

- a. From the scatterplot, there is a definite curvature in the relation between Durability and Concentration. A straight-line model would not appear to be appropriate.
- b. The coefficient of determination,  $R^2$ , measures the strength of the linear (straight-line) relation only. A straight-line model does not adequately describe the relation between Durability and Concentration. This is indicated by the small percentage of the variation, 11.6%, in the values of Durability explained by the model containing just a linear relation with Concentration.

## Chapter 12: Multiple Regression and the General Linear Model

12.1 a.  $y_j = \mu_{i_j} + \epsilon_j$ , with  $i_j = 1, 2, \dots, 12$

$$b. \quad x_1 = \begin{cases} 1 & \text{if } V_1 = \text{1st Level} \\ 0 & \text{if } \text{Otherwise} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if } V_2 = \text{1st Level} \\ 0 & \text{if } \text{Otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if } V_2 = \text{2nd Level} \\ 0 & \text{if } \text{Otherwise} \end{cases} \quad x_4 = \begin{cases} 1 & \text{if } V_3 = \text{1st Level} \\ 0 & \text{if } \text{Otherwise} \end{cases}$$

$$y_j = \beta_o + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \beta_4 x_{4j} + \epsilon_j, \text{ with } j = 1, \dots, n$$

12.3 The model is given here:

$$y_i = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \beta_7 x_{7i} + \beta_8 x_{8i} + \beta_9 x_{9i} + \epsilon_i$$

with  $x_1 = x_1, x_2 = x_2, x_3 = x_3, x_4 = x_1^2, x_5 = x_2^2,$   
 $x_6 = x_3^2, x_7 = x_1 x_2, x_8 = x_1 x_3, x_9 = x_2 x_3$

12.5 a. For the model of the  $i$ th observation in the experiment:

$$y_i = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i,$$

$\mu_{jk}$ , the mean of the population consisting of Treatment # $j$  and Location # $k$ ;  $j = 1, 2, 3$ ;  $k = 1, 2$  :

Location		
Treatment	1	2
1	$\mu_{11} = \beta_o$	$\mu_{12} = \beta_o + \beta_3$
2	$\mu_{21} = \beta_o + \beta_1$	$\mu_{22} = \beta_o + \beta_1 + \beta_3$
3	$\mu_{31} = \beta_o + \beta_2$	$\mu_{32} = \beta_o + \beta_2 + \beta_3$

We can thus interpret the  $\beta$ 's as follows:

$\beta_o = \mu_{11}$ , mean of TRT #1, LOC #1

$\beta_1 = \mu_{21} - \mu_{11}$  or  $\beta_1 = \mu_{22} - \mu_{12}$ ,

difference of mean of TRT #2 and TRT #1 at a given Location

$\beta_2 = \mu_{31} - \mu_{11}$  or  $\beta_2 = \mu_{32} - \mu_{12}$ ,

difference of mean of TRT #3 and TRT #1 at a given Location

$\beta_3 = \mu_{12} - \mu_{11}$  or  $\beta_3 = \mu_{22} - \mu_{21}$  or  $\beta_3 = \mu_{32} - \mu_{31}$ ,

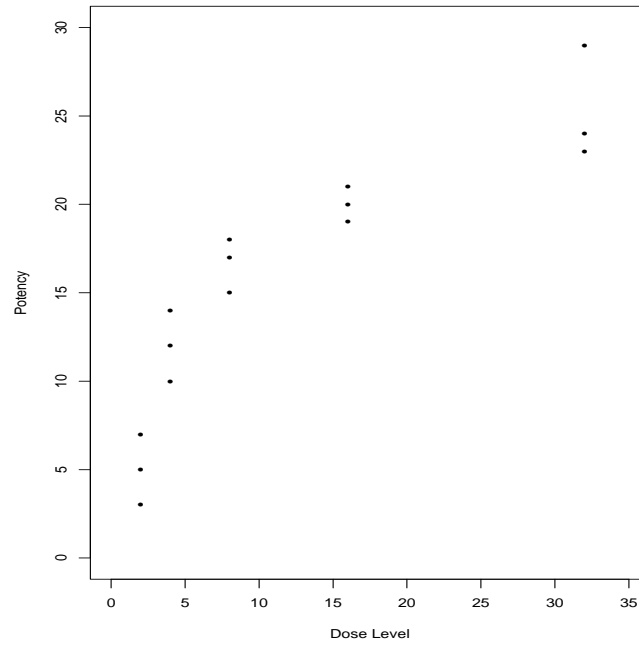
difference of mean of LOC #2 and LOC #1 for a given Treatment

