Regression Analysis

Cesar Acosta Ph.D.

Department of Industrial and Systems Engineering
University of Southern California

CORRELATION

The ***coefficient of correlation*** *can be used* to test for a linear relationship between two variables.

The range of the coefficient of correlation is [–1, +1]

- If $r = -1$ (negative association)
- If $r = +1$ (positive association)
- If $r = 0$ (no association)

REGRESSION ANALYSIS

Regression analysis is useful to find a relationship between a response and a set of predictors

The relation can be used to predict the value of the response

Response variable: **Y**
predictors : $\mathbf{X_1, X_2, \ldots, X_k}$

REGRESSION ANALYSIS

Regression analysis is useful to find a relationship between a response and a set of predictors

Two Regression Models

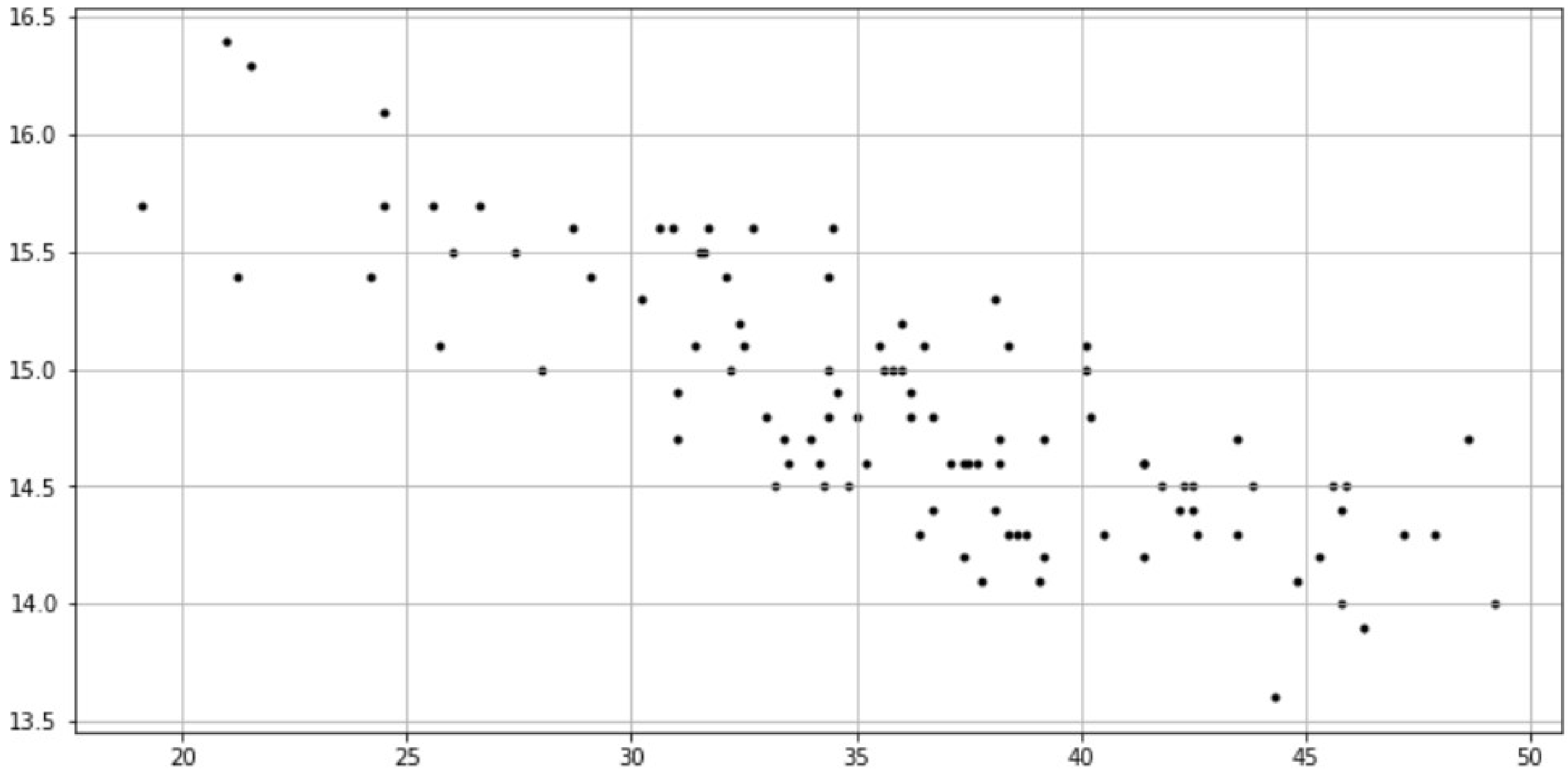
- Simple linear regression (SLR)

- Multiple linear regression (MLR)

