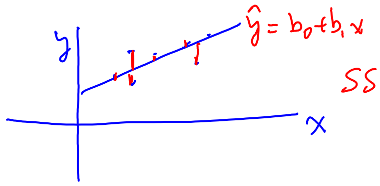
2013-11-28-C02

Thursday, November 28, 2013
10:24 AM

Review



$$SSE = \sum [y_i - \hat{y}_i]^2$$
$$= \sum (y_i - (b_0 + b_1 x_i))^2$$

Make SSE as small as possible & find b_0, b_1

Sol'n $\quad b_1 = \dfrac{SS_{xy}}{SS_{xx}}, \quad SS_{xy} = \sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)$

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

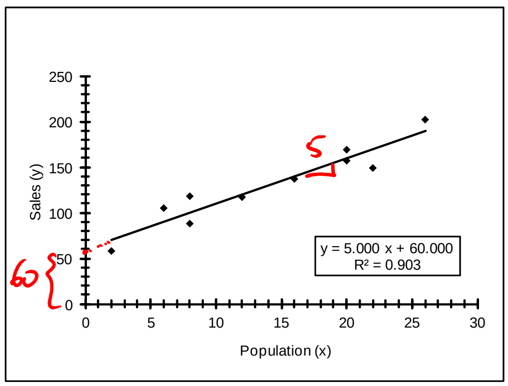
$$b_0 = \bar{y} - b_1 \bar{x}, \quad \bar{y} = \frac{1}{n}\sum y_i, \quad \bar{x} = \frac{1}{n}\sum x_i$$

Ex. Harvey's

http://profs.degroote.mcmaster.ca/ads/parlar/courses/q600/ChapterComments/documents/Harveys.xls

Student Population (x)	Monthly Sales (y)
2	58
6	105
8	88
8	118
12	117
16	137
20	157
20	169
22	149
26	202

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Harveys.xls>



i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	2	58	116	4	3364
2	6	105	630	36	⋮
⋮		⋮	⋮		⋮
10	26	202	5252	676	
	$\sum x_i = 140$	$\sum y_i = 1300$	$\sum x_i y_i = 21{,}040$	$\sum x_i^2 = 2528$	$\sum y_i^2 = 184{,}730$
	$\bar{x} = 14$	$\bar{y} = 130$			

$$SS_{xy} = 21{,}040 - \frac{140 \times 1300}{10} = 2840$$

$$SS_{xx} = 2528 - \frac{(140)^2}{10} = 568$$

$$SS \quad - \quad = 15{,}730$$

$$b_1 = \frac{2840}{568} = 5$$

$$b_0 = 130 - 5 \times 14 = 60$$

$$SS_{xy} = 21{,}040 - \frac{140 \times 1300}{10} = 2840$$
$$SS_{xx} = 2528 - \frac{(140)^2}{10} = 568$$
$$SS_{yy} = \qquad = 15{,}730$$

$$b_1 = \frac{2840}{568} = 5$$

$$b_0 = 130 - 5 \times 14 = 60$$

$$\hat{y} = \underset{\uparrow b_0}{60} + \underset{\uparrow b_1}{5}x$$

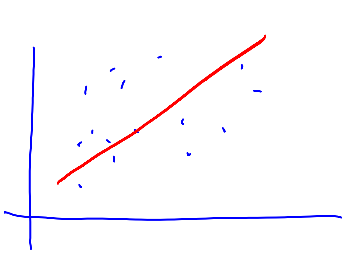
variables	coefficients
Intercept	60.0000
Population (x)	5.0000

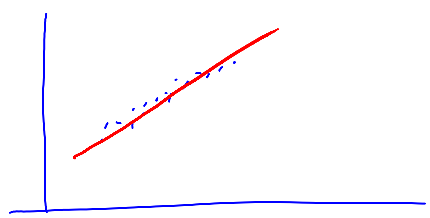
Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Harveys.xls>

$b_1 = 5$: marginal revenue

$x = 10$: $\hat{y} = 60 + 5 \times 10 = 110$
$x = 11$: $\hat{y} = 60 + 5 \times 11 = 115$ $\Big\}\ 5$

$b_0 = 60$: $x = 0$: $\hat{y} = 60 + 5 \times 0 = 60$ (Careful!)





