3703-3J04 Probability & Statistics for (Civil) Engineering

C01 - Lecture 30

Today LINEAR REGRESSION

X & Y quantities, related linearly, (or at least you suspect this to be true.)
non-deterministic i.e. no formula

We model this with $Y = \beta_0 + \beta_1 X + \varepsilon$

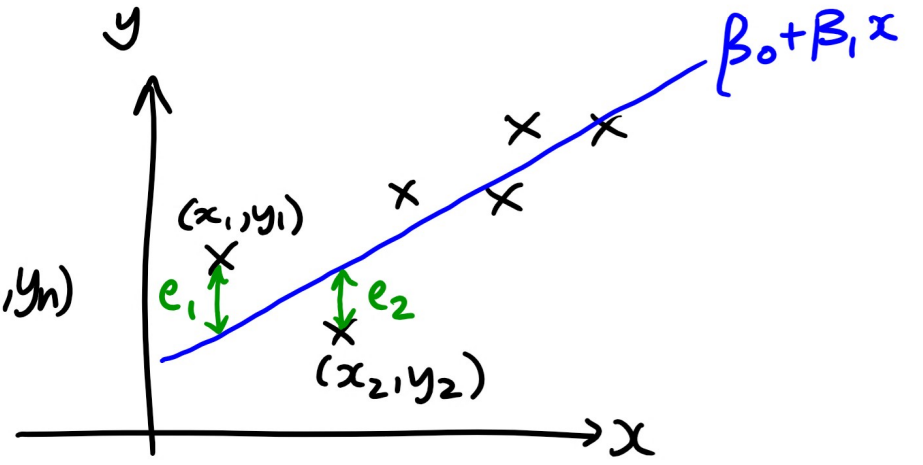
- β_0, β_1: regression coefficients
- X: regressor
- ε: random variable = error, $E(\varepsilon) = 0$

So for a given value x of X
$E(Y) = \beta_0 + \beta_1 x$

β_0, β_1 derived from data! How?

Think scatter plot:

n pairs of observations
$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$



Each y_i modelled by an r.v. $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ for each i. (where ε_i is, for each i, an r.v. with same distribution as ε.)

Want "line of best fit" $y = \beta_0 + \beta_1 x$

To find β_0, β_1 minimise $L = \sum_{i=1}^{n} \varepsilon_i^2$

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

The resulting estimates for β_0, β_1 are called least squares estimates, notated $\hat{\beta}_0, \hat{\beta}_1$ & then the line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is the least squares regression line.

Allows us to estimate y given an x-value.

How to find $\hat{\beta}_0, \hat{\beta}_1$?

Want to minimise $L = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

for varying β_0, β_1.

At minimum of L, we must have $\frac{\partial L}{\partial \beta_0} = 0 = \frac{\partial L}{\partial \beta_1}$

i.e. $-2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 = -2\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \qquad \sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

↑ Least squares normal equations ↑ 2 constraints. ≡

Solving this system of linear equations:

Notation

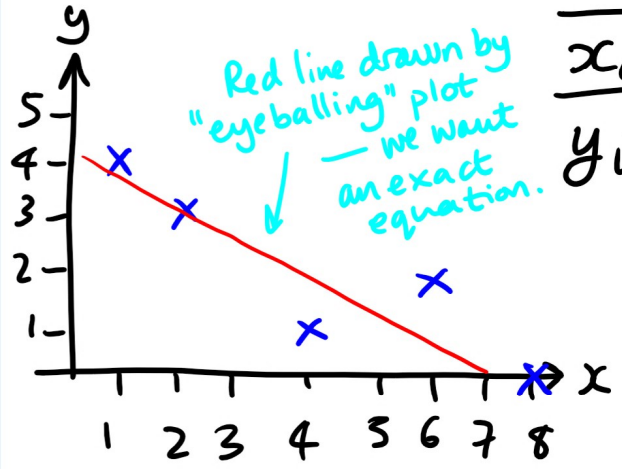
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \quad \begin{matrix} \} =: S_{xy} \\ \} =: S_{xx} \end{matrix}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\text{So } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Example Find the least squares regression line for data:

i	1	2	3	4	5
x_i	1	2	4	6	8
y_i	4	3	1	2	0



Solution

x	y	x^2	y^2	xy
1	4	1	16	4
2	3	4	9	6
4	1	16	1	4
6	2	36	4	12
8	0	64	0	0
Σ 21	10	121	30	26

$$\bar{x} = \frac{21}{5} = 4.2 \qquad \bar{y} = \frac{10}{5} = 2$$

$$S_{xy} = \sum xy - n\bar{x}\bar{y}$$
$$= 26 - 5(4.2)(2)$$
$$= -16$$

$$S_{xx} = \sum x^2 - n\bar{x}^2$$
$$= 121 - 5(4.2)^2$$
$$= 32.8.$$

So $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-16}{32.8} = \underline{-0.49}$.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 2 - (-0.49)(4.2) = \underline{4.05}$$

So least squares regression line is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$$\underline{\hat{y} = 4.05 - 0.49x}.$$

We model $Y = \beta_0 + \beta_1 X + \varepsilon$ ← random error

(or ε_i for a fixed i, if you prefer)

For each i the value ε takes is called the residual (of the data pt (x_i, y_i)), denoted e_i

It equals $y_i - \hat{y}_i$

actual data point ↗ ↖ expected pt given by the line

e.g. in above

$$e_3 = y_3 - \hat{y}_3$$
$$= 1 - (\hat{\beta}_0 + \hat{\beta}_1 x_3)$$
$$= 1 - (4.05 - 0.49(4))$$
$$= \underbrace{1}_{\text{observed value}} - \underbrace{2.09}_{\text{expected value from line (i.e. on line)}} = -1.09$$

So really what we have are n random variables $\varepsilon_1, \dots, \varepsilon_n$ which have same underlying distribution with $E(\varepsilon_i) = 0$, which take values $e_1, \dots, e_n$.

(↳ $E(Y_i) = \beta_0 + \beta_1 E(x_i)$)

We <u>assume</u> $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$ i.e. the errors ε_i are uncorrelated.

<u>How do we estimate $\sigma^2 = V(\varepsilon_i)$?</u>

We use the unbiased estimator

$$\hat{\sigma}^2 = \frac{1}{n-2} SS_E$$ ← SS_E is the <u>sums of squares error</u>

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Since this is tedious to compute, we have a shortcut: $SS_E = SS_T - SS_R$

"Total sum of squares" ↑ "Regression sum of squares"

$$s_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \qquad = \hat{\beta}_1 s_{xy}$$

i.e. $$SS_E = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy}$$

Example In example above, find $\hat{\sigma}^2$.

Solution $$\hat{\sigma}^2 = \frac{1}{n-2} SS_E$$

$$= \frac{1}{5-2}\left(\Sigma y^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy}\right)$$

$$= \frac{1}{3}\left(30 - 5(2)^2 - (-0.49)(-16)\right)$$

$$= \underline{\underline{0.72}}.$$