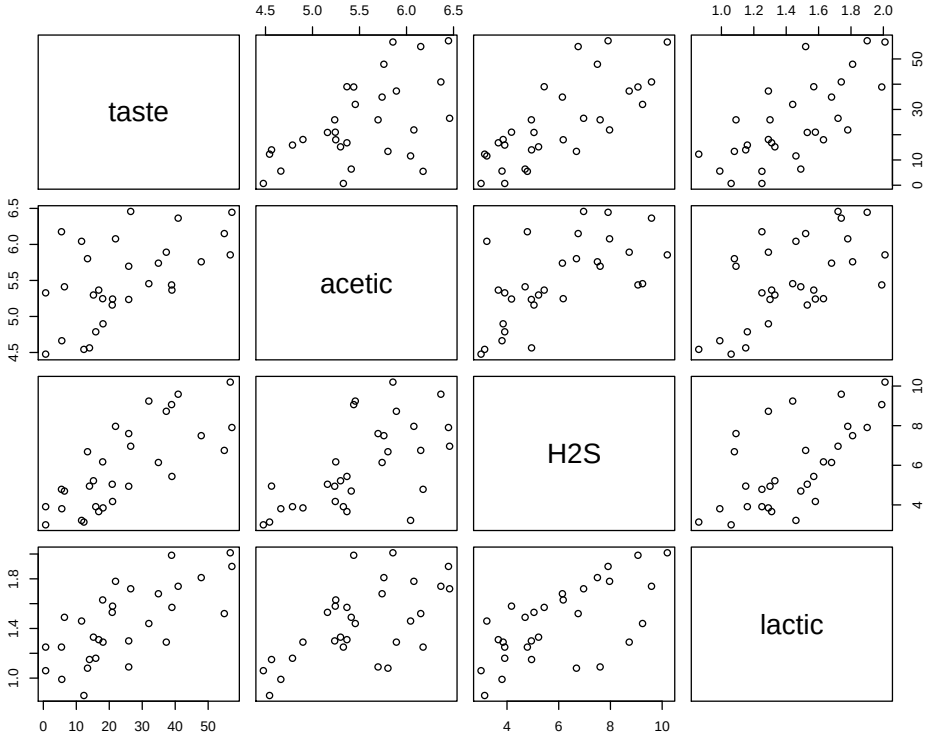
Solutions for 7.5 Exercises

1. **Cheese**

(a)
```
> cheese <- read.csv("cheese.csv")
> plot(cheese)
```



(b)
```
> taste.lm1 <- lm(taste ~ acetic, data = cheese)
> summary(taste.lm1)

Call:
lm(formula = taste ~ acetic, data = cheese)

Residuals:
    Min      1Q  Median      3Q     Max
-29.642  -7.443   2.082   6.597  26.581

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -61.499     24.846  -2.475  0.01964 *
acetic        15.648      4.496   3.481  0.00166 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 13.82 on 28 degrees of freedom
Multiple R-squared: 0.302,  Adjusted R-squared: 0.2771
F-statistic: 12.11 on 1 and 28 DF,  p-value: 0.001658
```

```
> taste.lm2 <- lm(taste ~ H2S, data = cheese)
> summary(taste.lm2)

Call:
lm(formula = taste ~ H2S, data = cheese)

Residuals:
    Min      1Q  Median      3Q     Max
-15.426  -7.611  -3.491   6.420  25.687

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -9.7868     5.9579  -1.643    0.112
H2S           5.7761     0.9458   6.107 1.37e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 10.83 on 28 degrees of freedom
Multiple R-squared: 0.5712, Adjusted R-squared: 0.5558
F-statistic: 37.29 on 1 and 28 DF,  p-value: 1.374e-06

> taste.lm3 <- lm(taste ~ lactic, data = cheese)
> summary(taste.lm3)

Call:
lm(formula = taste ~ lactic, data = cheese)

Residuals:
     Min       1Q   Median       3Q      Max
-19.9439  -8.6839  -0.1095   8.9998  27.4245

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -29.859     10.582  -2.822  0.00869 **
lactic        37.720      7.186   5.249 1.41e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 11.75 on 28 degrees of freedom
Multiple R-squared: 0.4959, Adjusted R-squared: 0.4779
F-statistic: 27.55 on 1 and 28 DF,  p-value: 1.405e-05
```

From these results, we conclude that there is a statistically significant association (at the 0.05 significance level) between taste and each of the three chemical levels.