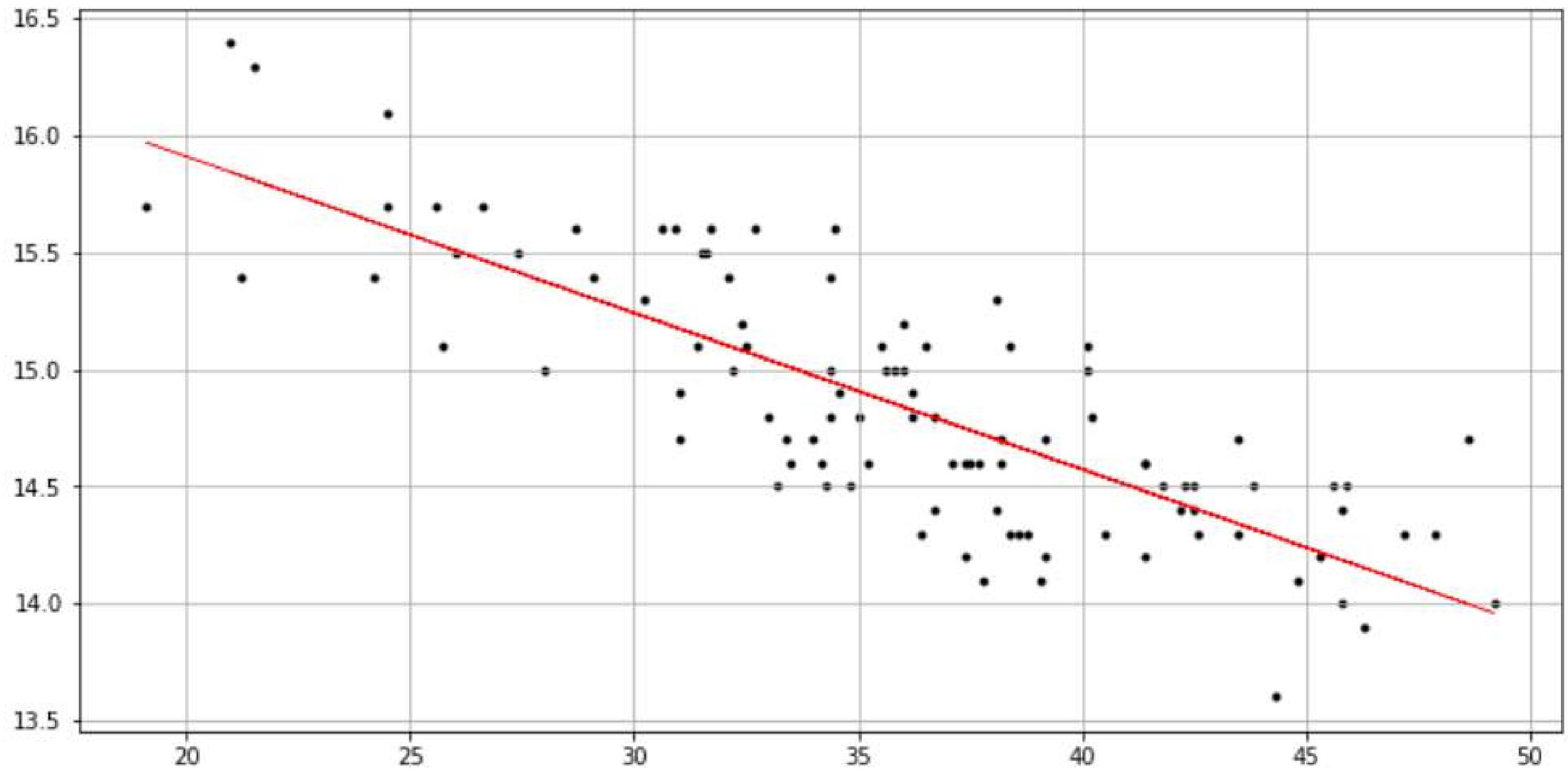
OLS line

Least Squares line
(OLS)

