3Y03 - Probability and Statistics for Engineering

WS19 Lecture 31

Last Time

More on Linear Regression

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

We assume $E(\varepsilon) = 0$, call $V(\varepsilon) = \sigma^2$

Estimator $\hat{\sigma}^2 = SS_E/(n-2)$

where $SS_E = SS_T - SS_R$

$$\underbrace{SS_E}_{\sum(y_i-\hat{y}_i) = \sum e_i^2} = \underbrace{SS_T}_{\sum y_i^2 - n\bar{y}^2} - \underbrace{SS_R}_{\hat{\beta}_1 S_{xy}}$$

Estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ have: $E(\hat{\beta}_0) = \beta_0$, $E(\hat{\beta}_1) = \beta_1$

as well as: $V(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$, $V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$

11.4 Tests in Simple Linear Regression

e.g. $H_0 : \beta_1 = (\beta_1)_0 \leftarrow (\beta_1)_0 = 0$ says No linear relationship

$H_1 : \beta_1 \neq (\beta_1)_0$

We assume $\varepsilon \sim \underline{\underline{N}}(0, \sigma^2)$

For given x_i values $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$

($\hat{\beta}_1$ linear comb. of normal r.v.s for fixed x_is)

Under $H_0 : \beta_1 = (\beta_1)_0$, $\frac{\hat{\beta}_1 - (\beta_1)_0}{\sqrt{\hat{\sigma}^2/S_{xx}}} = T_0$

$T_0 \sim t_{n-2}$ - distribution

comes from $\hat{\sigma}^2 = \frac{1}{n-2} SS_E$

If H_1 2-sided as above, critical region

$$|t_0| > t_{\frac{\alpha}{2}, n-2}$$

Example - see sheet

11.5 Confidence Intervals

(will go back to 11.4 later)

From above, a $100(1-\alpha)\%$ C.I. for β_1 is given by

$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}, n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

Example — see sheet.

Want to use the regression line to

① understand underlying distribution of Y (in terms of X)

② predict future y-values (in terms of X).

① Can estimate expected y-value ("response") at a given $X = x_0$ with $\left.\hat{\beta}_0 + \hat{\beta}_1 x_0\right\}$ estimator $\hat{\mu}_{Y|x_0}$

the mean of Y at $X = x_0$

call this $E(Y|x_0)$ or $\mu_{Y|x_0}$

Need $E(\hat{\mu}_{Y|x_0}) = E(\hat{\beta}_0) + E(\hat{\beta}_1)\,x_0$

$$= \beta_0 + \beta_1 x_0 = \mu_{Y|x_0}$$

$$V(\hat{\mu}_{Y|x_0}) = \sigma^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)$$

and $\hat{\mu}_{Y|x_0}$ is normally distributed.

We get
$$\frac{\hat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}\text{-distr.}$$

So a $100(1-\alpha)\%$ C.I. for $\mu_{Y|x_0}$ (expected value of Y at $X = x_0$) is given by

$$\underbrace{\hat{\mu}_{Y|x_0}}_{\hat{\beta}_0 + \hat{\beta}_1 x_0} \pm t_{\frac{\alpha}{2}, n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

<u>Example</u> - see sheet

11.6 Prediction of New Observations

Point Estimator: $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$

(Y_0 : actual observation at $X = x_0$ that we're trying to predict)

Need now to take possible error into account to predict a single observation:

$$e_{\hat{p}} = Y_0 - \hat{Y}_0$$

stands for "prediction" here → $\hat{p}$

Y_0 ↑ actual observation

$\hat{Y}_0$ ← our prediction

$$E(e_{\hat{p}}) = 0 \quad , \quad V(e_{\hat{p}}) = V(Y_0) + V(\hat{Y}_0)$$

$$= \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$$

$$= \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$$

So

$$\frac{\overbrace{Y_0 - \hat{Y}_0}^{e_{\hat{p}}}}{\sqrt{\sigma^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}\text{-distribution}$$

