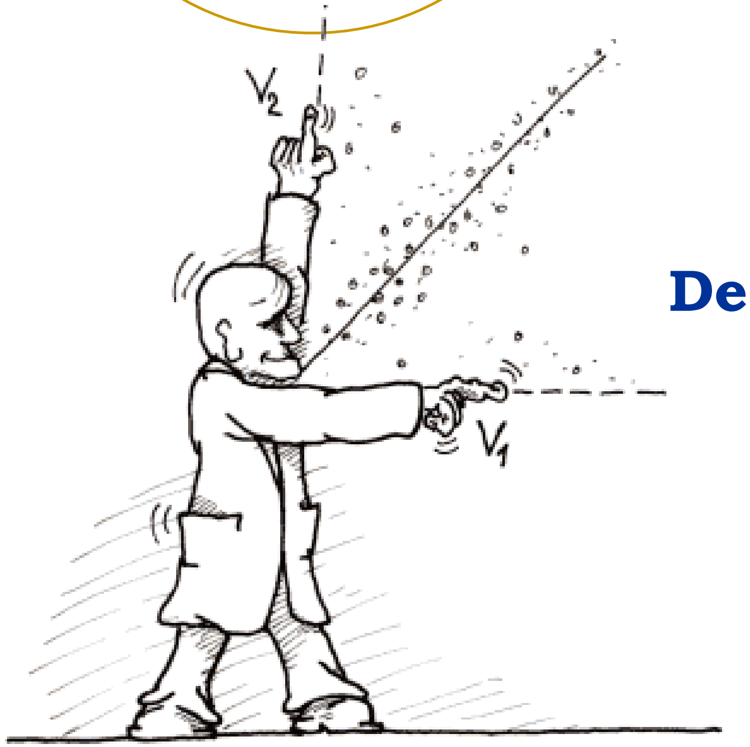
MC3020 – Association between Two Variables



T. Mayooran,
Department of Interdisciplinary Studies,
Faculty of Engineering,
University of Jaffna.
Email: mayooran@eng.jfn.ac.lk

Association between Two Variables

- Correlation Coefficient

- Liner Regression

General Definitions:

There are two types of relationships between variables:

- **Simple relationship-** Simple linear regression
- **Multiple relationship** - Multiple linear regression

In a **simple relationship,** there are two variables an **independent variable(X),** also called an explanatory variable or a predictor variable, and a **dependent variable(Y),** also called a response variable.

A simple relationship analysis is called *simple regression*, and there is one independent variable that is used to predict the dependent variable.

For example, a manager may wish to see whether the number of years the salespeople have been working for the company has anything to do with the amount of sales they make. This type of study involves a simple relationship, since there are only two variables—years of experience and amount of sales

In a **multiple relationship**, called multiple regression, two or more independent variables are used to predict one dependent variable.

For example, an educator may wish to investigate the relationship between a student's success in college and factors such as the number of hours devoted to studying, the student's GPA, and the student's high school background. This type of study involves several variables.

Simple relationships can also be positive or negative. A **positive relationship** exists when both variables increase or decrease at the same time.

For instance, a person’s height and weight are related; and the relationship is positive, since the taller a person is, generally, the more the person weighs.

In a **negative relationship,** as one variable increases, the other variable decreases, and vice versa.

For example, if you measure the strength of people over 60 years of age, you will find that as age increases, strength generally decreases.

Scatter Plots and Correlation

A scatter plot is a graph of the ordered pairs (x, y) of numbers consisting of the independent variable x and the dependent variable y.

The independent variable x is plotted on the horizontal axis, and the dependent variable y is plotted on the vertical axis.

The scatter plot is a visual way to describe the nature of the relationship between the independent and dependent variables.

