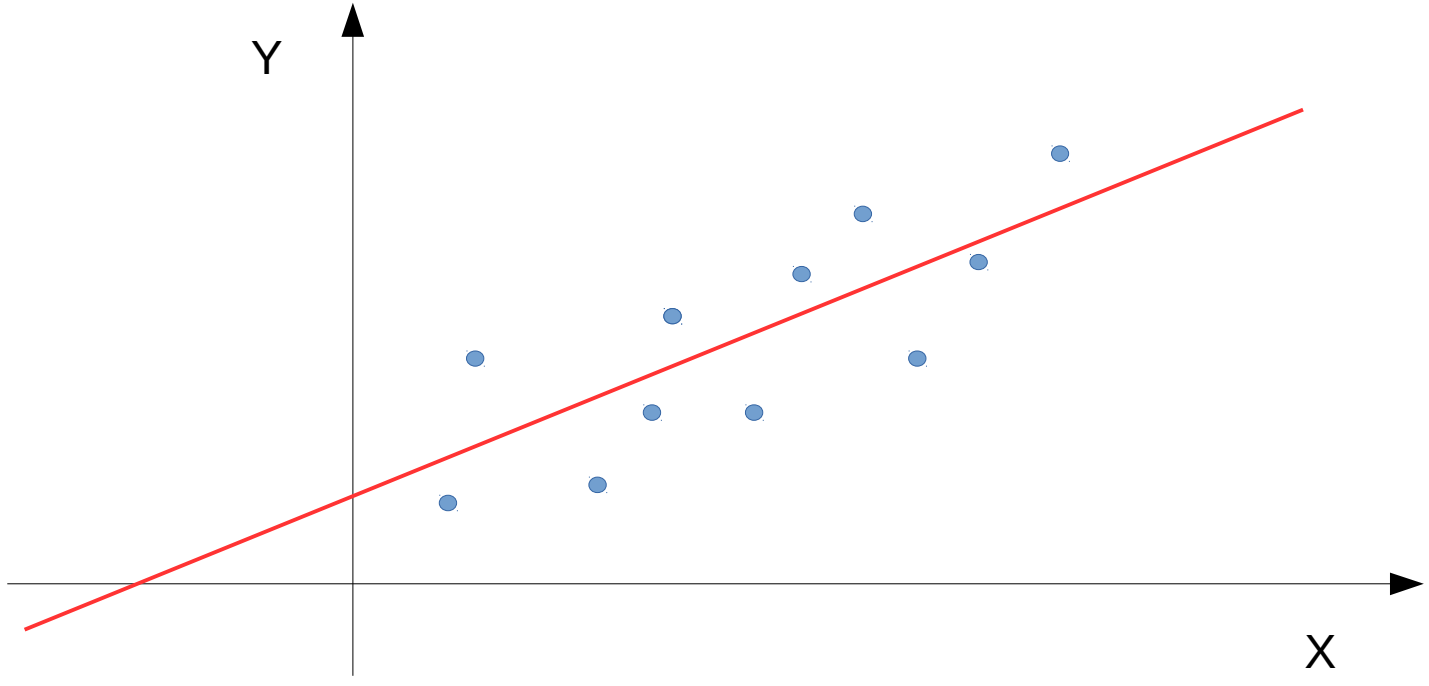
Linear Regression

- Find a pattern to build a model based on it

- A linear Equation:

$$Y = \beta_0 + \beta_1 X$$

  - β_1 is is the coefficient of the independent variable (slope)
  - β_0 is the constant term or the y intercept.