Residuals (errors) & SSE

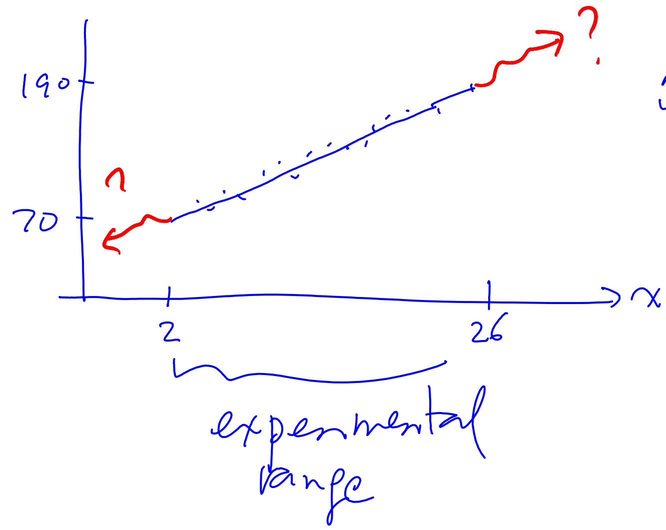
i	x_i	y_i	$\hat{y}_i = 60 + 5x_i$	error $y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	2	58	70	−12	144
2					
⋮	⋮		⋮	⋮	⋮
10	26	202	190	12	144

$$SSE = \sum (y_i - \hat{y}_i)^2 = 1530$$

ANOVA table	
Source	*SS*
Regression	14,200.0000
Residual	1,530.0000
Total	15,730.0000

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Harveys.xls>

Caveat



$\cancel{x = 50}$

$\cancel{\hat{y} = 60 + 5 \times 50}$

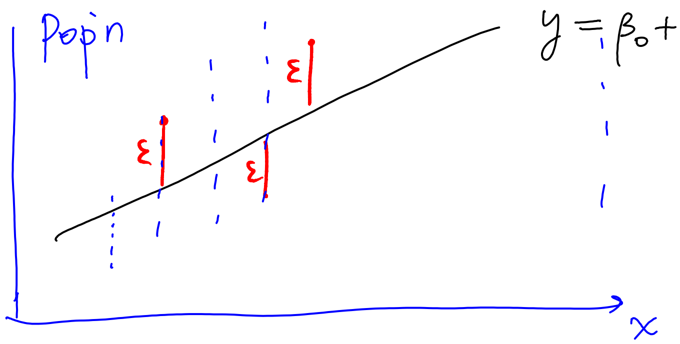
$\cancel{= 310?}$

not reliable

d) Standard error

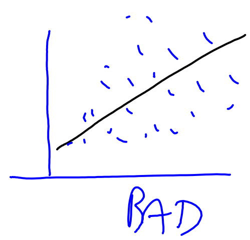
True model

$y = \beta_0 + \beta_1 x + \varepsilon$

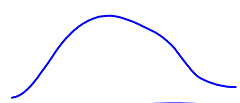


Assumptions about ε

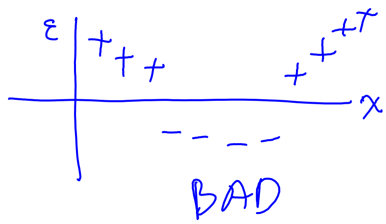
1. ε terms have zero mean
2. " " " constant variance for all x