b.  $\mu_{22} - \mu_{32} = (\beta_o + \beta_1 + \beta_3) - (\beta_o + \beta_2 + \beta_3) = \beta_1 - \beta_2$

Yes, the difference is the same for Location #1.

12.7 a. A scatterplot of the data is given here:

Plot of Drug Potency versus Dose Level



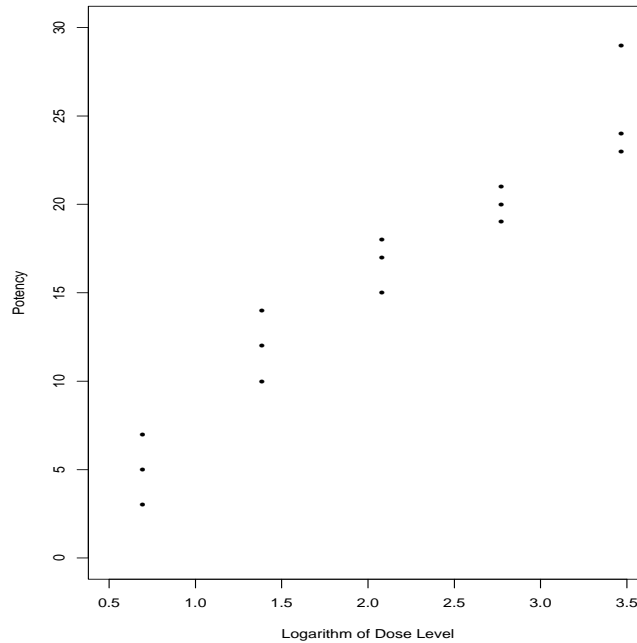
b.  $\hat{y} = 8.667 + 0.575x$

12.8 a. The logarithm of the dose levels are given here:

Dose Level (x)	2	4	8	16	32
$\ln(x)$	0.693	1.386	2.079	2.773	3.466

A scatterplot of the data is given here:

Plot of Drug Potency versus Log Dose Level



b.  $\hat{y} = 1.2 + 7.021\ln(x)$

12.9 a.  $\hat{y} = 326.39 + 136.10 * Promo - 61.18 * Devel - 43.70 * Research$

b.  $s_{\epsilon} = \sqrt{MS(Residual)} = \sqrt{656.811614} = 25.628$

12.17 a.  $MS(Regression) = 159.67$  and  $MS(Residual) = 7.00$

b.  $F = 22.81$

c.  $p - value = Pr(F_{3,8} \geq 22.81) < 0.0001$

d. In testing  $H_o : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$  versus  $H_a : \text{at least one } \beta_i \neq 0$ , the p-value provides strong evidence to reject  $H_o$  and conclude that the independent variables x,w, and v, as a group, have at least some degree of relationship with the dependent variable, y.

e. A 95% C.I. for  $\beta_1$  is given by  $\hat{\beta}_1 \pm (t_{.025,8})(SE(\hat{\beta}_1)) \Rightarrow 5.000 \pm (2.306)(6.895) \Rightarrow (-10.90, 20.90)$

12.20 The t-value is  $t_{.05,19} = 1.729 \Rightarrow 90\%$  C.I.'s for  $\beta_i$  are given by:

Air Miles: (0.0841, 0.5003)

