Chapter 13: Introduction to multiple regression

Learning objectives

After studying this chapter you should be able to:

1. construct a multiple regression model and analyse model output
2. differentiate between independent variables and decide which ones to include in the regression model, and determine which independent variables are more important in predicting a dependent variable
3. incorporate categorical and interactive variables in a regression model
4. detect collinearity

13.1 (a) Holding constant the effect of X_2, for each increase of one unit in X_1, the response variable Y is estimated to increase a mean of 8 units. Holding constant the effect of X_1, for each increase of one unit in X_2, the response variable Y is estimated to decrease an average of 3 units.

(b) The Y intercept 10 is the estimate of the mean value if X_1 and X_2 are both 0.

13.2 (a) Holding constant the effect of X_2, for each increase of one unit in X_1, the response variable Y is estimated to increase an average of 10 units. Holding constant the effect of X_1, for each increase of one unit in X_2, the response variable Y is estimated to decrease an average of 15 units.

(b) The Y intercept 100 is the estimate of the mean value of Y if X_1 and X_2 are both 0.

13.3 (a) $\hat{Y} = 0.02686 + 0.79116X_1 + 0.60484X_2$

(b) For a given measurement of the change in impact properties over time (holding constant the effect of X_2), each increase in one unit in forefoot shock-absorbing capability, X_1, is estimated to result in a mean increase in the long-term ability to absorb shock of 0.79116 units. For a given forefoot shock-absorbing ability (holding constant the effect of X_1), each increase of one unit in measurement of the change in impact properties over time is estimated to result in a mean increase in the long-term ability to absorb shock by 0.60484 units.

13.4 (a) $\hat{Y} = 9.8156 + 1.47693X_1 + 0.10344X_2$

(b) Holding the effect of stock market rate, X_2, constant, for each increase of 1% in unemployment rate, is estimated to result in a mean increase in the retirement rate by 1.477%. Holding the effect of unemployment rate, X_1, constant, for each increase of 1% in the stock market return rate, is estimated to result in a mean increase in the retirement rate by 0.1034%.

(c) The interpretation of b_0 has no practical meaning here because it would have been the estimated mean retirement rate when there were no unemployment rate and no stock market return rate. However, this does not allow for any other influence on retirement rate.

(d) $\hat{Y} = 9.8156 + 1.47693(6) + 0.10344(5) = 19.1944\%$

(e) $15.2758 \le \mu_{Y|X} \le 23.1130$

(f) $6.4057 \le Y_X \le 31.98312$

13.5 (a) The predicted sign for b_1 will be positive since the higher the GDP we would expect higher CO_2 emissions. Similarly, the predicted sign for b_2 will be positive; the

higher the population density we would expect higher CO_2 emissions.

(b) $\hat{Y} = 151.3935 + 204.6397X_1 - 0.1639X_2$

(c) For a given population density, each increase of GDP US$1 trillion is estimated to result in a mean increase of CO_2 emission of 204.6397 million metric tonnes. For a given GDP, each increase in population density of 1 person per kilometre is estimated to result in a mean decrease of 0.1639 million metric tonnes.

(d) The interpretation of b_0 has no practical meaning here because it would have been the estimated mean CO_2 emission for a country with zero GDP and population density.

(e) $\hat{Y} = 151.3935 + 204.6397(1) - 0.1639(50) = 347.8404$ million metric tonnes

(f) $-1677.83 \le \mu_{Y|X} \le 1030.133$

(g) $-1677.83 \le Y_X \le 2373.51$

13.6

a) PHStat output:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	1.1592	1.2719	0.9114	0.3667
alcohol	0.4962	0.1094	4.5378	0.00000
chlorides	-9.6331	3.6818	-2.6164	0.01199

$\hat{Y} = 1.1592 + 0.4962\ X_1 - 9.6331\ X_2$

(b) For a given amount of chlorides, each increase of one percent in alcohol is estimated to result in a mean increase in quality rating of 0.4962. For a given alcohol content, each increase of

one unit in chlorides is estimated to result in the mean decrease in quality rating of 9.6331.

(c) The interpretation of b_0 has no practical meaning here because it would have meant the estimated mean quality rating when a wine has 0 alcohol content and 0 amount of chlorides.

(d) $\hat{Y} = 1.1592 + 0.4962\,(10) - 9.6331(.08) = 5.3510.$

(e) $5.0635 \le \mu_{Y|X} \le 5.6386$

(f) $3.5484 \le Y_X \le 7.1536$

(g) The model uses both alcohol content (%) and the g) The model uses both alcohol content (%) and the amount of chlorides to predict wine quality. This may produce a better model than if only one of these independent variables is included.

13.7 (a) $\hat{Y} = 39.45 - 0.0003X_1 + 0.1526X_2$

(b) For a given CPI, each increase in GDP/capita of $1 is estimated to result in a mean decrease in the percentage of very happy citizens of 0.0003 percentage points. For a given GDP/capita, each one-point increase in CPI is estimated to result in a mean increase in the percentage of very happy citizens by 0.1526 percentage points.

(c) The interpretation of b_0 has no practical meaning here because it would have been the estimated mean percentage of very happy citizens when GDP/capita and CPI were zero.

(d) $\hat{Y} = 0.3945 - 0.000003(35000) + 0.0015(75) = 40.03$

(e) $37.28 \le \mu_{Y|X} \le 42.77$

(f) $30.33 \le Y_X \le 49.72$

13.8

(a) $\hat{Y} = 156.4 + 13.081\,X_1 + 16.795\,X_2$

(b) For a given amount of newspaper advertising, each increase of $1000 in radio advertising is estimated to

result in a mean increase in sales of $13,081. For a given amount of radio advertising, each increase of $1000

in newspaper advertising is estimated to result in the mean increase in sales of $16,795.

(c)When there is no money spent on radio advertising and newspaper advertising, the estimated mean amount of sales is $156,430.44.

(d) According to the results of (b), newspaper advertising is more effective as each increase of $1000 in newspaper advertising will result in a higher mean increase in sales than the same amount of increase in radio advertising.

13.9 (a) $MSR = SSR/k = 55/2 = 27.5$

$MSE = SSE/(n - k - 1) = 145/18 = 8.06$

(b) $F = MSR/MSE = 27.5/8.06 = 3.41$

(c) $F = 3.41 < F_{U(2,18)} = 3.555$. Do not reject H_0. There is no evidence of significant linear relationship.

(d) $R^2 = \frac{SSR}{SST} = \frac{55}{190} = 0.2895$

28.95% of the variation in y is explained by the model

(e) $R^2_{adj} = 1 - \left[(1 - R^2)\frac{n-1}{n-k-1}\right] = 1 - \left[(1 - 0.2895)\frac{21-1}{21-2-1}\right] = 0.2105$

13.10 (a) $MSR = SSR/k = 185/2 = 92.5$
$MSE = SSE/(n-k-1) = 315/10 = 31.5$

(b) $F = MSR/MSE = 92.5/31.5 = 2.9365$

(c) $F = 2.9365 > F_{U(2,8)} = 4.46$. Do not reject H_0. There is not sufficient evidence of a significant linear relationship.

(d) $R^2 = \frac{SSR}{SST} = \frac{185}{500} = 0.37$

(e) $R_{adj}{}^2 = 1 - \left[(1 - R^2)\frac{n-1}{n-k-1}\right] = 1 - \left[(1 - 0.37)\frac{11-1}{11-2-1}\right] = 0.2125$

13.11 (a) 72% of the total variability in enjoyment can be explained by length of stay after adjusting for the number of predictors and sample size. 78% of the total variability in enjoyment can be explained by average income after adjusting for the number of predictors and sample size. 68% of the total variability in enjoyment can be explained by both length of stay and average income after adjusting for the number of predictors and sample size.

(b) Model 2 is the best predictor of enjoyment because it has the highest adjusted R^2.

(c) The regression coefficients are needed to explain the relationships, especially the *b* slope coefficients. For example, if the beta of the length of stay is negative and the beta for average income is negative, it might be that the circle trams appeal most to the tourists with little time and not much money.

13.12 (a) $F = 97.69 > F_{U(2,15-2-1)} = 3.89$. Reject H_0. There is evidence of a significant linear relationship with at least one of independent variables.

(b) *p*-value is virtually zero. The probability of obtaining an *F* test statistic of 97.69 or larger is virtually zero if H_0 is true.

(c) $R^2 = \frac{SSR}{SST} = \frac{12.61020}{13.3847} = 0.9421$. So, 94.21% of the variation in the long-term ability to absorb shock can be explained by variation in forefoot absorbing capability and variation in midsole impact.

(d) $R_{adj}{}^2 = 1 - \left[(1 - R^2)\frac{n-1}{n-k-1}\right] = 1 - \left[(1 - 0.9421)\frac{15-1}{15-2-1}\right] = 0.93245$

13.13 (a) $F = MSR/MSE = 6.39 < F_{U(2,11)} = 3.98$. Reject H_0 and conclude there is evidence of a significant relationship.

(b) *p*-value = 0.014. The probability of obtaining an *F* test statistic of 6.39 or larger is 0.014 if H_0 is true.

(c) $R^2 = \frac{SSR}{SST} = \frac{9594740}{17855076} = 0.5374$. So, 53.74% of the variation in CO_2 emissions can be explained by variation in GDP and population density.

(d) $R^2_{adj} = 1 - \left[(1 - R^2)\frac{n-1}{n-k-1}\right] = 1 - \left[(1 - 0.5374)\frac{14-1}{14-2-1}\right] = 0.4533$

13.14 (a) $F = MSR/MSE = 124.5622939/31.21739546 = 3.990157 > F_{U(2,12)} = 3.89$. Reject H_0 and conclude that there is enough evidence of a significant linear relationship.

(b) *p*-value = 0.047. The probability of obtaining an *F* test statistic of 3.9902 or larger

is 0.047 if H_0 is true.

(c) $R^2 = \frac{SSR}{SST} = \frac{249.12}{623.73} = 0.3994$. So, 39.94% of the variation in retirement rates can be explained by the variation in unemployment rates and stock market returns.

(d) $R_{adj}{}^2 = 1 - \left[(1 - R^2)\frac{n-1}{n-k-1}\right] = 1 - \left[(1 - 0.3994)\frac{15-1}{15-2-1}\right] = 0.2993$

13.15 (a) $F = MSR/MSE = 95.841/17.401 = 5.508 < F_{U(2,10)} = 4.10$. Reject H_0 and conclude that there is enough evidence of a significant linear relationship.

(b) p-value = 0.024. The probability of obtaining an F test statistic of 5.508 or larger is 0.024 if H_0 is true.

(c) $R^2 = \frac{SSR}{SST} = \frac{191.683}{365.692} = 0.5242$. So, 52.42% of the variation in very happy citizens can be explained by GDP/capita and CPI.

(d) $R^2_{adj} = 1 - \left[(1 - R^2)\frac{n-1}{n-k-1}\right] = 1 - \left[(1 - 0.5242)\frac{13-1}{13-2-1}\right] = 0.4290$

13.16

(a) Partial PHStat output:

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	27.2241	13.6120	17.3963	0.0000
Residual	47	36.7759	0.7825		
Total	49	64.0000			

$F_{STAT} = MSR / MSE = 17.3963$

Since p-value = 0.0000 < 0.05, reject H_0. There is evidence of a significant linear relationship.

(b) p-value = 0.0000. The probability of obtaining an F test statistic of 17.3963 or larger is 0.0000 if H_0 is true.

(c) $R^2 = \frac{SSR}{SST} = \frac{27.2241}{64} = 0.4254$

S 2 he variation in quality rating can be
o . o
, 5 f
4
4 % t

explained by variation in the percentage of alcohol and variation in chorides.

(d) $R^2_{adj} = 1 - \left[(1 - R^2) \frac{n-1}{n-k-1} \right] = 0.4009$

13.17

$MSR = SSR / k = 2,028,033 / 2 = 1,014,016$

$MSE = SSE /(n - k - 1) = 479,759.9 / 19 = 25,251$

(a)

$F_{STAT} = MSR / MSE = 1,014,016 / 25,251 = 40.16$

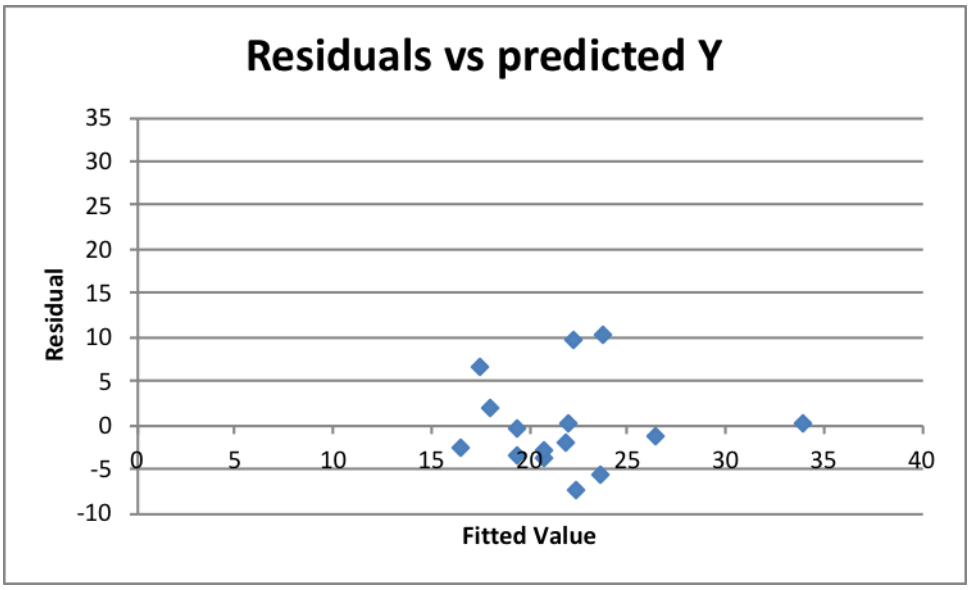
$F_{STAT} = 40 > F_{\alpha} = 3.52$. Reject H_0. There is evidence of a significant linear relationship.