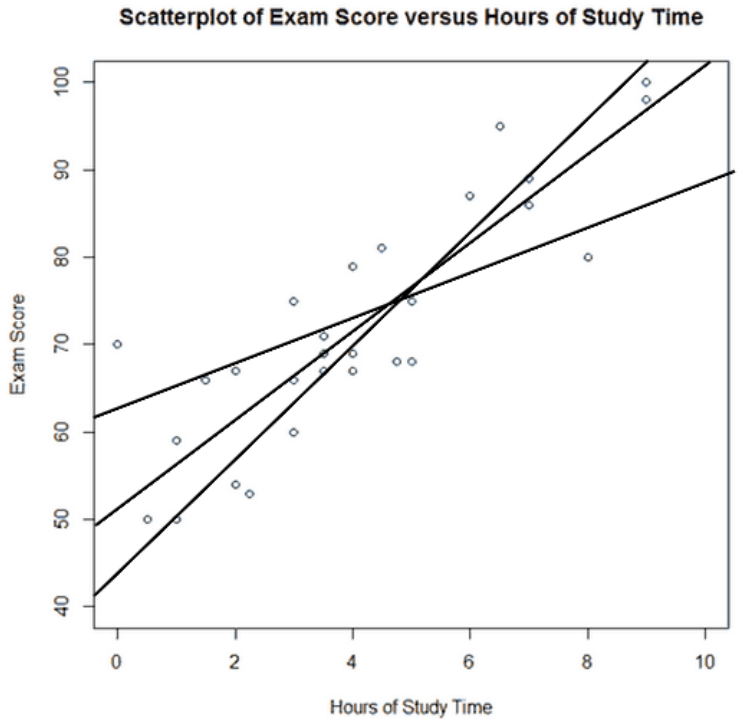


Simple linear regression

The equation for the simple linear regression line is given by

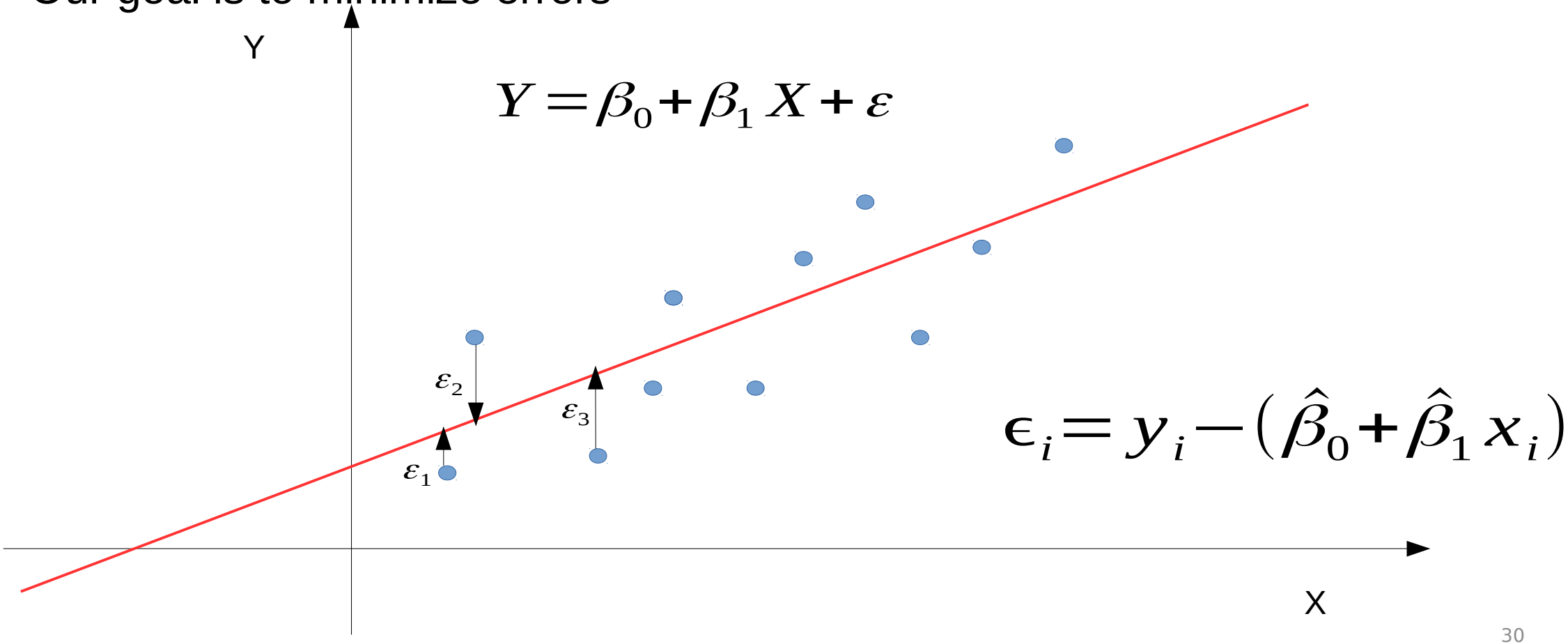
$$y = \beta_0 + \beta_1 x$$

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)