Population: (1.2302, 1.8319)

12.21 a.  $\hat{y} = 7.20439 + 1.36291 METAL + 0.30588 TEMP + 0.01024 WATTS - 0.00277 METXTEXP$

b. The results of the various t-tests are given here:

$H_o$	$H_a$	T.S. $t$	Conclusion
$\beta_o = 0$	$\beta_o \neq 0$	$t = 0.41$	$p - value = 0.6855$ Fail to Reject $H_o$
$\beta_1 = 0$	$\beta_1 \neq 0$	$t = 1.47$	$p - value = 0.1559$ Fail to Reject $H_o$
$\beta_2 = 0$	$\beta_2 \neq 0$	$t = 0.19$	$p - value = 0.8522$ Fail to Reject $H_o$
$\beta_3 = 0$	$\beta_3 \neq 0$	$t = 2.16$	$p - value = 0.0427$ Reject $H_o$
$\beta_4 = 0$	$\beta_4 \neq 0$	$t = -0.04$	$p - value = 0.9717$ Fail to Reject $H_o$

c.  $t_{.025,20} = 2.086 \Rightarrow 95\%$  C.I. on  $\beta_4$  is given by  
 $-0.00277 \pm (2.086)(0.07722) \Rightarrow (-0.164, 0.158)$

12.23 a.  $R^2 = 0.6978$

b.  $F = \frac{[SSReg.,Complete - SSReg.,Reduced]/(k-g)}{SSResidual,Complete/[n-(k+1)]} = 3.12$   
with  $df = 2, 20 \Rightarrow p - value = Pr(F_{2,20} \geq 3.12) = 0.066 \Rightarrow$   
Fail to reject  $H_o$ .

12.25 In the complete model, we want to test  $H_o : \beta_1 = \beta_2 = 0$  versus  $H_a : \beta_1 \neq 0$  and/or  $\beta_2 \neq 0$ .  
The F-statistic has the form:

$$F = \frac{[SSReg.,Complete - SSReg.,Reduced]/(k-g)}{SSResidual,Complete/[n-(k+1)]} = 24.84$$

with  $df = 2, 17 \Rightarrow p - value = Pr(F_{2,17} \geq 24.84) < 0.0001 \Rightarrow$   
Reject  $H_o$ .

12.28 a.  $\hat{y} = 50.0195 + 6.64357x_1 + 7.3145x_2 - 1.23143x_1^2 - 0.7724x_1x_2 - 1.1755x_2^2$

b.  $\hat{y} = 70.31 - 2.676x_1 - 0.8802x_2$

c. For the Complete model:  $R^2 = 86.24\%$   
For the Reduced model:  $R^2 = 58.85\%$

d. In the complete model, we want to test

$H_o : \beta_3 = \beta_4 = \beta_5 = 0$  versus  $H_a : \text{at least one of } \beta_3, \beta_4, \beta_5 \neq 0$ .

The F-statistic has the form:

$$F = \frac{[SSReg.,Complete - SSReg.,Reduced]/(k-g)}{SSResidual,Complete/[n-(k+1)]} = 9.29$$

with  $df = 3, 14 \Rightarrow p - value = Pr(F_{3,14} \geq 9.29) = 0.0012 \Rightarrow$   
Reject  $H_o$ .

12.29 The predicted y-value at  $x=3, w=1, v=6$  is  $\hat{y} = 33.000$  with 95% P.I.: (21.788, 44.212).

12.30 a. For the second order model, the 95% P.I. are  
(54.7081, 65.1439) for  $x_1 = 3.5$  and  $x_2 = 0.35$   
(57.0829, 67.6529) for  $x_1 = 3.5$  and  $x_2 = 2.5$

b. For the first order model, the 95% P.I. are  
(50.0280, 65.6986) for  $x_1 = 3.5$  and  $x_2 = 0.35$   
(51.0525, 66.4345) for  $x_1 = 3.5$  and  $x_2 = 2.5$