We can define a so-called 100(1−α)% prediction interval for future observation Y_0 (at $X = x_0$):

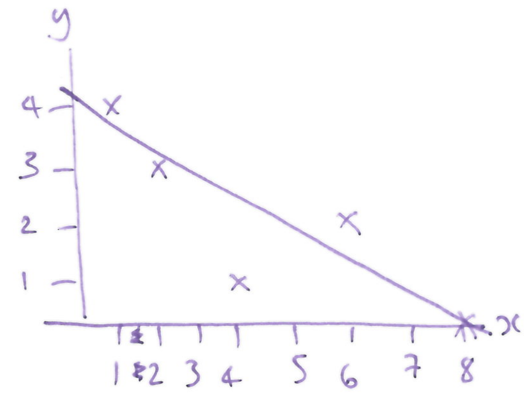
$$\hat{y}_0 \pm t_{\frac{\alpha}{2}, n-2} \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

$\hat{\beta}_0 + \hat{\beta}_1 x_0$

extra margin of error when trying to predict a single observation

<u>Example</u> – see sheet.

Least Squares Regression Example



$n=5$

i	x	y	x^2	y^2	xy
1	1	4	1	16	4
2	2	3	4	9	6
3	4	1	16	1	4
4	6	2	36	4	12
5	8	0	64	0	0
SUM	21	10	121	30	26

$\bar{x} = \frac{21}{5} = 4.2 \quad \bar{y} = \frac{10}{5} = 2$

$S_{xy} = \sum xy - n\bar{x}\bar{y} = 26 - 5(4.2)(2) = -16$

$S_{xx} = \sum x^2 - n\bar{x}^2 = 121 - 5(4.2)^2 = 32.8$

same

$S_{yy} = \sum y^2 - n\bar{y}^2 = 30 - 5(2)^2$

$=$

$SS_T = S_{yy} = 10$

$SS_R = \hat{\beta}_1 S_{xy} = (-0.49)(-16) \; \cancel{= 8} = 7.84$

$SS_E = SS_T - SS_R = \cancel{10 - 8} \quad 10 - 7.84 = 2.16$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-16}{32.8} = -0.49$ slope

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 2 - (-0.49)(4.2) = 4.05$ intercept

$\underline{\hat{y} = 4.05 - 0.49x}$

$\hat{\sigma}^2 = \frac{1}{n-2} SS_E = \frac{1}{5-2}\left(10 - \cancel{(-0.49)(-16)}^{7.84}\right) = \frac{1}{3}(2.16) = 0.72$

$R^2 =$

$R =$

Tests for β_1:

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$\alpha = 0.05$

$t_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = \frac{-0.49 - 0}{\underbrace{\sqrt{0.72/32.8}}_{0.14}} = -3.31$ vs. $t_{\frac{\alpha}{2}, n-2} = t_{0.025,3} = -3.182$

More extreme so reject H_0 - yes, a linear rel.

$f_0 =$

95% CI for β_1: $\hat{\beta}_1 \pm t_{0.025,3}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = -0.49 \pm 3.182(0.14) = -0.49 \pm 0.47 = (-0.96, -0.02)$

Notice 0 not in here: another way to reject $H_0: \beta_1 = 0$ for $H_1: \beta_1 \neq 0$

95% CI for Y at $x_0=3$:

$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{0.025,3}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)} = (4.05 - 0.49(3)) \pm 3.182\sqrt{0.72\left(\frac{1}{5} + \boxed{\frac{(3-4.2)^2}{32.8}}\right)} = (1.25, 3.91)$

95% PI for Y_0 at $x_0=3$:

" " " $\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$

$= (4.05 - 0.49(3)) \pm 3.182\sqrt{0.72\left(1 + \frac{1}{5} + \frac{(3-4.2)^2}{32.8}\right)} = (-0.43, 5.59).$