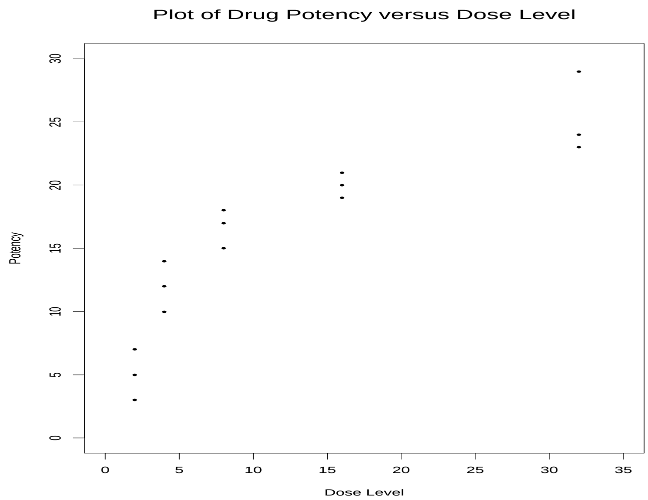
Chapter 12: Multiple Regression

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



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

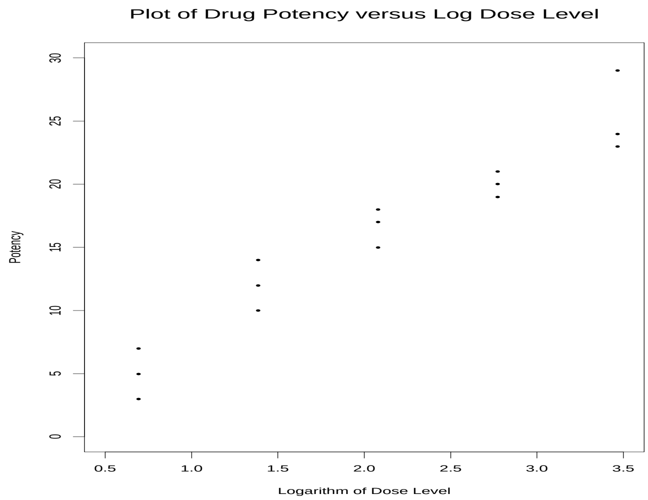
c. From the scatterplot, there appears to be curvature in the relation between Potency and Dose Level. A quadratic or cubic model may provide an improved fit to the data.

d. The quadratic model provides the better fit. The quadratic model has a much lower MS(Error), its R^2 value is 11% larger, the quadratic term has a p-value of 0.0062 which indicates that this term is significantly different from 0, however, the residual plot still has a distinct curvature as was found in the residual plot for the linear model.

12.2 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

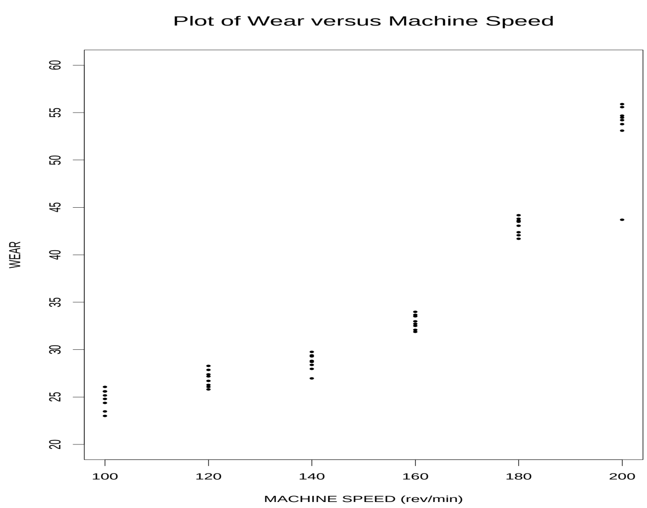
A scatterplot of the data is given here:



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

c. The model using $ln(x)$ provides a better fit based on the scatterplot, the decrease in MS(Error) over the quadratic model, increase in R^2, and the residual plot appears to be a random scatter of points about the horizontal line, whereas there was curvature in the residual plot from the fit of the quadratic model.

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



b. Since there is curvature in the relation, a quadratic model would be a possibility.

c. The quadratic model gives a better fit. For the quadratic model the residual plots displays a slight pattern in the residuals but not as distinct as found in the residual plot from the linear model. The R^2 values from the quadratic and cubic models are nearly identical but are about 10% higher than the value from the linear model. The cubic term has p-value=0.1794 which would indicate that the cubic term is not making a significant contribution to the fit of the model above just using a model having linear and quadratic terms.

d. There is one data value (Machine Speed=200, Wear=43.7) which is definitely an outlier. Considering the variability in the Wear values at each Machine Speed, it is possible that there is an error in recording the Wear value and in fact it should be 53.7. A check with the lab personnel would need to be done.

