Lecture 07 Oct 06

Linear Regression Models

- scatterplot Y vs x
- transformation ($\log$, $\sqrt{\ }$, ...)

Random Data

$\{(x_i, Y_i), i = 1, \dots, n\}$

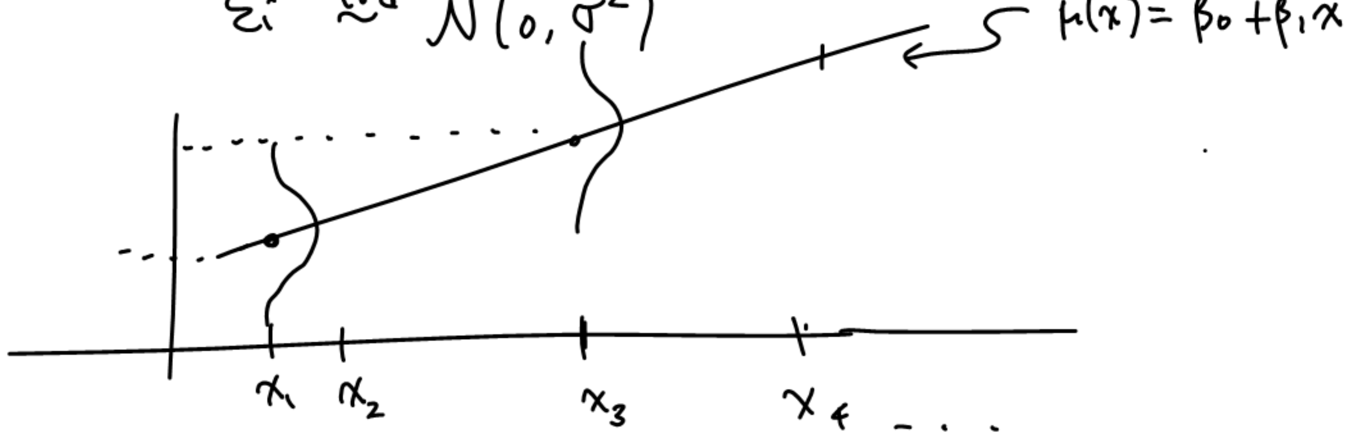
Observed data

$\{(x_i, y_i), i = 1, \dots, n\}$

treat $\{x_i, i = 1, \dots, n\}$ fixed & known.

Model:

$$\left.\begin{aligned} Y_i &= \mu(x_i) + \varepsilon_i, \quad \text{where} \\ \mu(x_i) &= \beta_0 + \beta_1 x_i \\ \varepsilon_i &\overset{iid}{\sim} N(0, \sigma^2) \end{aligned}\right\}$$



- $Y_i | x_i \sim N(\mu(x_i) = \beta_0 + \beta_1 x_i, \ \sigma^2)$
- $Var(Y_i | x_i) = \sigma^2 \quad \forall i$ constant across all values of x

- For a fixed covariate x^*, 99% of all values of the outcome variable Y fall in the interval $[\mu(x^*) - 3\sigma, \mu(x^*) + 3\sigma]$

- Mean function $\mu(x_i)$

$$E(Y_i | x_i) = \mu(x_i) = \beta_0 + \beta_1 x_i \leftarrow$$

Interpretation of β_0

When $x_i = 0 \Rightarrow \mu(x_i) = \beta_0 \leftarrow$

$$\therefore \beta_0 = E(Y | x = 0)$$

Rem: $z_i = x_i - \bar{x}$

$$E(Y_i | z_i) = \beta_0 + \beta_1 z_i$$

$$\beta_0 = E(Y_i | z_i = 0) = E(Y_i | x_i = \bar{x})$$

Interpretation of β_1

$$E(Y_i | x_i) = \mu(x_i) = \beta_0 + \beta_1 x_i$$

$$E(Y_i | x_i = a+1) = \beta_0 + \beta_1 (a+1)$$

$$E(Y_i | x_i = a) = \beta_0 + \beta_1 a$$

The least squares estimates for β_0, β_1:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}$$

$$= \sum w_i Y_i \quad \text{where} \quad w_i = \frac{x_i - \bar{x}}{\sum (x_j - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

The estimate for σ^2 is:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2}{n - 2}$$

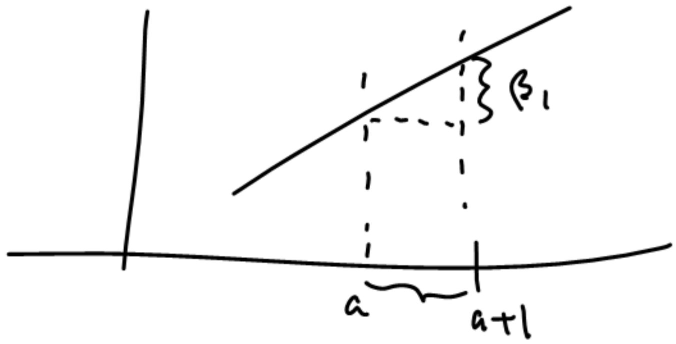
↖ # parameters in $\mu(x_i)$

Inference on β_1

Test $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

$$\Rightarrow \quad E(Y_i \mid x_i = a+1) - E(Y_i \mid x_i = a)$$

$$= \quad \mu(a+1) - \mu(a) \quad = \quad \beta_1$$



β_1 = change in the mean value of the dist'n of the outcome variable for every unit increase in the independent variable.