(c)
```
> taste.lm4 <- lm(taste ~ acetic + H2S + lactic, data = cheese)
> summary(taste.lm4)

Call:
lm(formula = taste ~ acetic + H2S + lactic, data = cheese)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-17.391  -6.612  -1.009   4.908  25.449

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
acetic        0.3277     4.4598   0.073  0.94198
H2S           3.9118     1.2484   3.133  0.00425 **
lactic       19.6705     8.6291   2.280  0.03108 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared: 0.6518, Adjusted R-squared: 0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

The multiple regression is highly significant and accounts for 65% of the variability in taste. However, not all of the terms are needed. In particular, the P-value for acetic acid is 0.94 which implies that it should be removed from the model.

```
> taste.lm5 <- lm(taste ~ H2S + lactic, data = cheese)
> summary(taste.lm5)

Call:
lm(formula = taste ~ H2S + lactic, data = cheese)

Residuals:
    Min      1Q  Median      3Q     Max
-17.343  -6.530  -1.164   4.844  25.618

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -27.592      8.982  -3.072  0.00481 **
H2S            3.946      1.136   3.475  0.00174 **
lactic        19.887      7.959   2.499  0.01885 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 9.942 on 27 degrees of freedom
Multiple R-squared: 0.6517, Adjusted R-squared: 0.6259
F-statistic: 25.26 on 2 and 27 DF,  p-value: 6.551e-07
```

Removing acetic acid from the model results in a lower residual standard error and a higher adjusted R^2. Both lactic acid and H_2S are still significant in this new model. It is the best model.

(Using the *step*() function in R should give the same result — check this.)

(d)
```
> cor(cheese)
           taste    acetic       H2S    lactic
taste  1.0000000 0.5495393 0.7557523 0.7042362
```

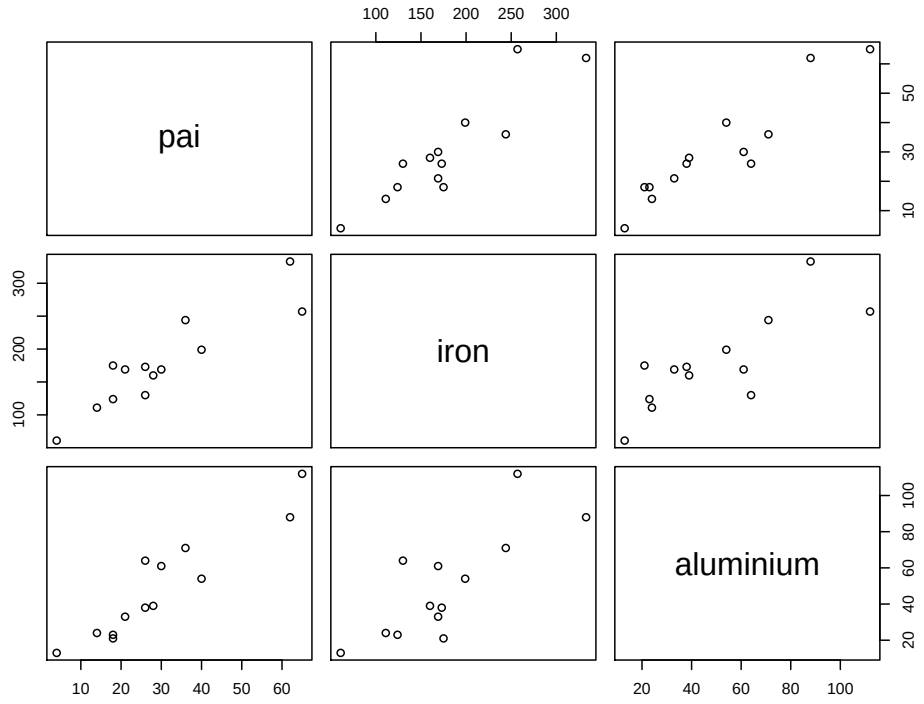
```
acetic 0.5495393 1.0000000 0.6179559 0.6037826
H2S    0.7557523 0.6179559 1.0000000 0.6448123
lactic 0.7042362 0.6037826 0.6448123 1.0000000
```

The relatively high correlations between taste and each of the explanatory variables are consistent with the findings in part (b). There is also high correlation amongst the explanatory variables, which explains why not all three were required in the model in part (c). Given the values of H_2S and lactic acid, we can obtain quite a reasonable estimate of the value of acetic acid The concentration of acetic acid has little additional information about taste, beyond that of H_2S and lactic acid.

(e) Taste is predicted to increase by 3.95 units for each additional unit of H_2S concentration, provided lactic acid concentration remains constant.

2. **Soil Data**

(a)
```
> soil <- data.frame(pai = c(4,18,14,18,26,26,21,30,28,36,65,62,40),
+ iron = c(61,175,111,124,130,173,169,169,160,244,257,333,199),
+ aluminium = c(13,21,24,23,64,38,33,61,39,71,112,88,54))

