STAT 231

Nov 18, 2016

a

Tutorial Quiz: 30th

Syllabus — end of today

Roadmap

* 5 min recap of SLRM.
* Least Square Estimates and their properties
* Confidence Intervals / Testing of Hypotheses for α, β, σ.

<u>Set-up</u>:

Y = response variate (trying to "explain" the variability of Y)

X = explanatory variable. (given, not random)

<u>Objective</u>: To try to estimate the relationship between X and Y.

Based on ?

$(x_1, y_1), \ldots (x_n, y_n) \rightarrow$ Sample we have drawn.

We make assumptions on how X and Y are related.

SLRM:

$$Y_i \sim G(\alpha + \beta X, \sigma^2)$$

$i = 1, \ldots n.$

The mean response is a linear function of the explanatory variable.

Standard Gaussian model $\mu(x) = \alpha + \beta x$

$$Y_i \sim G(\mu, \sigma)$$

$i = 1, \ldots,$

Notes:

$$Y_i \sim G(\mu, \sigma)$$

$$\boxed{Y_i = \underbrace{\mu}_{\text{DETERMINISTIC PART}} + \underbrace{R_i}_{\text{RANDOM PART}},} \quad R_i \sim G(0, \sigma)$$

degrees of freedom = n − # of unknowns in the determinish part.

Regression model

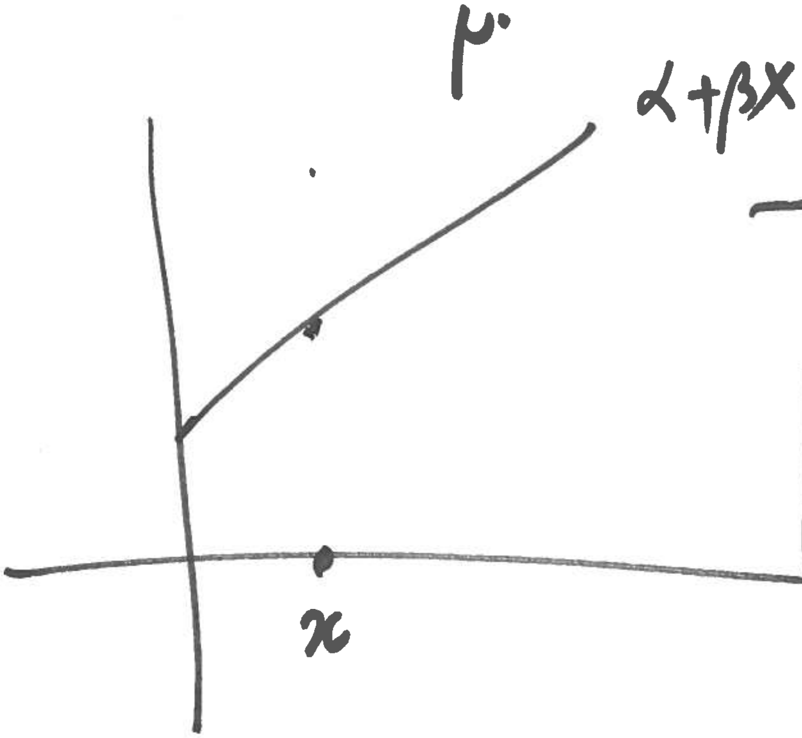
$$Y_i \sim G(\overbrace{\alpha + \beta x_i}^{\mu},\ \sigma)$$

$i = 1, \ldots, n$

$$\Downarrow$$

$$Y_i = \alpha + \beta X_i + R_i$$

$$R_i \sim G(0, \sigma)$$



$$\boxed{\text{Relevant } df = n - 2}$$

ESTIMATES

Method of Least Squares

Method of Max. Likelihood.

The estimates of α and β are the same for both methods.

$$\left.\begin{aligned} \hat{\alpha} &= \bar{y} - \hat{\beta}\,\bar{x} \\ \hat{\beta} &= S_{xy}/S_{xx} \end{aligned}\right\} \text{Least Square Estimates}$$

$$\hat{\sigma}^2 = \frac{1}{n}\left[S_{yy} - \hat{\beta}\,S_{xy}\right]$$

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

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

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

Given a data set, we can always estimate $\hat{\alpha}, \hat{\beta}, \hat{\sigma}$ ($\hat{\beta}$ should be estimated first)

$$\hat{\sigma}^2 = \boxed{\cancel{S}^2 = \frac{1}{n-2}\left[S_{yy} - \hat{\beta} S_{xy}\right]}$$

Sample Variance.

The Least Square Equation:

$$\boxed{\hat{y} = \hat{\alpha} + \hat{\beta} x.}$$

(best estimate for the linear relationship between X and Y).

$\cancel{\hat{\alpha} = 0.9}$

$$\left.\begin{aligned} \hat{\beta} &= 0.9 \\ \hat{\alpha} &= 10 \end{aligned}\right\}$$

$$\boxed{y = 10 + 0.9X}$$

y = STAT 231

X = STAT 230

Confidence Interval for β.

Fact 1:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{S_{xx}}.$$

$$= \boxed{\frac{\sum (x_i - \bar{x}) y_i}{S_{xx}}} = \frac{\sum x_i (y_i - \bar{y})}{S_{xx}}$$

$$= \frac{\sum x_i y_i - n \bar{x} \bar{y}}{S_{xx}}.$$

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

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

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

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

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

$$= \sum(x_i - \bar{x})y_i - \bar{y}\left[\sum x_i - \sum \bar{x}\right]$$

$$- \bar{y}\left[n\bar{x} - n\bar{x}\right]$$

$$0$$

$$\hat{\beta} = \sum a_i y_i$$

where $a_i = \dfrac{(x_i - \bar{x})}{S_{xy}}$

$$\boxed{\hat{\beta} = \sum_{i=1}^{n} a_i y_i} \qquad a_i = \frac{x_i - \bar{x}}{S_{xx}}$$

Each individual $\hat{\beta}$ can be thought of as outcome of a r.v. $\tilde{\beta}$.

What is the distribution of $\hat{\tilde{\beta}}$?

$$\tilde{\beta} = \sum a_i Y_i$$

Since Y_i's are all Gaussian,

$\tilde{\beta}$s are also Gaussian.

<u>Theorem</u>: $\boxed{\tilde{\beta} \sim G\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right)}$

$$\frac{\tilde{\beta} - \beta}{\frac{\sigma}{\sqrt{S_{xx}}}} = Z$$

$$\boxed{\frac{\tilde{\beta} - \beta}{S / \sqrt{S_{xx}}} \sim T_{n-2}}$$

95%

$n = 20$

(Row = 18
Column = 0.975)

Step 1: To find t^*

Step 2 $P(-t^* < T < t^*) = 0.95$

$$P\left(-t^* < \frac{\tilde{\beta} - \beta}{S / \sqrt{S_{xx}}} < t^*\right) = 0.95$$

Coverage Interval:

$$\left(\tilde{\beta} \pm t^{\alpha} \frac{s}{\sqrt{s_{xx}}} \right)$$

Confidence Interval

$$\boxed{\hat{\beta} \pm t^{\alpha} \frac{s}{\sqrt{s_{xx}}}}$$

Confidence Interval for β

$$H_0: \beta = \beta_0$$

$$D = \left| \frac{\hat{\beta} - \beta_0}{S/\sqrt{S_{xx}}} \right|$$ Test Statistic

Calculate $d = \left| \dfrac{\hat{\beta} - \beta_0}{\dfrac{s}{\sqrt{S_{xx}}}} \right|$

p-value: $P(D \geq d)$

$$= P(|T_{n-2}| \geq d)$$

$(x_1, y_1), \ldots \quad (x_n, y_n).$

Clicker question 1

All regression lines must pass through $(\bar{x}, \bar{y})$

(a) True 61%

(b) False. 39%

$$y = \hat{\alpha} + \hat{\beta} x.$$

$$= \hat{\alpha} + \hat{\beta} \bar{x}$$

$$= \bar{y} - \cancel{\hat{\beta} \bar{x}} + \cancel{\hat{\beta} \bar{x}}$$

$$\bar{y} = \hat{\alpha} + \hat{\beta} \bar{x}$$

Clicker Question 2

If $r_{xy} = 0$, then $\hat{\beta} = 0$

(a) True $\longrightarrow$ 74%

(b) False.