3Y03 - Probability and Statistics for Engineering

WS19 Lecture 32

Last Time

Even more on Linear Regression $Y = \beta_0 + \beta_1 X + \varepsilon$

→ Tests
↳ on β_1 slope (t-test for linearity)

→ Confidence Intervals
↳ for β_1 slope and mean of Y at x_0

→ Prediction Intervals
↳ for future value of Y at x_0

All assume $\varepsilon \sim N(0, \sigma^2)$ and use estimate $\hat{\sigma}^2 = \frac{1}{n-2} SS_E$ & t_{n-2} - distribution

where SS_E = "Error Sum of Squares" $\sum e_i^2$ = SS_T "Total Sum of Squares" $= S_{yy} = \sum y_i^2 - n\bar{y}^2$ $- SS_R$ "Regression Sum of Squares" $= \hat{\beta}_1 S_{xy}$

"Variability of Y unexplained by regression" ← due error

"Total variability of Y" SS_T = SS_E + SS_R ← "Amount of variability of Y accounted for by regression line"

$\sum (y_i - \bar{y})^2$ $\sum (y_i - \hat{y}_i)^2$ $\sum (\hat{y}_i - \bar{y})^2$

ANOVA = Analysis of Variance Approach

$$SS_T = SS_E + SS_R$$

Also can show:

$$E\left(\frac{1}{n-2} SS_E\right) = \sigma^2 \quad \text{(from before)} \qquad E(SS_R) = \sigma^2 + S_{xy}$$

So we say SS_E has $n-2$ d.o.f. & SS_R has 1 d.o.f.

d.o.f. must balance so SS_T has $n-1$ d.o.f.

It can be shown that, under $H_0 : \beta_1 = 0$,

($H_1 : \beta_1 \neq 0$)

$$F_0 = \frac{MS_R}{MS_E} := \frac{SS_R / 1}{SS_E / (n-2)}$$

(MS_R: mean squares (regression); MS_E: mean squares (error))

$\sim$ "F-distribution with numerator d.o.f. = 1, & denominator d.o.f. = $n-2$"

written $F_{1, n-2}$

The bigger F_0 is the more of variability is explained by the regression so more evidence for linear rel.

i.e. when testing $H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$ (2-sided)

reject H_0 at sig. level α if $f_0 > f_{\alpha, 1, n-2}$

Look up α f-table, column 1, row $n-2$

Example – see sheet

Notice : $T_0 = \dfrac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$ so $T_0^2 = \dfrac{\hat{\beta}_1^2 S_{xx}}{\hat{\sigma}^2} = \dfrac{\hat{\beta}_1 S_{xy}}{SS_E / n-2} = \dfrac{SS_R}{SS_E/n-2} = F_0$.

So for 2-sided test on $H_0: \beta_1 = 0$, t-test & f-test are equivalent (but for 1-sided, need to use t-test).

11.7 Adequacy of the Regression Model

Assumptions made ① there is a linear rel.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where $E(\varepsilon_i) = 0$ $(V(\varepsilon_i) = \sigma^2)$

② $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$

For tests on β_1 (& C.I., P.I.) : ③ $\varepsilon_i \sim N(0, \sigma^2)$

Goal check assumptions

③ ε_i has Normal distr.: make a prob. plot of observed values i.e. $e_1, \ldots, e_n$

Recall: Reorganize $e_{(1)} < \ldots < e_{(n)}$

& plot $e_{(j)}$ against z_j where

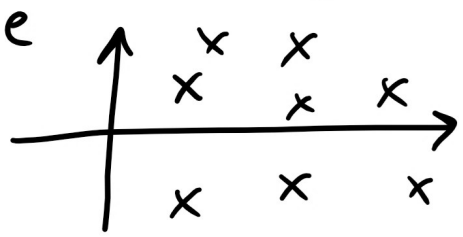
$$\Phi(z_j) = P(Z < z_j) = \frac{j - 0.5}{n}.$$

If correct i.e. ε_i Normal $\rightarrow$ get a straight line.

(2) $Cov(\varepsilon_i, \varepsilon_j) = 0$ $i \neq j$: plot e_i against y_i, or x_i or time

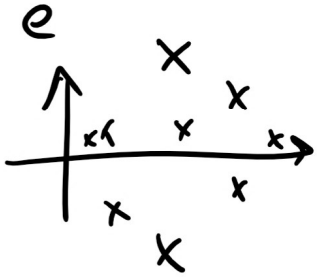
(none of these should influence error)

So should get:

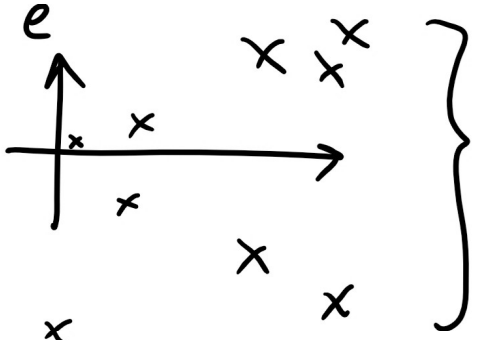


evenly spread either side of horizontal axis.