3. " " is normal



4. " " are independent



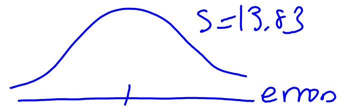
$$SSE = \sum (y_i - \hat{y}_i)^2$$

$$s^2 = \frac{SSE}{n-2}$$: sample var for ε (standard error) $s = \sqrt{s^2}$

Ex. Harvey's

$$s^2 = \frac{1530}{10-2} = 191.25$$

$$s = \sqrt{s^2} = 13.83$$

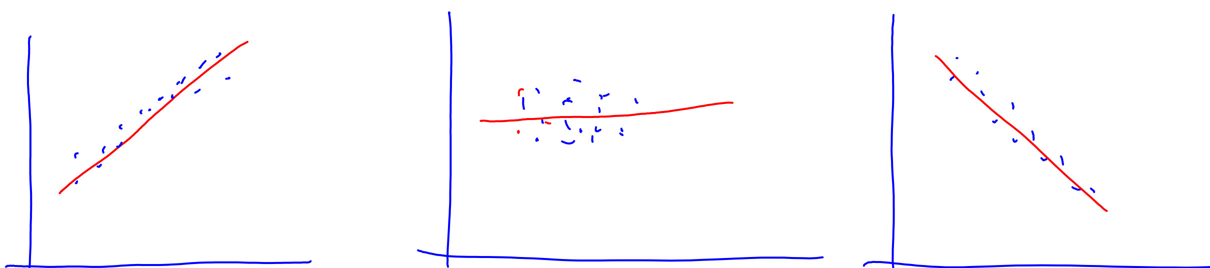


Std. Error	13.829

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Harveys.xls>

$y = \beta_0 + \beta_1 x + \varepsilon$

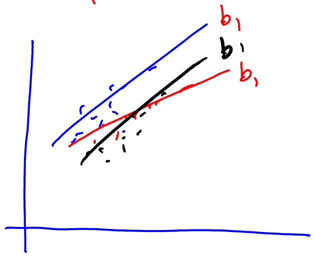
e) How significant is the model?



significant $\beta_1 > 0$ | not significant $\beta_1 \simeq 0$ | significant $\beta_1 < 0$

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$



$$S_{b_1}^2 = \frac{s^2}{SS_{xx}}, \quad S_{b_1} = \frac{s}{\sqrt{SS_{xx}}}$$

If $H_0: \beta_1 = 0$ is true, then $t = \frac{b_1}{S_{b_1}}$, $df = n-2$

$100(1-\alpha)\%$ CI for β_1: $[b_1 \mp t_{\alpha/2} S_{b_1}]$

Ex. Harvey's

$\hat{y} = b_0 + b_1 x = 60 + 5x$

$$\hat{y} = b_0 + b_1 x = \underset{b_0}{60} + \underset{b_1}{5} x$$

$$b_1 = 5, \quad SS_{XX} = 568, \quad S = 13.83$$

$$S_{b_1} = \frac{13.83}{\sqrt{568}} = 0.58$$

Regression output		
variables	*coefficients*	*std. error*
Intercept	60.0000	9.2260
Population (x)	5.0000	0.5803

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Harveys.xls>

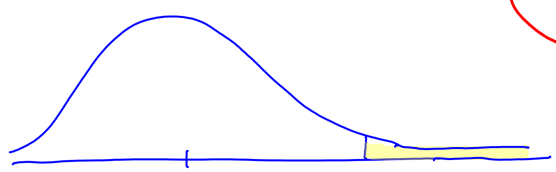
$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

$$t = \frac{b_1}{S_{b_1}} = \frac{5}{.58} = 8.62, \quad df = 10 - 2 = 8$$

Reject H_0

$\alpha = .05$



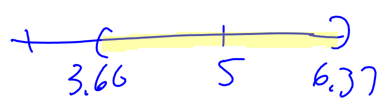
$t_{.025} = 2.306$

95% CI for β_1

$$\left(b_1 \mp t_{\alpha/2} \cdot S_{b_1}\right) = \left[5 \mp 2.306\,(.58)\right]$$

