Solution to Additional Exercise 2.8

The data consist of sets of physiological measurements for 60 men in the Los Angeles Heart Study. Our interest here is in predicting weight from the other variables.

y = weight (pounds),
x_1 = age (years),
x_2 = systolic blood pressure (mm of mercury),
x_3 = diastolic blood pressure (mm of mercury),
x_4 = cholesterol (mg per dl),
x_5 = height (inches).

All predictors are quantitative, so our initial model is the (full) multiple linear regression,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i,$$

where the errors $\varepsilon_1, \ldots, \varepsilon_{60}$ 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 between all variables in the model.

	weight	age	syst bp	dias bp	chol
age	0.036				
syst bp	0.293	0.369			
dias bp	0.352	0.328	0.840		
chol	-0.014	0.183	0.170	0.196	
height	0.345	-0.249	0.018	0.052	-0.092

It is seen that none of the predictors are particularly strongly correlated with the outcome (weight); therefore, models with rather low predictive power must be anticipated. Among the predictors, there is only a high correlation between the two blood pressures: biologically sensible, but still indicating considerable collinearity between them.

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
weight = - 112 + 0.029 age + 0.020 syst bp + 0.727 dias bp - 0.0210 chol
         + 3.25 height

Predictor      Coef  SE Coef      T      P    VIF
Constant    -112.50    89.56  -1.26  0.214
age          0.0291   0.2840   0.10  0.919  1.270
syst bp      0.0197   0.3039   0.06  0.949  3.508
dias bp      0.7274   0.4892   1.49  0.143  3.453
chol       -0.02103  0.04859  -0.43  0.667  1.063
height        3.248    1.241   2.62  0.011  1.095

S = 21.0861   R-Sq = 23.4%   R-Sq(adj) = 16.3%
PRESS = 29533.8   R-Sq(pred) = 5.76%
```

```
Analysis of Variance

Source            DF          SS          MS         F        P
Regression         5      7330.4      1466.1      3.30    0.011
Residual Error    54     24009.6       444.6
Total             59     31340.0

Source       DF      Seq SS
age           1        40.7
syst bp       1      2829.4
dias bp       1      1240.4
chol          1       172.2
height        1      3047.7

Unusual Observations
Obs        age     weight         Fit      SE Fit    Residual    St Resid
  2       35.0     216.00      173.69        7.28       42.31        2.14R
  5       61.0     182.00      187.29       12.09       -5.29       -0.31 X
 18       40.0     225.00      168.36        3.62       56.64        2.73R
 19       51.0     247.00      189.70        9.08       57.30        3.01R

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

Comments to the fit of the full multiple regression model:

- most regression coefficients are (partially) non-significant, the model is overall only moderately significant ($P = 0.01$) as a predictor of weight, and the predictive power is low ($R^2 = 0.23$), in particular when evaluated by the PRESS approach ($R^2 = 0.06$; the large reduction in R^2 could indicate an unstable prediction caused by noise in the prediction equation),

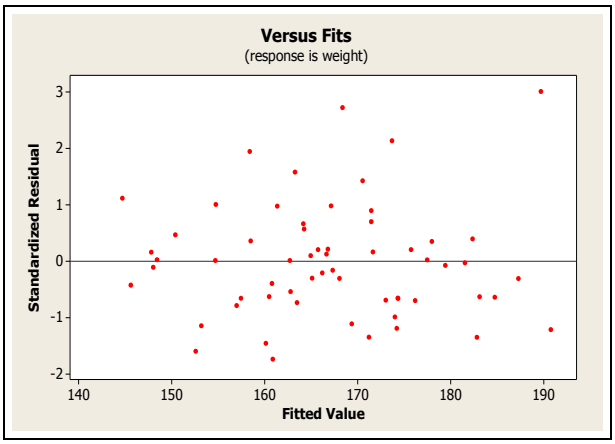
- only height is (partially) significant, but some effect of collinearity is seen by the fact that diastolic blood pressure has the strongest univariate (unconditional) association with weight, but is not the most significant predictor in the combined model,

