



Pearson

Chapter 16: Multiple regression model building

After studying this chapter you should be able to:

1. utilise quadratic terms in a regression model
2. calculate and use transformed variables in a regression model
3. examine the effect of each observation on the regression model
4. construct a regression model using either the stepwise or the best-subsets approach
5. recognise the many pitfalls involved in developing a multiple regression model

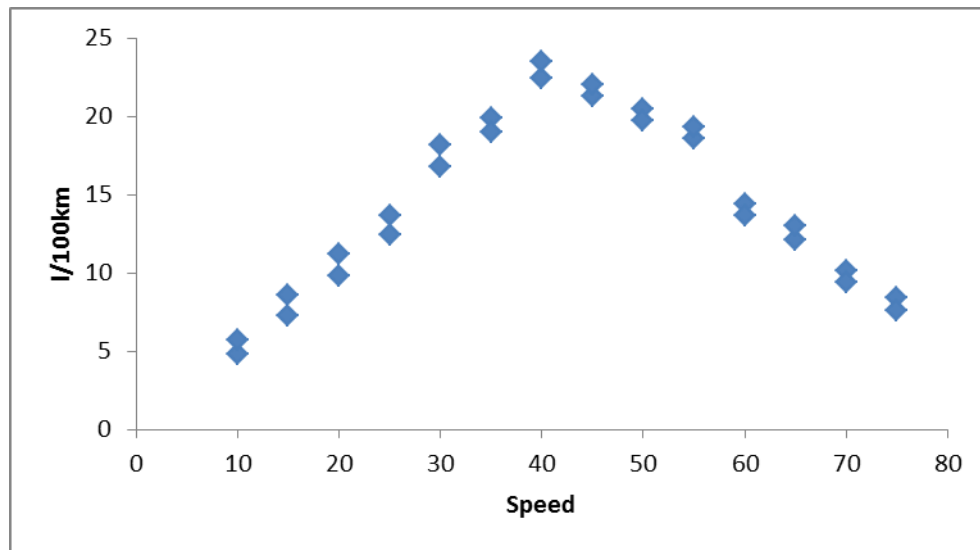
16.1 (a) $\hat{Y} = 10 + 2X_{1i} - 0.5X_{1i}^2 = 10 + 2(4) - 0.5(4)^2 = 10$

(b) $t_{calc} = 2.35 > t_{27} = 2.0518$ with 27 degrees of freedom. Reject H_0 . The quadratic term is significant.

(c) $t_{calc} = 1.17 < t_{27} = 2.0518$ with 27 degrees of freedom. Do not reject H_0 . The quadratic term is not significant.

(d) $\hat{Y} = 10 - 3X - 0.5X^2 = 10 - 3(2) - 0.5(2)^2 = 2$

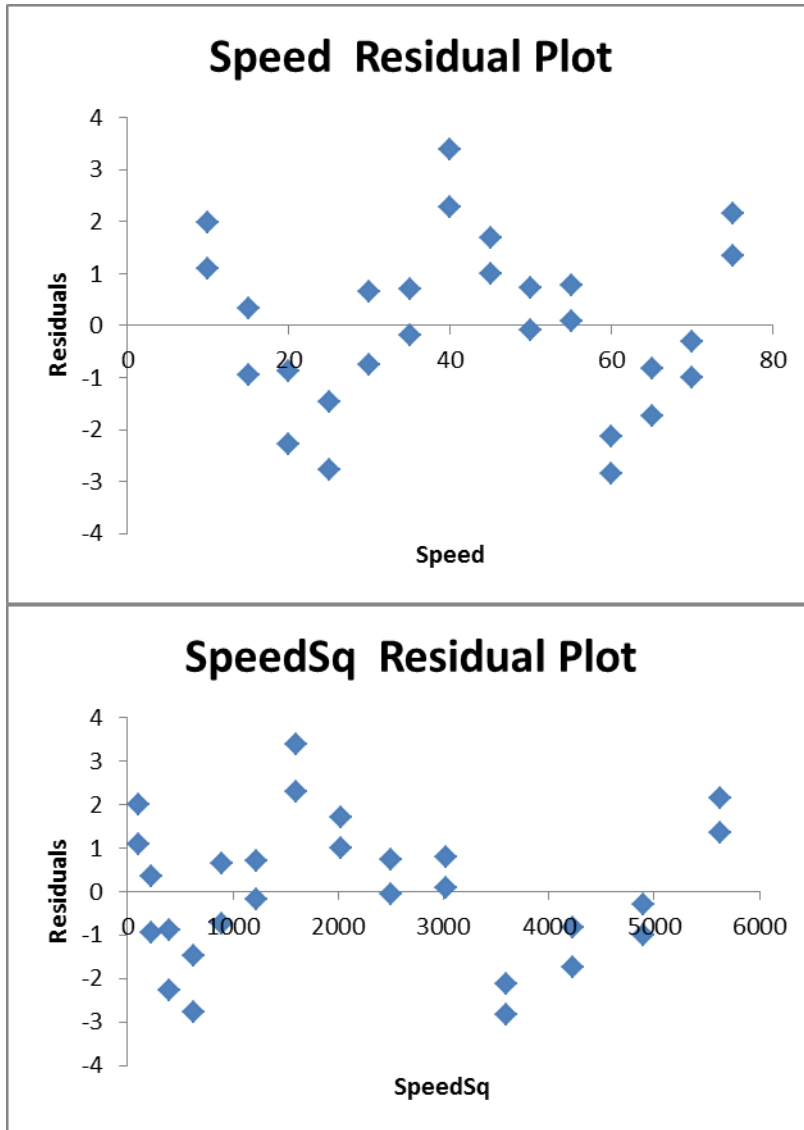
16.2 (a)



(b) $\hat{Y} = -7.555 + 1.2717X - 0.0145X^2$

(c) $\hat{Y} = -7.555 + 1.2717(90) - 0.0145(90)^2 = 224.358$

(d)



A residual analysis indicates no strong pattern.

- (e) $H_0 = \beta_1 = \beta_2 = 0$
 $H_1 = \text{at least one } \beta_j \neq 0$
 $F = 141.4596 > F_{2,25} = 3.39$. Reject H_0 . There is a significant quadratic relationship between litres per 100 kilometres and speed.
- (f) $H_0 = \text{including the quadratic effect does not significantly improve the model } (\beta_2 = 0)$.
 $H_1 = \text{including the quadratic effect significantly improves the model } (\beta_2 \neq 0)$.
 $t_{calc} = -16.6325 > t_{25} = 2.0595$. Reject H_0 . The quadratic effect is significant. Therefore, the quadratic model is a better fit than the linear regression model.

(g)

The coefficient of multiple determination R^2 represents the proportion of variation in Y that is explained by the variation in the independent variables. Therefore, 91.88% of the variation in litres per 100 kilometres is explained by the quadratic relationship between litres per 100 kilometres and speed.

$$(h) \quad R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n - 1)}{(n - k - 1)} \right]$$

$$R^2_{adj} = 1 - \left[\frac{(1 - 0.9188)(27 - 1)}{(27 - 2 - 1)} \right]$$

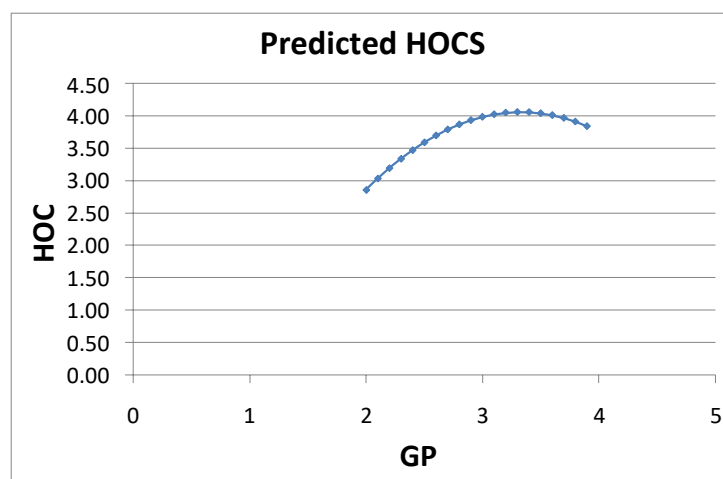
$$R^2_{adj} = 1 - 0.088 = 0.912$$

16.3

(a)