> plot(soil)
```



The relationships between the response `pai` and both variables `iron` and `aluminium` appears quite linear, as does the relationship between `iron` and `aluminium`.

(b)
```
> soil.lm.iron <- lm(pai ~ iron, data = soil)
> summary(soil.lm.iron)
```

```
Call:
lm(formula = pai ~ iron, data = soil)

Residuals:
     Min       1Q   Median       3Q      Max
-11.3196  -3.3714   0.3174   2.1030  16.9700

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.61117    6.00940  -1.766    0.105
iron          0.22818    0.03168   7.202 1.75e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.694 on 11 degrees of freedom
Multiple R-squared:  0.825,        Adjusted R-squared:  0.8091
F-statistic: 51.86 on 1 and 11 DF,  p-value: 1.749e-05
```

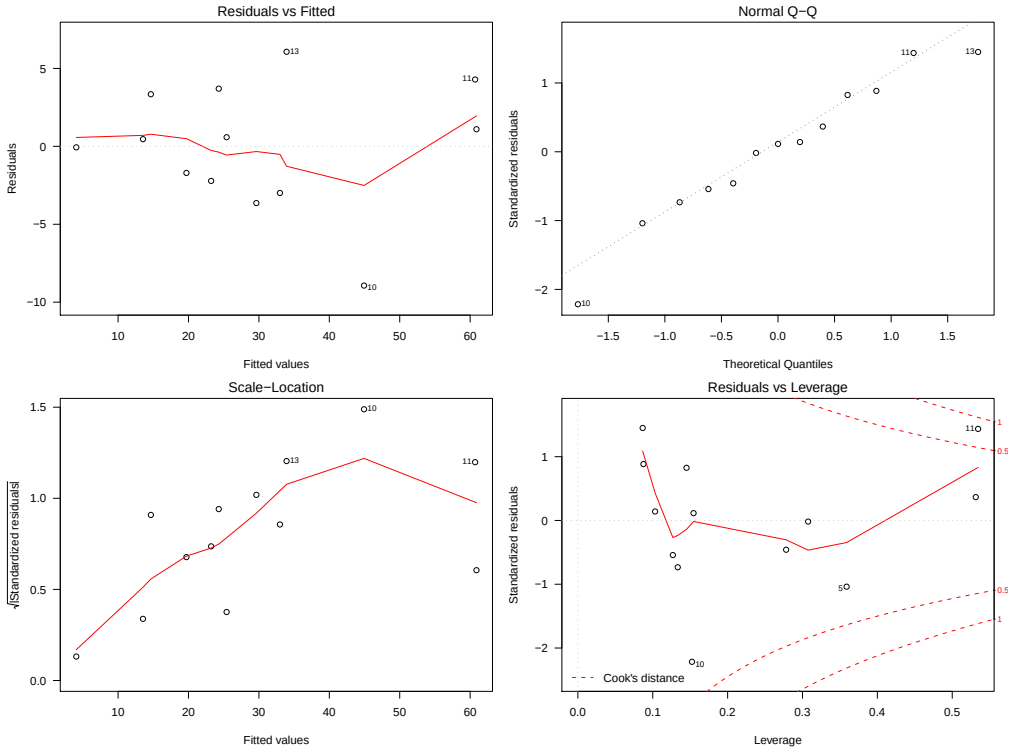
```
> soil.lm.1 <- lm(pai ~ iron + aluminium, data = soil)
> summary(soil.lm.1)
```