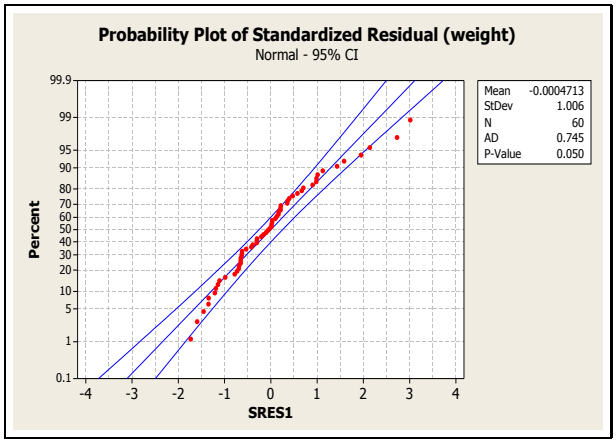
- the estimated correlation matrix for the 6 parameter estimates (intercept and 5 regression coefficients), as computed in Minitab by the `corrmat` macro (the `estat cve,corr` command in Stata gives the same result and is easier to use), is

  ```
   1.00000  -0.28436  -0.02483  -0.00318  -0.18629  -0.96166
  -0.28436   1.00000  -0.17635  -0.04601  -0.10408   0.26833
  -0.02483  -0.17635   1.00000  -0.81260   0.01269  -0.00343
  -0.00318  -0.04601  -0.81260   1.00000  -0.10090  -0.08730
  -0.18629  -0.10408   0.01269  -0.10090   1.00000   0.07173
  -0.96166   0.26833  -0.00343  -0.08730   0.07173   1.00000
  ```

  comments: apart from a strong correlation (-0.96) between $\hat{\beta}_0$ (intercept) and $\hat{\beta}_5$ (which is not alarming), the strongest correlation is, as expected, between the regression coefficients for the two blood pressures (-0.81), whereas the remaining parameter estimates are only weakly correlated; the VIF-values show a similar picture,

- before proceeding further with the analysis, we examine the residuals (below).





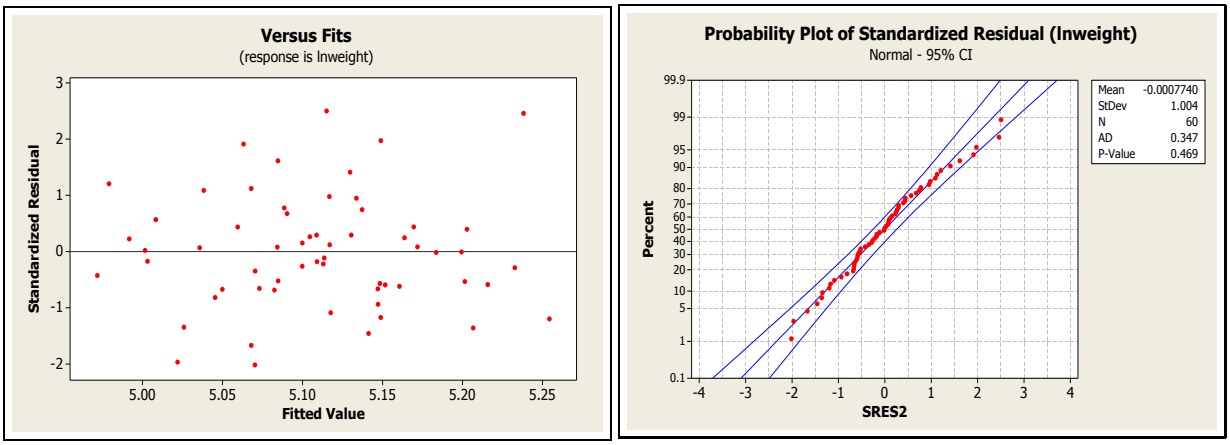
The plot of residuals versus fitted values does not look too bad, but clearly the residuals extend further on the positive than negative side. This is reflected in the curved shape of the normal plot; apparently, the residuals are skewed to the right. The Anderson-Darling test for normal distribution of the residuals has a P-value of 0.05 (other tests give values as low as 0.02); recall that due to the dependence of the residuals, the normality tests are only approximate and indicative. We conclude that the residuals do not look quite satisfactory, but they do not look terrible either. Before making any decision on the need for transformation, we look at the regression diagnostics:

Row	SRES1	TRES1	HI1	COOK1	DFITS1
1	0.90025	0.89864	0.045846	0.006490	0.19698
2	2.13824	2.21416	0.119234	0.103158	0.81467
3	0.66799	0.66453	0.033282	0.002560	0.12330
4	-0.73086	-0.72766	0.117192	0.011818	-0.26512
5	-0.30630	-0.30371	0.328628	0.007654	-0.21249
6	0.35321	0.35033	0.109845	0.002566	0.12306
7	-0.65484	-0.65134	0.067087	0.005139	-0.17467
8	1.00866	1.00883	0.093127	0.017413	0.32328
9	-0.65255	-0.64904	0.045881	0.003413	-0.14233
10	-1.45334	-1.46883	0.033927	0.012363	-0.27526
11	-0.20460	-0.20278	0.059818	0.000444	-0.05115
12	0.16662	0.16511	0.085446	0.000432	0.05047
13	0.01557	0.01542	0.113269	0.000005	0.00551
14	-1.34785	-1.35836	0.109514	0.037237	-0.47636
15	0.16318	0.16170	0.131054	0.000669	0.06280
16	-1.21101	-1.21637	0.205390	0.063178	-0.61841
17	0.70336	0.70003	0.166126	0.016426	0.31245
18	2.72658	2.90886	0.029486	0.037645	0.50703
19	3.01105	3.27016	0.185428	0.343976	1.56024
20	-0.10442	-0.10346	0.162112	0.000352	-0.04551
...					
60	-0.42208	-0.41885	0.271039	0.011040	-0.25540

The most extreme deletion residual, $t = 3.27$ for observation 19, is not significant at the 5% level ($P = 0.11$). The values of the other diagnostics for this observation are somewhat extreme (both relative to the other observations and the guidelines for interpretation). The leverage of observation 5 exceeds slightly the less strict cut-off of $3p/n = 18/60 = 0.3$, but does not seem unusual in any way. We conclude that except for obs. no. 19 there are no suspect observations in the data, and even for obs. 19 there is no evidence (at the 5% level) that it is outlying. Its values for Cook's D and DFits

are however beyond the textbook cut-offs for being noteworthy ($4/n$ and 1, respectively), so there is an indication that this observation is influential.

The question of the skewed residuals remains. A Box-Cox analysis in Minitab (`General Regression` menu) or Stata (`boxcox` command) yields an optimal λ-value of $\hat{\lambda} = -0.65$, however with a very wide 95% confidence interval (Minitab: $(-2.02, 0.69)$, Stata: $(-\infty, 0.685)$). Minitab suggests the power transformation with $\lambda = -0.5$, corresponding to inverse square-root transformation. Because the CI easily includes 0, it seems perhaps more natural to use a log-transformation which has the advantage of simpler interpretations. We show the residual plots for the log-transformed data as well.



Evidently the normal plot looks better, and the normality tests are no longer significant. However, from a practical point of view it is not clear whether the improved residuals warrant the transformation. Probably one would want to perform both analyses to check the robustness of the models and results.

To keep this solution at a reasonable length, the two above suggested model modifications are only discussed briefly:

- *log-transformation of the outcome*: The final statistical model has the same predictors as for the untransformed data, and at approximately the same significance. The predictions obtained from transformed and untransformed data are different, but they give similar results for the present data. The same conclusions hold for the inverse square-root transformation.