scatterplot



OLS line



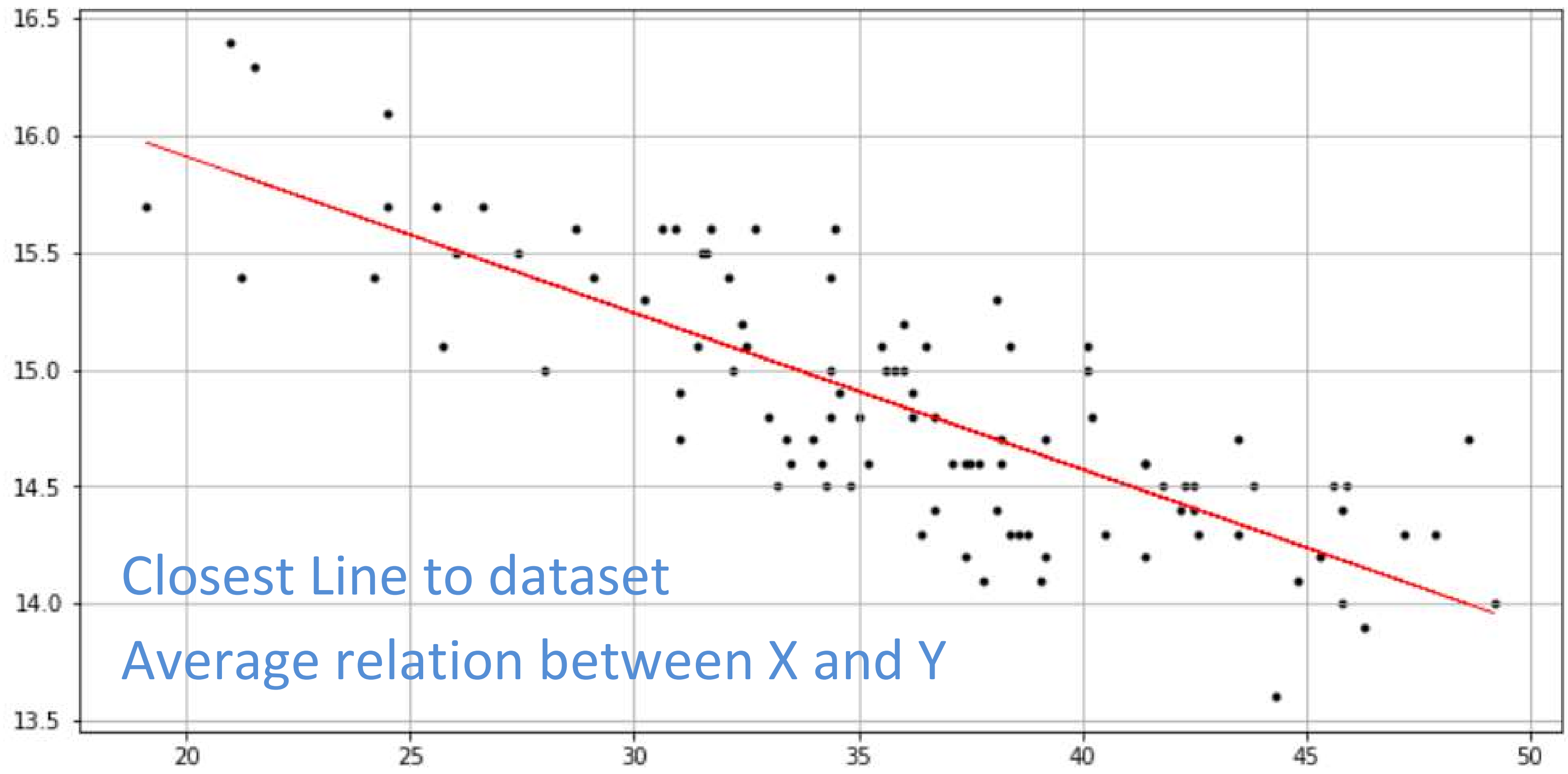
OLS line

What is the OLS line?

OLS line

- X, Y not random variables
- No statistics
- Not a regression line

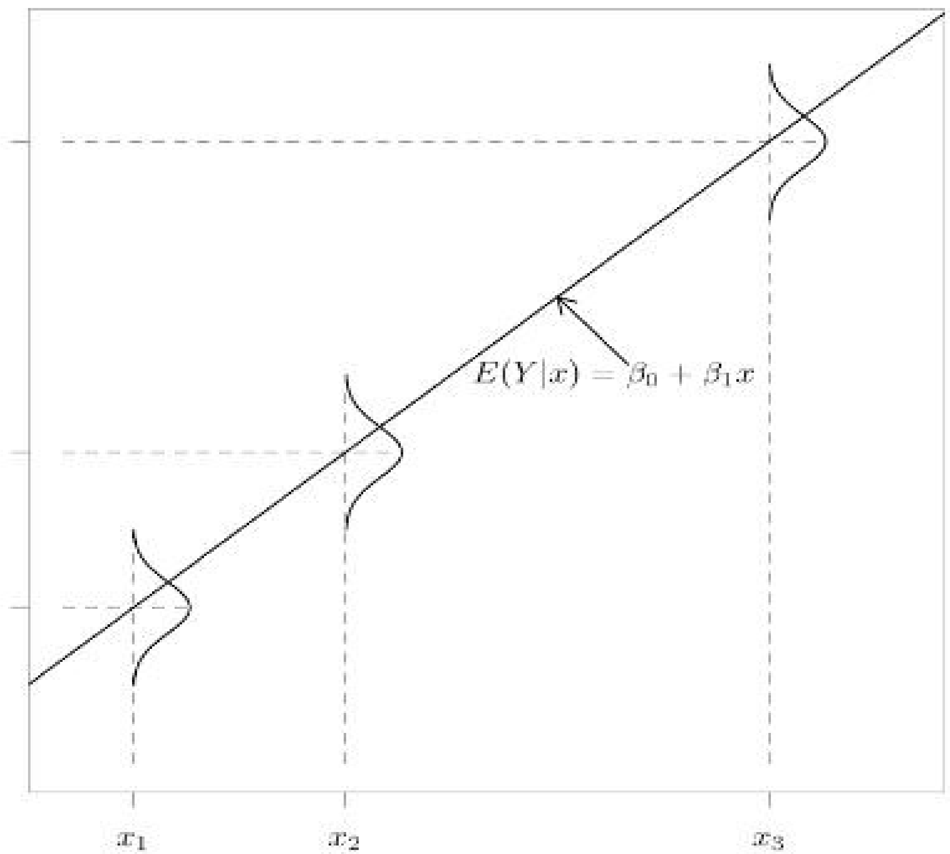
OLS line


16.5
16.0
15.5
15.0
14.5
14.0
13.5
20
25
30
35
40
45
50
Closest Line to dataset
Average relation between X and Y

