Math 3080 § 1. **Grasshopper Example: Influence Measures** Name: Example
Treibergs April 4, 2016

This note discusses influence measures: how to detect if there are points that have undue influence over the estimated model. Such points may be outliers, those with huge residuals, or they may be located away from the other data points, those with huge leverage.

Data is taken from Walpole, Myers, Myers, Ye, *Probability and Statistics for Engineers and Scientists,* 7th ed., Prentice Hall, 2002. A study in the VPI Department of Etymology made experimental runs for two methods for capturing grasshoppers, drop net and sweep net. The data gives average number of grasshoppers caught in a set of field quadrants on a given date. The goal is to estimate the number of grasshopper caught using only the sweep net method y, which is less costly in terms of the drop net catch x_1 and the height of the vegetation x_2. Is the fourth data point valid?

We fit a linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where ϵ is a vector of IID $N(0, \sigma^2)$ normal variables. If $\hat{\beta}_j$ denotes the least squares estimated coefficient, then the fitted point is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i}$$

and the residual is

$$e_i = y_i - \hat{y}_i$$

The residual is not quite an estimate for the error because it satisfies $\sum e_i = 0$ and $\sum x_i e_i = 0$. But for the purposes of diagnostics, any residual two or more standard deviations away from zero is an outlier.

The way to find outliers is to look at standardized residuals.

$$r_i = \frac{e_i}{s_{e_i}} = \frac{e_i}{s\sqrt{1 - h_{ii}}}$$

Here the estimator for σ^2 is

$$s^2 = MSE = \frac{1}{n-k-2}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

and h_{ii} is the diagonal of the hat matrix H, so called because $\hat{y} = X\hat{\beta} = Hy$. It is

$$H = X(X^T X)^{-1} X^T$$

where X is the design matrix whose first column is ones and the remaining columns are the vectors $x_1, \ldots, x_k$. h_{ii} is the leverage. Its presence as the standard error of e_i says that points far from $\bar{x}$ have large influence and the fitted value has smaller variability. Indeed, if there is one regressor $k = 1$ then

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}$$

The hat values always satisfy $1/n \le h_{ii} \le 1$ and $\sum_{i=1}^{n} h_{ii} = k + 1$. As a result, any data point whose hat diagonal value is large, *i.e.,*well above $(k+1)/2$ is in a position in the data set where the variance of $\hat{y}_i$ is relatively large and the variance of the residual is relatively small.

To detect undue influence, *jackknife* statistics are used. "Jackknife" means to *leave-out-one*, that an estimate be computed by omitting the ith observation. One example is the *Studentized residual*

$$t_i = \frac{e_i}{s_{(-i)}\sqrt{1 - h_{ii}}}$$

where $s_{(-i)}$ is the estimate of the standard error of ith fit excluding the ith point and estimating s from the other data points. Here the estimator for σ^2 is

$$s^2_{(-i)} = MSE = \frac{1}{n-k-1}\sum_{\substack{j=1\\ j\neq i}}^{n}(y_j - \hat{y}_{j(-i)})^2$$

Studentized residuals are used in the same way as standardized residuals. t_i follows a t-distribution with $n-k-2$ degrees of freedom assuming the model is correct. The extreme residuals tend to be farther out for studentized residuals than standardized residuals. The studentized residuals should not be used for simultaneously testing if there are any outliers at aqll locations. Rather, this statistic highlights data points where the error of fit is larger than by chance.

The "difference of fits," or `dffits` for short is the standardized difference in fit between including the ith point or not.

$$\texttt{dffits}_i = \frac{\hat{y}_i - \hat{y}_{i(-i)}}{s_{(-i)}/\sqrt{h_{ii}}}$$

where $s_{(-i)}$ is the estimator of σ^2 computed without the ith point and $\hat{y}_{i(-i)}$ is the estimated fit computed without the ith point. The point is excessively influential if $|\texttt{difffits}_i| > 2k/n$.

