2013-12-04-C01

Wednesday, December 04, 2013
1:42 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	3
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	2
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
8	381.2	0.55	3103	4	3 1/2	1	1	18.4	6	1
9	422.5	1.57	3859	4	5	1	0	17.9	4	5
10	427.3	0.78	4318	4	5	1	1	16.1	7	5
11	380.6	0.47	2931	4	3 1/2	1	1	12.8	2	4
12	439.6	1.43	3926	4	5	0	0	20.0	5	5
13	249.8	0.30	2206	3	1 1/2	1	0	21.2	4	1
14	248.0	0.72	2401	3	2 1/2	0	1	12.1	5	3
15	337.3	0.61	3152	3	4	0	0	5.9	6	3
16	376.5	0.38	2863	4	2 1/2	0	0	19.2	7	3
17	320.4	0.57	2683	3	3	1	0	7.2	6	3

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

$x = 2800$ SqFt

$\hat{y} = 55 + 0.09(2800)$

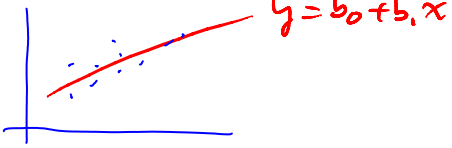
$= \$314,575$

Use more variables

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\RealEstateData_002-3.xls>

Ch.12 Multiple regression

Ch.11 x y



Ex. Marketing econ

x_1 : Your price

x_2: Competitor's price

y: demand

x_1	x_2	y
⋮	⋮	⋮

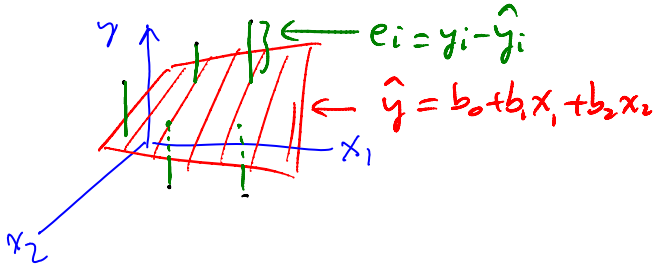
$k = 2$

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

a) The model

Ch.11 True model $Y = \beta_0 + \beta_1 X + \varepsilon$ (line)

Ch.12 " " $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ (plane)



Problem: find b_0, b_1, b_2 so that

$SSE = \sum_{i=1}^{n} e_i^2$ is minimized

MegaStat will do most of the work

Ex. R the trucking

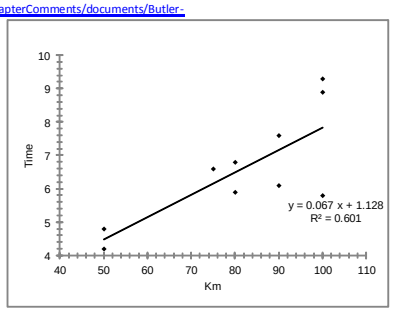
Ex. Butler trucking

Initially, travel time vs. distance

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-2.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-2.xls>

$\hat{y} = 1.1285 + 0.067x$

$x = 85$

$\hat{y} = 1.1285 + 0.067(85)$

$= 6.83$

Model may not be adequate ($r^2 = 0.6$)

A second variable 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

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Butler-x1-x2_001-2.xls>

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	0.0561	1.3267	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

b_0 b_1 b_2 s_{b_0} s_{b_1}

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_001-2.xls>

$y = 0.0561 + 0.0562 x_1 + 0.7635 x_2$

$x_1 = 85$, $x_2 = 3$

$\hat{y} = 7.1$ hn

$x_1 = x_2 = 0$

$\hat{y} = 0.036$

"multi-collinearity" is to be avoided

Correlation Matrix			
		Km	Deliveries
	Km	1.000	
	Deliveries	.280	1.000
		10	sample size

OK ✓ small corr

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Butler-x1-x2_001-2.xls>

b) Standard error

Ch.11 $s^2 = \frac{SSE}{n-2}$, $SSE = \sum (y_i - \hat{y}_i)^2$

$k=1$ $s = \sqrt{s^2}$: standard error

Now (Ch.12) k

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

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

Ex. $SSE = 5.051$, $n = 10$, $k = 2$, $n-(k+1) = 7$

$s^2 = \frac{5.051}{7} = .721$

$s = \sqrt{s^2} = .849$

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

c) Significance of a variable

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

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

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

$H_a: \beta_1 \neq 0$ | $t = \frac{b_1 - 0}{s_{b_1}}$ 

In Ch.12 s_{b_1} is not easy → Use stat

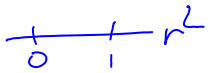
p-value

$H_0: \beta_0 = 0$, $H_a: \beta_0 \neq 0$ | $s_{b_0} = 1.326$, $t = \frac{b_0}{s_{b_0}} = \frac{.0367}{1.326} = .028$ | .9787 | Accept 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 $r^2 = \frac{\text{explained var}}{\text{total var}}$ 

Ch.12 $R^2 = \frac{\text{expl. var}}{\text{total var}} = \frac{18.95}{24} = 0.79$

Adjusted R^2 ($\bar{R}^2$)

