2013-12-05-C02

Thursday, December 05, 2013
11:12 AM

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
8	381.2	0.55	3103	4	3 1/2	1	1	18.4	6	4
9	422.5	1.57	3859	4	5	1	0	17.9	4	5

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

$\hat{y} = 55 + .09x$

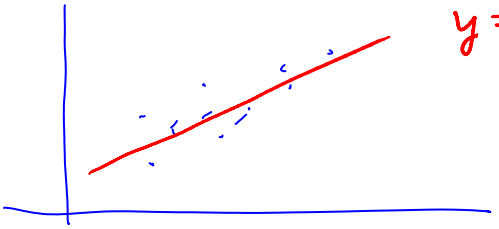
$x = 2,500$

$\hat{y} = \$314,575$

$r^2 = .65$

Ch.12 Multiple regression

Ch.11 x y



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

Ex. x_1: lot size
x_2: sqrft
y: price

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

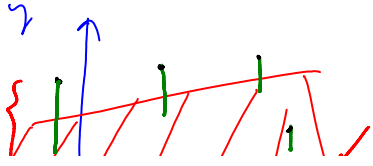
a) The model

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

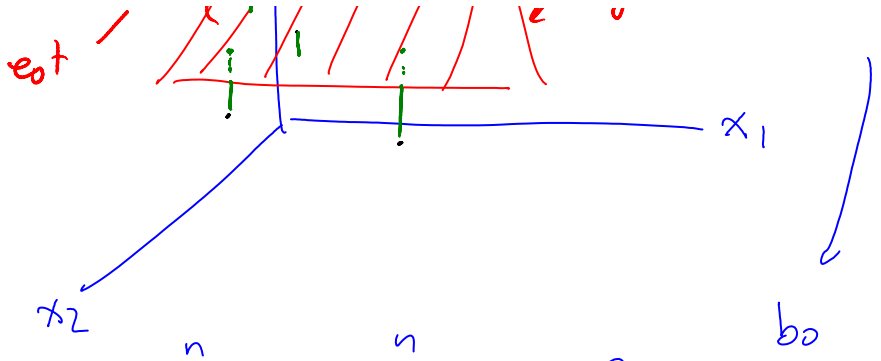
Ch.12 " " $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ multiple

obs

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



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



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

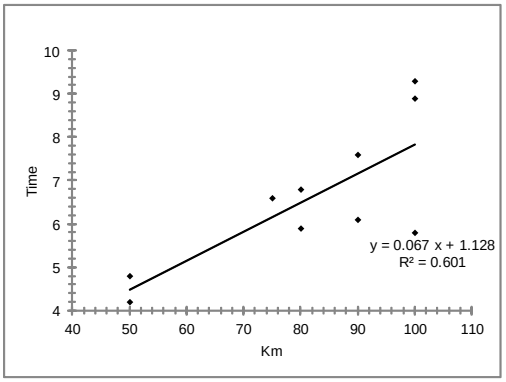
b_0
b_1
b_2

Ex. Butler trucking

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*
b_0 Intercept	1.1285	1.6123	0.700	.5038	-2.5896	4.8466
b_1 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>

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

$x = 85$

$\hat{y} = 6.83$ hr

Model may not be adequate!

Consider now,

x_2 : # deliveries

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

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

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

Pasted from <file:///C:\DOCUME~1\parlar\LOCALS~1\Temp\Butler-x1-x2_001.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	0.0367	1.3262	0.028	.9787	-3.0994	3.1727	
Deliveries	0.7639	0.3053	2.502	.0409	0.0419	1.4858	
Km	0.0562	0.0156	3.592	.0088	0.0192	0.0931	
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.xls>

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

$x_1 = 85, \quad x_2 = 3$

$\hat{y} = 7.10$ hrs

"Multi-collineanity" is to be avoided

OK

Correlation Matrix			
		Km	*Deliveries*
	Km	1.000	
	Deliveries	.280	1.000

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

b) Standard error

Ch.11 $s^2 = \frac{SSE}{n-2}$, $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

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

Ch.12 $s^2 = \frac{SSE}{n-(k+1)}$

k: # ind't vars

$s = \sqrt{s^2}$

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

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

$s = \sqrt{.721} = .849$

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

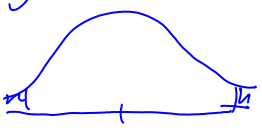
c) Significance of a variable

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

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

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

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



Ch.12

$H_0: \beta_0 = 0$
$H_a: \beta_0 \neq 0$

$s_{b_0} = 1.326$, $t = \frac{b_0}{s_{b_0}} = \frac{.0267}{1.326} = .028$

p-value .9787 "Accept" H_0

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

$s_{b_1} = .0156$, $t = \frac{b_1}{s_{b_1}} =$ $= 3.592$ $.0088$ β_{est} H_0

$H_0: \beta_2 = 0$
$H_a: \beta_2 \neq 0$

$s_{b_2} = .3053$, $t = \frac{b_2}{s_{b_2}} =$ $= 2.502$ $.0409$ β_{est} H_0 $(\alpha = .05)$