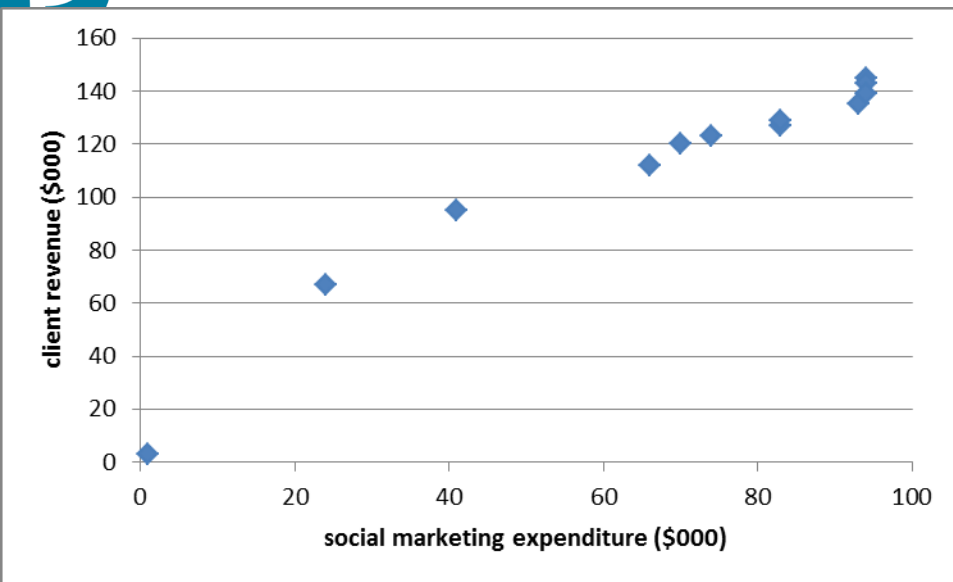
GPA	Predicted HOCS	GPA	Predicted HOCS
2	2.8600	3	3.9900
2.1	3.0342	3.1	4.0282
2.2	3.1948	3.2	4.0528
2.3	3.3418	3.3	4.0638
2.4	3.4752	3.4	4.0612
2.5	3.5950	3.5	4.0450
2.6	3.7012	3.6	4.0152
2.7	3.7938	3.7	3.9718
2.8	3.8728	3.8	3.9148
2.9	3.9382	3.9	3.8442
		4	3.7600

(b)



- (c) The curvilinear relationship suggests that HOCS increases at a decreasing rate. It reaches its maximum value of 4.0638 at GPA = 3.3 and declines after that as GPA continues to increase.
- (d) An r^2 of 0.07 and an adjusted r^2 of 0.06 tell you that GPA has very low explanatory power in explaining the variation in HOCS. You can tell that the individual HOCS scores will have scattered quite widely around the curvilinear relationship plotted in (b) and discussed in (c).

16.4 (a)

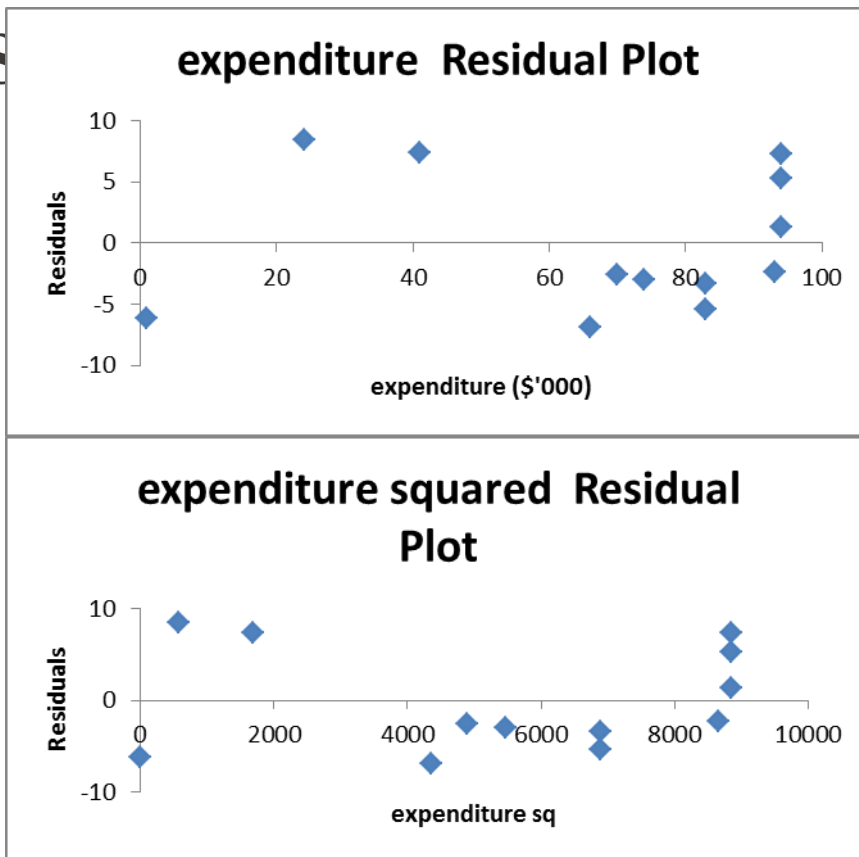


(b) $\hat{Y} = 6.7380 + 2.4198X - 0.0109X^2$

(c) $\hat{Y} = 6.7380 + 2.4198(75) - 0.0109(5625) = 126.8039 = \126803.90



Pearson



A residual analysis indicates no strong pattern.

- (e) $H_0 = \beta_1 = \beta_2 = 0$
 $H_1 = \text{at least one } \beta_j \neq 0$
 $F = 225.6689 > F_{2,9} = 4.26$. Reject H_0 . There is a significant quadratic relationship between client revenue and social marketing expenditure.
- (f) The p -value = virtually zero. It indicates that the probability of having an F test statistic of at least 225.6689 when β_1 and $\beta_2 = 0$ is virtually zero.
- (g) $H_0 = \text{including the quadratic effect does not significantly improve the model } (\beta_2 = 0)$.
 $H_1 = \text{including the quadratic effect significantly improves the model } (\beta_2 \neq 0)$.
 $t_{calc} = -4.5303 < t_9 = 2.2622$. Reject H_0 . The quadratic effect is significant and therefore the quadratic model is a better fit than the linear regression model.
- (h) The p -value = 0.0014. It indicates that the probability of having a t test statistic of less than -4.5303 when $\beta_2 = 0$ is 0.0014.
- (i) The coefficient of multiple determination R^2 represents the proportion of variation in Y that is explained by the variation in the independent variables. Therefore, 98.04% of the variation in client revenue is explained by the quadratic relationship between client revenue and social marketing expenditure.



arson

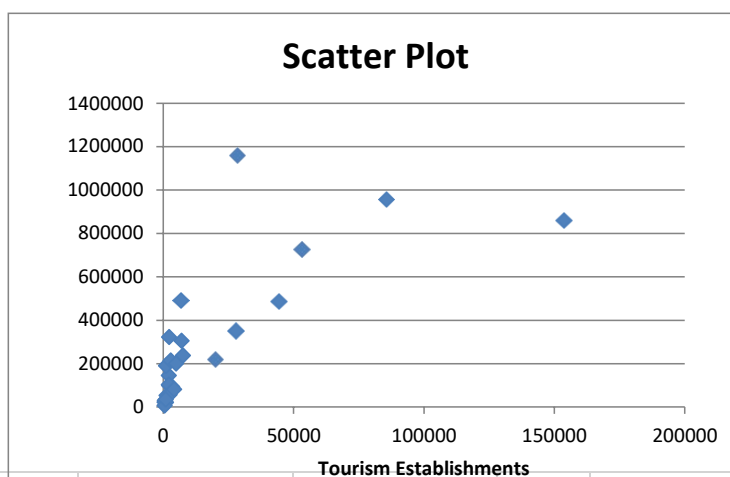
$$(j) \quad R_{adj}^2 = 1 - \left[(1 - R^2) \frac{(n-1)}{(n-k-1)} \right]$$

$$R_{adj}^2 = 1 - \left[(1 - 0.9804) \frac{(12-1)}{(12-2-1)} \right] = 0.9761$$

16.5

(a)

(b)



