Applied Statistics I
(IMT224β/AMT224β)

Department of Mathematics
University of Ruhuna

A.W.L. Pubudu Thilan

Linear Regression

- The goal of a correlation analysis was to quantify the strength of the linear relationship between the variables, whereas regression expresses the relationship in the form of an equation.
- Both correlation and linear regression assume that the relationship between the variables is linear.

Difference between correlation and regression

- In students taking a Mathematics and English test, we could use correlation to determine whether students who are good at Mathematics tend to be good at English.
- Regression can be used to determine whether the marks in English can be predicted for given marks in Mathematics.

Why do we need regression?

- The purpose of running the regression is to find a formula that fits the relationship between the variables.
- Then you can use that formula to predict values for the dependent variable when only the independent variables are known.
- **Eg:** A doctor could prescribe the proper dose based on a person's body weight.

Independent and dependent variables

- Independent and dependent variables are related to one another.
- The independent part is what you, the experimenter, changes or enacts in order to do your experiment.
- The dependent variable is what changes when the independent variable changes.
- The dependent variable depends on the outcome of the independent variable.

Independent and dependent variables
Examples

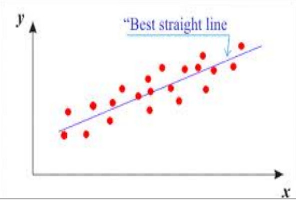
1. The profit made by the manufacturing unit, which is dependent on the sales volumes of the company. Here sales volume is the independent variable and profit is the dependent variable.
2. You are interested in how stress affects heart rate in humans. Your independent variable would be the stress and the dependent variable would be the heart rate.

Extraneous variables

- The independent and dependent variables are not the only variables present in many experiments.
- Any variables in your experiment that are not part of your manipulation is called extraneous variables.
- They are factors you haven't controlled.
- Extraneous variables affect your results, but usually they affect all your conditions equally and so they do not create any biases in your results.

Simple linear regression

- When there is only one independent (predictor, explanatory) variable, the prediction method is called **simple regression**.
- In **simple linear regression**, the predictions of y when plotted as a function of x form a straight line.



- The relationship between the dependent variable and the explanatory variable is represented by the following equation:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\ \beta_0 &- \text{constant term} \\ \beta_1 &- \text{coefficient of explanatory variable} \\ \epsilon_i &- \text{error term} \end{aligned}$$

The model

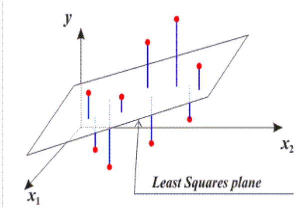
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, 2, 3, ..., n$$

is used with the following assumptions,

1. $E(\epsilon_i) = 0$.
2. The distribution of the errors in prediction of the value of y is constant regardless of the value of x.
   $\Rightarrow var(\epsilon_i) = \sigma^2$
3. Errors in prediction of the value of y are all independent of one another.
   $\Rightarrow cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$

Multiple linear regression

- In multiple linear regression, a linear combination of two or more explanatory variables is used to explain the variation in a response.
- When there are more than one explanatory variable, the method is quite similar, but instead of a scatterplot in two dimensions, we have to imagine a space with as many dimensions as there are variables.



The relationship between the dependent variable and the p explanatory variables is represented by the following equation:

$$\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \epsilon_i \\
\beta_0 &- \text{constant term} \\
\beta_i (i \neq 0) &- \text{coefficients relating the } p \text{ explanatory variables} \\
\epsilon_i &- \text{error term}
\end{aligned}$$

Remark

- Multiple linear regression can be thought of an extension of simple linear regression.
- Simple linear regression can be thought of as a special case of multiple linear regression, where $p = 1$.

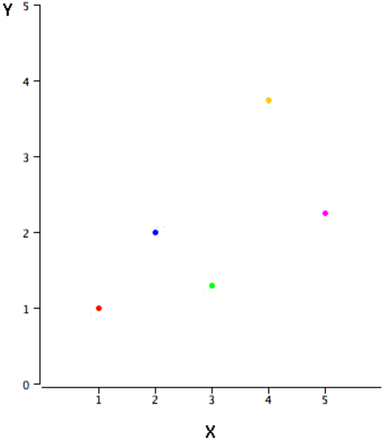
Find the regression line for following data.

X	Y
1.00	1.00
2.00	2.00
3.00	1.50
4.00	3.75
5.00	2.25

Illustrative example
Cont...