12.4 The Minitab output for the two models is given here:

```
Model I:   Regression Analysis: wear versus speed, speed_sq, conc


The regression equation is
wear = 60.5 - 0.705 speed + 0.00328 speed_sq + 8.88 conc

Predictor        Coef     SE Coef          T        P       VIF
Constant       60.477       5.512      10.97    0.000
speed        -0.70507     0.07551      -9.34    0.000     106.5
speed_sq    0.0032768   0.0002505      13.08    0.000     106.5
conc            8.875       2.499       3.55    0.001       1.0

S = 1.732       R-Sq = 97.4%     R-Sq(adj) = 97.2%

Analysis of Variance

Source            DF          SS          MS         F         P
Regression         3      4877.7      1625.9    542.22     0.000
Residual Error    44       131.9         3.0
Total             47      5009.6

Source       DF      Seq SS
speed         1      4326.8
speed_sq      1       513.1
conc          1        37.8

Unusual Observations
Obs      speed       wear         Fit      SE Fit    Residual    St Resid
 42        200     43.700      52.309       0.609      -8.609       -5.31R

R denotes an observation with a large standardized residual

Durbin-Watson statistic = 1.92
```

```
Model II:  Regression Analysis: wear versus speed, speed_sq, conc,
                                          speed*conc,speed_sq*conc


The regression equation is
wear = 42.3 - 0.421 speed + 0.00224 speed_sq + 69.5 conc - 0.949 speed_conc
           + 0.00345 speed_sq*conc

Predictor        Coef     SE Coef          T        P       VIF
Constant        42.28       17.01       2.49    0.017
speed         -0.4205      0.2351      -1.79    0.081    1064.7
speed_sq    0.0022422   0.0007800       2.87    0.006    1064.7
conc            69.54       53.77       1.29    0.203     477.4
speed_co      -0.9485      0.7435      -1.28    0.209    3118.0
speed_sq     0.003449    0.002467       1.40    0.169    1627.3

S = 1.705       R-Sq = 97.6%     R-Sq(adj) = 97.3%

Analysis of Variance

Source            DF          SS          MS        F        P
Regression         5     4887.53      977.51   336.22    0.000
Residual Error    42      122.11        2.91
Total             47     5009.64

Source       DF      Seq SS
speed         1     4326.79
speed_sq      1      513.10
conc          1       37.81
speed_co      1        4.15
speed_sq      1        5.68

Unusual Observations
Obs      speed       wear         Fit      SE Fit    Residual    St Resid
 42        200     43.700      51.419       0.773      -7.719       -5.08R

R denotes an observation with a large standardized residual

Durbin-Watson statistic = 2.12
```

The model $y = \beta_o + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \epsilon$ appears to be the better of the two models. The R^2-values are nearly the same (97.2% versus 97.3%). The mean squared error values are nearly the same (3.0 versus 2.91). The tests of significance for β_4 and β_5 are highly nonsignificant (p-values=0.209 and 0.169).

12.5 a. No, the two independent variables, Air Miles and Population, do not appear to be severly collinear, based on the correlation (-0.1502) and the scatterplot.

b. There are two potential leverage points in the Air Miles direction (around 300 and 350 miles). In addition, there is one possible leverage point in the population direction; this point has a value above 200.

12.6 We can answer the question by testing $H_o : \beta_1 = \beta_2 = 0$ versus $H_o : \beta_1 \neq 0$ and/or $\beta_2 \neq 0$. $F = 39.3378$ with p-value ≈ 0.0000. Therefore, there is substantial evidence to support H_a and conclude that the two variables have some value in predicting revenue.

12.7 a. For reduced model: $R^2 = 0.2049$

b. For complete model: $R^2 = 0.7973$

c. With INCOME as the only independent variable, there is a dramatic decrease in R^2 to a relatively small value. Thus, we can conclude that INCOME does not provide an adequate fit of the model to the data.

12.8 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:

$F = \frac{[SSReg.,Complete - SSReg.,Reduced]/(k-g)}{SSResidual,Complete/[n-(k+1)]} = \frac{[2.65376-0.68192]/(3-1)}{0.67461/[21-4]} = 24.84$

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

Reject H_o. There is substantial evidence to conclude that $\beta_2 \neq 0$ and/or $\beta_3 \neq 0$. Based on the F-test, omitting BUSIN and COMPET from the model has substantially changed the fit of the model. Dropping one or both of these independent variables from the model will result in a decrease in the predictive value of the model.

12.9 a. $R^2 = 0.979566 \Rightarrow$ 97.96% of the variation in Rating Score is accounted for the model containing the three independent variables.

b. $F = \frac{0.979566/3}{(1-0.979566)/496} = 7925.76$ This value is slightly smaller than the value on the output due to rounding-off error.

c. The p-value associated with such a large F-value would be much less than 0.0001 and hence there is highly significant evidence that the model containing the three independent variables provides predictive value for the Rating Score.

12.10 a. For the reduced model: R^2 is 89.53% which is a reduction of 8.43 percentage points from the complete model's R^2 of 97.96%.

b. In the complete model, we want to test $H_o : \beta_1 = \beta_3 = 0$ versus $H_a : \beta_1 \neq 0$ and/or $\beta_3 \neq 0$.

For the reduced model,

$SS(Regression, Reduced) = (R^2_{Reduced})(SS(Total) = (.895261)(99379.032) = 88970.17157$

The F-statistic has the form:

$F = \frac{[SSReg.,Complete - SSReg.,Reduced]/(k-g)}{SSResidual,Complete/[n-(k+1)]} = \frac{[97348.339-88970.17157]/(3-1)}{2030.693/[500-4]} = 1023.19$

with $df = 2, 496 \Rightarrow$ p-value $= Pr(F_{2,496} \geq 1023.19) < 0.0001 \Rightarrow$

Reject H_o. There is substantial evidence to conclude that $\beta_1 \neq 0$ and/or $\beta_3 \neq 0$. Based on the F-test, omitting Age and Debt Fraction from the model has substantially changed the fit of the model. Dropping one or both of these independent variables from the model will result in a decrease in the predictive value of the model.

12.11 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 : at least one of $\beta_3, \beta_4, \beta_5 \neq 0$.

The F-statistic has the form:
$F = \frac{[SSReg.,Complete - SSReg.,Reduced]/(k-g)}{SSResidual,Complete/[n-(k+1)]} = \frac{[448.193-305.808]/(5-2)}{71.489/[20-6]} = 9.29$
with $df = 3, 14 \Rightarrow$ p-value $= Pr(F_{3,14} \geq 9.29) = 0.0012 \Rightarrow$
Reject H_o. There is substantial evidence to conclude that at least one of $\beta_3, \beta_4, \beta_5 \neq 0$. Based on the F-test, omitting the second order terms from the model has substantially changed the fit of the model. Dropping one or more of these independent variables from the model will result in a decrease in the predictive value of the model.