(b) p-value < 0.001. The probability of obtaining an F test statistic of 40.16 or larger is less than 0.001 if H_0 is true.

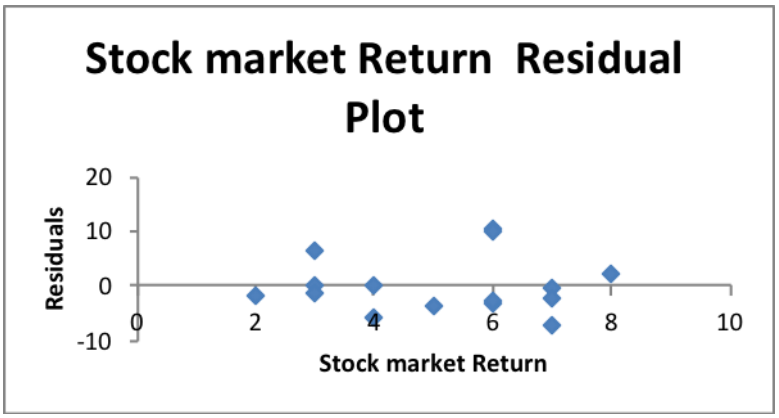
(c) $R^2 = \frac{SSR}{SST} = \frac{2028033}{2507793} = 0.8087$ 80.87% of the variation in sales can be explained by variation in radio per advertising.

(d) $R^2_{adj} = 1 - \left[(1 - R^2)\frac{n-1}{n-k-1}\right] = 0.7886$

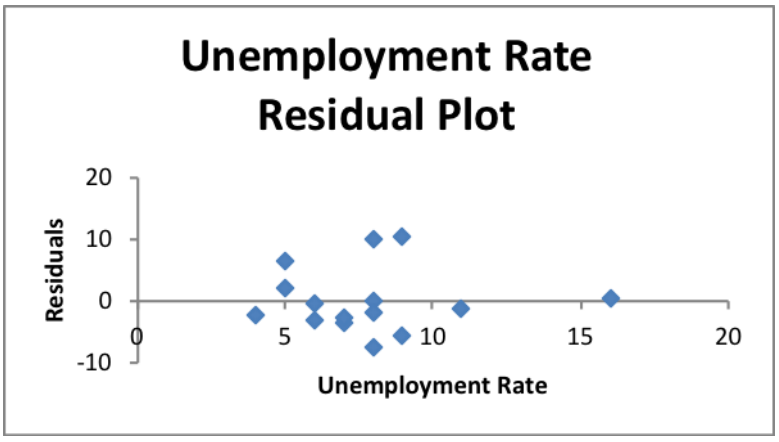
13.18



From the plot residuals versus predicted *Y*, we can see that it does not show a pattern for different predicted values of *Y* and the model appears to be adequate.

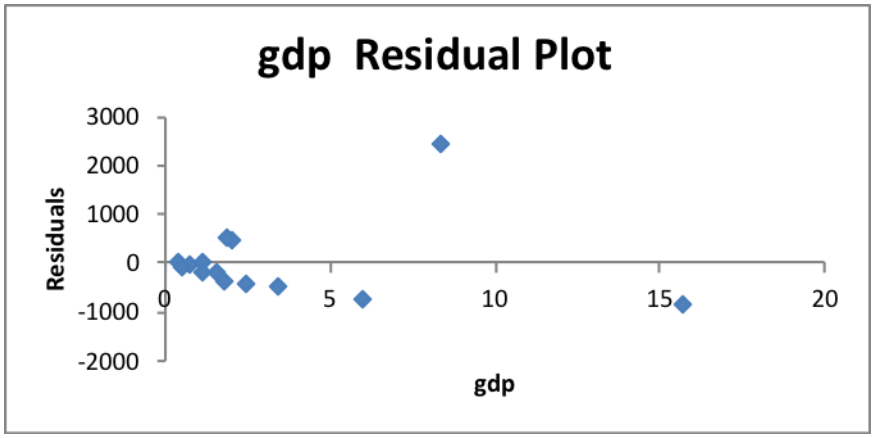


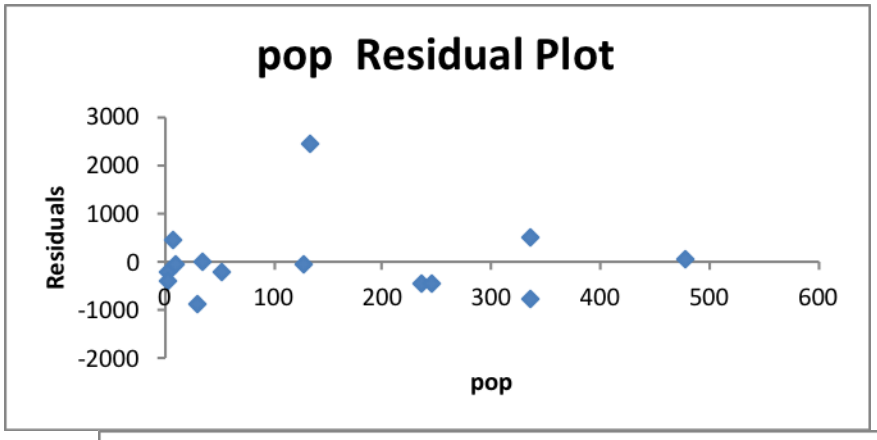
There is no evidence of a pattern in the residual versus stock market return.

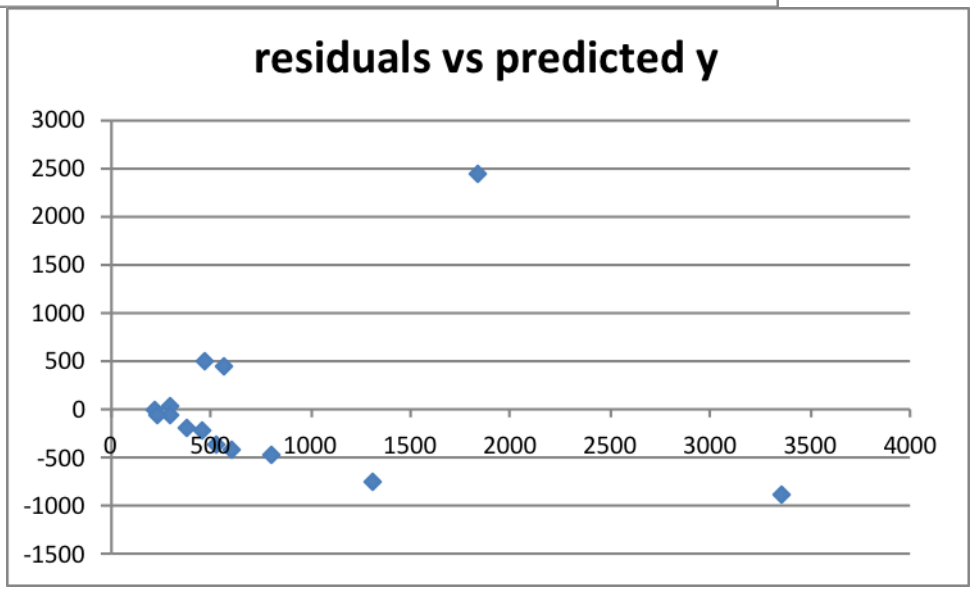


There is no evidence of a pattern in the residual against unemployment rate.

13.19 Excel output

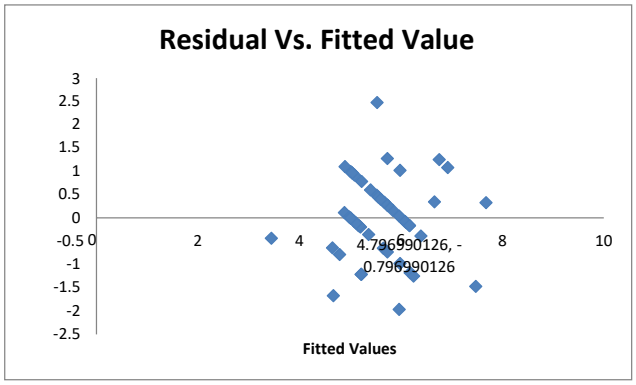


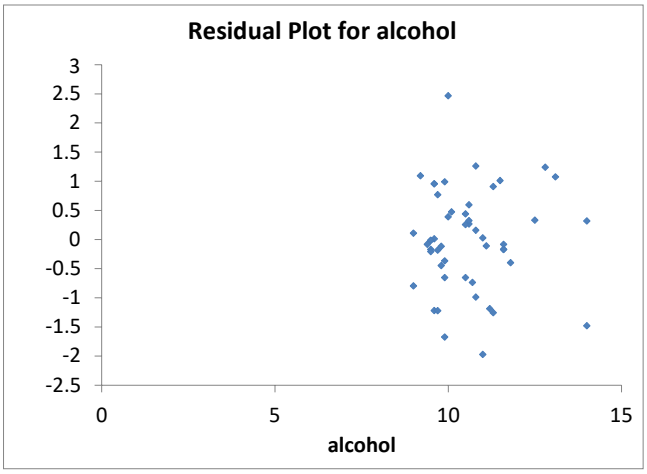


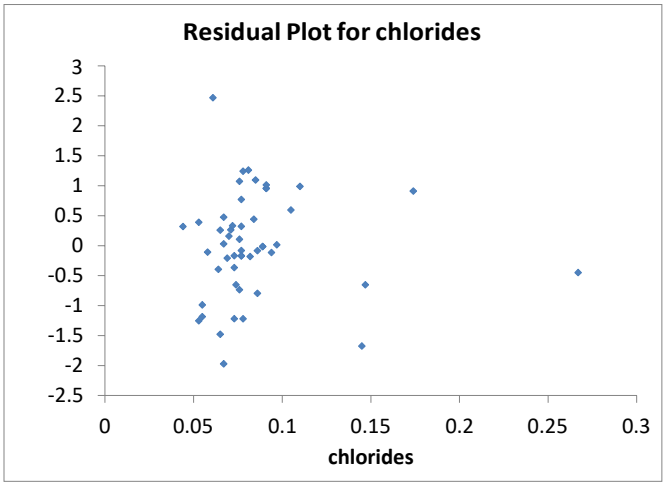


All the plots above show random pattern and there is evidence that the model is adequate.

13.20

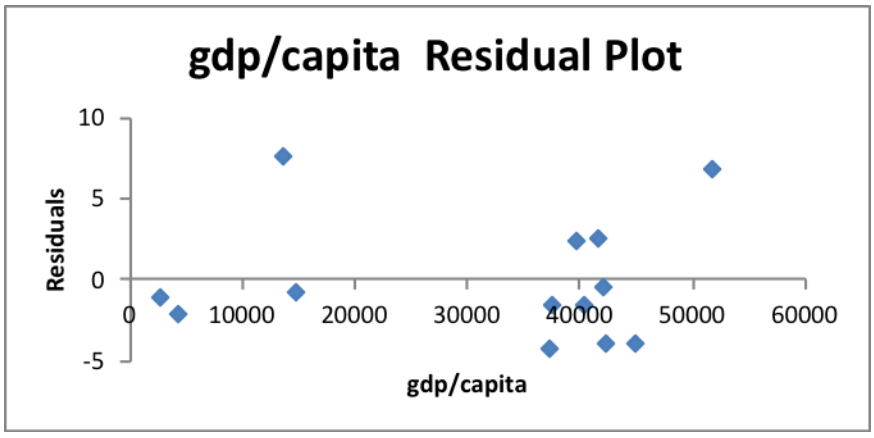


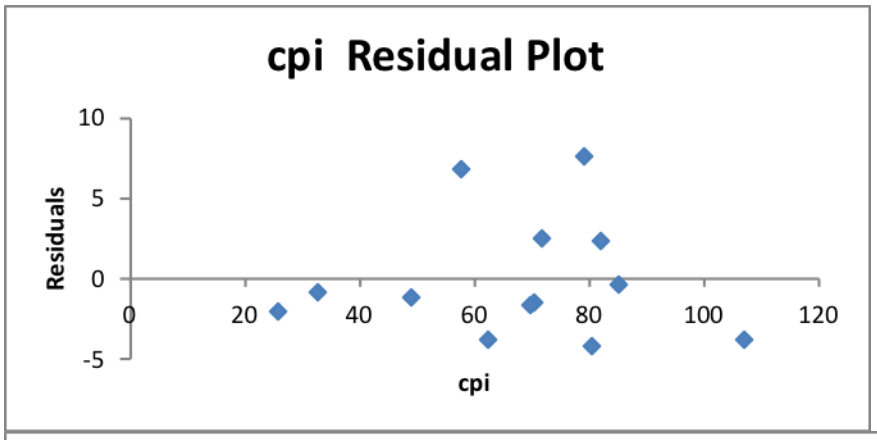


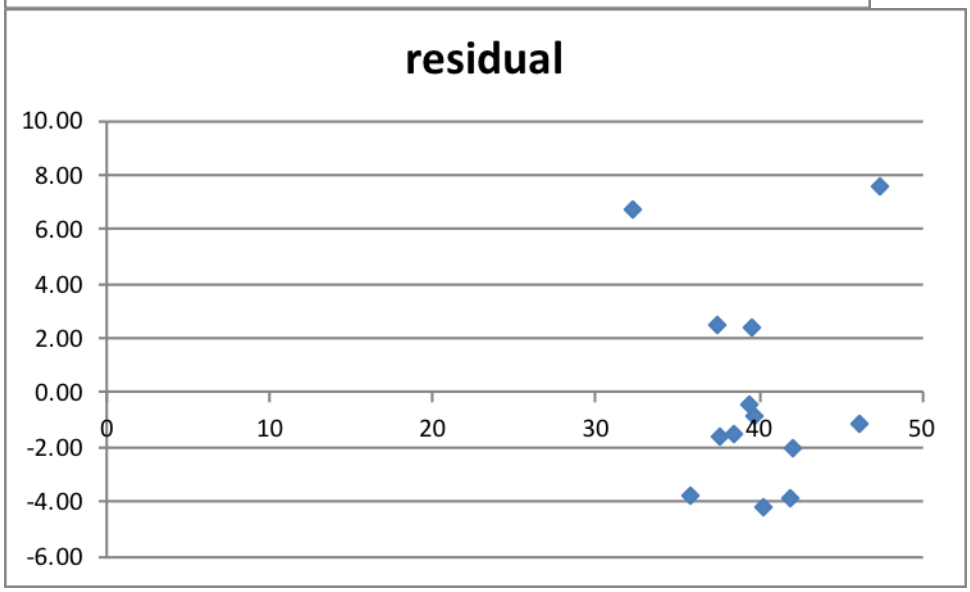


The residual plots do not reveal any specific pattern.