- *analysis without obs. no. 19*: The analysis no longer supports the same final model as in the full dataset: the diastolic blood pressure is replaced by the systolic blood pressure, and this variable is not clearly significant ($P = 0.069$ on original scale, $P = 0.046$ on logarithmic scale). Thus, apparently the predictive effect of the diastolic blood pressure depends critically on one observation!

Continuing the analysis of the full dataset with the untransformed outcome, the model reductions immediately suggested are to eliminate age, systolic blood pressure and cholesterol, either in one step or in several steps. (Recall, that the table does not provide evidence that eliminating several of them will turn out to be a nonsignificant model reduction, although it quite likely will.) The Minitab listing for the model with only two predictors is given below:

```
The regression equation is
weight = - 118 + 0.741 dias bp + 3.26 height

Predictor       Coef    SE Coef        T        P      VIF
Constant     -117.57      81.27    -1.45    0.153
dias bp       0.7410     0.2571     2.88    0.006    1.003
height         3.262      1.158     2.82    0.007    1.003

S = 20.5612   R-Sq = 23.1%   R-Sq(adj) = 20.4%
PRESS = 27374.9   R-Sq(pred) = 12.65%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      7242.6      3621.3      8.57    0.001
Residual Error    57     24097.4       422.8
Total             59     31340.0

Source       DF      Seq SS
dias bp       1      3886.2
height        1      3356.4

Unusual Observations
Obs    dias bp     weight         Fit      SE Fit    Residual    St Resid
  2         70     216.00      172.45        6.47       43.55        2.23R
 18         82     225.00      168.29        2.68       56.71        2.78R
 19        110     247.00      189.04        7.78       57.96        3.05R
 60         80     138.00      143.98        8.21       -5.98       -0.32 X

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

The list of unusual observations includes the previous ones, plus observation 60 with a moderately high leverage (nothing to worry about).

We test the model reduction from the full model (F) to this reduced model (R) by an F-statistic, as follows

$$F = \frac{[\text{SSE(R)} - \text{SSE(F)}]/[\text{DFE(R)} - \text{DFE(F)}]}{\text{MSE(F)}} = \frac{[24097.4 - 24009.6]/3}{444.6} = 0.066$$

which under the null hypothesis corresponding to model (R) follows an F-distribution with parameters (3,54). The observed F-value is very close to zero and far from significant ($P > 0.5$). There is absolutely no evidence in these data to indicate that the age, systolic blood pressure, and cholesterol level add something to the prediction by diastolic blood pressure and height alone. The VIF-values are essentially equal to 1, indicating no problems with collinearity. Indeed, the estimates for the two regression coefficients are almost uncorrelated (which means that they can be interpreted independently).

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

Best Subsets Regression: weight versus age, syst bp, ...

Vars	R-Sq	R-Sq(adj)	C-p	S	age	syst bp	diast bp	chol	height
1	12.4	10.9	5.7	21.756			X		
1	11.9	10.4	6.1	21.818					X
1	8.6	7.0	8.5	22.228		X			
2	23.1	20.4	0.2	20.561			X		X
2	20.1	17.3	2.3	20.958		X			X
2	13.5	10.5	7.0	21.810	X				X
3	23.4	19.3	2.0	20.709			X	X	X
3	23.1	19.0	2.2	20.743		X	X		X
3	23.1	19.0	2.2	20.743	X		X		X
4	23.4	17.8	4.0	20.894	X		X	X	X
4	23.4	17.8	4.0	20.896		X	X	X	X
4	23.1	17.5	4.2	20.930	X	X	X		X
5	23.4	16.3	6.0	21.086	X	X	X	X	X

It is seen that Mallows C_p (and the adjusted R^2) point to the selected model as the best one. Both forwards and backwards (stepwise) selection methods with all five predictors also lead to the same model. The same conclusions hold for both the alternative analyses discussed above.