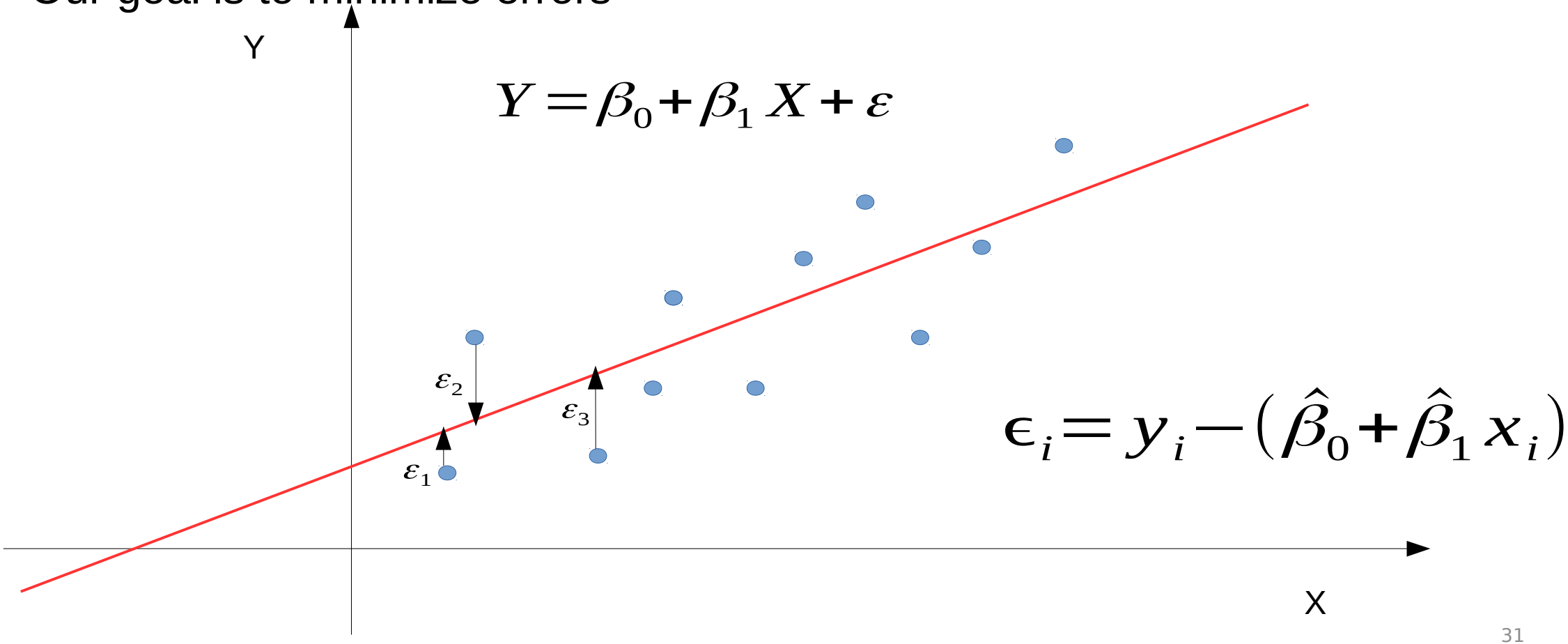
Linear Regression

- An error is the **distance** between actual value and model value
- Our goal is to minimize errors