13.21

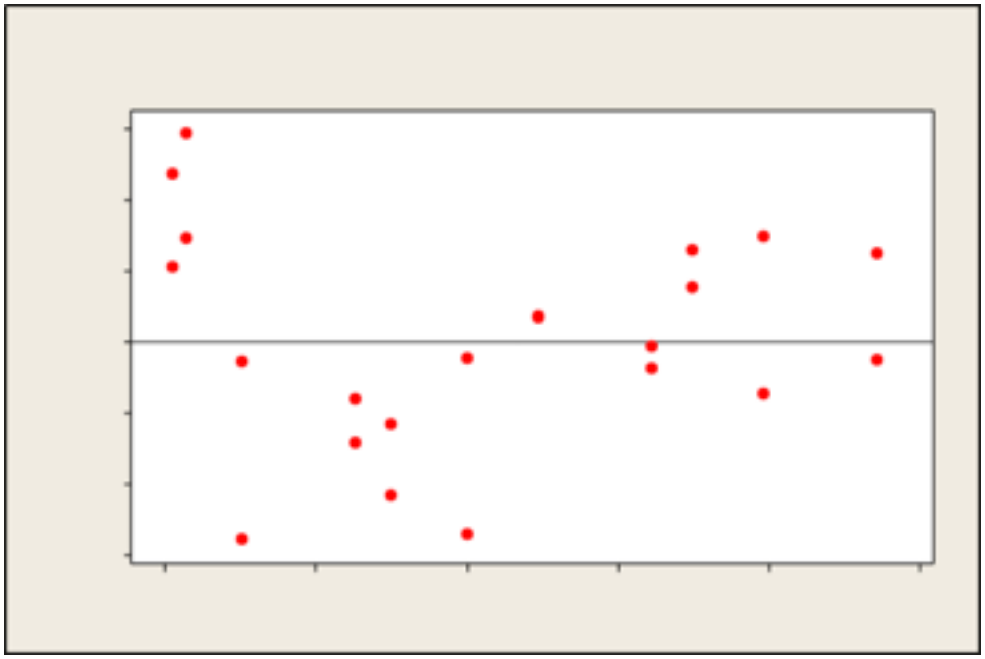


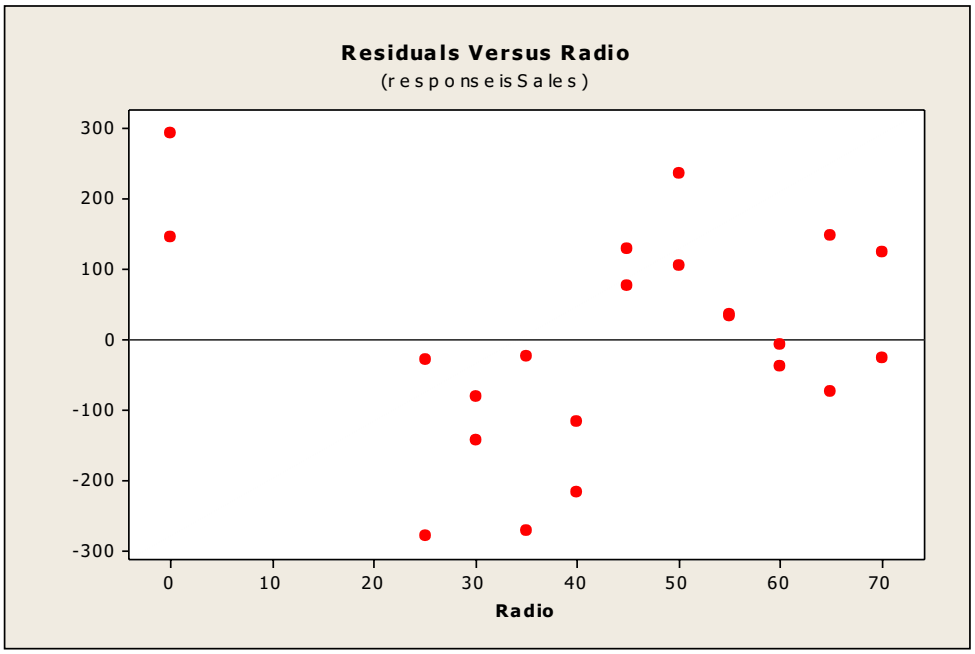


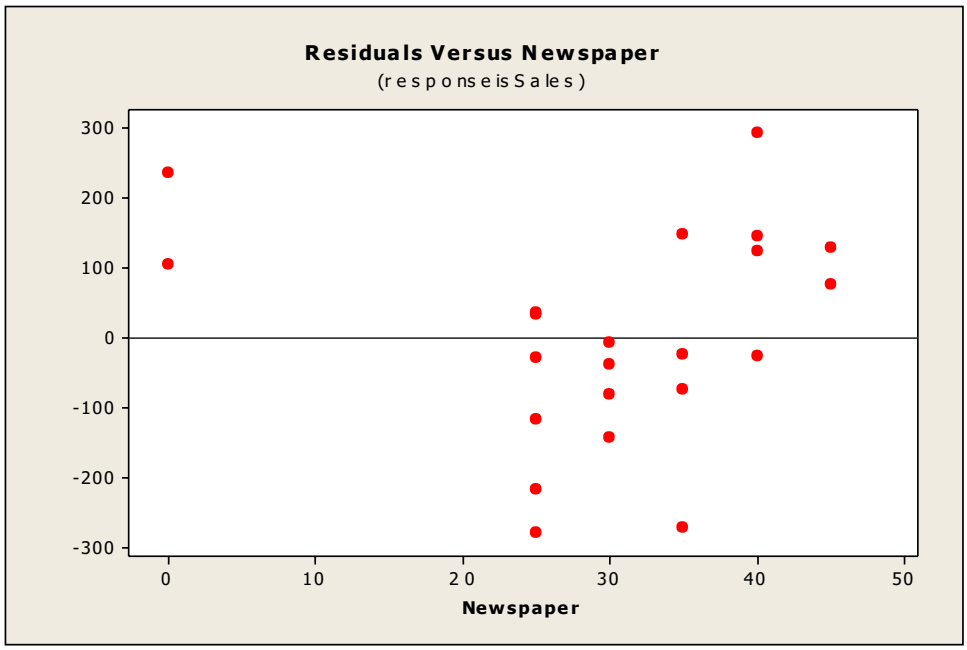


All the plots above show random pattern and there is no evidence in the plot to suggest a non-linear relationship. There is evidence that the model is adequate.

13.22







There appears to be a quadratic relationship in the plot of the residuals against the fitted value and both radio and newspaper advertising. Thus, quadratic terms for each of these explanatory models should be considered for inclusion in the model. The normal probability plot suggests that the distribution of the residuals is very close to a normal distribution.

13.23 (a) The slope of X_2 in terms of t statistic is 4 which is larger than the slope of X_1 in terms of t statistic which is 2.8.

(b) 95% confidence interval of β_1: $b_1 \pm t_{n-k-1} S_{\beta_1}$, $7 \pm 2.018(2.5)$

$1.955 \le \beta_1 \le 12.045$

(c) For X_1 : $t = b_1/S_{b1} = 7/2.5 = 2.8 > t_{42} = 2.018$ with 42 degrees of freedom for $\alpha = 0.05$. Reject H_0. There is evidence that the variable X_1 contributes to a model already containing X_2.
For X_2 : $t = b_2/S_{b2} = 6/1.5 = 4 > t_{42} = 2.018$ with 42 degrees of freedom for $\alpha = 0.05$. Reject H_0. There is evidence that the variable X_2 contributes to a model already containing X_1.
On the basis of these results, both variables X_1 and X_2 should be included in the model.

13.24 (a) 95% confidence interval of β_1: $b_1 \pm t_{n-k-1} S_{\beta_1}$, $0.79 \pm 2.17(0.063)$

$0.6533 \le \beta_1 \le 0.926$

(b) For X_1 : $t = b_1/S_{b1} = 0.79/0.063 = 12.57 > t_{12} = 2.17$ with 12 degrees of freedom for $\alpha = 0.05$. Reject H_0. There is evidence that the variable X_1 contributes to a model already containing X_2.
For X_2 : $t = b_2/S_{b2} = 0.605/0.071 = 8.43 > t_{12} = 2.17$ with 12 degrees of freedom for $\alpha = 0.05$. Reject H_0. There is evidence that the variable X_2 contributes to a model already containing X_1.
On the basis of these results, both variables X_1 and X_2 should be included in the model.

13.25 (a) 95% confidence interval of β_1: $b_1 \pm t_{n-k-1} S_{\beta_1}$, $1.477 \pm 2.17(0.566)$

$0.244 \le \beta_1 \le 2.71$

(b) For X_1 : $t = b_1/S_{b1} = 1.477/0.566 = 2.61 > t_{12} = 2.17$ with 12 degrees of freedom for α = 0.05. Reject H_0. There is evidence that the variable X_1 contributes to a model already containing X_2.
For X_2 : $t = b_2/S_{b2} = 0.103/0.891 = 0.1161 < t_{12} = 2.17$ with 12 degrees of freedom for α = 0.05. Do not reject H_0. There is not enough evidence that the variable X_2 contributes to a model already containing X_1.
On the basis of these results, only variables X_1 should be included in the model.

13.26 (a) 95% confidence interval of β_1: $b_1 \pm t_{n-k-1} S_{\beta_1}$,

$204.6397 \pm 2.201(57.4754)$

$78.1371 \le \beta_1 \le 331.1422$

(b) For X_1 : $t = b_1/S_{b1} = 204.6397/57.4754 = 3.5605 > t_{11} = 2.201$ with 11 degrees of freedom for α = 0.05. Reject H_0. There is evidence that the variable X_1 contributes to a model already containing X_2.
For X_2 : $t = b_2/S_{b2} = -0.1639/1.5509 = -0.1057 > t_{11} = -2.201$ with 11 degrees of freedom for α = 0.05. Do not reject H_0. There is not enough evidence that the variable X_2 contributes to a model already containing X_1.
On the basis of these results, only variables X_1 should be included in the model.

13.27

(a) PHStat output:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	1.15592	1.2719	0.91144	0.36667
alcohol	0.49626	0.1094	4.53788	0.00000
chlorides	-9.	3.6818	-2.	0.0

	6 3 3 1		6 1 6 4	1 1 9

$0.2762 \le \beta_1 \le 0.7162$

For X_1 : $t = b_1/S_{b1} = 0.4962/0.1094 = 4.5378 > t_{47} = 2.0117$ with 47 degrees of freedom for α = 0.05. Reject H_0. There is evidence that the X_1 alcohol contributes to a model already containing X_2 chlorides.
For X_2 : $t = b_2/S_{b2} = -9.6331/3.6818 = -2.6164 < t_{47} = -2.0117$ with 47 degrees of freedom for α = 0.05. Reject H_0. There is evidence that the variable X_2 chlorides contributes to a model already containing X_1 alcohol.Both variables should be included in the model.

13.28 (a) 95% confidence interval of β_1: $b_1 \pm t_{n-k-1} S_{\beta_1}$, $-0.0003 \pm 2.2281(0.00009)$

$-0.0005 \leq \beta_1 \leq -0.0001$

(b) For X_1 : $t = b_1/S_{b1}$ = -0.0003/0.00009 = –3.3174 < t_{10} = –2.2281 with 10 degrees of freedom for α = 0.05. Reject H_0. There is evidence that the variable X_1 contributes to a model already containing X_2.
For X_2 : $t = b_2/S_{b2}$ = 0.1526/0.0702 = 2.1729 < t_{10} = –2.2281 with 10 degrees of freedom for α = 0.05. Do not reject H_0. There is not enough evidence that the variable X_2 contributes to a model already containing X_1.
On the basis of these results, only X_1 should be included in the model.

