Review:

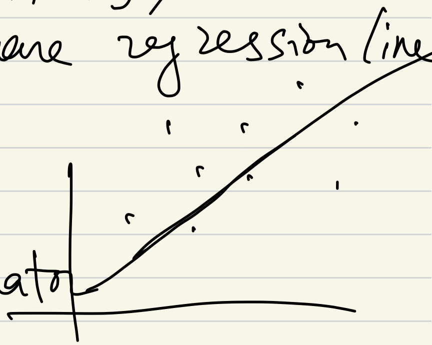
Notation: $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \overset{①}{=} \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \overset{②}{=} \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \overset{③}{=} \sum y_i^2 - n\bar{y}^2$$

$(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$

$\Rightarrow$ fit ~~a~~ simple least square regression line



$\hat{y} = b_0 + b_1 x$

b_0 and b_1 are the estimator of β_0 & β_1 where $y = \beta_0 + \beta_1 x + \varepsilon$ is the true regression model

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

why ② is true?

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i y_i - x_i \bar{y} - y_i \bar{x} + \bar{x}\bar{y})$$

$$= \sum x_i y_i - \sum x_i \bar{y} - \sum y_i \bar{x} + \sum \bar{x}\bar{y}$$

$$- \bar{y} \cdot n\bar{x} - \cancel{\bar{x} n \bar{y}} + \cancel{n\bar{x}\bar{y}}$$

For any inference on slope β_1:

$$T = \frac{b_1 - \beta_1}{s/\sqrt{S_{xx}}} \sim t(n-2)$$

where $s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{S_{yy} - b_1 S_{xy}}{n-2}}$

mean square error

sum of squared error

① ex: Find a 95% CI for regression slope β_1

$$b_1 \pm t_{\frac{\alpha}{2}}^{(n-2)} \cdot \frac{s}{\sqrt{S_{xx}}}$$

② H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

under H_0: $\frac{b_1}{s/\sqrt{S_{xx}}} \sim t(n-2)$

Inference on ~~the~~ mean response $\mu_{Y|x_0}$ and on a single response $Y_0|x_0$.

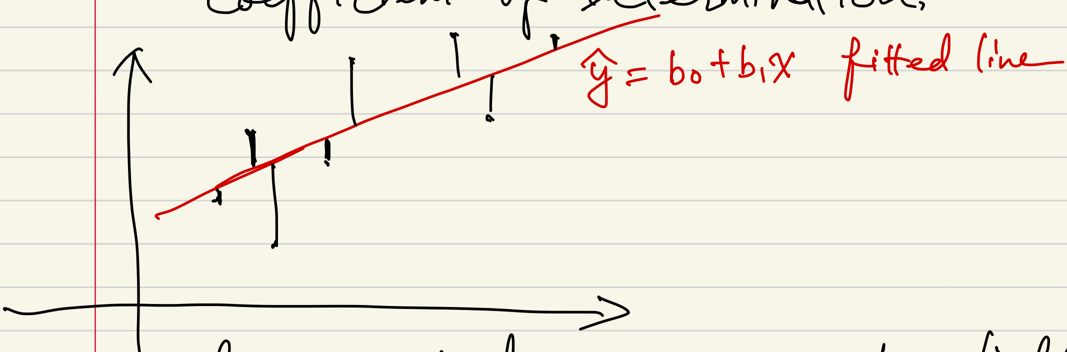
$$T = \frac{\hat{Y}_0 - \mu_{Y|x_0}}{S \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim T(n-2)$$

and:

$$T = \frac{\hat{Y}_0 - Y_0}{S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim T(n-2)$$

A measure of quality of fit:

Coefficient of Determination.



Before we introduce x as an input variable, we see a lot variation in y.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \; (= S_{yy})$$

(total sum of squares)

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

sum of square error

error after introducing the regression model.

$$R^2 = \text{Coefficient of Determination} = 1 - \frac{SSE}{SST}$$

total error.

$$R^2 = r^2$$

r: correlation coefficient of x_i & y_i

$-1 \leq r \leq 1.$

{0.1}.

Ex: Survey 200 Dem. 150 Rep. 150 Indep

Law	Dem	Rep	Indep	
For	85 ??	70	62	214
Against	92	62	67	222
undesided	25	18	21	64
	200	150	150	500

Ho: For each opinion (for, against, or decided) the 3 group of people have the same proportion.

H1.. Not

under H_0: $P_{D,for} = P_{rep,for} = P_{Indep,for}$

against

undecide ...

$\hat{P}_{for} = \frac{214}{500}$ $\hat{P}_{against} = \frac{222}{500}$ $\hat{P}_{undecided} = \frac{64}{500}$

$e_{Dem.for} = 200 * \frac{214}{500} =$

$$\chi^2 = \sum_{i=1}^{n} \frac{(e_i - O_i)^2}{e_i} \sim \chi^2(n-1)$$

Goodness-of-fit

$e_i \geq 5$