5.3

October 16, 2021

[75]:
```
library("dplyr")
```

[76]:
```
rawData = read.table("Table5.11.txt", header=TRUE, sep='\t')
```

[77]:
```
head(rawData)
```

A data.frame: 6 × 4

	Quarter <int>	Date <dbl>	Sales <dbl>	PDI <dbl>
1	1	64	37.0	109
2	2	64	33.5	115
3	3	64	30.8	113
4	4	64	37.9	116
5	1	65	37.4	118
6	2	65	31.6	120

[78]:
```
theData <- rawData %>%
    mutate(q1 = 1 * (Quarter == 1),
           q2 = 1 * (Quarter == 2),
           q3 = 1 * (Quarter == 3),
           q4 = 1 * (Quarter == 4),
           w = 1 * (Quarter == 4 | Quarter == 1),
           s = 1 * (Quarter == 2 | Quarter == 3),
           year = Date)
```

[79]:
```
head(theData)
```

A data.frame: 6 × 11

	Quarter <int>	Date <dbl>	Sales <dbl>	PDI <dbl>	q1 <dbl>	q2 <dbl>	q3 <dbl>	q4 <dbl>	w <dbl>	s <d
1	1	64	37.0	109	1	0	0	0	1	0
2	2	64	33.5	115	0	1	0	0	0	1
3	3	64	30.8	113	0	0	1	0	0	1
4	4	64	37.9	116	0	0	0	1	1	0
5	1	65	37.4	118	1	0	0	0	1	0
6	2	65	31.6	120	0	1	0	0	0	1

[80]:
```
lmfit <- lm(Sales ~ q1 + q2 + q3 + PDI + year , data = theData)
```

[81]:
```
lmfit2 <- lm(Sales ~ w + PDI + year, data = theData)
```

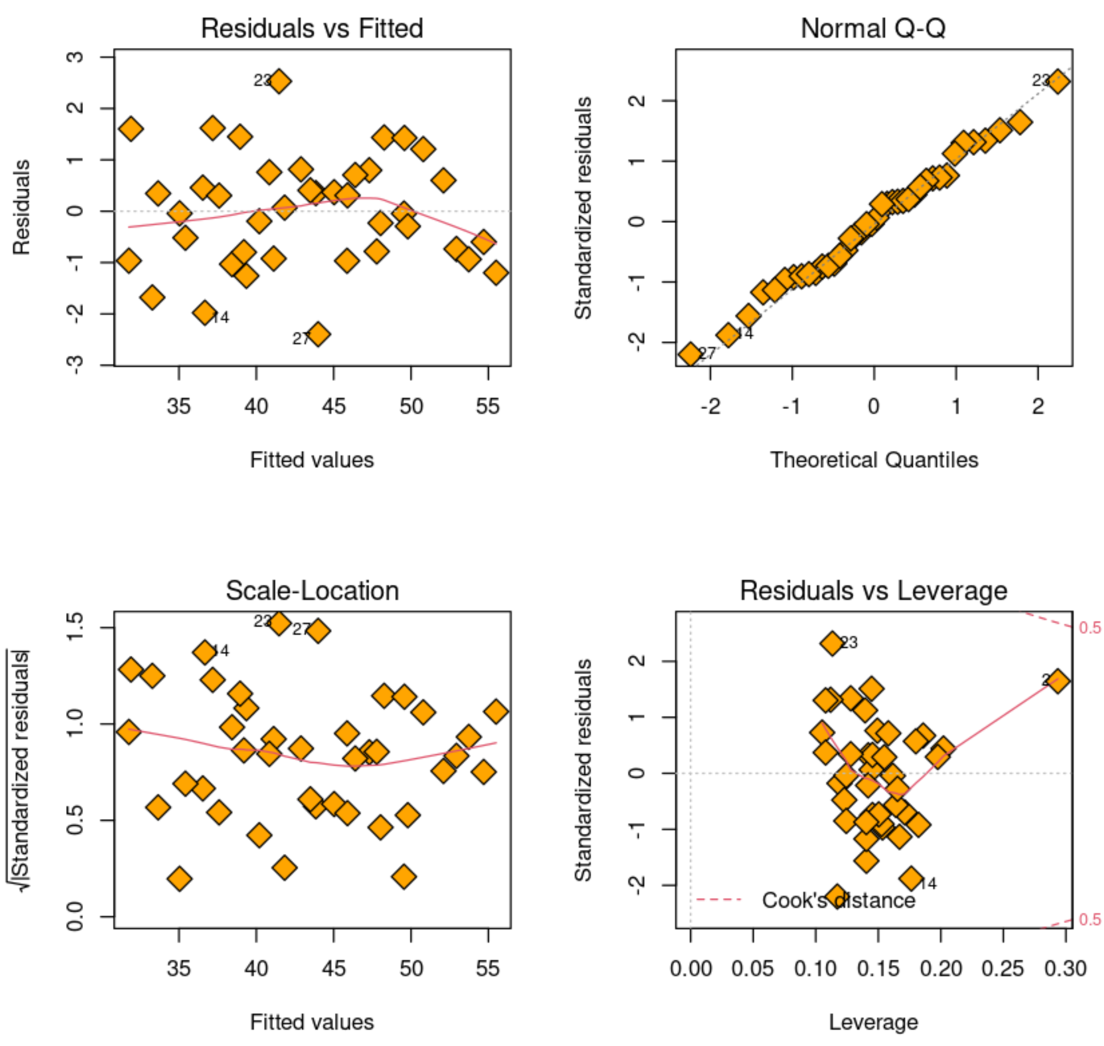
[82]:
```
summary(lmfit)
```

```
Call:
lm(formula = Sales ~ q1 + q2 + q3 + PDI + year, data = theData)

Residuals:
     Min       1Q   Median       3Q      Max
-2.39593 -0.82911  0.01367  0.71883  2.53010

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.90562   36.79196  -0.677    0.503
q1           -0.16463    0.70781  -0.233    0.817
q2           -5.57146    0.58344  -9.549 3.75e-11 ***
q3           -5.45030    0.53889 -10.114 8.72e-12 ***
PDI           0.12713    0.06694   1.899    0.066 .
year          0.74615    0.69160   1.079    0.288
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.159 on 34 degrees of freedom
Multiple R-squared:  0.9737,	Adjusted R-squared:  0.9698
F-statistic: 251.8 on 5 and 34 DF,  p-value: < 2.2e-16
```

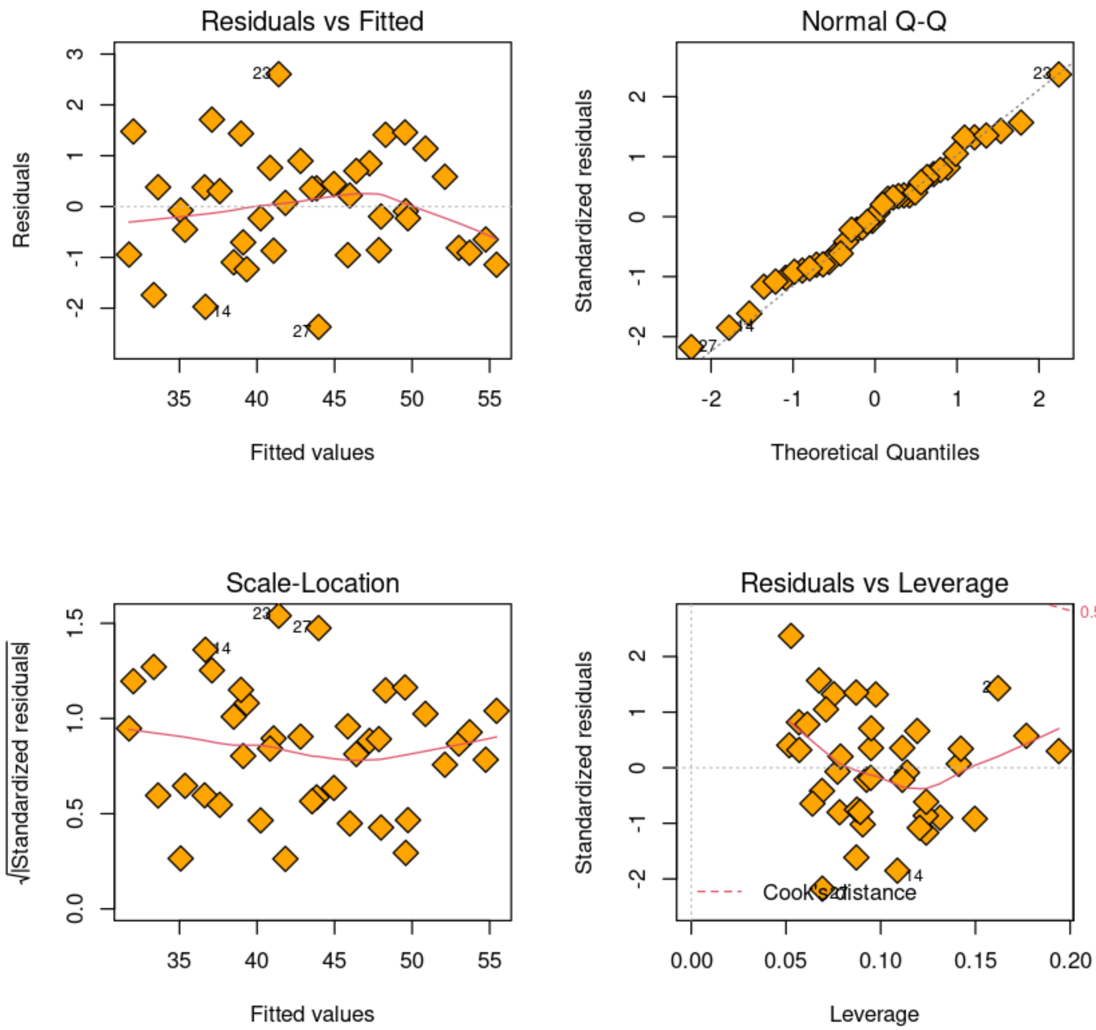
[83]:
```
summary(lmfit2)
```

```
Call:
lm(formula = Sales ~ w + PDI + year, data = theData)

Residuals:
     Min       1Q   Median       3Q      Max
-2.36790 -0.85939 -0.00198  0.71528  2.60341

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.72168   26.16019  -0.907  0.37055
w             5.43467    0.35747  15.203  < 2e-16 ***
PDI           0.13933    0.04703   2.963  0.00538 **
year          0.62065    0.48780   1.272  0.21141
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.128 on 36 degrees of freedom
Multiple R-squared:  0.9736,	Adjusted R-squared:  0.9714
F-statistic: 443.1 on 3 and 36 DF,  p-value: < 2.2e-16
```