Example

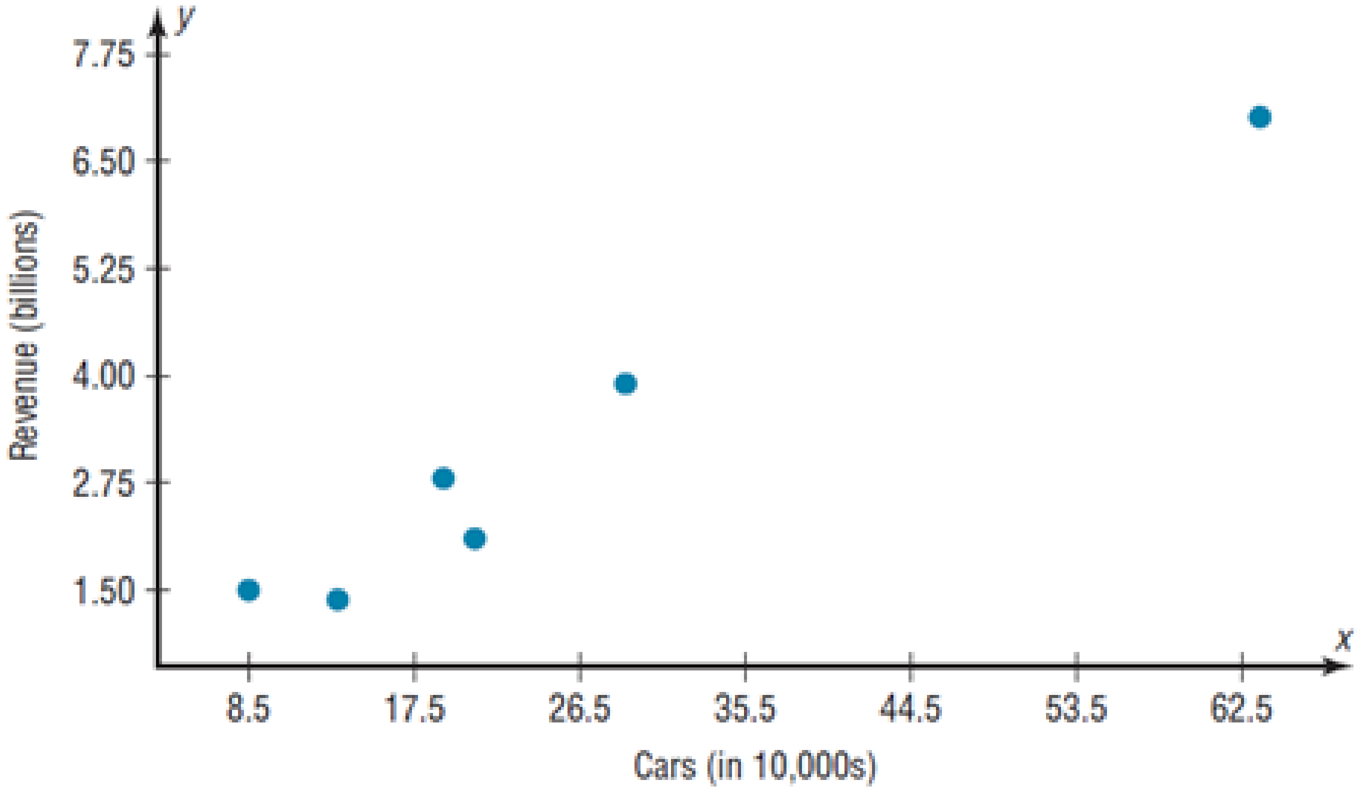
Construct a scatter plot for the data shown for car rental companies in the United States for a recent year.

Company	Cars (in ten thousands)	Revenue (in billions)
A	63.0	7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

Solution

Step 1 Draw and label the x and y axes.

Step 2 Plot each point on the graph, as shown in following slide.


y
7.75
6.50
5.25
4.00
2.75
1.50
Revenue (billions)
x
8.5
17.5
26.5
35.5
44.5
53.5
62.5
Cars (in 10,000s)

Correlation Coefficient

The correlation coefficient computed from the sample data measures the strength and direction of a linear relationship between two variables.

The symbol for the sample correlation coefficient is r.

The symbol for the population correlation coefficient is ρ (Greek letter rho).

Let X and Y be two different measurements on an individual subject. The population correlation coefficient is measured as

$$\rho = \frac{\text{Covariance between X and Y}}{\sqrt{\text{Variance}(X) * \text{Variance}(Y)}} \; ; -1 \leq \rho \leq +1$$

The Covariance between X and Y is the mean of the product of the deviations from the respective means.

The range of the correlation coefficient is from -1 to 1.

$\rho = 0$ indicates that there is no linear relationship between X and Y.

$\rho = +1$ indicates that there is a perfect positive linear relationship between X and Y.

$\rho = -1$ indicates that there is a perfect negative linear relationship between X and Y.

Other values are interpreted as how strong the relationship is depending on how close the value is to -1 or +1.

Let $(X_1, Y_1), (X_2, Y_2), \quad . . \quad . (X_n, Y_n)$ be n pairs of measurements on X and Y. Then the sample correlation coefficient is computed as,

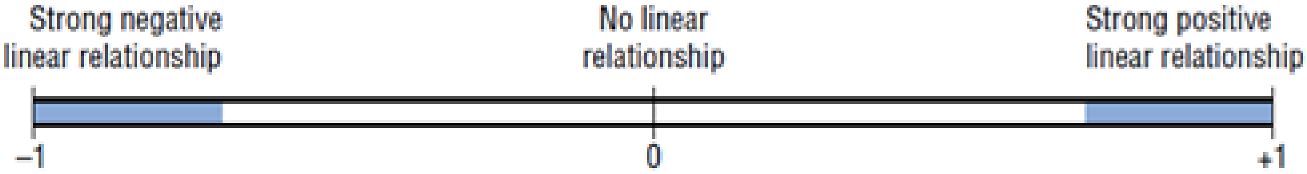
$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 * \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$= \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right)}}$$

If there is a strong positive linear relationship between the variables, the value of r will be close to 1.

If there is a strong negative linear relationship between the variables, the value of r will be close to -1.

When there is no linear relationship between the variables or only a weak relationship, the value of r will be close to 0. See following figure.