```
Call:
lm(formula = pai ~ iron + aluminium, data = soil)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9352 -2.2182  0.4613  3.3448  6.0708

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.35066    3.48467  -2.109 0.061101 .
iron         0.11273    0.02969   3.797 0.003504 **
aluminium    0.34900    0.07131   4.894 0.000628 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.379 on 10 degrees of freedom
Multiple R-squared:  0.9485,        Adjusted R-squared:  0.9382
F-statistic: 92.03 on 2 and 10 DF,  p-value: 3.634e-07
```

(c)
```
> predict(soil.lm.iron, newdata = data.frame(iron = 150), interval = "prediction")
      fit      lwr      upr
1 23.6152 5.938094 41.29230
> predict(soil.lm.1, newdata = data.frame(iron = 150,
+        aluminium=mean(soil$aluminium)), interval = "prediction")
       fit      lwr      upr
1 26.76768 16.48159 37.05377
```

The predicted value is similar, but the prediction interval is wider, which is not surprising for a model which doesn't fit as well.

(d) `> par(mfrow = c(2, 2), las = 1, mar = c(4, 4, 2, 1))`
`> plot(soil.lm.1)`



There is some evidence that the variance increases with the fitted values, but with only 13 points it is difficult to be sure. The Cook's distance plot suggests that point 11 has a substantial impact on the model, but it looks OK on the other plots.

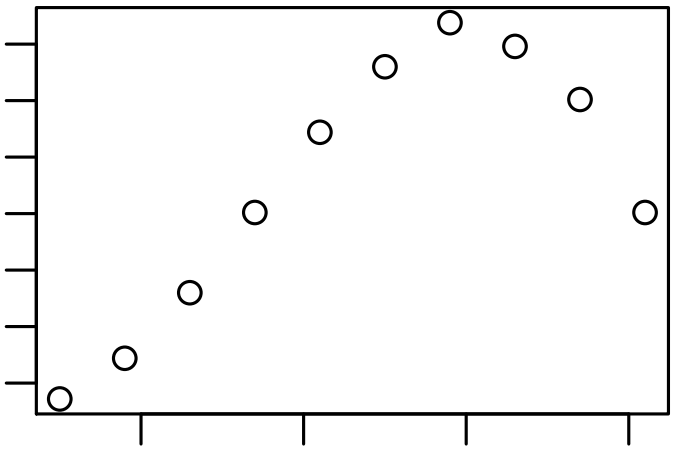
(e) `> step(soil.lm.1, ~.)`

```
Start:  AIC=40.99
pai ~ iron + aluminium

             Df Sum of Sq    RSS    AIC
<none>                    191.79 40.989
- iron        1    276.49 468.28 50.594
- aluminium   1    459.43 651.22 54.881

Call:
lm(formula = pai ~ iron + aluminium, data = soil)

Coefficients:
(Intercept)         iron    aluminium
    -7.3507       0.1127       0.3490
```

Omission of either explanatory variable substantially increases AIC, so both need to remain in the model.

3. **Girths of horses**

(a)
```
> girth <- data.frame(weight = c(2.5, 4.5, 6.5, 8.5, 10.5, 12.5,
+     14.5, 16.5, 18.5, 20.5), stretch = c(0.0136, 0.0172, 0.023,
+     0.0301, 0.0372, 0.043, 0.0469, 0.0448, 0.0401, 0.0301))

> par(mar = c(1, 1, 1, 1))
> plot(stretch ~ weight, data = girth)
```



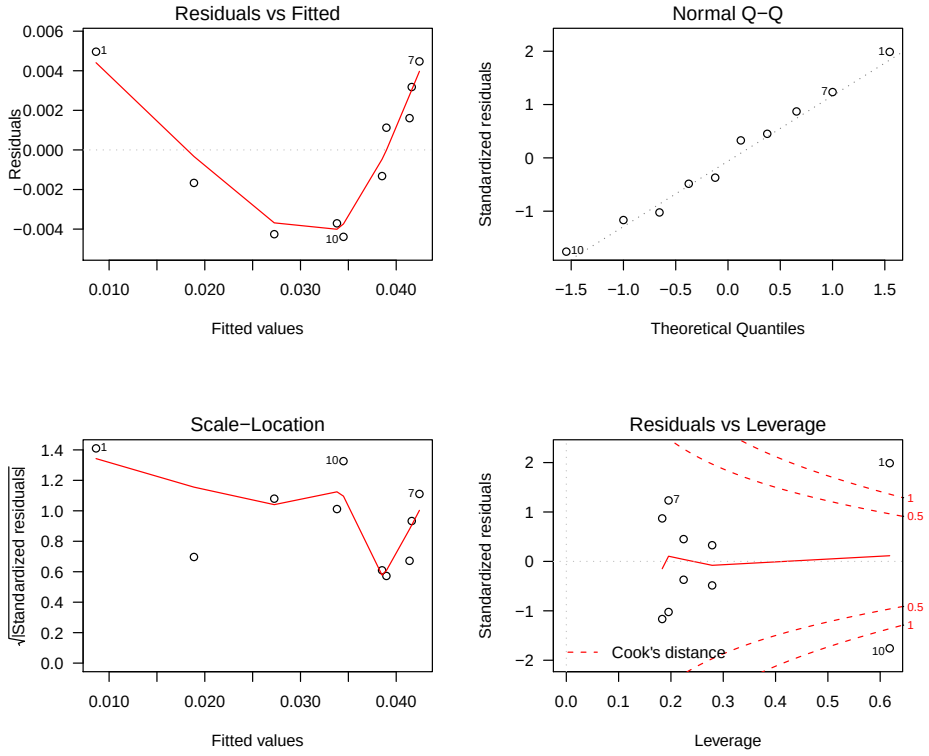
```
> cor(girth)

           weight   stretch
weight  1.0000000 0.7404735
stretch 0.7404735 1.0000000
```

$r = 0.740$. This is of limited usefulness here, because it measures the strength of the *linear association*, and the relationship is not linear.

(b)
```
> girth.lm.2 <- lm(stretch ~ weight + I(weight^2), data = girth)
> summary(girth.lm.2)

Call:
lm(formula = stretch ~ weight + I(weight^2), data = girth)

Residuals:
       Min         1Q     Median         3Q        Max
-4.391e-03 -3.198e-03 -9.848e-05  2.786e-03  4.964e-03

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.738e-03  5.216e-03  -1.292 0.237465
```

```
weight       6.725e-03  1.035e-03   6.496 0.000336 ***
I(weight^2) -2.299e-04  4.396e-05  -5.230 0.001212 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 0.00404 on 7 degrees of freedom
Multiple R-squared: 0.908,  Adjusted R-squared: 0.8817
F-statistic: 34.53 on 2 and 7 DF,  p-value: 0.0002365
```

```
> par(mfrow = c(2, 2), las = 1)
> plot(girth.lm.2)
```



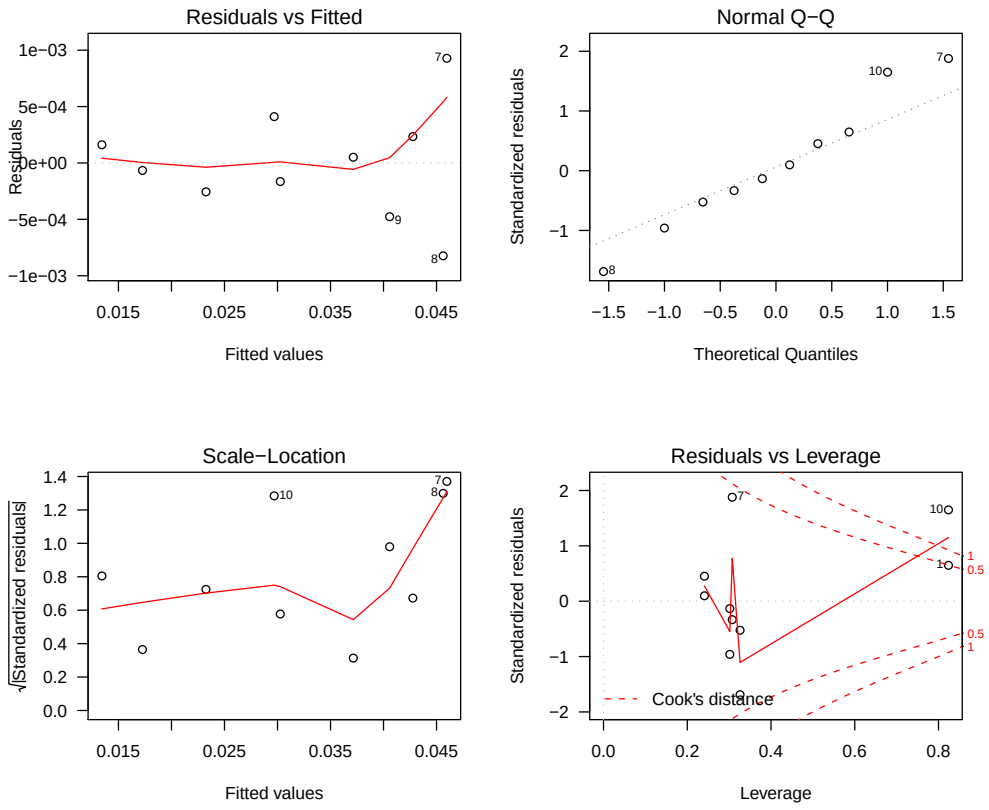
The quadratic term is highly significant, but the residuals suggest a cubic term is needed.