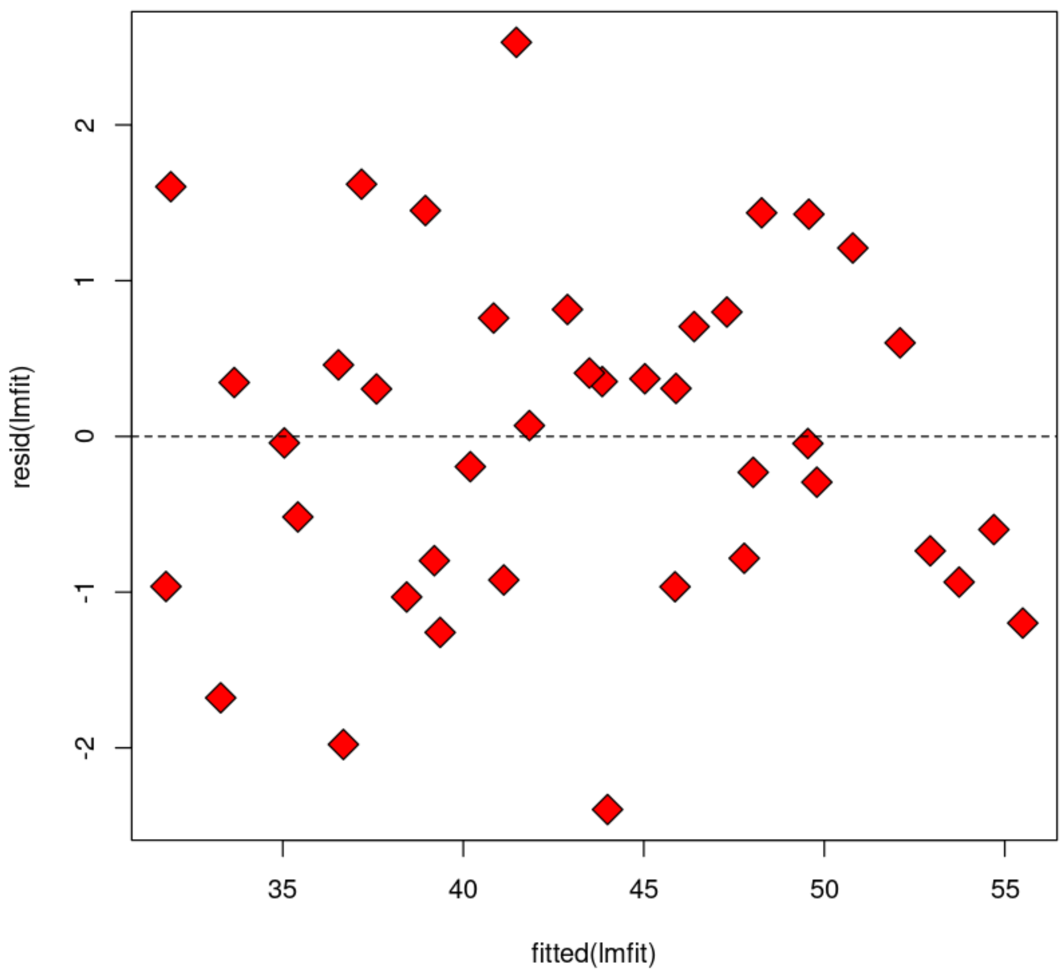
[84]:
```
par(mfrow=c(2,2))
plot(lmfit, pch=23 ,bg='orange',cex=2)
```



[85]:
```
par(mfrow=c(2,2))
plot(lmfit2, pch=23 ,bg='orange',cex=2)
```

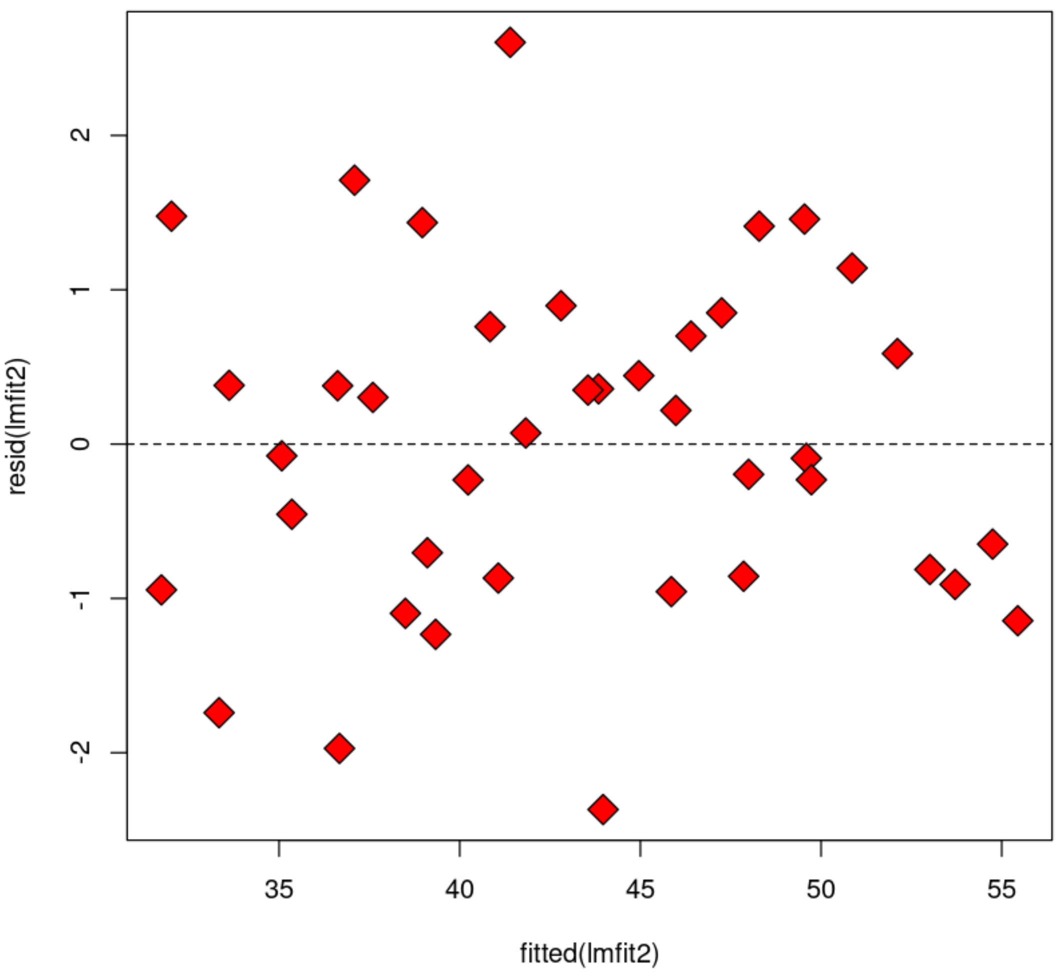


[86]:
```
plot(fitted(lmfit), resid(lmfit), pch=23, bg='red', cex=2)
abline(h=0, lty=2)
```



[87]:

```
plot(fitted(lmfit2), resid(lmfit2), pch=23, bg='red', cex=2)
abline(h=0, lty=2)
```



Using the additional seasonal variable, the model is expanded to be

$$S_t = \beta_0 + \beta_1 \mathrm{PDI}_t + \beta_2 Z_t + \varepsilon_t$$

where Z_t is the zero-one variable described above and β_2 is a parameter that measures the seasonal effect. Note that the model above can be represented by the two models (one for the cold weather quarters where $Z_t = 1$) and the other for the warm quarters where $Z_t = 0$):

Winter season:

$$S_t = (\beta_0 + \beta_2) + \beta_1 \mathrm{PDI}_t + \varepsilon_t$$

Summer season:

$$S_t = \beta_0 + \beta_1 \mathrm{PDI}_t + \varepsilon_t$$

Thus, the model represents the assumption that sales can be approximated by a linear function of PDI, in one line for the winter season and one for the summer season. The lines are parallel; that is, the marginal effect of changes in PDI is the same in both seasons. The level of sales, as reflected by the intercept, is different in each season.

The regression results are summarized in Table above and the index plot of the standardized residuals is shown in Figure above. We see that all indications of the seasonal pattern have been removed. Furthermore, the precision of the estimated marginal effect of PDI increased. The confidence interval is now ($\$186,520,$, $\$210,880$). Also, the seasonal effect has been quantified and we can say that for a fixed level of PDI the winter season brings between $\$4,734,109$ and $\$6,194,491$ over the summer season (with 95% confidence).