13.29 (a) $13.0807 \pm 2.093\ (1.7594)$

$9.398 \leq \beta_1 \leq 16.763$

(b) For X_1 : $t = b_1/S_{b1}$ = 13.0807/1.7594 = 7.43 > t_{19} = 2.093 with 19 degrees of freedom for α = 0.05. Reject H_0. There is evidence that the X_1 radio advertsising contributes to a model already containing X_2 newspaper.
For X_2 : $t = b_2/S_{b2}$ = 1.7649/0.379 = 4.66 > t_{19} = 2.093 with 19 degrees of freedom for α = 0.05. Reject H_0. There is evidence that the variable X_2 newspaper advertising contributes to a model already containing X_1 radio advertising.Both variables should be included in the model.

13.30 (a) For X_1: $SSR(X_1 \mid X_2) = SSR(X_1 and X_2) - SSR(X_2) = 60 - 25 = 35$

$F = \dfrac{SSR(X_1 \mid X_2)}{MSE} = \dfrac{35}{120/18} = 5.25 > F_{U(1,18)} = 4.41$ with 1 and 18 degrees of freedom and α = 0.05. Reject H_0. There is evidence that the variable X_1 contributes to a model already containing X_2.

For X_2: $SSR(X_2 \mid X_1) = SSR(X_1 and X_2) - SSR(X_1) = 60 - 45 = 15$

$F = \dfrac{SSR(X_2 \mid X_1)}{MSE} = \dfrac{15}{120/18} = 2.25 < F_{U(1,18)} = 4.41$ with 1 and 18 degrees of freedom and α = 0.05. Do not reject H_0. There is no sufficient evidence that the variable X_2 contributes to a model already containing X_1. Since variable X_2 does not significantly contribute to a model in the presence of X_1, only variable X_1 should be included and a simple linear regression model is developed.

(b) $R^2_{Y1,2} = \dfrac{SSR(X_1 \mid X_2)}{SST - SSR(X_1 and X_2) + SSR(X_1 \mid X_2)} = \dfrac{35}{180 - 60 + 35} = 0.2258$

Holding constant the effect of variable X_2, 22.58% of the variation in Y can be explained by the variation in variable X_1.

$R^2_{Y2,1} = \dfrac{SSR(X_2 \mid X_1)}{SST - SSR(X_1 and X_2) + SSR(X_2 \mid X_1)} = \dfrac{15}{180 - 60 + 15} = 0.1111$

Holding constant the effect of variable X_1, 11.11% of the variation in Y can be explained by the variation in variable X_2.

13.31 (a) For X_1: $SSR(X_1 \mid X_2) = SSR(X_1 and X_2) - SSR(X_2) = 234 - 97 = 137$

$F = \dfrac{SSR(X_1 \mid X_2)}{SSE} = \dfrac{137}{346/13} = 5.147 > F_{U(1,13)} = 4.67$ with 1 and 13 degrees of freedom and α = 0.05. Reject H_0. There is evidence that the variable X_1 contributes to a model already containing X_2.

For X_2:

$SSR(X_2 \mid X_1) = SSR(X_1 \text{ and } X_2) - SSR(X_1) = 234 - 124 = 110$

$F = \dfrac{SSR(X_2 \mid X_1)}{SSE} = \dfrac{110}{346/13} = 4.133 > F_{U(1,13)} = 4.67$

with 1 and 13 degrees of freedom and α = 0.05. Do not reject H_0. There is no sufficient evidence that the variable X_2 contributes to a model already containing X_1. Since variable X_2 does not significantly contribute to a model in the presence of X_1, based on these results only variable X_1 should be included and a simple linear regression model is developed.

(b) $R^2_{Y1,2} = \dfrac{SSR(X_1 \mid X_2)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_1 \mid X_2)} = \dfrac{137}{580 - 234 + 137} = 0.2836$

Holding constant the effect of variable X_2, 28.36% of the variation in Y can be explained by the variation in variable X_1.

$R^2_{Y2,1} = \dfrac{SSR(X_2 \mid X_1)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_2 \mid X_1)} = \dfrac{110}{580 - 234 + 110} = 0.2412$

Holding constant the effect of variable X_1, 24.12% of the variation in Y can be explained by the variation in variable X_2.

13.32 (a) For X_1: $SSR(X_1|X_2) = SSR(X_1 \text{ and } X_2) - SSR(X_2) = 9594740.095 - 75100.326 = 9519639.769$

$F = \dfrac{SSR(X_1|X_2)}{MSE} = \dfrac{9519639.769}{8260336.027/11} = 12.677 > F_{U(1,11)} = 4.84$ with 1 and 11 degrees of freedom and α = 0.05. Reject H_0. There is evidence that the variable X_1 contributes to a model already containing X_2.

For X_2: $SSR(X_2|X_1) = SSR(X_1 \text{ and } X_2) - SSR(X_1) = 9594740.095 - 9586357.683 = 8382.4125$

$F = \dfrac{SSR(X_2|X_1)}{MSE} = \dfrac{8382.4125}{8260336.027/11} = 0.0112 < F_{U(1,11)} = 4.84$ with 1 and 11 degrees of freedom and α = 0.05. Do not reject H_0. There is no sufficient evidence that the variable X_2 contributes to a model already containing X_1. Since variable X_2 does not significantly contribute to a model in the presence of X_1, only variable X_1 should be included and a simple linear regression model is developed.

(b)

$$R^2_{Y1,2} = \frac{SSR(X_1|X_2)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_1|X_2)}$$

$$= \frac{9519639.769}{17855076.12 - 9594740.095 + 9519639.769} = 0.5354$$

Holding constant the effect of variable X_2, 53.54% of the variation in Y can be explained by the variation in variable X_1.

$$R^2_{Y2,1} = \frac{SSR(X_2|X_1)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_2|X_1)}$$

$$= \frac{8382.4125}{17855076.12 - 9594740.095 + 8382.4125} = 0.0010$$

Holding constant the effect of variable X_1, 0.10% of the variation in Y can be explained by the variation in variable X_2.

13.33 (a) For X_1: $SSR(X_1 \mid X_2) = SSR(X_1 and X_2) - SSR(X_2) = 249.125 - 36.48 = 212.65$

$F = \dfrac{SSR(X_1 \mid X_2)}{MSE} = \dfrac{212.65}{31.22} = 6.81 > F_{U(1,13)} = 4.67$ with 1 and 13 degrees of freedom and α = 0.05. Reject H_0. There is evidence that the variable X_1 contributes to a model already containing X_2.

For X_2: $SSR(X_2 \mid X_1) = SSR(X_1 and X_2) - SSR(X_1) = 249.125 - 248.704 = 0.421$

$F = \dfrac{SSR(X_2 \mid X_1)}{MSE} = \dfrac{0.421}{31.22} = 0.013 < F_{U(1,13)} = 4.67$ with 1 and 13 degrees of freedom and α = 0.05. Do not reject H_0. There is no sufficient evidence that the variable X_2 contributes to a model already containing X_1. Since variable X_2 does not significantly contribute to a model in the presence of X_1, only variable X_1 should be included and a simple linear regression model is developed.

(b) $R^2_{Y1,2} = \frac{SSR(X_1 \mid X_2)}{SST - SSR(X_1 and X_2) + SSR(X_1 \mid X_2)} = \frac{212.65}{623.73 - 249.125 + 212.65} = 0.3621$ Holdin

g constant the effect of variable X_2, 36.21% of the variation in Y can be explained by the variation in variable X_1.

$$R^2_{Y2,1} = \frac{SSR(X_2 \mid X_1)}{SST - SSR(X_1 and X_2) + SSR(X_2 \mid X_1)} = \frac{0.421}{623.73 - 249.125 + 0.421} = 0.0011$$

Holding constant the effect of variable X_1, 0.11% of the variation in Y can be explained by the variation in variable X_2

13.34 (a) For X_1: $SSR(X_1 \mid X_2) = SSR(X_1 \text{ and } X_2) - SSR(X_2) =$ 191.683 – 0.189 = 191.494

$F = \frac{SSR(X_1 \mid X_2)}{MSE} = \frac{191.494}{17.401} = 11.0048 > F_{U(1,10)} = 4.96$ with 1 and 10 degrees of freedom and α = 0.05. Reject H_0. There is sufficient evidence that the variable X_1 contributes to a model already containing X_2.

For X_2: $SSR(X_2 \mid X_1) = SSR(X_1 \text{ and } X_2) - SSR(X_1) =$ =365.692 – 109.526 = 82.157

$F = \frac{SSR(X_2 \mid X_1)}{MSE} = \frac{82.157}{17.401} = 4.72 < F_{U(1,10)} = 4.96$ with 1 and 10 degrees of freedom and α = 0.05. Do not reject H_0. There is no sufficient evidence that the variable X_2 contributes to a model already containing X_1. Since variable X_2 does not significantly contribute to a model in the presence of X_1, based on these results only variable X_1 should be included and a simple linear regression model is developed.

(b) $R^2_{Y2,1} = \frac{SSR(X_1 \mid X_2)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_1 \mid X_2)} = \frac{191.494}{365.692 - 191.683 + 191.494} = 0.5239$

Holding constant the effect of variable X_2, 52.39% of the variation in Y can be explained by the variation in variable X_1.

$$R^2_{Y2,1} = \frac{SSR(X_2 \mid X_1)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_2 \mid X_1)} = \frac{82.157}{365.692 - 191.683 + 82.157} = 0.3207$$

Holding constant the effect of variable X_1, 32.07% of the variation in Y can be explained by the variation in variable X_2.

13.35

(a) For X_1: $SSR(X_1 \mid X_2) = SSR(X_1 \text{ and } X_2) - SSR(X_2) = = 27.2241 - 11.1119 = 16.1122$

$$F = \frac{SSR(X_1 \mid X_2)}{MSE} = \frac{16.1122}{0.7825} = 20.5916$$

with 1 and 47 degrees of freedom, and p-value = 0.0000. Reject H_0. There is sufficient evidence that the variable percentage alcohol contributes to a model already containing chorides.

For X_2: $SSR(X_2 \mid X_1) = SSR(X_1 \text{ and } X_2) - SSR(X_1) = = 27.2241 - 21.8677 = 5.3564$

$$F = \frac{SSR(X_2 | X_1)}{MSE} = \frac{5.3564}{0.7825} = 0.68455$$

with 1 and 47 degrees of freedom and p-value = 0.0119. Reject H_0. There is enough evidence that the variable chorides contributes to a model already containing percentage alcohol. Since both percentage alcohol and chlorides make a significant contribution to the model in the presence of the other, the most appropriate regression model for this data set should include both percentage alcohol and chlorides.

(
b
)

$$R^2_{Y2,1} = \frac{SSR(X_1 | X_2)}{SST \square SSR(X_1 \text{ and } X_2) + SSR(X_1 | X_2)} = \frac{16.1122}{64 \square 27.2241 + 16.1122}$$

= 0.3046. Holding constant the effect of chlorides, 30.46% of the variation in quality rating can be explained by the variation in percentage alcohol.

$R^2_{Y2,1} = \frac{SSR(X_2 | X_1)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_2 | X_1)} = \frac{5.3564}{64 - 27.2241 + 5.3564}$ =$= 0.1271$. Holding constant the effect of percentage alcohol, 12.71% of the variation in quality rating can be explained by the variation in chlorides.

13.36

(a) For X_1: $SSR(X_1 | X_2) = SSR(X_1 \text{ and } X_2) - SSR(X_2) = 2,028,033 - 632,259.4 = 1,395,773.6$

$$F = \frac{SSR(X_1 | X_2)}{MSE} = \frac{1395773.6}{479,759.9/19} = 55.28 > F_{(1,19)} = 4.381$$

Reject H_0. There is evidence that the variable X_1 contributes to a model already containing X_2.

For X_2: For X_2: $SSR(X_2 | X_1) = SSR(X_1 \text{ and } X_2) - SSR(X_1) = = 2{,}028{,}033 - 1{,}216{,}940 = 811{,}093$

$F = \frac{SSR(X_2 | X_1)}{MSE} = \frac{811,093}{479,759.9} = 32.12 > F_{U(1,19)} = 4.381$ with 1 and 19 degrees of freedom 9. Reject H_0.

There is enough evidence that the variable X_2 contributes to a model already containing X_1.
Since both variables make a significant contribution to the model in the presence of the other, the most appropriate regression model for this data set should include both variables.

(b) $R^2_{Y1,2} = \frac{SSR(X_1 | X_2)}{SST \square SSR(X_1 \text{ and } X_2) + SSR(X_1 | X_2)} = \frac{1,395,773.6}{2,507,793 \square 2,028,033 + 1,395,773.6} = 0.7442$

Holding constant the effect of newspaper advertising, 74.42% of the variation in *Y* can be explained by the variation in radio advertising.

$$R^2_{Y2,1} = \frac{SSR(X_2 \mid X_1)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_2 \mid X_1)} = \frac{811,093}{2,507,793 - 2,028,033 + 811,093} = 0.6283$$

Holding constant the effect of radio advertising, 62.83% of the variation in *Y* can be explained by the variation in newspaper advertising.

13.37 (a) Holding constant the effect of X_2, the estimated mean value of the dependent variable will increase by 5 units for each increase of one unit of X_1.

(b) Holding constant the effect of X_1, the presence of the condition represented by X_2 = 1 is estimated to increase the mean value of the dependent variable by 0.5 units.

(c) t = 2.67 > t_{32} = 2.037. Reject H_0. The presence of X_2 makes a significant contribution to the model.