NOT

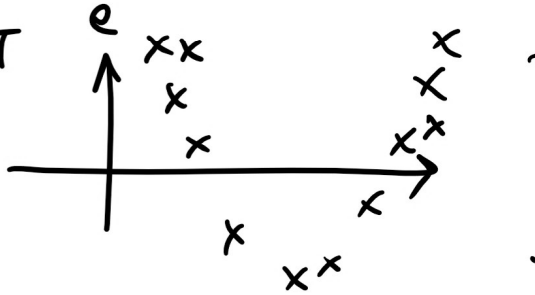


OR



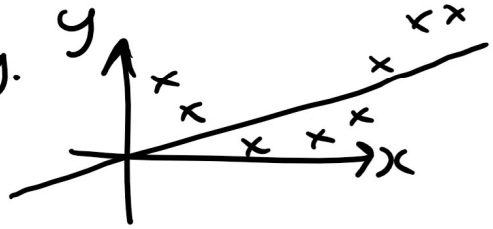
spread not even – error may depend on y_i, or x_i, or time

and NOT



suggests relationship NOT linear

e.g.



(3) Linear rel. between X & Y

(a) t-test on β_1

(b) ANOVA f-test on β_1

(c) ANOVA:

The coefficient of determination $R^2 = \frac{SS_R}{SS_T}$

- proportion of total variation in Y explained by regression $\in [0,1]$

- bigger : more likely to be a linear relationship

Example — see sheet.

11.8 Correlation

↳ assume X & Y jointly distributed

Recall underlying correlation coefficient: $\rho = \dfrac{Cov(X,Y)}{\sigma_X \sigma_Y}$

Under our assumptions it turns out that

$$\beta_1 = \frac{\sigma_Y}{\sigma_X}\rho$$ i.e. NO linear rel. exactly when $\rho = 0$.

Estimator for ρ :

$$R = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

$$= \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Note : $R^2 = \ldots = (\text{coeff. of det.})^2$ ⊛ see below

Example : $R = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = -16 / \sqrt{(32.8)(10)} = -0.88$

This says X & Y are very negatively correlated.

↳ Now added to separate sheet.

$$\circledast \quad (R)^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \frac{S_{xy}}{S_{xx}} \cdot \frac{S_{xy}}{S_{yy}} = \hat{\beta}_1 \cdot SS_T = \frac{\hat{\beta}_1 S_{xy}}{SS_T} = \frac{SS_R}{SS_T} = R^2$$

↑ sample correlation coefficient

↑ coefficient of determination

Observe, though, that we can't get R from R^2 without knowing something more:

Phew! $R^2 = R^2$!

in our running example on the separate sheet, $R^2 = 0.784$.

So we know $R = \sqrt{0.784}$ or $-\sqrt{0.784}$

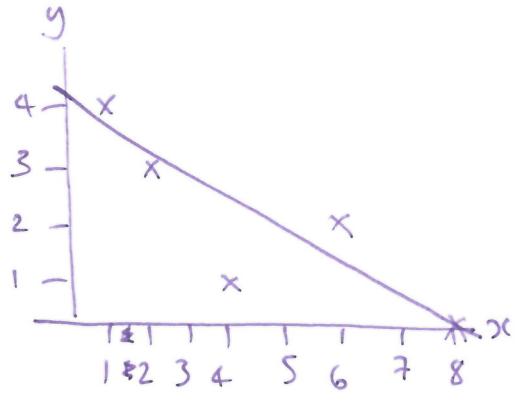
i.e. 0.88 or -0.88

In order to conclude $R = -0.88$ this way we would need to know something more e.g. that the slope of the line $\hat{\beta}_1$ is negative (and here we do know that; $\hat{\beta}_1 = -0.49$).

Least Squares Regression Example

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

$n=5$



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

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

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

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

$SS_T = s_{yy} = 10$

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

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

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

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

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

$R^2 = \frac{SS_R}{SS_T} = \frac{7.84}{10} = 0.784$

$f_0 = \frac{SS_R/1}{SS_E/n-2} = \frac{7.84}{2.16/3} = 10.8$

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

$R = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \frac{-16}{\sqrt{(32.8)10}} = -0.88$

$f_{0.05,1,3} = 10.13$ so reject H_0

Tests for β_1:

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$\alpha = 0.05$

$t_0 = f_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2/s_{xx}}} = \frac{-0.49 - 0}{\sqrt{0.72/32.8}} = \frac{-0.49}{0.14} = -3.31$

$t_{\frac{\alpha}{2},n-2} = t_{0.025,3} = -3.182$

more extreme so reject H_0 - yes, a linear rel.

95% CI for β_1: $\hat{\beta}_1 \pm t_{0.025,3}\sqrt{\frac{\hat{\sigma}^2}{s_{xx}}} = -0.49 \pm 3.182(0.14) = -0.49 \pm 0.47 = (-0.96, -0.02)$

→ Notice 0 not in here: another way to reject $H_0: \beta_1 = 0$ for $H_1: \beta_1 \neq 0$

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

$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{0.025,3}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}\right)} = (4.05 - 0.49(3)) \pm 3.182\sqrt{0.72\left(\frac{1}{5} + \frac{(3-4.2)^2}{32.8}\right)} = (1.25, 3.91)$

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

" " " $\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}\right)}$

$= (4.05 - 0.49(3)) \pm 3.182\sqrt{0.72\left(1 + \frac{1}{5} + \frac{(3-4.2)^2}{32.8}\right)} = (-0.43, 5.59)$