Regression line

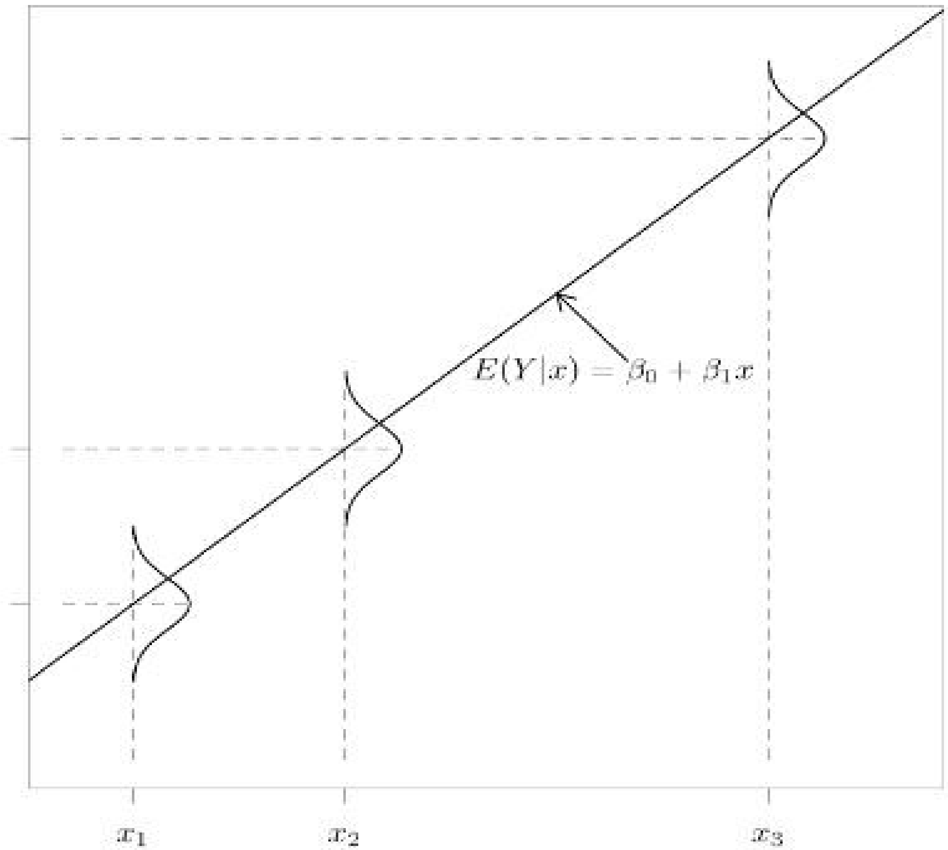
Regression relation

- Y is random variable
- X is not random
- Mean of Y varies with X



Regression relation

- This is an unknown relation
- We will try to estimate it from an OLS line
- Obtained from a random sample



Regression relation

- At each value of X, say *x,*

  Y is a normal random variable

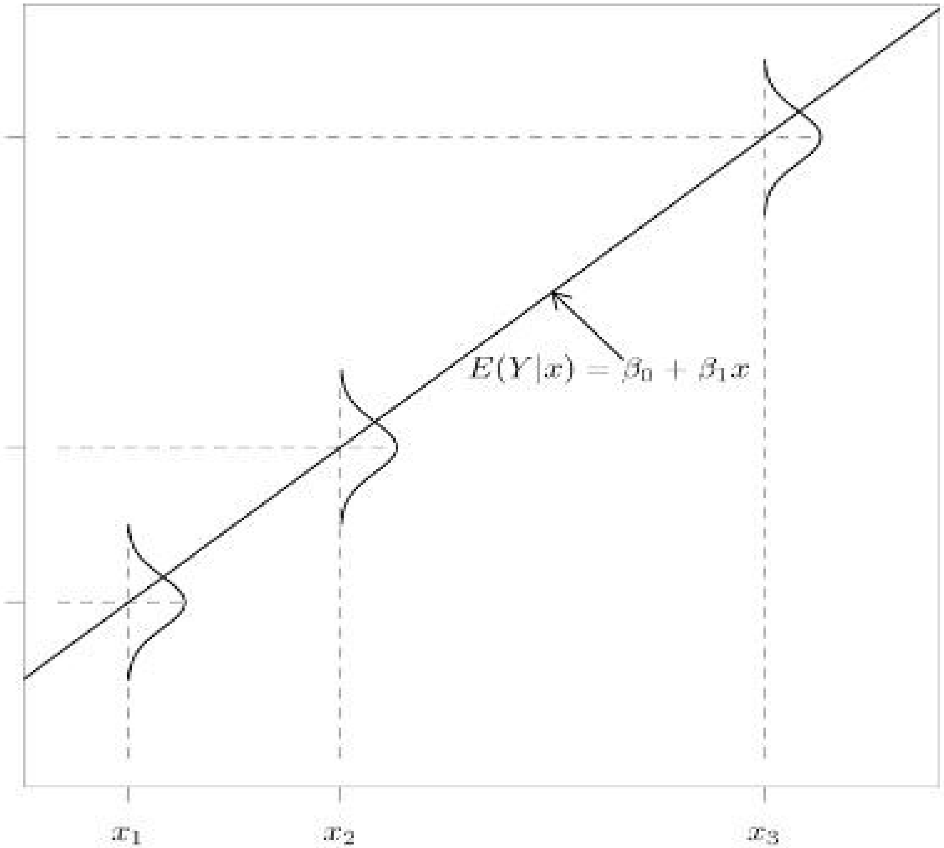
- with mean that changes with X

  $$\beta_0 + \beta_1 x$$

- There is one rv. Y for each X
- All rv. Y with same variance σ^2

  $$Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

- All Y variables are independent

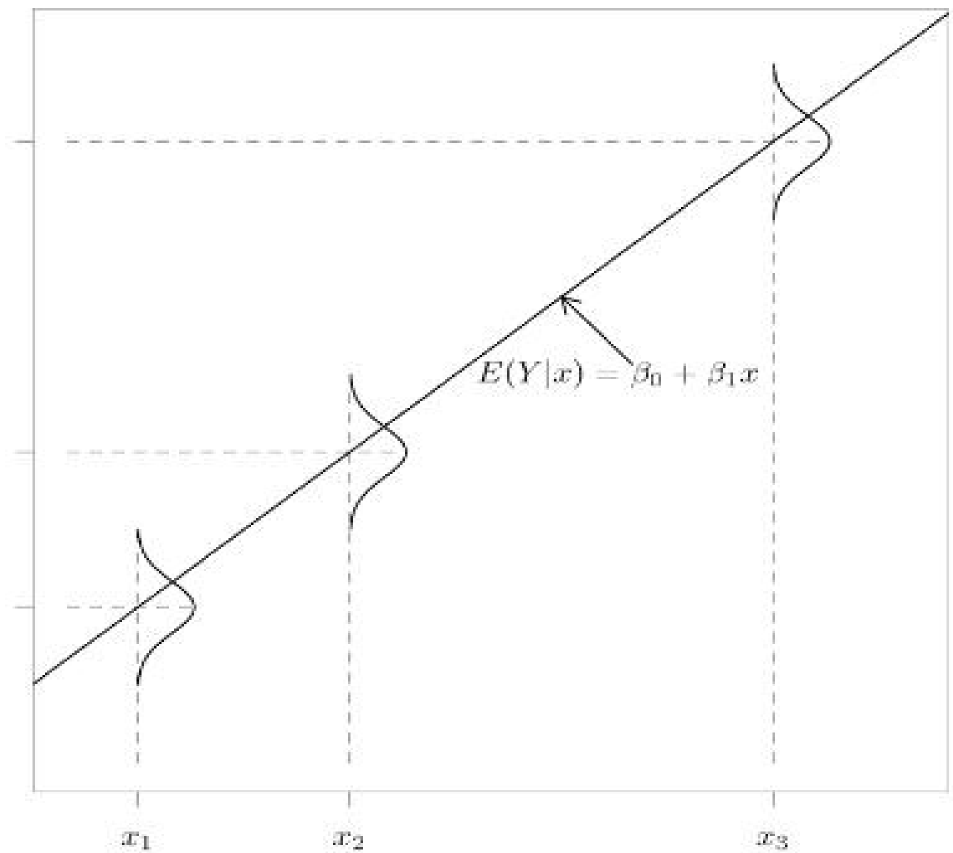


Regression relation

- Notice that the line is

$$E[Y] = \beta_0 + \beta_1 x$$

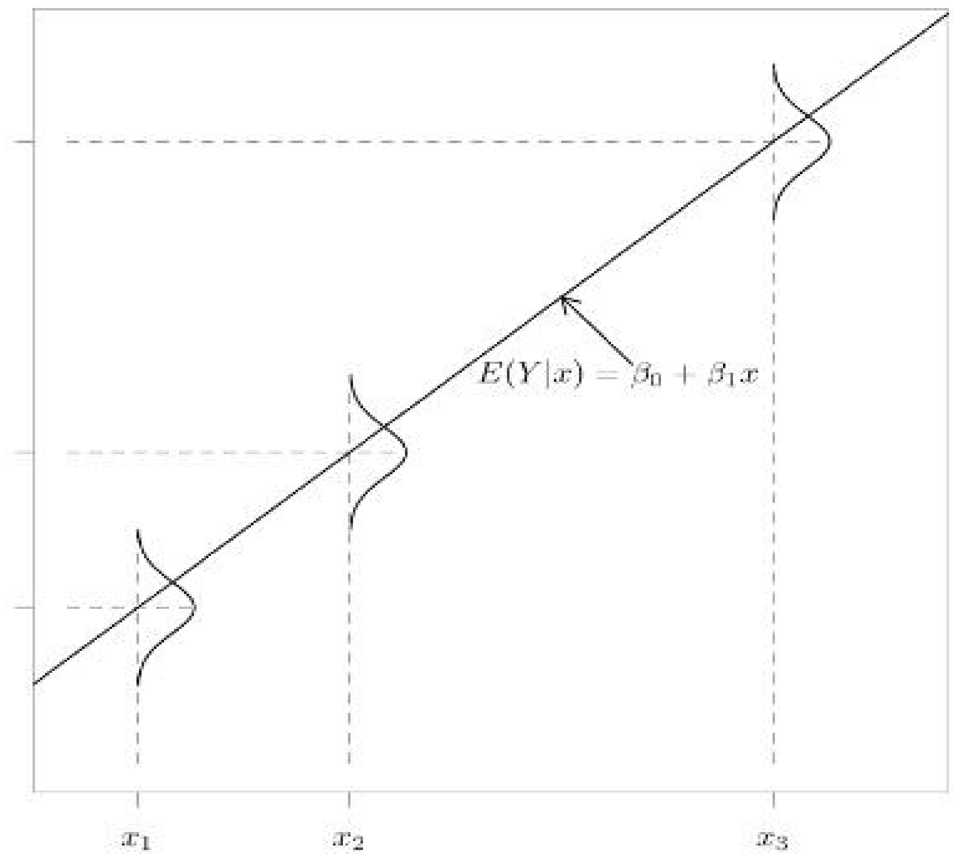
(no random term)