13.38 (a) First develop a multiple regression model using X_1 as the variable amount spent on advertising and X_2 a dummy variable with X_2 = 1 if the music has air-time on ratio. If the dummy variable coefficient is significantly different from zero, you need to develop a model with the interaction term X_1X_2 to make sure that the coefficient of X_1 is not significantly different if X_2 = 0 or X_2 = 1.

(b) If the music receives air-time on radio, the sales would be estimated to be 0.30 greater than had the same amount been spent on advertising, but without air-time on radio.

13.39 (a) $\hat{Y} = -6.456 - 199.1872X_1 + 69.4149X_2$

X_1 = Location (east = 0), X_2 = Number of rooms

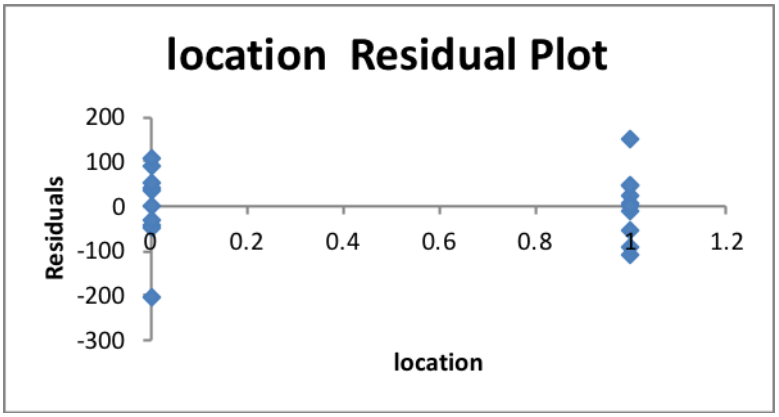
(b) Holding the effect of neighbourhood constant, for each additional room, the selling price is estimated to increase by a mean of \$69,414 000. For a given number of rooms, a western city side is estimated to decrease the selling price over an eastern city location by \$199,187.

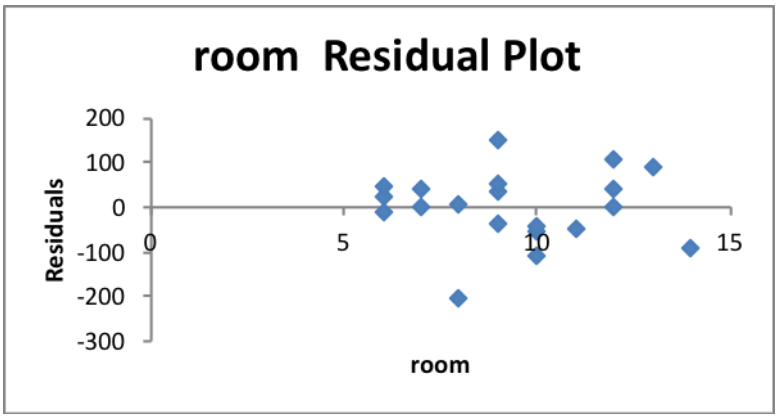
(c) $\hat{Y} = -6.456 - 199.1872(0) + 69.4149(9) = \$618{,}278$

$554.4951 \le Y_{X=X_i} \le 682.0602$

$428.005 \le \mu_{Y|X=X_i} \le 808.5503$

(d)





The models appear to be adequate from residual analysis.

(e) $F = 71.1155 > F_{U(2,17)} = 3.59$. Reject H_0. There is evidence of a relationship between selling price and the two independent variables.

(f) For X_1: $t = -4.622 < -t_{17} = -2.110$. Reject H_0. Location makes a significant contribution and should be included in the model.
For X_2: $t = 7.5143 > t_{17} = 2.110$. Reject H_0. Number of rooms makes a significant contribution and should be included in the model.

(g) $-290.1025 \le \beta_1 \le -108.272, 49.9251 \le \beta_2 \le 88.9047$

(h) $R^2 = \dfrac{SSR}{SST} = 0.9451$. So, 94.51% of the variation in selling price is explained by variation in the two independent variables.

(i) $R^2_{adj} = 0.8932$

(j) $R^2_{Y1,2} = \dfrac{SSR(X_1 \mid X_2)}{SST - SSR(X_1 and X_2) + SSR(X_1 \mid X_2)} = 0.5569$. Holding constant the effect of number of rooms, 55.69% of selling price can be explained by location.

$R^2_{Y2,1} = \dfrac{SSR(X_2 \mid X_1)}{SST - SSR(X_1 and X_2) + SSR(X_2 \mid X_1)} = 0.7686$. Holding constant the effect of location, 76.86% of selling price can be explained by number of rooms.

(k) The slope of selling price with number of rooms is the same regardless of whether the house is located in the east or west of Melbourne.

(l) $\hat{Y} = -354.5434 - 274.4005X_1 + 102.5660X_2 - 48.2717X_1X_2$.
For X_1X_2 the p-value = 0.01. Reject H_0. It is likely that the size of houses and location interact, and the two independent variables are not independent of each other.

(m) The regression results indicate that the two variables model can be used as both independent variables do contribute to an explanation of selling price. However, there is a risk of lack of independence between the two independent variables and the model could be run with just number of rooms (the higher partial coefficient of determination).

13.40 (a) $\hat{Y} = 16.5613 + 0.1529X_1 - 6.9086X_2$

(b) Holding constant effect of traffic lights for each additional one-unit increase in traffic volume, accidents increase by 0.1529 units.

For a given number of traffic volume, an intersection with traffic lights is estimated to decrease the accidents over a no traffic lights location by a mean of 6.9 units.

(c) $\hat{Y} = 16.5613 + 0.1529(18) - 6.9086(1) = 12.4053$

$1.3684 \le Y_{X=X_i} \le 2.4156$

$1.6880 \le \mu_{Y|X=X_i} \le 2.096$

(d) The model appears to be adequate from residual analysis.

(e) $F = 4.5799 > F_{2,12} = 3.89$. Reject H_0. There is evidence of a relationship between accidents and two independent variables.

(f) For X_1: $t = 1.0538 < 2.179$. Do not reject H_0. Likely that traffic volume does not contribute significantly to explaining accidents and could be excluded from the model.
For X_2: $t = 2.6160 > 2.179$. Reject H_0. Traffic lights do make significant contribution to the explanation of accidents and should be included in the model.

(g) $-0.1633 \le \beta_1 \le 0.4691, -12.6083 \le \beta_2 \le -1.2087$

(h) $R^2 = \frac{SSR}{SST} = 0.6579$. So, 65.79% of the variation in accidents are explained by variation in the two independent variables.

(i) $R^2_{adj} = 0.3384$. Note that the sample size is small so adjusted R^2 is much lower than R^2.

(j) $R^2_{Y1,2} = \frac{SSR(X_1 \mid X_2)}{SST - SSR(X_1 and X_2) + SSR(X_1 \mid X_2)} = 0.0847$. Holding traffic lights constant, 8.47% of the variation in accidents can be explained by traffic volume.

$R^2_{Y2,1} = \frac{SSR(X_2 \mid X_1)}{SST - SSR(X_1 and X_2) + SSR(X_2 \mid X_1)} = 0.3676$. Holding traffic volume constant 36.76% of the variation in accidents can be explained by traffic lights.

(k) The assumption for the slope on volume of traffic is that there is no change in the variable traffic lights.

(l) $\hat{Y} = 4.2182 + 0.6533X_1 + 8.3687X_2 - 0.6344X_1X_2$
For X_1X_2 the p-value = 0.0716. Do not reject H_0 at 5% significance level. It is unlikely that traffic volume and traffic lights interact.

(m) The regression result indicates that the model should only contain one independent variable traffic lights. Traffic volume is not significantly related to accidents. Because of low partial R^2 of 36.76% for traffic lights there must be some other variables to explain the variation in accidents that have not been measured.

13.41

PHStat output:

Regression Statistics						
Multiple R	0.5068					
R Square	0.25					

	68					
Adjusted R Square	0.2415					
Standard Error	1.0509					
Observations	100					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	37.0257	18.5129	16.76177	0.0000	
Residual	97	107.1343	1.1045			
Total	99	144.1600				
	Coeffic	*Standar*	*t Stat*	*P-value*	*Lower 9*	*Upper 9*

	i e n t s	*d E r r o r*			*5 %*	*5 %*
Intercept	0.9342	0.8770	1.0652	0.2894	-0.80644	2.67447
alcohol	0.4652	0.0820	5.67622	0.00000	0.30255	0.62278
Type of Wine	-0.2577	0.2102	-1.2258	0.22232	-0.67495	0.15955

(a) $\hat{Y} = 0.9342 + 0.4652\,X - 0.25\,77\,X$

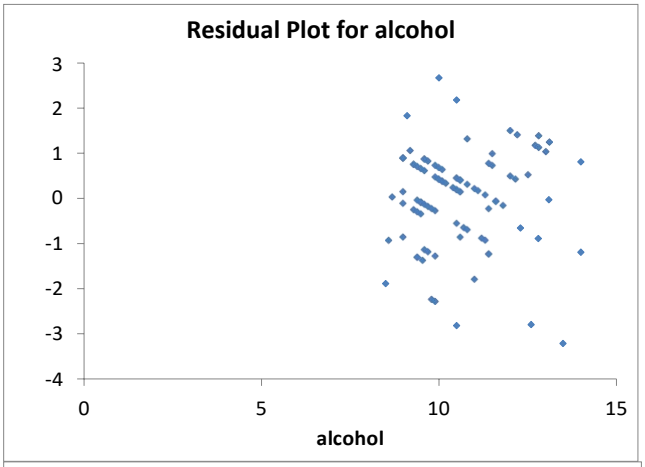
(b) Holding constant the effect of the type of wine, for each additional % increase in alcohol content, wine quality is estimated to increase by a mean of 0.4652. For a given amount of alcohol content, a white wine is estimated to have a 0.2577 higher mean quality than a red wine.

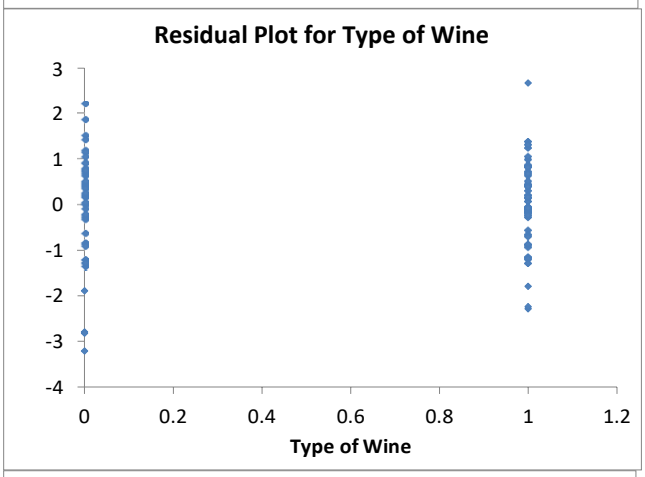
(c) $\hat{Y} = 0.9342 + 0.4652(10) - 0.2577(1) = 5.3283$

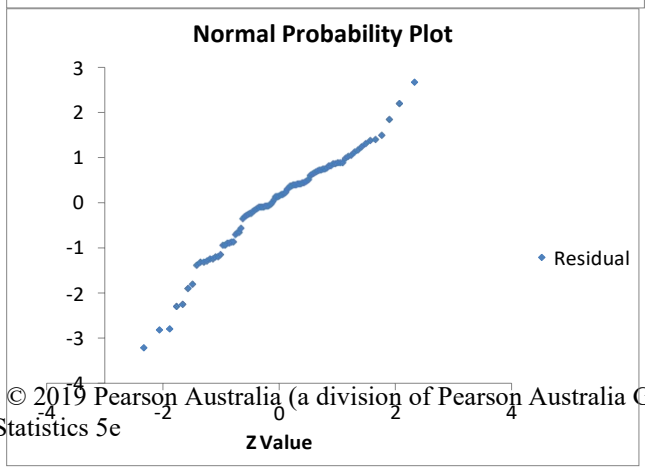
$3.2196 \le Y_{X=X_i} \le 7.4371$

$5.0184 \le \mu_{Y|X=X_i} \le 5.6382$

PHStat output:







(d) Based on a residual analysis, there is not any obvious pattern in the residual plots but the normal probability plot indicates departure from the normality assumption.

(e) F_{STAT} = 16.7617 with a p-value = 0.0000. Reject H_0. There is evidence of a relationship between quality and percentage of alcohol and the type of wine

(f) For X_1: t_{STAT} = 5.6762 with a *p*-value = 0.0000. Reject H_0. Alcohol content makes a significant contribution and should be included in the model.

For X_2: t_{STAT} = -1.2258 with a *p*-value = 0.2232. Do not reject H_0. The type of wine does not make a significant contribution and should not be included in the model. Only alcohol content should be kept in the model.

(g) $0.3025 \le \beta_1 \le 0.6278$, $-0.6749 \le \beta_2 \le 0.1595$

(h) The slope here takes into account the effect of the other predictor variable, type of wine, while the solution for Problem 13.4 did not.

(i) $R^2 = 0.2568$. So, 25.68% of the variation in quality can be explained by variation in alcohol content and variation in the type of wine.

(j) $R^2_{adj} = 0.2415$

0.2568 while $R^2 = 0.3417$ in Problem 13.16 (a).

(k) $R^2_{Y1,2} = 0.2493$. Holding constant the effect of wine type, 24.93% of the variation in quality can be explained by variation in alcohol content.

