CHAPTER 2

Many excellent computer programs are available for plotting the data and for carrying out the regression calculations. Here we use S-Plus, R, Minitab, SAS, and SPSS. Most programs work the same and it is not difficult to switch from one program to the other. Most packages are spreadsheet programs. You enter the data into the various columns of a spreadsheet and use simple commands to carry out the operations. The results (fitted values, residuals, …) can be stored in unused columns of the worksheet. Many options are available within all programs. You need to consult the on-line help for detailed discussion and examples.

The Minitab software is very easy to use. Minitab works like a spreadsheet program. We enter the data into columns of a spreadsheet and use the tabs: Stat > Regression > Regression. We specify the response variable and the explanatory (regressor) variables and execute the regression command. The output provides the estimates, standard errors, t-ratios and probability values. It displays the ANOVA table and the coefficient of determination. The output (residuals and fitted values) can be stored in unused columns of the worksheet.

A note on computing with R

R is a free software which is available through the internet; it can be downloaded from http://cran.us.r-project.org/. It is very similar to the commercial package S-Plus. R is a language and an environment for statistical computing and graphics. It can be used with Windows 95 or later versions, a variety of Unix and Linux platforms, and Apple Macintosh (OS versions later than 8.6).

The most convenient way to use R is at a graphics work station running a windowing system. We have used R on UNIX machines to solve several of the exercises, and the following discussion assumes this set-up. If you are running R under Windows, you will need to make some minor adjustments.

R issues the prompt **" >"** whenever it expects input commands. Let us assume that the UNIX shell prompt is %. You can start the R program with the command **%R**. Then R will return with a banner line, and R commands may be issued at this point. The command

>help.start()

starts the HTML interface for on-line help, using the web browser that is available at your computer. You can use the mouse to explore features of the help facility. The command for quitting an R session is

>q()

At this point you will be asked whether you want to save the data from your R session.

R has an extensive help facility. You can get information on any specific function – for example the natural logarithm – by typing

>help(log) or >?log

R is case-sensitive, so x and X refer to different variables. R operates on named data structures. Data can be entered at the terminal or can be read from an external file. Entering the elements of a vector x – consisting of the four numbers 2, 4, 5, and 7 – one uses the R command

>x <- c(2,4,5,7) or >x = c(2,4,5,7)

This is an assignment statement using the function c(). Notice that the assignment operator "<-" (which is the same as the "=" operator) consists of the two characters < ("less than") and - ("minus") and points to the object receiving the value of the expression. For simplicity we use "=".

For the exercises in this book we read the data from an external file (a text file in UNIX). In exercise 2.6, for example, we have modified the file **hooker** so that the first four lines are as follows:

Temp AP
210.8 29.211
210.2 28.559
208.4 27.972

The first line of the file specifies a name for each variable in the data frame. The subsequent lines include the values for each variable. To read an entire data frame, we use the command

>hook = read.table("hooker",header=T)

The filename **hooker** is in quotes; header =T indicates that the first line includes the names of the variables. The commands

>Temp = hook[,1]; >AP=hook[,2]

define the first column of the matrix "hook" as Temp and the second column as AP. The statement

>LnAP = 100*log(AP)

results in a transformation of the variable AP; log(AP) is the natural log of AP.

The function for fitting simple or multiple linear regression models is lm(). For instance, a simple linear regression of Temp on LnAP can be fit by issuing the command

>hookfit = lm(Temp~LnAP)

The output object from the lm() command, "hookfit", is a fitted model object. Information about the fitted model can be extracted from this file. For example,

>summary(hookfit)

prints a comprehensive summary of the results of the regression analysis including the estimated coefficients, their standard errors, t–values and p-values (see the solution to exercise 2.6).

The command

>anova(hookfit)

supplies the analysis of variance (ANOVA) table. The command

>plot(LnAP,Temp)

plots Temp (the y-coordinate) against LnAP (the x-coordinate). A graphics window opens automatically. The fitted line can be superimposed on the scatter plot by issuing the command

>abline(hookfit)

The command

>qqnorm(hookfit$residuals)

leads to a normal probability plot of the residuals where "residuals" is in the fitted model object "hookfit".

Our discussion has focused on the free software package R. Note that the commands and the output of S-Plus are pretty much the same.

In subsequent chapters (Chapters 4 - 8) we consider multiple linear regression models. These models can be fit quite easily with R (and S-Plus). Suppose we have data in the vectors y, x1, x2 and x3. We can fit a multiple linear regression of y on x1, x2, and x3 by using the command

>mregfit=lm(y~x1+x2+x3)

Information about the model is in the fitted model object "mregfit". Note that an intercept term is included by default. One can restrict the intercept to be zero through

>mulregfit=lm(y~x1+x2+x3-1)

The above commands can be fine-tuned according to specific requirements. Many other commands are available to perform various statistical analyses and plots (such as residual analysis, leverages, Cook's D, various residual plots). This note is meant as a brief introduction to R. You should use the on-line help mentioned above to obtain more details.