A second jackknife statistic is the change in the $\hat{\beta}_i$'s if the ith point is omitted.

$$\texttt{dfbeta}_j = \hat{\beta}_j - \hat{\beta}_{j(-i)}$$

The Cook's Distance tells the same information as the studentized residual.

$$D_i = \frac{e_i^2}{s^2 k}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] = \frac{r_i^2}{k}\left[\frac{h_{ii}}{1-h_{ii}}\right]$$

This statistic is the product of the square of the standardized residual and the leverage factor. **R**© will plot D_i versus i or draw the D level lines in the leverage-standardized residual plane. the Cool's Distance is considered large if $D_i > 1$.

The influence statistics flag extreme data points, that have a big influence on the fit. Such data points should be checked with whatever resources possible. Regression may be run without the extreme points for comparison. However, deleting points from regression data should not be done indiscriminately.

In conclusion. the studentized residual for the fourth data point is 7.08 which is huge.

Data Set Used in this Analysis :

```
# M3080 - 1       Grasshopper data       Apr.4, 2016
#
# From Walpole, Myers, Myers, Ye, "Probability and Statistics for
# Engineers and Scientists," 7th ed., Prentice Hall, 2002
# From a study in VPI Department of Etymology. Experimental runs
# were made for two methods for capturing grasshoppers, drop net and
# sweep net. The data gives average number of grasshoppers caught in
# a set of field quadrants on a given date.  The goal is to estimate
# the number of grasshopper catch using only the sweep net method, which is
# less costly. Is the fourth data point valid?
# y = drop net catch
# x1 = sweep net catch
# x2 = plant height (cm)
"y" "x1" "x2"
18     4.15476  52.705
8.875  2.02381  42.069
2       .15909  34.766
20     2.32812  27.622
2.375   .25521  45.879
2.75    .57292  97.472
3.3333  .70139 102.062
1       .13542  97.79
1.3333  .12121  88.265
1.75    .10937  58.737
4.125   .5625   42.386
12.875 2.45312  31.274
5.375   .045312 31.75
28     6.6875   35.401
4.75    .86979  64.516
1.75    .14583  25.241
.1333   .01562  36.354
```