Linear Regression

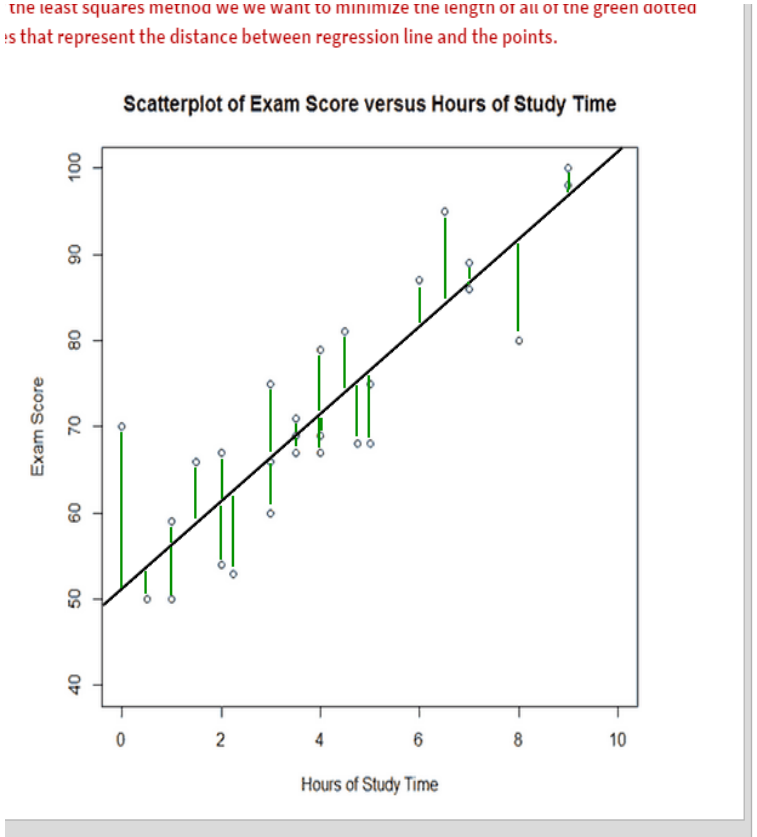
- An error is the **distance** between actual value and model value
- Our goal is to minimize errors



How to find the regression line that best fits the data

- We want to **minimize** the vertical distance between each of the points and the regression line.
- The most widely used method, **least-squares method**, aims to minimize the **sum of the squares of the distances** between the points and the regression line.
- Using the least-squares method, you can calculate the equation using just the correlation between the variables and each variable's mean and standard deviation.

Simple linear regression fits a straight line through the set of data points in such a way that makes the sum of squared residuals of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible.