5_4

October 16, 2021

[45]:
```
library("dplyr")
```

[46]:
```
rawData = read.table("Table5.12.txt", header=TRUE, sep='\t')
```

[48]:
```
theData <- rawData %>%
    mutate(r2 = 1 * (region == 2),
           r3 = 1 * (region == 3),
           r4 = 1 * (region == 4),
          year = y)
```

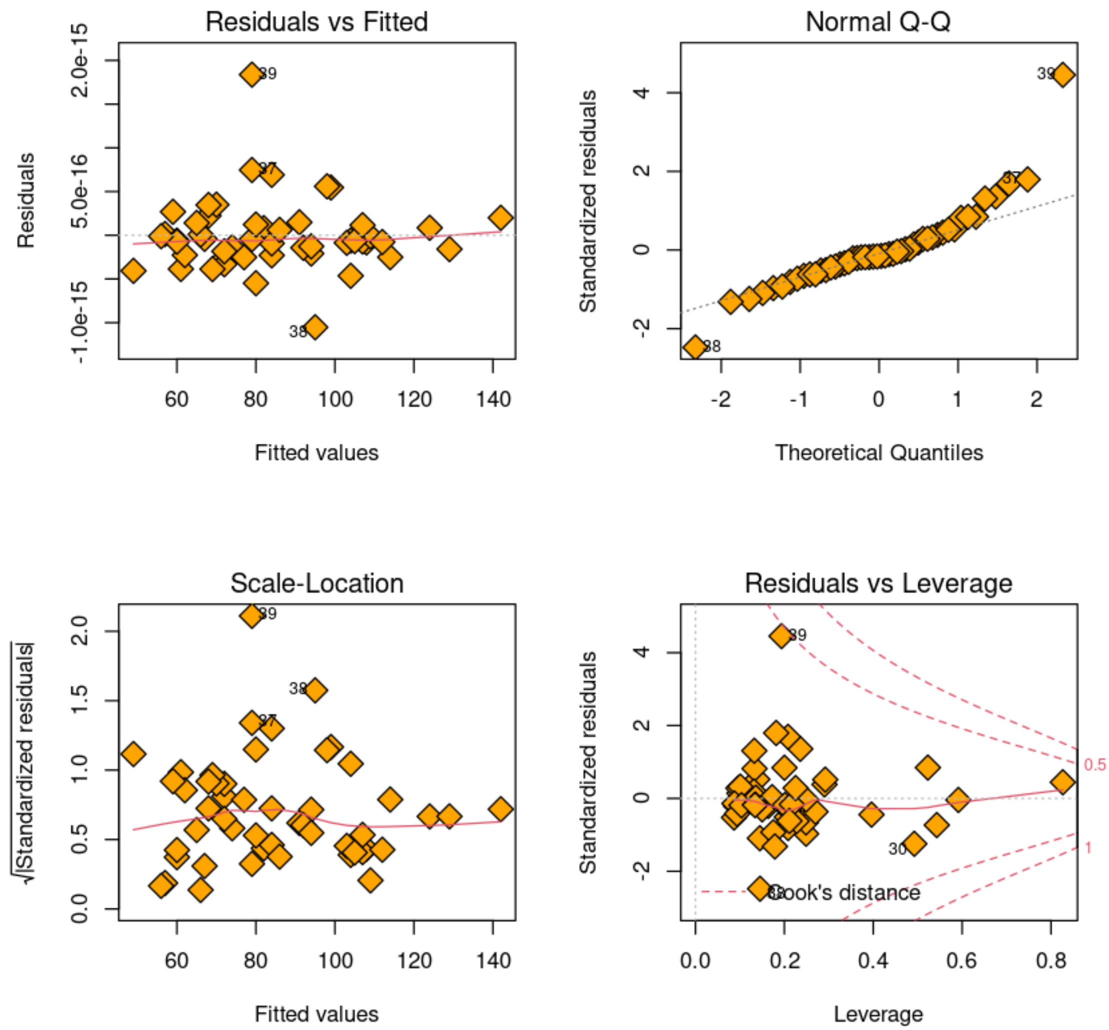
[51]:
```
rawData$region <- as.factor(rawData$region)
```

[52]:
```
rawData$year <- rawData$y
```

[53]:
```
lmfit <- lm(y ~ x1 + x2 + x3 + r2 + r3 + r4 + year + x1*year + x2*year +
 ↪x3*year, data = theData)
```

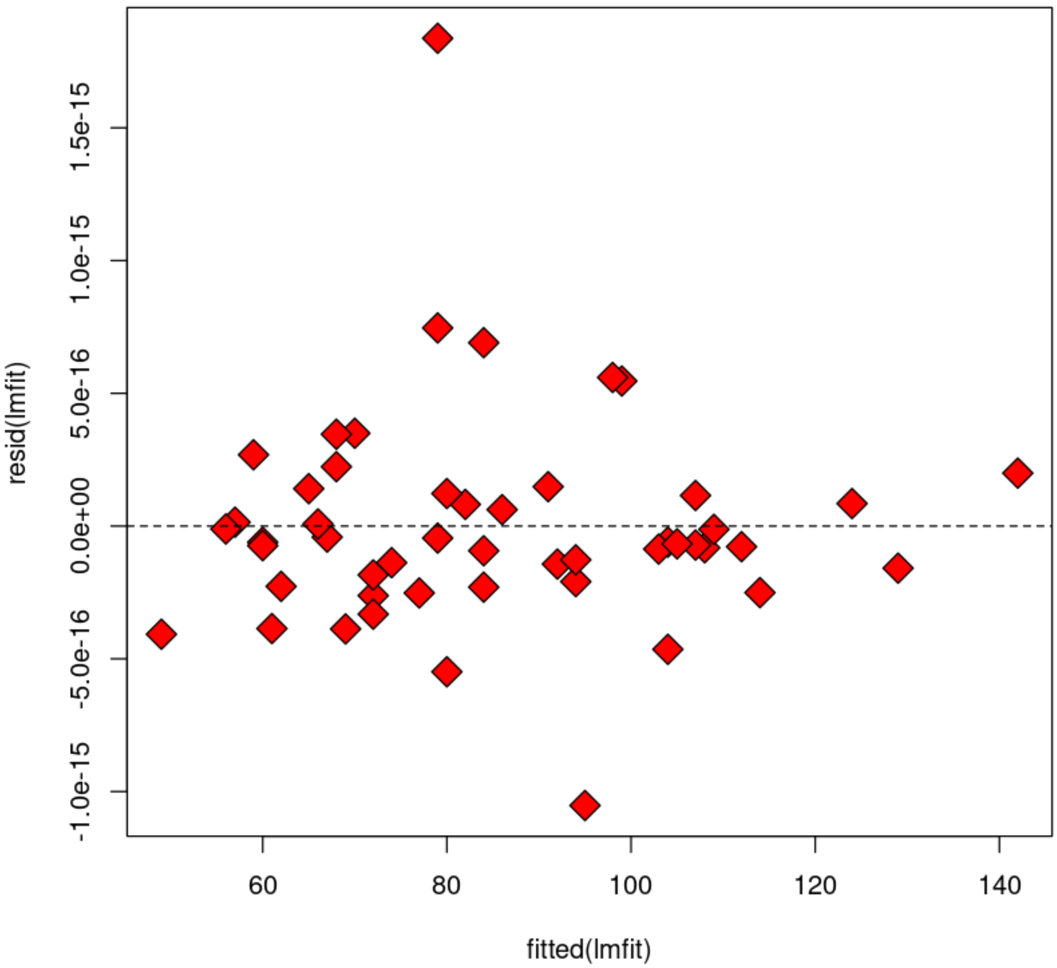
Fit a model to the data. In your model, include X1, X2, X3, year, region indicator variables, and the interaction effects to allow the coefficients ("slopes") for X1, X2, X3 to vary by year.

[56]:
```
par(mfrow=c(2,2))
plot(lmfit, pch=23 ,bg='orange',cex=2)
```



In the plot of the standardized residuals vs $\hat{y}$ below, we see that the model violates the homoscedastic assumption on the errors. In fact, as the value of $\hat{y}$ increase, the variance of the residuals increases. This is reinforced in the Q-Q plot where we see that the residuals have greater dispersion than a normally distributed errors should have.

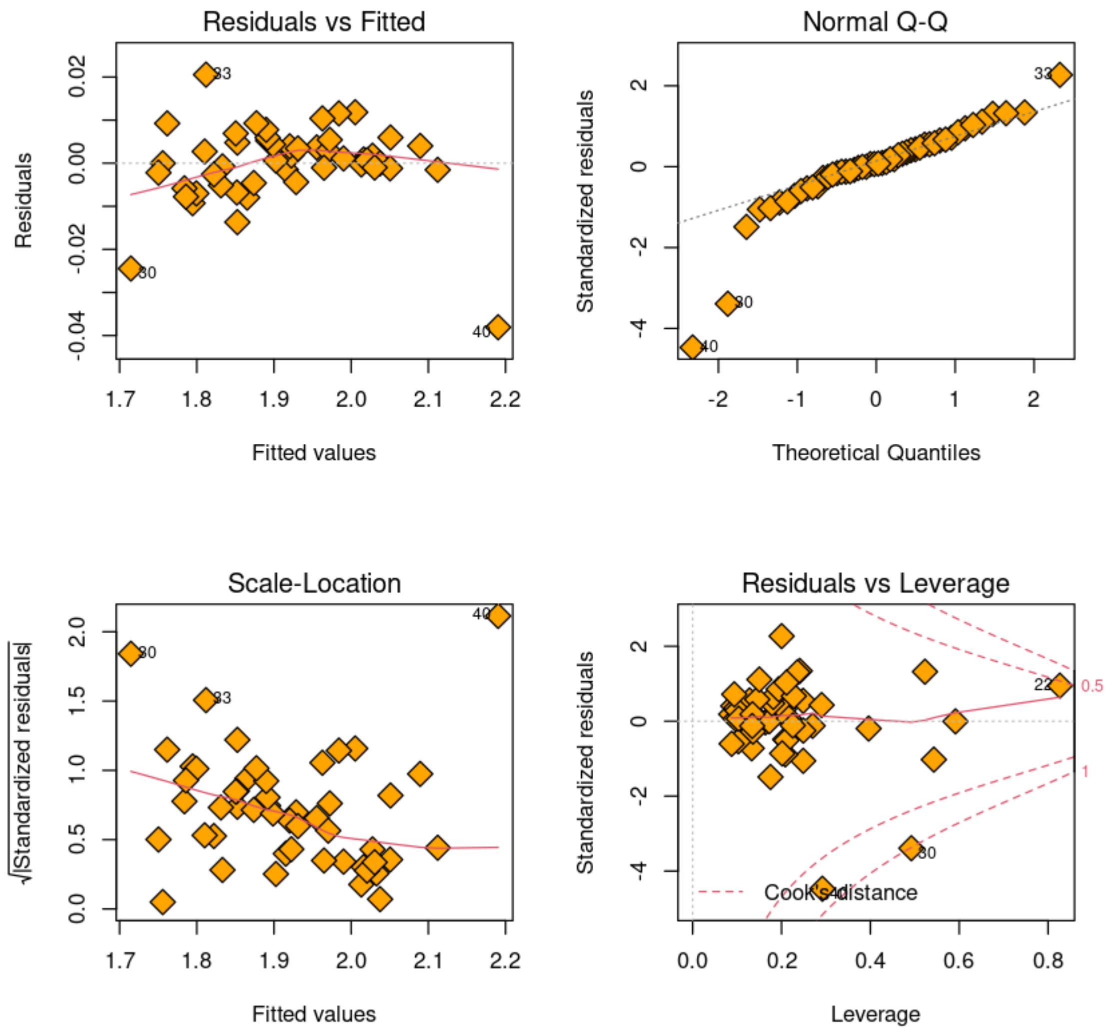
[58]:
```
plot(fitted(lmfit), resid(lmfit), pch=23, bg='red', cex=2)
abline(h=0, lty=2)
```