- 12.31 a.  $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1x_2 + \beta_5x_1x_3 + \epsilon$ , where  
 $x_1 = \log(\text{dose})$

$$x_2 = \begin{cases} 1 & \text{if Product B} \\ 0 & \text{if Products A or C} \end{cases} \quad x_3 = \begin{cases} 1 & \text{if Product C} \\ 0 & \text{if Products A or B} \end{cases}$$

$\beta_0 = y$ -intercept for Product A regression line

$\beta_1 =$  slope for Product A regression line

$\beta_2 =$  difference in  $y$ -intercepts for Products A and B regression lines

$\beta_3 =$  difference in  $y$ -intercepts for Products A and C regression lines

$\beta_4 =$  difference in slopes for Products A and B regression lines

$\beta_5 =$  difference in slopes for Products A and C regression lines

- b.  $y = \beta_0 + \beta_1x_1 + \beta_2x_1x_2 + \beta_3x_1x_3 + \epsilon$

- 12.32 a. The Minitab output for fitting the Complete and Reduced models is given here:

Regression Analysis: y versus x1, x2, x3, x1\*x2, x1\*x3

The regression equation is

$$y = 7.31 + 3.30 x1 - 2.15 x2 - 4.35 x3 - 1.50 x1*x2 - 2.28 x1*x3$$

Predictor	Coef	SE Coef	T	P
Constant	7.3072	0.2103	34.75	0.000
x1	3.3038	0.2186	15.11	0.000
x2	-2.1548	0.2974	-7.25	0.000
x3	-4.3486	0.2974	-14.62	0.000
x1*x2	-1.5004	0.3092	-4.85	0.003
x1*x3	-2.2795	0.3092	-7.37	0.000

S = 0.3389      R-Sq = 98.8%      R-Sq(adj) = 97.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	55.293	11.059	96.30	0.000
Residual Error	6	0.689	0.115		
Total	11	55.982			

Regression Analysis: y versus x1, x2, x3

The regression equation is

$$y = 6.59 + 2.04 x1 - 1.30 x2 - 3.05 x3$$

Predictor	Coef	SE Coef	T	P
Constant	6.5894	0.5131	12.84	0.000
x1	2.0438	0.3519	5.81	0.000
x2	-1.3000	0.6679	-1.95	0.087
x3	-3.0500	0.6679	-4.57	0.002

S = 0.9446      R-Sq = 87.2%      R-Sq(adj) = 82.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	48.844	16.281	18.25	0.001
Residual Error	8	7.138	0.892		
Total	11	55.982			

In the complete model:  $y = \beta_o + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1x_2 + \beta_5x_1x_3 + \epsilon$ , the test of equal slopes is a test of the hypotheses:

$$H_o : \beta_4 = 0, \beta_5 = 0 \text{ versus } H_a : \beta_4 \neq 0 \text{ and/or } \beta_5 \neq 0$$

Under  $H_o$ , the reduced model becomes  $y = \beta_o + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \epsilon$

$$F = 28.08 \Rightarrow p\text{-value} = Pr(F_{2,6} \geq 28.08) = 0.0009 \Rightarrow$$

b. Reject  $H_o$

c. In the complete model, a test of equal intercepts is a test of the hypotheses:

$$H_o : \beta_2 = 0, \beta_3 = 0 \text{ versus } H_a : \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0$$

Under  $H_o$ , reduced model becomes  $y = \beta_o + \beta_1x_1 + \beta_4x_1x_2 + \beta_5x_1x_3 + \epsilon$

Obtain the SS's from the reduced model and then conduct the F-test as was done in part a.

12.33 a. For testing  $H_o : \beta_1 = 0$  versus  $H_a : \beta_1 \neq 0$ , the p-value for the output is  $p\text{-value} = 0.0427$ . Thus, at the  $\alpha = 0.05$  level we can reject  $H_o$

b. From the output,  $\hat{p}(24) = 0.765$  with 96% C.I. (0.437, 0.932).

