3Y03 - Probability and Statistics for Engineering

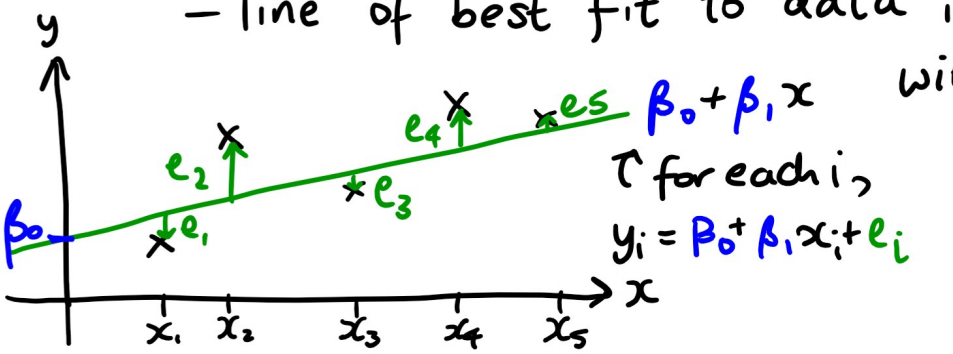
WS19 Lecture 30

Yesterday

Least Squares Regression Line

- (For now) assume X, Y random variables related linearly
- Model this as $Y = \beta_0 + \beta_1 X + \varepsilon$ ← random error, $E(\varepsilon) = 0$.
- Least Squares Regression Line:
  - line of best fit to data i.e. estimate β_0 and β_1 with $\hat{\beta}_0$ and $\hat{\beta}_1$ using data
    ↳ the values of β_0 and β_1 that minimize:

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



↑ for each i, $y_i = \beta_0 + \beta_1 x_i + e_i$

Recall:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}$$

($\sum x_i = n\bar{x}$, $\sum y_i = n\bar{y}$, $\frac{1}{n}(\sum x_i)^2 = n\bar{x}^2$)

$$= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \quad \begin{matrix} =: S_{xy} \\ =: S_{xx} \end{matrix}$$

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

$$\hat{\beta}_0 = \frac{1}{n}\sum y_i - \frac{\hat{\beta}_1}{n}\sum x_i = \bar{y} - \hat{\beta}_1 \bar{x}$$

Example — see separate sheets

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

Points by points $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ with $E(\varepsilon_i) = 0$

So we have n random variables $\varepsilon_1, \ldots, \varepsilon_n$ with same underlying distribution & which take values $\underbrace{e_1, \ldots, e_n}_{\text{residuals}}$

& $e_i = y_i - \hat{y}_i$ (actual value, line value)

$= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

So e.g. $e_3 = y_3 - \hat{y}_3 = y_3 - (\hat{\beta}_0 + \hat{\beta}_1 x_3)$ ($x_3 = 4$, $y_3 = 1$ in our example ($i=3$); in attached example)

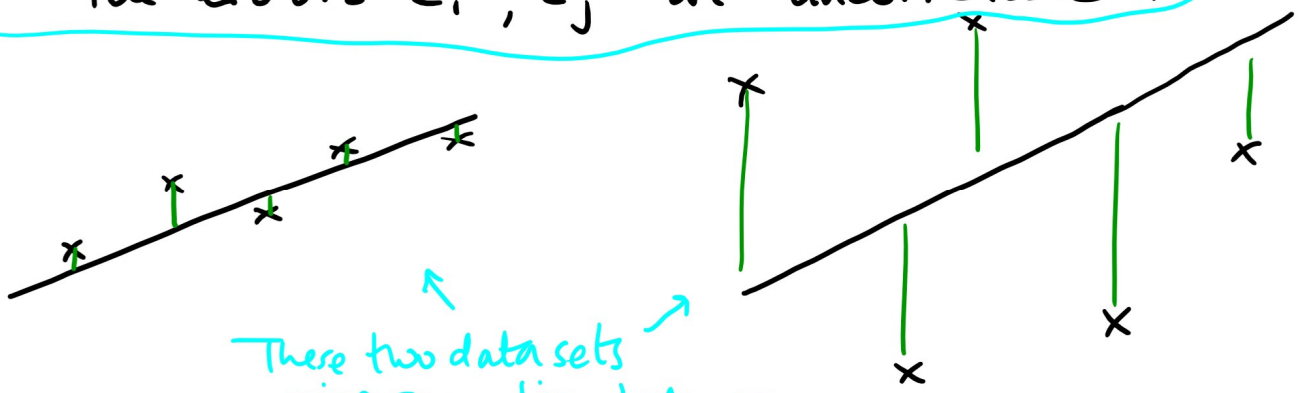
$= 1 - (4.05 - 0.49(4))$

$= 1 - 2.09 = -1.09.$

(1 ↑ actual value; 2.09 ↑ line value)

We assume $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$ i.e.

the errors $\varepsilon_i, \varepsilon_j$ are uncorrelated.



These two data sets give same line, but very different residuals.

Want to estimate $\sigma^2 = V(\varepsilon_i)$.

