Lecture 4

Monday, January 24, 2022 10:53 AM

Admin

① pages → BlueJeans.

- Extend single mean:

$z = X_2 - X_1$

indep.

σ_{M_z}

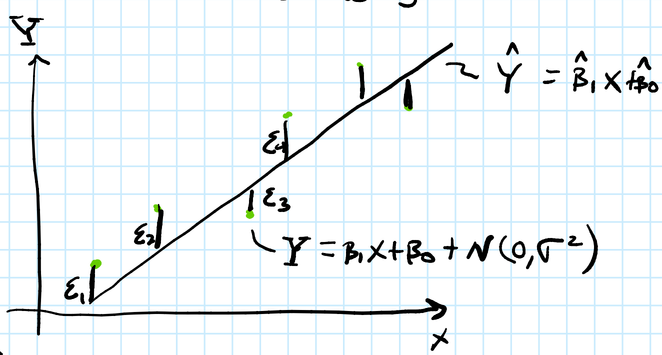
- Ordinary Least Squares Regression

① specify a least squares linear fit

② Discuss parameter variance / hypothesis testing.

Fitting a line to data

- In many experiments, we want to determine if there is a "significant" relationship between two variables



→ age & memory in a mouse

→ position & voltage of a sensor

- Frequently want to ask if it is worth continuing to collect data

→ we need a formal test to tell us if observed relationship is real.

Step 1: Find the best fit

Step 2: Ask if "significant"?
⇒ Hypothesis testing.

$$SSE = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2$$

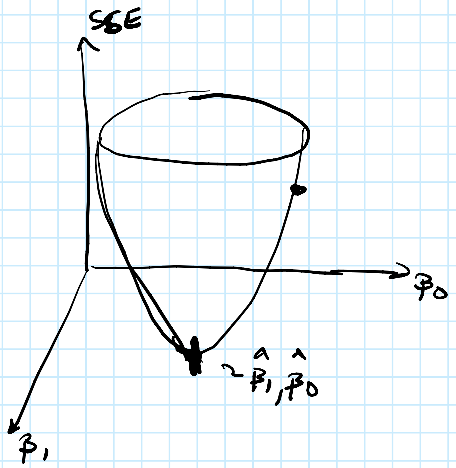
sum of

$SSE = \sum_{i=1} \quad = \sum_{i=1} (\quad)$

sum of squared errors.

$$\hat{\beta}_0, \hat{\beta}_1 = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$$

$\hat{Y}_i = \hat{\beta}_1 X + \hat{\beta}_0$



Define : $\mu_X = \frac{1}{N} \sum_{i=1}^{N} X_i$; $\mu_Y = \frac{1}{N} \sum_{i=1}^{N} Y_i$

$$\Rightarrow SSE = \sum_{i=1}^{N} \Big(Y_i \underbrace{- \beta_1 X_i - \beta_0}_{\text{model}} \Big)^2$$

$$= \sum_{i=1}^{N} \left[(Y_i + \mu_Y - \mu_Y) - \beta_0 - \beta_1 (X_i + \mu_X - \mu_X) \right]^2$$

$$= \sum_{i=1}^{N} \left[(\mu_Y - \beta_0 - \beta_1 \mu_X) + (Y_i - \mu_Y - \beta_1 X_i + \beta_1 \mu_X) \right]^2$$

$$= \sum_{i=1}^{N} \left[(\mu_Y - \beta_0 - \beta_1 \mu_X)^2 \right] + \sum_{i=1}^{N} \left[(\beta_1 X_i - \beta_1 \mu_X - Y_i + \mu_Y) \right]^2$$

$$\underbrace{+ 2 \sum_{i=1}^{N} (\mu_Y - \beta_0 - \beta_1 \mu_X) \big(\overbrace{\beta_1 X_i - \beta_1 \mu_X}^{=0} \overbrace{- Y_i + \mu_Y}^{=0} \big)}_{=0}$$

$$= N (\mu_Y - \beta_0 - \beta_1 \mu_X)^2$$
$$+ \beta_1^2 \sum_{i=1}^{N} (X_i - \mu_X)^2 - 2\beta_1 \sum_{i=1}^{N} (X_i - \mu_X)(Y_i - \mu_Y)$$
$$+ \sum_{i=1}^{N} (Y - \mu_Y)^2$$