- Simple linear regression consists of finding the best-fitting straight line through the points.
- The best-fitting line is called a **regression line**.

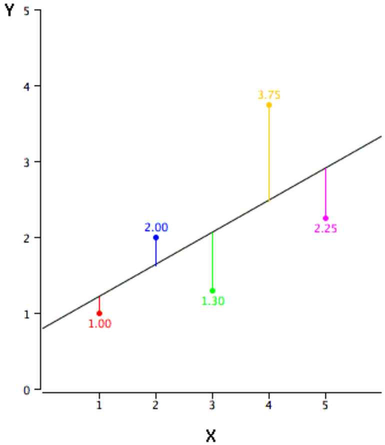
Illustrative example
Cont...



Illustrative example
Cont...

- The black diagonal line in Figure is the regression line and consists of the predicted score on Y for each possible value of X.

Illustrative example
Cont...



Illustrative example
Cont...

- The vertical lines from the points to the regression line represent the errors of prediction.
- The red point is very near the regression line; its error of prediction is small.
- The yellow point is much higher than the regression line and therefore its error of prediction is large.

Illustrative example
Cont...

- The error of prediction for a point is the value of the point minus the predicted value (the value on the line).
- For example, the first point has a y of 1.00 and a predicted $\hat{y}$ of 1.21. Therefore its error of prediction is -0.21.
- The below Table shows the predicted values $(\hat{y}_i)$ and the errors of prediction $(y_i - \hat{y}_i)$.

Illustrative example
Cont...

x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1.00	1.00	1.210	-0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	-0.760	0.578
4.00	3.75	2.485	1.265	1.600
5.00	2.25	2.910	-0.660	0.436

Illustrative example
Cont...

- The most commonly used criterion for the **best fitting line** is the line that minimizes the sum of the squared errors of prediction.
- That is the criterion that was used to find the line in Figure.
- The last column in Table shows the squared errors of prediction.
- The sum of the squared errors of prediction shown in Table is lower than it would be for any other regression line.

Least square method

- In the least square method, we minimize the sum of square of differences of observed y_i and $\hat{y}_i$.
- Let us consider n pairs (x_i, y_i) for $i = 1, 2, 3, ..., n$.
- The liner regression model is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and we wish to compute a line of the form $\hat{y} = \beta_0 + \beta_1 x$ and is termed as regression line of y on x.

Least square method

Parameter estimation

- In the regression line of y on x, that is $\hat{y} = \beta_0 + \beta_1 x$, the unknown parameters β_0 and β_1 should be estimated.
- As an estimation method we use **least square method**.

- Let

$$
\begin{aligned}
y_i &- \text{observed value} \\
\hat{y}_i &- \text{estimated value} \\
\epsilon_i &= y_i - \hat{y}_i \\
S &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\epsilon_i^2
\end{aligned}
$$

- Our intension is to estimate values of β_0 and β_1 by minimizing S.

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\
\hat{y}_i &= \beta_0 + \beta_1 x_i \\
S &= \sum_{i=1}^{n} [y_i - \beta_0 - \beta_1 x_i]^2 \\
\frac{\partial S}{\partial \beta_0} &= 2\sum_{i=1}^{n} [y_i - \beta_0 - \beta_1 x_i](-1) \\
\frac{\partial S}{\partial \beta_1} &= 2\sum_{i=1}^{n} [y_i - \beta_0 - \beta_1 x_i](-x_i)
\end{aligned}
$$

- Equating these partial derivatives to zero we get normal equations.

$$
\begin{aligned}
\frac{\partial S}{\partial \beta_0} &= 0 \\
\Rightarrow 2\sum_{i=1}^{n}\left[y_i - \beta_0 - \beta_1 x_i\right](-1) &= 0 \\
\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i &= 0 \qquad (1)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial S}{\partial \beta_1} &= 0 \\
\Rightarrow \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 &= 0 \qquad (2)
\end{aligned}
$$

Least square method

Parameter estimation⇒Cont...

$$(1) \times \sum_{i=1}^{n} x_i \quad \Rightarrow$$

$$\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - n\beta_0 \sum_{i=1}^{n} x_i - \beta_1 \left(\sum_{i=1}^{n} x_i\right)^2 = 0 \qquad (3)$$

$$(2) \times n \quad \Rightarrow$$

$$n\sum_{i=1}^{n} x_i y_i - n\beta_0 \sum_{i=1}^{n} x_i - \beta_1 n \sum_{i=1}^{n} x_i^2 = 0 \qquad (4)$$