Regression relation

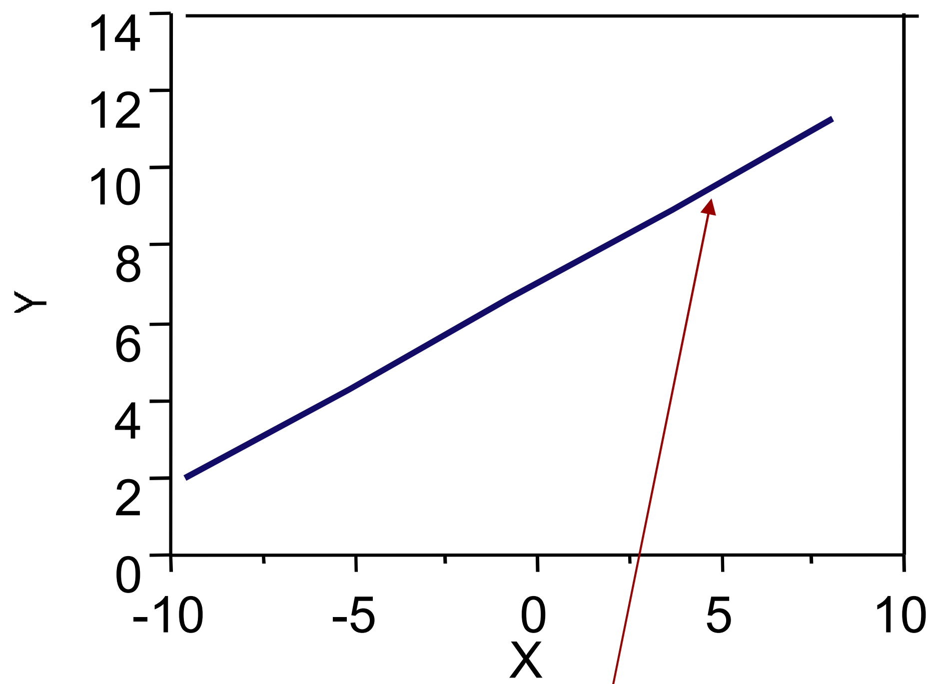
- The mean of Y changes with X

- The regression relation is between X (not random) and the mean of variable Y



Regression relation

This is an unknown relation



True population line (unknown) $E[Y] = \beta_0 + \beta_1 x$

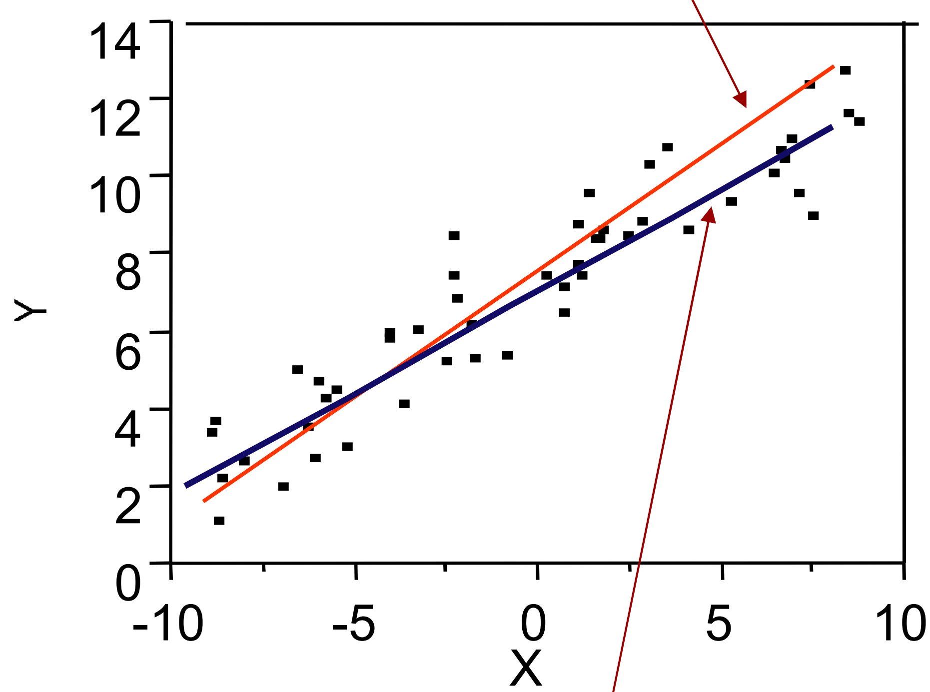
Regression relation

This is an unknown relation

We will try to estimate it

from a random sample

Estimated (least squares) line $b_0 + b_1 x$



True population line (unknown) $E[Y] = \beta_0 + \beta_1 x$

Regression assumptions

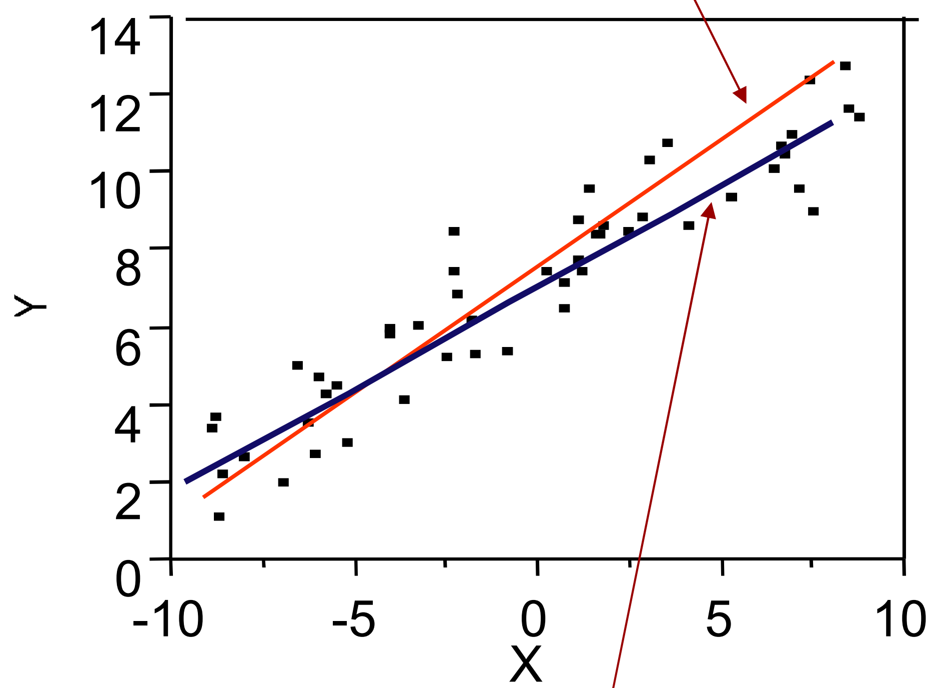
$Y_1, Y_2, \ldots, Y_n$ are random vars.