12.35 a.  $\hat{y} = -1.320 + 5.550 \text{ EDUC} + 0.885 \text{ INCOME} + 1.925 \text{ POPN} - 11.389 \text{ FAMSIZE}$   
(57.98) (2.702) (1.308) (1.371) (6.669)

b.  $R^2 = 96.2\%$  and  $s_\epsilon = 2.686$

12.37 a.  $R^2 = 94.2\%$

b. In the complete model, we want to test

$$H_o : \beta_2 = \beta_3 = 0 \text{ versus } H_a : \text{at least one of } \beta_2, \beta_3 \neq 0.$$

The F-statistic has the form:

$$F = 1.89$$

$$\text{with } df = 2, 7 \Rightarrow p\text{-value} = Pr(F_{2,7} \geq 1.89) = 0.2206 \Rightarrow$$

Fail to reject  $H_o$ .

12.39 a.  $F = \frac{0.894477/4}{(1-0.894477)/(43-5)} = 80.53$  with  $df=4,38$ . The  $p\text{-value} = Pr(F_{4,38} \geq 80.53) < 0.0001 \Rightarrow$  Reject  $H_o : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

b. Using  $\alpha = 0.01$ , none of the p-values for testing  $H_o : \beta_i = 0$  versus  $H_a : \beta_i \neq 0$ , .0999, .0569, .5954, and .3648, respectively, are less than 0.01.

12.41 a. The regression model is

$$\hat{y} = -16.8198 + 1.47019x_1 + .994778x_2 - .0240071x_3 - .01031x_4 - .000249574x_5$$

$$s_\epsilon = 3.39011$$

b. Test  $H_o : \beta_3 = 0$  versus  $H_a : \beta_3 \neq 0$ . From output,  $t = -1.01$  with  $p\text{-value}=0.3243$ . Thus, there is not substantial evidence that the variable  $x_3 = x_1x_2$  adds predictive value to a model which contains the other four independent variables.

12.43 a.  $\hat{y} = 0.8727 + 2.548 \text{ SIZE} + 0.220 \text{ PARKING} + 0.589 \text{ INCOME}$   
(1.946) (1.201) (0.155) (0.178)

b. The interpretation of coefficients is given here:

Coefficient	Interpretation
$\hat{\beta}_0 = y\text{-Intercept}$	The estimated average daily sales for the population of stores having 0 Size, 0 Parking, 0 Income
$\hat{\beta}_1 = \hat{\beta}_{SIZE}$	The estimated change in average Daily Sales per unit change in SIZE, for fixed values of PARKING and INCOME
$\hat{\beta}_2 = \hat{\beta}_{PARKING}$	The estimated change in average Daily Sales per unit change in PARKING, for fixed values of SIZE and INCOME
$\hat{\beta}_3 = \hat{\beta}_{INCOME}$	The estimated change in average Daily Sales per unit change in INCOME, for fixed values of SIZE AND PARKING

c.  $R^2 = 0.7912$  and  $s_\epsilon = 0.7724$

12.44 The results of the various tests are given here:

$H_o$	$H_a$	T.S.	$p\text{-value}$	Conclusion
$\beta_1 = \beta_2 = \beta_3 = 0$	at least one $\neq 0$	$F = 15.16$	0.0002	Reject $H_o$
$\beta_o = 0$	$\beta_o \neq 0$	$t = 0.449$	0.662	Fail to Reject $H_o$
$\beta_1 = 0$	$\beta_1 \neq 0$	$t = 2.122$	0.055	Fail to Reject $H_o$
$\beta_2 = 0$	$\beta_2 \neq 0$	$t = 1.418$	0.182	Fail to Reject $H_o$
$\beta_3 = 0$	$\beta_3 \neq 0$	$t = 3.310$	0.006	Reject $H_o$