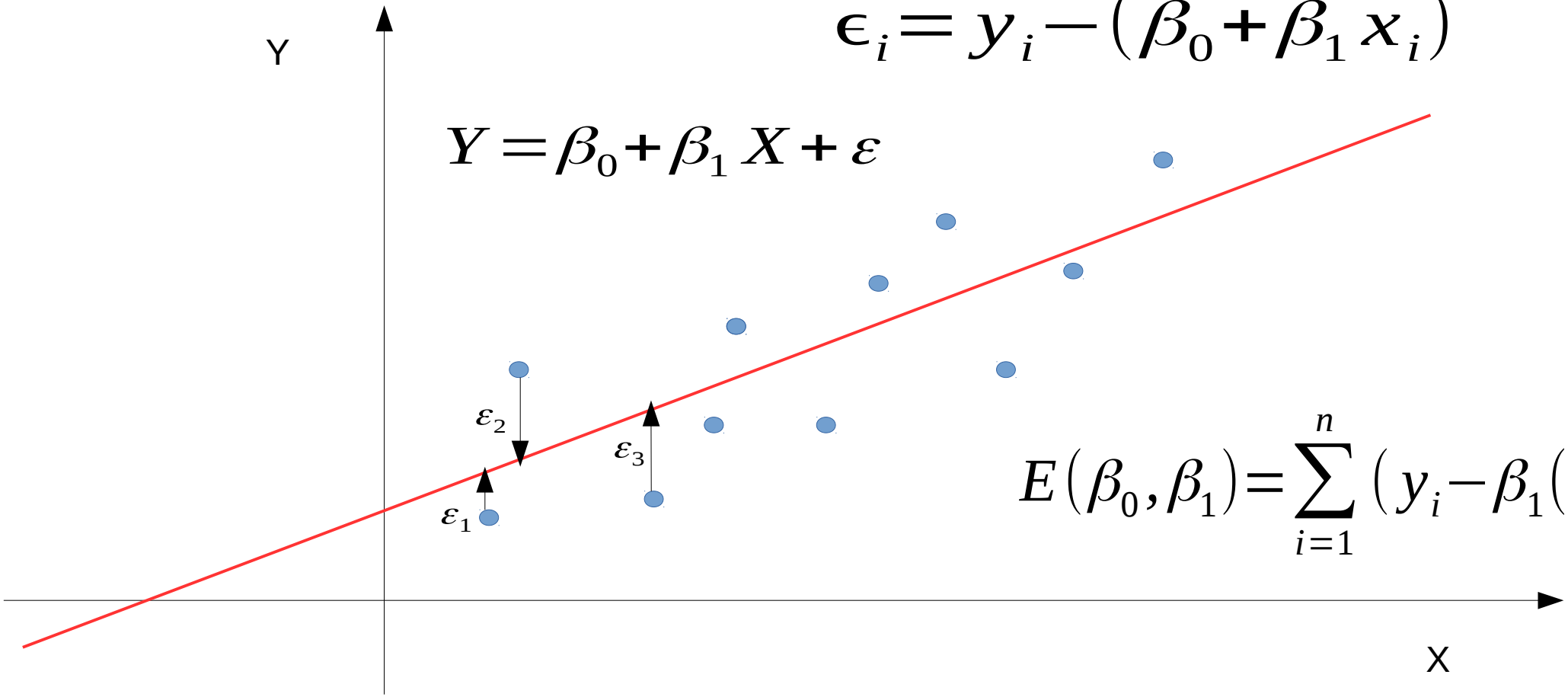


Linear Regression

- An error is the **distance** between actual value and model value

$$\epsilon_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$



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

Equation for the least-squares regression line

The equation for the simple linear regression line is given by

$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x$

$\hat{y}$ is the expected or predicted value of y for a given value of x
x is the explanatory or independent variable
$\widehat{\beta_0}$ is the least-squares estimates of β_0 (the intercept)
$\widehat{\beta_1}$ is the least-squares estimates of β_1 (the slope)

In the least-squares regression, the estimates of β_0 and β_1 are:
$\widehat{\beta_1} = r\frac{S_y}{S_x}$ and $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$

r = correlation coefficient, S_x = the sample standard deviation of x, S_y = the sample standard deviation of y, $\bar{x}$ = sample mean of x, $\bar{y}$ = sample mean of y

The equation for $\widehat{\beta_0}$ ensures that the least-squares regression line always passes through the "center of mass" point ($\bar{x}$, $\bar{y}$)

Simple linear regression

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)

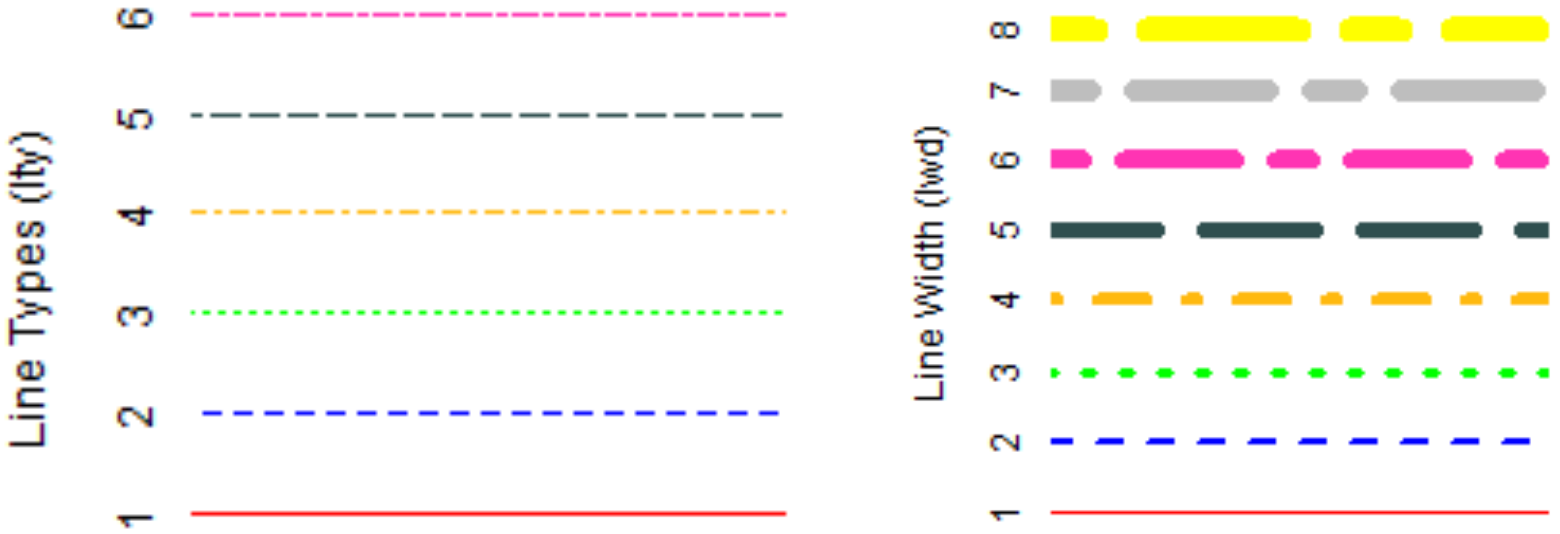
> **lm(data$responsevariable~data$explanatory)**
> **abline**(a=intercept, b=slope)