(l) $R^2_{Y2,1} = 0.0153$. Holding constant the effect of alcohol content, 1.53% of the variation in quality can be explained by variation in wine type.

(m) The slope of alcohol content is the same regardless of whether the wine is red or white.

(n) PHStat output:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	7.7000	0.2000	38.5000	0.0000	7.2999	8.1001
Alcohol	0.4636	0.0816	5.6762	0.0000	0.3001	0.6271
Wine	-0.2500	0.2040	-1.2258	0.2232	-0.6581	0.1581

	n t s					
Intercept	1.6780	1.1448	1.4658	0.1460	-0.5944	3.9503
alcohol	0.3947	0.1076	3.6667	0.0004	0.1810	0.6083
Type of Wine	-2.0309	1.7669	-1.1494	0.2533	-5.5382	1.4764
alcohol X Type of Wine	0.1678	0.1660	1.0107	0.3147	-0.1617	0.4973

Since the t_{STAT} for the significance of X_1X_2 has a p-value = 0.3147, do not reject H_0.

s a contribution to the model.

(o) The one-variable model should be used.

(p) Only the alcohol content is significant in predicting the wine quality.

13.42 (a) $\hat{Y} = 23.8016 - 0.5052X_1 - 2.955X_2 + 0.4528X_1X_2$

Where X_1: Unemployment rate, X_2: Stock market return, X_1 X_2 = interaction between unemployment rate and stock market return.

For X_1 X_2: the p-value = 0.3187 > 0.05. Do not reject H_0. There is not enough evidence that the interaction term makes a contribution to the model.

(b) Since there is not enough evidence of any interaction effect between unemployment rates and stock market return, the model in problem 13.4 should be used.

13.43

(a) $\hat{Y} = -1293.3105 + 43.6600\,X + 56.9335\,X - 0.8430\,X$.

where X_1 = radio advertisement, X_2 = newspaper advertisement, $X_3 = X_1X_2$

For X_1X_2: the p-value is 0.0018 < 0.05. Reject H_0. There is enough evidence that the interaction term makes a contribution to the model.

(b) Since there is enough evidence of an interaction effect between radio and newspaper advertisement, the model in this problem should be used.

13.44 (a) $\hat{Y} = 186.595 + 185.701X_1 - 0.879X_2 + 0.293X_1X_2$

Where X_1: GDP, X_2: population density, X_1 X_2 = interaction between GDP and population density

For X_1 X_2: the p-value = 0.685 > 0.05. Do not reject H_0. There is not enough evidence that the interaction term makes a contribution to the model.

(b) Since there is not enough evidence of any interaction effect between GDP and population density, the model in problem 13.5 should be used.

13.45 (a) $\hat{Y} = 29.08 + 0.0001X_1 + 0.37X_2 - 0.0000008X_1X_2$

Where X_1:GDP/capita, X_2: CPI, X_1 X_2 = interaction between GDP/capita and CPI

For X_1 X_2: the p-value = 0.0499 < 0.05. Reject H_0. There is enough evidence that the interaction term makes a contribution to the model.

(b) Since there is enough evidence of an interaction effect between GDP/capita and CPI, the model including this interaction term should be used.

13.46 (a)

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-63.98133	16.7997	-3.80855	0.0008
Proficiency	1.12558	0.15899	7.08688	0.0000
Classroom	-22.28877	4.31544	-5.16499	0.0000
Online	8.	4.310	1.	0.

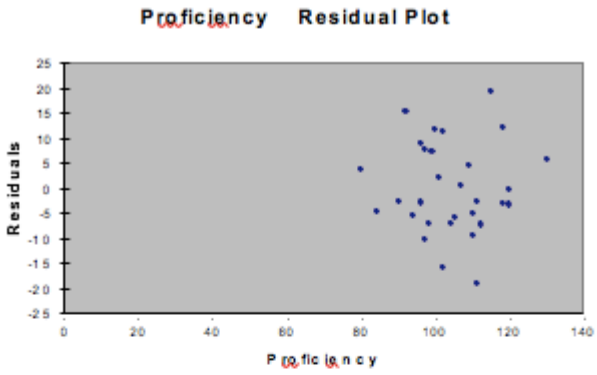
	0 8 8 0	3	8 7 6 5	0 7 1 9

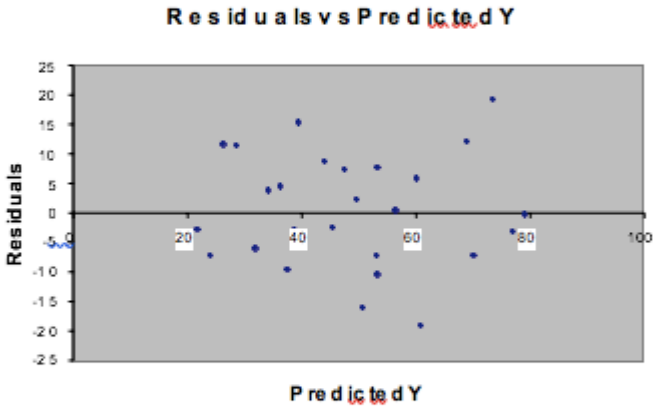
where X_1 = proficiency exam, X_2 = classroom dummy, X_3 = online dummy

(b) Holding constant the effect of training method, for each point increase in proficiency exam score, the end-of-training exam score is estimated to increase by a mean of 1.1258 points. For a given proficiency exam score, the end-of-training exam score of a trainee who has been trained by the classroom method will have an estimated mean score that is 22.2887 points below a trainee that has been trained using the courseware app method. For a given proficiency exam score, the end-of-training exam score of a trainee who has been trained by the online method will have an estimated mean score that is 8.0880 points above a trainee that has been trained using the courseware app method

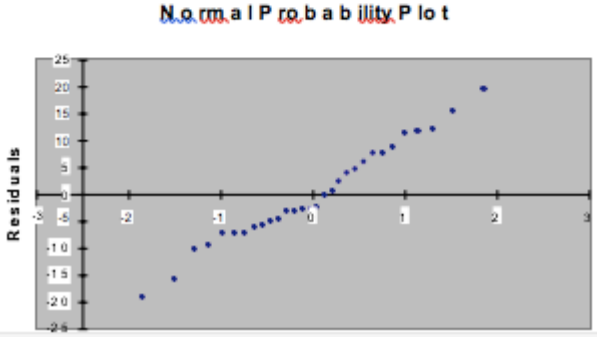
(c) $\hat{Y} = -63.9813 + 1.1258\left(100\right) = 48.5969$

(d)





There appears to be a quadratic effect from the residual p



(e) F_{STAT} = 31.77 with 3 and 26 degrees of freedom. The p-value is virtually 0. Reject H_0 at 5% level of significance. There is evidence of a relationship between end-of-training exam score and the independent variables.

(f) For X_1: t_{STAT} = 7.0868 and the p-value is virtually 0. Reject H_0. Proficiency exam score makes a significant contribution and should be included in the model.
For X_2: t_{STAT} = -5.1649 and the p-value is virtually 0. Reject H_0. The classroom dummy makes a significant contribution and should be included in the model.
For X_3: t = 1.876 and the p-value = 0.07186. Do not reject H_0. There is not sufficient evidence to conclude that there is a difference in the online method and the courseware app method on the mean end-of-training exam scores.
Base on the above result, the regression model should use the proficiency exam score and the classroom dummy variable.

(g) $0.7992 \le \beta_1 \le 1.4523$, $-31.1591 \le \beta_2 \le -13.4182$, $-0.7719 \le \beta_3 \le 16.9480$

(h) $R^2_{Y123} = 0.7857$. 78.57% of the variation in the end-of-training exam score can be explained by the proficiency exam score and the various training methods.

(i) $R^2_{adj} = 0.7610$

(j) $R^2_{Y1,23} = 0.6589$. Holding constant the effect of training method, 65.89% of the variation in end-of-training exam score can be explained by variation in the proficiency exam score.

$R^2_{Y2,13} = 0.5064$. Holding constant the effect of proficiency exam score, 50.64% of the variation in end-of-training exam score can be explained by the difference between classroom and courseware app methods.

$R^2_{Y3,12} = 0.1193$. Holding constant the effect of proficiency exam score, 11.93% of the variation in end-of-training exam score can be explained by the difference between online and courseware app methods.

(k) The slope of end-of-training exam score with proficiency score is the same regardless of the training method.

(l) Let $X_4 = X_1X_2$, $X_5 = X_1X_3$.

$H_0: \beta_4 = \beta_5 = 0$ There is no interaction among X_1, X_2 and X_3.

H_1: At least one of β_4 and β_5 is not zero.

There is interaction among at least a pair of X_1, X_2 and X_3.

$$STAT \quad F = \frac{SSR(X_5, X_1 \mid X_2, X_3, X_4)}{MSE(X_1, X_2, X_3, X_4, X_5)} = \frac{[SSR(X_2, X_3, X_4, X_5) - SSR(X_2, X_3)]/2}{MSE(X_1, X_2, X_3, X_4, X_5)}$$

= H_0. The interaction terms do not make a significant contribution to the model.

0.81 (m) The regression model should use the proficiency exam score and the classroom dummy variable.

13.47 $VIF = \frac{1}{1-0.35} = 1.54$

13.48 $VIF = \frac{1}{1-0.5} = 2.0$

13.49 Collinearity is potentially significant and you need to consider whether all the independent variables are needed.

13.50 $R_1^2 = 0.1767$, $VIF = \frac{1}{1-0.1767} = 1.215$

$R_2^2 = 0.1767$, $VIF = \frac{1}{1-0.1767} = 1.215$

There is no reason to suspect the existence of collinearity.

13.51 $R_1^2 = 0.0035$, $VIF = \frac{1}{1-0.0035} = 1.0035$

$R_2^2 = 0.0035$, $VIF = \frac{1}{1-0.0035} = 1.0035$

There is no reason to suspect the existence of collinearity.

13.52 $R_1^2 = 0.759$, $VIF = \frac{1}{1-0.759} = 4.15$

$R_2^2 = 0.759$, $VIF = \frac{1}{1-0.759} = 4.15$

There is no reason to suspect the existence of collinearity.

13.53 $R_1^2 = 0.184$, $VIF = \frac{1}{1-0.184} = 1.225$

$R_2^2 = 0.184$, $VIF = \frac{1}{1-0.184} = 1.225$

There is no reason to suspect the existence of collinearity.

13.54 $VIF = \frac{1}{1-0.3181} = 1.466$

There is no reason to suspect the existence of collinearity.

13.55 In the case of the simple linear regression model, the slope b_1 represents the change in the estimated mean of Y per unit change in X and does not take into account any other variables. In the multiple linear regression model, the slope b_1 represents the change in the estimated mean of Y per unit change in X_1, taking into account the effect of all the other independent variables.

ion model involves a simultaneous test of whether any of the independent variables are significant. Testing the contribution of each independent variable tests the contribution of the independent variable after accounting for the effect of the other independent variables in the mo partial determination measures the proportion of variation in *Y* explained by a particular *X* variable holding constant the effect of other independent variables in the model. The coefficient of multiple determination measures the proportion of variation in *Y* explained by all the *X* variables included in the model.

13.58 The coefficient of partial determination measures the proportion of variation in *Y* explained by a particular *X* variable holding constant the effect of other independent variables in the model. The coefficient of multiple determination measures the proportion of variation in *Y* explained by all the *X* variables included in the model.

13.59 To test for collinearity, we can use *VIF* values. If a set of independent variables is uncorrelated, each VIF_j is equal to 1. If the set is highly correlated, then a VIF_j might even exceed 10. Snee recommends using alternatives to least square regression if maximum VIF_j exceeds 5.

13.60 A dummy variable will be included to represent a categorical independent variable. One category is coded as 0 and the other category of the variable is coded as 1.

13.61 You test whether the interaction term in the regression model makes a significant contribution to the regression model. If it makes a significant contribution, then it should be included in the model. We can't interpret them separately since the two variables interact. The effect of an independent variable on the response variable *Y* is dependent on the value of a second variable.

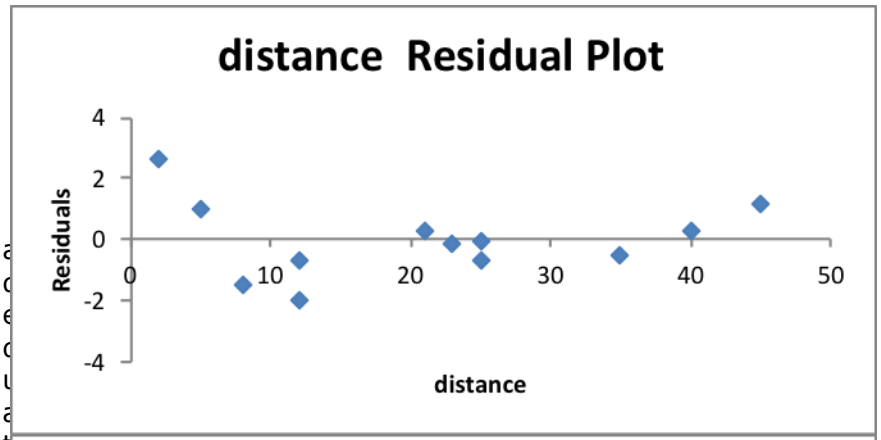
13.62 Start with an F test of the overall significance of the model. If you are able to reject this null hypothesis you should test the contribution of each individual x variable using t tests.