$$+ \sum_{i=1} (Y - \mu_Y)$$

$$= (\text{Algebra}) - \text{Skipp } 4\text{-}1$$

$$SSE = \underbrace{N(\mu_Y - \beta_0 - \beta_1 \mu_X)^2}_{1}$$

$$+ \underbrace{\left[\beta_1 - \frac{\sum (x_i - \mu_X)(Y_i - \mu_Y)}{\sum (x_i - \mu_X)^2}\right]^2 \sum_{i=1}^{N} (x_i - \mu_X)^2}_{2}$$

$$+ \underbrace{\sum (Y_i - \mu_Y)^2 \left\{ 1 - \frac{\left[\sum (x_i - \mu_X)(Y_i - \mu_Y)\right]^2}{\sum (x_i - \mu_X)^2 \sum (Y_i - \mu_Y)^2} \right\}}_{3}$$

- our goal is to minimize SSE via β_1, β_0

$$1 \Rightarrow 0 = (\mu_Y - \hat{\beta}_0 - \hat{\beta}_1 \mu_X)$$

$$2 \Rightarrow 0 = \hat{\beta}_1 - \frac{\sum (x_i - \mu_X)(Y_i - \mu_Y)}{\sum (x_i - \mu_X)^2}$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum (x_i - \mu_X)(Y_i - \mu_Y)}{\sum (x_i - \mu_X)^2}$$

$$\hat{\beta}_0 = \mu_Y - \hat{\beta}_1 \mu_X$$

$\Rightarrow$ $\hat{\beta}_0, \hat{\beta}_1$ are our estimates of the parameters β_0, β_1.

$\rightarrow$ we can <u>estimate</u> parameters of trendline.

<u>Determining significance of a parameterization</u>

- How do we know if the relationship between x & Y is significant?

- Hypothesis test:

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$\rightarrow$ we need to determine the variance of $\hat{\beta}_1$

→ we need to determine the variance of $\hat{\beta}_1$

$Y_i = \beta_0 + \beta_1 x_i + N(0, \sigma^2)$

→ start w/ mean:

in every term

$$E[\hat{\beta}_1] = E\left[\frac{\sum (x_i - \mu_x)(Y_i - \mu_Y)}{\sum (x_i - \mu_x)^2}\right] = \frac{\sum [(x_i - \mu_x)(E(Y_i - \mu_Y))]}{\sum (x_i - \mu_x)^2}$$

$$= \underbrace{\frac{\sum (x_i - \mu_x)}{\sum (x_i - \mu_x)^2}}\, E[Y_i - \mu_Y]$$

$[\beta_0 + \beta_1 x_i + N(0, \sigma^2) - \mu_Y]$

$\beta_0 + \beta_1 \mu_X$

$$= \{ \quad \} \underbrace{E[\cancel{\beta_0} + \beta_1 x_i - \cancel{\beta_0} - \beta_1 \mu_X]}_{E[\beta_1 (x_i - \mu_x)]}$$

$$= \frac{\sum (x_i - \mu_x) E[\beta_1 (x_i - \mu_x)]}{\sum (x_i - \mu_x)^2}$$

$$= \frac{\sum \beta_1 (x_i - \mu_x)^2}{\sum (x_i - \mu_x)^2}$$

$$= \beta_1 \frac{\sum (x_i - \mu_x)^2}{\sum (x_i - \mu_x)^2}$$

$\Rightarrow E[\hat{\beta}_1] = \beta_1$

\# random part

$$\mathrm{var}[\hat{\beta}_1] = \mathrm{Var}\left[\frac{\sum (x_i - \mu_x)(Y_i - \mu_Y)}{\sum (x_i - \mu_x)^2}\right]$$

\#

$\mathrm{Var}(aX) = a^2 \mathrm{Var}(X)$

in every term

$$= \frac{\sum [(x_i - \mu_x)^2 \mathrm{Var}(Y_i - \mu_Y)]}{\left[\sum (x_i - \mu_x)^2\right]^2} = \frac{\sum (x_i - \mu_x)^2}{\left[\sum (x_i - \mu_x)^2\right]^2} \mathrm{Var}(\cancel{Y_i - \mu_Y})$$