12.45 a.  $\hat{y} = 102.708 - .833 \text{ PROTEIN} - 4.000 \text{ ANTIBIO} - 1.375 \text{ SUPPLEM}$

b.  $s_\epsilon = 1.70956$

c.  $R^2 = 90.07\%$

12.46 a. When PROTEIN=15%, ANTIBIO=1.5%, SUPPLEM=5%,

$$\hat{y} = 102.708 - .83333(15) - 4.000(1.5) - 1.375(5) = 77.333$$

c. The 95% C.I. on the mean value of TIME when PROTEIN=15%, ANTIBIO=1.5%, SUPPLEM=5% is given on the output: (76.469, 78.197)

12.47 a.  $\hat{y} = 89.8333 - 0.83333 \text{ PROTEIN}$

b.  $R^2 = 0.5057$

c. In the complete model, we want to test

$$H_o : \beta_2 = \beta_3 = 0 \text{ versus } H_a : \text{at least one of } \beta_2, \beta_3 \neq 0.$$

The F-statistic has the form:

$$F = 27.84$$

with  $df = 2, 14 \Rightarrow p\text{-value} = Pr(F_{2,14} \geq 27.84) < 0.0001 \Rightarrow \text{Reject } H_o.$

There is substantial evidence to conclude that at least one of  $\beta_2, \beta_3 \neq 0$ .

12.52 SAS output is given here:

The REG Procedure  
 Model: MODEL1  
 Dependent Variable: y SALARY

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	39.31705	13.10568	13.11	<.0001
Error	63	62.95698	0.99932		
Corrected Total	66	102.27403			

Root MSE	0.99966	R-Square	0.3844
Dependent Mean	29.36418	Adj R-Sq	0.3551
Coeff Var	3.40435		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	25.53758	0.64279	39.73	<.0001
x1	NUMEXPL	1	0.00389	0.00172	2.27	0.0269
x2	MARGIN	1	0.09567	0.03652	2.62	0.0110
x3	IPCOST	1	0.21657	0.06917	3.13	0.0026

- $\hat{y} = 25.53758 + 0.00389 \text{ NUMEXPL} + 0.09567 \text{ MARGIN} + 0.21657 \text{ IPCOST}$
- The F-statistic from the AOV on the output is  $F = 13.11$  with  $p\text{-value} < 0.0001$ . There is highly significant evidence that the independent variables as a group provide some predictive value for estimating salary.
- Using the  $p\text{-values}$  from the output, all three independent variables have relatively small p-values (0.0269, 0.0110, 0.0026). Thus, each of the three independent variables provide added predictive value to a model containing only two of the independent variables.

- 12.53
- $R^2 = 0.3844 = 38.44\%$
  - The SAS output is given here:

The SAS System

Dependent Variable: y SALARY

Analysis of Variance

Source	DF	Sum of	Mean	F Value	Pr > F
		Squares	Square		
Model	1	3.66167	3.66167	2.41	0.1251
Error	65	98.61236	1.51711		
Corrected Total	66	102.27403			

Root MSE	1.23171	R-Square	0.0358
Dependent Mean	29.36418	Adj R-Sq	0.0210
Coeff Var	4.19461		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	29.08407	0.23484	123.85	<.0001
x1	NUMEXPL	1	0.00326	0.00210	1.55	0.1251

$R^2$  has decreased dramatically to 0.0358=3.58%.

c. In the complete model, we want to test

$H_o : \beta_2 = \beta_3 = 0$  versus  $H_a : \text{at least one of } \beta_2, \beta_3 \neq 0$ .

The F-statistic has the form:

$$F = \frac{[39.31706 - 3.66167]/(3-1)}{62.95698/[67-4]} = 17.84$$

with  $df = 2, 63 \Rightarrow p\text{-value} = Pr(F_{2,63} \geq 17.84) < 0.0001 \Rightarrow \text{Reject } H_o$ .

There is substantial evidence to conclude that at least one of  $\beta_2, \beta_3 \neq 0$ . Based on the F-test, omitting MARGIN and/or IPCOST from the model would substantially changed the fit of the model. Dropping MARGIN and IPCOST from the model will result in a large decrease in the predictive value of the model.