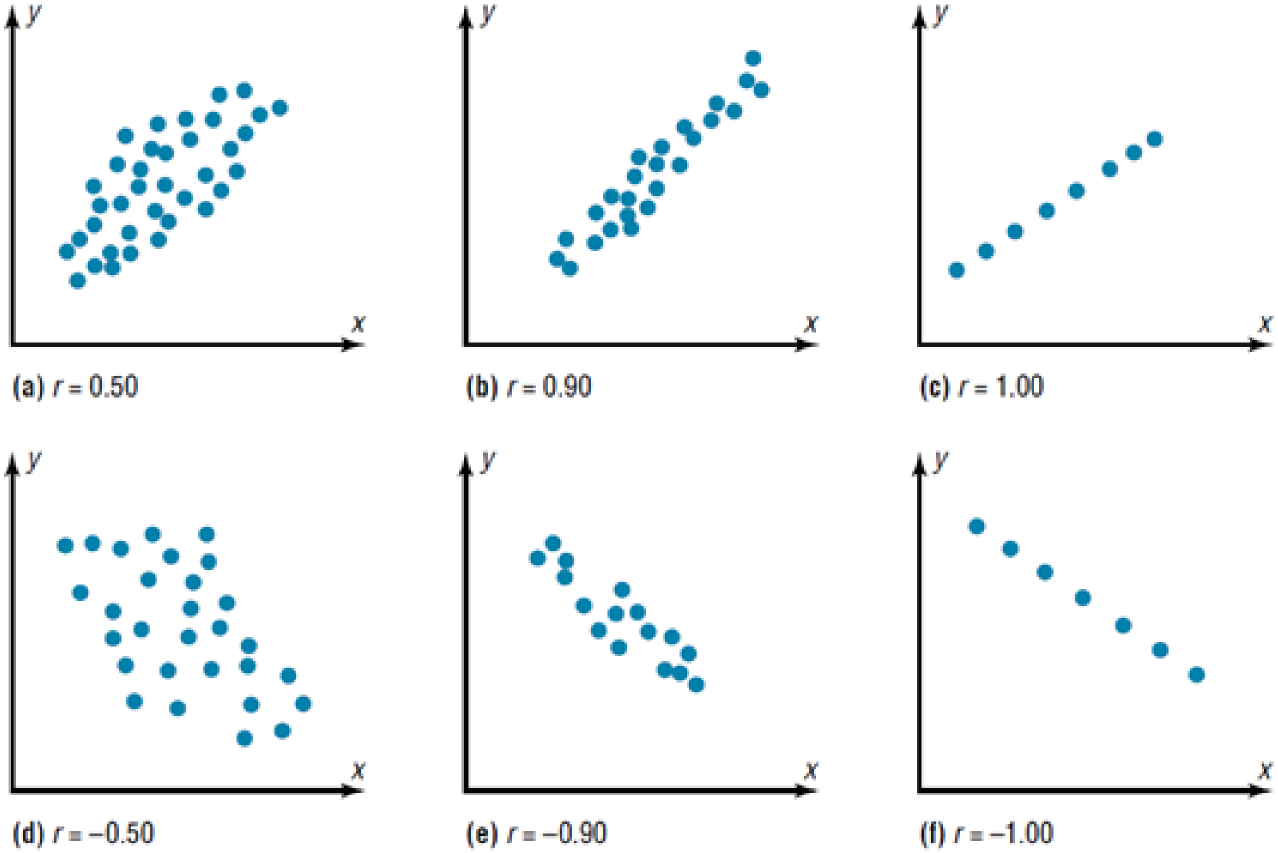


The graphs in following figure show the relationship between the correlation coefficients and their corresponding scatter plots.

Notice that as the value of the correlation coefficient increases from 0 to +1 (parts a, b, and c), data values become closer to an increasingly stronger relationship.

As the value of the correlation coefficient decreases from 0 to -1 (parts d, e, and f), the data values also become closer to a straight line.

There are no units associated with r, and the value of r will remain unchanged if the x and y values are switched.



(a) $r = 0.50$

(b) $r = 0.90$

(c) $r = 1.00$

(d) $r = -0.50$

(e) $r = -0.90$

(f) $r = -1.00$

Example 1:

A college administers all its courses a student evaluation questionnaire. For a random sample of 12 courses the accompanying table and the data file student evaluation show both the average student ratings of the instructor (on a scale of 1 to 5), and the average expected grades of the students (on a scale from A = 4 to F = 0). Find the sample correlation coefficient between instructor ratings and expected grades.

Instructor ratings	2.8	3.7	4.4	3.6	4.7	3.5	4.1	3.2	4.9	4.2	3.8	3.3
Expected grades	2.6	2.9	3.3	3.2	3.1	2.8	2.7	2.4	3.5	3	3.4	2.5

Instructor ratings (X)	Expected grades (Y)	XY	X^2	Y^2
2.8	2.6	7.28	7.84	6.76
3.7	2.9	10.73	13.69	8.41
4.4	3.3	14.52	19.36	10.89
3.6	3.2	11.52	12.96	10.24
4.7	3.1	14.57	22.09	9.61
3.5	2.8	9.8	12.25	7.84
4.1	2.7	11.07	16.81	7.29
3.2	2.4	7.68	10.24	5.76
4.9	3.5	17.15	24.01	12.25
4.2	3.0	12.6	17.64	9.0
3.8	3.4	12.92	14.44	11.56
3.3	2.5	8.25	10.89	6.25
$\sum_{i=1}^{12} X_i = 46.2$	$\sum_{i=1}^{12} Y_i = 35.4$	$\sum_{i=1}^{12} X_i Y_i = 138.09$	$\sum_{i=1}^{12} X_i^2 = 182.22$	$\sum_{i=1}^{12} Y_i^2 = 105.86$

$$\bar{X} = \frac{\sum_{i=1}^{12} X_i}{12} = \frac{46.2}{12} = 3.85 \qquad \bar{Y} = \frac{\sum_{i=1}^{12} Y_i}{12} = \frac{35.4}{12} = 2.95$$

$$\sum_{\mathrm{i=1}}^{12} \mathrm{X_i Y_i} = 138.09 \qquad \sum_{\mathrm{i=1}}^{12} X_{\mathrm{i}}^2 = 182.22 \qquad \sum_{\mathrm{i=1}}^{12} Y_{\mathrm{i}}^2 = 105.86$$

$$r = \frac{\sum_{i=1}^{n} X_i Y_i \; - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 \; - n\bar{X}^2\right)\left(\sum_{i=1}^{n} Y_i^2 \; - n\bar{Y}^2\right)}}$$

$$= \frac{138.09 - 12 * 3.85 * 2.95}{\sqrt{(182.22 - 12 * 3.85^2)(105.86 - 12 * 2.95^2)}} = 0.7217$$