$df = 8$

$$= \left[3.66,\ 6.33\right)$$

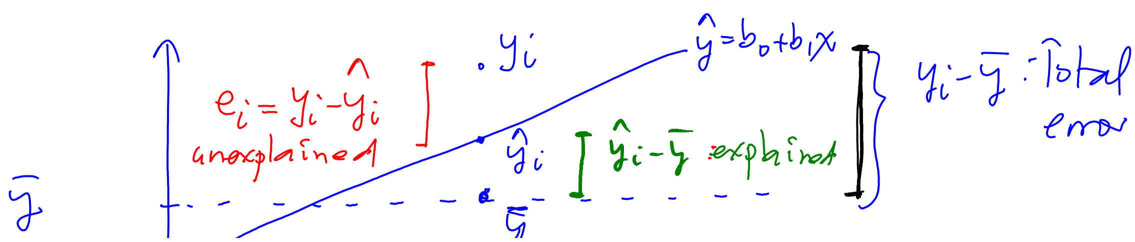


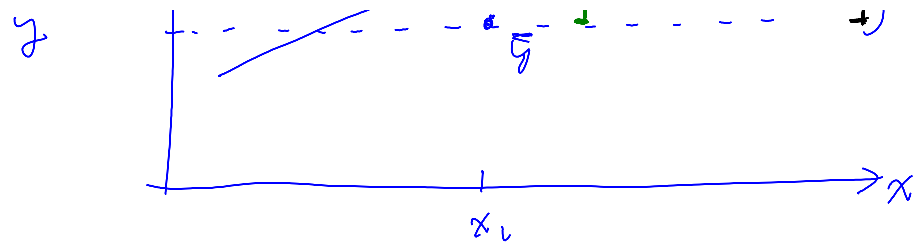
Regression output					confidence interval	
variables	*coefficients*	*std. error*	*t (df=8)*	*p-value*	*95% lower*	*95% upper*
Intercept	60.0000	9.2260	6.503	.0002	38.7247	81.2753
Population (x)	5.0000	0.5803	8.617	2.55E-05	3.6619	6.3381

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Harveys.xls>

f) Simple coefficient of determination (r^2)

" " " Correlation (r)





Variation: $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$

Total unexpl explained

$$\sum (y_i - \bar{y})^2 = \underbrace{\sum (y_i - \hat{y}_i)^2}_{} + \sum (\hat{y}_i - \bar{y})^2$$

Total SS Unexplained SS SSE Explained SS

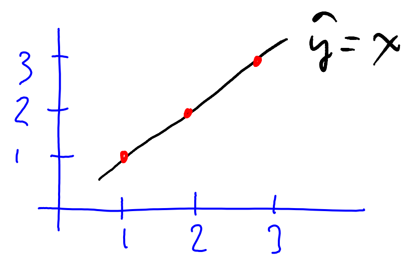
Def. r^2: coeff. of determination

$$r^2 = \frac{\text{explained var}}{\text{total var}} \qquad 0 \le r^2 \le 1$$

r^2	0.903

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Harveys.xls>

Exercise



Show $r^2 = 1$

Coefficient of correlation r

$r = \sqrt{r^2}$ Same as $r = \dfrac{S_{xy}}{S_x S_y}$

$-1 \le r \le +1$

Ex. Harvey's unexpl. expl. Total

i	x_i	y_i	$\hat{y}_i$	$\bar{y}$	$(y_i-\hat{y}_i)^2$	$(\hat{y}_i-\bar{y})^2$	$(y_i-\bar{y})^2$
			Too time consuming				
					1530	14,200	15,730

$$r^2 = \frac{14,200}{15,730} = .903$$

ANOVA table	
Source	*SS*
Regression	14,200.0000
Residual	1,530.0000
Total	15,730.0000