independent (*independence)*

normal (*normality)*

with same variance
(*constant variance)*

Estimated (least squares) line $b_0 + b_1 x$



True population line (unknown) $E[Y] = \beta_0 + \beta_1 x$

Example

COEFFICIENTS TABLE

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.248727	0.182093	94.72	<2e-16 ***
Odometer	-0.066861	0.004975	-13.44	<2e-16 ***

GOODNESS OF FIT - MEASURES

Residual standard error: 0.3265 on 98 degrees of freedom
Multiple R-squared: 0.6483, Adjusted R-squared: 0.6447
F-statistic: 180.6 on 1 and 98 DF, p-value: < 2.2e-16

Example

COEFFICIENTS TABLE

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	17.248727	0.182093	94.72	<2e-16	***
Odometer	-0.066861	0.004975	-13.44	<2e-16	***

GOODNESS OF FIT - MEASURES

Residual standard error: 0.3265 on 98 degrees of freedom
Multiple R-squared: 0.6483, Adjusted R-squared: 0.6447
F-statistic: 180.6 on 1 and 98 DF, p-value: < 2.2e-16

The variance σ^2 is estimated to be $0.3265^2 = 0.1066$

Example

COEFFICIENTS TABLE

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.248727	0.182093	94.72	<2e-16 ***
Odometer	-0.066861	0.004975	-13.44	<2e-16 ***

GOODNESS OF FIT - MEASURES

Residual standard error: 0.3265 on 98 degrees of freedom
Multiple R-squared: 0.6483, Adjusted R-squared: 0.6447
F-statistic: 180.6 on 1 and 98 DF, p-value: < 2.2e-16

The variance σ^2 is estimated to be $0.3265^2 = 0.1066$

64.83% of variation of Y is explained by X

Goodness of Fit

The least squares method will always produce a straight line,

even if there is no relationship between the variables, or

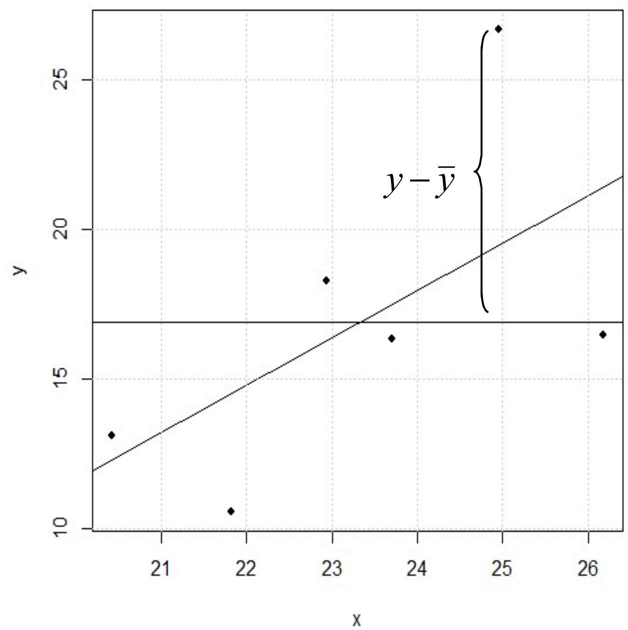
if the relationship is other than linear

Hence, in addition to determining the coefficients of the least squares line, we need to assess it,

to see how well it "fits" the data.

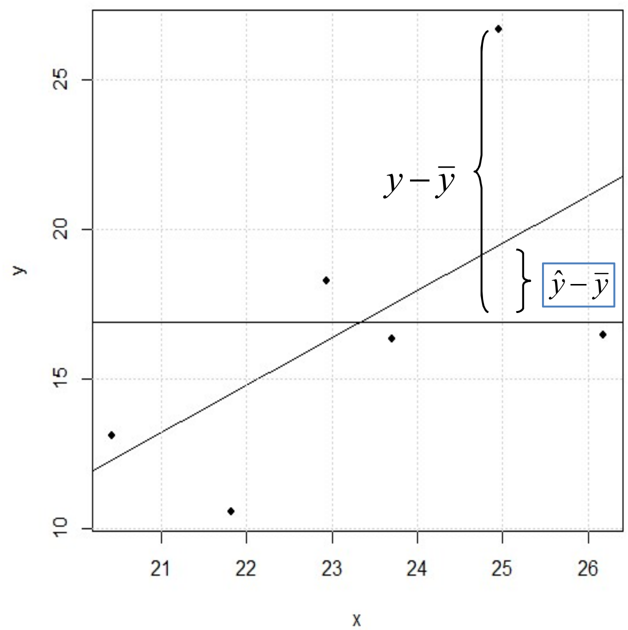
Goodness of Fit

$y-\bar{y}$ change in y (as x increases)



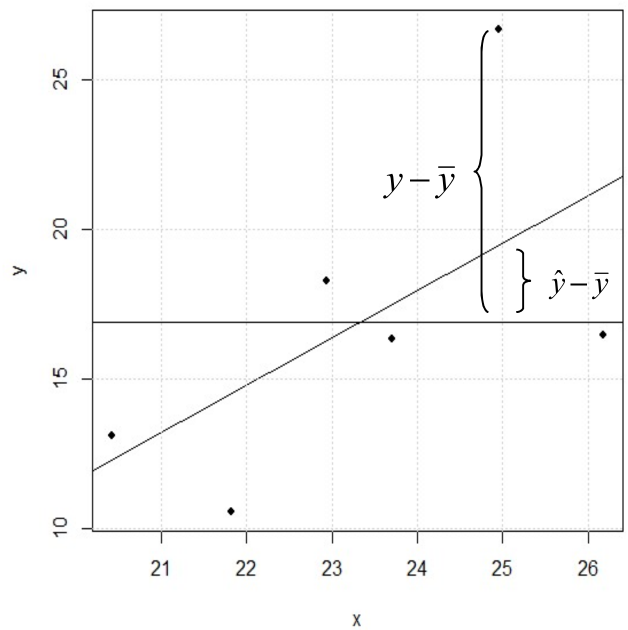
Goodness of Fit

$y - \bar{y}$	change in y	(as x increases)
$\hat{y} - \bar{y}$	change in y explained by $\hat{y}$	(as x increases)



