Chapter 11. Simple Linear Regression and Correlation.

Y: output v., response v., dependent v. . . – .

$X_1, X_2, \cdots X_n$, input v, explanatory var, indep. v. . .

$$\underbrace{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n}_{\text{deterministic relation.}}$$

Linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

$\beta_0, \beta_1, \cdots \beta_n$ — parameters, constants,

$x_1, x_2, \cdots x_n$ — inputs

ε: random error. $\sim N(0, \sigma^2)$

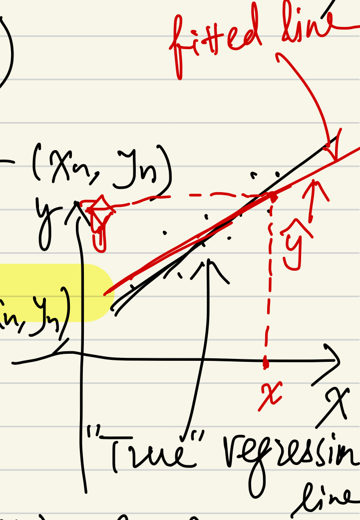
$$Y \sim N(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n, \sigma^2)$$

Simple linear regression model:

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

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

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



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

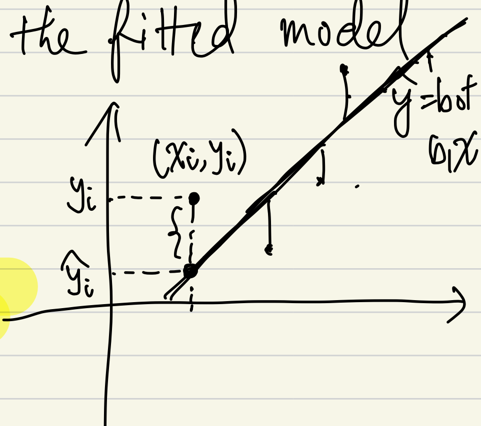
fit a line $\boxed{\hat{y} = b_0 + b_1 x}$

to est. $\boxed{E(Y) = \beta_0 + \beta_1 x}$

$E(Y) = \beta_0 + \beta_1 x$

$\hat{\beta}_0 = b_0 \quad \hat{\beta}_1 = b_1$

§3. Least squares and the fitted model



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

$$S.S.E. = \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i)\right)^2$$

sum of squared errors

Find b_0 & b_1 to minimize S.S.E.

$$\frac{d(S.S.E)}{d\,b_0} = \sum_{i=1}^{n} 2\left(y_i - (b_0 + b_1 x_i)\right)(-1) = 0$$

$$\frac{d(SSE)}{d\,b_1} = \sum_{i=1}^{n} \cancel{2}\left(y_i - (b_0 + b_1 x_i)\right)(\cancel{-}x_i) = 0$$

$$\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0$$

$$\cancel{n}\bar{y} - \cancel{n} b_0 - b_1 \cancel{n}\bar{x} = 0 \qquad b_0 = \bar{y} - b_1\bar{x}$$

$$\sum\left(x_i y_i - b_0 x_i - b_1 x_i^2\right) = 0$$

$$\sum\left(x_i y_i - (\bar{y} - b_1\bar{x})x_i - b_1 x_i^2\right) = 0$$

$$\sum\left(x_i y_i - \bar{y}x_i + b_1\bar{x}x_i - b_1 x_i^2\right) = 0$$

$$\sum x_i y_i - n\bar{x}\bar{y} + n b_1 \bar{x}^2 - b_1 \sum x_i^2 = 0$$

$$\sum b_1\bar{x}\,x_i = b_1\bar{x}\sum x_i = n b_1 \bar{x}^2 \qquad \sum \bar{y}\,x_i = \bar{y}\sum x_i = n\bar{x}\bar{y}$$

$$b_1\left(\Sigma x_i^2 - n\bar{x}^2\right) = \Sigma x_i y_i - n\bar{x}\bar{y}$$

$$b_1 = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$

statistic

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

x	123	55	100	75	159	109	48	138	164	28
y	76	62	66	58	88	70	37	82	88	43

$n=10$, $\bar{x}=99.9$, $\bar{y}=67$

$\Sigma x^2 = 119969$ $\quad \Sigma y^2 = 47670$

$\Sigma xy = 74058$

b_0

b_1

$$b_1 = \frac{74058 - 10 * 99.9 * 67}{119969 - 10 * 99.9^2}$$

$= 0.3533$

$b_0 = 67 - 0.3533 * 99.9$

$= 31.71$

$$\hat{y} = 31.71 + 0.3533x$$

If $x = 110$, what's the est. mean response y? $\hat{y} = 31.71 + 110 * .353$

b_1 (b_0) as an estimator for β_1 (β_0), what is its sampling dist?