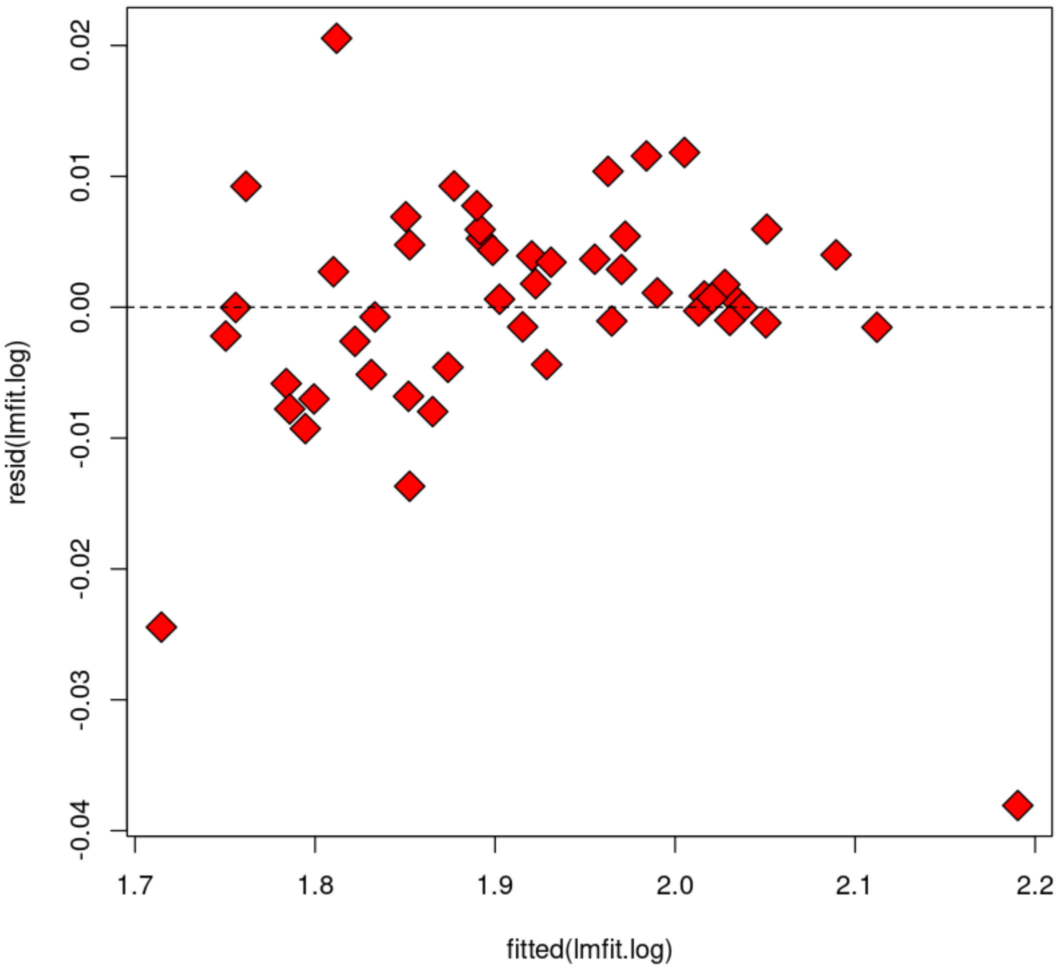
The problems discovered in part above can sometimes be alleviated by changing the response from y to $\log(y)$. Create a new response variable: $\log(y)$. Fit the same model in part above except using the new response variable, $\log(y)$.

[54]:
```
theData <- theData %>%
    mutate(logy = log(y, 10))
lmfit.log <- lm(logy ~ x1 + x2 + x3 + r2 + r3 + r4 + year + x1*year + x2*year +
 ↪x3*year, data = theData)
```

[55]:
```
par(mfrow=c(2,2))
plot(lmfit.log, pch=23 ,bg='orange',cex=2)
```



[57]:
```
plot(fitted(lmfit.log), resid(lmfit.log), pch=23, bg='red', cex=2)
abline(h=0, lty=2)
```



From the figures below, we summarize our diagnostics. - The linearity assumption seems valid. From the standardized residuals vs. $\hat{y}$, we see that there is no obvious pattern to the errors. - From the same plot, we also no longer see any heteroscedasticity. However, we do see one residual that has a large, negative value. This is observation number 30 . - The QQ plot suggests that a normal model for the errors is reasonable. - However, we do see a few points of high leverage and a point of high influence. The observation with the largest leverage and largest influence are observation number 30.

Test the overall effects of X_1, X_2, X_3 on Y. Specify the hypothesis to be tested, the test used and your conclusions at the 5% significance level. Denote the model by

$$\begin{aligned}\log_{10}(y) = &\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 r_2 + \beta_5 r_3 + \beta_6 r_4 + \\ &\beta_7 \text{ year } + \beta_8 x_1 \text{ year } + \beta_9 x_2 \text{ year } + \beta_{10} x_3 \text{ year } + \epsilon\end{aligned}$$

- Hypothesis tests:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_8 = \beta_9 = \beta_{10} = 0$$

$$H_a : \text{at least one of the coefficients is non-zero}$$

- Test used: F-test.

	F Statistic	P-value
Observation 30 kept:	21.526	$< 2.2\text{e} - 16$
Observation 30 removed:	22.804	$< 2.2\text{e} - 16$

- Conclusion: Reject the null hypothesis at $\alpha = 0.05$. That is, the terms X_1, X_2, X_3 should not be removed from the model.

Test whether the effects of X_1, X_2, X_3 remain unchanged over time. Specify the hypothesis to be tested, the test used and your conclusions at the 5% significance level. One appropriate way to perform this test is to add indicator variables for the years. Then, we can consider the following model

$$\begin{aligned}\log_{10}(y) =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 r_2 + \beta_5 r_3 + \beta_6 r_4 + \\ & \gamma_1 z_{1970} + \gamma_2 z_{1975} + \delta_1 x_1 z_{1970} + \delta_2 x_1 z_{1975} + \\ & \delta_3 x_2 z_{1970} + \delta_4 x_2 z_{1975} + \delta_5 x_3 z_{1970} + \delta_6 x_3 z_{1975} + \epsilon\end{aligned}$$

where z_{1970} is 1 if the year is 1970 and 0 otherwise; and z_{1975} is 1 if the year is 1975 and 0 otherwise.

There is some vagueness in the problem. We may test if the overall effects of X_1, X_2, X_3 remain unchanged over time or we may test if X_1 's effect remains unchanged, X_2 's effect remains unchanged or if X_3 's effect remains unchanged over time. If each test is done individually, then the Bonferroni correction should be used. For the solution, we will test if the overall effects of X_1, X_2, X_3 remain unchanged over time.

- Hypothesis tests:

$$H_0 : \delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = \delta_6 = 0$$

$$H_a : \text{at least one of the coefficients is non-zero}$$

- Test used: F-test.

	F Statistic	P-value
Observation 30 kept:	4.641	0.0002549
Observation 30 removed:	4.5213	0.0003321

- Conclusion: Reject the null hypothesis at $\alpha = 0.05$. That is, the effects of X_1, X_2 and X_3 do change over time.

Based on findings in above section decide whether separate regressions by year interval need to be reported. Report coefficients for X variables separately by year (i.e. fit a separate model for each year). We know from section above that there is statistical evidence to include the year coefficients in the model. Therefore, this suggests that we should separate the regressions by year.

- For 1960

	term	estimate	std.error	statistic	p.value
1	(Intercept)	$1.565311\text{e} + 00$	$1.071567\text{e} - 01$	14.6076792	$2.812078\text{e} - 18$
2	x1	$1.858449\text{e} - 04$	$2.840047\text{e} - 05$	6.5437268	$5.912989\text{e} - 08$
3	x2	$-6.868906\text{e} - 05$	$1.768153\text{e} - 04$	-0.3884792	$6.995783\text{e} - 01$
4	x3	$-1.008785\text{e} - 04$	$6.849121\text{e} - 05$	-1.4728677	$1.480694\text{e} - 01$
5	r2	$8.998835\text{e} - 02$	$2.496525\text{e} - 02$	3.6045441	$8.071982\text{e} - 04$
6	r3	$5.045615\text{e} - 02$	$2.768262\text{e} - 02$	1.8226651	$7.531234\text{e} - 02$
7	r4	$1.656219\text{e} - 01$	$2.514287\text{e} - 02$	6.5872322	$5.111370\text{e} - 08$

- For 1970

	term	estimate	std.error	statistic	p.value
1	Intercept)	1.5733348166	1.597552e − 01	9.8484095	1.369590e − 12
2	x1	0.0001543246	2.311512e − 05	6.6763496	3.793105e − 08
3	x2	0.0009415367	4.078470e − 04	2.3085539	2.584097e − 02
4	x3	−0.0002082683	7.592779e − 05	−2.7429778	8.841579e − 03
5	$r2$	0.0045121391	2.596437e − 02	0.1737820	8.628524e − 01
6	$r3$	−0.0187511395	2.871853e − 02	−0.6529284	5.172784e − 01
7	$r4$	0.0613218921	3.010727e − 02	2.0367805	4.785873e − 02

- For 1975

	term	estimate	std.error	statistic	p.value
1	Intercept)	$1.501650e + 00$	1.912740e − 01	7.85077770	7.720540e − 10
2	x1	9.276766e − 05	1.788934e − 05	5.18563809	5.499489e − 06
3	x2	1.571523e − 03	4.895898e − 04	3.20987611	2.512489e − 03
4	x3	1.708977e − 06	7.290820e − 05	0.02344012	9.814076e − 01
5	$r2$	−1.977881e − 02	2.489658e − 02	−0.79443892	4.313029e − 01
6	$r3$	−1.871280e − 02	2.541307e − 02	−0.73634572	4.655187e − 01
7	$r4$	3.032602e − 02	2.698805e − 02	1.12368320	2.673818e − 01

- new response variable model

	term	estimate	std.error	statistic	p.value
1	(Intercept)	1.560530e + 00	9.986319e − 02	15.62668225	2.078498e − 32
2	x1	2.042802e − 04	2.303646e − 05	8.86769053	3.255906e − 15
3	x2	7.361055e − 05	1.744044e − 04	0.42206817	6.736273e − 01
4	x3	−1.656538e − 04	6.596496e − 05	−2.51123953	1.317552e − 02
5	$r2$	2.493975e − 02	1.506686e − 02	1.65527185	1.001255e − 01
6	$r3$	8.085911e − 04	1.603168e − 02	0.05043707	9.598465e − 01
7	$r4$	8.832339e − 02	1.590520e − 02	5.55311238	1.378712e − 07
8	year	7.248962e − 03	1.230775e − 02	0.58897541	5.568337e − 01
9	x1 : year	−7.166606e − 06	1.446879e − 06	−4.95314682	2.086481e − 06
10	x2 : year	6.556682e − 05	2.719311e − 05	2.41115563	1.720834e − 02
11	x3 : year	6.016874e − 06	5.870579e − 06	1.02492002	3.071813e − 01

Compare the estimated coefficients X_1, X_2, X_3 for different models, Show that the coefficient estimates for in new response variable model which change the response from y to log(y) can be used to find the coefficients in above section.

Given the model used in new response variable model which change the response from y to log(y) with an ordinal year variable it would not be possible to recover the coefficients in above section. However, for the model

$$\begin{aligned}\log_{10}(y) =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 r_2 + \beta_5 r_3 + \beta_6 r_4 + \\ & \gamma_1 z_{1970} + \gamma_2 z_{1975} + \delta_1 x_1 z_{1970} + \delta_2 x_1 z_{1975} + \\ & \delta_3 x_2 z_{1970} + \delta_4 x_2 z_{1975} + \delta_5 x_3 z_{1970} + \delta_6 x_3 z_{1975} + \epsilon\end{aligned}$$

with nominal year variables (indicator variables) this would be possible.

5_7

October 16, 2021

[1]:
```python
import os
```

[4]:
```python
from pystata import config
config.init('mp')
```

```
  ___  ____  ____  ____  ____ ©
 /__    /   ____/   /   ____/      17.0
___/   /   /___/   /   /___/       MP-Parallel Edition

 Statistics and Data Science       Copyright 1985-2021 StataCorp LLC
                                   StataCorp
                                   4905 Lakeway Drive
                                   College Station, Texas 77845 USA
                                   800-STATA-PC        https://www.stata.com
                                   979-696-4600        stata@stata.com

Stata license: Unlimited-user 4-core network, expiring 13 Mar 2022
Serial number: 501709301094
  Licensed to: fbsh


Notes:
      1. Unicode is supported; see help unicode_advice.
      2. More than 2 billion observations are allowed; see help obs_advice.
      3. Maximum number of variables is set to 5,000; see help set_maxvar.
```

[45]:
```
%%stata
use https://stats.idre.ucla.edu/stat/stata/examples/chp/p148, clear
    xi i.fertiliz
    gen F1 = 0
    gen F2 = 0
    gen F3 = 0
    replace F1 = 1 if fertiliz == 1
    replace F2 = 1 if fertiliz == 2
    replace F3 = 1 if fertiliz == 3
    list
    regress yield F1 F2 F3
```

```
test F1 F2 F3
test F1=F2=F3
gen Fs = F1 + F2 + F3
regress yield Fs
```

```
. use https://stats.idre.ucla.edu/stat/stata/examples/chp/p148, clear

.     xi i.fertiliz
i.fertiliz        _Ifertiliz_1-4      (naturally coded; _Ifertiliz_1 omitted)

.     gen F1 = 0

.     gen F2 = 0

.     gen F3 = 0

.     replace F1 = 1 if fertiliz == 1
(10 real changes made)

.     replace F2 = 1 if fertiliz == 2
(10 real changes made)

.     replace F3 = 1 if fertiliz == 3
(10 real changes made)

.     list

     +-------------------------------------------------------------------+
     | yield   fertiliz   _Ifert~2   _Ifert~3   _Ifert~4   F1   F2   F3 |
     |-------------------------------------------------------------------|
  1. |    31          1          0          0          0    1    0    0 |
  2. |    34          1          0          0          0    1    0    0 |
  3. |    34          1          0          0          0    1    0    0 |
  4. |    34          1          0          0          0    1    0    0 |
  5. |    43          1          0          0          0    1    0    0 |
     |-------------------------------------------------------------------|
  6. |    35          1          0          0          0    1    0    0 |
  7. |    38          1          0          0          0    1    0    0 |
  8. |    36          1          0          0          0    1    0    0 |
  9. |    36          1          0          0          0    1    0    0 |
 10. |    45          1          0          0          0    1    0    0 |
     |-------------------------------------------------------------------|
 11. |    27          2          1          0          0    0    1    0 |
 12. |    27          2          1          0          0    0    1    0 |
 13. |    25          2          1          0          0    0    1    0 |
 14. |    34          2          1          0          0    0    1    0 |
```

```
15. |    21          2          1          0          0    0    1    0 |
    |------------------------------------------------------------------|
16. |    36          2          1          0          0    0    1    0 |
17. |    34          2          1          0          0    0    1    0 |
18. |    30          2          1          0          0    0    1    0 |
19. |    32          2          1          0          0    0    1    0 |
20. |    33          2          1          0          0    0    1    0 |
    |------------------------------------------------------------------|
21. |    36          3          0          1          0    0    0    1 |
22. |    37          3          0          1          0    0    0    1 |
23. |    37          3          0          1          0    0    0    1 |
24. |    34          3          0          1          0    0    0    1 |
25. |    37          3          0          1          0    0    0    1 |
    |------------------------------------------------------------------|
26. |    28          3          0          1          0    0    0    1 |
27. |    33          3          0          1          0    0    0    1 |
28. |    29          3          0          1          0    0    0    1 |
29. |    36          3          0          1          0    0    0    1 |
30. |    42          3          0          1          0    0    0    1 |
    |------------------------------------------------------------------|
31. |    33          4          0          0          1    0    0    0 |
32. |    27          4          0          0          1    0    0    0 |
33. |    35          4          0          0          1    0    0    0 |
34. |    25          4          0          0          1    0    0    0 |
35. |    29          4          0          0          1    0    0    0 |
    |------------------------------------------------------------------|
36. |    20          4          0          0          1    0    0    0 |
37. |    25          4          0          0          1    0    0    0 |
38. |    40          4          0          0          1    0    0    0 |
39. |    35          4          0          0          1    0    0    0 |
40. |    29          4          0          0          1    0    0    0 |
    +------------------------------------------------------------------+

.     regress yield F1 F2 F3

      Source |       SS           df       MS      Number of obs   =        40
-------------+----------------------------------   F(3, 36)        =      5.14
       Model |      362.6         3  120.866667   Prob > F        =    0.0046
    Residual |      845.8        36  23.4944444   R-squared       =    0.3001
-------------+----------------------------------   Adj R-squared   =    0.2417
       Total |     1208.4        39  30.9846154   Root MSE        =    4.8471

------------------------------------------------------------------------------
       yield | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          F1 |        6.8   2.167692     3.14   0.003     2.403717    11.19628
          F2 |         .1   2.167692     0.05   0.963    -4.296283    4.496283
          F3 |        5.1   2.167692     2.35   0.024     .7037167    9.496283
```

```
       _cons |       29.8    1.53279    19.44   0.000     26.69136    32.90864
------------------------------------------------------------------------------

.     test F1 F2 F3

 ( 1)  F1 = 0
 ( 2)  F2 = 0
 ( 3)  F3 = 0

       F(  3,    36) =    5.14
            Prob > F =    0.0046

.     test F1=F2=F3

 ( 1)  F1 - F2 = 0
 ( 2)  F1 - F3 = 0

       F(  2,    36) =    5.16
            Prob > F =    0.0107

.     gen Fs = F1 + F2 + F3

.     regress yield Fs

      Source |       SS           df       MS      Number of obs   =        40
-------------+----------------------------------   F(1, 38)        =      4.19
       Model |         120         1         120   Prob > F        =    0.0476
    Residual |      1088.4        38  28.6421053   R-squared       =    0.0993
-------------+----------------------------------   Adj R-squared   =    0.0756
       Total |      1208.4        39  30.9846154   Root MSE        =    5.3518

------------------------------------------------------------------------------
       yield | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          Fs |          4   1.954213     2.05   0.048     .0439032    7.956097
       _cons |       29.8   1.692398    17.61   0.000     26.37392    33.22608
------------------------------------------------------------------------------

.
```

1. Shown above
2. Shown above. Fit the model $y_{ij} = \mu_0 + \mu_1 F_{i1} + \mu_2 F_{i2} + \mu_3 F_{i3} + \epsilon_{ij}$
3. Test the hypothesis that none of the three types of fertilizer has an effect on corn crops. Specify the hypothesis to be tested, the test used, and the conclusions at the 5% significance level. Our null and alternative hypotheses are: $H_0 : \mu_1 = \mu_2 = \mu_3 = 0$ versus $H_1 : \mu_j \neq 0$ for any $j = 1, 2, 3$. F test shown above. From the output, we can see that the p-value for the F test is $0.0046 < 0.05$. Therefore, we reject the null and conclude that there is plenty

of evidence to reject the null that the fertilizers have, on average, no effect. At least one fertilizer differs from the control group.

4. Test the hypothesis that the three types of fertilizer have equal effects on corn crop. Specify the hypothesis to be tested, the test used and the conclusions at the 5% significance level. To test for equal fertilizer effect, our null and alternative hypotheses are: $H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_1 : \mu_j \neq \mu_k$ for any $j, k = 1, 2, 3$ and $j \neq k$. The p-value for this test is 0.0107. So, there is strong evidence to suggest that there are some differences between different fertilizers in terms of the mean yield.

5. To test whether a common estimate of the fertilizer effect, call it μ_F, is actually different from zero. The null and alternative are $H_0 : \mu_F = 0$ versus $H_1 : \mu_F \neq 0$. From the output, the p-value for the t-test of $\mu_F = 0$ is 0.048 and we compare it to the significance level of 0.05 so the common fertilizer effect is significantly different from that of the control.

4.3

October 16, 2021

1. Linear relation of each predictor with the response; perhaps the existence of outliers or influential points.
2. Independence among predictors.
3. Normality assumption of the residuals.
4. Linearity; constant variance; and uncorrelation of residuals.
5. Each observation has approximately equal influence.