(4)-(3)⇒

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \\
&= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\
\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x}
\end{aligned}$$

Note: $\hat{\beta}_0$ and $\hat{\beta}_1$ are termed as least square estimates . The regression line is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}).\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

$$\hat{\beta}_1 = r\frac{\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

$$\hat{\beta}_1 = \frac{rS_y}{S_x} \text{ in the case of sample}$$

$$\hat{\beta}_1 = \frac{r\sigma_y}{\sigma_x} \text{ in the case of population.}$$

$$\text{where } \sigma_x = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \overline{x})^2}{N}} \qquad \sigma_y = \sqrt{\sum_{i=1}^{N} \frac{(y_i - \overline{y})^2}{N}}$$

$$S_x = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}} \qquad S_y = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \overline{y})^2}{n-1}}.$$

Example 1

(a) Fit a least square line to the following data.

(b) Estimate y if $x = 7$.

x_i	y_i
1	2
2	5
3	3
4	8
5	7

Example 1
Solution

x_i	y_i	$x_i y_i$	x_i^2
1	2	2	1
2	5	10	4
3	3	9	9
4	8	32	16
5	7	35	25
15	25	88	55

Example 1
Solution⇒Cont...

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \dfrac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \\
&= \frac{88 - \dfrac{15 \times 25}{5}}{55 - \dfrac{(15)^2}{5}} \\
&= 1.3
\end{aligned}
$$

Example 1

Solution⇒Cont...

$$\begin{aligned}\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1\overline{x} \\ &= \frac{25}{5} - 1.3\frac{15}{5} \\ &= 1.1\end{aligned}$$

The equation of least square line becomes $y = 1.1 + 1.3x$.

The value of y when $x = 7$ is $1.1 + 1.3 \times 7 = 10.2$.

Example 2

The data on age and days after arthroscopic shoulder surgery before being able to return to thier sport, for 10 weight lifters is as follows.

Age	33	31	32	28	33	26	34	32	28	27
Recovery time	6	4	4	1	3	3	4	2	3	2

(i) Determine the least squares equation to predict recovery time for injured athletes when age is given?

(ii) Estimate the recovery time for an injured athlete if his age is 29.

The corresponding table is,

x_i	y_i	$x_i y_i$	x_i^2
33	6	198	1089
31	4	124	961
32	4	128	1024
28	1	28	784
33	3	99	1089
26	3	78	676
34	4	136	1156
32	2	64	1024
28	3	84	784
27	2	54	729
304	32	993	9316

Example 2

Solution⇒Cont...

(i)

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \dfrac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \\
&= \frac{993 - \dfrac{304 \times 32}{10}}{9316 - \dfrac{(304)^2}{10}} \\
&= \frac{993 - 972.8}{9316 - 9241.6} \\
&= \frac{20.2}{74.4} \\
&= 0.2715
\end{aligned}
$$

$$\begin{aligned}\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1\overline{x} \\ &= \frac{32}{10} - (0.2715) \times \frac{304}{10} \\ &= -5.0536\end{aligned}$$

The least squares equation of recovery time on age is

recovery time= 0.2715 age -5.0536.

(ii) The recovery time for injured athletes when age is 29,

recovery time= 0.2715 age -5.0536=0.2715 $\times$ 29-5.0536=2.82

Approximately three days are needed for recovery.

Coefficient of determination

- It is used to check the suitability of the model.
- The coefficient of determination is defined as,

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

- If $R^2 \to 1$ the regression line is appropriate for the given set of data.
- Otherwise it is necessary to find some other models.

Coefficient of determination
Properties

1. $0 \leq R^2 \leq 1$.
2. $R^2 = r^2$.
3. $R^2 = \dfrac{\hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$.

Coefficient of determination

Properties⇒Proof of (3)

$$
\begin{aligned}
R^2 &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \text{ by definition.} \\
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\
\overline{y} &= \hat{\beta}_0 + \hat{\beta}_1 \overline{x} \\
\hat{y}_i - \overline{y} &= \hat{\beta}_1 (x_i - \overline{x}) \\
R^2 &= \frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}
\end{aligned}
$$

Example 1

The data below give details of x and y such that

x – water content of skew on 1^{st} may in a river

y – water yield in the river after 3-months from that day

The data have been recorded for 10 years.

(a) Obtain the regression line of y on x.

(b) Find the coefficient of determination.

(c) Discuss the suitability of the model.