Testing Hypothesis for population correlation:

Step I

The significance of the correlation coefficient can be tested by writing the hypotheses in one of the three forms below:

Case 1:$H_0 : \rho \geq 0 \quad H_1 : \rho < 0$ (Left tailed)

Case 2:$H_0 : \rho \leq 0 \quad H_1 : \rho > 0$ (Right tailed)

Case 3: $H_0 : \rho = 0 \quad H_1 : \rho \neq 0$ (two tailed)

Step II

The statistic,

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

will have t-distribution with $(n-2)$ degrees of freedom.

Step III

Calculate the critical values and Identify critical region.

Step IV

Statistical Conclusion- Reject the H_0 or Do not reject H_0

Step V General Conclusion

Example 2:

Test the significance of the correlation coefficient found in Example 1, Use $\alpha = 0.05$

Simple Linear Regression

The study in which the linear relationship is obtained is known as regression analysis. Let Y be dependent variable and X be independent variable. That is, value of Y depends on the value of X. The relationship can be of any nature or functional form but here we only consider the linear relationship,

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

Where β_0 is the intercept coefficient and β_1 is the slope coefficient. is the random fluctuation from the line that follows normal distribution with mean zero and variance.

For a random sample of size *n* , the sum of squared errors can be written as

$$SSE = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

The SSE is minimized when β_1 is estimated as,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)}$$

and β_0 is estimated as,

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are known as the least square estimates for β_0 andβ_1 respectively.

Then the least square estimate of the regression line,

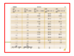
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Example 3:

A college administers all its courses a student evaluation questionnaire. For a random sample of 12 courses the accompanying table and the data file student evaluation show both the average student ratings of the instructor (on a scale of 1 to 5), and the average expected grades of the students (on a scale from A = 4 to F = 0).

Instructor ratings	2.8	3.7	4.4	3.6	4.7	3.5	4.1	3.2	4.9	4.2	3.8	3.3
Expected grades	2.6	2.9	3.3	3.2	3.1	2.8	2.7	2.4	3.5	3	3.4	2.5

1. Determine the least-square equation in predicting the Expected grades using their Instructor ratings.



2. Interpret the slope coefficient in the regression line in the context of the problem.

3. Use this regression line, predict the Expected grade's scale for a Instructor rating 4.8.

Use your calculator Casio 991ES/MS, to verify these calculated values.

Testing Hypothesis for slope coefficient:

The significance of the regression line can be tested by writing the hypotheses in one of the three forms below:

Case 1:$H_0 : \beta_1 \geq 0 \quad H_1 : \beta_1 < 0$

Case 2:$H_0 : \beta_1 \leq 0 \quad H_1 : \beta_1 > 0$

Case 3: $H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$

The statistic,

$$T = \frac{\hat{\beta}_1 - 0}{S_e \Big/ \sqrt{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}}$$

will have t-distribution with $n-2$ degrees of freedom, where,

$$S_e = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2}}$$

the standard error of estimate.

Testing Hypothesis for intercept coefficient:

The significance of the regression line can be tested by writing the hypotheses in one of the three forms below:

Case 1:$H_0 : \beta_0 \geq 0 \quad H_1 : \beta_0 < 0$

Case 2:$H_0 : \beta_0 \leq 0 \quad H_1 : \beta_0 > 0$

Case 3: $H_0 : \beta_0 = 0 \quad H_1 : \beta_0 \neq 0$

Similar tests involving , the intercept coefficient can also be obtained. Where the test statistic is used as:

$$T = \frac{\hat{\beta}_0 - 0}{S_e \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}}}$$

which has t-distribution with $n - 2$ degrees of freedom.

Confidence Interval Estimation:

$(1-\alpha) * 100\ \%$ Confidence Interval for β_1 can be computed as,

$$\hat{\beta}_1 \mp t_{\alpha/2, n-2} * \frac{S_e}{\sqrt{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}}$$

$(1-\alpha) * 100\ \%$ Confidence Interval for β_0 can be computed as,

$$\hat{\beta}_0 \mp t_{\alpha/2, n-2} * S_e \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}}$$

$(1-\alpha)*100\,\%$ Confidence Interval for mean response at $x = x_0, y(x_0) = \beta_0 + \beta_1 x_0$ can be computed as,

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \mp t_{\alpha/2,n-2} * S_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}}$$

$(1-\alpha)*100\,\%$ Prediction Interval for predicted response at $x = x_0, y(x_0) = \beta_0 + \beta_1 x_0$ can be computed as,

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \mp t_{\alpha/2,n-2} * S_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}}$$

Example:

Consider example 3,
Determine the standard error and construct 90% Confidence Interval for the predicted response at $x = 4$.