12.54 SAS output is given here:

The SAS System

Pearson Correlation Coefficients, N = 67

	SALARY	NUMEXPL	MARGIN	IPCOST
SALARY	1.00000	0.18922	0.50447	0.50727
NUMEXPL	0.18922	1.00000	0.00884	-0.10725
MARGIN	0.50447	0.00884	1.00000	0.53134
IPCOST	0.50727	-0.10725	0.53134	1.00000

Only the correlation between MARGIN and IPCOST, 0.53134, is of a moderate size. The other two correlations among the independent variables is very small in magnitude (.10725 and .00884). Therefore, collinearity does not appear to be a problem.

- 12.58 a. Holding any three of the four predictor variables constant, we would expect that a unit increase in the remaining variable would result in an increase in sales. Thus, we would expect that each of the four partial slopes would be positive.
- b. SAS output is given here:

Source		DF	Sum of Squares	Mean Square	F Value	Pr > F
Model		4	17785	4446.15952	116.68	<.0001
Error		47	1790.93411	38.10498		
Corrected Total		51	19576			
Root MSE			6.17292	R-Square	0.9085	
Dependent Mean			112.96923	Adj R-Sq	0.9007	
Coeff Var			5.46425			

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	-10.08918	7.35341	-1.37	0.1766
x1	TITLES	1	0.44914	0.25412	1.77	0.0836
x2	FOOTAGE	1	0.30836	0.18508	1.67	0.1024
x3	IBMBASE	1	0.06956	0.05110	1.36	0.1799
x4	APLBASE	1	-0.01260	0.08495	-0.15	0.8827

The partial slopes are all positive with the exception of the partial slope for APLBASE. The negative sign on the partial slope for APLBASE should not be of great concern because the standard error for the coefficient is much large in magnitude than the estimate of the coefficient. Therefore, the independent variable, APLBASE, provides very little additional predictive value given the other three independent variables in the model.

- c. With  $t_{0.025,47} = 2.012$  and  $SE(\hat{\beta}_{TITLES}) = 0.25412$ , a 95% C.I. for the coefficient associated with TITLES is given by  $0.44914 \pm (2.012)(0.25412) \Rightarrow (-0.062, 0.960)$ .
- 12.59 a. To test  $H_o : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$  versus  $H_o : \text{at least one } \beta_i \neq 0$ , we can use the  $F$ -test from the SAS output:  $F = 116.68$  with  $p\text{-value} < 0.0001$ . Yes, we can reject

$H_o$  and conclude that at least one of the independent variables has predictive value for SALES.

- b. The p-values for test  $H_o : \beta_i = 0$  versus  $H_a : \beta_i \neq 0$  for  $i = 1, 2, 3, 4$  are 0.0836, 0.1024, 0.1799, and 0.8827. All of these values are relatively large (greater than  $\alpha = 0.05$ ). Therefore, none of the independent variables add significant predictive value to a model that already contains the other three independent variables.

12.60 SAS output is given here:

```

The SAS System

The CORR Procedure

Pearson Correlation Coefficients, N = 52

      SALES      TITLES      FOOTAGE      IBMBASE      APLBASE
SALES      1.00000      0.94905      0.92648      0.92941      0.92476
TITLES      0.94905      1.00000      0.95623      0.96593      0.97277
FOOTAGE      0.92648      0.95623      1.00000      0.91200      0.93608
IBMBASE      0.92941      0.96593      0.91200      1.00000      0.94629
APLBASE      0.92476      0.97277      0.93608      0.94629      1.00000

```

All of the correlations between pairs of independent variables are very large (greater than 0.9). Thus the collinearity problem is very severe for this regression modelling. All four variables are related to the size of the store. For a store that is growing rapidly, all four of the independent variables would tend to increase in a similar fashion. Note also since this is time series data on a given store, the 52 values of the five variables would likely to be autocorrelated and hence violate the independence condition that is required in using the regression techniques for testing and model fitting.