Use the unbiased estimator $\hat{\sigma}^2 = \frac{1}{n-2} \underbrace{SS_E}$

SS_E : sums of squares error

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

Tedious to compute!

Shortcut: $SS_E = SS_T - SS_R$

"Total sum of squares" ↑ SS_T

"Regression sum of squares" ↑ SS_R

$$S_{yy} = SS_T = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

$$SS_R = \hat{\beta}_1 S_{xy}$$

i.e. $$\hat{\sigma}^2 = \frac{1}{n-2} SS_E = \frac{1}{n-2}\left(SS_T - SS_R\right)$$

$$\left[= \frac{1}{n-2}\left(S_{yy} - \hat{\beta}_1 S_{xy}\right)\right]$$

Example – see sheet

11.3 Properties of the least Squares Estimators $\hat{\beta}_0, \hat{\beta}_1$

For fixed x, $Y = \beta_0 + \beta_1 x + \varepsilon$ where $E(\varepsilon) = 0$

$$\& so \quad E(Y) = \beta_0 + \beta_1 x$$

$$V(Y) = V(\varepsilon) = \sigma^2$$

Find $E(\hat{\beta}_1)$, $V(\hat{\beta}_1)$:

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

$$E(\hat{\beta}_1) = E\left(\quad \downarrow \quad \right) = \frac{\sum_{i=1}^{n} x_i E(Y_i) - n\bar{x}E(\bar{Y})}{S_{xx}}$$

(xs fixed)

$$= \frac{\sum_{i=1}^{n} x_i(\beta_0 + \beta_1 x_i) - n\bar{x}(\beta_0 + \beta_1 \bar{x})}{S_{xx}}$$

$$= \frac{\beta_0(\sum x_i - n\bar{x}) + \beta_1 \overbrace{(\sum x_i^2 - n\bar{x}^2)}^{S_{xx}}}{S_{xx}}$$

($\sum x_i - n\bar{x} \to 0$)

$$= \beta_1$$

i.e. $\hat{\beta}_1$ is an unbiased estimator for β_1

$$V(\hat{\beta}_1) = V\left(\frac{S_{xy}}{S_{xx}}\right) = \ldots = \frac{\sigma^2}{S_{xx}}.$$

$E(\hat{\beta}_0)$, $V(\hat{\beta}_0)$:

$$E(\hat{\beta}_0) = E(\bar{Y}) - \bar{x}E(\hat{\beta}_1) = E(\bar{Y}) - \beta_1\bar{x}$$

($E(\hat{\beta}_1) = \beta_1$ from above.)

$$= \beta_0 + \beta_1\bar{x} - \beta_1\bar{x}$$

$$= \beta_0$$

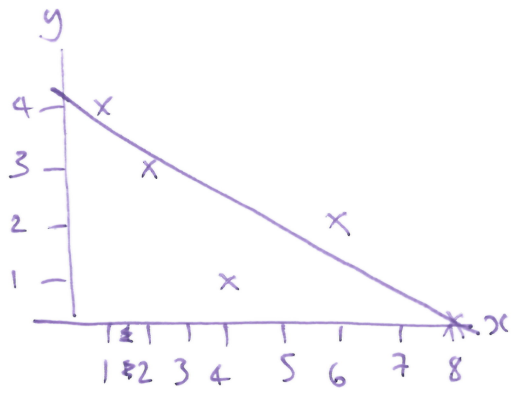
So $\hat{\beta}_0$ is an unbiased estimator for β_0.

& $V(\hat{\beta}_0) = \ldots = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$.

Since we don't know σ^2, we estimate with $\hat{\sigma}^2 = \frac{1}{n-2} SS_E$ & so we get the estimated standard error of $\hat{\beta}_1$ & $\hat{\beta}_0$:

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}, \quad se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}.$$

Least Squares Regression Example



$n=5$

i	x	y	x^2	y^2	xy
1	1	4	1	16	4
2	2	3	4	9	6
3	4	1	16	1	4
4	6	2	36	4	12
5	8	0	64	0	0
SUM	21	10	121	30	26

$\bar{X} = \frac{21}{5} = 4.2 \quad \bar{y} = \frac{10}{5} = 2$

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

$s_{xx} = \Sigma x^2 - n\bar{x}^2$
$= 121 - 5(4.2)^2 = 32.8$

$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \frac{-16}{32.8} = -0.49$ slope

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

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

same

$s_{yy} = \Sigma y^2 - n\bar{y}^2$
$= 30 - 5(2)^2$

$SS_T = s_{yy} = 10$

$\hat{\sigma}^2 = \frac{1}{n-2} SS_E = \frac{1}{5-2}(10 - (-0.49)(-16)) = \frac{1}{3}(2.16)$
$= \frac{1}{3}(7.84) = 0.72$

$SS_R = \hat{\beta}_1 s_{xy} = (-0.49)(-16) = 7.84$

$SS_E = SS_T - SS_R = 10 - 7.84 = 2.16$

$R^2 =$

$R =$

Tests for β_1: $t_0 =$

$f_0 =$

95% CI for β_1:

95% CI for Y at $x_0=3$:

95% PI for Y_0 at $x_0=3$: