2013-12-03-EC01

Tuesday, December 03, 2013
6:11 PM

y x

No.	Price	LotSize	SqrFt	Bedrooms	Bathrooms	Pool?	Basement?	Distance	Rating	SubDiv
1	480.1	1.72	3985	5	4 1/2	1	1	17.0	6	5
2	397.8	0.77	3564	4	4	1	0	12.6	6	4
3	307.4	0.46	2914	3	3	0	1	8.5	6	3
4	413.0	1.56	3413	4	3 1/2	0	0	18.4	4	1
5	389.3	0.50	3627	4	2 1/2	0	0	17.5	1	4
6	353.3	1.21	3727	3	3 1/2	0	1	17.6	4	2
7	331.0	0.38	2304	4	2 1/2	0	0	16.4	4	4

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\RealEstateData_002-1.xls>

124 ,

$\hat{y} = b_0 + b_1 x = 55 + 0.09x$

$x = 2,800$, $\hat{y} = \$314,579$

$r^2 = 0.69$

Ch.12 Multiple regression

Ch.11 x y

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

Ch.12 x_1 x_2 y

x_1 ourprice, x_2 other's price, y sales

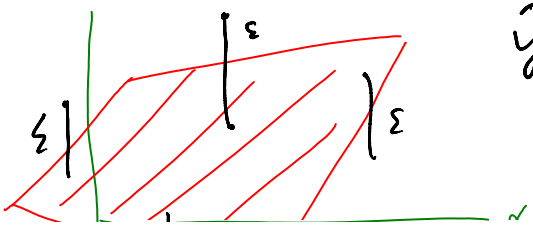
$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$

a) The Model

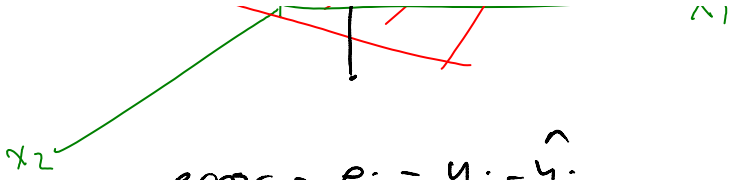
True model

in Ch.11 $Y = \beta_0 + \beta_1 X + \varepsilon$

Ch.12 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$



$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$



$$\text{error} = e_i = y_i - \hat{y}_i$$

$$= y_i - (b_0 + b_1 x_1 + b_2 x_2)$$

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

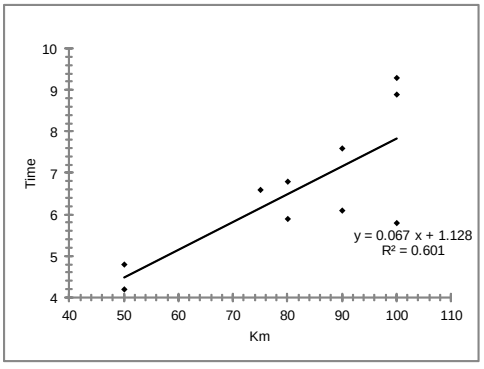
MegaStat does all work!

Ex. Butler trucking Co

http://profs.degroote.mcmaster.ca/ads/parlar/courses/q600/ChapterComments/documents/Butler-Simple.xls

	x1	y
	Km	Time
1	100	9.3
2	50	4.8
3	100	8.9
4	100	5.8
5	50	4.2
6	80	6.8
7	75	6.6
8	80	5.9
9	90	7.6
10	90	6.1

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



Regression Analysis						
	r^2	0.601	n	10		
	r	0.776	k	1		
	Std. Error	1.094	Dep. Var.	**Time**		
ANOVA table						
Source	*SS*	*df*	*MS*	*F*	*p-value*	
Regression	14.4340	1	14.4340	12.07	.0084	
Residual	9.5660	8	1.1957			
Total	24.0000	9				
Regression output					*confidence interval*	
variables	*coefficients*	*std. error*	*t (df=8)*	*p-value*	*95% lower*	*95% upper*
Intercept	1.1285	1.6123	0.700	.5038	-2.5896	4.8466
Km	0.0671	0.0193	3.474	.0084	0.0226	0.1117

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

$r^2 = 0.6$

$\hat{y} = 1.13 + 0.067 x_1$

$x_1 = 85$ km

$\hat{y} = 1.13 + 0.067(85)$

$= 6.83$

$x_1 = 0 \rightarrow \hat{y} = 1.13$??

⇒ Model may not be adequate!

Include now x_2 = #deliveries made

http://profs.degroote.mcmaster.ca/ads/parlar/courses/q600/ChapterComments/documents/Butler-x1-x2_001.xls

	x1	x2	y
	Km	Deliveries	Time
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	5.8
5	50	2	4.2
6	80	1	6.8
7	75	3	6.6
8	80	2	5.9
9	90	3	7.6
10	90	2	6.1

$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Butler-x1-x2_001-1.xls>

(avoid multi-collinearity!) distance & gasoline use / age & experience

Regression Analysis						
	R^2	0.790				
	Adjusted R^2	0.729	n	10		
	R	0.889	k	2		
	Std. Error	0.849	Dep. Var.	**Time**		
ANOVA table						
Source	*SS*	*df*	*MS*	*F*	*p-value*	
Regression	18.9499	2	9.4749	13.13	.0043	
Residual	5.0501	7	0.7214			
Total	24.0000	9				
Regression output					*confidence interval*	
variables	*coefficients*	*std. error*	*t (df=7)*	*p-value*	*95% lower*	*95% upper*
Intercept	0.0367	1.3262	0.028	.9787	-3.0994	3.1727
Km	0.0562	0.0156	3.592	.0088	0.0192	0.0931
Deliveries	0.7639	0.3053	2.502	.0409	0.0419	1.4858

$\hat{y} = 0.0367 + 0.0562 x_1 + 0.7639 x_2$

$x_1 = 85$ km

$x_2 = 3$ deliv.

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Butler-x1-x2_001-1.xls>

$\hat{y} = .0367 + .0562(85) + .7639(3) = 7.11$

Predicted values for: Time							
			95% Confidence Interval		*95% Prediction Interval*		
Km	*Deliveries*	*Predicted*	*lower*	*upper*	*lower*	*upper*	*Leverage*
85	3	7.1021	6.4074	7.7968	4.9769	9.2273	0.120

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Butler-x1-x2_001-1.xls>

http://profs.degroote.mcmaster.ca/ads/parlar/courses/q600/ChapterComments/documents/MultReg.pdf

b) Standard error

In Ch.11

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

$$s^2 = \frac{SSE}{n-2}$$

s: standard error $= \sqrt{s^2}$

In Ch.12 (now)

k var's (ind't) $x_1, x_2, \ldots, x_k$

$$s^2 = \frac{SSE}{n-(k+1)}$$

$$s = \sqrt{\frac{SSE}{n-(k+1)}}$$

↑ std error

Ex. Butler $k=2$ x_1, x_2

$n=10$, $k=2$, $SSE = 5.051$

$n-(k+1) = 10-3 = 7$ $\quad s^2 = \frac{5.051}{7} = .721$

$s = .849$

c) Significance of a variable

In Ch.11 $\quad Y = \beta_0 + \beta_1 X + \varepsilon$

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

$$s_{b_1} = \frac{s}{\sqrt{S_{xx}}}, \quad S_{xx} = \sum (x_i - \bar{x})^2$$

$$t = \frac{b_1 - 0}{s_{b_1}}$$

In Ch.12, s_{b_1}, etc. is not easy. Just use Megastat output

Ex. Butler

p-value

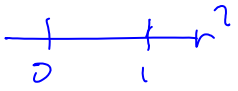
$H_0: \beta_0 = 0$, $H_a: \beta_0 \neq 0$ | $s_{b_0} = 1.326$, $t = \frac{b_0}{s_{b_0}} = \frac{.0367}{1.326} = .028$ | .9787 | Don't reject H_0