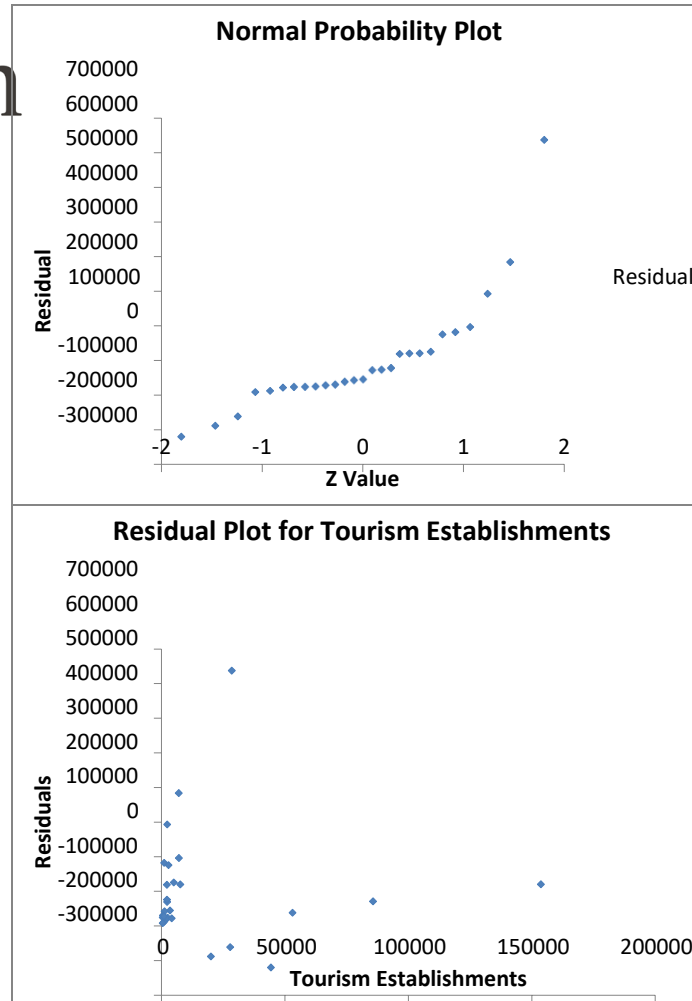
Regression Statistics		Tourism Establishments				
Multiple R	0.8459					
R Square	0.7155					
Adjusted R Square	0.6918					
Standard Error	173172.0845					
Observations	27					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	1809871444435.5500	904935722217.7750	30.1760	0.0000	
Residual	24	719725700749.6360	29988570864.5682			
Total	26	2529597145185.1900				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	87251.7358	41921.1371	2.0813	0.0482	730.7611	173772.7104
Tourism Establishments	17.4863	2.9851	5.8578	0.0000	11.3254	23.6473
Tourism Establishments Sq	-0.0001	0.0000	-3.8046	0.0009	-0.0001	0.0000

$$\hat{Y} = 87251.7358 + 17.4863 X - 0.0001 X^2$$

$$(c) \quad \hat{Y} = 87251.7358 + 17.4863(3000) - 0.0001(3000)^2 = \mathbf{138972.3308}$$



Pearson



A residual analysis reveals potential violation of the equal variance assumption and the normality assumption.

- (e) $F_{STAT} = 30.1760$ with a p -value of essentially 0. Reject H_0 . The overall model is significant.
- (f) The p -value = 0.0000 indicates that the probability of having an F -test statistic of at least 30.1760 when $\beta_1 = 0$ and $\beta_2 = 0$ is essentially 0.
- (g) $t_{STAT} = -3.8046$ with a p -value = 0.0009. Reject H_0 . The quadratic effect is significant.
- (h) The p -value = 0.0009 indicates that the probability of having a t -test statistic with an absolute value of at least 3.8046 when $\beta_2 = 0$ is 0.0009.
- (i) $r_{Y,12}^2 = 0.7155$. So, 71.55% of the variation in taxes can be explained by the quadratic relationship between the number of jobs generated in the travel and tourism industry in 2012 and the number of establishments.



(j)

$$r_{adj}^2 = 0.6918 \cdot r^2$$

arson



(k) There is a significant quadratic relationship between the number of jobs generated in the travel and tourism industry in 2012 and the number of establishments that provide overnight accommodation for tourists.

Pearson

16.6

- (a) $\hat{Y} = 9.0356 + 0.85204\sqrt{X_1}$
- (b) $\hat{Y} = 9.0356 + 0.85204\sqrt{85} = 16.89$
- (c) The residual analysis indicates a clear quadratic term. The model does not adequately fit the data.
- (d) $t = 1.3537 > t_{26} = 2.055$. Fail to reject H_0 . The model does not provide a significant relationship.
- (e) $R^2 = 0.0658$. So, 6.58% of the variation on litres per 100 kilometres can be explained by variation in the square root of highway speed.
- (f) $R^2_{adj} = 0.0299$
- (g) The quadratic regression model in 16.2 is superior to this model. R^2_{adj} is far less than the model for 16.2.

16.7

- (a) $\ln \hat{Y} = 2.3882 + 0.0045X_1$
- (b) $\ln \hat{Y} = 1.9708 + 0.0040(85) = 2.77$ $\hat{Y} = e^{2.77} = 15.97/100$ km
- (c) The residual analysis indicates a clear quadratic pattern. The model does not adequately fit the data.
- (d) $F = 1.218 > F_{1,26} = 4.23$. Fail to reject H_0 . The model does not provide a significant relationship.
- (e) $R^2 = 0.04475$. 4.475% of the variation in litres per 100 kilometres can be explained by variation in the natural logarithm of speed.
- (f) $R^2_{adj} = 0.008$.
- (g) The quadratic regression model in 16.2 is superior to 16.6 and this model. $R^2_{adj} = 0.008$ is less than both the other models. The model in 16.2 with the highest R^2_{adj} is the most appropriate model to use.

16.8

- (a) $\ln \hat{Y} = 11.413 + 0.0000234X$
- (b) $\ln \hat{Y} = 11.413 + 0.0000234(5000) = 11.53$ $\hat{Y} = e^{11.53} = 101794.52$
- (c) The residual analysis does not indicate any clear patterns.
- (d) $t = 3.6164 > t_{25} = 2.0595$. Reject H_0 . The model provides a significant relationship.
- (e) $R^2 = 0.3435$. 34.35% of the variation of the log of tourism employment is explained by the number of tourism establishments.
- (f) $R^2_{adj} = 0.3172$.
- (g) We cannot directly compare R^2 values as there is a different dependent variable.

16.9

- (a) $\hat{Y} = 6390.1439 + 2814.4438\sqrt{X}$
- (b) $\hat{Y} = -10.3422 + 15.4973\sqrt{5000} = 205401.37$
- (c) The residual analysis does not indicate clear patterns.
- (d) $t = 7.556 > t_{25} = 2.0595$. Reject H_0 . The model provides a significant relationship.
- (e) $R^2 = 0.6955$. So, 69.55% of the variation in tourism employment can be explained by variation in the square root of the number of tourism establishments.
- (f) $R^2_{adj} = 0.6833$
- (g) This model here is the best, based upon R^2_{adj} and residual analysis.



16.10

<i>Observation</i>	<i>Predicted Emission</i>	<i>Residuals</i>	<i>Standard Residuals</i>	<i>h_i</i>	<i>Stud_t</i>	<i>Cook_d</i>	<i>t test</i>	<i>Hat test</i>	<i>Cook's test</i>
	18.0278	1.972205	0.381265	0.247177	0.40966	0.027171	No	No	No
	23.72864	10.27136	1.985652	0.114677	2.366037	0.247231	Yes	No	No
	21.94137	0.058628	0.011334	0.177088	0.011567	1.44E-05	No	No	No
	19.29784	-3.29784	-0.63754	0.097828	-0.63167	0.020937	No	No	No
	22.2517	9.748295	1.884533	0.088544	2.151245	0.162225	Yes	No	No
	20.67133	-3.67133	-0.70974	0.075133	-0.69695	0.018962	No	No	No
	21.83793	-1.83793	-0.35531	0.308384	-0.39815	0.034881	No	No	No
	23.52175	-5.52175	-1.06746	0.09565	-1.0894	0.057114	No	No	No
	19.40128	-0.40128	-0.07758	0.142931	-0.0776	0.000502	No	No	No
	17.51057	6.489426	1.254531	0.344031	1.575389	0.539284	No	No	No
	26.37217	-1.37217	-0.26527	0.194796	-0.27455	0.009061	No	No	No
	33.86028	0.139721	0.027011	0.662903	0.043075	0.001824	No	Yes	No
	16.44742	-2.44742	-0.47313	0.207118	-0.49697	0.031607	No	No	No
	20.77477	-2.77477	-0.53642	0.082927	-0.5245	0.01216	No	No	No
	22.35515	-7.35515	-1.42189	0.160812	-1.57932	0.197858	No	No	No

Since none of the observations are flagged by all three diagnostic tests, there is no strong evidence that any of them are influential.

