STA 302 / 1001

Note Title 10/6/2011

The test will be on Tues. Oct. 25 10:10 - 11:40 in EX 200

It will cover material to the end of lecture on Thurs. Oct. 13

More details to come...

Reminder Tomorrow Andriy in RW 109 13:10 - 15:00

Exercise <u>from last</u> class:

Use $\sum \hat{e}_i \hat{y}_i = 0$ to show

$$r_{\hat{e}_i, \hat{y}_i} = 0 \quad , \quad r_{\hat{e}_i, x_i} = 0$$

<u>Crime Example</u>

<u>Moral</u>: high R^2, low p-value for test $H_0: \beta_1 = 0$ do not tell you whether regression model is <u>appropriate</u>

Need to look at data (scatterplot)

Prediction

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

Least Squares line: $\hat{y} = b_0 + b_1 x$

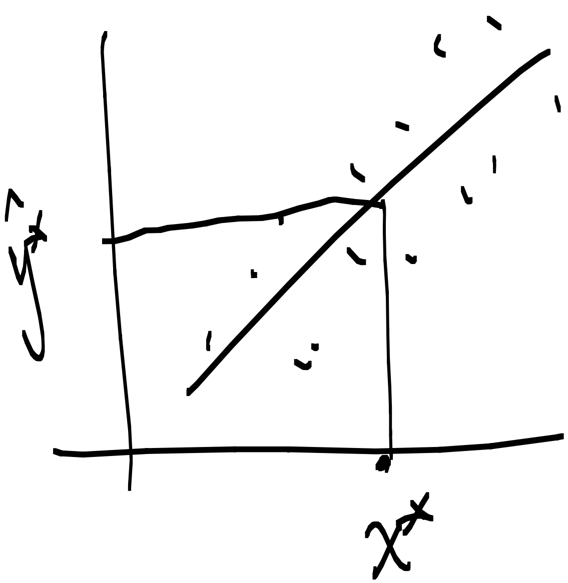
$Var(\hat{\beta}_1) = \sigma^2 / S_{xx}$, $Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$

$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \bar{x} / S_{xx}$

$\hat{y}$ is an estimate of the mean of the distribution of the y's when $X = x$

$\hat{y}$ is an est. of $E(Y \mid X = x) = \underline{\beta_0 + \beta_1 x}$

Want an indication of precision of this est.



$$\hat{y}^* = b_0 + b_1 x^*$$

Estimators:

$$\text{Var}(\hat{y}^* \mid X = x^*)$$

$$= \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x^*)$$

$$= \text{Var}(\hat{\beta}_0) + x^{*2}\,\text{Var}(\hat{\beta}_1) + 2x^*\,\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$$

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

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

Est Var of $E(Y \mid X = x^*)$: $s^2 \left\{ \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right\}$

For a given value of x, x^*

A 95% CI for $E(Y \mid X = x^*) = \beta_0 + \beta_1 x^*$

$$\hat{y}^* \pm t_{n-2, .025}\, s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}, \quad \text{where } s = \sqrt{\sum \hat{e}_i^2 / (n-2)}$$

Example pre MP CFCs. Interested in May 1, 1987

time $= 1987 + 4/12 = x^*$

$$\hat{y}^* = -19064 + 9.71152\,(1987 + 4/12)$$

$$= 236.0274$$

s.e. of $E(\hat{y}^*$ at $x^*) = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xy}}}$