$H_0: \beta_1 = 0$, $H_a: \beta_1 \neq 0$ | $s_{b_1} = .0156$, $t = \frac{b_1}{s_{b_1}} = 3.592$ | .0088 | Reject H_0

$H_0: \beta_2 = 0$, $H_a: \beta_2 \neq 0$ | $s_{b_2} = .3053$, $t = \frac{b_2}{s_{b_2}} = 2.502$ | .0409 | Reject H_0

d) R^2 & adjusted R^2

Ch.11 Coeff. of det. $r^2 = \frac{\text{expl. var}}{\text{total var}}$



Ch.12 $R^2 = \frac{\text{expl. var.}}{\text{total var}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{18.95}{24} = .79$

Adjusted R^2 (#tea drinkers & house price)

$$\bar{R}^2 = \left(R^2 - \frac{k}{n-1}\right)\left(\frac{n-1}{n-(k+1)}\right)$$

$k = 2$, $n = 10$

$$= \left(.79 - \frac{2}{9}\right)\left(\frac{9}{7}\right)$$

$$= .73$$

d) The overall F test.

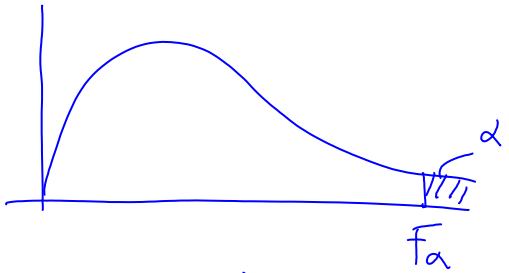
How significant is the overall model?

Ch.11 $H_0: \beta_1 = 0$, $H_a: \beta_1 \neq 0$ } t-test

Ch.12 $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_a: at least one $\neq 0$

This too needs F distribution



(df_1, df_2)

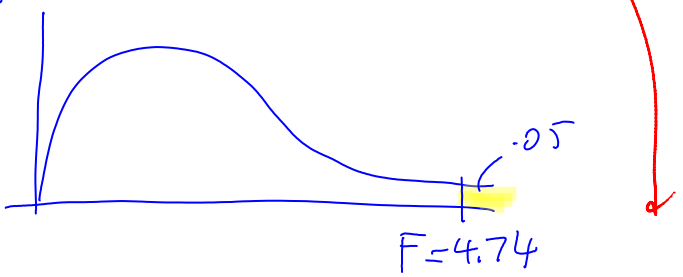
In Ch.11 $F(\text{model}) = \dfrac{\text{expl. var}/1}{(\text{unexpl. var})/(n-2)}$ $df_1 = 1$, $df_2 = n-2$

In Ch.12 $F(\text{model}) = \dfrac{(\text{expl. var})/k}{(\text{unexpl. var})/[n-(k+1)]}$, $df_1 = k$, $df_2 = n-(k+1)$

If $F(\text{model}) > F_\alpha \rightarrow$ reject H_0

Ex. Butte $F(\text{model}) = \dfrac{18.94/2}{5.05/7} = 13.13$ $n=10$, $k=2$

If $\alpha = .05$, $F^{(2,7)}_{.05} = 4.74$



reject H_0

We reject H_0: $R_1 = R_2 = 0$

We reject $H_0: \beta_1 = \beta_2 = 0$

$\Rightarrow$ The model is significant

f) Dummy variables to model qualitative indt vars

Ex. Butler

Indt vars

X_1: km
X_2: #deliveries
} quantitative vars

X_3: type of pickup truck / van } qualitative

Dep. " : y: time

小心

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

$$X_3 = \begin{cases} 0 & \text{if pickup} \\ 1 & \text{van} \end{cases}$$

$X_3 = 1$ (van) $\quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 + \varepsilon$

$X_3 = 0$ (pickup) $\quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

β_3: diff in travel time between a van & pickup

http://profs.degroote.mcmaster.ca/ads/parlar/courses/q600/ChapterComments/documents/Butler-x1-x2-Dummy.xls

	x1	x2	x3	y
	Km	Deliveries	Truck type	Time
1	100	4	1 van	9.3
2	50	3	0 truck	4.8
3	100	4	1	8.9
4	100	2	0	5.8
5	50	2	0	4.2
6	80	1	1	6.8
7	75	3	1	6.6
8	80	2	0	5.9
9	90	3	0	7.6