16.11

<i>Observation</i>	<i>Predicted emission</i>	<i>Residual</i>	<i>hi</i>	<i>Stud_t</i>	<i>Cook_d</i>	<i>Hat test</i>	<i>t test</i>	<i>Cook's test</i>
	1840.275412	2,425.76	0.181384	8.588	1.060474428	no	yes	no
	3355.265466	-877.24	0.762805	-2.67	6.947020552	yes	yes	yes
	561.3083952	438.87	0.141735	0.554	0.024675892	no	no	no
	472.7739058	490.71	0.194705	0.643	0.048136377	no	no	no
	1315.87294	-754.66	0.225743	-1.04	0.142795886	no	no	no
	808.685565	-458.18	0.097207	-0.56	0.016670476	no	no	no
	304.3976324	34.30	0.436311	0.053	0.001075796	yes	no	no
	462.0410228	-220.34	0.155167	-0.28	0.007027675	no	no	no
	295.0947749	-53.76	0.164843	-0.07	0.000454879	no	no	no
	224.1418807	4.91	0.153256	0.006	3.42854E-06	no	no	no
	610.6208019	-413.92	0.105474	-0.51	0.015036999	no	no	no
	384.3232706	-193.56	0.122784	-0.24	0.003980501	no	no	no
	230.8918401	-65.22	0.109291	-0.08	0.000390177	no	no	no
	523.2870923	-357.67	0.149297	-0.45	0.017571869	no	no	no

No observation is flagged by all tests as being influential.

Analysis re-performed without influential data. Since none of the observations are flagged by all three diagnostic tests, there is no strong evidence that any of the remaining data are influential.



Pearson

Observation	Predicted Quality	Residual	h	student t	Cook D	t test	hat test	Cook test
1	5.525891816	0.474108	0.030716	0.546126	0.004416	no	no	no
2	6.398411466	-0.39841	0.043837	-0.46165	0.004455	no	no	no
3	4.893321595	0.106678	0.062633	0.124584	0.000457	no	no	no
4	4.988773819	0.011226	0.03463	0.012917	2.79E-06	no	no	no
5	6.759794591	1.240205	0.092437	1.506841	0.091636	no	no	no
6	5.231061475	0.768939	0.034376	0.892074	0.013003	no	no	no
7	5.0465727	0.953427	0.034007	1.110955	0.019776	no	no	no
8	5.170344626	-0.17034	0.043049	-0.19694	0.0008	no	no	no
9	5.73548282	0.264517	0.02299	0.302827	0.001028	no	no	no
10	5.082187321	-0.08219	0.044375	-0.09505	0.000192	no	no	no
11	5.407955826	0.592044	0.02755	0.682066	0.006175	no	no	no
12	5.656958661	-0.65696	0.022129	-0.75559	0.006106	no	no	no
13	5.844365403	0.155635	0.023494	0.178108	0.000364	no	no	no
14	5.560627192	0.439373	0.020114	0.503129	0.002482	no	no	no
15	5.090641717	0.909358	0.18272	1.153121	0.099887	no	yes	no
16	5.736941803	-0.73694	0.021154	-0.84849	0.007345	no	no	no
17	5.743656983	0.256343	0.026906	0.294043	0.001131	no	no	no
18	5.208877214	-0.20888	0.045917	-0.2419	0.001283	no	no	no
19	5.988862606	-0.98886	0.034022	-1.1534	0.021283	no	no	no
20	5.677683938	0.322316	0.020884	0.368773	0.001387	no	no	no
21	6.187361479	-1.18736	0.036323	-1.3954	0.032766	no	no	no
22	5.016214276	-0.01621	0.037124	-0.01868	6.25E-06	no	no	no
23	5.97251428	0.027486	0.026087	0.031486	1.26E-05	no	no	no
24	5.012417064	0.987583	0.034452	1.152131	0.021497	no	no	no
25	6.927935039	1.072065	0.112993	1.310127	0.084058	no	no	no
26	6.668719318	0.331281	0.073121	0.38963	0.005155	no	no	no
27	5.0465727	0.953427	0.034007	1.110955	0.019776	no	no	no
28	5.182895741	-0.1829	0.032502	-0.21031	0.000695	no	no	no
29	5.989442345	1.010558	0.03509	1.180119	0.022927	no	no	no
30	4.905872709	1.094127	0.049036	1.290673	0.037601	no	no	no
31	7.480522118	-1.48052	0.189116	-1.93099	0.274713	no	yes	no
32	6.10883732	-0.10884	0.032774	-0.12513	0.000248	no	no	no
33	5.611131155	0.388869	0.046464	0.45117	0.0045	no	no	no
34	5.738400787	1.261599	0.020728	1.474187	0.021091	no	no	no
35	6.17393112	-0.17393	0.034934	-0.20024	0.000676	no	no	no
36	5.219969344	-1.21997	0.039621	-1.43795	0.037741	no	no	no
37	5.221428328	-1.22143	0.033932	-1.43531	0.032385	no	no	no
38	5.368843499	-0.36884	0.03117	-0.42444	0.002712	no	No	no
39	6.256252491	-1.25625	0.039418	-1.48252	0.039813	no	No	no
40	5.116922696	-0.11692	0.029223	-0.13418	0.000256	no	No	no



arson

41	6.087232798	-0.08723	0.036257	-0.10046	0.000177	no	No	no
42	4.675256923	-1.67526	0.081215	-2.06334	0.146614	yes	No	no
43	4.655990629	-0.65599	0.085169	-0.78035	0.023591	no	No	no
44	5.016214276	-0.01621	0.037124	-0.01868	6.25E-06	no	No	no
45	6.17393112	-0.17393	0.034934	-0.20024	0.000676	no	No	no
46	5.97251428	-1.97251	0.026087	-2.39331	0.064902	yes	No	no
47	5.53406598	2.465934	0.037801	3.122916	0.147094	yes	No	no
48	3.450388285	-0.45039	0.573945	-0.78515	0.110941	no	No	no
49	7.682818203	0.317182	0.193676	0.399999	0.012936	no	Yes	no
50	4.796990126	-0.79699	0.057739	-0.93681	0.023514	no	No	no

No observation is flagged by all tests as being influential.

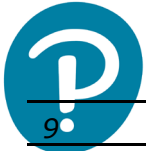
13

Observation	<i>Hi</i>	<i>Stud_t</i>	<i>Cook_d</i>	<i>Hat test</i>	<i>t test</i>	<i>Cook's test</i>
	0.39175	3.503247	1.774455	no	yes	no
	0.344334	-0.33097	0.028452	no	no	no
	0.114163	0.620158	0.023865	no	no	no
	0.399828	-0.63806	0.130305	no	no	no
	0.10992	0.651784	0.025163	no	no	no
	0.133582	-0.103	0.000817	no	no	no
	0.389995	2.75405	1.378803	no	yes	no
	0.281909	-0.23319	0.010616	no	no	no
	0.368244	-1.25789	0.398152	no	no	no
	0.087988	-0.38193	0.006935	no	no	no
	0.105201	-0.40399	0.00944	no	no	no
	0.108234	-1.12388	0.068056	no	no	no
	0.164938	-1.05384	0.098716	no	no	no

Since none of the observations are flagged by all three diagnostic tests, there is no strong evidence that any of them are influential.

16.14

Observation	<i>Predicted Sales</i>	<i>Sales</i>	<i>Residual</i>	<i>h</i>	<i>studt</i>	<i>cook</i>	<i>hat test</i>	<i>t test</i>	<i>cook test</i>
1	828.2415577	973	144.7584	0.292389	1.118013	0.13	no	no	no
2	828.2415577	1119	290.7584	0.292389	2.510081	0.54	yes	yes	no
3	903.3294104	875	-28.3294	0.098132	-0.1879	0.00	no	no	no
4	903.3294104	625	-278.329	0.098132	-2.0356	0.15	no	no	no
5	1052.709206	910	-142.709	0.066755	-0.95154	0.03	no	no	no
6	1052.709206	971	-81.7092	0.066755	-0.53629	0.01	no	no	no
7	1202.089001	931	-271.089	0.060799	-1.92424	0.09	no	no	no
8	1202.089001	1177	-25.089	0.060799	-0.16303	0.00	no	no	no



Pearson

9	1099.539625	882	-217.54	0.055995	-1.48895	0.05	no	no	n
10	1099.539625	982	-117.54	0.055995	-0.7732	0.02	no	no	n
11	1500.848591	1628	127.1514	0.125148	0.872466	0.04	no	no	n
12	1500.848591	1577	76.15141	0.125148	0.515937	0.01	no	no	n
13	810.4644823	1044	233.5355	0.356373	2.018837	0.44	yes	no	n
14	810.4644823	914	103.5355	0.356373	0.826626	0.09	no	no	n
15	1295.749839	1329	33.25016	0.069023	0.217134	0.00	no	no	n
16	1295.749839	1330	34.25016	0.069023	0.223681	0.00	no	no	n
17	1445.129634	1405	-40.1296	0.080129	-0.26379	0.00	no	no	n
18	1445.129634	1436	-9.12963	0.080129	-0.05991	0.00	no	no	n
19	1594.509429	1521	-73.5094	0.116655	-0.49537	0.01	no	no	n
20	1594.509429	1741	146.4906	0.116655	1.006684	0.05	no	no	n
21	1743.889225	1866	122.1108	0.178602	0.864406	0.05	no	no	n
22	1743.889225	1717	-26.8892	0.178602	-0.18688	0.00	no	no	n

Since none of the observations are flagged by all three diagnostic tests, there is no strong evidence that any of them are influential.