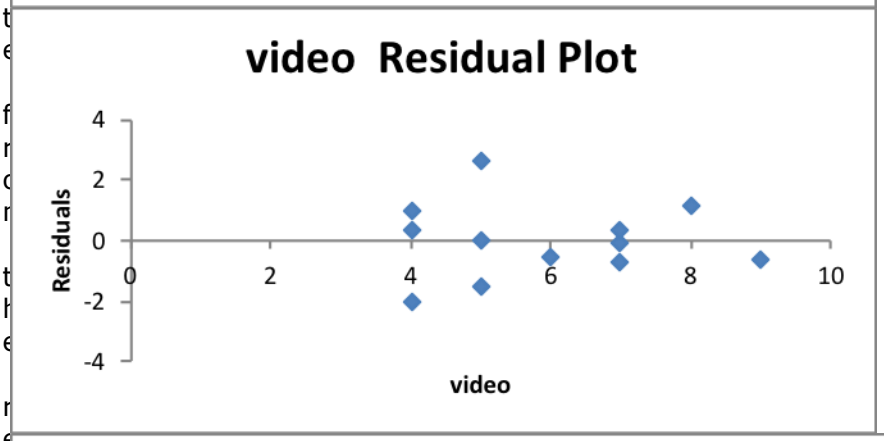
13.63 (a) $\hat{Y} = 16.0906 - 0.1437X_1 - 1.0947X_2$

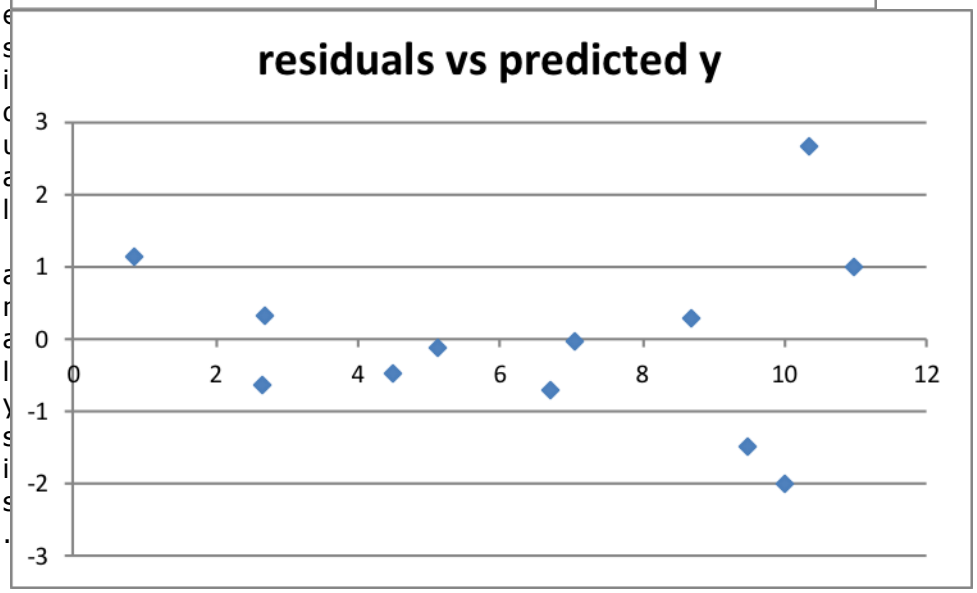
(b) Holding constant the effect of video access, for each one km increase in distance from campus, lecture attendance decreases by 0.1437 classes per session. Holding constant the effect of distance to campus, for each access to video lecture recordings, lecture attendance decreases by 1.0947 classes per session.

(c) $\hat{Y} = 16.0906 - 0.1437(5) - 1.0947(3) = 12.0880$

(d)







The model appears to be adequate from the residual analysis.

(e) $F = 34.09 > F_{2,9} = 4.26$. Reject H_0. There is evidence of a relationship between lecture attendance and the two independent variables.

(f) p-value is virtually zero. The probability of obtaining an F test statistic of 34.09 or larger is virtually zero if H_0 is true.

(g) $R^2 = 0.8834$. So, 88.34% of the variation in lecture attendance is explained by the variation in the two independent variables.

(h) 0.8575

(i) For X_1: $t = -3.9087 < -2.2622$. Reject H_0. There is sufficient evidence that distance to campus
For X_2: $t = -3.5972 < -2.2622$. Reject H_0. there is suficient evidence that video lecture access

(j) For X_1: p-value = 0.0036, which means the probability of obtaining a t statistic of −3.9087 or s
For X_2: p-value = 0.0058, which means the probability of obtaining a t statistic of −3.9087 or s

(k) VIF = 1.5248. The measure of collinearity is below 5 and indicates the two independent variab

13.64

(a)

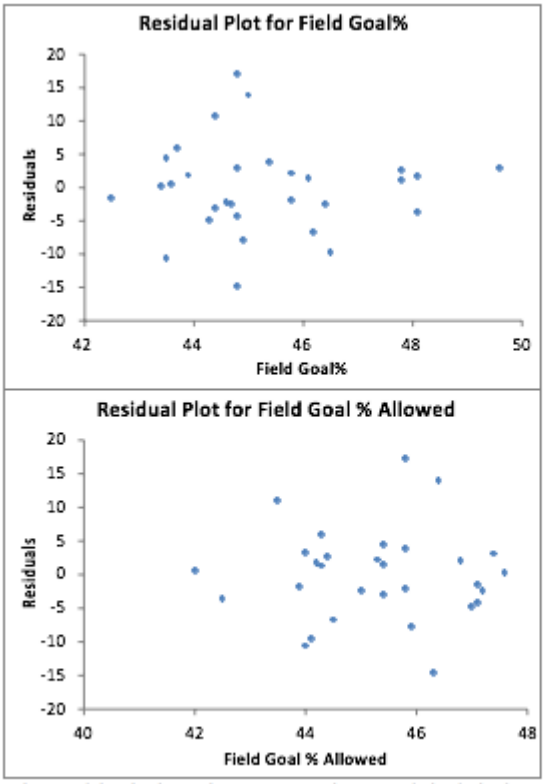
$\hat{Y} = 61.8354 + 3.8250X_1 - 4.2881X_2$

where X_1 = field goal %, X_2 = opponent field goal %

(b) For a given opponent field goal %, each increase of 1% in field goal % increases the estimated mean number of wins by 3.8250. For a given field goal %, each increase of 1% in opponent field goal % decreases the estimated mean number of wins by 4.2881.

(c) $\hat{Y} = 61.8354 + 3.8250(45) - 4.2881(44) = 45.2825$

(d)



The residual plots do not reveal potential violations of the assumptions.

(e) H_0: $\beta_1 = \beta_2 = 0$ H_1: Not all $\beta_j = 0$ for $j = 1, 2$

$F = MSR/MSE = 34.0700$

p-value is essentially 0 < 0.05. Reject H_0 at 5% level of significance. There is evidence of a significant linear relationship between number of wins and the two explanatory variables.

(f) p-value is virtually 0. The probability of obtaining an F test statistic equal to or larger than 34.0700 is is true. virtually 0 if H_0

(g) $R^2 = SSR/SST = 0.7162$. So, 71.62% of the variation in number of wins can be explained by variation in field goal % for the team and opponent field goal %.

(h) $R^2_{adj} = 0.6952$.

(i) For X_1: $t_{STAT} = b_1 / s_b = 4.4077$ and p-value = 0.0001< 0.05, reject H_0. There is evidence that the variable X_1 contributes to a model already containing X_2.

For X_2: $t_{STAT} = b_2 / s_b = -4.3061$ and p-value = 0.0002 < 0.05, reject H_0. There is evidence that the variable X_2 contributes to a model already containing X_1. Both variables X_1 and X_2 should be included in the model.

(j) For X_1: p-value = 0.0001. The probability of obtaining a t test statistic that differs from 0 by 4.4077 or more in either direction is 0.01% if X_1 is insignificant.
For X_2: p-value is virtually 0. The probability of obtaining a t test statistic that differs from 0 by 4.3061 or more in either direction is 0.02% if X_2 is insignificant.

(k) $R^2_{Y1,2} = 0.4185$. Holding constant opponent field goal %, 41.85% of the variation in number of wins can be explained by variation in field goal% for the team.

$R^2_{Y2,1} = 0.4071$. Holding constant the effect of field goal % for the team, 40.71% of the

variation in number of wins can be explained by variation in opponent field goal %.

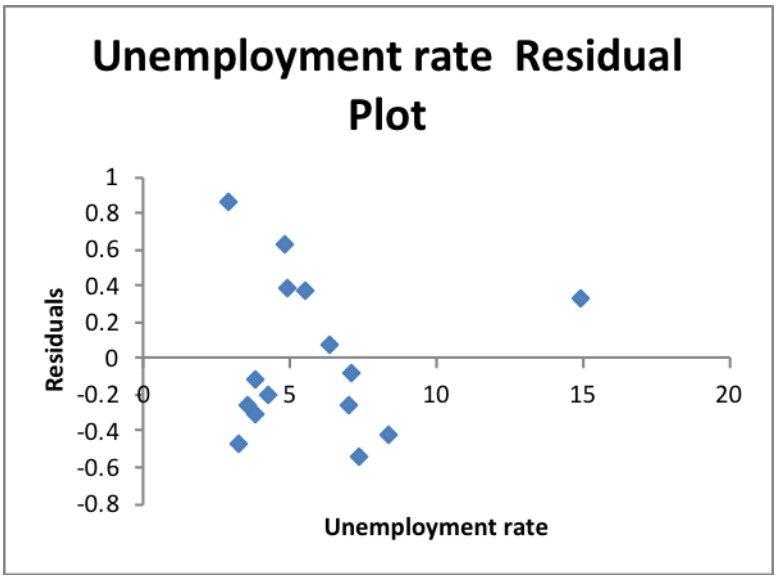
13.65 (a) $\hat{Y} = 0.809 - 0.008X_1 + 0.019X_2$

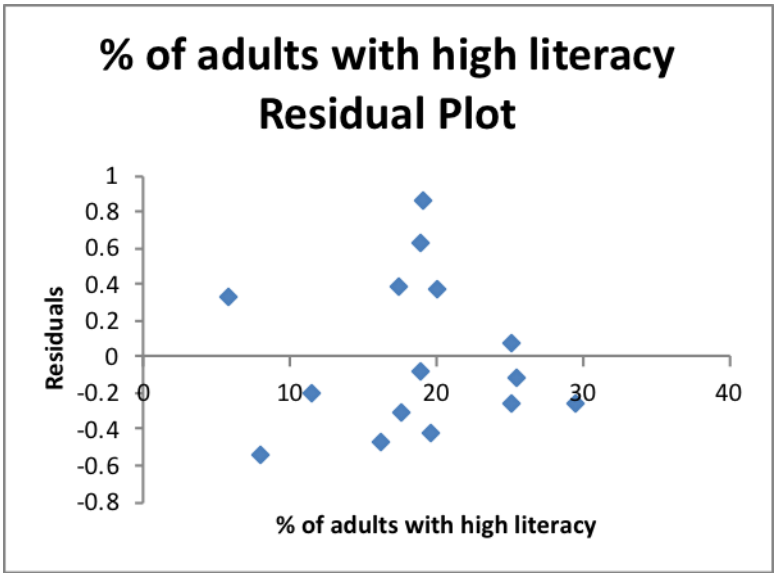
(b) Holding constant the effect of unemployment rate for a 1% increase in the percentage of adults with high literacy, robbery rate decreases by 0.008 units. Holding constant the effect of percentage of adults with high literacy for a 1% increase in unemployment, robbery rate increases by 0.019 units.

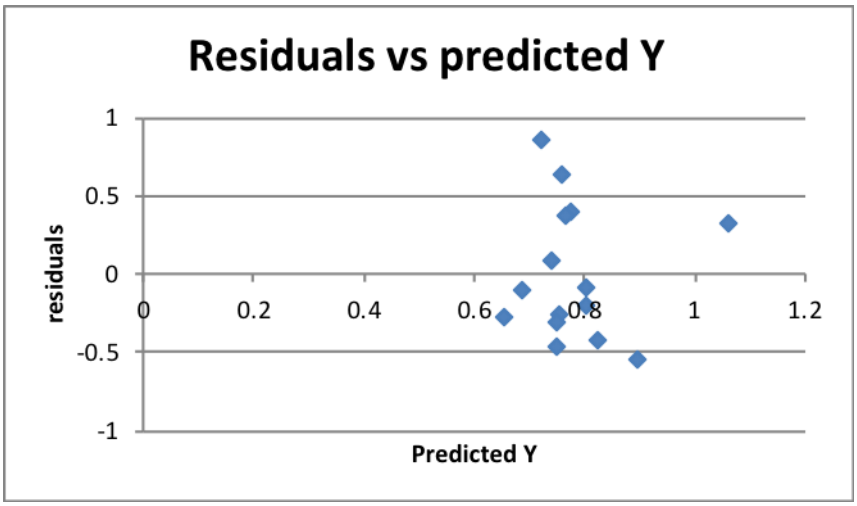
(c) $\hat{Y} = 0.809 - 0.008(20) + 0.019(7) = 0.782$

(d) R^2 = 0.048. So, 4.8% of the variation in robbery rate is explained by the variation in the two independent variables (unemployment and percentage of adults with high literacy).

(e)







There is some indication in the residual plot of a non-linear relationship.

(f) $F = 0.301 < F_{(2,12)} = 3.89$. Do not reject H_0. There is no evidence of a relationship between robbery rate and two independent variables (unemployment rate and percentage of adults with high literacy).