Instructor ratings (X)	Expected grades (Y)	XY	X^2	Y^2	$\hat{Y} = 1.3569 + 0.4138 * X$	$e = Y - \hat{Y}$	$(Y - \hat{Y})^2$
2.8	2.6	7.28	7.84	6.76	2.5155	0.0845	0.0071
3.7	2.9	10.73	13.69	8.41	2.8880	0.0120	0.0001
4.4	3.3	14.52	19.36	10.89	3.1776	0.1224	0.0150
3.6	3.2	11.52	12.96	10.24	2.8466	0.3534	0.1249
4.7	3.1	14.57	22.09	9.61	3.3018	-0.2018	0.0407
3.5	2.8	9.8	12.25	7.84	2.8052	-0.0052	0.0000
4.1	2.7	11.07	16.81	7.29	3.0535	-0.3535	0.1249
3.2	2.4	7.68	10.24	5.76	2.6811	-0.2811	0.0790
4.9	3.5	17.15	24.01	12.25	3.3845	0.1155	0.0133
4.2	3.0	12.6	17.64	9.0	3.0949	-0.0949	0.0090
3.8	3.4	12.92	14.44	11.56	2.9293	0.4707	0.2215
3.3	2.5	8.25	10.89	6.25	2.7224	-0.2224	0.0495
$\sum_{i=1}^{12} X_i = 46.2$	$\sum_{i=1}^{12} Y_i = 35.4$	$\sum_{i=1}^{12} X_i Y_i = 138.09$	$\sum_{i=1}^{12} X_i^2 = 182.22$	$\sum_{i=1}^{12} Y_i^2 = 105.86$		$\sum_{i=1}^{12} e \approx 0$	$\sum_{i=1}^{12} (Y - \hat{Y})^2 = 0.6850$

Interpretation of the Coefficients

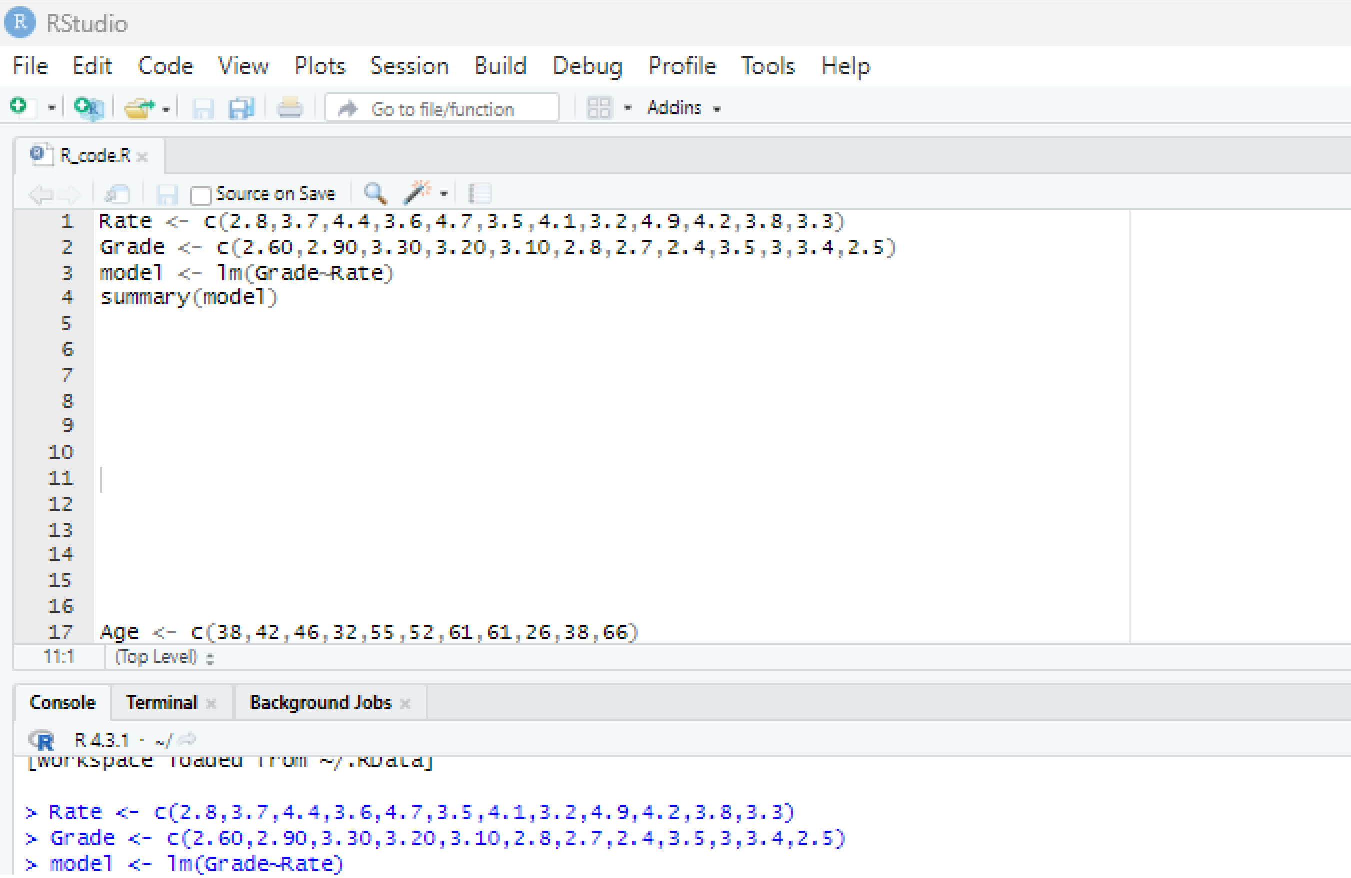
$\hat{\beta}_0$ is the point where the regression line crosses the vertical axis. It is really the value of when x = 0. However in general if x = 0 is not within the range of the given X values the interpretation of β_0 is meaningless as is the case here. On examination of the X values, we see that 0 is not with its range. When the range of the observed explanatory variable values does not include 0, it is best to think of the intercept as an "anchor" for the least squares line.