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

$$S_{yy} = \sum (y_i - \bar{y})^2$$

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

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

$$= \underbrace{\sum x_i y_i - n\bar{x}\bar{y}}_{} - \cancel{n\bar{x}\bar{y}} + \cancel{n\bar{x}\bar{y}}$$

$$S_{xx} = \sum x_i^2 - n\bar{x}^2$$

$$\sum x_i \bar{y} = \bar{y} \sum x_i = \bar{y} n\bar{x}$$

§11.5

To determine the dist of b_1:

$$b_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{S_{xx}} = \frac{\sum (x_i - \bar{x}) y_i}{S_{xx}} \quad ✓$$

$$\text{RHS} = \sum (x_i y_i - \bar{x} y_i)$$
$$= \sum x_i y_i - \bar{x} \sum y_i$$
$$= \sum x_i y_i - n\bar{x}\bar{y} = \text{LHS}$$

$$\boxed{b_1 = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} y_i \qquad \text{Recall: } y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)}$$

b_1 is a linear combination of n indep normal r.v's $y_1, y_2, \ldots, y_n$.

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

$$E(b_1) = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} (\beta_0 + \beta_1 x_i)$$

$$= \sum_{i=1}^{n} \frac{(X_i - \bar{X})}{S_{XX}} \cdot \beta_0 + \sum_{i=1}^{n} \frac{X_i - \bar{X}}{S_{XX}} \cdot \beta_1 X_i$$

$$= \frac{\beta_0}{S_{XX}} \cancel{\sum_{i=1}^{n} (X_i - \bar{X})} + \frac{\beta_1}{S_{XX}} \sum_{i=1}^{n} (X_i - \bar{X}) \cdot X_i$$

0

$$= 0 + \frac{\beta_1}{\cancel{S_{XX}}} \left[\sum_{i=1}^{n} X_i^2 - \bar{X} \sum X_i \right]$$

$$= \sum X_i^2 - n\bar{X}^2$$

$$= \cancel{S_{XX}}$$

$$\sum (X_i - \bar{X}) = \sum X_i - n\bar{X} = 0$$

$$= \beta_1$$

$$Var(b_1) = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S_{XX}} \right)^2 \sigma^2$$

$$= \frac{\sigma^2}{S_{XX}^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{\sigma^2}{S_{XX}}$$

continued: $b_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$

$$\frac{b_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim Z.$$

$\hat{y}_i = b_0 + b_1 x_i$ estimates for β_0 & β_1

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

$\sigma = ?$ $$S = \sqrt{\frac{SSE}{n-2}}$$

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

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

$100(1-\alpha)\%$ CI for β_1:

$$\left(b_1 - t_{\frac{\alpha}{2}}^{(n-2)} \cdot \frac{s}{\sqrt{S_{xx}}},\ b_1 + t_{\frac{\alpha}{2}}^{(n-2)} \frac{s}{\sqrt{S_{xx}}}\right)$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = S_{yy} - b_1 S_{xy}$$

x	123	55	100	75	159	109	48	138	164	28
y	76	62	66	58	88	70	37	82	88	43

$n=10$, $\bar{x} = 99.9$, $\bar{y} = 67$

$\Sigma x^2 = 119969$ $\quad \Sigma y^2 = 47670$

$\Sigma xy = 74058$

b_0

b_1

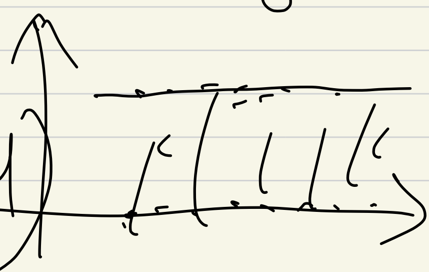
$$b_1 = \frac{74058 - 10 * 99.9 * 67}{119969 - 10 * 99.9^2} = 0.3533$$

$b_0 = 67 - 0.3533 * 99.9$
$= 31.71$

$$\hat{y} = 31.71 + 0.3533x$$

① Find a 95% C.I. on the regression slope.

(If $\beta_1 = 0$, means no useful regression model)



② Test $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$ at $\alpha = 0.05$.

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

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

$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$

①
$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{S_{yy} - b_1 S_{xy}}{n-2}}$$

$$= \sqrt{\frac{(47670 - 10 * 67^2) - 0.3533(74058 - 10 * 99.9 * 67)}{8}}$$

$$= \sqrt{\frac{2780 - 0.3533 * 7125}{8}} = 5.735$$

$$\sqrt{S_{xx}} = \sqrt{\sum X_i^2 - n\bar{x}^2} = \sqrt{119969 - 10 * 99.9^2}$$
$$\approx 142.$$

$$b_1 \pm t_{0.025}(8) \cdot \frac{s}{\sqrt{S_{xx}}} =$$

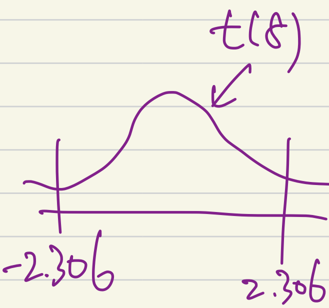
$$= 0.3533 \pm 2.306 * \frac{5.735}{142}$$

