Exercises 3

1. The data in the accompanying R output relate to a study of the quantity of vitamin B_2 in turnip green (from Anderson R.L. and Bancroft T.L. (1959), Statistical Theory in Research, McGraw-Hill). Three explanatory variables are:

   x_1 = radiation in relative gram claeries per minute during the preceding half day of sunlight (coded by dividing by 100),
   x_2 = average soil moisture tension (coded by dividing by 100),
   x_3 = air temperature in degrees Farenheit (coded by dividing by 10),

   and the response variable is

   Y = milligrams of vitamin B_2 per gram of turnip green.

   Go through the R output and explain what you can deduce from it as follows:

   (a) Comment on the matrix plot of the data.
   (b) State algebraically in terms of unknown parameters the form of model 1.
   (c) What do you deduce from the P-value associated with the 'F-statistic' for model 1?
   (d) From the value R^2, and the residuals vs. fitted and normal plot of standardised residuals only, comment on whether model 1 is an adequate fit to the data. Explain what output prompted the fitting of model 2.
   (e) Proceeding as in (d), comment on the fit of model 2.
   (f) What can you deduce from the t-test associated with the coefficient of `I(x2^2)` in the output for model 2? (`I(x2^2)` in R means that variable x_2^2 has been included.)
   (g) Is there any variable that you could drop from model 2? Explain your answer.
   (h) Write down the *fitted* form of model 3 and give reasons why you could choose model 3 as an adequate model (for this model, also comment on the other residual plots).

2. Let
$$Z_i = \mathbf{u}_i^T \boldsymbol{\beta}_Z + d_i,\ i = 1, \dots, N,$$
   d_i iid with $Ed_i = 0,\ \text{var}(d_i) = \sigma_Z^2$.
   Let $Y_i = aZ_i + \mathbf{u}_i^T \mathbf{b},\ \mathbf{x}_i = \mathbf{A}^T \mathbf{u}_i$, where $a \in I\!R,\ \mathbf{b} \in I\!R^p$, $\mathbf{A}$ invertible $p \times p$-matrix.

   (a) Show that the linear model
$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta}_Y + e_i,\ i = 1, \dots, N,$$
   holds with
$$\boldsymbol{\beta}_Y = \mathbf{A}^{-1}(a\boldsymbol{\beta}_Z + \mathbf{b}),\ Ee_i = 0,\ \text{var}(e_i) = \sigma_Y^2 = a^2 \sigma_Z^2.$$
   (b) Show that for the LS-estimators $\hat{\boldsymbol{\beta}}_Y,\ \hat{\boldsymbol{\beta}}_Z$ of $\boldsymbol{\beta}_Y,\ \boldsymbol{\beta}_Z$:
$$\hat{\boldsymbol{\beta}}_Y = \mathbf{A}^{-1}(a\hat{\boldsymbol{\beta}}_Z + \mathbf{b}).$$
   Hint: using the matrix notation introduced for linear regression in the notes, and defining $\mathbf{U}$ and $\mathbf{Z}$ by analogy, you get
$$\mathbf{X} = \mathbf{U}\mathbf{A},\ \mathbf{Y} = a\mathbf{Z} + \mathbf{U}\mathbf{b}.$$
   (Why?)
   (c) Show that for $\hat{\sigma}_Y^2,\ \hat{\sigma}_Z^2$ estimating $\sigma_Y^2,\ \sigma_Z^2$:
$$\hat{\sigma}_Y^2 = a^2 \hat{\sigma}_Z^2.$$