2.1

(a) 95^{th} percentile = 10 + 3(1.645) = 14.93; 99^{th} percentile = 10 + 3 (2.326) = 16.98

(b) $t(0.95;10) = 1.812$; $t(0.95;25) = 1.708$; $t(0.99;10) = 2.764$; $t(0.99;25) = 2.485$

(c) $\chi^2(0.95;1) = 3.84$; $\chi^2(0.95;4) = 9.49$; $\chi^2(0.95;10) = 18.31$

$\chi^2(0.99;1) = 6.63$; $\chi^2(0.99;4) = 13.28$; $\chi^2(0.99;10) = 23.21$

(d) $F(0.95;2,10) = 4.10$; $F(0.95;4,10) = 3.48$; $F(0.99;2,10) = 7.56$;

$F(0.99;4,10) = 5.99$

2.2 Computer programs can be used to calculate the percentiles. Or, they can be looked up in the tables given in the appendix. The rounding errors are due to the number of digits displayed in various tables (and programs).

(a) $z(0.95) = 1.645$; $\chi^2(0.90;1) = 2.706$: $(1.645)^2 = 2.706$

$z(0.975) = 1.96$; $\chi^2(0.95;1) = 3.841$: $(1.96)^2 = 3.841$

$z(0.99) = 2.326$; $\chi^2(0.98;1) = 5.412$: $(2.326)^2 = 5.412$

$z(0.995) = 2.576$; $\chi^2(0.99;1) = 6.635$: $(2.576)^2 = 6.635$

(b) $t(0.95;4) = 2.132$; $F(0.90;1,4) = 4.545$: $(2.132)^2 = 4.545$

$t(0.975;4) = 2.776$; $F(0.95;1,4) = 7.709$: $(2.776)^2 = 7.709$

$t(0.99;10) = 2.764$; $F(0.98;1,10) = 7.638$: $(2.764)^2 = 7.638$

$t(0.995,10) = 3.169$; $F(0.99;1,10) = 10.044$: $(3.169)^2 = 10.044$

2.3 Correlation = 0.816; R^2 = 0.867; Estimated equation: $\hat{\mu} = 3 + 0.5x$

Same (linear regression) results for all four data sets. However, scatter plots in Figure 4.10 of the text show that linear regression is only appropriate for first data set. The correlation coefficients and the least squares estimates can be obtained by computer programs such as S-Plus, R, Minitab, SPSS, Minitab and others.

2.4

(a) Scatter plot shows an approximate linear relationship

(b) $\hat{\beta}_1 = 40/12.8 = 3.125$; $\hat{\beta}_0 = 13 - (3.125)(4.2) = -0.125$

(c) Fitted equation: $\hat{\mu} = -0.125 + 3.125x$

(d) $\hat{\mu}(x = 5) = -0.125 + 3.125(5) = 15.5$

(e)

X = Sales People	Y = Cars Sold	Fitted Value	Residual
6	20	18.625	1.375
6	18	18.625	-0.625
4	10	12.375	-2.375
2	6	6.125	-0.125
3	11	9.250	1.750

(f) $s^2 = 11/3 = 3.67$

(g) 95% confidence interval for β_1 : $3.125 \pm (3.182)(0.5352)$ or (1.42, 4.83). Since zero is not in this interval, we reject $\beta_1 = 0$.

(h) Significant relationship between the number of cars sold and the number of sales people. Number of cars sold increases as the number of sales people increases.

(i) If you know (can predict) sales, you can solve the equation in (c) to obtain the number of sales people that are required. However, only five weeks of data was available to estimate the model. Also, we do not know whether this period is representative for the whole year. Advisable to collect more data before using this model for decision making.

2.5 Minitab Output:

```
The regression equation is
Cars Sold = - 0.12 + 3.12 Sales People

Predictor         Coef      SE Coef         T        P
Constant        -0.125       2.406      -0.05    0.962
Sales People    3.1250      0.5352      5.84    0.010

S = 1.915       R-Sq = 91.9%     R-Sq(adj) = 89.2%

Analysis of Variance

Source           DF          SS          MS         F        P
Regression        1      125.00      125.00     34.09    0.010
Residual Error    3       11.00        3.67
Total             4      136.00
```

2.6
(a) Scatter plot (not shown here) indicates that a linear model is not appropriate. A quadratic component or a transformation are needed.
(b) Scatter plot confirms linear relationship between y = TEMP and x = 100ln(AP).
(c) R (S-Plus) output from the function 'lm':

```
                  Value  Std. Error    t value  Pr(>|t|)
(Intercept)     49.2684      1.1990    41.0925    0.0000
100ln(AP)        0.4782      0.0040   119.0838    0.0000

Residual standard error: s = 0.4016 with 29 degrees of freedom
Multiple R-Squared: 0.998
F-statistic: 14,180 with 1 and 29 degrees of freedom; the p-value is 0
```

(c) Estimated equation: $\hat{\mu} = 49.268 + 0.478\ln(\text{AP})$; $R^2 = 0.998$; $s = \sqrt{\text{MSE}} = 0.402$.
The model is appropriate since there is small random scatter around the fitted line;

(d) (i) $\hat{\beta}_1 = 0.4782$ and $\text{s.e.}(\hat{\beta}_1) = 0.0040$. Since t(0.975;29) = 2.045, a 95% confidence interval for β_1: 0.4782 – 2.045(0.0040), 0.4782 + 2.045(0.0040), or (0.470 , 0.486)

(ii) $\hat{\mu} = 49.268 + 0.478(100\ln(25)) = 203.195$;

$$\text{s.e.}(\hat{\mu}) = \sqrt{s^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right]} = \sqrt{(0.402)^2/31 + (0.0040)^2(321.888 - 298.041)^2} = 0.1196$$

95% confidence interval:
[203.195 – 2.045 (0.1196), 203.195 + 2.045 (0.1196)], or (202.950, 203.440)

(e) Estimates and standard errors of β_0 and β_1 change by factor of 5/9.

2.7

(a) $\hat{\beta} = \bar{y} = \sum y_i / n$; $s^2 = \sum (y_i - \bar{y})^2 / (n-1)$

(b) (i) Prediction interval is wider
(ii) 99% percent prediction interval is wider
(iii) Calculation error

2.8 Minitab output:

```
The regression equation is
Revenue = 32 + 0.263 Cars

Predictor        Coef     SE Coef          T        P
Constant         31.9       185.2       0.17    0.867
Cars          0.26251     0.03930       6.68    0.000

S = 264.0       R-Sq = 84.8%     R-Sq(adj) = 82.9%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1     3109923     3109923     44.62    0.000
Residual Error     8      557529       69691
Total              9     3667452
```

(a) Estimated equation: $\hat{\mu} = 31.9 + 0.2625x$; t-ratio($\hat{\beta}_1$) = 0.2625/0.0393 = 6.68; p-value = 0.0002; number of cars sold is a significant predictor variable.

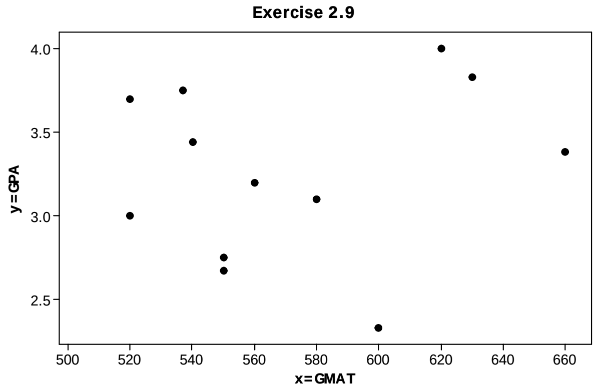
(b) 95% confidence interval for β_1: $0.2625 \pm (2.306)(0.0393)$ or (0.172, 0.353)

(c) $R^2 = 0.848$

(d) Standard deviation of y after factoring in x is $s = \sqrt{MSE} = 264.0$; standard deviation of y (without factoring x) is 638.3531.

(e) $\hat{\mu}(x = 1187) = 343.5$

2.9 The scatter plot of y = GPA against x = GMAT score shows considerable variability.



The Minitab regression output is given below:

```
The regression equation is
GPA = 2.16 + 0.00193 x=GMAT

Predictor      Coef   SE Coef     T      P
Constant      2.158     2.014  1.07  0.309
GMAT     0.001931  0.003510  0.55  0.594

S = 0.532633   R-Sq = 2.9%   R-Sq(adj) = 0.0%

Analysis of Variance

Source          DF      SS      MS     F      P
Regression       1  0.0858  0.0858  0.30  0.594
Residual Error  10  2.8370  0.2837
Total           11  2.9228
```

(a) Estimated equation: $\hat{\mu} = 2.158 + 0.0019\text{x}$; R^2 = 0.029; the model explains only 2.9% of the variability in y; not much of a relationship over the limited range of GMAT scores; other factors may be more important

(b) $\hat{\mu}(\text{x} = 540) = 2.158 + 0.001931(40) = 3.23$

(c) t-ratio($\hat{\beta}_1$) = 0.001931/0.00351 = 0.55; p-value = 0.594; conclude $\beta_1 = 0$

2.10

(a) Prediction at weight 2000 is 0.5598 + (0.001024)(2000) = 2.6078. Since n is large and the estimation error can be ignored, s.e(prediction error) = s = $\sqrt{0.066}$ = 0.2569. Thus, an approximate 95% prediction interval is 2.6078 $\pm$ (1.96)(0.2569), or (2.104, 3.111). Note that 1.96 is from the standard normal table.

(b) The prediction at weight 1500 is 0.5598 + (0.001024)(1500) = 2.0958. Thus, an approximate 95% prediction interval is 2.09 $\pm$ (1.96)(0.2569) = (1.592, 2.599)

2.11

$$\frac{1}{R^2} = \frac{SST}{SSR} = \frac{SSR + SSE}{SSR} = 1 + \frac{SSE}{SSR} = 1 + \frac{n-p-1}{p}\frac{1}{F}$$

Hence, $R^2 = \left[1 + \frac{n-p-1}{pF}\right]^{-1}$.

2.12

(a) $\hat{\beta}_1 = \sum x_i y_i \big/ \sum x_i^2$; $s^2 = \sum (y_i - \hat{\beta}_1 \bar{y})^2 \big/ (n-1)$

(b) $\sum e_i x_i = 0$, but not necessarily $\sum e_i = 0$

(c) $V(\hat{\beta}_1) = \frac{1}{\left[\sum x_i^2\right]^2} \sigma^2 \sum x_i^2 = \sigma^2 \frac{1}{\left[\sum x_i^2\right]}$

2.13

(a) Estimated equation: $\hat{\mu} = 0.520x$; $s^2 = 46.2/16 = 2.89$;

$\hat{\beta}_1 = 0.520$; s.e.$(\hat{\beta}_1) = 0.0132$; 95% confidence interval: (0.492, 0.548)

(b) Estimated equation: $\hat{\mu} = 0.725 + 0.498x$; $\hat{\beta}_0 = 0.725$; s.e.$(\hat{\beta}_0) = 1.549$;

$\hat{\beta}_0$ / s.e.$(\hat{\beta}_0) = 0.725/1.549 = 0.47$; p-value = 0.65; conclude $\beta_0 = 0$

2.14 Minitab output:

```
The regression equation is
y = - 0.228 + 0.995 x

Predictor        Coef     SE Coef          T        P
Constant      -0.2281      0.1378      -1.65    0.137
x            0.994757    0.005219     190.59    0.000

S = 0.2067      R-Sq = 100.0%    R-Sq(adj) = 100.0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      1552.2      1552.2  36322.72    0.000
Residual Error     8         0.3         0.0
Total              9      1552.6
```

(a) Fitted equation: $\hat{\mu} = -0.228 + 0.995x$

(b) 95% confidence interval for β_0: $-0.2281 \pm (2.306)(0.1378)$ or $(-0.546, 0.090)$

(c) 95% confidence interval for β_1: $0.9948 \pm (2.306)(0.005219)$ or $(0.983, 1.007)$

(d) (i) Test $\beta_0 = 0$: 95% confidence interval for β_0 covers 0;

(ii) Test $\beta_1 = 0$: 95% confidence interval for β_1 covers 1

(e) Minitab output

```
The regression equation is
y = 0.987 x

Predictor        Coef     SE Coef          T        P
Noconstant
x            0.987153    0.002704     365.09    0.000

S = 0.2258

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      6796.2      6796.2 133292.08    0.000
Residual Error     9         0.5         0.1
Total             10      6796.7
```

95% confidence interval for β_1: $0.9872 \pm (2.262)(0.002704)$ or $(0.981, 0.993)$; does not cover 1

(e) Restriction $\beta_0 = 0$. The estimate of β_1 depends on the estimate of β_0. Thus the estimates of β_1 with β_0 restricted at 0 and with unrestricted β_0 are not necessarily the same.

2.15 R output:

```
Residual Standard Error = 4.5629
R-Square = 0.6767
F-statistic (df=1, 5) = 10.4657
p-value = 0.0231

            Estimate   Std.Err   t-value   Pr(>|t|)
Intercept    68.4459   12.9270    5.2948     0.0032
x            -0.4104    0.1268   -3.2351     0.0231

ANOVA
Source          DF        SS        MS      F      P
Regression       1    217.90    217.90  10.47  0.023
Residual Error   5    104.10     20.82
Total            6    322.00
```

(a) Estimated equation: $\hat{\mu} = 68.45 - 0.41x$; $R^2 = 0.677$; $s = 4.563$.

F-statistic = 10.47; p-value = 0.023; reject $\beta_1 = 0$

(b) $\text{s.e.}(\hat{\beta}_0) = 12.93$; $\hat{\beta}_0 / \text{s.e.}(\hat{\beta}_0) = 68.45/12.93 = 5.29$; p-value = 0.003

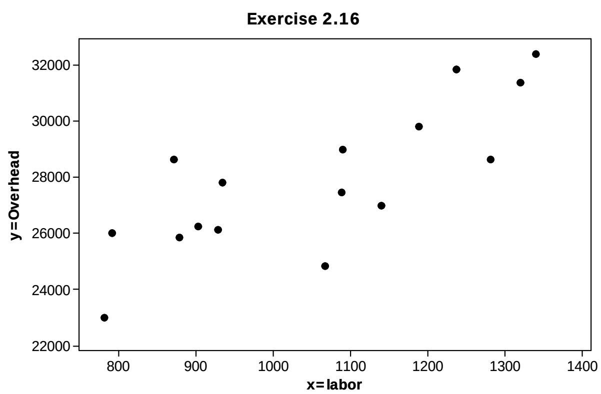
$\text{s.e.}(\hat{\beta}_1) = 0.127$; $\hat{\beta}_1 / \text{s.e.}(\hat{\beta}_1) = -0.41/0.127 = -3.23$; p-value = 0.023;

reject $\beta_0 = 0$ and $\beta_1 = 0$ at the 5 percent significance level.
99% confidence interval for β_1: (-0.92, 0.11).

(c) $\hat{\mu}(\mathrm{x} = 100) = 27.41$; $\mathrm{s.e.}(\hat{\mu}(\mathrm{x} = 100)) = 1.73$;
95% confidence interval: (22.97,31.86).

(d) $\hat{\mu}(\mathrm{x} = 84) = 33.98$; $\mathrm{s.e.}(\hat{\mu}(\mathrm{x} = 84)) = 2.76$;
95% confidence interval: (26.88, 41.07).
Note that $\bar{x}$ =101 and $s.e.(\hat{\mu}_0)$ is smallest when $x_0 = \bar{x}$. As x_0 moves away from $\bar{x}$, $s.e.(\hat{\mu}_0)$ becomes larger and the corresponding confidence interval becomes wider.

2.16 The scatterplot of overhead against labor hours shows a linear relationship




```
The regression equation is
Overhead = 16310 + 11.0 Labor

Predictor    Coef  SE Coef     T      P
Constant    16310     2421  6.74  0.000
Labor      10.982    2.268  4.84  0.000

S = 1645.61   R-Sq = 62.6%   R-Sq(adj) = 60.0%

Analysis of Variance

Source          DF         SS        MS      F      P
Regression       1   63517077  63517077  23.46  0.000
Residual Error  14   37912232   2708017
Total           15  101429309
```

The fitted values are the estimates of the expected total departmental overhead; they can be used as the predictions of the total departmental overhead for these given labor hours. Prediction intervals can be calculated. For example, for a new month with

$x_i = 1{,}000$ labor hours, the prediction is $\hat{y}_i = 428$ and the 95% prediction interval is (23645, 30939).

2.17

(a) The scatter plot shows that length (y) increases with increasing width (x).

```
Residual Standard Error = 4.295
R-Square = 0.9555
F-statistic (df=1, 8) = 171.7821
p-value = 0

            Estimate  Std.Error    t-value    Pr(>|t|)
Intercept   -46.4359    13.4161    -3.4612      0.0086
Width (x)     1.7924     0.1368    13.1066      0.0000
```

(b) Estimated equation: $\hat{\mu} = -46.44 + 1.792x$;

95% confidence interval for β_0 : (-77.37, -15.50);

95% confidence interval for β_1: (1.48, 2.11).

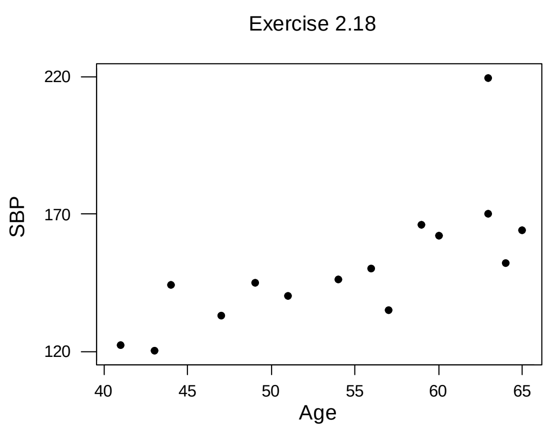
(c) Good fit; $R^2 = 0.956$

(d) $\hat{\mu}(x = 100) = 132.8$; 95% prediction interval: (122.39,143.22)

(e) Strong linear relationship

2.18

(a) The plot of SBP against age indicates that there is a linear relationship between SBP and age.



(b) Estimated equation: $\hat{\mu} = 33.31 + 2.168x$;

(c) Analysis of variance

Source	DF	SS	MS	F	P
Regression	1	4361.5	4361.5	14.58	0.002
Residual Error	13	3889.4	299.2		
Total	14	8250.9			

(d) F = 14.58; p-value = 0.002; reject $\beta_1 = 0$

(e) s.e.$(\hat{\beta}_1) = 0.568$; $\hat{\beta}_1 / \text{s.e.}(\hat{\beta}_1) = 2.168/0.568 = 3.82$; same p-value = 0.002; reject $\beta_1 = 0$

(f) Individual with $x = 63$ and $y = 220$ unusual. Estimates and standard errors change; R^2 increases. See R output shown below.

```
Residual Standard Error = 8.9007
R-Square = 0.7019
F-statistic (df=1, 12) = 28.2562
p-value=2e-04

           Estimate  Std.Error   t-value   Pr(>|t|)
Intercept  58.9876    16.6075    3.5519    4e-03
Weight      1.6244     0.3056    5.3157    2e-04

ANOVA
Source           DF        SS      MS      F      P
Regression        1    2238.5  2238.5  28.26  0.000
Residual Error   12     950.7    79.2
Total            13    3189.2
```

2.19 R Output:

```
Residual Standard Error = 0.1512
R-Square = 0.9496
F-statistic (df=1, 4) = 75.4083
p-value = 0.001

            Estimate   Std.Error   t-value   Pr(>|t|)
Intercept    3.7073      0.0955   38.8347      0.000
Mol.weight  -0.0123      0.0014   -8.6838      0.001
```

(a) Estimated equation: $\hat{\mu} = 3.707 - 0.0123x$; $R^2 = 0.950$

(b) F-statistic = 75.41; p-value = 0.001; reject $\beta_1 = 0$ at the 0.01 significance level. Significant linear relationship.

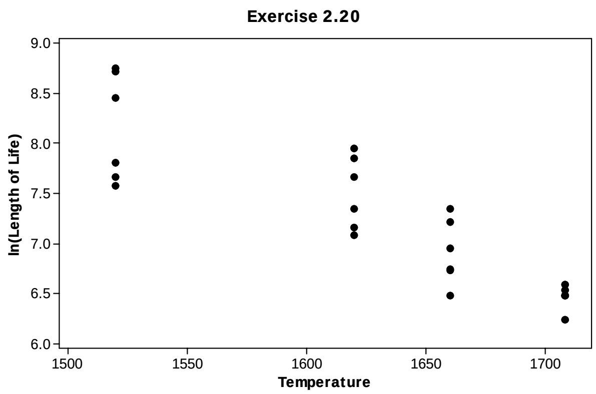
(c) Response is average of 3 observations. Use of individual values would improve the sensitivity of the analysis.

(d) No; molecular weight 200 far outside the region of experimentation; one does not know whether the linear relationship will continue to hold.

2.20

(a) Scatterplot of y = length of life against x = temperature shows: (i) length of life decreases with increasing temperature; (ii) variability in y is related to the level of y.

(b) Logarithmic transformation, ln(y), goes a long way toward stabilizing the variability.



(c) Minitab output

```
The regression equation is
ln(Life) = 22.1 - 0.00911 temp

Predictor       Coef   SE Coef      T      P
Constant      22.084     1.773  12.46  0.000
temp       -0.009110  0.001088  -8.37  0.000

S = 0.368943   R-Sq = 76.1%   R-Sq(adj) = 75.0%
```

```
Analysis of Variance
Source          DF       SS      MS      F      P
Regression       1   9.5347  9.5347  70.05  0.000
Residual Error  22   2.9946  0.1361
Total           23  12.5293
```

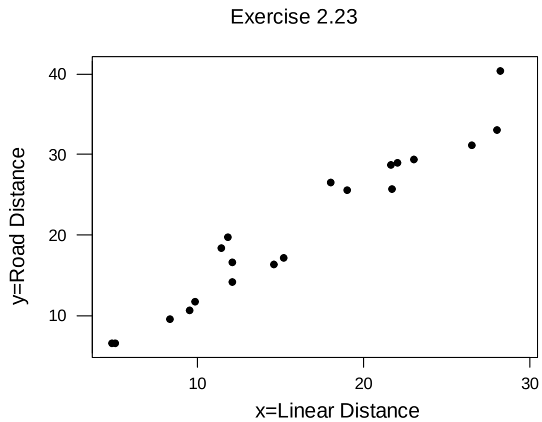
2.21 Plot of the chemical test against the magnetic test (not shown) indicates a linear relationship. Results of fitting a linear regression model are given below (R output):

```
Residual Standard Error = 3.4636
R-Square = 0.5372
F-statistic (df=1, 51) = 59.2056
p-value = 0
          Estimate  Std.Err   t-value  Pr(>|t|)
Intercept   8.9565   1.6523   5.4205     0
Mag Test    0.5866   0.0762   7.6945     0
```

Estimated equation: $\hat{\mu} = 8.957 + 0.587x$; R^2 = 0.537; F = 59.21; reject $\beta_1 = 0$
Significant linear relationship between the tests. However, variability large and predictive power low.

2.22 Plot of y (memory retention) against x (time) shows a nonlinear (exponentially decaying) pattern. Graphs of ln(y) against x and ln(y) against ln(x) show similar patterns. Plot of y against ln(x) shows a linear pattern.
Estimated equation: $\hat{\mu} = 0.846 - 0.079\ln(x)$; R^2 = 0.990; good model

2.23 The graph of road distance against linear distance shows an approximate linear relationship

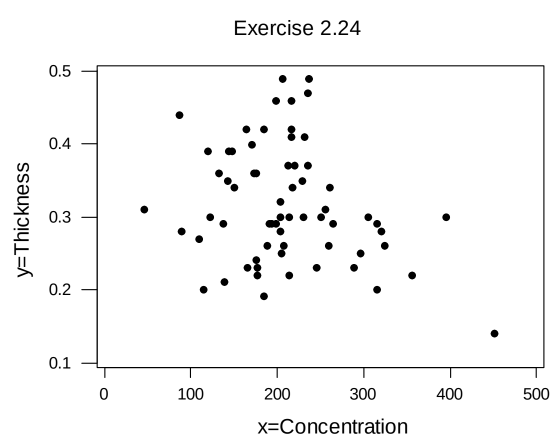


.

Estimated equation: $\hat{\mu} = 0.375 - 0.000279x$; $R^2 = 0.939$; s = 2.436;
$t(\hat{\beta}_1) = 0.379/1.26943 = 16.67$; p-value 0.000; conclude that $\beta_1 > 0$. Interesting fact that the confidence interval for β_1 does not cover one; $1.269 \pm (2.10)(0.076)$ or (1.109, 1.429)

```
The regression equation is
y=Road = 0.38 + 1.27 x=Linear

Predictor         Coef     SE Coef          T        P
Constant         0.379       1.344       0.28    0.781
x=Linear       1.26943     0.07617      16.67    0.000

S = 2.436       R-Sq = 93.9%     R-Sq(adj) = 93.6%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      1648.3      1648.3    277.73    0.000
Residual Error    18       106.8         5.9
Total             19      1755.1
```

2.24 The graph of concentration against thickness shows considerable scatter. Also the first egg with concentration = 452 and thickness = 0.14 is unusual and somewhat different from the rest (more on outlying cases in Chapter 6).



Estimated equation: $\hat{\mu} = 0.375 - 0.000279x$; $R^2 = 0.064$ small;
$t(\hat{\beta}_1) = -0.000279/0.000135 = -2.07$ with p-value 0.042 is barely significant at the 0.05 significance level.
Without the first case, the estimated equation is: $\hat{\mu} = 0.357 - 0.000184x$; $R^2 = 0.025$ is

small; $t(\hat{\beta}_1) = -0.000184/0.000146 = -1.26$ with p-value = 0.214. We conclude that $\beta_1 = 0$.

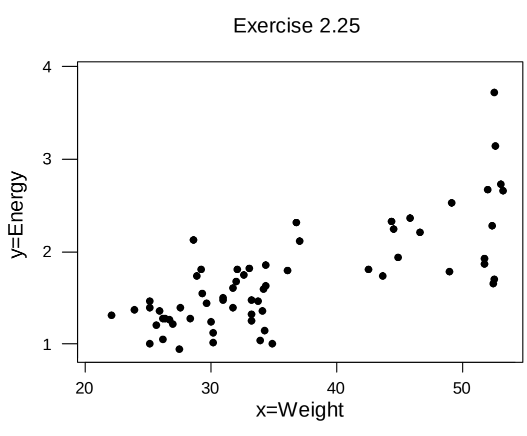
With all observations:

```
The regression equation is
Thickness = 0.375 -0.000279 Concentration

Predictor        Coef     SE Coef          T        P
Constant      0.37494     0.02990      12.54    0.000
Concentr   -0.0002790   0.0001345      -2.07    0.042

S = 0.07848     R-Sq = 6.4%      R-Sq(adj) = 4.9%

Analysis of Variance
Source            DF          SS          MS         F        P
Regression         1    0.026493    0.026493      4.30    0.042
Residual Error    63    0.388021    0.006159
Total             64    0.414514
```

With the first observation omitted:

```
The regression equation is
Thickness = 0.357 -0.000184 Concentration

Predictor        Coef     SE Coef          T        P
Constant      0.35700     0.03174      11.25    0.000
Concentr   -0.0001838   0.0001464      -1.26    0.214

S = 0.07761     R-Sq = 2.5%      R-Sq(adj) = 0.9%
```

2.25 The scatter plot of energy requirement against weight shows a linear relationship.



Estimated equation: $\hat{\mu} = 0.133 - 0.0434x$; $R^2 = 0.563$; s = 0.3662;

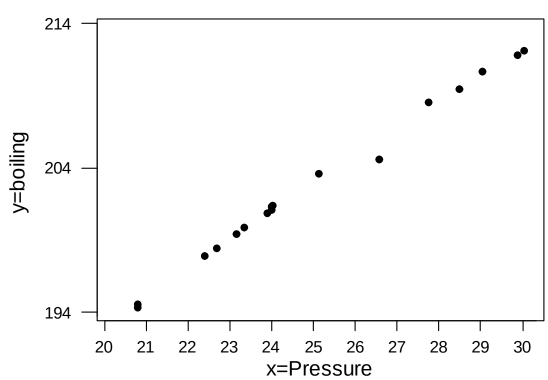
$t(\hat{\beta}_1) = 0.04342 / 0.004857 = 8.94$ with p-value 0.000 is significant; we conclude that $\beta_1 > 0$ and that weight has a significant influence. Energy requirement increases by 0.0434 Mcal/Day for each kg of body weight.

The 11th observation (weight = 52.6; y = 3.73) should be scrutinized it is the observation that seems somewhat different from the pattern exhibited by the majority of the cases (more on outlying cases in Chapter 6).

```
The regression equation is
Energy = 0.133 + 0.0434 Weight

Predictor        Coef     SE Coef          T        P
Constant       0.1329      0.1804       0.74    0.464
Weight       0.043416    0.004857       8.94    0.000

S = 0.3662      R-Sq = 56.3%     R-Sq(adj) = 55.6%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      10.718      10.718     79.91    0.000
Residual Error    62       8.316       0.134
Total             63      19.034
```

2.26 The scatter plot of boiling point against barometric pressure shows a strong linear relationship.

Figure 2.26



Estimated equation: $\hat{\mu} = 155.296 + 1.902x$; $R^2 = 0.994$; s = 0.444;

$t(\hat{\beta}_1) = 1.90178 / 0.03676 = 51.74$ with p-value 0.000; we conclude $\beta_1 > 0$; barometric pressure has a significant influence on boiling point. The boiling point

increases by 1.92 degrees F when barometric pressure increases by one inch of mercury.
The observation y = 204.6, x = 26.57 should be scrutinized as it seems different from the pattern that is exhibited by the rest (more on outlying cases in Chapter 6).

```
The regression equation is
boiling = 155 + 1.90 Pressure

Predictor        Coef     SE Coef          T        P
Constant      155.296       0.927     167.47    0.000
Pressure      1.90178     0.03676      51.74    0.000

S = 0.4440      R-Sq = 99.4%     R-Sq(adj) = 99.4%

Analysis of Variance
Source            DF          SS          MS         F        P
Regression         1      527.82      527.82   2677.11    0.000
Residual Error    15        2.96        0.20
Total             16      530.78
```

The data set in Exercise 2.6 includes cases where barometric pressure < 20. The graph with both data sets (not given) shows that the estimated models are quite similar.

2.27

(a) Response y = takeup(kg). Scatter plot indicates a linear relationship. R output:

```
Residual Standard Error = 3.3945
R-Square = 0.9858
F-statistic (df=1, 22) = 1530.289    p-value = 0

              Estimate  Std.Error    t-value    Pr(>|t|)
Intercept      -9.8960     1.6887    -5.8602       0
x               0.0753     0.0019    39.1189       0
```

y = Takeup(kg): $\hat{\mu} = -9.896 + 0.0753x$; R^2 = 0.986; F = 1,530.3; reject $\beta_1 = 0$

(b) Response y = takeup(kg). Scatter plot indicates a linear relationship. R output:

```
Residual Standard Error = 0.3952
R-Square = 0.703
F-statistic (df=1, 22) = 52.068
p-value = 0
           Estimate  Std.Error    t-value   Pr(>|t|)
Intercept    4.7372     0.1966    24.0973      0
 x           0.0016     0.0002     7.2158      0
```

y = Takeup(%): $\hat{\mu} = 4.737 + 0.00162x$; R^2 = 0.703; F = 52.07; reject $\beta_1 = 0$

Both models fit well. However, the first one seems to be better (larger R^2).