10	90	2	1	6.1

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Butler-x1-x2-Dummy-2.xls>

Regression Analysis						
	R^2	0.858				
	Adjusted R^2	0.787	n	10		
	R	0.926	k	3		
	Std. Error	0.753	Dep. Var.	**Time**		
ANOVA table						
Source	*SS*	*df*	*MS*	*F*	*p-value*	
Regression	20.5969	3	6.8656	12.10	.0059	
Residual	3.4031	6	0.5672			
Total	24.0000	9				
Regression output					*confidence interval*	
variables	*coefficients*	*std. error*	*t (df=6)*	*p-value*	*95% lower*	*95% upper*
Intercept	0.5222	1.2100	0.432	.6811	-2.4385	3.4829
Km	0.0464	0.0150	3.092	.0213	0.0097	0.0831
Deliveries	0.7102	0.2725	2.606	.0403	0.0433	1.3771
Truck type	0.9000	0.5281	1.704	.1393	-0.3923	2.1923

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Butler-x1-x2-Dummy-2.xls>

b_3

$\hat{y} = .522 + .0464 X_1 + .7102 X_2 + (.9) X_3$

km del. type

$b_3 = .9$: estimated diff between travel time for van & pickup is .9hr = 54 min (van takes longer)

$R^2 = .858$

	p
β_0 intercept	.68
β_1 km	.02 } reject $\beta_1 = 0$, $\beta_2 = 0$
β_2 del	.04 }
β_3 type	.13 ← type is not significant

one km & del. are taken into account

Exclude truck type

Ex. Real estate data

Four var's X: lotsize

x_2: SqrFt

x_3: bedroom

x_4: bathroom

y: price

http://profs.degroote.mcmaster.ca/ads/parlar/courses/q600/ChapterComments/documents/RealEstateData_003.xls

	y	x_1	x_2	x_3	x_4
No.	Price	LotSize	SqrFt	Bedrooms	Bathrooms
1	480.1	1.72	3985	5	4 1/2
2	397.8	0.77	3564	4	4
3	307.4	0.46	2914	3	3
4	413.0	1.56	3413	4	3 1/2
5	389.3	0.50	3627	4	2 1/2
6	353.3	1.21	3727	3	3 1/2
7	331.0	0.38	2304	4	2 1/2
8	381.2	0.55	3103	4	3 1/2
9	422.5	1.57	3859	4	5

$n = 124$

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\RealEstateData_003.xls>

Regression Analysis						
		R^2 0.931				
	Adjusted R^2	0.929	n	124		
	R	0.965	k	4		
	Std. Error	19.759	Dep. Var.	**Price**		
ANOVA table						
Source	*SS*	*df*	*MS*	*F*	*p-value*	
Regression	629,435.8198	4	157,358.9550	403.07	3.66E-68	
Residual	46,457.8989	119	390.4025			
Total	675,893.7187	123				
Regression output					*confidence interval*	
variables	*coefficients*	*std. error*	*t (df=119)*	*p-value*	*95% lower*	*95% upper*
Intercept	17.4104	9.9691	1.746	.0833	-2.3294	37.1502
LotSize	12.2789	6.1685	1.991	.0488	0.0645	24.4932
SqrFt	0.0194	0.0052	3.721	.0003	0.0091	0.0297
Bedrooms	55.0359	3.1618	17.407	1.67E-34	48.7753	61.2965
Bathrooms	20.5564	3.3757	6.089	1.43E-08	13.8721	27.2407

$R^2 = .93$

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\RealEstateData_003.xls>

$$\hat{y} = 17.41 + 12.27x_1 + 0.02x_2 + 55.03x_3 + 20.55x_4$$

$x_1 = 1$
$x_2 = 2800$
$x_3 = 6$
$x_4 = 4$

$\rightarrow \hat{y} = \$496,457$

(as opposed to $314,575 with SqFt only)