   These results are called "linear equivariance" (or "affine equivariance") and mean that the LS-estimator behaves automatically correctly if the response and predictors are linearly transformed.

```
> #    exercises 3, question 1, R output

> # Data frame turnip contains data on x1,x2,x3,y
> turnip
     x1    x2  x3     y
1  1.76 0.070 7.8 110.4
2  1.55 0.070 8.9 102.8
3  2.73 0.070 8.9 101.0
4  2.73 0.070 7.2 108.4
5  2.56 0.070 8.4 100.7
6  2.80 0.070 8.7 100.3
7  2.80 0.070 7.4 102.0
8  1.84 0.070 8.7  93.7
9  2.16 0.070 8.8  98.9
10 1.98 0.020 7.6  96.6
11 0.59 0.020 6.5  99.4
12 0.80 0.020 6.7  96.2
13 0.80 0.020 6.2  99.0
14 1.05 0.020 7.0  88.4
15 1.80 0.020 7.3  75.3
16 1.80 0.020 6.5  92.0
17 1.77 0.020 7.6  82.4
18 2.30 0.020 8.2  77.1
19 2.03 0.474 7.6  74.0
20 1.91 0.474 8.3  65.7
21 1.91 0.474 8.2  56.8
22 1.91 0.474 6.9  62.1
23 0.76 0.474 7.4  61.0
24 2.13 0.474 7.6  53.2
25 2.13 0.474 6.9  59.4
26 1.51 0.474 7.5  58.7
27 2.05 0.474 7.6  58.0

> # Model 1
> turnip.lm1<-lm(y~x1+x2+x3, turnip)
> summary(turnip.lm1)

Call:
lm(formula = y ~ x1 + x2 + x3, data = turnip)
Residuals:
    Min      1Q  Median      3Q     Max
-22.097  -3.790   1.956   5.486  16.980

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   82.069     19.831   4.138 0.000399 ***
x1             2.276      3.657   0.623 0.539736
x2           -77.831      9.354  -8.321 2.17e-08 ***
x3             1.640      2.933   0.559 0.581466
---
Residual standard error: 9.876 on 23 degrees of freedom
Multiple R-Squared: 0.7549,     Adjusted R-squared: 0.7229
F-statistic: 23.61 on 3 and 23 DF,  p-value: 3.306e-07
> # Residual plots attached (the commands have been edited out)
```

```
> # Model 2
> turnip.lm2<-update(turnip.lm1, .~.+I(x2^2), turnip)
> summary(turnip.lm2)

Call:
lm(formula = y ~ x1 + x2 + x3 + I(x2^2), data = turnip)

Residuals:
    Min      1Q  Median      3Q     Max
-11.889  -3.490  -0.632   2.772  13.957

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   119.571     13.676   8.743 1.31e-08 ***
x1             -3.367      2.438  -1.381   0.1811
x2            542.504    100.526   5.397 2.03e-05 ***
x3             -5.026      2.109  -2.383   0.0263 *
I(x2^2)     -1209.047    195.603  -6.181 3.20e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.104 on 22 degrees of freedom
Multiple R-Squared: 0.9104,     Adjusted R-squared: 0.8941
F-statistic:  55.9 on 4 and 22 DF,  p-value: 3.282e-11

> # Residual plots attached (commands have been edited out)

> # Model 3

> turnip.lm3<-update(turnip.lm2, .~.-x1, turnip)
> summary(turnip.lm3)

Call:
lm(formula = y ~ x2 + x3 + I(x2^2), data = turnip)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9663  -3.4432  -0.8141   4.2950  13.2652

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   120.627     13.922   8.665 1.06e-08 ***
x2            490.414     95.006   5.162 3.12e-05 ***
x3             -5.716      2.089  -2.736   0.0118 *
I(x2^2)     -1107.853    184.910  -5.991 4.14e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.223 on 23 degrees of freedom
Multiple R-Squared: 0.9027,     Adjusted R-squared:  0.89
F-statistic: 71.09 on 3 and 23 DF,  p-value: 8.747e-12

> # Residual plots attached (commands have been edited out)
```