- 12.61 From the model we have  $R^2 = 0.9085$ . The correlation between SALES and TITLES is 0.94905. It has square  $(0.94905)^2 = 0.9007$ . This corresponds to the  $R^2$  for the one variable model relating SALES to just TITLES. Thus, the one variable model has nearly the same  $R^2$  as the four variable model. The SAS output for the one variable model is given here:

```

The SAS System

Dependent Variable: y SALES

Analysis of Variance

Source          DF          Sum of          Mean
                DF          Squares          Square  F Value  Pr > F
Model              1          17632          17632  453.55  <.0001
Error             50      1943.77196      38.87544
Corrected Total   51          19576

```

Root MSE	6.23502	R-Square	0.9007
Dependent Mean	112.96923	Adj R-Sq	0.8987
Coeff Var	5.51922		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	-10.88845	5.87976	-1.85	0.0700
x1	TITLES	1	0.84633	0.03974	21.30	<.0001

In the complete model, we want to test

$H_o : \beta_2 = \beta_3 = \beta_4 = 0$  versus  $H_a : \text{at least one of } \beta_2, \beta_3, \beta_4 \neq 0$ .

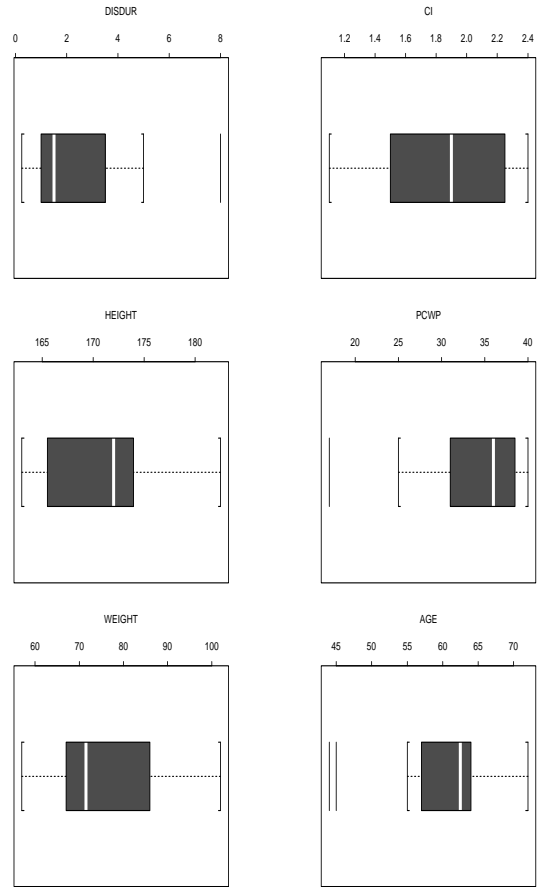
The F-statistic has the form:

$$F = \frac{[17785 - 17632] / (4 - 1)}{1790.93411 / [52 - 5]} = 1.34$$

with  $df = 3, 47 \Rightarrow p\text{-value} = Pr(F_{3,47} \geq 1.34) = 0.2727 \Rightarrow \text{Fail to reject } H_o$ .

There is not substantial evidence to conclude that at least one of  $\beta_2, \beta_3, \beta_4 \neq 0$ . Based on the F-test, omitting FOOTAGE, IBMBASE, and APLBASE from the model would not substantially changed the fit of the model. Dropping FOOTAGE, IBMBASE, and APLBASE from the model will not result in a large decrease in the predictive value of the model. These variables are so highly correlated with TITLES that they essentially add no additional predictive value once SALES is modelled by TITLES. Also, there is the possibility that the autocorrelation is confusing the issue. To truly relate SALES to the four independent variables we would need data from 52 different computer stores during at fixed time period.

- 12.63 a. The box plots of the data are given here:



b. The scatterplots of the data are given here:

