Solution to Additional Exercise 2.7

The data consist of sets of 5 chemical measurements on 32 crude oil samples. Our interest is in predicting y (the percentage of gasoline obtained) from the 4 other variables.

y = percentage of gasoline obtained,
x_1 = gravity,
x_2 = vapor pressure,
x_3 = temperature at which 10% of the crude oil is vaporized,
x_4 = temperature at which 100% of the crude oil is vaporized.

Our initial model is the (full) multiple linear regression,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i,$$

where the errors $\varepsilon_1, \ldots, \varepsilon_{32}$ are assumed independent and identically distributed (i.i.d.) and normally distributed $\mathrm{N}(0, \sigma^2)$. Before starting the regression analyses, we compute the simple correlations (with P-values) between the variables in the model.

	y	x1	x2	x3
x1	0.246			
	0.174			
x2	0.384	0.621		
	0.030	0.000		
x3	-0.315	-0.700	-0.906	
	0.079	0.000	0.000	
x4	0.712	-0.322	-0.298	0.412
	0.000	0.073	0.098	0.019

It is seen that only x_4 is strongly correlated with weight, but among the predictors correlations between x_1, x_2 and x_3 are reasonably strong, in particular the value of -0.91 between x_2 and x_3 should be noted.

The following Minitab output for the multiple regression model gives parameter estimates, the ANOVA table and a list of "unusual" observations.

```
The regression equation is
y = - 6.8 + 0.227 x1 + 0.554 x2 - 0.150 x3 + 0.155 x4

Predictor        Coef     SE Coef          T        P
Constant        -6.82       10.12      -0.67    0.506
x1            0.22725     0.09994       2.27    0.031
x2             0.5537      0.3698       1.50    0.146
x3           -0.14954     0.02923      -5.12    0.000
x4           0.154650    0.006446      23.99    0.000

S = 2.234       R-Sq = 96.2%     R-Sq(adj) = 95.7%

Analysis of Variance
```

```
Source            DF          SS          MS         F        P
Regression         4     3429.27      857.32    171.71    0.000
Residual Error    27      134.80        4.99
Total             31     3564.08

Source       DF      Seq SS
x1            1      216.26
x2            1      309.85
x3            1       29.21
x4            1     2873.95

Unusual Observations
Obs         x1          y         Fit      SE Fit    Residual    St Resid
 29       40.0     30.400      25.779       0.541       4.621        2.13R

R denotes an observation with a large standardized residual
```

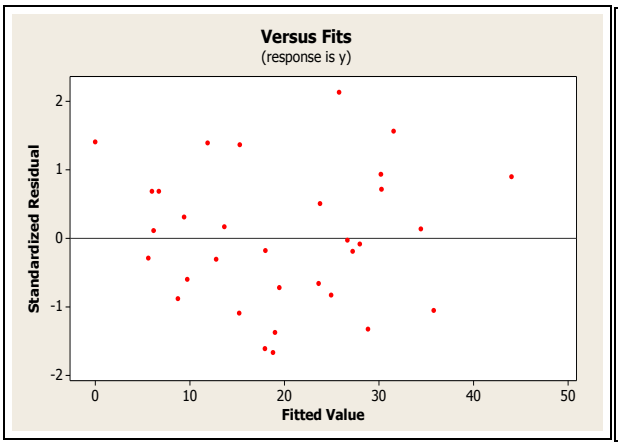
Comments to the fit of the full multiple regression model:

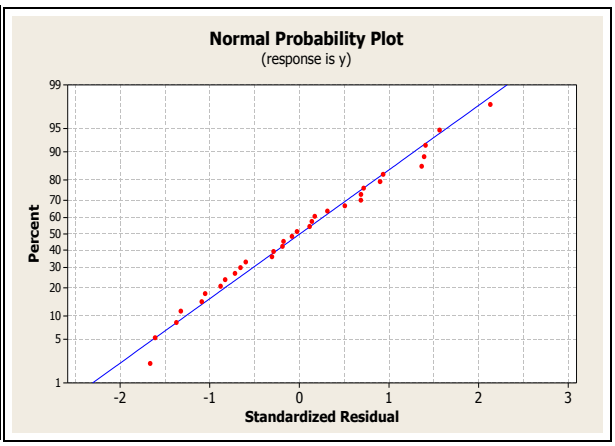
- only the regression coefficient for x_2 is non-significant, the model is overall strongly significant ($P = 0.01$) as a predictor of gasoline yield, and the predictive power is high ($R^2 = 0.96$),
- the estimated correlation matrix for the 5 parameter estimates (intercept + 4 regression coefficients), as computed by the `corrmat` macro, is

```
 1.00000  -0.66825  -0.71588  -0.90903   0.01726
-0.66825   1.00000   0.03687   0.38165   0.04262
-0.71588   0.03687   1.00000   0.84609  -0.19439
-0.90903   0.38165   0.84609   1.00000  -0.30912
 0.01726   0.04262  -0.19439  -0.30912   1.00000
```

  comments: apart from strong correlations (up to -0.91) between $\hat{\beta}_0$ (intercept) and the other regressions coefficients (which is not alarming), the strongest correlation is, as expected, between the regression coefficients for x_2 and x_3 (0.85), whereas the remaining parameter estimates are only weakly correlated,

- before proceeding further with the analysis, we examine the residuals. The plot of residuals versus fitted values (next page) looks fine, and so does the normal plot and normality tests are far from significant.





- finally, a list of regression diagnostics (standardized residuals, deletion residuals, leverage and Cook's distance):

Row	SRES1	TRES1	HI1	COOK1
1	-0.87919	-0.87537	0.134002	0.0239214
2	-1.60913	-1.66069	0.036123	0.0194076
3	0.68743	0.68056	0.149440	0.0166055
4	-0.59872	-0.59146	0.181359	0.0158824
5	0.68696	0.68009	0.282893	0.0372337
6	1.40816	1.43555	0.195570	0.0964151
7	-0.28780	-0.28285	0.116989	0.0021948
8	-0.30533	-0.30014	0.285065	0.0074345
9	0.31195	0.30668	0.201405	0.0049086
10	-1.66423	-1.72394	0.072971	0.0436031
11	-0.18774	-0.18435	0.049523	0.0003673
12	0.17113	0.16802	0.111398	0.0007342
13	1.39213	1.41794	0.175522	0.0825164
14	0.11385	0.11175	0.139028	0.0004186
15	-0.17655	-0.17335	0.090137	0.0006176
16	-0.65717	-0.65011	0.213362	0.0234277
17	0.50798	0.50089	0.159762	0.0098131
18	-1.32332	-1.34286	0.082260	0.0313926
19	1.56497	1.61049	0.076282	0.0404503
20	1.36540	1.38868	0.072894	0.0293166
21	-0.82714	-0.82216	0.124807	0.0195127
22	-0.71864	-0.71204	0.180781	0.0227929
23	-1.08968	-1.09363	0.251935	0.0799790
24	-1.37312	-1.39711	0.106660	0.0450226
25	0.93543	0.93319	0.168152	0.0353761
26	0.13818	0.13564	0.223213	0.0010973
27	0.71780	0.71120	0.188141	0.0238802
28	-1.05163	-1.05378	0.129884	0.0330169
29	2.13163	2.29367	0.058576	0.0565440
30	-0.02655	-0.02606	0.289723	0.0000575
31	-0.08244	-0.08091	0.152098	0.0002438
32	0.90072	0.89747	0.300045	0.0695544

  The most extreme residual, $t = 2.13$ for observation 29, is not at all extreme in a dataset of this size, and the outlier test based on the deletion residual and a Bonferroni correction is far from significant at the 5% level ($P = 0.96$). The highest leverage value is 0.30 which is just below the lower comparison value ($2p/n = 10/32 = 0.31$). All values of Cook's distance are very small, and none is markedly higher than the others. We conclude that the residuals and diagnostics do not indicate any problems with the model's assumptions.

In the full model, only one model reduction is suggested: to drop the variable x_2 which has a non-significant effect and was previously noted to be strongly associated with x_3. The resulting model is shown below; only minor changes are observed in the estimates and model statistics.

```
The regression equation is
y = 4.03 + 0.222 x1 - 0.187 x3 + 0.157 x4

Predictor        Coef     SE Coef          T        P
Constant        4.032       7.223       0.56    0.581
```

```
x1            0.2217      0.1021       2.17    0.038
x3          -0.18657     0.01592     -11.72    0.000
x4          0.156527    0.006462      24.22    0.000

S = 2.283       R-Sq = 95.9%     R-Sq(adj) = 95.5%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3      3418.1      1139.4    218.51    0.000
Residual Error    28       146.0         5.2
Total             31      3564.1

Source       DF      Seq SS
x1            1       216.3
x3            1       142.1
x4            1      3059.7

Unusual Observations
Obs         x1          y         Fit      SE Fit    Residual    St Resid
 29       40.0     30.400      25.634       0.544       4.766        2.15R

R denotes an observation with a large standardized residual.
```

For illustration we show also a listing of model comparison statistics for the best models of different sizes (from the `Best Subsets` menu in Minitab).

```
                                                 x x x x
Vars   R-Sq    R-Sq(adj)       C-p         S     1 2 3 4

   1   50.6         49.0     324.5    7.6588           X
   1   14.8         11.9     580.6    10.064       X
   1    9.9          6.9     615.0    10.345         X
   1    6.1          2.9     642.5    10.564   X
   2   95.2         94.9       8.2    2.4255         X X
   2   89.6         88.9      48.1    3.5713       X   X
   2   75.8         74.1     146.6    5.4518   X       X
   2   15.4          9.5     578.2    10.199       X X
   2   14.8          8.9     582.5    10.235   X X
   3   95.9         95.5       5.2    2.2835   X   X X
   3   95.5         95.0       8.2    2.3951       X X X
   3   92.6         91.8      29.2    3.0792   X X   X
   3   15.6          6.5     578.6    10.366   X X X
   4   96.2         95.7       5.0    2.2344   X X X X
```

It is seen that both Mallows C_p and the adjusted R^2 point to the full model as the best model, but we prefer to remove the non-significant effect of x_2 (because it *is* non-significant and because of the high correlation in the parameter estimates).

Normally, the analysis would stop here by drawing conclusions based on the selected model, but there is a problem with the data and model that has not shown up in the model checking statistics examined so far. The data only take 10 different sets of values of the predictors (x_1, x_2, x_3). In a dataset of 32 observations this is hardly a coincidence. Following the discussion of Atkinson (1985), "these 10 sets of values characterizeten different crudes, which were then subjected to experimentally controlled distillation conditions, varying in number from 2 to 4 per crude". (Admittedly, such an interpretation

is very hard to give based on the minimal information provided about the data.) Before discussing the implications of this finding, a few words on how it could be detected. First, listing the data sorted by x_1 (or x_2 or x_3) would show the replicates of these variables. Second, plots of these variables against each other would show 10 instead of 32 points. And third, plots of the residuals against any of the 3 predictors would also show the limited number of different values of the predictor.
Inspection of the data shows that the variable x_3 takes exactly 10 different values, corresponding to these 10 sets of values. Therefore, we can fit a model with a separate mean for each of the 10 sets by fitting x_3 as a categorical variable.

```
Factor     Type Levels Values
x3        fixed     10 190 210 217 220 231 236 267 274 284 316

Analysis of Variance for y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
x3          9     769.76    1685.56     187.28   53.05  0.000
x4          1    2720.18    2720.18    2720.18  770.57  0.000
Error      21      74.13      74.13       3.53
Total      31    3564.08

Term          Coef   SE Coef        T      P
Constant   -33.708     1.947   -17.31  0.000
x4        0.158730  0.005718    27.76  0.000

Unusual Observations for y

Obs         y       Fit      SE Fit  Residual   St Resid
  2   14.4000   17.5360      1.1208   -3.1360     -2.08R
  6    2.8000   -0.3757      1.0604    3.1757      2.05R
 19   34.9000   31.5042      1.1071    3.3958      2.24R

R denotes an observation with a large standardized residual.

Means for Covariates

Covariate      Mean     StDev
x4            332.1     69.76

Least Squares Means for y

 x3      Mean     StDev
190    32.549    0.9495
210    24.275    1.1259
217    27.782    1.1334
220    21.154    0.9398
231    21.519    1.0936
236    20.436    1.0865
267    15.036    0.9416
274    13.063    1.1006
284     9.805    1.3658
316     4.436    1.1353
```

We can now test our full multiple regression model against this larger model (a test for *lack of fit*.) Our null hypothesis is that the full multiple regression describes equally well as the larger model;

effectively this means that the 10 levels in our larger model are modelled well by the linear regression equation involving variables x_1, x_2 and x_3. We compute an F-test as follows,

$$F = \frac{[\text{SSE(R)} - \text{SSE(F)}]/[\text{DFE(R)} - \text{DFE(F)}]}{\text{MSE(F)}} = \frac{[134.80 - 74.13]/[27 - 21]}{3.53} = 2.86$$

which corresponds to $P = 0.034$ in a $F(6, 21)$-distribution. Therefore, the multiple regression model does not describe the data quite as well as the model with separate levels for each of the 10 predictor sets for (x_1, x_2, x_3). Model checking procedures for this model show no problem with the model's assumptions.

On another note, Box-Cox analyses of both the multiple regression model and the general linear model (with a categorical variable and a regressor) show strong evidence for the need of transformation ($P = 0.006$ and $P < 0.001$ for the test of $\lambda = 1$). The optimal values of λ are around 0.69 for both models, with an approximate 95% CI for the latter model ranging from 0.54 to 0.83. Therefore, a convenient choice of transformation is $y^{2/3}$. It seems surprising that the evidence in favor of such a "mild" transformation is so clear, in particular as our model checking procedures did not reveal any problems. Atkinson (1985) notes the largest values of Cook's distance to be *smaller* than expected from the model, using more advanced methods for model checking than discussed in this course. One useful rule of thumb, though, is to always consider the possibility of transforming outcomes that are restricted in their range (percentages are between 0 and 100%, and several of the data values are close to zero).