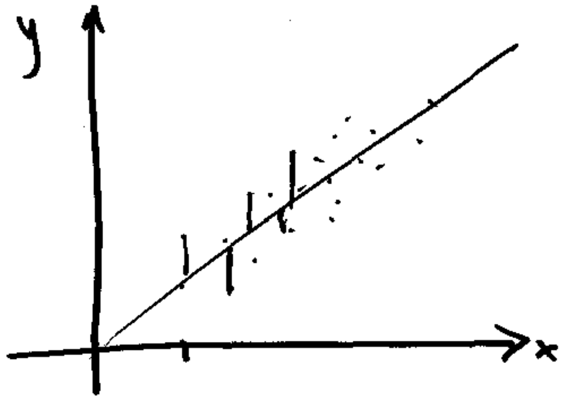
Stat 451
3-6-18
1



Find the "best" line

$i = 1, \dots, n \qquad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

We will find the line that minimizes the sum of squares of the vertical errors. Least Squares Method

2

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

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

$$\frac{\partial SSE}{\partial \beta_0} = \sum_{i=1}^{n} \cancel{2}(y_i - \beta_0 - \beta_1 x_i)(-\cancel{1}) \stackrel{set}{=} 0 \quad (1)$$

$$\frac{\partial SSE}{\partial \beta_1} = \sum_{i=1}^{n} \cancel{2}(y_i - \beta_0 - \beta_1 x_i)(\cancel{-}x_i) \stackrel{set}{=} 0 \quad (2)$$

③

$$(1)\quad \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0$$

$$\bar{y} - \beta_0 - \beta_1 \bar{x} = 0$$

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

$$(2)\quad \sum_{i=1}^{n} y_i x_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} y_i x_i - (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

④

$$\sum_{i=1}^{n} y_i x_i - \bar{y} \sum_{i=1}^{n} x_i + \beta_1 \bar{x} \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\underbrace{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}_{SS_{XY}} = \beta_1 \Bigg(\underbrace{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}_{SS_{XX}} \Bigg)$$

$$\boxed{\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

slope intercept

5

Now, your prediction line is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Note: This line goes through the center of mass $(\bar{x}, \bar{y})$

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 \bar{x} = \bar{y} \checkmark$$

6

Let's redo this task, using Maximum Likelihood.

Assume that $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent, identically distributed $N(0, \sigma^2)$

then $y_i = N(\beta_0 + \beta_1 x_i, \sigma^2)$

and the y_i's are independent

To begin the MLE process, write the joint prob density for $y_1, \ldots, y_n$

7

$$\prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i-\beta_0-\beta_1 x_i}{\sigma}\right)^2}$$

$$= \sigma^{-n}(2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}$$

= the likelihood function

Take ln of this

$$l(\beta_0,\beta_1) = -n\ln\sigma - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2$$

$$\frac{\partial l}{\partial \beta_0} = -\frac{1}{\cancel{2}\sigma^2}\sum_{i=1}^{n}\cancel{2}(y_i-\beta_0-\beta_1 x_i)(\cancel{-1}) \overset{set}{=} 0 \qquad (1)$$

8

$$\frac{\partial l}{\partial \beta_1} = -\frac{1}{\cancel{2}\sigma^2}\sum_{i=1}^{n}\cancel{2}(y_i-\beta_0-\beta_1 x_i)(\cancel{-}x_i) \overset{set}{=} 0 \qquad (2)$$

These are exactly the same 2 equations from the Least Squares Method.

So the MLE's of β_0 & β_1 are the same as the L.S. estimators.

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

9

What if the y-intercept has to be 0?

$$y_i = \beta_1 x_i + \varepsilon_i$$

$$\varepsilon_i = y_i - \beta_1 x_i$$

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

$$\frac{dSSE}{d\beta_1} = \sum_{i=1}^{n} \cancel{2}(y_i - \beta_1 x_i)(\cancel{-}x_i) \overset{set}{=} 0$$

10

$$\sum_{i=1}^{n} y_i x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

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

2-sample confidence intervals

(11)

Population 1 has mean μ_1
" 2 " " μ_2

We want to estimate $\mu_1 - \mu_2$ + create a confidence interval

Know: $\bar{X}_1$ can be used to estimate μ_1
For large n_1, $\bar{X}_1 \sim N(\mu_1, \sigma_1^2/n_1)$

(12)

Also, $\bar{X}_2$ can be used to estimate μ_2
and for large n_2, $\bar{X}_2 \sim N(\mu_2, \sigma_2^2/n_2)$

So $\bar{X}_1 - \bar{X}_2$ can be used to estimate $\mu_1 - \mu_2$

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

provided that $\bar{X}_1$ & $\bar{X}_2$ are <u>independent</u>

(13)

$$\therefore \quad \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$$P\left(-1.96 < \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < 1.96\right) = .95$$

$$\bar{X}_1 - \bar{X}_2 - 1.96\sqrt{\cdots} < \mu_1 - \mu_2 < \bar{X}_1 - \bar{X}_2 + 1.96\sqrt{\cdots}$$

(14)

C.I. for $\mu_1 - \mu_2$: $\underbrace{\bar{X}_1 - \bar{X}_2}_{\text{estimat}} \pm \underbrace{Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}_{\text{margin of error}}$