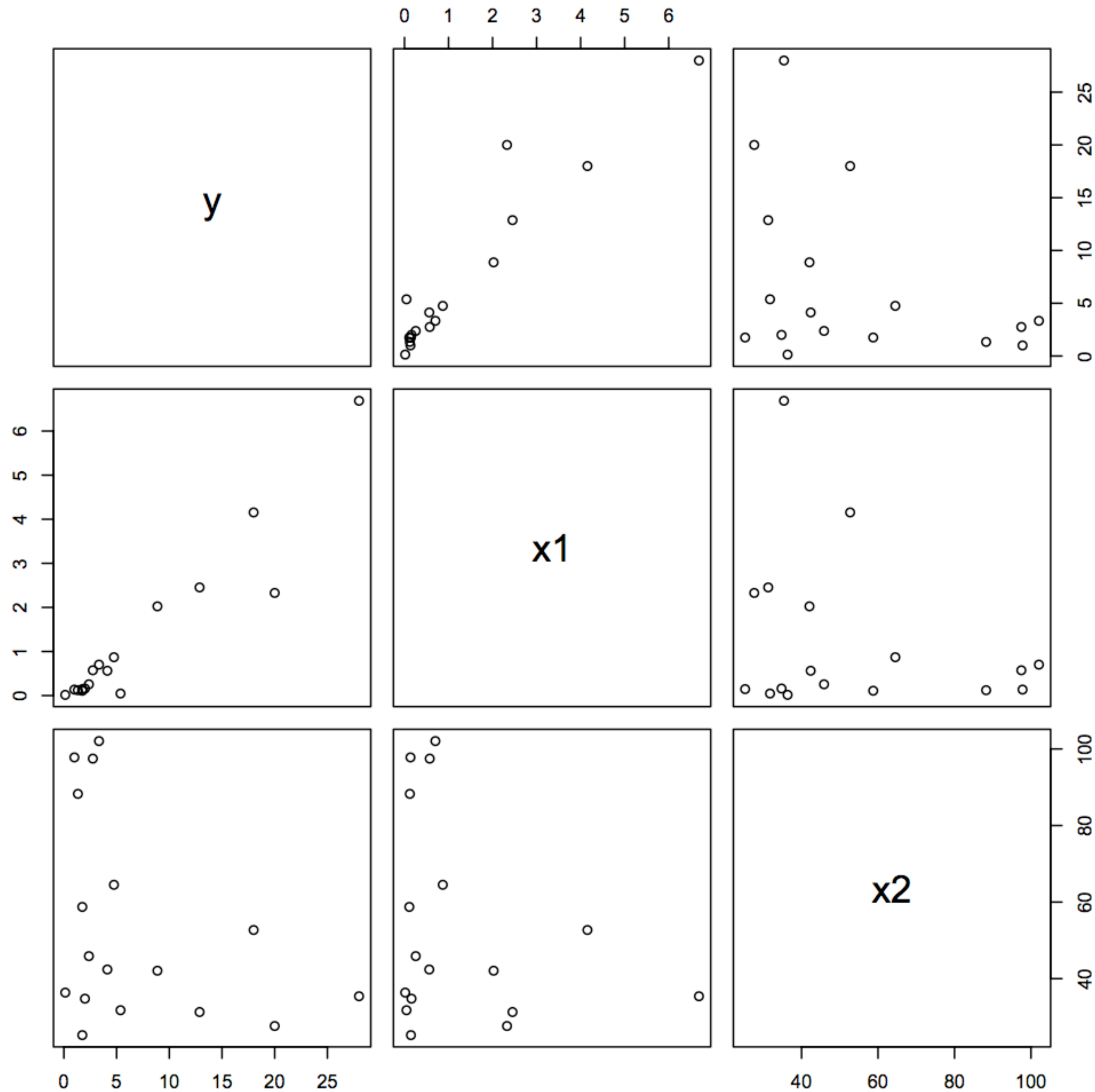
R Session:

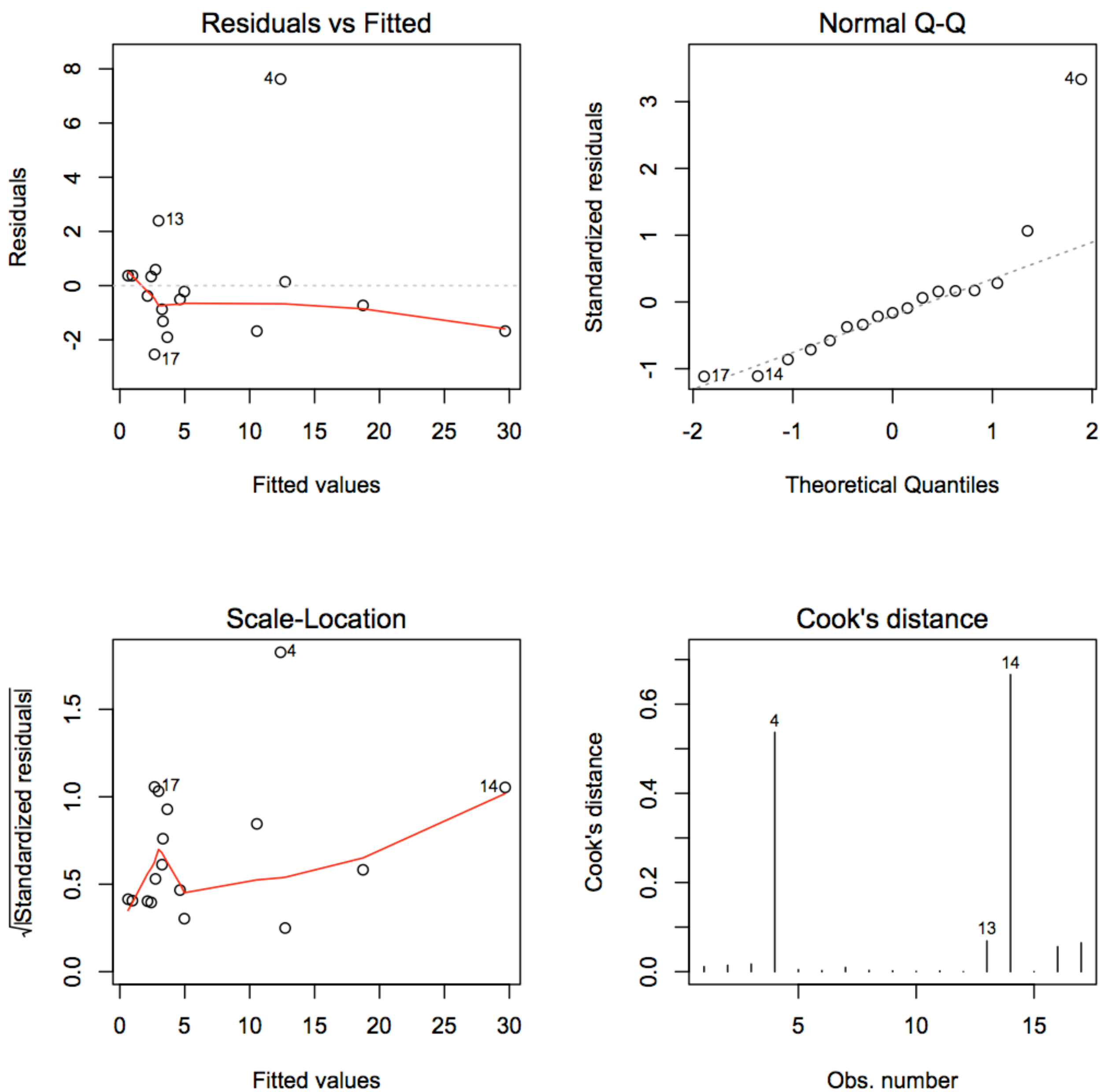
```
R version 2.10.1 (2009-12-14)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.31 (5538) powerpc-apple-darwin8.11.1]

[Workspace restored from /Users/andrejstreibergs/.RData]

> tt=read.table("M3083GrasshopperData.txt", header=T)
> attach(tt)
> tt
         y       x1      x2
1  18.0000 4.154760  52.705
2   8.8750 2.023810  42.069
3   2.0000 0.159090  34.766
4  20.0000 2.328120  27.622
5   2.3750 0.255210  45.879
6   2.7500 0.572920  97.472
7   3.3333 0.701390 102.062
8   1.0000 0.135420  97.790
9   1.3333 0.121210  88.265
10  1.7500 0.109370  58.737
11  4.1250 0.562500  42.386
12 12.8750 2.453120  31.274
13  5.3750 0.045312  31.750
14 28.0000 6.687500  35.401
15  4.7500 0.869790  64.516
16  1.7500 0.145830  25.241
17  0.1333 0.015620  36.354

> ###################  PAIRS PLOT OF ALL VARIABLES  ###################
> pairs(tt)
```