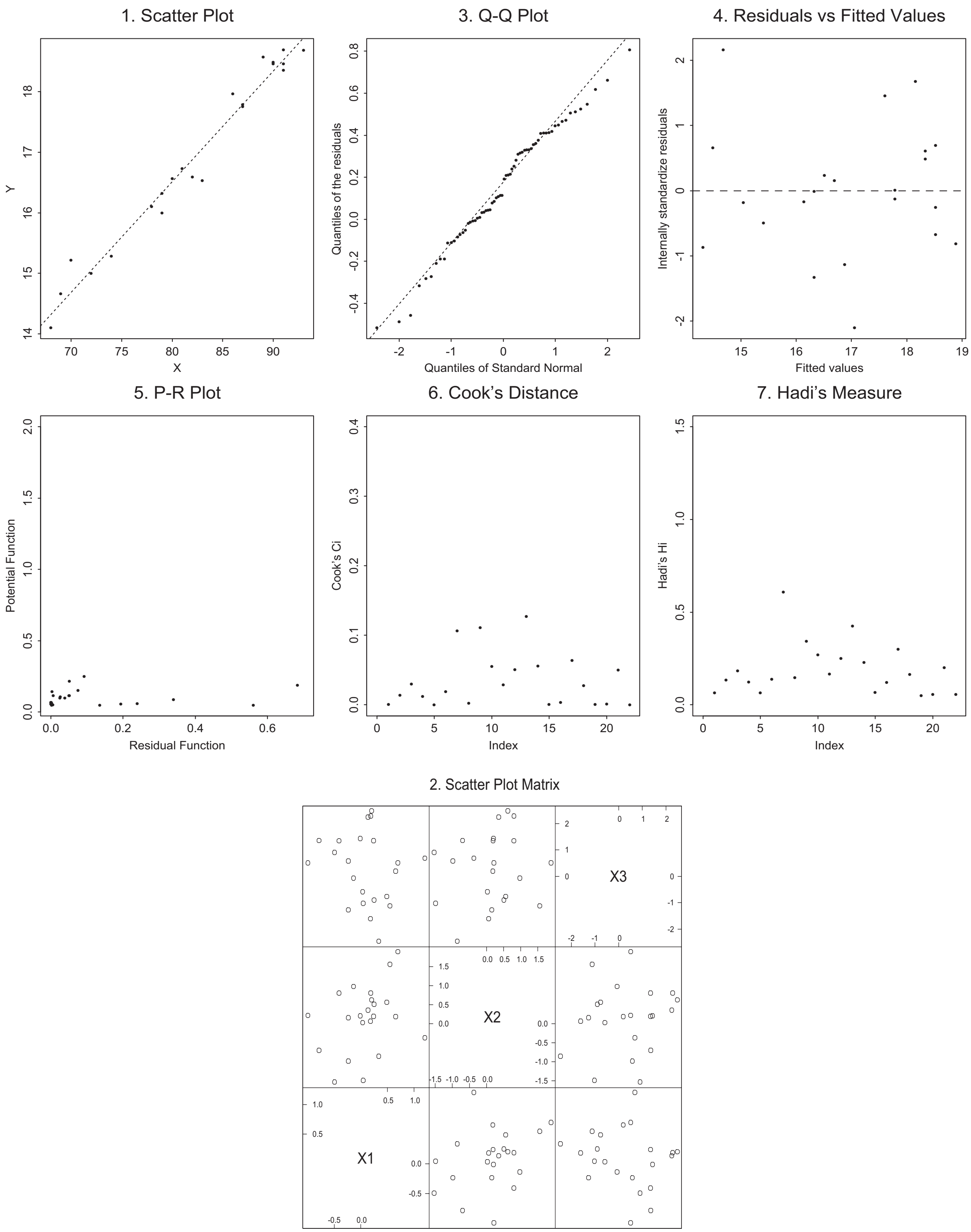


Figure 1: Graphs with assumptions satisfied.

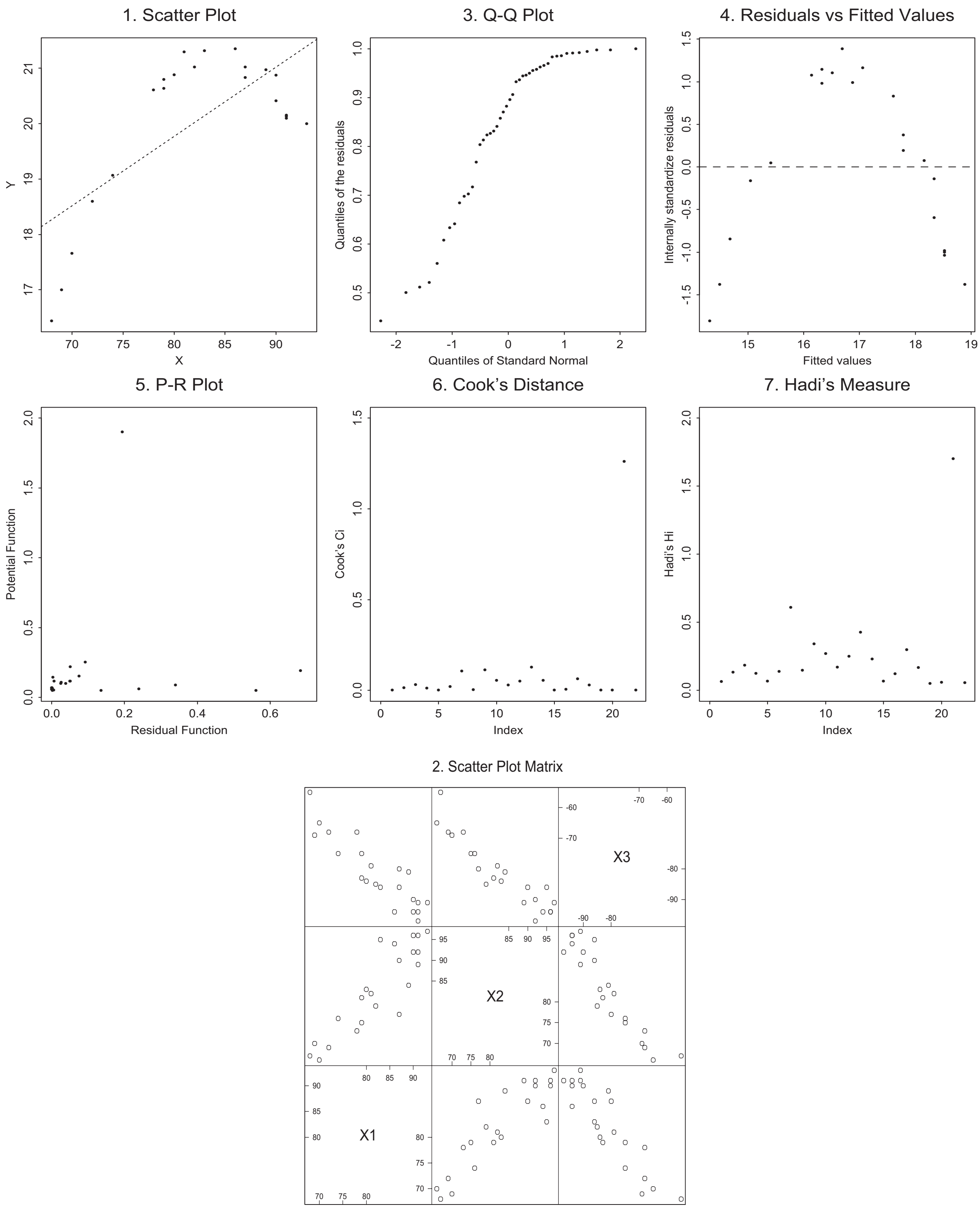


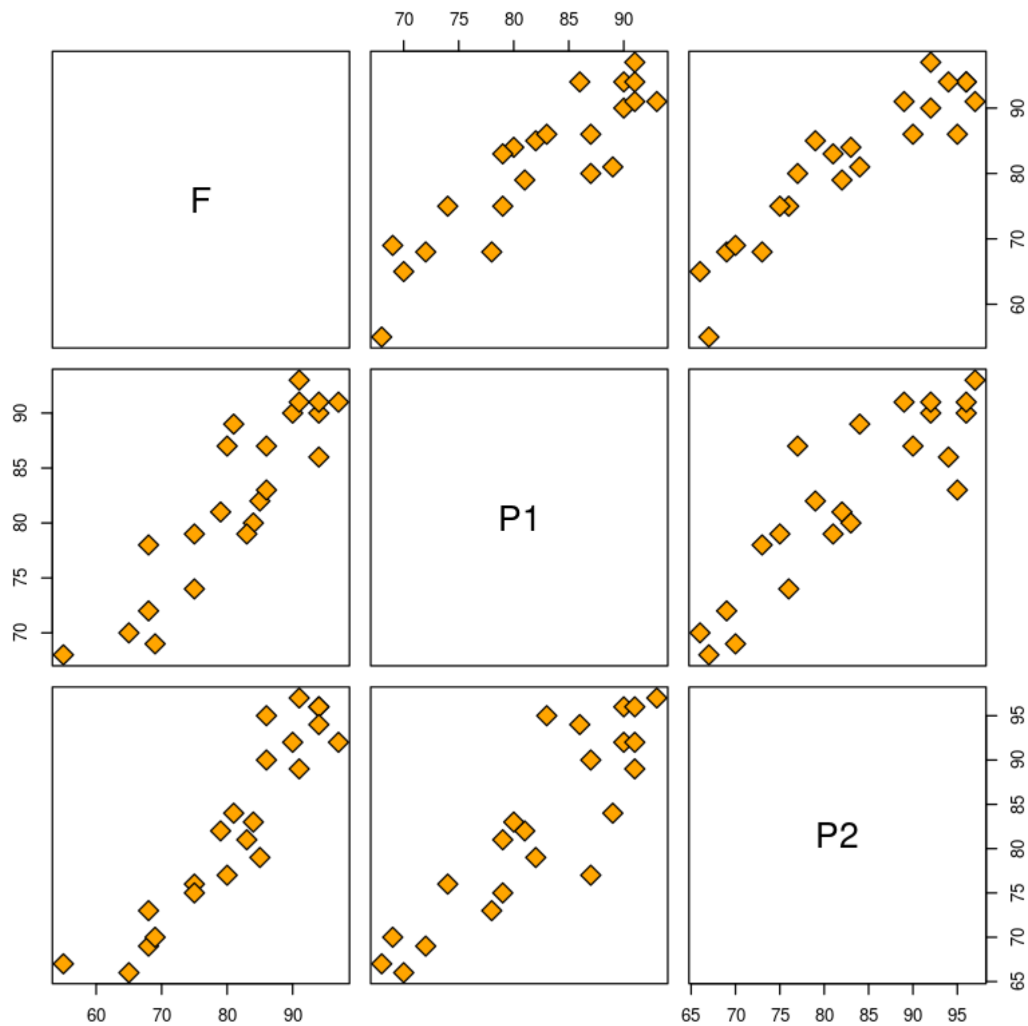
Figure 2: Graphs when assumptions are violated.