Example 1
Cont...

x_i	y_i
23.1	10.5
32.8	16.7
31.8	18.2
32.0	17.0
30.4	16.3
24.0	10.5
39.5	23.1
24.2	12.4
52.5	24.9
37.9	22.8

Example 1
Solution

x_i	y_i	$x_i y_i$	x_i^2
23.1	10.5	242.55	533.61
32.8	16.7		
31.8	18.2		
32.0	17.0		
30.4	16.3		
24.0	10.5		
39.5	23.1		
24.2	12.4		
52.5	24.9		
37.9	22.8		
328.2	172.4	6044.49	11483.4

Example 1
Solution⇒Cont...

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \\
&= \frac{6044.49 - \frac{328.2 \times 172.4}{10}}{11483.4 - \frac{(328.2)^2}{10}} \\
&= \frac{386.322}{711.876} \\
&= 0.543
\end{aligned}
$$

$$
\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{10} x_i}{10} = \frac{328.2}{10} = 32.82 \\
\overline{y} &= \frac{\sum_{i=1}^{10} y_i}{10} = \frac{172.4}{10} = 17.24 \\
\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} \\
&= 17.24 - (0.543) \times 32.82 \\
&= -0.58
\end{aligned}
$$

The regression line of y on x is $y = -0.58 + 0.543x$.

Example 1

Solution⇒Cont...

x_i	y_i	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
23.1	10.5	94.4784	45.4276
32.8	16.7		
31.8	18.2		
32.0	17.0		
30.4	16.3		
24.0	10.5		
39.5	23.1		
24.2	12.4		
52.5	24.9		
37.9	22.8		
328.2	172.4	711.876	240.364

Example 1
Solution⇒Cont...

$$\begin{aligned}\hat{\beta}_1 &= 0.543\\ R^2 &= \frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(x_i-\overline{x})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}\\ &= \frac{(0.543)^2(711.876)}{240.367}\\ &= 0.87\end{aligned}$$

The value is closed to 1. Therefore the obtained line is suitable.

Example 2

In an effort to determine the effective duration of a tranquilizer for animals, the concentration of the substance in blood samples taken at various times after the injection is as follows.

Elapsed time (hours)	1	2	3	6	12	18
Concentration	1.8	1.4	1.2	0.9	0.5	0.1

(a) Determine the least squares equation that relates the concentration to the elapsed time after the injection.

(b) Check the suitability of the model.

(c) Estimate the concentration of the substance in a blood sample taken after four hours of the injection.

Example 2
Solution

(a) The corresponding table is,

x_i	y_i	$x_i y_i$	x_i^2
1	1.8	1.8	1
2	1.4	2.8	4
3	1.2	3.6	9
6	0.9	5.4	36
12	0.5	6.0	144
18	0.1	1.8	324
42	5.9	21.4	518

Example 2
Solution

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \dfrac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \\
&= \frac{21.4 - \dfrac{42 \times 5.9}{6}}{518 - \dfrac{(42)^2}{6}} \\
&= \frac{21.4 - 41.3}{518 - 294} \\
&= -\frac{19.9}{224} \\
&= -0.0888393
\end{aligned}
$$

Example 2
Solution

$$
\begin{aligned}
\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} \\
&= \frac{5.9}{6} - (-0.0888393) \times \frac{42}{6} \\
&= 1.6052
\end{aligned}
$$

The least squares equation that relates the concentration to the elapsed time after the injection is

concentration=1.6052-0.0888393 time.

Example 2
Solution

(b) To check the suitability of model, we have to find value of coefficient of determination.

x_i	y_i	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
1	1.8	36	0.6724
2	1.4	25	0.1764
3	1.2	16	0.0484
6	0.9	1	0.0064
12	0.5	25	0.2304
18	0.1	121	0.7744
		224	1.9084

Example 2
Solution

$$\begin{aligned}
\hat{\beta}_1 &= -0.0888393 \\
R^2 &= \frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \\
&= \frac{(-0.0888393)^2(224)}{1.9084} \\
&= 0.9264
\end{aligned}$$

The value is closed to 1. Therefore the obtained model is suitable to predict concentration of the substance in a blood sample given elapsed time.

(c) The concentration of the substance in a blood sample taken after four hours of the injection is,

$$\begin{aligned} \text{concentration} &= 1.6052 - 0.0888393\,\text{time} \\ &= 1.6052 - 0.0888393 \times 4 \\ &= 1.2498 \end{aligned}$$

Thank You