```
> girth.lm.3 <- lm(stretch ~ weight + I(weight^2) + I(weight^3),
+     data = girth)
> summary(girth.lm.3)
```

```
Call:
lm(formula = stretch ~ weight + I(weight^2) + I(weight^3), data = girth)

Residuals:
       Min          1Q      Median          3Q         Max
-8.230e-04 -2.330e-04 -7.576e-06  2.159e-04  9.283e-04

Coefficients:
```

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.344e-02  1.366e-03   9.833 6.37e-05 ***
weight      -1.330e-03  4.765e-04  -2.791   0.0315 *
I(weight^2)  5.919e-04  4.653e-05  12.721 1.45e-05 ***
I(weight^3) -2.382e-05  1.336e-06 -17.836 2.00e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 0.0005938 on 6 degrees of freedom
Multiple R-squared: 0.9983, Adjusted R-squared: 0.9974
F-statistic:  1172 on 3 and 6 DF,  p-value: 1.081e-08
```

```
> par(mfrow = c(2, 2), las = 1)
> plot(girth.lm.3)
```



The cubic term makes a further significant improvement, and the residuals don't suggest a 4th power. The fit is amazingly good (too good?), judging from the R^2 and the residual standard deviation.

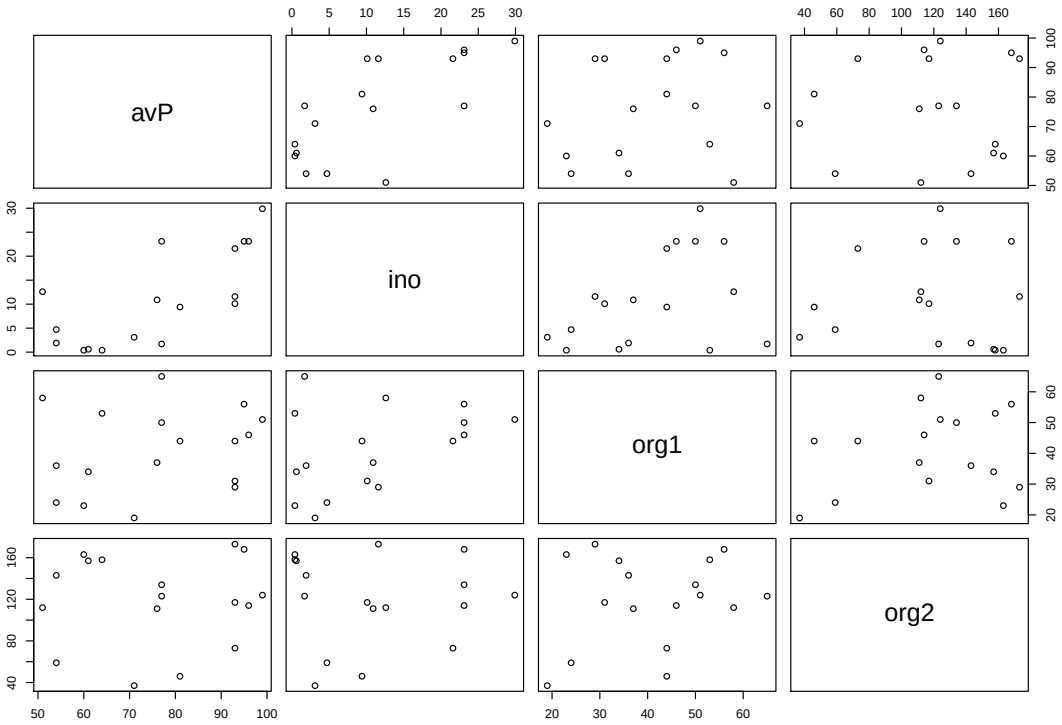
(c)
```
> predict(girth.lm.3, newdata = data.frame(weight = 10.5), interval = "confidence")
         fit        lwr       upr
1 0.03714918 0.03643587 0.0378625
```