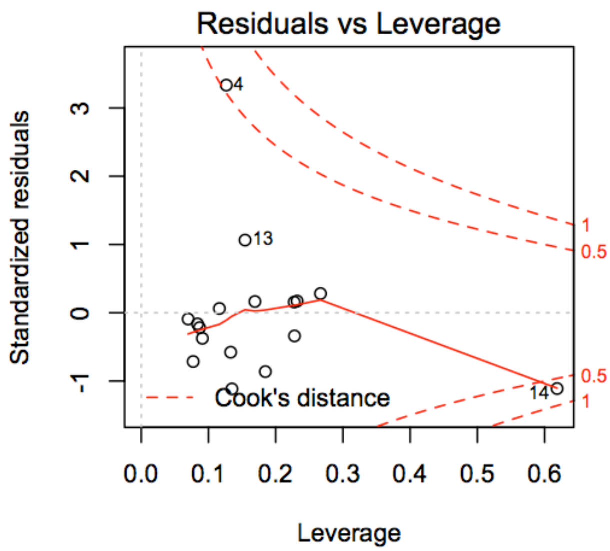
```
> ##################  RUN LINEAR MODEL  ##############################
> l1=lm(y~x1+x2); anova(l1); summary(l1)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
x1         1 931.72  931.72 155.7034 5.656e-09 ***
x2         1  17.77   17.77   2.9698    0.1068
Residuals 14  83.78    5.98
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5379 -1.3164 -0.3808  0.3676  7.6233

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.10141    1.57663   2.601   0.0209 *
x1           4.04184    0.34923  11.574 1.49e-08 ***
x2          -0.04108    0.02384  -1.723   0.1068
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.446 on 14 degrees of freedom
Multiple R-squared: 0.9189,Adjusted R-squared: 0.9073
F-statistic: 79.34 on 2 and 14 DF,  p-value: 2.303e-08

> ############# PLOT USUAL SIX DIAGNOSTICS  ########################
> opar = par(mfrow=c(2,2))
> plot(l1, which=1:4)
> plot(l1, which=5:6)
> par(opar)
>
```

```
> ###############  RESIDUALS  ######################################
> M=matrix(c(rstandard(l1),rstudent(l1)),ncol=2)
> colnames(M)=c("Standard.Resid.","Student.Resid.")
> M
      Standard.Resid. Student.Resid.
 [1,]     -0.33932862    -0.32833825
 [2,]     -0.71420925    -0.70112056
 [3,]     -0.57793768    -0.56367940
 [4,]      3.33447774     7.08285597
 [5,]     -0.37447254    -0.36267167
 [6,]      0.15663416     0.15106888
 [7,]      0.28133251     0.27186837
 [8,]      0.17170749     0.16563598
 [9,]      0.16485604     0.15901367
[10,]     -0.16260561    -0.15683886
[11,]     -0.21768360    -0.21012106
[12,]      0.06222734     0.05997206
[13,]      1.06455885     1.07005849
[14,]     -1.11012779    -1.12018497
[15,]     -0.09191635    -0.08859955
[16,]     -0.86198642    -0.85359096
[17,]     -1.11548727    -1.12611916

> #############  INFLUENCE DIAGNOSTICS  ############################
> influence.measures(l1)
Influence measures of
 lm(formula = y ~ x1 + x2) :

     dfb.1_   dfb.x1    dfb.x2   dffit  cov.r   cook.d    hat inf
1   0.03880 -0.15364 -0.037122 -0.1784 1.5782 0.011333 0.2280
2  -0.10095 -0.05797  0.062053 -0.2028 1.2109 0.014233 0.0772
3  -0.20152  0.12451  0.137311 -0.2208 1.3400 0.017085 0.1330
4   1.83173  0.64810 -1.623791  2.6958 0.0125 0.536912 0.1265   *
5  -0.08751  0.06209  0.043702 -0.1147 1.3331 0.004678 0.0910
6  -0.04094  0.00267  0.068731  0.0820 1.6084 0.002409 0.2275
7  -0.09108  0.01452  0.142689  0.1639 1.6744 0.009593 0.2666   *
8  -0.03970 -0.00871  0.073024  0.0910 1.6158 0.002967 0.2319
9  -0.02211 -0.01269  0.051119  0.0717 1.4944 0.001843 0.1691
10 -0.02117  0.02471 -0.000799 -0.0474 1.3554 0.000805 0.0837
11 -0.05246  0.02823  0.030186 -0.0648 1.3538 0.001501 0.0868
12  0.01284  0.00720 -0.011097  0.0218 1.4121 0.000170 0.1163
13  0.42702 -0.26731 -0.303153  0.4573 1.1466 0.068993 0.1544
14  0.24097 -1.32048 -0.048380 -1.4266 2.4839 0.666313 0.6186   *
15 -0.00236  0.00251 -0.008266 -0.0243 1.3402 0.000211 0.0697
16 -0.39336  0.21949  0.302619 -0.4062 1.3006 0.056082 0.1846
17 -0.39866  0.26885  0.262465 -0.4448 1.0920 0.064697 0.1349
```