12.12 a. $y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \epsilon$, where
$x_1 = \log(\text{dose})$

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

$\beta_0 = y-$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. $y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$

12.13 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: $y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \epsilon$, the test of equal slopes is a test of the hypotheses:

$H_o : \beta_4 = 0, \beta_5 = 0$ versus $H_o : \beta_4 \neq 0$ and/or $\beta_5 \neq 0$

Under H_o, the reduced model becomes $y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$

$F = \frac{(55.293-48.844)/(5-3)}{0.689/6} = 28.08 \Rightarrow$ p-value $= Pr(F_{2,6} \geq 28.08) = 0.0009$

b. Reject H_o and conclude there is significant evidence that the slopes of the three regression lines (one for each Drug Product) are different.

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

$H_o : \beta_2 = 0, \beta_3 = 0$ versus $H_o : \beta_2 \neq 0$ and/or $\beta_3 \neq 0$

Under H_o, reduced model becomes $y = \beta_o + \beta_1 x_1 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \epsilon$

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

12.14 a. For testing $H_o : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$, the p-value for the output is p-value = 0.0427. Thus, at the $\alpha = 0.05$ level we can reject H_o and conclude there is significant evidence that the probability of Completing the Task is related to Experience.

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

12.15 a. For testing $H_o : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$, the p-value for the output is p-value < 0.0001. Thus, we can reject H_o and conclude there is significant evidence that the probability of Tumor Development is related to Amount of Additive .

b. From the output, $\hat{p}(100) = 0.827$ with 95% C.I. (0.669, 0.919).

12.16 The output for fitting a regression model having TravTime regressed on Miles and TravDir is given here.

```
Regression Analysis: TravTime versus Miles, TravDir


The regression equation is
TravTime = 0.632 + 0.00191 Miles - 0.127 TravDir

Predictor         Coef     SE Coef          T        P
Constant       0.63182     0.03098      20.39    0.000
Miles       0.00191363  0.00002396      79.87    0.000
TravDir       -0.12743     0.01302      -9.79    0.000

S = 0.1898      R-Sq = 98.5%     R-Sq(adj) = 98.5%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      230.09      115.05   3194.94    0.000
Residual Error    97        3.49        0.04
Total             99      233.59

Source       DF      Seq SS
Miles         1      226.64
TravDir       1        3.45

Unusual Observations
Obs      Miles   TravTime         Fit      SE Fit    Residual    St Resid
 11       4501     9.5833      9.4999      0.0915      0.0834        0.50 X
 54       3784     7.0000      7.6182      0.0715     -0.6182       -3.52RX
 71       1395     4.0833      3.4288      0.0253      0.6546        3.48R
 75       2588     6.2500      5.8392      0.0513      0.4108        2.25R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

There are only slight changes in the fit of the model when the variable TravDir replaces the variable DirEffct in the model. The Error sum of squares of the fitted model increases from 2.51 to 3.49, the value of R^2_{adj} decreases from 98.9% to 98.5%, and the F-statistic for testing the overall fit decreases from 4472 to 3195. Thus, there are only trival changes to the fitted model.

12.17 The output for fitting a regression model having TravTime regressed on Miles, TravDir, DirEffct is given here.

```
Regression Analysis: TravTime versus Miles, TravDir, DirEffct

The regression equation is
TravTime = 0.645 + 0.00191 Miles - 0.0186 TravDir -0.000079 DirEffct

Predictor        Coef     SE Coef          T        P
Constant      0.64494     0.02635      24.47    0.000
Miles      0.00190703  0.00002034      93.74    0.000
TravDir      -0.01859     0.02065      -0.90    0.370
DirEffct  -0.00007890  0.00001265      -6.24    0.000

S = 0.1609      R-Sq = 98.9%     R-Sq(adj) = 98.9%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3     231.100      77.033   2975.60    0.000
Residual Error    96       2.485       0.026
Total             99     233.585

Source       DF      Seq SS
Miles         1     226.643
TravDir       1       3.449
DirEffct      1       1.008

Unusual Observations
Obs      Miles   TravTime         Fit      SE Fit    Residual    St Resid
 11       4501     9.5833      9.9759      0.1088     -0.3926      -3.31RX
 13       1746     3.9667      4.2873      0.0336     -0.3206      -2.04R
 39       1561     4.0833      3.7636      0.0232      0.3198       2.01R
 54       3784     7.0000      7.2269      0.0872     -0.2269      -1.68 X
 64       2477     5.3333      4.9406      0.0479      0.3927       2.56R
 71       1395     4.0833      3.4339      0.0215      0.6494       4.07R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.

Predicted Values for New Observations

New Obs     Fit     SE Fit         95.0% CI            95.0% PI
1        1.7145     0.0356   (  1.6438,  1.7852)  (  1.3874,  2.0416)