Goodness of Fit

$y - \bar{y}$	change in *y*	(as *x* increases)
$\hat{y} - \bar{y}$	change in *y* explained by	(as *x* increases)
$y - \hat{y}$	change in *y* not explained by	(as *x* increases)



Goodness of Fit

$y - \bar{y}$	change in *y*	(as *x* increases)
$\hat{y} - \bar{y}$	change in *y* explained by	(as *x* increases)
$y - \hat{y}$	change in *y* not explained by	(as *x* increases)

Goodness of Fit

$y-\overline{y}$	change in *y*	(as *x* increases)
$\hat{y}-\overline{y}$	change in *y* explained by	(as *x* increases)
$y-\hat{y}$	change in *y* not explained by	(as *x* increases)

$$SSTotal = \sum_{i=1}^{n}(y-\overline{y})^2$$

$$SSR = \sum_{i=1}^{n}(\hat{y}-\overline{y})^2$$

$$SSE = \sum_{i=1}^{n}(y-\hat{y})^2$$

$$SSTotal = SSR + SSE$$

Goodness of Fit

$y - \overline{y}$	change in *y*	(as *x* increases)
$\hat{y} - \overline{y}$	change in *y* explained by	(as *x* increases)
$y - \hat{y}$	change in *y* not explained by	(as *x* increases)

$$SSTotal = \sum_{i=1}^{n} (y - \overline{y})^2 \qquad MSTotal = \frac{SSTotal}{n-1}$$

$$SSR = \sum_{i=1}^{n} (\hat{y} - \overline{y})^2 \qquad MSR = \frac{SSR}{1}$$

$$SSE = \sum_{i=1}^{n} (y - \hat{y})^2 \qquad MSE = \frac{SSE}{n-2}$$

$$SSTotal = SSR + SSE \qquad MSTotal \neq MSR + MSE$$

Example – MEAN SQUARED ERROR (MSE)

COEFFICIENTS TABLE

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.248727	0.182093	94.72	<2e-16 ***
Odometer	-0.066861	0.004975	-13.44	<2e-16 ***

GOODNESS OF FIT - MEASURES

Residual standard error: 0.3265 on 98 degrees of freedom
Multiple R-squared: 0.6483, Adjusted R-squared: 0.6447
F-statistic: 180.6 on 1 and 98 DF, p-value: < 2.2e-16

The variance σ^2 is estimated to be 0.3265^2 = 0.1066 = MSE

64.83% of variation of Y is explained by X

Example – RESIDUAL STANDARD ERROR

COEFFICIENTS TABLE

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.248727	0.182093	94.72	<2e-16 ***
Odometer	-0.066861	0.004975	-13.44	<2e-16 ***

GOODNESS OF FIT - MEASURES

Residual standard error: 0.3265 on 98 degrees of freedom
Multiple R-squared: 0.6483, Adjusted R-squared: 0.6447
F-statistic: 180.6 on 1 and 98 DF, p-value: < 2.2e-16

The variance σ^2 is estimated to be $0.3265^2 = 0.1066 = MSE$

$S = 0.3265 = \sqrt{MSE}$ is average distance to regression line

Example – RESIDUAL STANDARD ERROR

$$SSTotal = \sum_{i=1}^{n} (y - \overline{y})^2 \qquad MSTotal = \frac{SSTotal}{n-1}$$

sample variance Y

$$SSR = \sum_{i=1}^{n} (\hat{y} - \overline{y})^2 \qquad MSR = \frac{SSR}{1}$$

$$SSE = \sum_{i=1}^{n} (y - \hat{y})^2 \qquad MSE = \frac{SSE}{n-2}$$

$$SSTotal = SSR + SSE \qquad MSTotal \neq MSR + MSE$$

Example – RESIDUAL STANDARD ERROR

$$SSTotal = \sum_{i=1}^{n} (y - \overline{y})^2 \qquad MSTotal = \frac{SSTotal}{n-1}$$

sample variance Y

$$SSR = \sum_{i=1}^{n} (\hat{y} - \overline{y})^2 \qquad MSR = \frac{SSR}{1}$$

$$SSE = \sum_{i=1}^{n} (y - \hat{y})^2 \qquad MSE = \frac{SSE}{n-2}$$

$$SSTotal = SSR + SSE \qquad MSTotal \neq MSR + MSE$$

R-SQUARED

R^2 is the fraction of changes in Y that is explained by X

$$R^2 = \frac{SSR}{SSTotal}$$

R-SQUARED

R^2 is the fraction of changes in Y that is explained by X

$$R^2 = \frac{SSR}{SSTotal}$$

$$= \frac{SSTotal - SSE}{SSTotal}$$

$$= 1 - \frac{SSE}{SSTotal}$$

R-SQUARED

R^2 is always between 0 and 1

0 means no changes in Y have been explained by X

1 means it has all been explained (a perfect fit to the data)

R^2 is also called

Coefficient of multiple determination,

Multiple R-squared

R-SQUARED

R^2 is always between 0 and 1

0 means no changes in Y have been explained by X

1 means it has all been explained (a perfect fit to the data)

R^2 is also called

Coefficient of multiple determination,

Multiple R-squared