- 16.15 (a) For the model that includes independent variables A and B , the value of C_p exceeds 3, the number of parameters, so this model does not meet the criterion for further consideration.
For the model that includes independent variables A and C , the value of C_p is less than or equal to 3, the number of parameters, so this model does meet the criterion for further consideration.
For the model that includes independent variables A , B and C , the value of C_p is less than or equal to 4, the number of parameters, so this model does meet the criterion for further consideration.
- (b) The inclusion of variable C in the model appears to improve the model's ability to explain variation in the dependent variable sufficiently to justify its inclusion in a model that contains only variables A and B .
- 16.16 (a)
$$C_p = \frac{(1 - R_k^2)(n - T)}{1 - R_T^2} - [n - 2(k + 1)] = \frac{(1 - 0.274)(26 - 7)}{1 - 0.623} - [26 - 2(2 + 1)]$$

$$= 16.59$$
- (b) C_p overwhelmingly exceeds $k + 1 = 3$, the number of parameters (including the Y -intercept), so this model does not meet the criterion for further consideration as the best model.
- 16.17 (a) Let Y = sales, X_1 = Wonderlic personnel test score, X_2 = Strong-Campbell interest inventory test score, X_3 = experience, X_4 = 1 with a degree in electrical engineering; 0 otherwise.
Based on a full regression model involving all of the variables:
All VIF s are less than 5. So there is no reason to suspect collinearity between any pair of variables.
The best-subset approach yields the following models to be considered:



arson

Partial output from the best-subsets selection:

Model	C_p	k	R^2	R^2_{adj}	Std. Error	Consider This Model?
X1X2X3X4	5	5	0.593228	0.552551222	11.74203	Yes
X1X2X4	3.101155	4	0.5922	0.562360664	11.6126	Yes
X2X3X4	3.001097	4	0.593217	0.563452639	11.59811	Yes
X2X4	1.101172	3	0.5922	0.572780469	11.47353	Yes

Partial PHStat2 output of the full regression model:

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	25.7683	13.9537	1.8467	0.0722
Wonder	-0.0134	0.4050	-0.0331	0.9737
SC	1.3514	0.1947	6.9407	0.0000
Experience	0.1682	0.5287	0.3180	0.7521
Engineer Dummy	7.2747	4.1011	1.7738	0.0837

Since the p -value for X_1 and X_3 are larger than 0.05, they do not have significant effect individually on sales. The best model should include both X_2 and X_4 .

PHStat2 output of the model with only X_2 and X_4 :

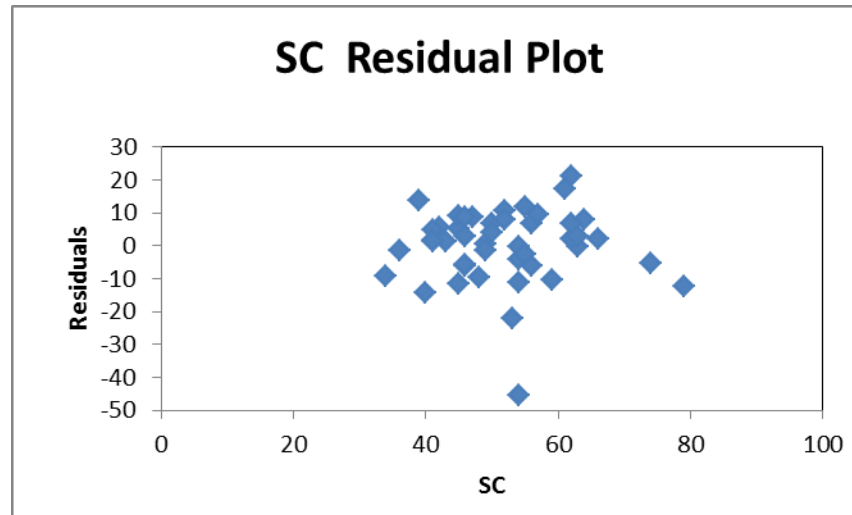
<i>Regression Statistics</i>	
Multiple R	0.7695
R^2	0.5922
R^2_{adj}	0.5728
Standard Error	11.4735
Observations	45

ANOVA					
	df	SS	MS	F	Significance F
Regression	2	8029.0413	4014.5207	30.4958	6.59784E-09
Residual	42	5528.9587	131.6419		
Total	44	13558			
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t_{calc}</i>	<i>P-value</i>	
Intercept	26.8910	9.7718	2.7519	0.0087	
SC	1.3408	0.1792	7.4824	0.0000	
Engineer Dummy	7.2869	3.9857	1.8282	0.0746	

The most appropriate model to predict sales is:

$$\hat{Y} = 26.8910 + 1.3408 X_2 + 7.2869 X_4$$

(b) With the exception of one single residual point at a value of -44.58 when $SC = 54$, there is no specific pattern in the residual plot. Influential analysis using the hat matrix elements, the deleted residuals and Cook's distance statistic does not reveal any observation that needs to be deleted.



According to the finding in (a), the company only needs to administer the Strong-Campbell test.

- (c) According to the model in (a), the variable X_4 helps predict sales and, hence, the idea of only hiring electrical engineers should be supported.
- (d) Prior selling experience (X_3) does not help predict sales according to the model chosen in (a).
- (e) The company only needs to administer the Strong-Campbell test to save time and money. It should hire only sales managers with an electrical engineering degree.

16.18

Let Y = Revenue, X_1 = Number of partners, X_2 = Number of Professionals, X_3 = MAS (%), $X_4 = 1$ for Southeast Region; 0 otherwise, $X_5 = 1$ for Gulf Coast Region; 0 otherwise.

Based on a full regression model involving all of the variables:

X_1 and X_2 have $VIFs > 5$ with X_2 having the highest $VIF = 6.9763$. All $VIFs < 5$ after dropping X_2 . So there is no reason to suspect collinearity between any pair of the remaining variables.

Let Y = Revenue, X_1 = Number of partners, X_2 = MAS (%), $X_3 = 1$ for Southeast Region; 0 otherwise, $X_4 = 1$ for Gulf Coast Region; 0 otherwise.

The best-subset approach yields the following models to be considered:

Partial PHStat output from the best-subsets selection:

Model	Cp	k+1	R Square	Adj. R Square	Std. Error	Consider This Model?
X1	2.1964	3	0.8237	0.8157	20.5374	Yes
X1X2	3.3349	4	0.8272	0.8151	20.5667	Yes
X1X2X3	3.2422	4	0.8276	0.8155	20.5441	Yes
X1X2X3X4	5.0000	5	0.8286	0.8122	20.7276	Yes



arson

Partial PHStat output of the full regression model:

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	-3.4970	6.5617	-0.5329	0.5969
Number of Partners	1.3536	0.1063	12.7379	0.0000
MAS (%)	0.7215	0.2507	2.8778	0.0063
SE	3.9856	8.0979	0.4922	0.6252
GC	-4.2658	7.3713	-0.5787	0.5659

Since the p -value for X_3 and X_4 are larger than 0.05, they do not have significant effect cont. individually on revenue. X_3 with the largest p -value is dropped.

Partial PHStat output of the regression model with X_3 dropped:

	Coefficients	Standard Error	t Stat	P-value
Intercept	-1.8107	5.5465	-0.3264	0.7457
Number of Partners	1.3705	0.0997	13.7456	0.0000
MAS (%)	0.6994	0.2445	2.8608	0.0065
GC	-6.1515	6.2416	-0.9856	0.3299

Since the p -value for X_4 is larger than 0.05, it does not have significant effect individually on revenue. X_4 with the largest p -value is dropped.