Values of Predictors for New Observations

New Obs     Miles   TravDir  DirEffct
1             500     -2.00     -1000
```

When DirEffct is included in a model with TravDir, the variable TravDir no longer provides a significant contribution to the fit of the model (p-value=0.37). Therefore, with DirEffct in the model, it is not necessary to also include the variable TravDir. This conclusion is further confirmed by comparing the models with Miles and just DirEffct to the model with Miles, DirEffct, and TravDir. The model without TravDir has the same value for R^2_{adj} as the model with TravDir, but a much larger value for the F-test of model significance (4471.71 vs 2975.60). From the output, the predicted travel time would be 1.7145 hours with a 95% prediction interval of (1.3874, 2.0416).

12.18 a. $\hat{y} = -1.320 + 5.550$ EDUC+0.885 INCOME+1.925 POPN-11.389 FAMSIZE

(57.98) (2.702) (1.308) (1.371) (6.669)

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

c. A value of 2.07 standard deviations from the predicted value is unusual since we would expect approximately 95% of all values to be within 2 standard deviations of the predicted. Thus, 2.07 standard deviations from the predicted is a moderately unusual point, but not a serious outlier.

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

H_o	H_a	T.S. t	p-value	Conclusion
$\beta_o = 0$	$\beta_o \neq 0$	$t = -0.02$	0.982	Fail to Reject H_o
$\beta_1 = 0$	$\beta_1 \neq 0$	$t = 2.05$	0.079	Fail to Reject H_o
$\beta_2 = 0$	$\beta_2 \neq 0$	$t = 0.68$	0.520	Fail to Reject H_o
$\beta_3 = 0$	$\beta_3 \neq 0$	$t = 1.40$	0.203	Fail to Reject H_o
$\beta_4 = 0$	$\beta_4 \neq 0$	$t = -1.71$	0.131	Fail to Reject H_o

None of the four independent variables appears to have predictive value given the remaining three variables have already been included in the model.

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

b. In the complete model, we want to test

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

The F-statistic has the form:

$F = \frac{[SSReg.,Complete - SSReg.,Reduced]/(k-g)}{SSResidual,Complete/[n-(k+1)]} = \frac{[1295.70-1268.48]/(4-2)}{50.51/[12-5]} = 1.89$

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

Fail to reject H_o. There is not substantial evidence to conclude that at least one of β_2, $\beta_3 \neq 0$. Based on the F-test, omitting INCOME and POPN from the model would not substantially changed the fit of the model. Dropping these independent variables from the model will not result in a large decrease in the predictive value of the model.

12.21 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-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.

12.22 a. Estimated complete model:

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

Estimated reduced model: $\hat{y} = 0.840085 + 1.01583x_1 + 0.0558262x_2$

b. In the complete model, we want to test
$H_o : \beta_3 = \beta_4 = \beta_5 = 0$ versus $H_a :$ at least one of $\beta_3, \beta_4, \beta_5 \neq 0$.