$$= 0.3533 \pm 0.093$$

$$= (0.2603, \ 0.4463)$$

② under H_0, $\frac{b_1 - \cancel{\beta_1}\ 0}{s/\sqrt{S_{xx}}} \sim t(n-2)$

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



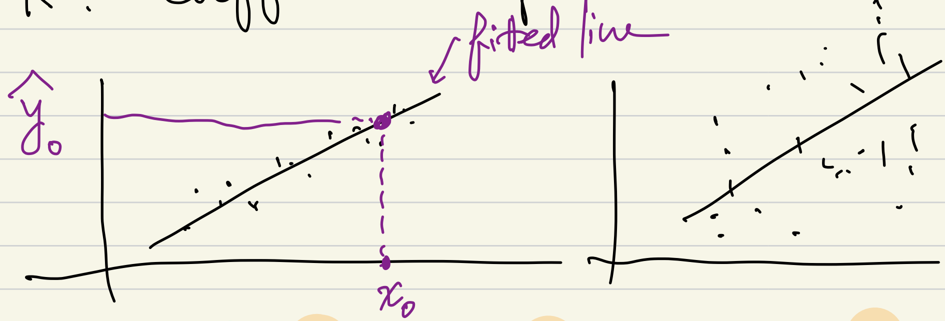
$$C = \left\{ \frac{b_1}{s/\sqrt{S_{xx}}} > 2.306 \text{ or } \frac{b_1}{s/\sqrt{S_{xx}}} < -2.306 \right\}$$

$$t_{obs} = \frac{0.3533}{5.735/142} = 8.75 \in C$$

Reject H_0 and conclude $\beta_1 \neq 0$.

A measure of quality of fit: R^2

R^2: coefficient of determination



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

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

sum of squares, total

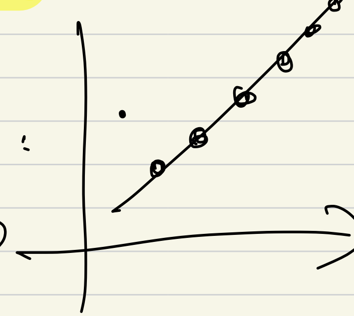
$\hat{y}_i = b_0 + b_1 x_i$

fitted line

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

$$R^2 = 1 - \frac{SSE}{SST}$$

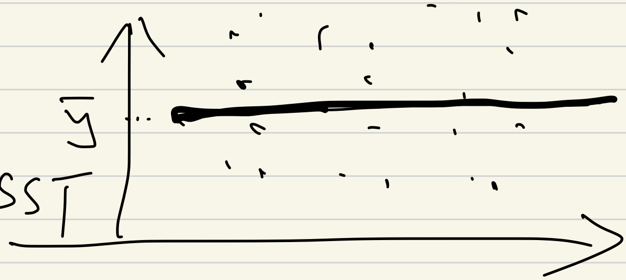
Extreme case:



$SSE = 0$

$R^2 = 1$

other extreme case:



$SSE = SST$

$R^2 = 0$

$$0 \le R^2 \le 1$$

percentage of the variations in y_i that's explaine by introducing x_i. (explained by the fitted regression model).

continue with previous example:

③ : $R^2 = ?$ $\quad 1 - \frac{SSE}{SST} = 1 - \frac{SSE}{S_{yy}}$

$SSE = S_{yy} - b_1 S_{xy} = 2780 - 0.3533 * 7125$
$= 262.7$

$S_{yy} = 2780$

$$R^2 = 1 - \frac{262.07}{2780} = 0.905$$

§ 11.6. Prediction

$$\boxed{\begin{aligned} R^2 &= 1 - \frac{S_{yy} - b_1 S_{xy}}{S_{yy}} \\ &= b_1 \cdot \frac{S_{xy}}{S_{yy}} \end{aligned}}$$

$$\begin{cases} \text{mean response} \\ \text{individual response} \end{cases}$$

a) $100(1-\alpha)\%$ confidence interval for mean response $\mu_{Y|x_0}$:

$$\left(\hat{y}_0 - t_{\frac{\alpha}{2}}^{(n-2)} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},\ \hat{y}_0 + t_{\frac{\alpha}{2}}^{(n-2)} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$$

b) $100(1-\alpha)\%$ prediction interval for a single response y_0 :

$$\left(\hat{y}_0 - t_{\frac{\alpha}{2}}^{(n-2)} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},\ \hat{y}_0 + t_{\frac{\alpha}{2}}^{(n-2)} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$$

④ 95% C.I. for mean response when $x=50$.

$$\hat{y}_0 = 31.71 + 0.3533\, \underset{\underset{50}{\parallel}}{x_0} = 49.38$$

$$\hat{y}_0 \pm t_{\frac{\alpha}{2}}^{(n-2)} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}$$

$$= 49.38 \pm 2.306 * 5.735 \cdot \sqrt{\frac{1}{10} + \frac{(50-99.9)^2}{20169}}$$

$$= 49.38 \pm 6.25$$

$$= (43.13,\ 55.63)$$

⑤ 95% prediction interval for an individual response when $x_0 = 50$

. . . . $(34.77, 63.99)$