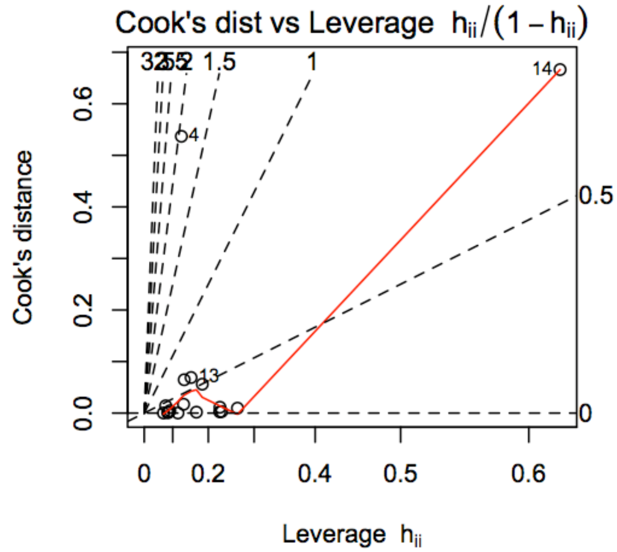
```
> #############  PLOT DFFITS, DFBETAS AND COOK'S DISTNCE  #################
> opar = par(mfrow=c(2,2))
> plot(rstandard(l1))
> plot(rstudent(l1))
> plot(dffits(l1),type="l")
> matplot(dfbetas(l1), type="l")
> lines(sqrt(cooks.distance(l1)),col=4)
> legend(6,1.8,c("beta0","beta1","beta2","Cook's"),fill=c(1,2,3,4))
> par(opar)

>
> ###  FOURTH OBSERVATION LOOKS INFLUENTIAL. RUN REGRESSION WITHOUT IT ####
>

> l1=lm(y~x1+x2, subset = -4); anova(l1); summary(l1)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
x1         1 830.30  830.30 626.0498 2.214e-12 ***
x2         1   5.23    5.23   3.9413   0.06863 .
Residuals 13  17.24    1.33
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Call:
lm(formula = y ~ x1 + x2, subset = -4)

Residuals:
     Min       1Q   Median       3Q      Max
-1.83910 -0.38524 -0.02918  0.13345  3.18053

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.74182    0.76667   3.576  0.00338 **
x1           3.93528    0.16510  23.836 4.11e-12 ***
x2          -0.02286    0.01151  -1.985  0.06863 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.152 on 13 degrees of freedom
Multiple R-squared: 0.9798,Adjusted R-squared: 0.9767
F-statistic:   315 on 2 and 13 DF,  p-value: 9.712e-12
```