The F-statistic has the form:
$F = \frac{[2546.03-2516.12]/(5-2)}{229.857/[26-6]} = 0.87$
with $df = 3, 20 \Rightarrow$ p-value $= Pr(F_{3,20} \geq 0.87) = 0.4730 \Rightarrow$
Fail to reject H_o. There is not substantial evidence to conclude that at least one of $\beta_3, \beta_4, \beta_5 \neq 0$. Based on the F-test, omitting $x_3, x_4,$ and x_5 from the model would not substantially changed the fit of the model. Dropping these independent variables from the model will not result in a large decrease in the predictive value of the model.

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

b. $s_\epsilon = 1.70956$

c. $R^2 = 90.07\%$

d. There is no collinearity problem in the data set. The correlations between the pairs of independent variables is 0 for each pair and the VIF values are all equal to 1.0. This total lack of collinearity is due to the fact that the independent variables are perfectly balanced. Each combination of PROTEIN and ANTIBIO values appear exactly three times in the data set. Each combination of PROTEIN and SUPPLEM occur twice, etc.

12.24 a. When PROTEIN=15%, ANTIBIO=1.5%, SUPPLEM=5%,
$\hat{y} = 102.708 - .83333(15) - 4.000(1.5) - 1.375(5) = 77.333$

b. There is no extrapolation for these values of the independent variables because these values represent the mean of the values in the data set. In other words, the prediction of y is occurring in the middle of the data set.

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)

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

b. $R^2 = 0.5057$

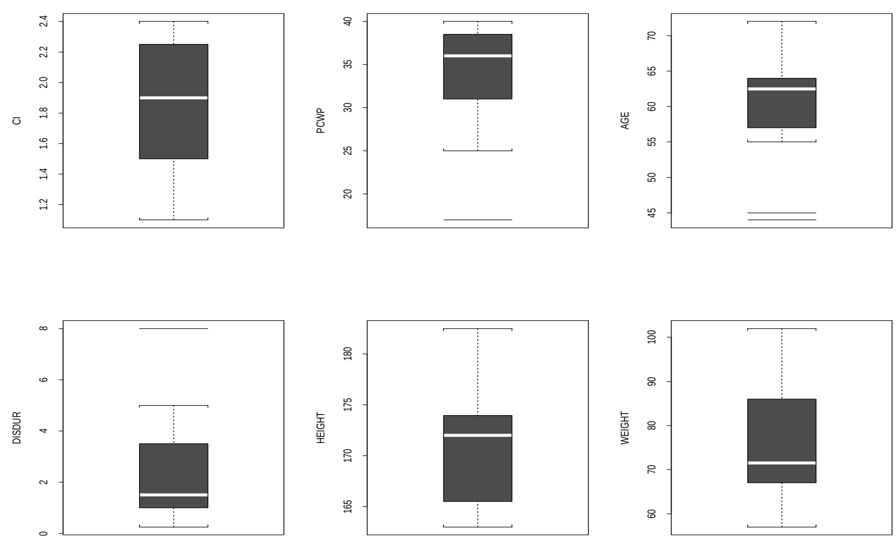
c. In the complete model, we want to test
$H_o : \beta_2 = \beta_3 = 0$ versus $H_a :$ at least one of $\beta_2, \beta_3 \neq 0$.
The F-statistic has the form:
$F = \frac{[371.083-208.333]/(3-1)}{40.9166/[18-4]} = 27.84$
with $df = 2, 14 \Rightarrow$ p-value $= Pr(F_{2,14} \geq 27.84) < 0.0001 \Rightarrow$ 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 x_2 and/or x_3 from the model would substantially changed the fit of the model. Dropping ANTIBIO and/or SUPPLEM from the model may result in a large decrease in the predictive value of the model.

12.26 The following points should be mentioned:

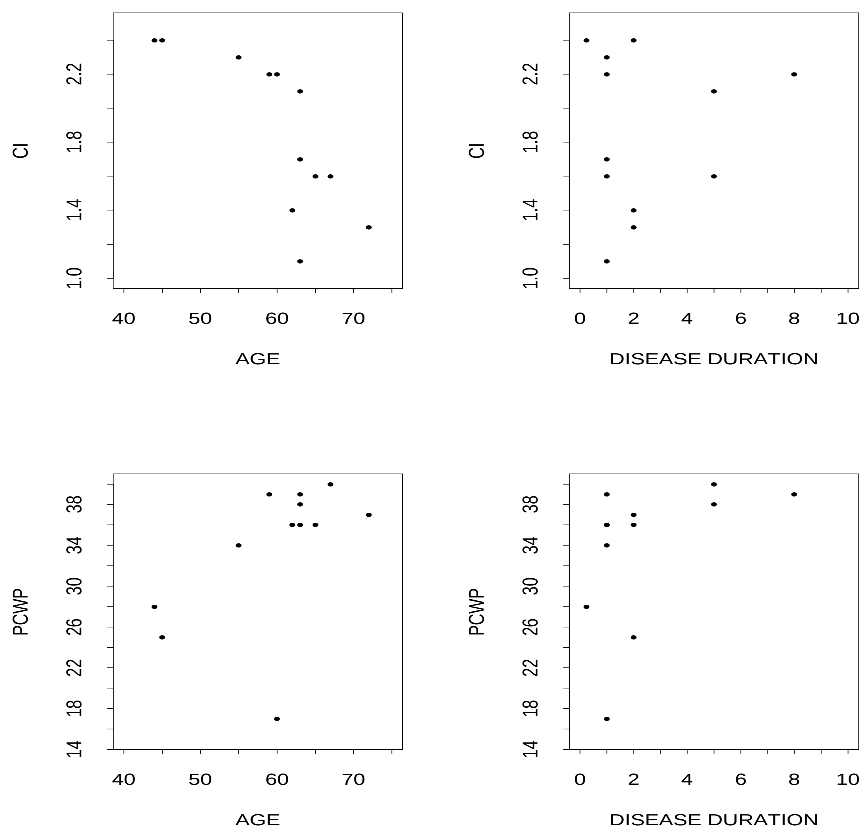
- Using the correlations between the dependent variable AvgOrder and the four independent variables, note that average order size has the strongest positive association with house price, followed by income, then education. There are negative associations between average order size and the two temperature variables.

- The regression model identifies house price as the most significant predictor. For fixed values of the other three variables, we estimate that a \$1 increase in house price is associated with an increase of \$0.06 in order size.
- Winter temperature also appears to have some predictive value in estimating average order size.
