Linear Regression

Linear regression

- Deals with relationship between two variables X and Y.
- Y is the variables whose "behavior" we wish to study (e.g., fuel efficiency in a car).
- X is the variable we believe would help explain the behavior of Y (e.g., the size of the car).

Regression model

- The simple linear regression model:

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

Y = Dependent/response variable
X = Independent/explanatory variable
(X is the predictor variable)
ε = Random error term
(captures unexplained variation in Y)
β_0 = Y-intercept
β_1 = Slope of line

Components of the models

Non random component = $\beta_0 + \beta_1 X$

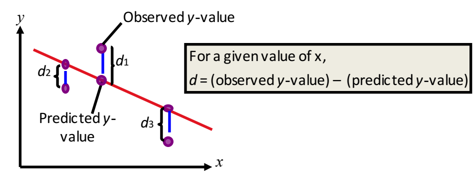
Random component = ε

- Since ε is a random variable, Y is also a random variable since Y, in part, depends on ε
- We wish to find the expected value of Y:

  $E(Y) = \beta_0 + \beta_1 X$
- As X increases, Y increases, on average, if $\beta_1 > 0$
- As X increases, Y decreases, on average, if $\beta_1 < 0$

Residuals

- After verifying that the linear correlation between two variables is significant, next we determine the equation of the line that can be used to predict the value of *y* for a given value of *x*.



Each data point d_i represents the difference between the observed *y*-value and the predicted *y*-value for a given *x*-value on the line. These differences are called **residuals**.

Regression Line

- A **regression line**, also called a **line of best fit**, is the line for which the sum of the squares of the residuals is a minimum.

The Equation of a Regression Line

The equation of a regression line for an independent variable *x* and a dependent variable *y* is

ŷ = mx + b

where $\hat{y}$ is the predicted *y*-value for a given *x*-value. The slope *m* and *y*-intercept *b* are given by

$$m = \frac{n\sum xy \quad (\sum x)(\sum y)}{n\sum x^2 \quad (\sum x)^2} \text{ and } b = \bar{y} \quad m\bar{x} = \frac{y}{n} \quad m\frac{x}{n}$$

where $\bar{y}$ is the mean of the y-values and $\bar{x}$ is the mean of the x-values. The regression line always passes through $(\bar{x}, \bar{y})$.

Regression Line

Example:

Find the equation of the regression line.

x	y
1	– 3
2	– 1
3	0
4	1
5	2

Continued.

Regression Line

Example:

Find the equation of the regression line.

x	y	xy	x^2	y^2
1	– 3	– 3	1	9
2	– 1	– 2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4

Continued.

Regression Line

Example:

Find the equation of the regression line.

x	y	xy	x^2	y^2
1	– 3	– 3	1	9
2	– 1	– 2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
$x = 15$	$y = \ 1$	$xy = 9$	$x^2 = 55$	$y^2 = 15$

Continued.

Regression Line

Example:

Find the equation of the regression line.

x	y	xy	x^2	y^2
1	– 3	– 3	1	9
2	– 1	– 2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
$x = 15$	$y = \ \ 1$	$xy = 9$	$x^2 = 55$	$y^2 = 15$

$$m = \frac{n\sum xy \quad (\sum x)(\sum y)}{n\sum x^2 \quad (\sum x)^2} = \frac{5(9) \quad (15)(\ \ 1)}{5(55) \quad (15)^2} = \frac{60}{50} = 1.2$$

Continued.

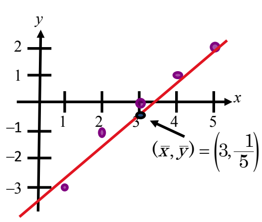
Regression Line

Example continued:

$$b = \bar{y} \quad m\bar{x} = \frac{\ \ 1}{5} \quad (1.2)\frac{15}{5} = \quad 3.8$$

The equation of the regression line is

$\hat{y} = 1.2x - 3.8.$



Regression Line

Example:

The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

a.) Find the equation of the regression line.

b.) Use the equation to find the expected test score for a student who watches 9 hours of TV.

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50
xy	0	85	164	222	285	340	380	420	348	455	525	500
x^2	0	1	4	9	9	25	25	25	36	49	49	100
y^2	9216	7225	6724	5476	9025	4624	5776	7056	3364	4225	5625	2500

$x = 54$ $\quad y = 908$ $\quad xy = 3724$ $\quad x^2 = 332$ $\quad y^2 = 70836$

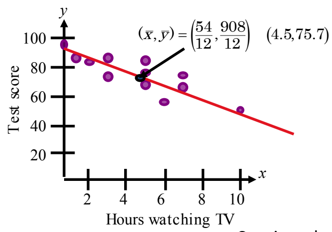
Regression Line

Example continued:

$$m = \frac{n\sum xy \quad (\sum x)(\sum y)}{n\sum x^2 \quad (\sum x)^2} = \frac{12(3724) \quad (54)(908)}{12(332) \quad (54)^2} \quad 4.067$$

$$b = \bar{y} \quad m\bar{x}$$
$$= \frac{908}{12} \quad (\quad 4.067)\frac{54}{12}$$
$$93.97$$

$\hat{y} = -4.07x + 93.97$



Continued.

Regression Line

Example continued:

Using the equation $\hat{y}$ = -4.07x + 93.97, we can predict the test score for a student who watches 9 hours of TV.

$$\hat{y} = -4.07x + 93.97$$
$$= -4.07(9) + 93.97$$
$$= 57.34$$

A student who watches 9 hours of TV over the weekend can expect to receive about a 57.34 on Monday's test.

Coefficient of Determination

- The **coefficient of determination R^2** is the ratio of the explained variation to the total variation. That is,

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

Example:

- The correlation coefficient for the data that represents the number of hours students watched television and the test scores of each student is $r \approx -0.831$. Find the coefficient of determination.

$$R^2 \approx (-0.831)^2$$
$$0.691$$

About 69.1% of the variation in the test scores can be explained by the variation in the hours of TV watched. About 30.9% of the variation is unexplained.

Regression hypothesis

- Regression equation: $Y = \beta_0 + \beta_1 X$
- H_0: $\beta_1 = 0$ (no regression relationship exists)
- H_1: $\beta_1 \neq 0$ (there is a regression relationship)

F-test

- F-test is a test for the entire regression
- The calculated F statistic is as follows:

$$F_{(k-1,n-k)} = \frac{MSR}{MSE} = \frac{\frac{SSR}{k-1}}{\frac{SSE}{n-k}}$$

$k = 2$ (for b_0 and b_1) and n = Sample size

- Decision rule: Reject H_0 if $F_{CALC} > F_{CV}$

Regression

Test of Significance

for β_1

(slope of regression line)

Hypothesis for Slope, β_1

- Regression model: $Y = \beta_0 + \beta_1 X + \varepsilon$
- H_0: $\beta_1 = 0$ (X has no impact on Y)
- H_1: $\beta_1 \neq 0$

t-test for β_1

- t-test is a test for only β_1 (not entire regression)

$$t_{n-k} = \frac{b_1}{Std\ error\ of\ b_1}$$

$$Std.\,error\ of\ b_1 : S(b_1) = \frac{\sqrt{MSE}}{\sqrt{\sum (X - \bar{X})^2}}$$

DF = n-k, where n = sample size and k = 2

- Decision rule: Reject H_0 if $t_{CALC} > t_{CV}$