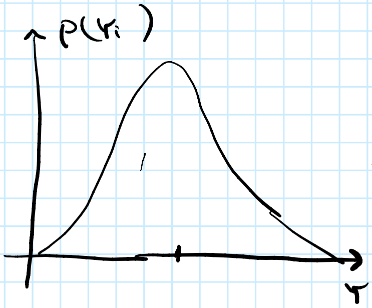
σ_Y^2

$= \sigma_Y^2$

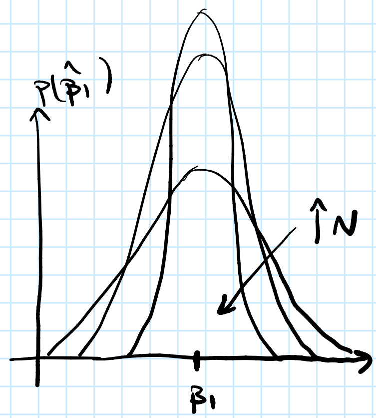
$$\left[\sum (x_i-\mu_x)^2\right] \qquad \left[\sum (x_i-\mu_x)^2\right] \sigma_Y^2$$

$$= \frac{\sigma_Y^2}{\sum (x_i-\mu_x)^2}$$

$$\Rightarrow \sigma_{\beta_1} = \frac{\sigma_Y}{\sqrt{\sum_{i=1}^{N} (x_i-\mu_x)^2}}$$

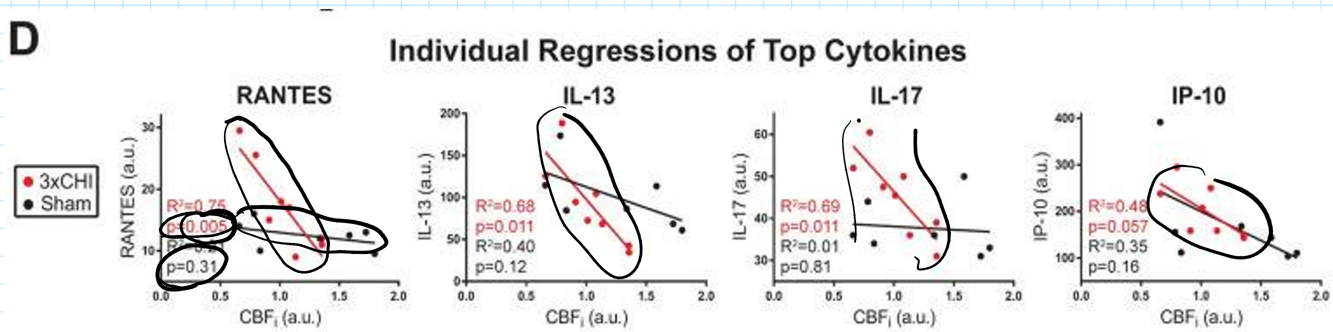


$\Rightarrow$

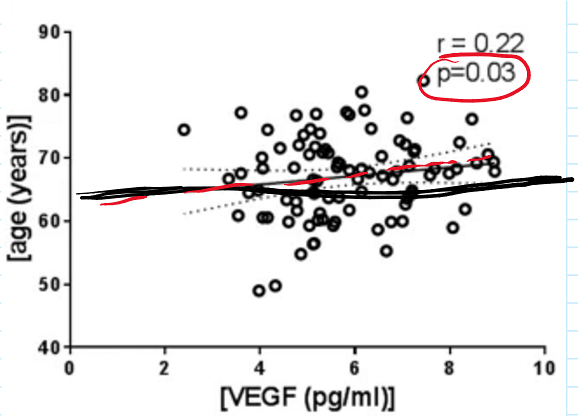


- See supp lecture 4-2.
- Hypothesis testing now takes on same form as for differences in mean.
- Problem 2: will need to derive expression for σ_{β_0}
  - Some examples.

(sankar, Pybus et al, 2019)



Chakraborty et al, 2018



r = 0.22
p=0.03
[age (years)]
[VEGF (pg/ml)]