- The replacing of the actual house price with 0 for those two houses having missing values was not a good idea. These two values should just be deleted from the analysis for any models containing house price. The inclusion of these two data points could be very influential in the fit of the model. Check the output for these points to see if they were identified as influential. In any case, they should not be included.
- There are several zip codes with large negative residuals. This indicates that the model is considerably overpredicting the order sizes for these zip codes. These zip codes should be investigated to see why they differ from the other data points. A more complex model may be needed in order to accomodate these zip codes.

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



b. The scatterplots of the data are given here:

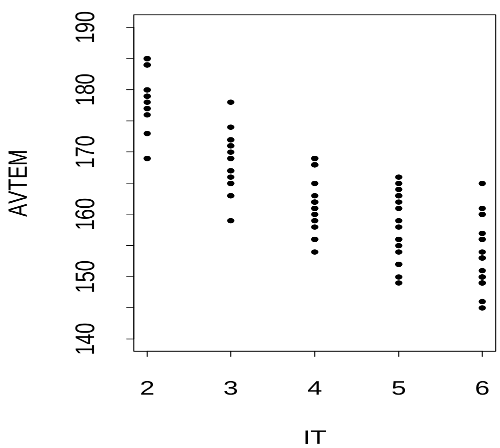


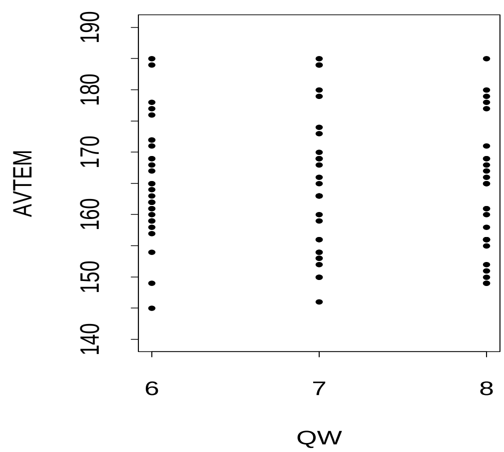
There appears to be somewhat of a negative correlation between Age and CI but a positive correlation between Age and PCWP. The relation between Disease Duration and CI is very weak but slightly positive. While the correlation between disease duration and PCWP is somewhat stronger.

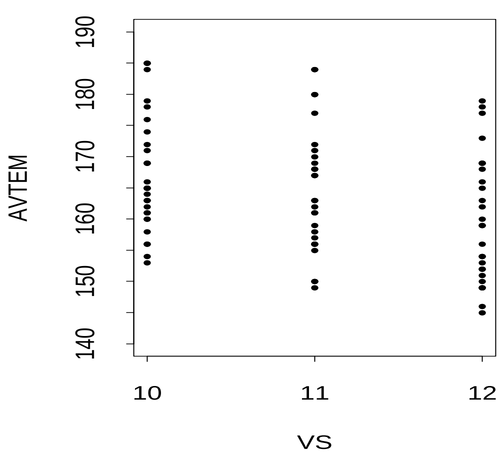
12.28 a. For both CI and PCWP, the addition of the interaction term between Age and Disease Duration (x_1x_2) did not appreciatively improve the fit of the model. The t-statistics for the interaction terms, which measures the effect of adding the interaction term to a model already containing the terms x_1 and x_2, were not significant. The p-values were 0.8864 for the CI model and 0.6838 for the PCWP model. The R^2 values for the two CI models were 67.39% for the model without the interaction term and 67.48% for the model with the interaction. Similarly, the R^2 values for the two PCWP models were 42% for the model without the interaction term and 43.27% for the model with the interaction. Thus for both the CI model and the PCWP model the interaction term adds very little predictive value to the model. The CI model has a p-value of 0.0021 for the Age term but only 0.1514 for Disease Duration.

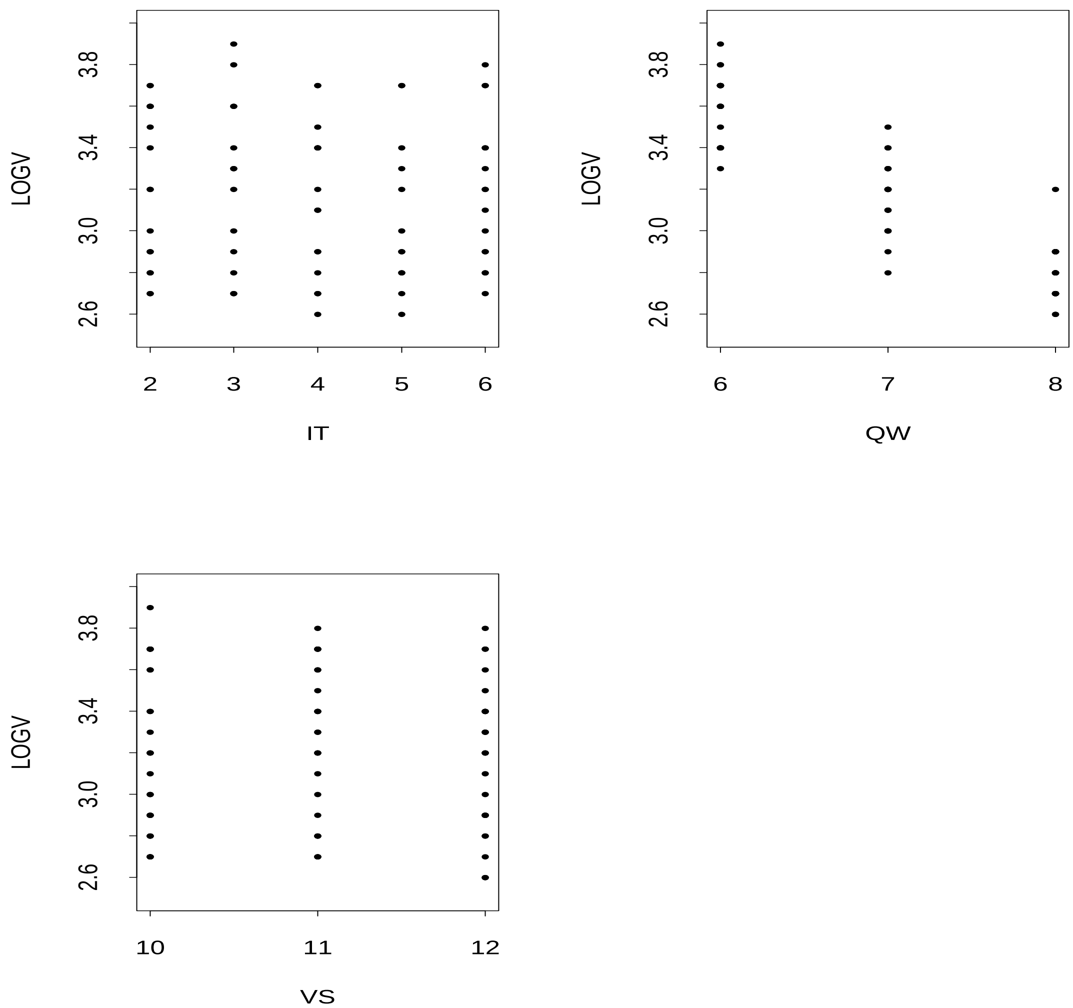
b. This confirms the correlation depictions in the two scatterplots for CI, where Age appears to have a greater correlation with CI than the correlation between Disease Duration and CI. A similar relationship is observed in the PCWP model, although in this model neither Age nor Disease Duration appears to provide much predictive value for PCWP.

12.29 a. 1. The scatterplots of the data are given here:









2. AVTEM tends to decrease as IT increases, but appears to remain fairly constant with increases in QW or VS.
3. LOGV tends to decrease as QW increases, but appears to remain fairly constant with increases in IT or VS.

b. SAS output is given here: with the following identification:

Variable	Notation	Variable	Notation
IT	x_1	IT*QW	x_7
QW	x_2	IT*VS	x_8
VS	x_3	VS*QW	x_9
I2	x_4	avtem	y_1
Q2	x_5	logv	y_2
V2	x_6		