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Harveys.xls>

Easier calc's

Recall $r = \dfrac{S_{xy}}{S_x S_y}$

$$S_{xy} = \frac{SS_{xy}}{n-1} = \frac{2840}{9} = 315.55$$

$$S_x = \sqrt{\frac{SS_{xx}}{n-1}} = \sqrt{\frac{568}{9}} = 7.94$$

$$S_y = \sqrt{\frac{SS_{yy}}{n-1}} = \sqrt{\frac{15730}{9}} = 41.80$$

$$r = \frac{315.55}{(7.94)(41.80)} = 0.95, \quad r^2 = (0.95) = 0.903$$

Regression Analysis		
	r^2	0.903
	r	0.950

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Harveys.xls>

Remark F-test approach to significance

$H_0: \beta_1 = 0$
$H_a: \beta_1 \neq 0$

$$F(\text{model}) = \frac{\text{explained var}/1}{(\text{unexpl. var})/(n-2)} = \frac{\sum(\hat{y}_i - \bar{y})^2/1}{\sum(y_i - \hat{y}_i)^2/(n-2)}$$

$df = (1, n-2)$ — two parameters b_0, b_1

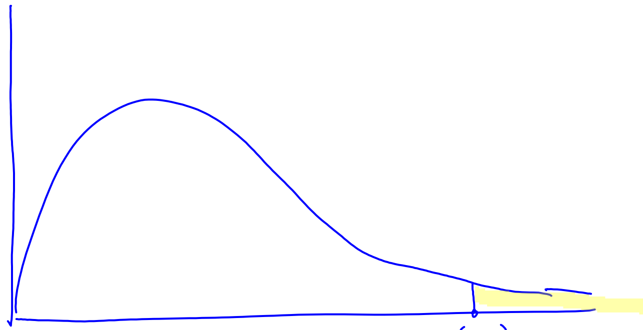
one ind't var (x)

$df = (1, 8)$

Our problem $F(\text{model}) = \dfrac{14{,}200/1}{1530/8} = 74.25$

Reject H_0!

$\alpha = .05$



$F_{.05}^{(1,8)} = 5.32$

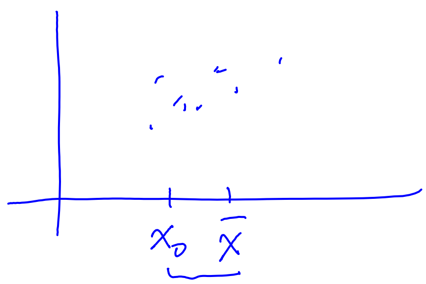
ANOVA table						
Source	*SS*	*df*	*MS*	*F*	*p-value*	
Regression	14,200.0000	1	14,200.0000	74.25	2.55E-05	
Residual	1,530.0000	8	191.2500			
Total	15,730.0000	9				

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Harveys.xls>

http://profs.degroote.mcmaster.ca/ads/parlar/courses/q600/ChapterComments/documents/Harveys-Solution-CI-PI-2013.pdf

8) Confidence interval for true mean &
prediction " " an individual value

. Distance value (DV)
(leverage)