Partial PHStat output of the regression model with X_4 dropped:

	Coefficients	Standard Error	t Stat	P-value
Intercept	-4.1412	5.0156	-0.8257	0.4135
Number of Partners	1.3698	0.0997	13.7439	0.0000
MAS (%)	0.7090	0.2442	2.9031	0.0058

The best model should include both X_1 and X_2

PHStat output of the model with only X_1 and X_2 :


Regression Statistics						
Multiple R	0.9076					
R Square	0.8237					
Adjusted R Square	0.8157					
Standard Error	20.5374					
Observations	47					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	86692.4957	43346.2479	102.7682	0.0000	
Residual	44	18558.6100	421.7866			
Total	46	105251.1057				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-4.1412	5.0156	-0.8257	0.4135	-14.2494	5.9671
Number of Partners	1.3698	0.0997	13.7439	0.0000	1.1689	1.5707
MAS (%)	0.7090	0.2442	2.9031	0.0058	0.2168	1.2012

The p -value of the t -test for the significance of individual independent variables are all < .05.

The first observation has a Studentized deleted residual $|t^*| = 4.2369 > t_{\alpha/2} = 1.6811$

with d.f. = 44, a hat matrix diagonal element $h_i = 0.5315 > 2(k+1)/n = 0.1277$

and a Cook's $D_i = 4.8999 > F_{\alpha} = 0.8010$ with d.f. = 3 and 44.



Hence, using the Studentized deleted residuals, hat matrix diagonal elements and Cook's distance statistic together, the first observation should be deleted from the data set.



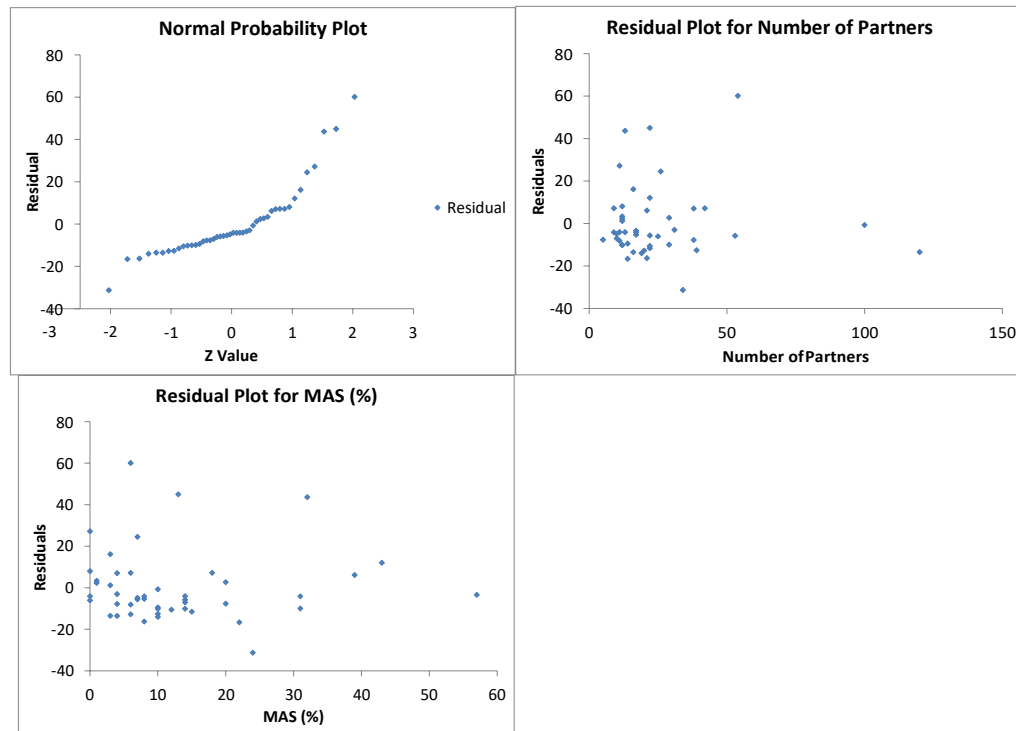
PHStat output with the first observation deleted: cont.

Pearson

Regression Statistics						
Multiple R	0.7959					
R Square	0.6334					
Adjusted R Square	0.6164					
Standard Error	17.4495					
Observations	46					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	2	22621.7672	11310.8836	37.1478	0.0000	
Residual	43	13092.7953	304.4836			
Total	45	35714.5625				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	5.7798	4.8624	1.1887	0.2411	-4.0262	15.5858
Number of Partners	1.0089	0.1201	8.4006	0.0000	0.7667	1.2512
MAS (%)	0.5402	0.2113	2.5565	0.0142	0.1140	0.9663

The most appropriate model to predict sales is

$$\hat{Y} = 5.7798 + 1.0089 X_1 + 0.5402 X_3$$



The normal probability plot suggests possibly departure from the normality assumption. However, because of the large sample size, the validity of the results is not seriously impacted. The residual plots do not reveal any specific pattern.

16.19 Using best subsets

Best subsets analysis

Intermediate Calculations	
R^2_T	0.399409
$1 - R^2_T$	0.600591
n	15
T	3
$n - T$	12

Model	C_p	$k+1$	R^2	R^2_{adj}	Std. Error
X1	1.0135	2	0.3987	0.3525	5.3711
X2	7.8115	2	0.0585	-0.0139	6.7211
X1X2	3.0000	3	0.3994	0.2993	5.5873

Using Stepwise

	df	SS	MS	F	Significance F
Regression	1	248.7041	248.7041	8.6211	0.0116
Residual	13	375.0293	28.8484		



14

623.7333

Pearson

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t_{calc}</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	10.5619	4.0923	2.5809	0.0228	1.7210	19.4028
Unemployment Rate	1.4493	0.4936	2.9362	0.0116	0.3829	2.5157

Based upon best subsets and stepwise techniques we should include only the unemployment rate as an explanatory variable.

16.20 Using best subsets

Best subsets analysis

Intermediate Calculations	
R^2T	0.968235
$1 - R^2T$	0.031765
n	21
T	3
$n - T$	18



arson

Model	Cp	k+1	R Square	Adj. Square	R Std. Error
X1	9.1995	2	0.9538	0.9513	43.0746
X2	80.4280	2	0.8281	0.8190	83.0648
X1X2	3.0000	3	0.9682	0.9647	36.6819

Using stepwise

Stepwise Regression Analysis

Table of Results for General Stepwise

price entered.

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	727224.1267	727224.1267	391.9454	0.0000
Residual	19	35253.0161	1855.4219		
Total	20	762477.1429			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t_{calc}</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	985.2239	29.7425	33.1251	0.0000	922.9721	1047.4758
price	-0.9649	0.0487	-19.7976	0.0000	-1.0670	-0.8629

income entered.

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	738257.0376	369128.5188	274.3305	0.0000
Residual	18	24220.1053	1345.5614		
Total	20	762477.1429			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t_{calc}</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	735.9766	90.6539	8.1185	0.0000	545.5198	926.4334
price	-0.7539	0.0846	-8.9122	0.0000	-0.9316	-0.5762
income	0.2116	0.0739	2.8635	0.0103	0.0564	0.3669

Based upon the output we should include both explanatory variables – price and income.

16.21 Using best subsets

Best subsets analysis



Pearson

Intermediate Calculations	
R^2T	0.524165
$1 - R^2T$	0.475835
n	13
T	3
$n - T$	10

Model	C_p	$k+1$	R^2	R^2_{adj}	Std. Error
X1	5.7214	2	0.2995	0.2358	4.8257
X2	12.0048	2	0.0005	-0.0903	5.7643
X1X2	3.0000	3	0.5242	0.4290	4.1714

Using stepwise

Stepwise Regression Analysis

Table of Results for General Stepwise

No variables could be entered into the model.