```
                           The SAS System

                          The REG Procedure
                       Dependent Variable: y1

MODEL 1:
                        Analysis of Variance

                                    Sum of         Mean
 Source                  DF        Squares       Square  F Value  Pr > F

 Model                    3     7660.94568   2553.64856   131.97  <.0001
 Error                   86     1664.17654     19.35089
 Corrected Total         89     9325.12222


         Root MSE              4.39896    R-Square     0.8215
         Dependent Mean      164.25556    Adj R-Sq     0.8153
         Coeff Var             2.67812



                        Parameter Estimates

                      Parameter       Standard
 Variable     DF       Estimate          Error    t Value    Pr > |t|

 Intercept     1      234.56106        7.63872      30.71      <.0001
 x1            1       -6.15000        0.32788     -18.76      <.0001
 x2            1       -0.67445        0.56822      -1.19      0.2385
 x3            1       -3.73340        0.56843      -6.57      <.0001
------------------------------------------------------------------------
```

MODEL 2:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	7941.21675	1323.53612	79.38	<.0001
Error	83	1383.90547	16.67356		
Corrected Total	89	9325.12222			

Root MSE	4.08333	R-Square	0.8516
Dependent Mean	164.25556	Adj R-Sq	0.8409
Coeff Var	2.48596		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	234.87853	7.16673	32.77	<.0001
x1	1	-6.18215	0.30447	-20.30	<.0001
x2	1	-0.72541	0.52761	-1.37	0.1729
x3	1	-3.81541	0.52812	-7.22	<.0001
x4	1	0.96451	0.24758	3.90	0.0002
x5	1	-0.29207	0.91332	-0.32	0.7499
x6	1	-1.04740	0.91355	-1.15	0.2549

MODEL 3:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	7683.85390	1280.64232	64.76	<.0001
Error	83	1641.26833	19.77432		
Corrected Total	89	9325.12222			

Root MSE	4.44683	R-Square	0.8240
Dependent Mean	164.25556	Adj R-Sq	0.8113
Coeff Var	2.70726		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	214.01181	58.56103	3.65	0.0005
x1	1	-0.53333	5.30316	-0.10	0.9201
x2	1	0.21831	7.91120	0.03	0.9781
x3	1	-2.60819	5.21968	-0.50	0.6186
x7	1	-0.29167	0.40594	-0.72	0.4745
x8	1	-0.32500	0.40594	-0.80	0.4256
x9	1	0.02498	0.70409	0.04	0.9718

MODEL 4:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	7968.16362	885.35151	52.20	<.0001
Error	80	1356.95860	16.96198		
Corrected Total	89	9325.12222			

Root MSE	4.11849	R-Square	0.8545
Dependent Mean	164.25556	Adj R-Sq	0.8381
Coeff Var	2.50737		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	203.41326	54.30065	3.75	0.0003
x1	1	-0.22505	4.91223	-0.05	0.9636
x2	1	1.72599	7.33803	0.24	0.8146
x3	1	-1.82023	4.83905	-0.38	0.7078
x4	1	0.97354	0.25005	3.89	0.0002
x5	1	-0.29587	0.92146	-0.32	0.7490
x6	1	-1.04984	0.92165	-1.14	0.2581
x7	1	-0.34034	0.37617	-0.90	0.3683
x8	1	-0.32500	0.37597	-0.86	0.3899
x9	1	-0.09944	0.65298	-0.15	0.8793

---

1. The R^2 values for the four models are 0.8215, 0.8516, 0.8240, and 0.8545. There is very little difference in the 4 values for R^2, therefore, the selected model would be the model with the fewest independent variables, Model 1.
2. This question is equivalent to testing in Model 2 the hypotheses:
   $H_o : \beta_4 = \beta_5 = \beta_6 = 0$ versus H_a : at least one of $\beta_4, \beta_5, \beta_6 \neq 0$
   $F = \frac{(7941.21675-7660.94568)/(6-3)}{1383.90547/83} = 5.60$ with $df = 3, 83$
   p-value $= Pr(F_{3,83} \geq 5.50) = 0.0015$
   We thus conclude that Model 2 is significantly different in fit than Model 1, that is, at least one of $\beta_4, \beta_5, \beta_6$ is not equal to 0 in Model 2.
3. This question is equivalent to testing in Model 3 the hypotheses:
   $H_o : \beta_4 = \beta_5 = \beta_6 = 0$ versus H_a : at least one of $\beta_4, \beta_5, \beta_6 \neq 0$
   $F = \frac{(7683.85390-7660.94568)/(6-3)}{1641.26833/83} = 0.39$ with $df = 3, 83$
   p-value $= Pr(F_{3,83} \geq 0.39) = 0.7605$
   We thus conclude that Model 3 is not significantly different in fit than Model 1, that is, we cannot reject the hypothesis that $\beta_4 = \beta_5 = \beta_6 = 0$ in Model 3.
4. This question is equivalent to testing in Model 4 the hypotheses:
   $H_o : \beta_4 = \beta_5 = \beta_6 = 0$ versus H_a : at least one of $\beta_4, \beta_5, \beta_6 \neq 0$
   $F = \frac{(7968.16362-7683.85390)/(9-6)}{1356.95860/80} = 5.59$ with $df = 3, 80$
   p-value $= Pr(F_{3,80} \geq 5.59) = 0.0016$
   We thus conclude that Model 4 is significantly different in fit than Model 3, that is, at least one of $\beta_4, \beta_5, \beta_6$ is not equal to 0 in Model 4.

5. This question is equivalent to testing in Model 4 the hypotheses:
$H_o : \beta_7 = \beta_8 = \beta_9 = 0$ versus H_a : at least one of $\beta_7, \beta_8, \beta_9 \neq 0$
$F = \frac{(7968.16362-7941.21675)/(9-6)}{1356.95860/80} = 0.53$ with $df = 3, 80$
p-value $= Pr(F_{3,80} \geq 0.53) = 0.5296$
We thus conclude that Model 4 is not significantly different in fit than Model 2, that is, we cannot reject the hypothesis that $\beta_7 = \beta_8 = \beta_9 = 0$ in Model 4.
6. I would select Model 2 because from the plots and various tests Model 4 is not significantly different from Model 2 whereas Model 2 is significantly different from Model 1. Model 2 includes the variables I2, Q2 and V2, at least one of which appears to significantly improve the fit of the model over Model 1. Model 4 is more complex than Model 2 but does not appear to provide much improvement in the fit over Model 2 ($R^2 = 0.8545$ versus 0.8516).

c. SAS output is given here: with the following identification:

Variable	Notation	Variable	Notation
IT	x_1	IT*QW	x_7
QW	x_2	IT*VS	x_8
VS	x_3	VS*QW	x_9
I2	x_4	avtem	y_1
Q2	x_5	logv	y_2
V2	x_6		

```
Dependent Variable: y2

MODEL 1:

                                 Sum of          Mean
Source                   DF     Squares        Square   F Value   Pr > F

Model                     3     9.87413       3.29138    160.33   <.0001
Error                    86     1.76543       0.02053
Corrected Total          89    11.63956