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

$k = 2$, $n = 10$

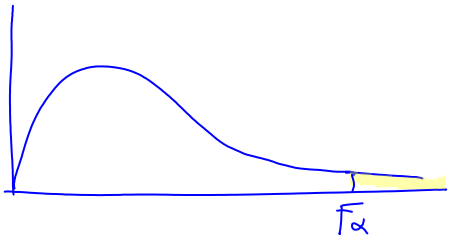
$$= \left(.79 - \frac{2}{9}\right)\left(\frac{9}{7}\right) = .73$$

e) Overall F test : significance of overall model

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

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

H_a: at least one $\neq 0$



F-statistic

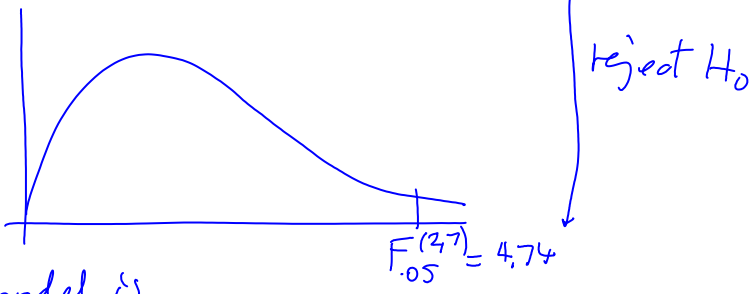
$$F(\text{model}) = \frac{(\text{explain. var})/k}{(\text{unexpl. var})/[n-(k+1)]}$$

$df_1 = k$, $df_2 = n-(k+1)$

" "

In our case, $n=10$, $k=2$

$$F(\text{model}) = \frac{18.95/2}{5.05/7} = \frac{9.474}{0.721} = 13.13$$



$\Rightarrow$ The model is significant! ✓

Multi-collinearity issue

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

x1	x2	x3	y
Km	Deliveries	Gas	Time
100	4	9.9	9.3
50	3	5.4	4.8
100	4	10.2	8.9
100	2	9.9	5.8
50	2	4.5	4.2
80	1	7.8	6.8
75	3	7.4	6.6
80	2	7.8	5.9
90	3	8.8	7.6
90	2	9.01	6.1

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

Correlation Matrix					
		Km	*Deliveries*	*Gas*	
	Km	1.000			
	Deliveries	.280	1.000		
	Gas	.992	.334	1.000	

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

	R^2	0.796					
	Adjusted R^2	0.694	n	10			
	R	0.892	k	3			
	Std. Error	0.903	Dep. Var.	**Time**			
ANOVA table							
Source	*SS*	*df*	*MS*	*F*	*p-value*		
Regression	19.1043	3	6.3681	7.80	.0171		
Residual	4.8957	6	0.8160				
Total	24.0000	9					
Regression output					*confidence interval*		
variables	*coefficients*	*std. error*	*t (df=6)*	*p-value*	*95% lower*	*95% upper*	
Intercept	-0.1582	1.4799	-0.107	.9183	-3.7793	3.4629	
Km	0.1161	0.1388	0.837	.4349	-0.2234	0.4556	
Deliveries	0.8368	0.3655	2.290	.0620	-0.0574	1.7310	
Gas	-0.6045	1.3896	-0.435	.6788	-4.0047	2.7958	

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

f) Dummy variables (qualitative)

Ex. Butler

x_1 : km
x_2 : #del. } quant

x_3 : type { van (1), pickup-truck (0) } qualit

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

$x_3 = 1$ (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

Ex.

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
100	4	1	9.3
50	3	0	4.8
100	4	1	8.9
100	2	0	5.8
50	2	0	4.2
80	1	1	6.8
75	3	1	6.6
80	2	0	5.9
90	3	0	7.6

90	2	1	6.1

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Butler-x1-x2-Dummy-3.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*
b_0 Intercept	0.5222	1.2100	0.432	.6811	-2.4385	3.4829
b_1 Km	0.0464	0.0150	3.092	.0213	0.0097	0.0831
b_2 Deliveries	0.7192	0.2725	2.606	.0403	0.0433	1.3771
b_3 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-3.xls>

$\hat{y} = 0.522 + 0.0464 x_1 + 0.71 x_2 + 0.90 x_3$

		p	
β_0:	Intercept	.6811	Don't reject $H_0: \beta_0 = 0$
β_1:	km	.02	reject $H_0: \beta_1 = 0$
β_2:	del	.04	reject $H_0: \beta_2 = 0$
β_3:	type	.13	Don't reject $\beta_3 = 0$

	Km	*Deliveries*	*Truck type*
Km	1.000		
Deliveries	.280	1.000	
Truck type	.419	.218	1.000

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Butler-x1-x2-Dummy-3.xls>

Ex. Real estate data

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

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

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

model is significant

	SqrFt	0.0194	0.0052	3.721	.0003	0.0091	0.0297
x_1	Bedrooms	55.0359	3.1618	17.407	1.67E-34	48.7753	61.2965
x_2	Bathrooms	20.5564	3.3757	6.089	1.43E-08	13.8721	27.2407
x_3							
x_4							

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

$\hat{y} = 17.41 + 12.27x_1 + 0.019x_2 + 55.03x_3 + 20.55x_4$

$x_1 = 1, \quad x_2 = 2800, \quad x_3 = 6, \quad x_4 = 4$

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

Q: Is there a multi-collinearity? Check!