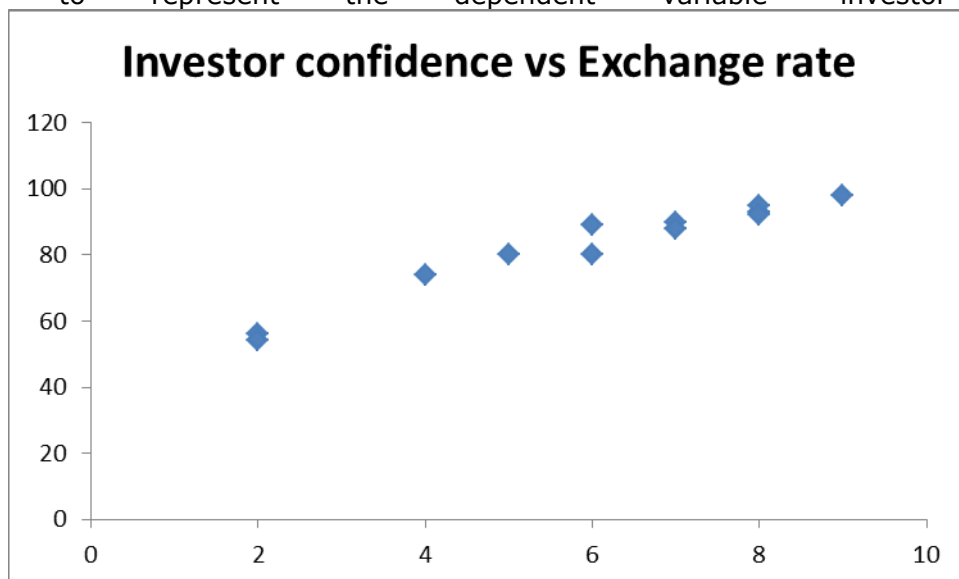
Based upon the output we should include both explanatory variables.

- 16.22 We can measure the influence of individual observations on the model by computing the Variance Inflationary Factor (VIF).
- 16.23 Stepwise regression and best-subsets regression both identify useful predictors during the exploratory stages of model building. Stepwise regression starts with one independent variable and adds variables until we reach a stage where no more significant variables can be added. In comparison, best-subsets regression evaluates all possible specifications, and we generally use the adjusted R^2 or C_p to arrive at the preferred model.
- 16.24 The C_p is used to determine which combination of independent X variables is best to use in a multiple regression model. The C_p statistic, defined in Equation 16.10, measures the differences between a fitted regression model and a true model, along with random error. When a regression model with k independent variables contains only random differences from a true model, the mean value of C_p is $k + 1$. The goal is to find models whose C_p is close to or less than $k + 1$.
- 16.25 There are several ways of validating a regression model:
- Collect new data and compare the results.
 - Compare the results of the regression model with previous results.
 - If the data set is large, split the data into two parts and cross-validate the results.



16.26 (a) For the scatter diagram: X-axis to represent the independent variable 'exchange rate' while Y-axis to represent the dependent variable 'investor confidence'.

arson



(b) $\hat{Y} = 2.6761 - 0.1095X + 0.0018X^2$

(c) $\hat{Y} = 2.6761 - 0.1095(85) + 0.0018(85)^2 = 6.1391$

(d) $H_0 = \beta_1 = \beta_2 = 0$

$H_1 =$ at least one $\beta_j \neq 0$

$F = 131.66 > F_{2,9} = 4.26$. Reject H_0 . There is a significant quadratic relationship between investor confidence and exchange rates.

(e) $H_0 =$ including the quadratic effect does not significantly improve the model ($\beta_2 = 0$).

$H_1 =$ including the quadratic effect significantly improves the model ($\beta_2 \neq 0$).

$t_{\text{calc}} = 2.1595 < t_9 = 2.2622$. Fail to reject H_0 . The quadratic effect is not significant. Therefore, the quadratic model is not a better fit than the linear regression model.

(f) The coefficient of multiple determination R^2 represents the proportion of variation in Y that is explained by the variation in the independent variables. Therefore, 96.69% of the variation in investor confidence is explained by the quadratic relationship between investor confidence and exchange rates.

(g) $R^2_{\text{adj}} = 1 - \left[(1 - R^2) \frac{(n-1)}{(n-k-1)} \right]$

$$R^2_{\text{adj}} = 1 - \left[(1 - 0.9669) \frac{(12-1)}{(12-2-1)} \right]$$

$$R^2_{\text{adj}} = 1 - 0.0405 = 0.9595$$

16.27

Since the variable *Rooms* is the sum of *Bathrooms*, *Bedrooms*, *Loft/Den* and *Finished Basement*, it is removed from the list of potential independent variable. Including it will introduce perfect collinearity.

An analysis of the linear regression model using PHStat with all of the remaining seven possible independent variables revealed that none of the variables have *VIF* values in excess of 5.0.

A best subsets regression produces the following potential models that have C_p values less than or equal to $k+1$.

Model	C_p	$k+1$	R Square	Adj. R Square	Std. Error
X1X2X3X4X5	5.657125	6	0.532645	0.490157971	yes
X1X2X3X5X6	5.78991	6	0.531509	0.488919347	yes
X1X2X3X4X5X6	6.074978	7	0.546173	0.49574803	yes
X1X2X3X4X5X6X7	8	8	0.546814	0.486959627	yes

where $X_1 = \text{Hot tub}$ (0 = No and 1 = Yes), $X_2 = \text{Lake View}$ (0 = No and 1 = Yes), $X_3 = \text{Bathrooms}$, $X_4 = \text{Bedrooms}$, $X_5 = \text{Loft/Den}$ (0 = No and 1 = Yes), $X_6 = \text{Finished Basement}$ (0 = No and 1 = Yes), $X_7 = \text{Acres}$.

Looking at the p -values of the t statistics for each slope coefficient of the model that includes

X_1 through X_7 reveals that *Bathrooms*, *Bedrooms*, *Loft/Den*, *Finished Basement* and *Acres* are not significant at 5% level of significance.

	Coefficients	Standard Error	t Stat	P -value
Intercept	56.86025281	75.83635651	0.749775641	0.456705242
Hot Tub	83.33704829	39.54169465	2.107574019	0.039815205
Lake View	188.1459004	46.55629756	4.041255648	0.000172817
Bathrooms	44.97723054	28.05985879	1.602902954	0.114899805
Bedrooms	31.7825021	24.30126221	1.307853963	0.196567371
Loft/Den	68.68786861	34.62218517	1.983926441	0.052453764
Finished Basement	42.40047126	35.03765953	1.210139942	0.231594833
Acres	8.214876036	30.00092343	0.273820773	0.7852867

Dropping *Acres* which has the highest p -value, the new regression indicates that *Bathrooms*, *Bedrooms* and *Finished Basement* are still not significant.

	Coefficients	Standard Error	t Stat	P -value
Intercept	63.04035916	71.77716439	0.878278763	0.383683708
Hot Tub	82.87369204	39.16564269	2.115979372	0.038973223
Lake View	190.4252197	45.41206354	4.193273877	0.000102798
Bathrooms	44.42625163	27.74686825	1.601126702	0.115183479
Bedrooms	31.82322517	24.09177116	1.320916796	0.192099589
Loft/Den	68.8302055	34.3204956	2.005513157	0.049930143
Finished Basement	43.67867801	34.42659968	1.268747957	0.209972962



Dropping *Finished Basement*, which has the largest p -value, the new regression indicates cont. that *Bathrooms* and *Bedrooms* are still insignificant.

arson

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	33.50312632	68.27210075	0.490729389	0.625570141
Hot Tub	98.1802218	37.46721666	2.620430087	0.011330891
Lake View	181.7931221	45.14769931	4.026630921	0.000174862
Bathrooms	52.45485906	27.1649841	1.930973303	0.058647073
Bedrooms	36.99281947	23.87596476	1.549374856	0.127027169
Loft/Den	79.72730905	33.41209396	2.386181158	0.020493138

Dropping *Bedrooms* next, which has the largest p -value, the new regression indicates that all the remaining variables become significant at 5% level of significance.

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	102.0014472	52.67086383	1.936582007	0.057848524
Hot Tub	89.12922256	37.4689494	2.378748911	0.020806217
Lake View	183.8349004	45.68930995	4.023586712	0.00017363
Bathrooms	77.69078822	22.01055013	3.529706789	0.000839855
Loft/Den	76.80439638	33.77336974	2.274111141	0.026811687

The best linear model is determined to be:

$$\hat{Y} = 102.0014 + 89.1292 X_1 + 183.8349 X_2 + 77.6908 X_3 + 76.8044 X_5$$

The overall model has $F_{STAT} = 14.7030$ (4 and 56 degrees of freedom) with a p -value that is

virtually 0. $r^2 = 0.5122$, $r^2_{adj} = 0.4774$.

The 52nd observation has a Studentized deleted residual $t^* = 5.7102 > t_{\alpha/2} = 1.6730$ with d.f. =

56, a hat matrix diagonal element $h_i = 0.1835 > 2(k+1)/n = 0.1639$ and a Cook's $D_i = 0.9370$

$> F_{\alpha} = 0.8809$ with d.f. = 5 and 56.