        Root MSE               0.14328     R-Square     0.8483
        Dependent Mean         3.19778     Adj R-Sq     0.8430

                        Parameter Estimates

                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|

Intercept     1        6.23345        0.24880      25.05      <.0001
x1            1        0.00667        0.01068       0.62      0.5341
x2            1       -0.40568        0.01851     -21.92      <.0001
x3            1       -0.02028        0.01851      -1.10      0.2764
```

MODEL 2:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	9.96474	1.66079	82.30	<.0001
Error	83	1.67482	0.02018		
Corrected Total	89	11.63956			

Root MSE	0.14205	R-Square	0.8561
Dependent Mean	3.19778	Adj R-Sq	0.8457

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.25908	0.24932	25.10	<.0001
x1	1	0.00632	0.01059	0.60	0.5525
x2	1	-0.40624	0.01835	-22.13	<.0001
x3	1	-0.02148	0.01837	-1.17	0.2457
x4	1	0.01047	0.00861	1.22	0.2274
x5	1	0.01043	0.03177	0.33	0.7436
x6	1	-0.05300	0.03178	-1.67	0.0991

MODEL 3:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	9.97345	1.66224	82.81	<.0001
Error	83	1.66610	0.02007		
Corrected Total	89	11.63956			

Root MSE	0.14168	R-Square	0.8569
Dependent Mean	3.19778	Adj R-Sq	0.8465

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.95482	1.86582	5.34	<.0001
x1	1	-0.21000	0.16896	-1.24	0.2174
x2	1	-0.81681	0.25206	-3.24	0.0017
x3	1	-0.35718	0.16630	-2.15	0.0347
x7	1	0.00083333	0.01293	0.06	0.9488
x8	1	0.01917	0.01293	1.48	0.1421
x9	1	0.03719	0.02243	1.66	0.1012

MODEL 4:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	10.05889	1.11765	56.57	<.0001
Error	80	1.58066	0.01976		
Corrected Total	89	11.63956			

Root MSE	0.14056	R-Square	0.8642
Dependent Mean	3.19778	Adj R-Sq	0.8489
Coeff Var	4.39568		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.83366	1.85328	5.31	<.0001
x1	1	-0.20686	0.16765	-1.23	0.2209
x2	1	-0.79658	0.25045	-3.18	0.0021
x3	1	-0.34633	0.16516	-2.10	0.0392
x4	1	0.00993	0.00853	1.16	0.2482
x5	1	0.01164	0.03145	0.37	0.7122
x6	1	-0.05187	0.03146	-1.65	0.1031
x7	1	0.00033702	0.01284	0.03	0.9791
x8	1	0.01917	0.01283	1.49	0.1392
x9	1	0.03547	0.02229	1.59	0.1154

---

1. The R^2 values for the four models are 0.8483, 0.8561, 0.8569, and 0.8642. There is very little difference in the 4 values for R^2, therefore, the selected model would be the model with the fewest independent variables, Model 1.
2. This question is equivalent to testing in Model 2 the hypotheses:
   $H_o : \beta_4 = \beta_5 = \beta_6 = 0$ versus H_a : at least one of $\beta_4, \beta_5, \beta_6 \neq 0$
   $F = \frac{(9.96474-9.87413)/(6-3)}{1.67482/83} = 1.50$ with $df = 3, 83$
   p-value $= Pr(F_{3,83} \geq 1.50) = 0.2206$
   We thus conclude that Model 2 is not significantly different in fit than Model 1, that is, we cannot reject the hypothesis that $\beta_4 = \beta_5 = \beta_6 = 0$ in Model 2.
3. This question is equivalent to testing in Model 3 the hypotheses:
   $H_o : \beta_4 = \beta_5 = \beta_6 = 0$ versus H_a : at least one of $\beta_4, \beta_5, \beta_6 \neq 0$
   $F = \frac{(9.97345-9.87413)/(6-3)}{1.66610/83} = 1.65$ with $df = 3, 83$
   p-value $= Pr(F_{3,83} \geq 1.65) = 0.1842$
   We thus conclude that Model 3 is not significantly different in fit than Model 1, that is, we cannot reject the hypothesis that $\beta_4 = \beta_5 = \beta_6 = 0$ in Model 3.

4. This question is equivalent to testing in Model 4 the hypotheses:
   $H_o : \beta_4 = \beta_5 = \beta_6 = 0$ versus H_a : at least one of $\beta_4, \beta_5, \beta_6 \neq 0$
   $F = \frac{(10.05889-9.97345)/(9-6)}{1.58066/80} = 1.44$ with $df = 3, 80$
   p-value $= Pr(F_{3,80} \geq 1.44) = 0.2373$
   We thus conclude that Model 4 is not significantly different in fit than Model 3, that is, we cannot reject the hypothesis that $\beta_4 = \beta_5 = \beta_6 = 0$ in Model 3.
5. This question is equivalent to testing in Model 4 the hypotheses:
   $H_o : \beta_7 = \beta_8 = \beta_9 = 0$ versus H_a : at least one of $\beta_7, \beta_8, \beta_9 \neq 0$
   $F = \frac{(10.05889-9.96474)/(9-6)}{1.58066/80} = 1.59$ with $df = 3, 80$
   p-value $= Pr(F_{3,80} \geq 1.59) = 0.5296$
   We thus conclude that Model 4 is not significantly different in fit than Model 2, that is, we cannot reject the hypothesis that $\beta_7 = \beta_8 = \beta_9 = 0$ in Model 4.
6. I would select Model 1 because from the plots and various tests for the following reasons. Model 2 and Model 3 are not significantly different from Model 1. Model 4 is more complex than Model 2 but does not provide much improvement in the fit over Model 2. Therefore, since the models are not significantly different, the R^2 values are nearly the same, and Model 1 is the model containing the fewest independent variables (hence the easiest to understand), I would select Model 1.