Residual $= 0.0372 - 0.03715 = 0.00005$.

4. **Phosphorus content in corn.**

(a) `> Pcorn <- read.csv("Pcorn.csv")`

`> plot(Pcorn)`



`> cor(Pcorn)`

```
             avP          ino      org1         org2
avP  1.00000000  0.720086602 0.2118376  0.029934165
ino  0.72008660  1.000000000 0.3989231 -0.006425396
org1 0.21183758  0.398923129 1.0000000  0.222479243
org2 0.02993417 -0.006425396 0.2224792  1.000000000
```

Available P seems to have a reasonably strong relationship with inorganic P, but little relationship with the two organic P measurements.

(b) `> Pcorn.lm <- lm(avP ~ ino + org1 + org2, data = Pcorn)`
`> step(Pcorn.lm, ~.)`

```
Start:  AIC=89.78
avP ~ ino + org1 + org2

       Df Sum of Sq    RSS    AIC
- org2  1     14.12 2101.3 87.891
- org1  1     38.78 2125.9 88.089
<none>              2087.2 89.776
- ino   1   2139.28 4226.5 99.770
```

```
Step:  AIC=87.89
avP ~ ino + org1

       Df Sum of Sq    RSS    AIC
- org1  1     29.95 2131.2 86.131
<none>              2101.3 87.891
+ org2  1     14.12 2087.2 89.776
- ino   1   2126.54 4227.8 97.776

Step:  AIC=86.13
avP ~ ino

       Df Sum of Sq    RSS    AIC
<none>              2131.2 86.131
+ org1  1     29.95 2101.3 87.891
+ org2  1      5.29 2125.9 88.089
- ino   1   2295.23 4426.5 96.556

Call:
lm(formula = avP ~ ino, data = Pcorn)

Coefficients:
(Intercept)          ino
     62.569        1.229


> Pcorn.lm1 <- lm(avP ~ 1, data = Pcorn)
> step(Pcorn.lm1, ~. + ino + org1 + org2)
Start:  AIC=96.56
avP ~ 1

       Df Sum of Sq    RSS    AIC
+ ino   1   2295.23 2131.2 86.131
<none>              4426.5 96.556
+ org1  1    198.64 4227.8 97.776
+ org2  1      3.97 4422.5 98.541

Step:  AIC=86.13
avP ~ ino

       Df Sum of Sq    RSS    AIC
<none>              2131.2 86.131
+ org1  1     29.95 2101.3 87.891
+ org2  1      5.29 2125.9 88.089
- ino   1   2295.23 4426.5 96.556

Call:
```