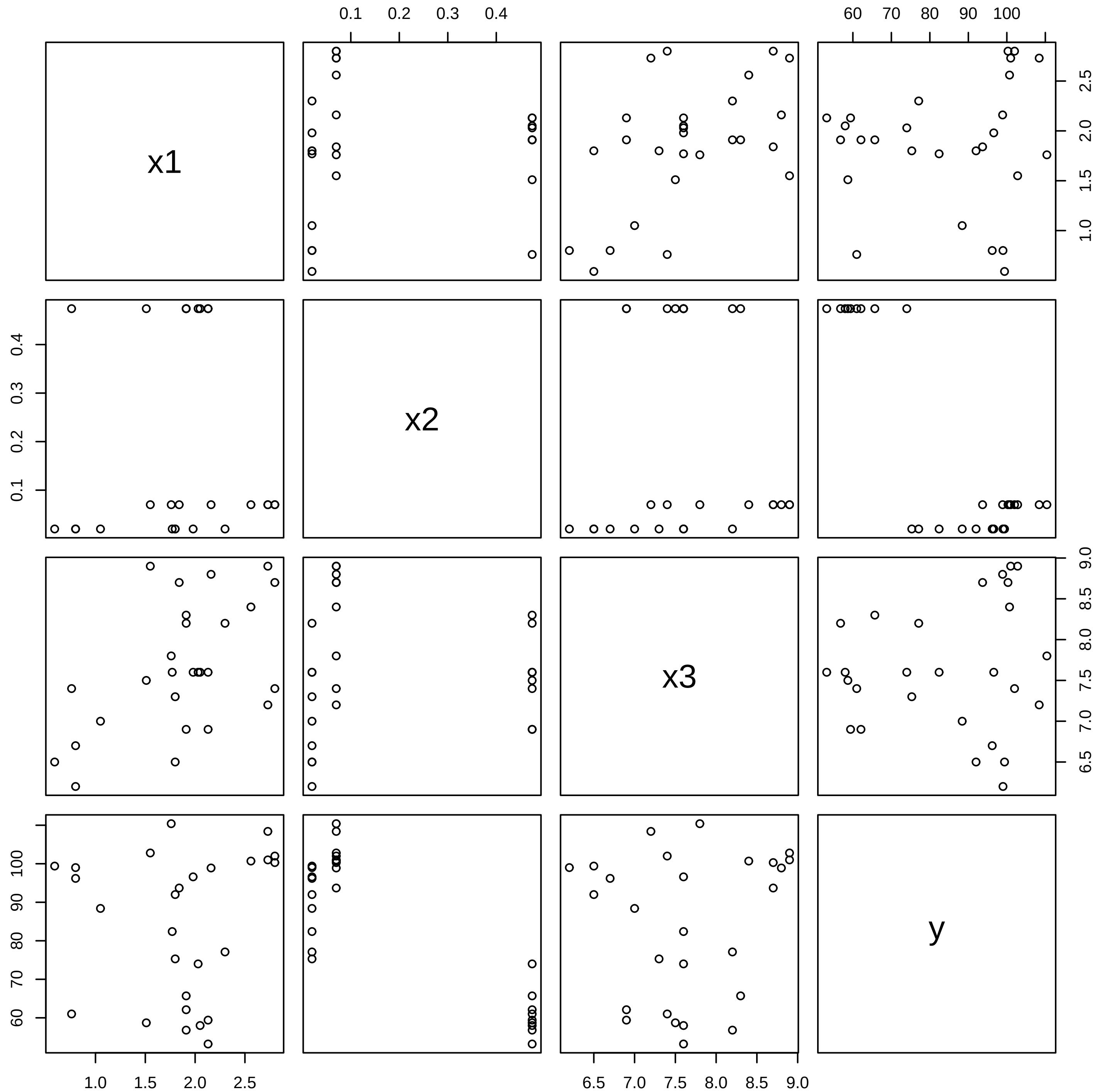

Figure 1: Matrix plot for turnip data.

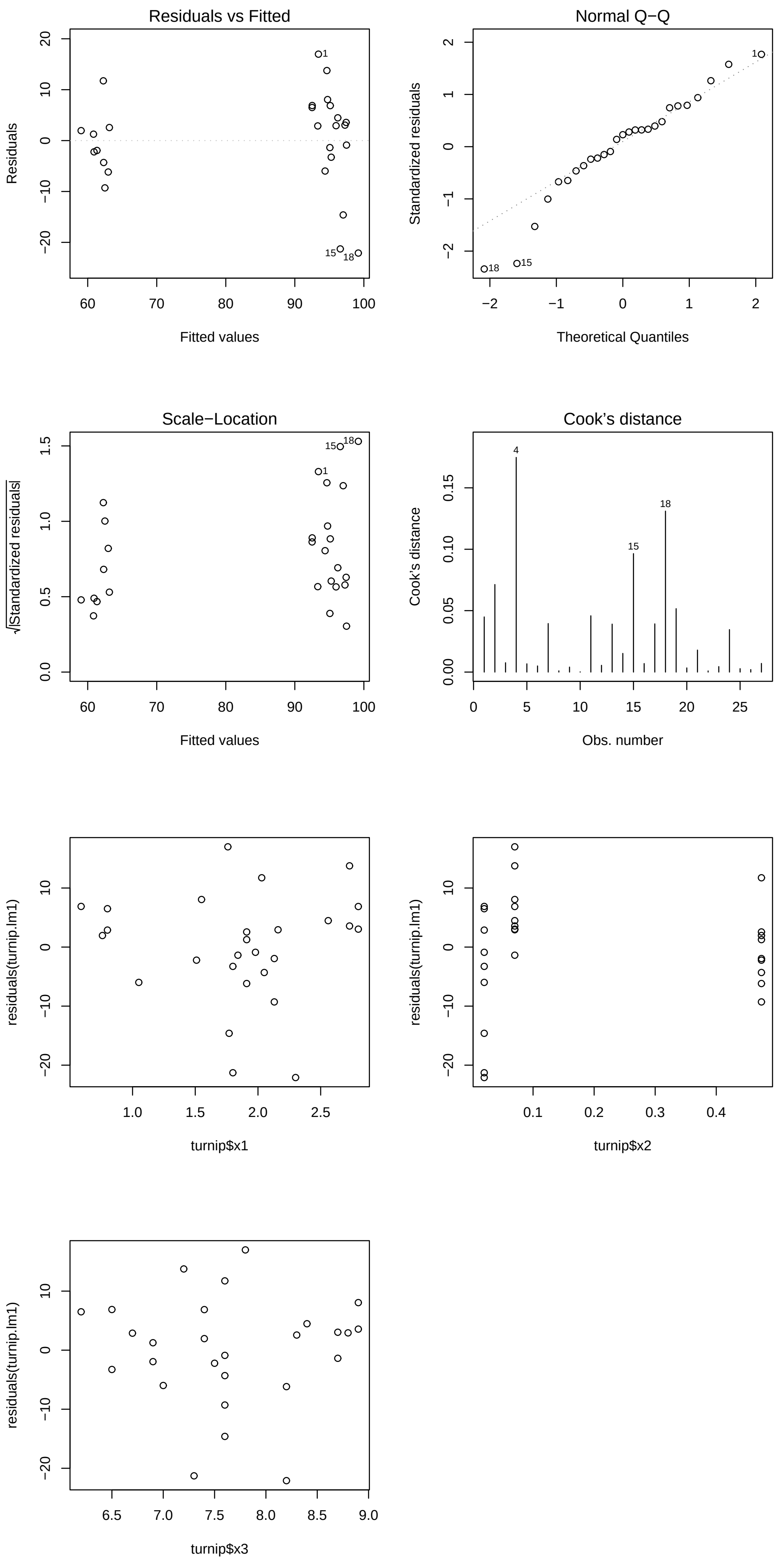

Figure 2: Residual plots for model 1.

