Lecture 9

June 10, 2019 6:47 PM

Note:

1. t-distribution. assumption: the population of the sample should be normally distributed

2. Recall $H_0: \mu_1 = \mu_2$
   $H_1: \mu_1 \neq \mu_2$ (two tail for two sample mean)

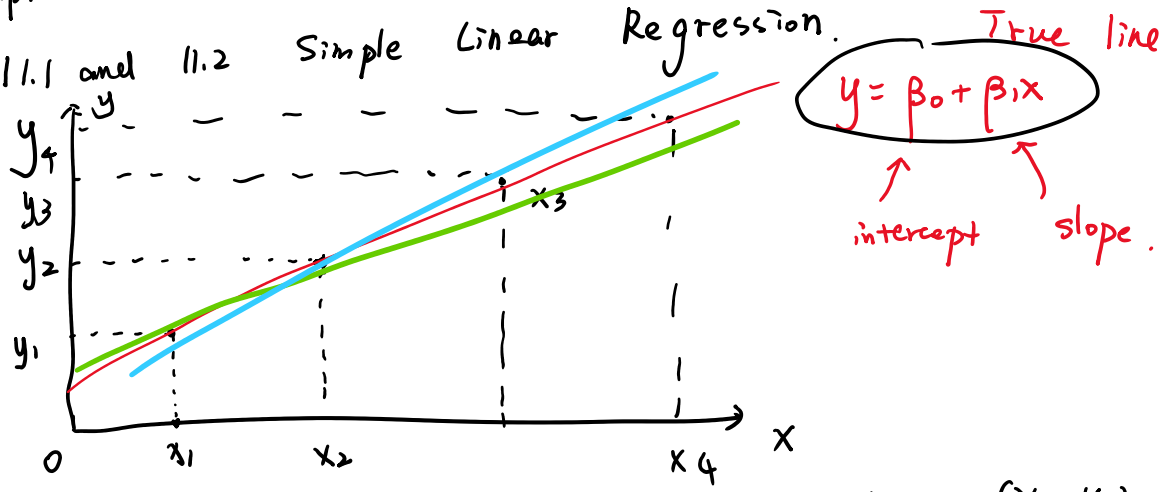
In general, when testing the above hypothesis.

We reject H_0 at significance level α if and only if the $100(1-\alpha)\%$ CI for $(\mu_1 - \mu_2)$ does not contain zero.

$\mu_1 - \mu_2$ 95% CI. $(-1, -0.5)$

Chapter 11.

11.1 and 11.2 Simple Linear Regression.



True line $y = \beta_0 + \beta_1 x$

intercept slope.

Suppose we have n pairs of observations (x_1, y_1) (x_2, y_2), (x_3, y_3) ... (x_n, y_n).

Estimate Line:

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ or $\hat{y} = b_0 + b_1 x$

$\hat{y}_1 \Rightarrow y_1$, $\hat{y}_2 \Rightarrow y_2$ ----

$\hat{\beta}_0 = b_0 \Rightarrow \beta_0$, $\hat{\beta}_1 = b_1 \Rightarrow \hat{\beta}_1$

Least square method:

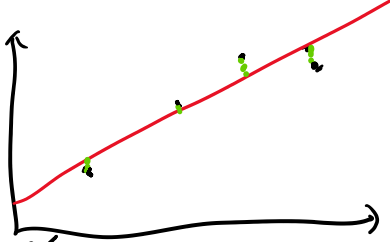
$e_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n$.

estimate line

Least square

$$e_i = y_i - \hat{y}_i \quad , \quad i = 1, 2, \ldots, n.$$

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



min L

or minimize the sum of differences' square.

$$L = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \quad \text{with two unknown } \hat{\beta}_0, \hat{\beta}_1$$

$$\frac{\partial L}{\partial \hat{\beta}_0} = 0 \qquad \frac{\partial L}{\partial \hat{\beta}_1} = 0$$

$$\Rightarrow \begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = \dfrac{S_{xy}}{S_{xx}} \end{cases}$$

where $S_{xy} = \sum_{i=1}^{n} x_i y_i - \dfrac{\sum x_i \sum y_i}{n}$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum x_i)^2}{n}$$

Least squares estimators of β_0 and β_1

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

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

Example:

x	1	2	4	7
y	5	3	2	1

Find $\hat{\beta}_1$, $\hat{\beta}_0$.

$\hat{\beta}_1$, $\hat{\beta}_0$ $\quad \sum x_i^2$ $\quad \sum xy$ $\quad \sum y^2$

No.	x	y	x^2	xy	y^2
1	1	5	1	5	25
2	2	3	4	6	9
3		2	16	8	4

1	1	3	4	6	9
2	2	2	16	8	4
3	4	1	49	7	1
4	7				
sum	$\Sigma x = 14$	$\Sigma y = 11$	$\Sigma x^2 = 70$	$\Sigma xy = 26$	$\Sigma y^2 = 39$

$$S_{xx} = \sum x_i^2 - \frac{(\Sigma x)^2}{n} = 70 - \frac{14^2}{4} = 21$$

$$S_{xy} = \sum xy - \frac{(\Sigma x)(\Sigma y)}{n} = 26 - \frac{14 \cdot 11}{4} = -12.5$$

$$\therefore \begin{cases} \hat{\beta}_1 = -\frac{25}{42} \\ \hat{\beta}_0 = \frac{11}{4} - \left(-\frac{25}{42}\right) \times \frac{14}{4} = \frac{29}{6} \end{cases}$$

$$\hat{y} = \frac{29}{6} - \frac{25}{42}x$$

whats the expected change in y if x is increased by ②?

$$\Delta y = 2 \cdot \left(\frac{-25}{42}\right)$$

Definition:

Residual: $e_i = y_i - \hat{y}_i$

$$\begin{aligned} e_1 &= y_1 - \hat{y}_1 \\ &= 5 - (\hat{\beta}_0 + \hat{\beta}_1 x_1) \\ &= 5 - \left(\frac{29}{6} - \frac{25}{42} \cdot 1\right) \\ &= 0.762 \end{aligned}$$

Assumptions: ε_i are R.V.

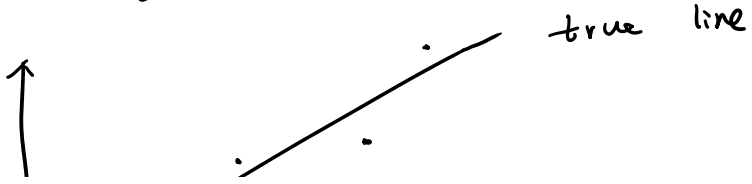
1. $E(\varepsilon_i) = 0$ where $\varepsilon_i = y_i - \beta_0 - \beta x_i$

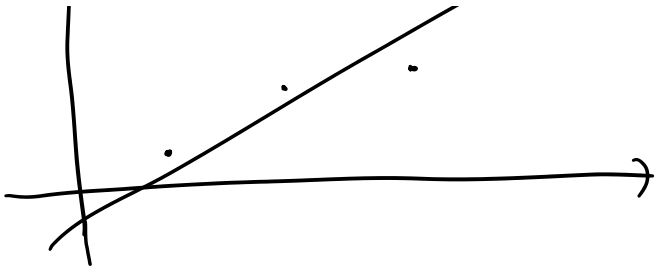
   or $y_i = \beta_0 + \beta x + \varepsilon_i$ ← e_i

2. $Var(\varepsilon_i) = \sigma^2$

3. $cov(\varepsilon_i, \varepsilon_j) = 0$ for any i and j.

   or ε_i are not correlated.





σ^2 estimating.

The sum of square error is defined by:

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

$\hat{\sigma}^2 = \frac{SSE}{n-2}$ which is an unbiased estimator of σ^2.

(similar (sample SD. $S = \frac{\sum (x_i - \bar{x})^2}{n-1}$ is an unbiased estimator of population SD σ))

The total sum of squares is defined by

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

then $SSE = SST - \hat{\beta}_1 S_{xy}$

$$= 39 - 4\left(\frac{11}{4}\right)^2 + \frac{25}{42}(-12.5)$$

$$= 1.3095$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1.3095}{4-2} = 0.65476$$

which is an estimator of $Var(\varepsilon) = \sigma^2$

11.3 Properties of Least Squares Estimators.

(i) $\hat{\beta}_1$ is an unbiased estimator of β_1.

ie. $E(\hat{\beta}_1) = \beta_1$

(ii) $Var(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{S_{xx}}$

(iii) $\hat{\beta}_1$ is a normal r.v. (since is a linear combination of normal R.V.)

(iii) $\hat{\beta}_1$ is a normal r.v. (since is a linear combination of normal R.V).

$[\varepsilon_i \sim N(0, \sigma^2)]$ for all i.

11.4 Hypothesis test for Least Square Estimator:

$y = \beta_0 + \beta_1 X + \varepsilon$ true line.

if β_1 is not significantly different from 0.

$\Leftrightarrow$ y and X are not linear correlated.

① Hypothesis: $H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$ (two tail)

② Test Stat:

$$t_{n-2} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

where $SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{Sxx}}$

$df = n-2$

③ ④ are same as usual.

Example (Cont.)

$n = 4 \quad df = n-2 = 4-2 = 2$

$$t_2 = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{\frac{\hat{\sigma}}{\sqrt{Sxx}}}$$

$$= \frac{\frac{-25}{42} - 0}{\frac{0.65476}{\sqrt{21}}}$$

$= -3.371$ = test stat.

Critical value at $\alpha = 5\%$ two tail

$t^*_{2, \frac{\alpha}{2}, \text{two tail}} = 4.303$

$t^*_{2,\frac{\alpha}{2},\text{two tail}} = 4.303$

$|\text{Test Stat}| < |\text{Critical Value}| \iff p\text{-value} > 0.05$

$\therefore$ do not reject H_0.

we can not say there is a significant difference between β_1 and zero.

Analysis of Variance (ANOVA). F-test.

We define the total sum of squares

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

the error sum of squares

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

the regression sum of squares

③ $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 \cdot S_{xy}$

Note: $SST = SSE + SSR$.

$E(\hat{\sigma}^2) = E\left(\frac{SSE}{n-2}\right) = \sigma^2$.

① Hypothesis: $H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$.

② Test Stat (F test) $\quad F_{1,2} \neq F_{2,1}$

$$F_{df_1, df_2} = \frac{SSR}{SSE/n-2}$$

where $df_1 = 1$

$df_2 = n-2$

$\boxed{F = t^2}$

where $df_1 = 1$

$df_2 = n-2$.

$\boxed{F = t^2}$

③ ④ as usual. F is always positive.

Example: (Cont.)

Hypothesis: $H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$.

Test Stat:

$$F_{1,2} = \frac{SSR}{SSE/n-2}$$

$$= \frac{\hat{\beta}_1 S_{xy}}{\hat{\sigma}^2}$$

$$= \frac{\frac{-25}{42} \cdot \left(\frac{-25}{2}\right)}{1.3095/2}$$

$= 11.36 = (t_2)^2 = (-3.371)^2$ ← t-statistic.

Critical value (at $\alpha = 5\%$)

$$F^*_{1,2,0.05} = 18.51$$

Since |Test Stat| < |Critical value|

∴ do not reject H_0.

Summary of ANOVA:

Source of Variation	SS	df	MS	F
Regression	SSR	1	SSR/1 = MSR	MSR/MSE
Error	SSE	n-2	SSE/n-2	

Error	SSE	$n-2$	$SSE/n-2$ = MSE
Total	SST	$n-1$	

11.5 Confidence Interval for the slope β_1.

Recall from section 11.3 that $E(\hat{\beta}_1)=\beta_1$, $Var(\hat{\beta}_1)=\frac{\sigma^2}{S_{xx}}$

∴ Find

$100(1-\alpha)\%$ CI for β_1

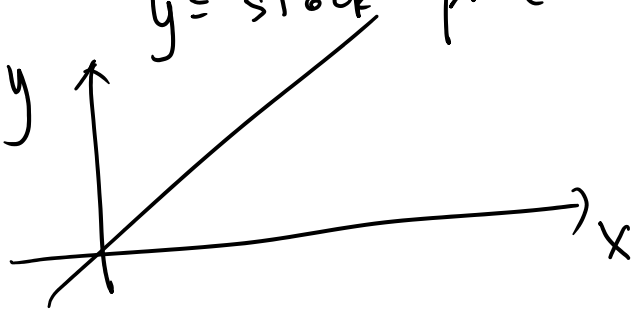
$$\hat{\beta}_1 \pm t^*_{n-2,\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

Example (Cont.):

95% CI: $(-1.355, 0.646)$

CI for mean response when $X = X_0$.

y = stock price x = gas price



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

For $X = X_0$, what's mean response of y.

A $100(1-\alpha)\%$ CI for mean of y when $X = X_0$. is

$$\hat{y}_0 \pm t^*_{n-2,\frac{\alpha}{2}} \cdot \hat{\sigma}\sqrt{\frac{1}{n}+\frac{(X_0-\bar{X})^2}{S_{xx}}}$$

where $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$

where $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$

Example: (Cont.) when $x = x_0 = 5$ is

$2.0556 \pm 2.44 \Rightarrow (-0.384, 4.4955)$

A $100(1-\alpha)\%$ CI for a **single observation** y_0 when $x = x_0$ is:

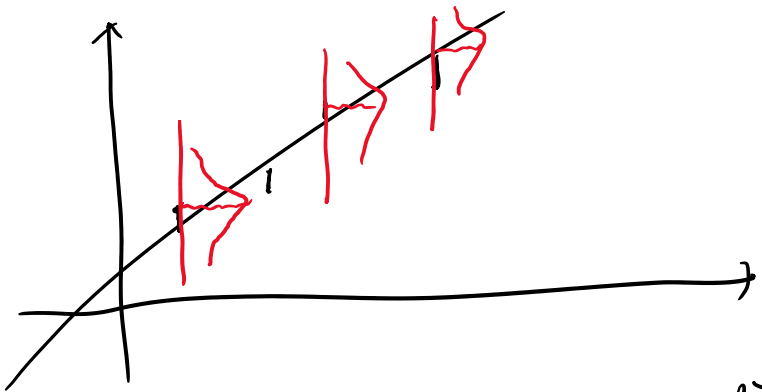
$$\hat{y}_0 \pm t^*_{n-2, \frac{\alpha}{2}} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Example:

$2.0556 \pm 4.2517 \Rightarrow (-2.196, 6.307)$

11.7 Adequacy of the Regression Model:

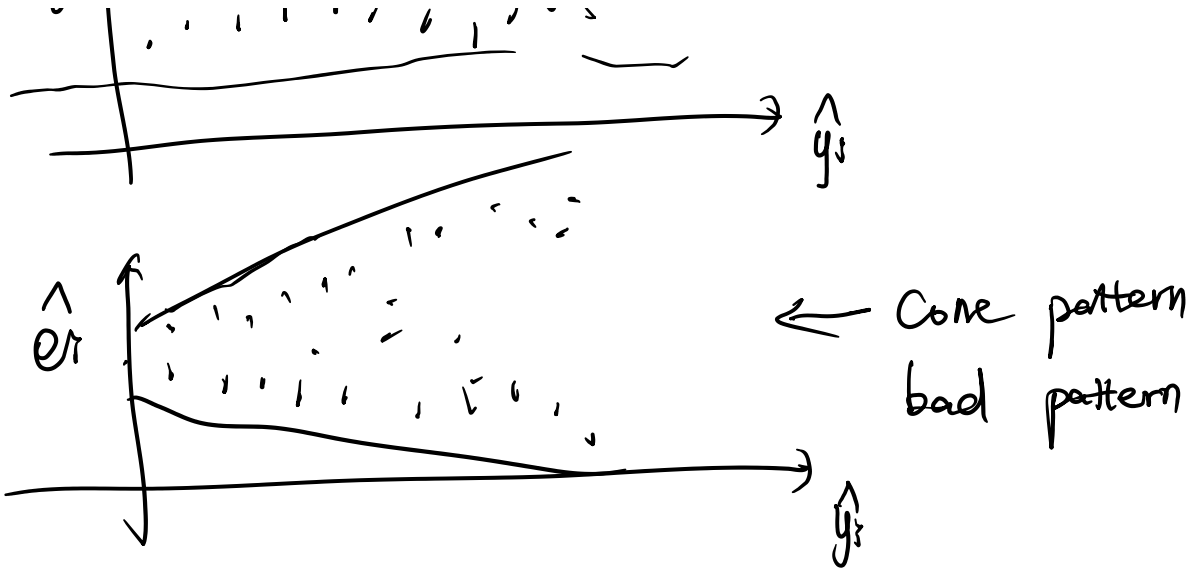
$\varepsilon_i \sim N(0, \sigma^2)$ ← we want to estimate.



$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \sim N(\ ,\ )$

If the equal variance assumption holds, σ^2 is constant

← good pattern



11.8 Correlation Coefficient.

Correlation coefficient:

$$R = r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot SST}}$$

Note:

(1) $-1 \le R \le 1$

(2) If R is close to 1, then there is a strong positive correlation between x and y.

(3) ---- -1 ---- negative ---- .

(4) $R = 0$, x and y are linearly independent.

Coefficient of determination:

$$r^2 = R^2 = \frac{SSR}{SST}$$

Test x and y are independent linear ($\hat{\beta}_1$ test)

$r = 0 \iff \beta_1 = 0$

Do a test: true value of r. ρ

R/r is an estimator of ρ.

① Hypothesis: $H_0: \rho = 0$

$H_a: \rho \neq 0$

$\Leftrightarrow$ same result with test β_1 (t and F)