Construct a procedure for testing $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

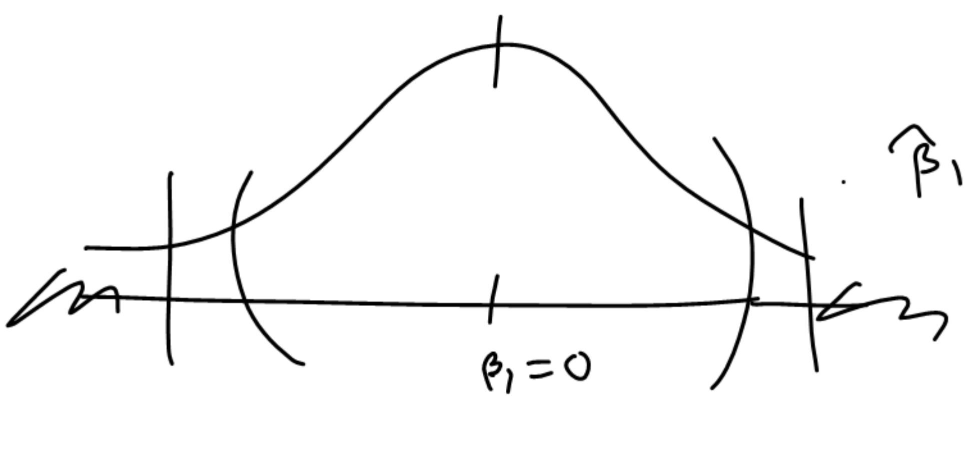
Note: When $\beta_1 = 0 \Rightarrow E(Y_i | x_i) = \beta_0$ Constant over all x

We use $\hat{\beta}_1$:

$$\hat{\beta}_1 = \sum w_i Y_i$$

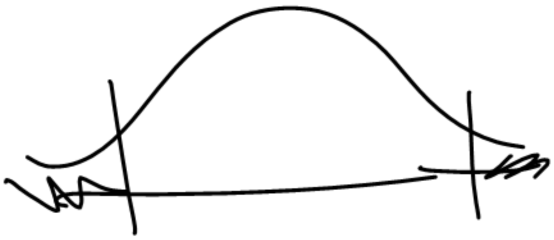
$$\sim N\left(E\hat{\beta}_1 = \beta_1,\ \mathrm{Var}\,\hat{\beta}_1 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right)$$

<u>Population</u>: $\{ (x_i, Y_i),\ i = 1, \dots, \infty \}$



Idea: Reject H_0 if $|\hat{\beta}_1|$ is "large".

We need a reference distn:



$$|\hat{\beta}_1| > 2 \cdot SD(\hat{\beta}_1)$$

$$> 2 \cdot \sqrt{\sigma^2 / \Sigma (x_i - \bar{x})^2}$$

Problem: the threshold $2\sqrt{\sigma^2 / \Sigma (x_i - \bar{x})^2}$ is not known since σ^2 is not known.

Results:

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 / \Sigma (x_i - \bar{x})^2\right)$$

$$\Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / \Sigma (x_i - \bar{x})^2}} \sim N(0, 1) \quad (A)$$

$$\hat{\sigma}^2 = \frac{\Sigma \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2}{n-2} \quad (B)$$

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{\Sigma \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2}{\sigma^2} \sim \chi^2_{(df = n-2)}$$

$$\frac{Y_i - \mu(x_i)}{\sigma} \sim N(0,1)$$

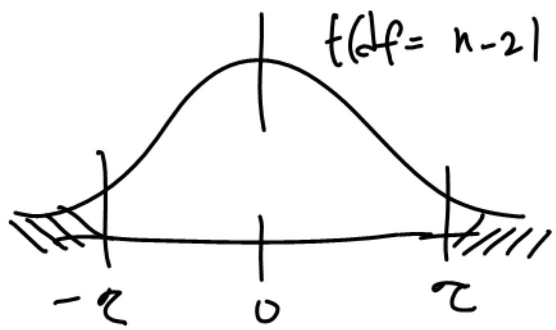
$\hat{\beta}_1$, $\hat{\sigma}^2$ statistically independent (C)

$$\Rightarrow \quad \frac{\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\cancel{\sigma^2}/\sum (x_i - \bar{x})^2}}}{\sqrt{\dfrac{\cancel{(n-2)}\hat{\sigma}^2}{\cancel{\sigma^2}} \Big/ \cancel{(n-2)}}} \sim t(df = n-2)$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2/\sum (x_i - \bar{x})^2}} \sim t(df = n-2)$$

$$T = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/\sum (x_i - \bar{x})^2}} \overset{H_0}{\sim} t(df = n-2)$$

$$T = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/\sum (x_i - \bar{x})^2}}$$