4_8

October 16, 2021

[50]:

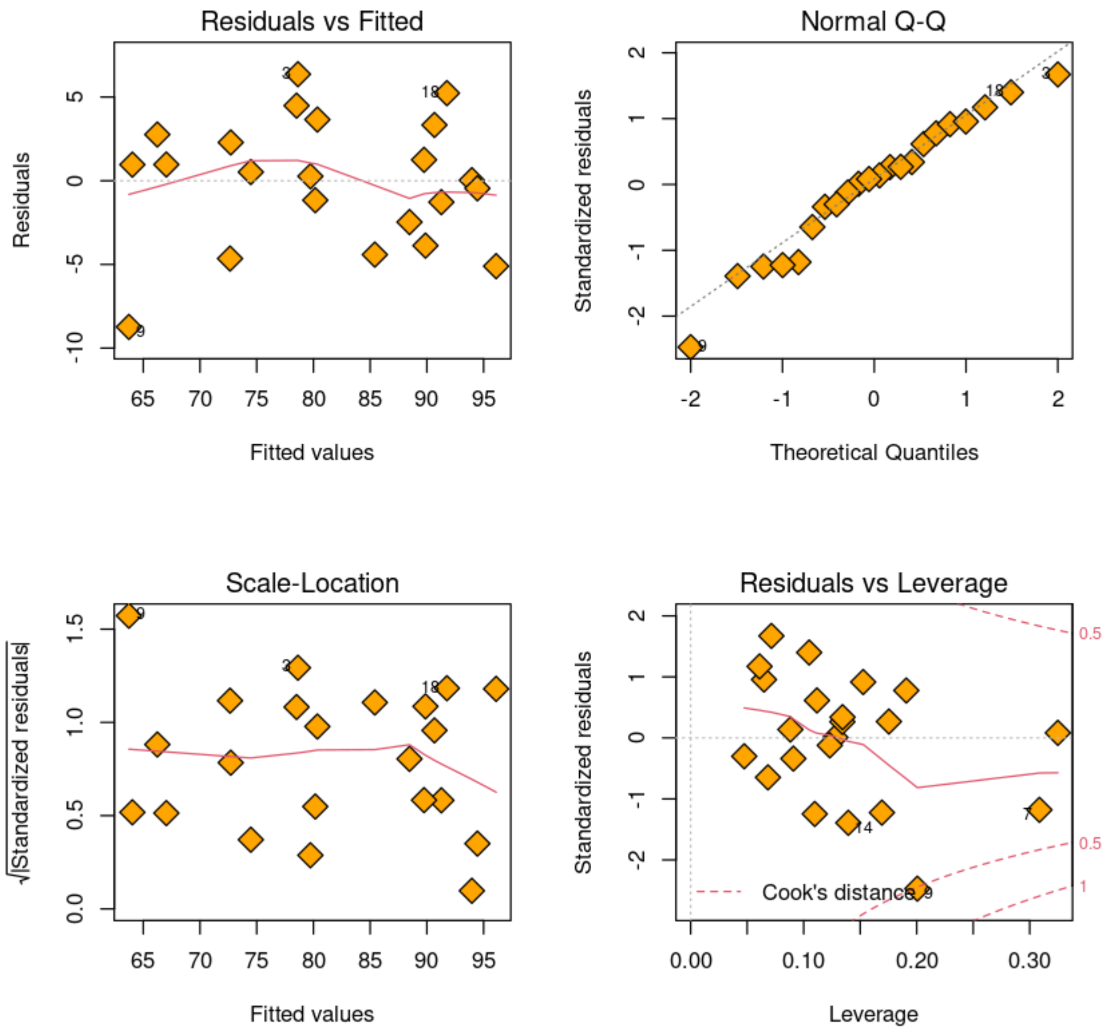
```
finals.table = read.table("Table3.10.txt", header=TRUE, sep='\t')
plot(finals.table, pch=23, bg='orange', cex=2)
```



[63]:
```
finals.lm = lm(F ~ P1 + P2, data=finals.table)
finals.lm2 = lm(F ~ P1, data=finals.table)
finals.lm3 = lm(F ~ P2, data=finals.table)
summary(finals.lm)
```

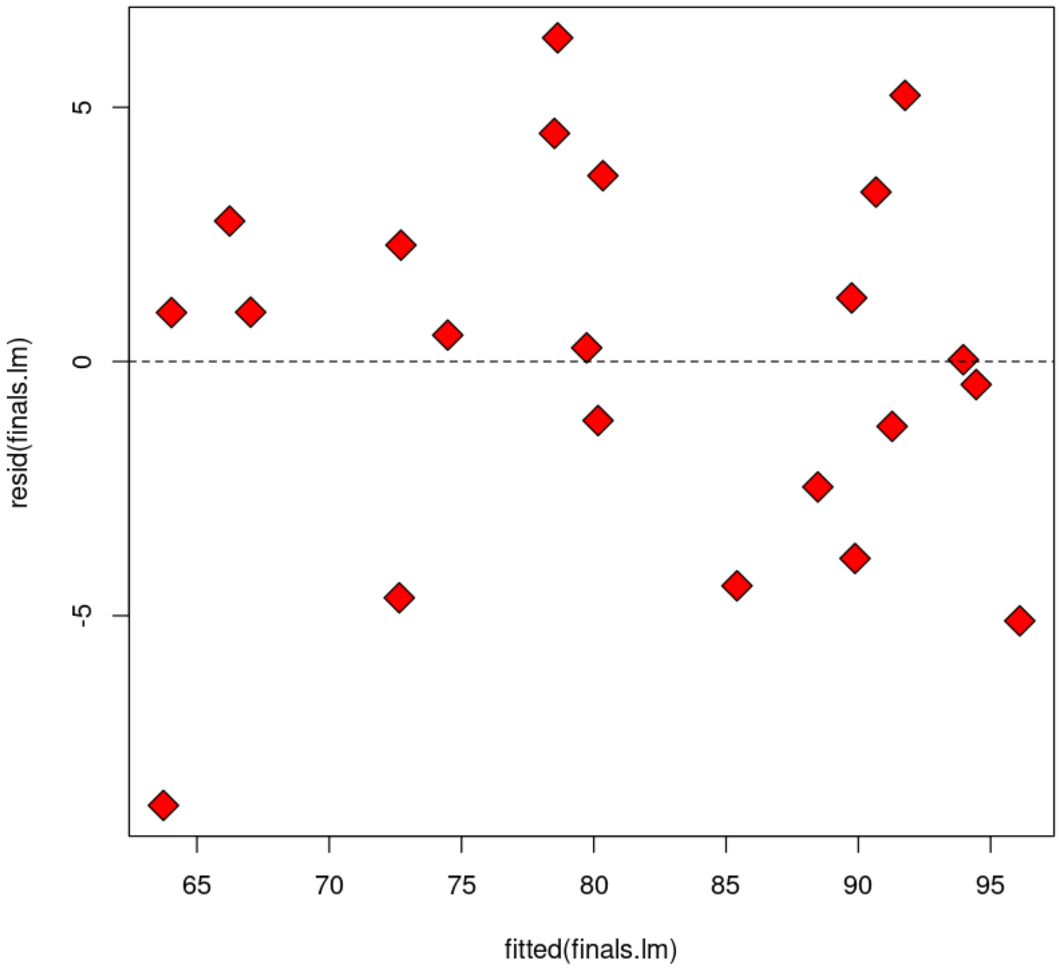
```
Call:
lm(formula = F ~ P1 + P2, data = finals.table)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7328 -2.1703  0.3938  2.6443  6.3660

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.5005     9.2356  -1.570  0.13290
P1            0.4883     0.2330   2.096  0.04971 *
P2            0.6720     0.1793   3.748  0.00136 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.953 on 19 degrees of freedom
Multiple R-squared:  0.8863,	Adjusted R-squared:  0.8744
F-statistic: 74.07 on 2 and 19 DF,  p-value: 1.069e-09
```

[52]:
```
par(mfrow=c(2,2))
plot(finals.lm, pch=23 ,bg='orange',cex=2)
```



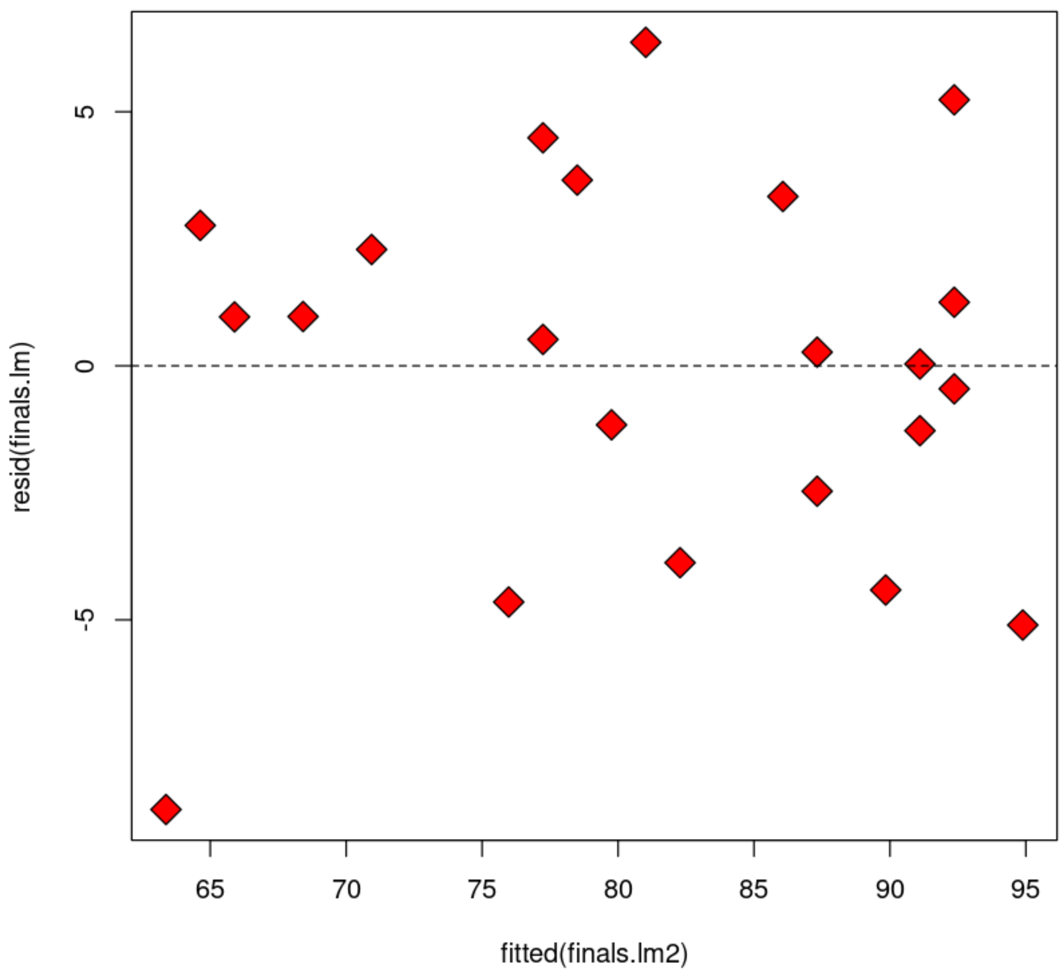
0.1 Model 1

[58]:
```
plot(fitted(finals.lm), resid(finals.lm), pch=23, bg='red', cex=2)
abline(h=0, lty=2)
```



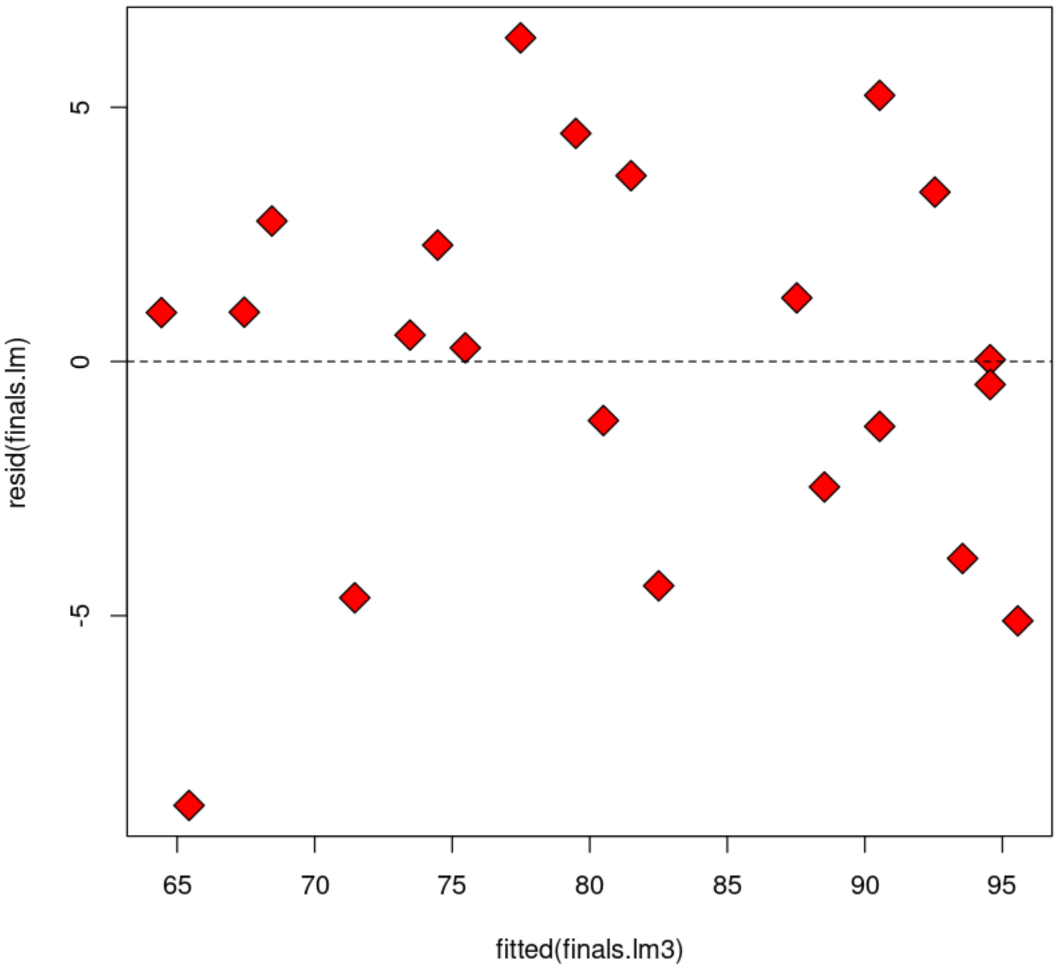
0.2 Model 2

[64]:
```
plot(fitted(finals.lm2), resid(finals.lm), pch=23, bg='red', cex=2)
abline(h=0, lty=2)
```



0.3 Model 3

[65]:
```
plot(fitted(finals.lm3), resid(finals.lm), pch=23, bg='red', cex=2)
abline(h=0, lty=2)
```



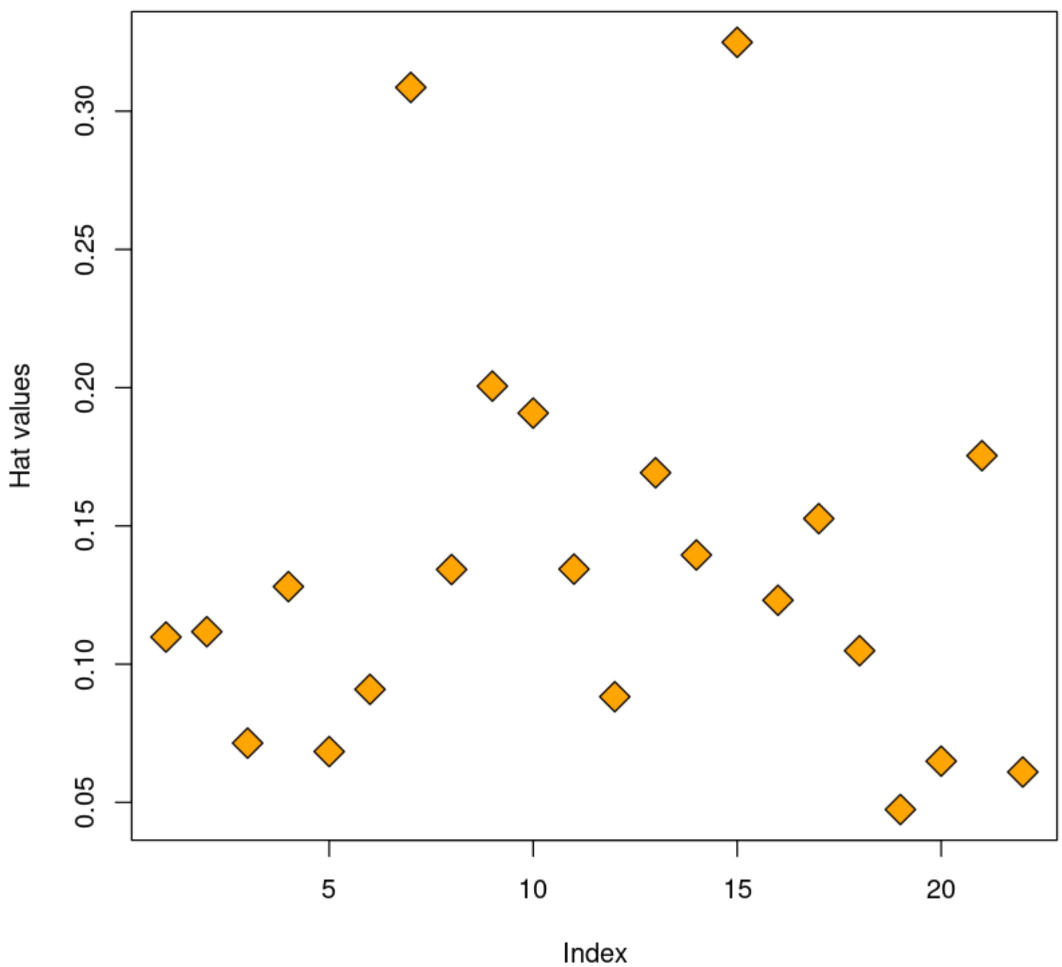
1 Outlying X values

leverage $_i = H_{ii} = \left(X\left(X^TX\right)^{-1}X^T\right)_{ii}$ ## Model 1

[59]:
```
plot ( hatvalues ( finals.lm ), pch = 23 , bg = 'orange' , cex = 2 , ylab =
→'Hat values' )
finals.table [ which ( hatvalues ( finals.lm )  >  0.3 ),]
```

A data.frame: 2 × 3

	F	P1	P2
	<int>	<int>	<int>
7	86	83	95
15	80	87	77



[60]:
```
X = rnorm(100)
Y = 2 * X + 0.5 + rnorm(100)
alpha = 0.1
cutoff = qt(1 - alpha / 2, 97)
sum(abs(rstudent(lm(Y~X))) > cutoff)
```

9

[61]:
```
# Bonferroni correction
X = rnorm(100)
Y = 2 * X + 0.5 + rnorm(100)
cutoff = qt(1 - (alpha / 100) / 2, 97)
sum(abs(rstudent(lm(Y~X))) > cutoff)
```

0

[62]:
```
library(car)
outlierTest(finals.lm)
```

```
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferroni p
9 -2.919642          0.0091487      0.20127
```

1.1 Model 2

[66]:
```
library(car)
outlierTest(finals.lm2)
```

```
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferroni p
9 -1.970607           0.063516           NA
```

1.2 Model 3

[67]:
```
library(car)
outlierTest(finals.lm3)
```

```
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferroni p
9 -3.225882          0.0044483     0.097862
```

[ ]:
```

```