> 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/

abline() function

- Control color using col=“color”
- Control the line type by lty = (There are 6 line types)
- Control the line width lwd = (it can be a >0 number, for example, lwd from 1 - 8 as follows:



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

An example - the least-squares regression line

National Unemployment Male Vs. Female

```
> 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
```

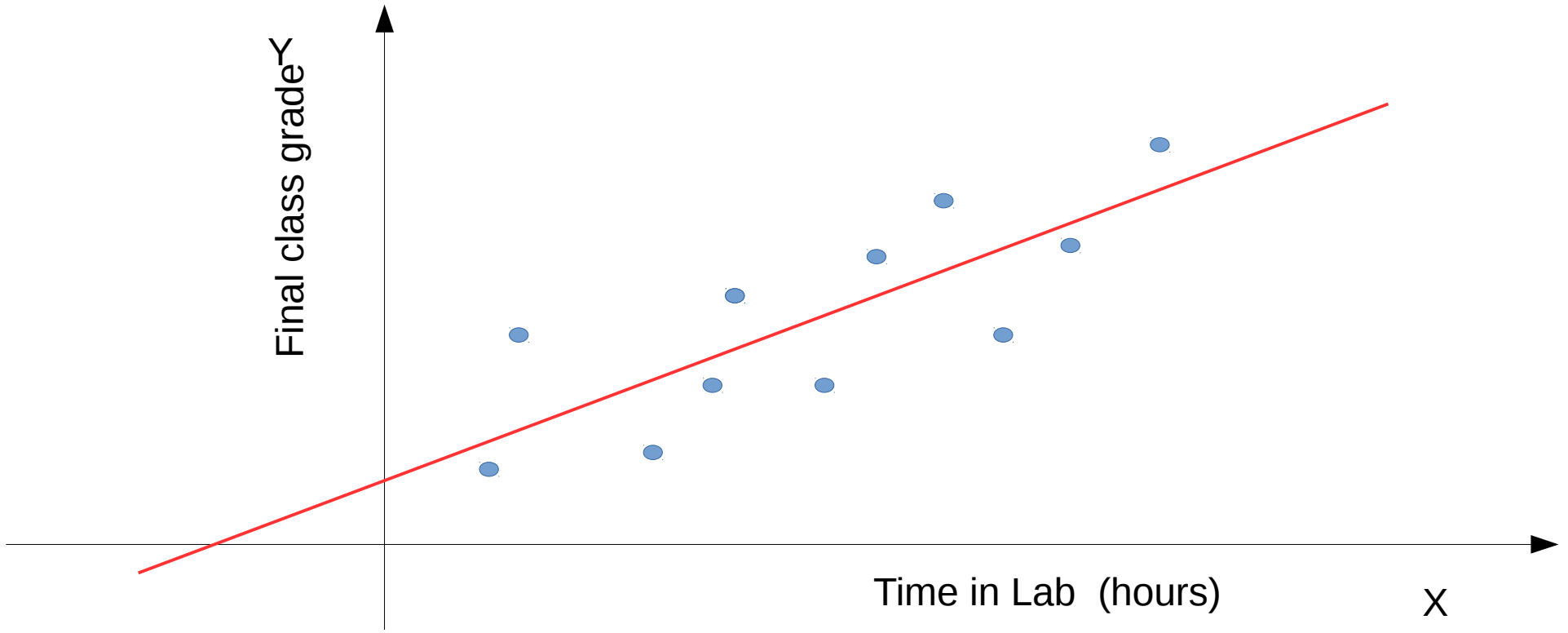
$\widehat{\beta_1}$=0.69 and $\widehat{\beta_0}$=1.43 thus $\hat{y} = 1.43 + 0.69x$

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

Reference: Statistical Abstract of the United States

Example

- Positive Relationship, positive slope



Example

- Negative Relationship, negative slope