Suppose we set $P(\text{Type I error}) = \alpha = .05$

$$\Rightarrow \quad \tau = t(0.975;\ df = n-2)$$

Decision Rule: Reject H_0 at level α if

$$|T| > \tau = t(.975, df = n-2)$$

$$\iff \quad T > \tau \quad \text{OR} \quad T < -\tau$$

Now we have the observed data:

$$\{(x_i, y_i),\ i = 1, \dots, n\}$$

$$\bar{x},\ \bar{y}$$

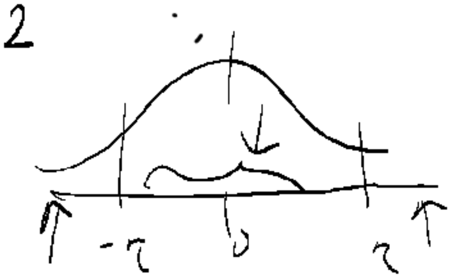
$$\Rightarrow \quad S_{xy} = \Sigma (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xx} = \Sigma (x_i - \bar{x})(x_i - \bar{x})$$

$$\hat{\beta}_{1,obs} = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_{0,obs} = \bar{y} - \hat{\beta}_{1,obs}\,\bar{x}$$

$$\hat{\sigma}^2_{obs} = \frac{\Sigma \left(y_i - (\hat{\beta}_{0,obs} + \hat{\beta}_{1,obs} x_i)\right)^2}{n-2}$$

$$\Rightarrow \quad T_{obs} = \frac{\hat{\beta}_{1,obs}}{\sqrt{\hat{\sigma}^2_{obs} / \Sigma (x_i - \bar{x})^2}}$$



So far:

- Model
- Estimators for the parameters
- Sampling behavior of $\hat{\beta}_1$

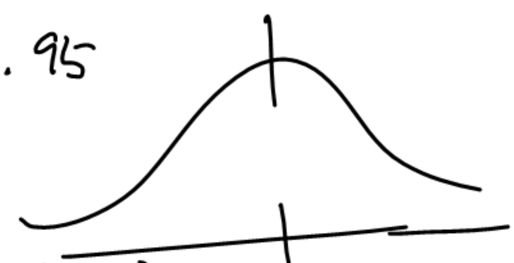
Next: Confidence intervals for β_1, $\mu(x^*)$

Let θ be an unknown parameter.

Let $I = [L, U]$ be a random interval estimator for θ.

The random interval I is a 95% confidence interval estimator for θ if:

$$P(\theta \in I) = 0.95$$



Let $Y_1, \dots, Y_n$ iid $N(\mu, \sigma^2 = 1)$

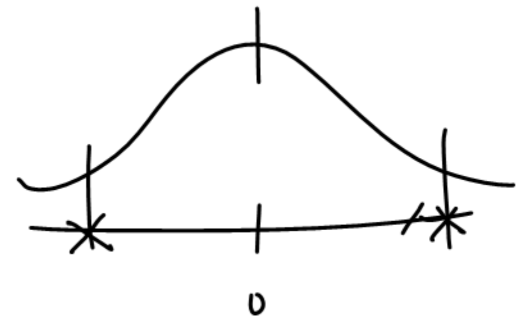
μ unknown

$\bar{Y} = \frac{1}{n} \Sigma Y_i$ is the LSE for μ $\quad \sigma^2 = 1$

MLE for μ

$$\bar{Y} \sim N(\mu, 1/n)$$

$$\Rightarrow \quad \frac{\bar{Y} - \mu}{\sqrt{1/n}} \sim N(0, 1)$$



$$\Rightarrow P\left(-1.96 < \frac{\bar{Y} - \mu}{\sqrt{1/n}} < +1.96\right) = 0.95$$

$$\vdots$$

$$P\left(L(\bar{Y}) < \mu < U(\bar{Y})\right) = 0.95$$

$$\underline{P\left(\mu \in [L(\bar{Y}), \; U(\bar{Y})]\right)} = 0.95$$

$$\Rightarrow P\left(\bar{Y} - 1.96\left(1/\sqrt{n}\right) < \mu < \bar{Y} + 1.96\left(1/\sqrt{n}\right)\right) = 0.95$$

$$\Rightarrow P\left(\mu \in \left[\bar{Y} - 1.96/\sqrt{n}, \; \bar{Y} + 1.96/\sqrt{n}\right]\right) = 0.95$$

A 95% CI <u>estimator</u> for μ is:

$$\left[\bar{Y} - 1.96/\sqrt{n}, \; \bar{Y} + 1.96/\sqrt{n}\right]$$

Observed data $\Rightarrow$ 95% CI <u>estimate</u> $\left[\bar{y} - 1.96/\sqrt{n}, \; \bar{y} + 1.96/\sqrt{n}\right]$.