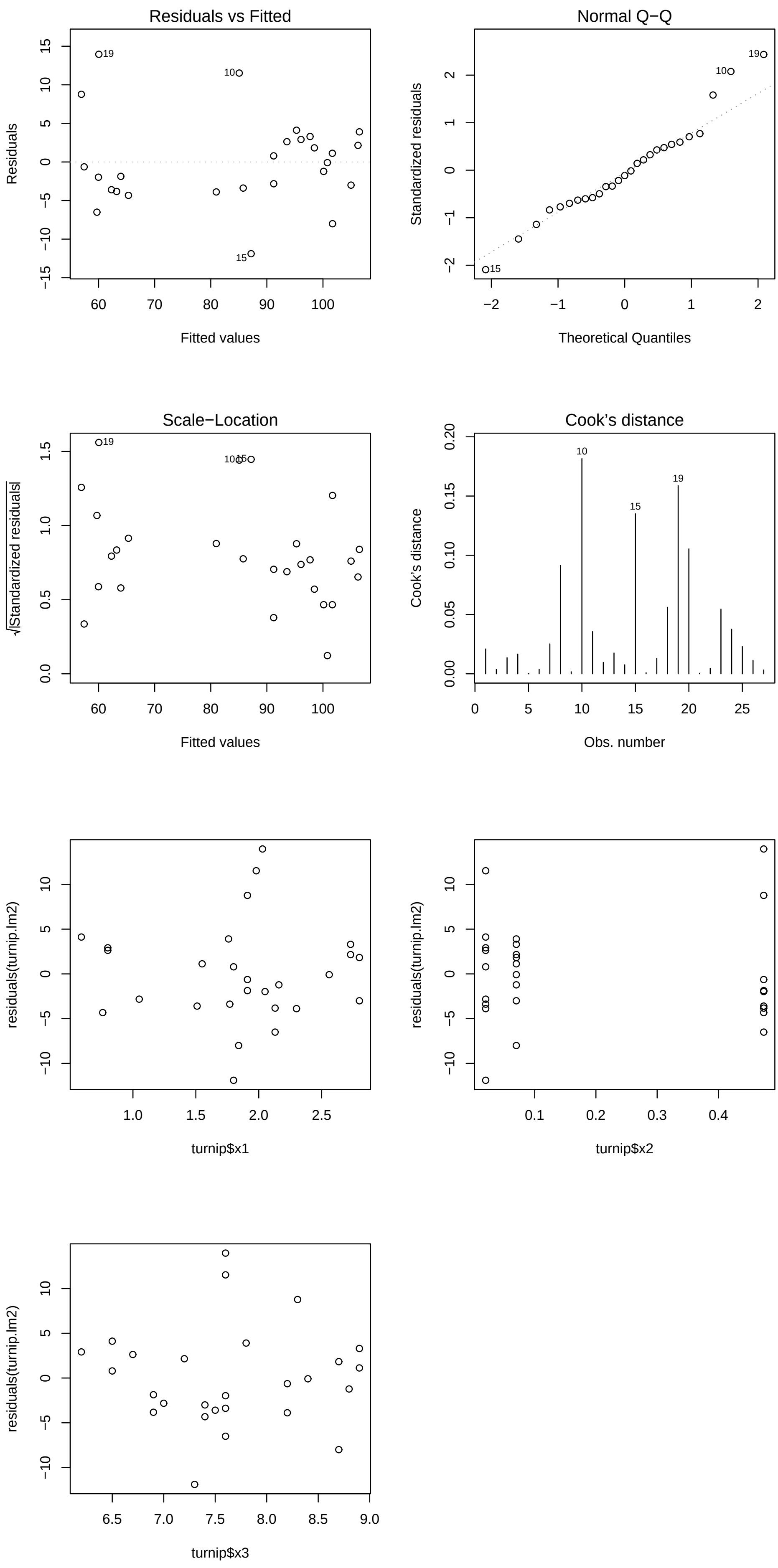

Figure 3: Residual plots for model 2.

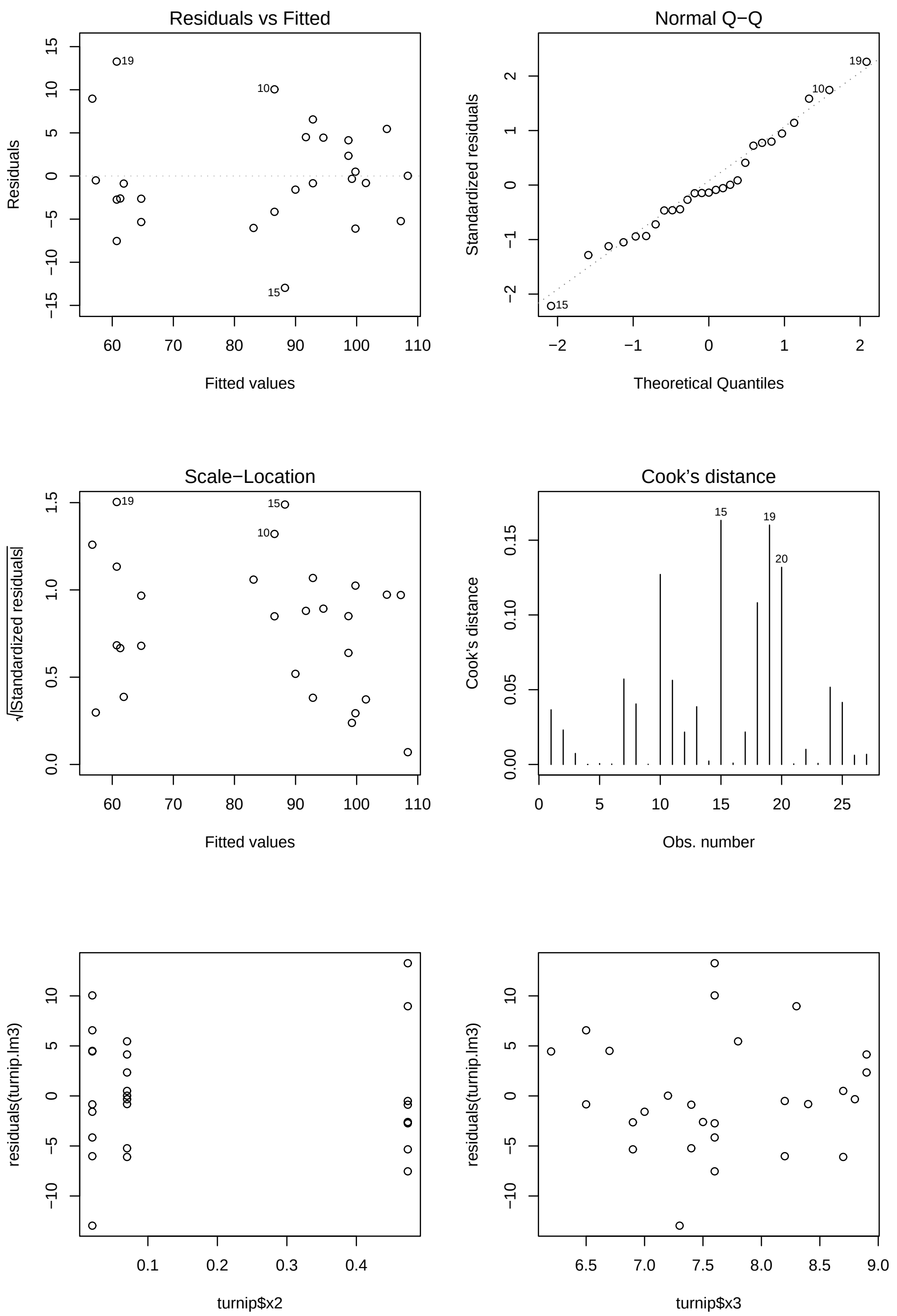

Figure 4: Residual plots for model 3.