y
x1
x2
0 1 2 3 4 5 6
0 5 10 15 20 25
40 60 80 100

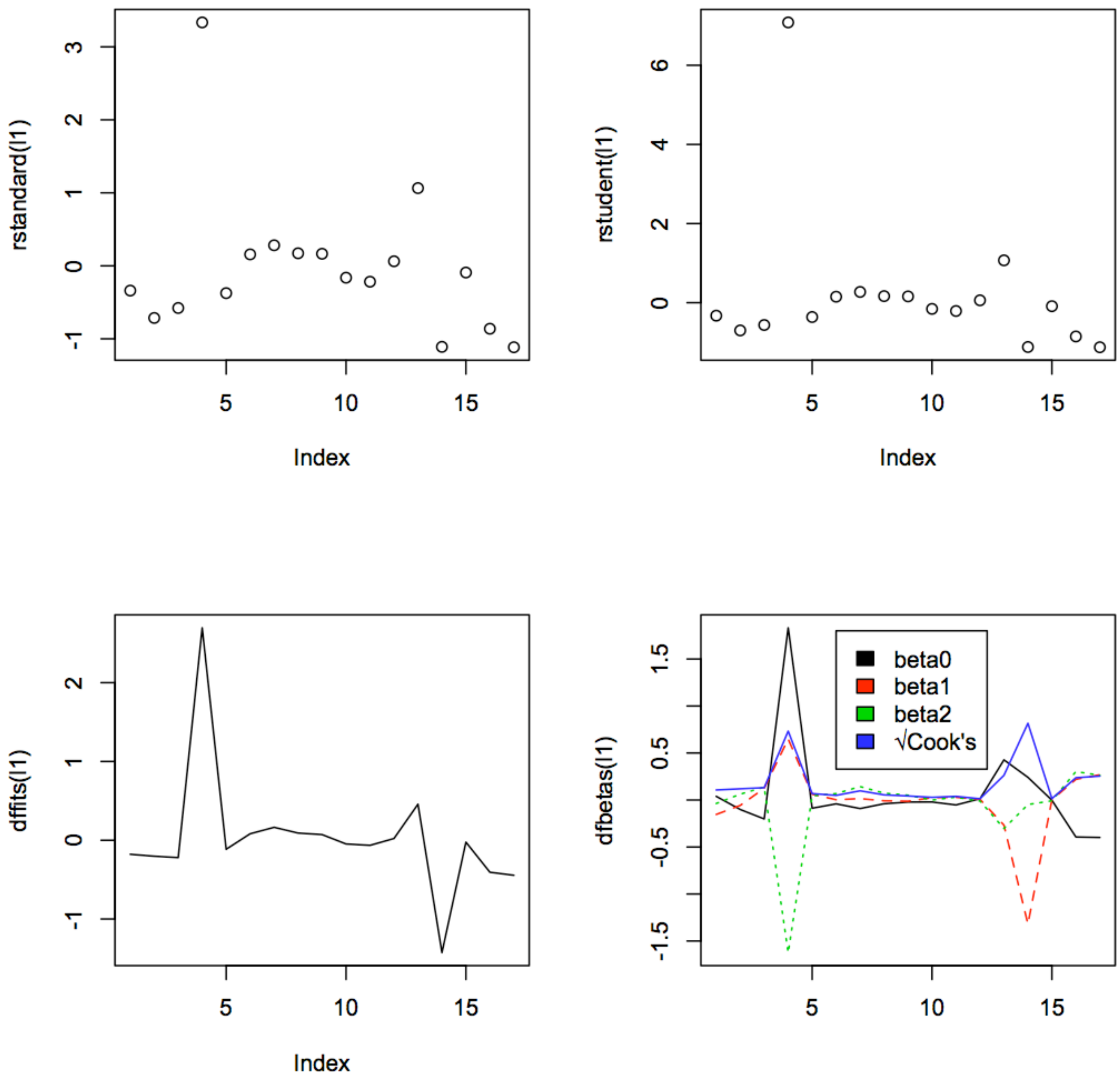

Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Cook's distance
Cook's distance
Obs. number



Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
4
13
14
1
0.5
0.5
1



Cook's dist vs Leverage $h_{ii}/(1-h_{ii})$
Cook's distance
Leverage h_{ii}
1.5
1
0.5
0
4
13
14



rstandard(l1)
Index
rstudent(l1)
Index
dffits(l1)
Index
dfbetas(l1)
beta0
beta1
beta2
√Cook's