$$= 2.40084\sqrt{\frac{1}{153} + \frac{(1987.33 - 1983.435)^2}{S_{xx}}}$$

$$= 0.2804754$$

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

$= 601906139 - 153(1983.435)^2$

or $S_{xx} = (n-1)\text{ var of } x = 152(14.169)$

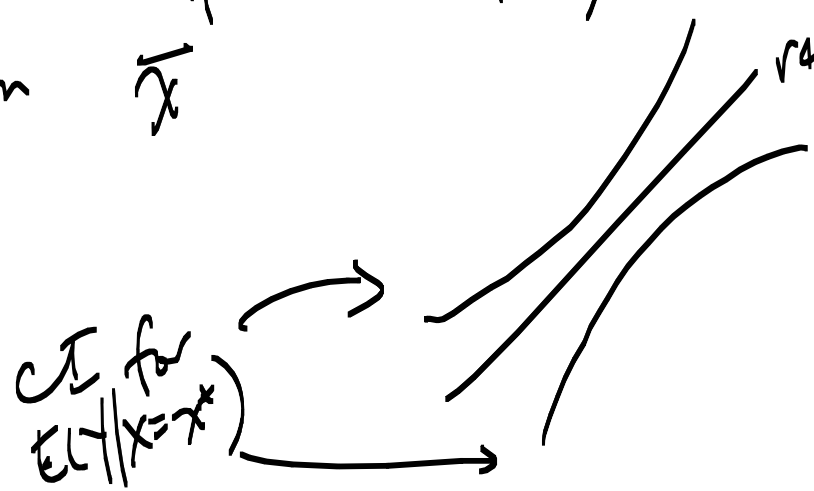
95% CI for $E(Y)$ at May 1, 1987

$$t_{151,\,.025} = 1.976$$

$$236.0274 \pm 1.976\,(0.28048)$$

$$= (235.47,\ 235.58)$$

Note that CI for $E(Y \mid X = x^*)$ is wider the further x^* is from $\bar{x}$



Suppose we want to use the regression line to predict a particular y^+ given $X = x^+$

"Prediction Interval" (<u>not</u> a confidence interval)

CIs are for parameters, this prediction interval is for a r.v.

Predicted value $\hat{y}^+ = b_0 + b_1 x^+$

Using $\hat{y}^+$ to forecast what actual value would be (if observed) y^+

Assume y^+ comes from model:

$$Y^+ = \beta_0 + \beta_1 x^+ + e^+$$

$$\begin{aligned}
&Var(Y^+ - \hat{Y}^+ \mid X = x^+) \\
&= Var(Y^+ \mid X = x^+) + Var(\hat{Y}^+ \mid X = x^+) \\
&\qquad + 2\,cov(Y^+, \hat{Y}^+ \mid X = x^+) \\
&= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x^+ - \bar{x})^2}{S_{xx}} \right) + 0 \\
&= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x^+ - \bar{x})^2}{S_{xx}} \right)
\end{aligned}$$

because Y^+ is a new observation, unrelate to original n observations

To est s.e. of prediction, put s^2 in for σ^2 and take $\sqrt{}$

95% Prediction Interval for y when $x = x^+$

$$\hat{y}^+ \pm t_{n-2,\,.025}\, s\sqrt{1 + \frac{1}{n} + \frac{(x^+ - \bar{x})^2}{S_{xx}}}$$

<u>Note</u> wider for CI for $E(Y|x)$

<u>Example</u> pre MP CFCs, time $= 1987 + 4/12 = x^+$

$$t_{151,\,.05} = 1.655$$

90% prediction interval when time = $1987 + 4/12$

(232.02, 240.04)

<u>SAS</u> on model statement options:

/ clm cli ;

CI for $E(Y)$ at each value of x in dataset

PI for y at each value of x in dataset

```
plot y*x / conf ;   ← plots CIs for E(Y)
plot y*x / pred ;   ← plots PIs
```

Dummy Variable Regression or Indicator Variable Regression

Model: $Y = \beta_0 + \beta_1 X + e$

Normally distributed.

No distributional assumptions on X

Distribution of $Y|X$ is Normally distributed

Dummy or Indicator Variable

- takes on values 0 or 1
- indicates which category observations are in

Example CFCs

Let $x_i = \begin{cases} 1 & i^{th} \text{ observation is before MP} \\ 0 & \text{" " " after MP} \end{cases}$

Model $Y_i = \beta_0 + \beta_1 x_i + e_i$

$E(Y_i) = \begin{cases} \beta_0 + \beta_1 & i\text{th observation is before MP} \\ \beta_0 & \text{" " " after MP} \end{cases}$

Can estimate $E(Y_i)$ by $\begin{cases} b_0 + b_1 & \text{if } x_i = 1 \\ b_0 & \text{if } x_i = 0 \end{cases}$

Can test whether mean of Y_i is same before & after MP by testing $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$

Equivalent to two-sample t-test assuming equal variances (consistent with Gauss Markov condition)

— End of Chapter 2 —