(g) $-0.06 \le \beta_1 \le 0.04, -0.083 \le \beta_2 \le 0.122$

(h) For X_1: $t = -0.35 > -2.179$. Do not reject H_0. Percentage of adults with high literacy does not contribute significantly to an explanation of robbery rate.
For X_2: $t = 0.418 < 2.179$. Do not reject H_0. It is likely that unemployment rate does not contribute to an explanation of sales. So, there must be other variables that better explained the robbery rate.

(i) $VIF = 1.35$.
The measure of collinearity is below 5 and indicates the two independent variables are independent of each other.

13.66 Excel output:

Regression Statistics	
Multiple R	0.7520
R Square	0.5655
Adjusted R Square	0.4785
Standard Error	0.9

	136					
Observations	19					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	16.2908	5.43033	6.50633	0.00049	
Residual	15	12.5192	0.83446			
Total	18	28.8100				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-18.69155	7.9789	-2.34266	0.03334	-35.6982	-1.68488
Viscosity	0.	0.008	1.	0.	-0.	0.

	0.121	2	4.8171	1.5911	0.053	0.2996
Pressure	0.0844	0.0414	2.04159	0.0592	-0.0037	0.17266
Plate Gap	0.5000	0.1379	3.62711	0.0025	0.2062	0.79388

The r^2 of the multiple regression is 0.5655. So 56.66% of the variation in tear rating can be explained by the variation of viscosity, pressure, and plate gap on the bag-sealing equipment. The *F* test statistic for the combined significant of viscosity, pressure, and plate gap on the bag- sealing equipment is 6.5063 with a *p*-value of 0.0049. Hence, at a 5% level of significance, there is enough evidence to conclude that viscosity, pressure, and plate gap on the bag-sealing equipment affect tear rating.
The *p*-value of the *t* test for the significance of viscosity is 0.1591, which is larger than 5%. Hence, there is not sufficient evidence to conclude that viscosity affects tear rating holding constant the effect of pressure and plate gap on the bag-sealing equipment.
The *p*-value of the *t* test for the significance of pressure is 0.0592, which is also larger than 5%. There is not enough evidence to conclude that pressure affects tear rating at 5% level of significance holding constant the effect of viscosity and plate gap on the bag-sealing equipment. The *p*-value of the *t* test for the significance of plate gap is 0.0025, which is smaller than 5%.
There is enough evidence to conclude that plate gap affects tear rating at 5% level of significance holding constant the effect of viscosity and pressure.

13.67 (a) $\hat{Y} = -5603.64 + 20028.74X_1 + 22075.85X_2$

(b) Holding constant the effect of economics major, for each one-point increase in WAM, the wage is estimated to increase in a mean of $20,028.74. For a given WAM, the wage of students with an economics major will have an estimated mean wage that is 22,075.85 above a wage of students with no economic major.

(c) $\hat{Y} = -5603.64 + 20028.74(3.5) + 22075.85(1) = \$86572.80$

(d) $F = 47.99 > F_{(2,7)} = 4.74$. There is evidence of a significant relationship between wage and WAM and economics major.

(e) *p*-value = 0. The *F* statistic is statistically significant.

(f) $R^2_{adj} = 0.9126$

13.68 (a) $\hat{Y} = 0.5610 - 0.1633X_1 + 0.2614X_2$

(b) Holding constant the effect of growth in exports, for each one-unit increase in manufacturing growth there is a mean decrease of 0.1633 in GDP.
Holding constant the effect of manufacturing growth for each one-unit increase in export growth, there is a mean increase of 0.2614 in GDP.

(c) $\hat{Y} = 0.5610 - 0.1633(1.2) + 0.2614(6.3) = 2.012$

(d) $R^2 = 0.3327$. So, 33.27% of the variation in GDP is explained by the variation in the two independent variables (manufacturing growth and exports).

(e) $R^2_{adj} = -0.009$. The sample size is too small for the R^2_{adj} to be reliable.

(f) For X_1: $t = -0.1710 > -3.5$. Do not reject H_0. It is likely that manufacturing growth does not contribute significantly to an explanation of GDP.
For X_2: $t = 0.689 < 3.5$. Do not reject H_0. It is likely that export growth does not contribute to an explanation of GDP.

(g) The regression does not explain GDP growth using either of the independent variables, and the sample size is possibly too small. The model theory must be strong enough too allow for the small sample size in that group of countries to make up a significant proportion of the population countries in that group.

13.69 (a) $\hat{Y} = 56.7688 - 0.7770X_1 - 1.5486X_2$

(b) Holding constant the effect school life expectancy for each one-year increase in the average age of females having their first child, the percentage of youth not in employment, education or training (NEET) decreases by 0.78 percentage points.
Holding constant the effect of mean birth age, for each one-year increase in school life expectancy, the percentage of youth NEET decreases by 1.55 years.

(c) $\hat{Y} = 56.7688 - 0.7770(25) - 1.5486(15) = 14.12$

(d) $F = 9.29 > F_{2,27} = 3.35$. Reject H_0. There is enough evidence of a significant relationship between the percentage of youth NEET and the two independent variables.

(e) $R^2 = 0.4077$. So, 40.77% of the variation in percentage of youth NEET is explained by the variation in the two independent variables.

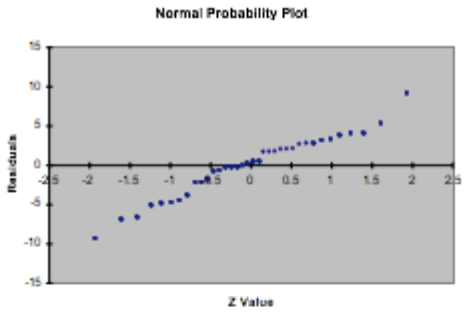
(f) $R^2_{adj} = 0.3342\ 0.3638$

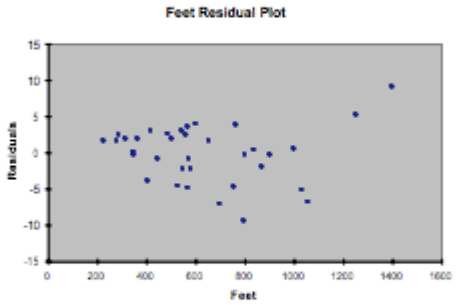
(g) For X_1: $t = -1.4340 > -2.0518$. Do not reject H_0. It is likely that average age of first-child birth does not contribute significantly to an explanation of percentage of youth NEET.
For X_2: $t = -1.5486 < -2.0518$. Reject H_0. It is likely that school life expectancy does contribute to an explanation of percentage of youth NEET at 5% significance level.

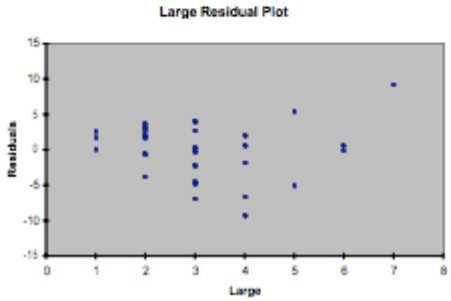
13.70 (a) $\hat{Y} = -3.9152 + 0.0319X_1 + 4.2228X_2$ where X_1 = amount of cubic feet moved and X_2 = number of pieces of large furniture

b) Holding constant the number of pieces of large furniture, for each additional cubic foot moved, the mean labor hours are estimated to increase by 0.0319. Holding constant the amount of cubic feet moved, for each additional piece of large furniture, the mean labor hours are estimated to increase by 4.2228.

(c) $\hat{Y} = -3.9152 + 0.0319(500) + 4.2228(2) = 20.4926$







(d)

Based on a residual analysis, the errors appear to be normally distributed. The equal cont.
variance assumption might be violated because the variances appear to be larger around
the center region of both independent variables. There might also be violation of the linearity assumption. A model with quadratic terms for both independent variables might be fitted.

(e) $F_{STAT} = 228.80$, p-value is virtually 0. Since p-value < 0.05, reject H_0. There is evidence of a significant relationship between labor hours and the two independent variables (the amount of cubic feet moved and the number of pieces of large furniture).

(f) The p-value is virtually 0. The probability of obtaining a test statistic of 228.80 or greater is virtually 0 if there is no significant relationship between labor hours and the two independent variables (the amount of cubic feet moved and the number of pieces of large furniture).

(g) $R^2_{Y12} = 0.9327$. 93.27% of the variation in labor hours can be explained by variation in the amount of cubic feet moved and the number of pieces of large furniture.

(h) $R^2_{adj} = 0.9287$

(i) For X_1: $t_{STAT} = 6.9339$, p-value is virtually 0. Reject H_0. The amount of cubic feet moved makes a significant contribution and should be included in the model.
For X_2: $t_{STAT} = 4.6192$, p-value is virtually 0. Reject H_0. The number of pieces of large furniture makes a significant contribution and should be included in the model.
Based on these results, the regression model with the two independent variables should be used.

(j) For X_1: $t_{STAT} = 6.9339$, p-value is virtually 0. The probability of obtaining a sample that will yield

a test statistic farther away than 6.9339 is virtually 0 if the amount of cubic feet moved does not make a significant contribution holding the effect of the number of pieces of large furniture constant.

For X_2: $t_{STAT} = 4.6192$, p-value is virtually 0. The probability of obtaining a sample that will yield a test statistic farther away than 4.6192 is virtually 0 if the number of pieces of large furniture does not make a significant contribution holding the effect of the amount of cubic feet moved constant.

(k) $.0226 \leq \beta_1 \leq 0.0413$. We are 95% confident that the mean labor hours will increase by somewhere between 0.0226 and 0.0413 for each additional cubic foot moved holding constant the number of pieces of large furniture. In Problem 13.44, we are 95% confident that the mean labor hours will increase by somewhere between 0.0439 and 0.0562 for each additional cubic foot moved regardless of the number of pieces of large furniture.

(l) $R^2_{Y1,2} = 0.5930$. Holding constant the effect of the number of pieces of large furniture, 59.3% of the variation in labor hours can be explained by variation in the amount of cubic feet moved.

$R^2_{Y2,1} = 0.3927$. Holding constant the effect of the amount of cubic feet moved, 39.27% of the variation in labor hours can be explained by variation in the number of pieces of large furniture.

(m) Both the number of cubic feet moved and the number of large pieces of furniture are useful in predicting the labor hours, but the cubic feet removed is more important.

13.71 Excel output:

Regression Statistics	
Multiple R	0.5707
R Square	0.3257
Adjusted R Square	0.2615
Standard Error	11.3

	6 2 1					
Observations	24					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	1309.5192	654.7596	5.0718	0.01600	
Residual	21	2711.0379	129.0970			
Total	23	4020.5571				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	18	71.48	0.	0.	-1	16

	.2892	49	25558	80006	30.37182	6.95022
Die Temperature	0.5976	0.463 9	1.2883	0.2117	-0.3671	1.5622
Die Diameter	-13.5108	4.638 6	-2.9127	0.0083	-23.1572	-3.8644

The r^2 of the multiple regression is 0.3257. So 32.57% of the variation in unit density can be explained by the variation of die temperature and die diameter.

The *F* test statistic for the combined significant of die temperature and die diameter is 5.0718 with a *p*-value of 0.0160. Hence, at a 5% level of significance, there is enough evidence to conclude that die temperature and die diameter affect unit density.

The *p*-value of the *t* test for the significance of die temperature is 0.2117, which is larger than 5%. Hence, there is not sufficient evidence to conclude that die temperature affects unit density holding constant the effect of die diameter.

The *p*-value of the *t* test for the significance of die diameter is 0.0083, which is smaller than 5%. There is enough evidence to conclude that die diameter affects unit density at 5% level of significance holding constant the effect of die temperature.

Excel output after dropping die temperature:

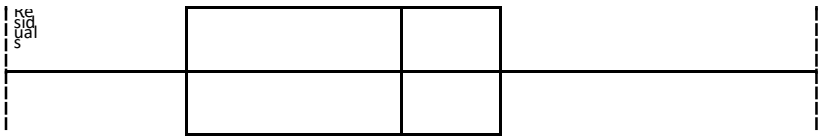
Regression Statistics						
Multiple R	0.5219					
R Square	0.2724					
Adjusted R Square	0.2393					
Standard Error	11.5					

	312					
Observations	24					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	1095.2557	1095.2557	8.2370	0.0089	
Residual	22	2925.3014	132.9682			
Total	23	4020.5571				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	107.	16.6438	6.48	0.00	73.4	142.

	9 2 6 7		4 5	0 0	0 9 5	4 4 3 9
Die Diam eter	- 1 3 . 5 1 0 8	4. 70 76	- 2 . 8 7 0 0	0 . 0 0 8 7	- 2 3 . 2 7 3 8	- 3 . 7 4 7 9

Die diameter still remains statistically significant at the 5% level of significance. Hence, only die cont. diameter needs to be used in the model.

The residual plot suggests that the equal variance assumption is likely violated.



The normal probability plot and the boxplot both suggest that the normal distribution assumption might be violated.

None of the observations have a Cook's $D_i > F_\alpha = 0.7155$ with d.f. = 2 and 22. Hence, using the Studentized deleted residuals, hat matrix diagonal elements and Cook's distance statistic together, there is insufficient evidence for removal of any observation from the model.

13.71 (a) 1 = 2917.00
2 = 3801
3 = 4384.5
4 = 8745
5 = 10064
The better choice would be 5.

(b) The problem in (a) is that the regression model is a constant linear growth trend and even if the values in (a) are within the range of the data, it would follow that sales would increase as both salary and commission increase. However, this may not be realistic in a real-world setting.