Hence, using the Studentized deleted residuals, hat matrix diagonal elements and Cook's distance statistic together, the 52nd observation should be deleted from the data set.

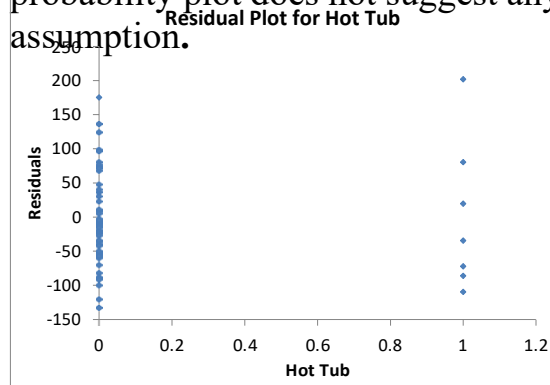


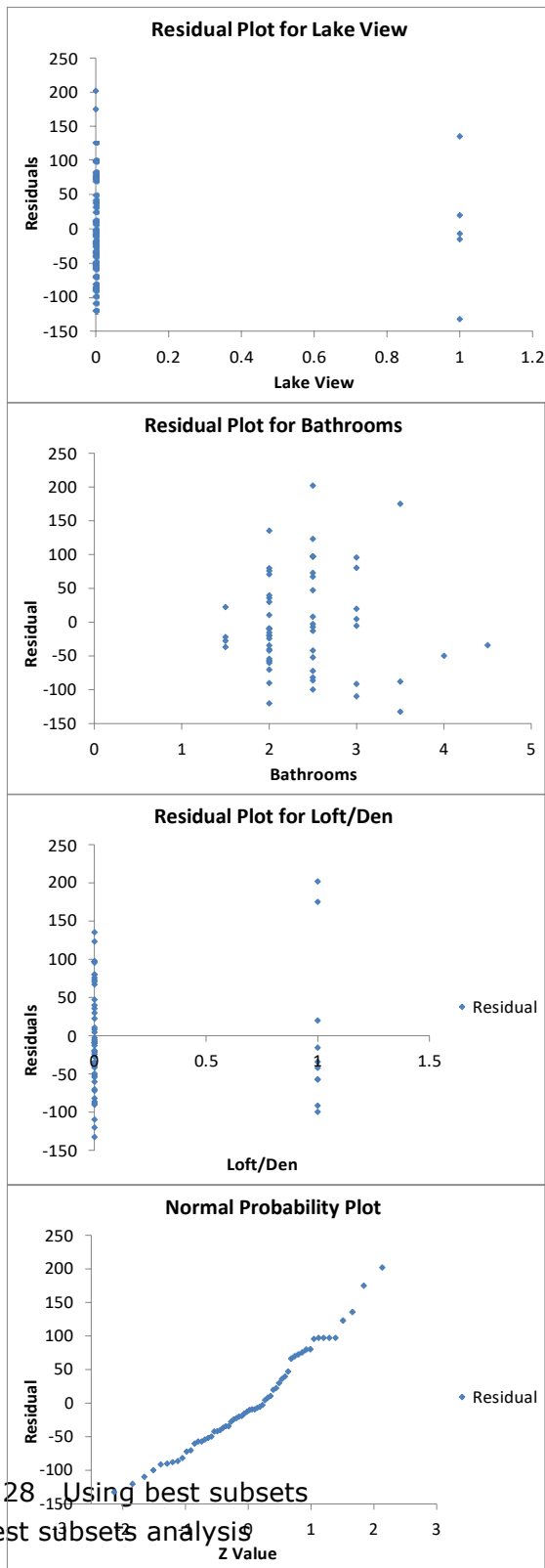
Pearson

PHStat output after deleting the 52nd observation. cont.

Regression Statistics						
Multiple R	0.7622					
R Square	0.5810					
Adjusted R Square	0.5505					
Standard Error	76.9111					
Observations	60					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	4	451063.3243	112765.8311	19.0634	0.0000	
Residual	55	325342.2590	5915.3138			
Total	59	776405.5833				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	88.6802	42.1756	2.1026	0.0401	4.1584	173.2020
Hot Tub	34.2776	31.4593	1.0896	0.2806	-28.7683	97.3235
Lake View	204.4737	36.7076	5.5703	0.0000	130.9101	278.0374
Bathrooms	85.2307	17.6472	4.8297	0.0000	49.8650	120.5965
Loft/Den	36.9265	27.8907	1.3240	0.1910	-18.9677	92.8208

A residual analysis does not reveal any strong patterns and the normal probability plot does not suggest any departure from the normality assumption.





16.28 Using best subsets
Best subsets analysis

Intermediate Calculations

R^2T	0.970118
$1 - R^2T$	0.029882
n	10
T	5
$n - T$	5

Model	C_p	$k+1$	R^2	R^2_{adj}	Std Error
X1	0.1637	2	0.9632	0.9586	177500.8425
X2	40.3816	2	0.7228	0.6882	486916.0372
X3	3.2951	2	0.9444	0.9375	217976.0978
X4	151.2612	2	0.0602	-0.0573	896585.6861
X1X2	1.7114	3	0.9659	0.9561	182661.4553
X1X3	1.6629	3	0.9662	0.9565	181885.4082
X1X4	1.9200	3	0.9646	0.9545	185968.4955
X2X3	4.3839	3	0.9499	0.9356	221309.7260
X2X4	41.8160	3	0.7262	0.6480	517351.3096
X3X4	4.3188	3	0.9503	0.9361	220449.2488
X1X2X3	3.3110	4	0.9683	0.9524	190256.2955
X1X2X4	3.5184	4	0.9670	0.9505	193935.6927
X1X3X4	3.2829	4	0.9684	0.9526	189751.7029
X2X3X4	5.6387	4	0.9543	0.9315	228170.1253
X1X2X3X4	5.0000	5	0.9701	0.9462	202220.9133

Using stepwise

Stepwise Regression Analysis

Table of Results for General Stepwise

price chicken (\$/kg) entered.

	df	SS	MS	F	Significance F
Regression	1	6590515245923.2000	6590515245923.2000	209.1792	0.0000
Residual	8	252052392626.8990	31506549078.3624		
Total	9	6842567638550.1000			

	Coefficients	Standard Error	t Stat	P -value	Lower 95%	Upper 95%
Intercept	7371855.9427	248124.5325	29.7103	0.0000	6799679.7448	7944032.1406
price chicken (\$/kg)	-293600.0036	20300.0352	-14.4630	0.0000	-340411.9687	-246788.0385

No other variables could be entered into the model. Stepwise ends.

The best model includes the price of chicken only as an explanatory variable.

16.29

Glencove:

Based on a full regression model involving all of the variables:
All *VIFs* are less than 5. So there is no reason to suspect collinearity between any pair of variables.

The best-subset approach yielded the following models to be

<u>Model considered:</u>	<u>Cp</u>	<u>k+1</u>	<u>R Square</u>	<u>Adj. R Square</u>	<u>Std. Error</u>	<u>Consider This Model?</u>
X1X2X3	2.1558	4	0.8424	0.8242	60.5007	Yes
X1X2X3X4	4.1117	5	0.8427	0.8175	61.6425	Yes
X1X2X3X5	3.2400	5	0.8484	0.8241	60.5180	Yes
X1X2X3X6	3.8887	5	0.8442	0.8192	61.3567	Yes
X1X2X3X4X5	5.1832	6	0.8488	0.8173	61.6903	Yes
X1X2X3X4X6	5.7825	6	0.8449	0.8125	62.4827	Yes
X1X2X3X5X6	5.1038	6	0.8493	0.8179	61.5846	Yes
X1X2X3X4X5X6	7.0000	7	0.8500	0.8108	62.7677	Yes

The stepwise regression approach reveals the following best model:

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	260.6791	66.3288	3.9301	0.0006
Property Size (acres)	362.8318	48.6233	7.4621	0.0000
House Size (square feet)	0.1109	0.0228	4.8682	0.0000
Age	-1.7543	0.5483	-3.1996	0.0036

The *p*-value of the individual slope coefficients indicate that all the remaining independent variables are significant individually.

Combining the results of both approaches, the most appropriate multiple regression model for predicting fair market value in Glencove is

$$\hat{Y} = 260.6791 + 362.8318 X_1 + 0.1109 X_2 - 1.7543 X_3$$