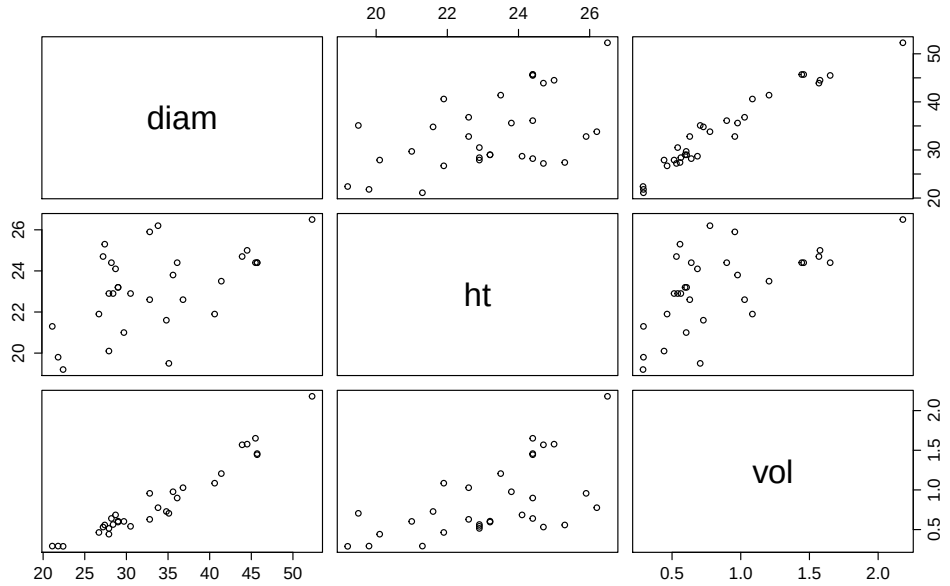


Figure 1: > plot(timber)

```
lm(formula = avP ~ ino, data = Pcorn)

Coefficients:
(Intercept)          ino
     62.569        1.229
```

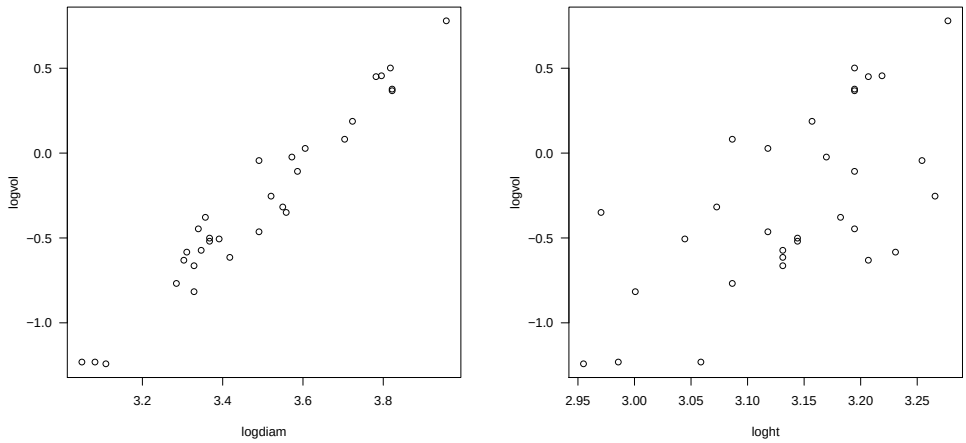
For either backward elimination or forward selection, the smallest AIC is for the model with only inorganic P, confirming the impressions from the graphs.

5. **Estimating timber volume**

```
> timber <- read.csv("timber.csv")
> plot(timber)
```

From looking at the graphs (and based on the geometry of the problem), taking logs of each variable might be a good place to start.

```
> timber$logvol = log(timber$vol)
> timber$logdiam = log(timber$diam)
> timber$loght = log(timber$ht)

> par(mfrow = c(1, 2), las = 1, mar = c(4, 6, 2, 1))
> plot(logvol ~ logdiam, data = timber)
> plot(logvol ~ loght, data = timber)
```



log(volume) vs log(diameter) looks very promising.

log(volume) vs log(height) has more noise, but also looks linear.

```
> timber.lm <- lm(logvol ~ logdiam + loght, data = timber)
> summary(timber.lm)

Call:
lm(formula = logvol ~ logdiam + loght, data = timber)

Residuals:
      Min        1Q    Median        3Q       Max
-0.171144 -0.048326  0.006864  0.061369  0.128931

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.6912     0.5481 -19.506  < 2e-16 ***
logdiam       1.9835     0.0751  26.411  < 2e-16 ***
loght         1.1078     0.2037   5.437  8.4e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 0.08148 on 28 degrees of freedom
Multiple R-squared: 0.9776, Adjusted R-squared: 0.976
F-statistic: 611.3 on 2 and 28 DF,  p-value: < 2.2e-16
```

$R^2 = 0.978$, which is larger than the R^2 for models fitted without transforming.