$\hat{\beta}_1 = 0.4138$ implies that, for every additional unit in Instructor rating (as a unit of those of the leading competitor), expected grade's scale (as a unit of those of the leading competitor) increases on average by 0.4138 unit.

If Coefficient of determination (R^2) = 0.850, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.

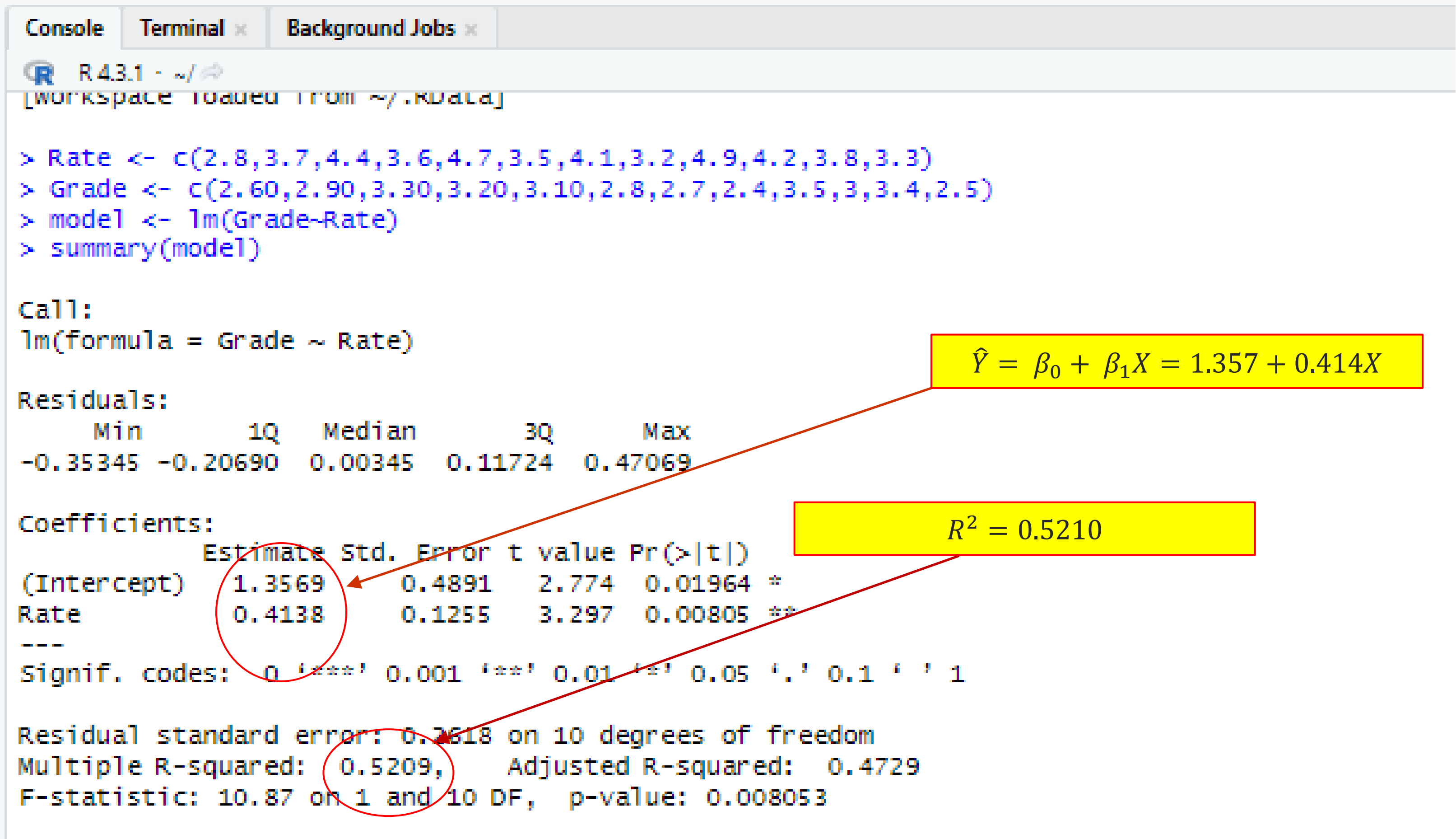
In our example, Coefficient of determination (R^2) = 0.5208, So we can say instructor rating (X) in the fitted model explains 52.08% of the total variation in expected grade(Y).

In R program



```
Rate <- c(2.8,3.7,4.4,3.6,4.7,3.5,4.1,3.2,4.9,4.2,3.8,3.3)
Grade <- c(2.60,2.90,3.30,3.20,3.10,2.8,2.7,2.4,3.5,3,3.4,2.5)
model <- lm(Grade~Rate)
summary(model)













Age <- c(38,42,46,32,55,52,61,61,26,38,66)
```

```
[Workspace loaded from ~/.RData]

> Rate <- c(2.8,3.7,4.4,3.6,4.7,3.5,4.1,3.2,4.9,4.2,3.8,3.3)
> Grade <- c(2.60,2.90,3.30,3.20,3.10,2.8,2.7,2.4,3.5,3,3.4,2.5)
> model <- lm(Grade~Rate)
```



```
[Workspace loaded from ~/.RData]

> Rate <- c(2.8,3.7,4.4,3.6,4.7,3.5,4.1,3.2,4.9,4.2,3.8,3.3)
> Grade <- c(2.60,2.90,3.30,3.20,3.10,2.8,2.7,2.4,3.5,3,3.4,2.5)
> model <- lm(Grade~Rate)
> summary(model)

Call:
lm(formula = Grade ~ Rate)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.35345 -0.20690  0.00345  0.11724  0.47069 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   1.3569     0.4891   2.774  0.01964 * 
Rate          0.4138     0.1255   3.297  0.00805 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2618 on 10 degrees of freedom
Multiple R-squared:  0.5209,	Adjusted R-squared:  0.4729 
F-statistic: 10.87 on 1 and 10 DF,  p-value: 0.008053
```

$\hat{Y} = \beta_0 + \beta_1 X = 1.357 + 0.414X$

$R^2 = 0.5210$

Example:

Using the following data consisting of seven different Fathers and their daughters' heights in inches, answer the questions below:

Father’s height	63.0	67.0	64.0	60.0	65.0	67.0	66.0
Daughter’s height	63.6	65.7	65.3	61.0	65.4	66.4	67.2

a. Obtain the least-square equation in predicting the Daughter’s height using their Father’s height.

b. What is the predicted height of the Daughter of a 59 inches tall Father?

c. Find the estimate of the standard error of the estimate (s_e).

d. Test whether the father’s height plays a significant role in predicting the daughter’s height. Use the 5% level of significance.

e. Construct the 95% prediction interval(PI) for the height of the daughter of a farther who is 62 inches tall

Example:

Based on Previous Example,

1. Compute the linear correlation coefficient between Daughter’s height and Father’s height.

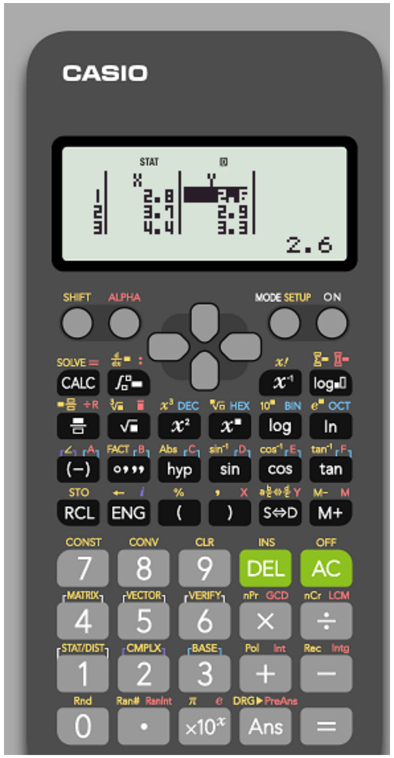
2. Use a 0.05 significance level to test for a linear correlation between Daughter’s height and Father’s height positive or not.

In Your Tutorial 6:

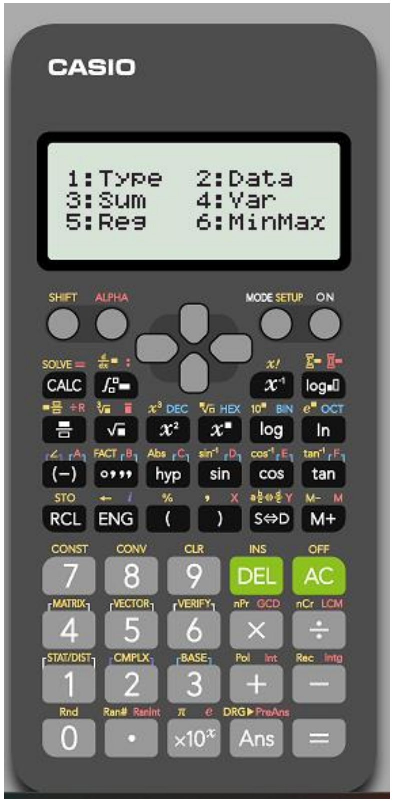
1. 3
2. 4

In Your Calculator:

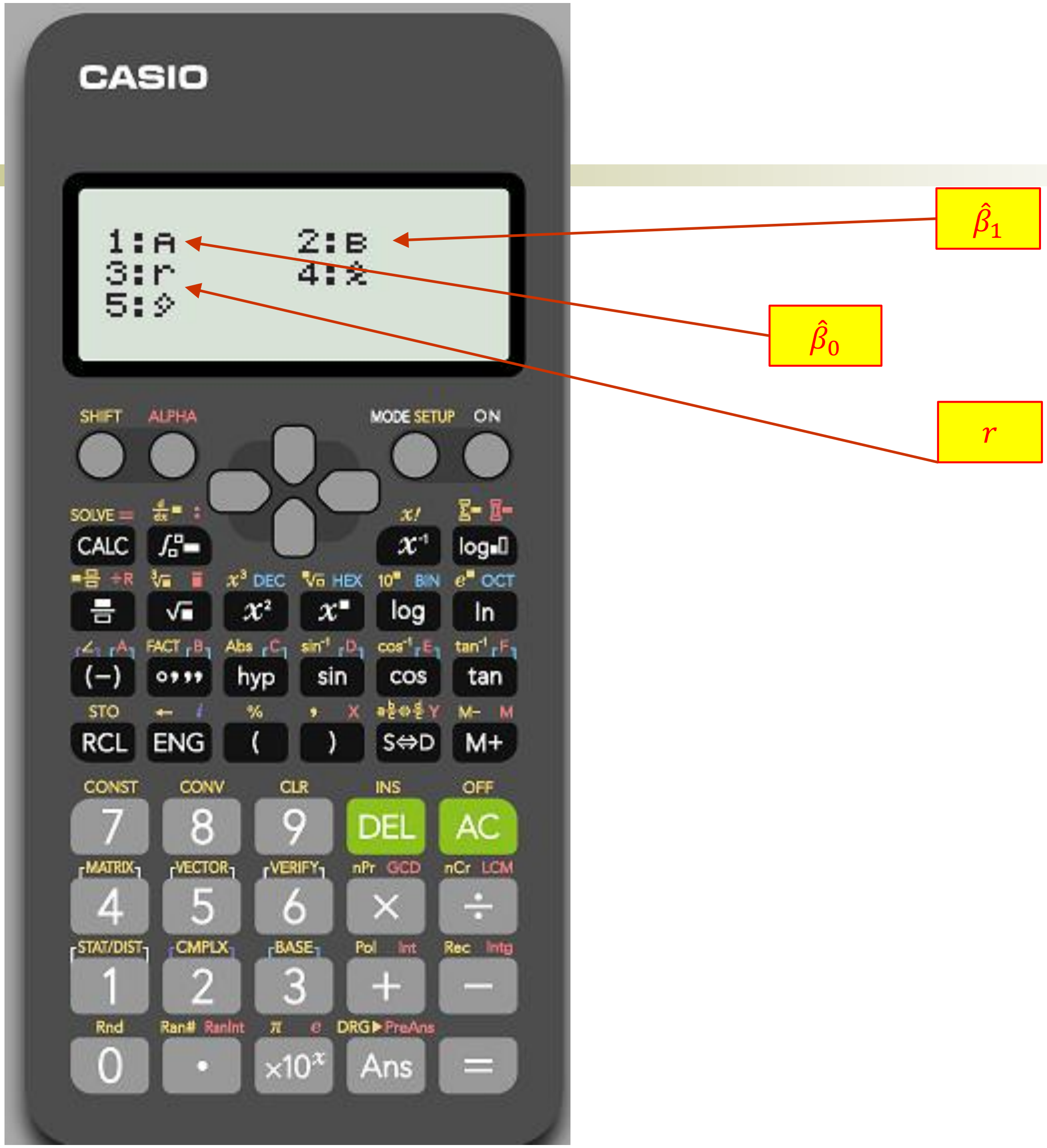
To find the equation of the regression line and correlation coefficient:



```
     STAT       D
 |   X  |   Y  |
1|  2.8 |  2.6 |
2|  3.7 |  2.9 |
3|  4.4 |  3.3 |
                 2.6
```



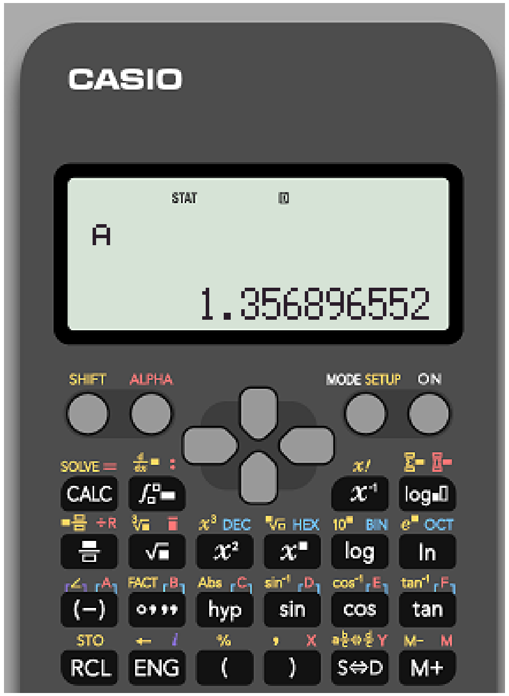
```
1:Type    2:Data
3:Sum     4:Var
5:Reg     6:MinMax
```



$\hat{\beta}_1$

$\hat{\beta}_0$

r



To find the fitted values($\hat{Y}$) of the regression:

1. Enter the **X** values in **X** and the **Y** values in **Y**.
2. Press **SHIFT** and press 1 and move to cursor to
3. Type your regression equation (Example **Y=1.3569+0.4138*L1**)
4. After press ENTER then you can find

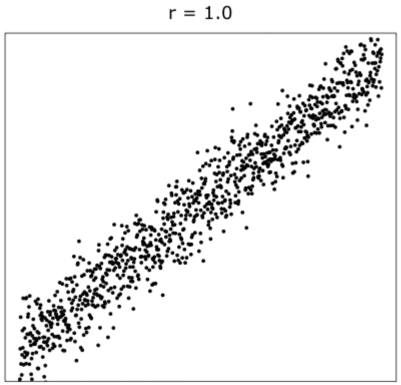
Multiple Linear regression

R program-based discussions

The slides will be updated soon!

Don’t hesitate to contact us if you have any questions about this course’s teaching contents. Also, don’t forget to check out the course page and Microsoft Team folder,

- course page Link: https://mayooran1987.github.io/MC3020/
- Course page’s QR code



- Microsoft Team folder link