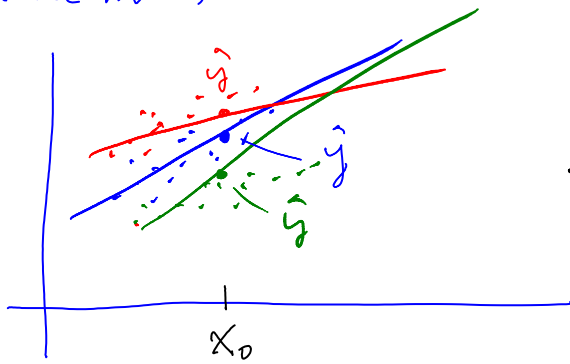


$$DV = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}$$

$\underbrace{x_0 \quad \bar{x}}$

$$DV = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}$$

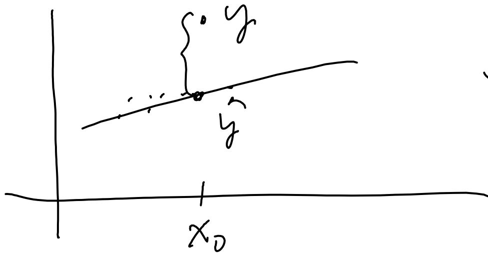
- CI for true mean



$$S_{\hat{y}} = S\sqrt{DV}$$

CI for true mean is $\left[\hat{y} \mp t_{\alpha/2}\, s\sqrt{DV}\right]$

- Pred. Int



$$S_{y-\hat{y}} = S\sqrt{1+DV}$$

less important

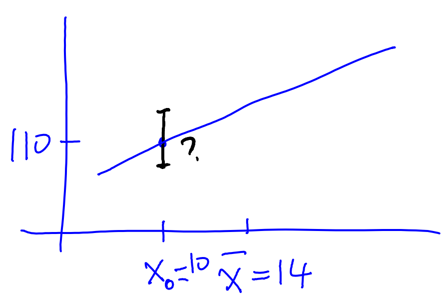
PI $\left[\hat{y} \mp t_{\alpha/2}\, S\sqrt{1+DV}\right)$

Ex. Harvey's

$x_0 = 10$, $\hat{y} = 60 + 5x = 110$, $n = 10$, $SS_{xx} = 568$

$SSE = 1530$

$df = 8$



$$DV = \frac{1}{10} + \frac{(10-14)^2}{568} = .128,$$

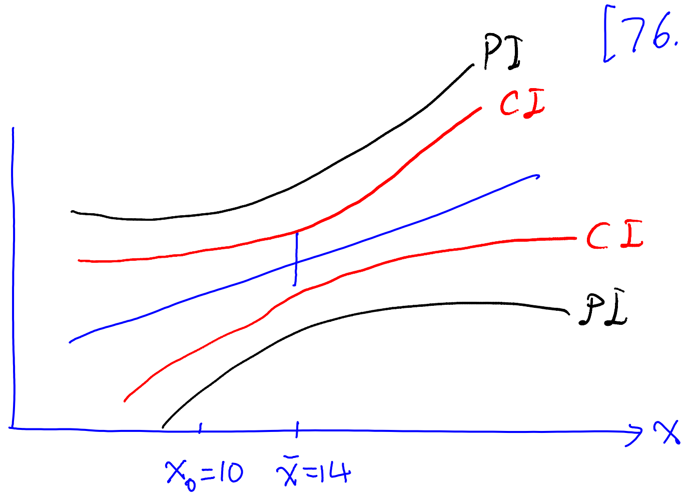
$$DV = \frac{1}{10} + \frac{(10-14)^2}{568} = .128 \quad ,$$

$$S^2 = \frac{SSE}{n-2} = 191.25$$

$$S = 13.83 \quad , \quad t_{.025} = 2.306$$

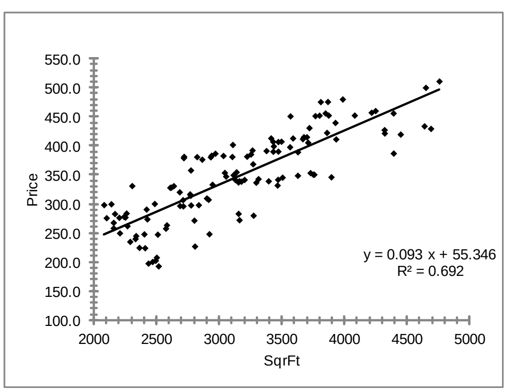
95% CI $\quad [110 \mp 2.306(13.83)\sqrt{.128}] = [98.59, 121.41]$

95% PI $\quad [76.13, 143.87]$



Ex. Real estate data

http://profs.degroote.mcmaster.ca/ads/parlar/courses/q600/ChapterComments/documents/RealEstate Data_002.xls



$x = 2800$

$$\hat{y} = 55.346 + 0.093(2800)$$
$$= \$314,579$$

$r^2 = .69$

$\rightarrow$ Use multiple regression!
Ch.12