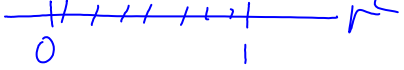
$\hat{y} = .0367 + .0562 x_1 + .7639 x_2$

$x_1 = 0$ km
$x_2 = 0$ del

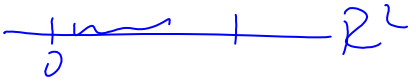
$\hat{y} = .0367 \simeq 0$

d) R^2 & adjusted R^2

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



Ch. 12 $R^2 = \frac{\text{explain. var}}{\text{total var}}$



Ex. $R^2 = \frac{18.95}{24} = .79$

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

$$\bar{R}^2 = \left(R^2 - \frac{k}{n-1}\right)\left(\frac{n-1}{n-(k+1)}\right), \quad k=2, \; n=10$$

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

$$= .73$$

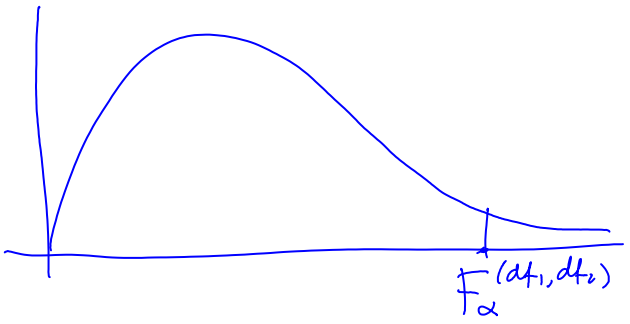
e) The overall F test (Significance of

overall model)

Ch.11 $H_0: \beta_1 = 0$ | t-test
$H_a: \beta_1 \neq 0$ | F-test → $F(\text{model}) = \dfrac{\text{expl.var}/1}{(\text{unexp. var})/(n-2)}$

Ch.12 $H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$
H_a: at least one $\neq 0$

$$F(\text{model}) = \frac{(\text{expl.var})/k}{(\text{unexp. var})/(n-(k+1))}, \quad df_1 = k, \quad df_2 = (n-(k+1))$$

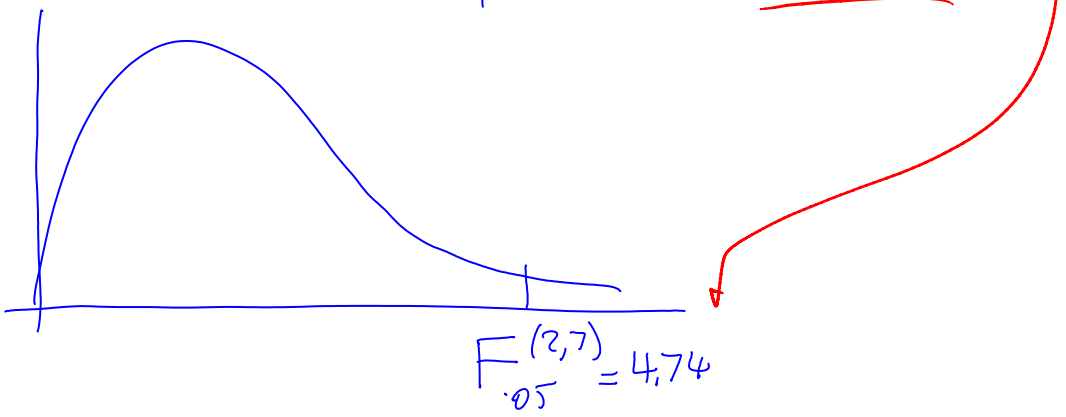


Ex. $k=2$, $n=10$

$$F(\text{model}) = \frac{18.94/2}{5.05/7} = \frac{9.47}{.72} = 13.13$$

$p = .0043$ reject H_0

$\alpha = .05$



Ex. Multi-collinearity problem

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>

Regression Analysis						
	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 for qualitative data

Ex. Butler

X_1: km
X_2: deliv
} quantitative

X_3: type of truck (van 1 / pickup 0) } qualit.

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

$\beta_3 = ?$

$X_3 = 1$ (van)

$X_3 = 0$ (pickup)

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

$- \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

β_3 : difference in travel time between a van and a 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
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.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	b_0 0.5222	1.2100	0.432	.6811	-2.4385	3.4829
Km	b_1 0.0464	0.0150	3.092	.0213	0.0097	0.0831
Deliveries	b_2 0.7102	0.2725	2.606	.0403	0.0433	1.3771
Truck type	b_3 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.xls>

Ex. Real estate data

	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

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
Bathrooms	20.5564	3.3757	6.089	1.43E-08	13.8721	27.2407
Bedrooms	55.0359	3.1618	17.407	1.67E-34	48.7753	61.2965

b_0 b_1 b_2 b_3 b_4

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

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

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

$\Rightarrow \quad \hat{y} = \$496,457$ as opposed to $314 000 with sqft = 2000 only