36-617: Applied Linear Models

Nonlinearity, Marginal Model Plots & Transformations

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Reading, HW, etc.

- Reading
  - This week: Sheather Ch 6
  - For next week: Begin Reading Ch 7
- HW 04 due tonight
- Quiz on Ch 6 today – you need a laptop!
- HW05 out later today; due next Monday as usu.
  - Information about Project 01 out soon.

Outline

- Graphical methods for detecting non-linearity
  - Casewise diagnostic plots
  - Added variable plots
    - Control for other predictors
    - Intended to illustrate linear regression coefficients
    - But can also reveal non-lineariy
  - Marginal model plots
    - Average over other predictors
    - New for today
- Example(s)
- Transformations – a brief summary

Some graphical techniques for detecting nonlinearity

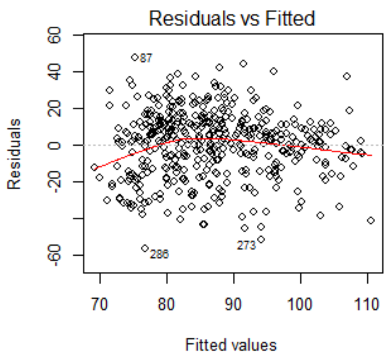
- Casewise diagnostic plots
  - Patterns in raw residual plot, and/or scale-location plot
  - Nonconstant variance and apparent non-normality can also be due to nonlinearity
- Added variable plot
  - Although the plot is intended to check for linear predictors it can also display nonlinearity
- Marginal model plot (we will focus on today)

Quick review…

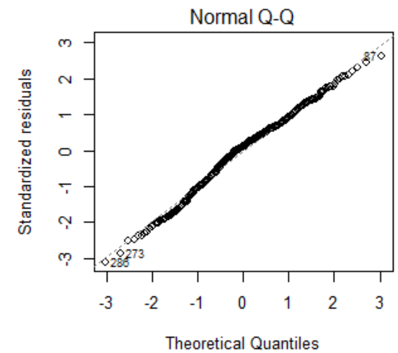
- Casewise diagnostic plots
- Added variable plots

- What do they look like when things go wrong?

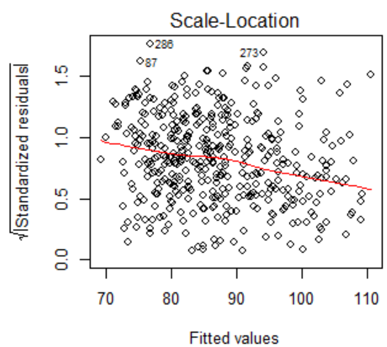
Casewise Diagnostics and Patterns



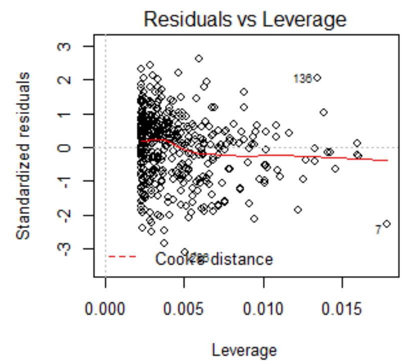
- Mean zero?
- Vertical patterns?
- Outliers?
- Functional dependence on $\hat{y}_i$?



- Normal?
- Outliers?
- Large enough sample?



- Constant variance?
- Vertical patterns?
- Outliers?
- Functional dependence on $\hat{y}_i$?

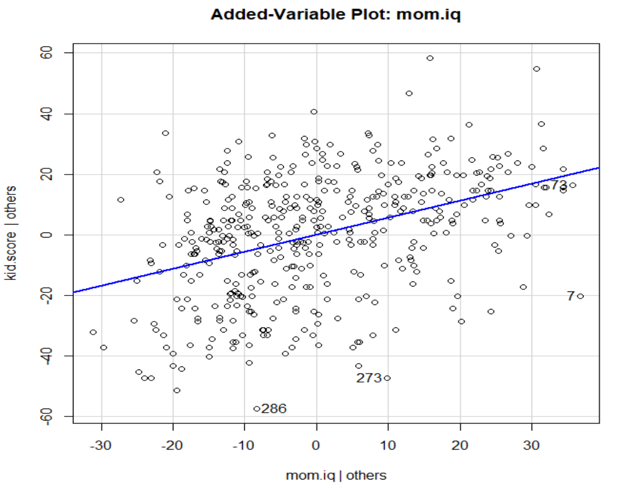


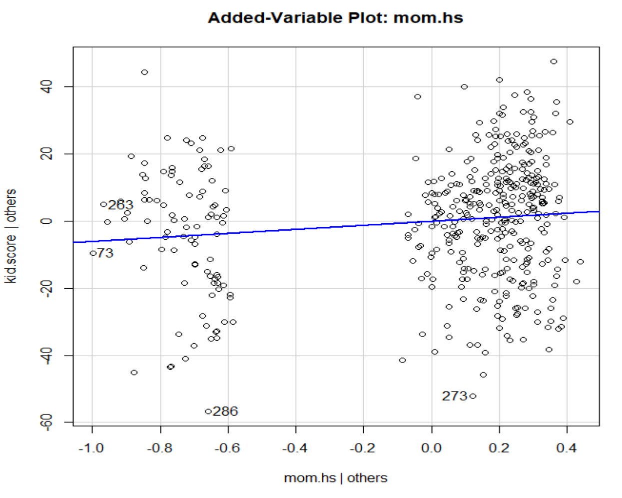
- NE & SE corners:
  - High leverage h_{ii}
  - High std resid r_i
- D_i > 0.5 or so?

- Generally these are conversation points
  - Could reveal things investigator cares about!
  - Otherwise, look for data collection/recording errors
- Delete data only with a good justification!

Added-Variable Plots – Example…

```
> library(car)
> lm.3

Call:
lm(formula = kid.score ~ mom.iq +
mom.hs, data = kidiq)

Coefficients:
(Intercept)    mom.iq    mom.hs
    25.7315    0.5639    5.9501

> avPlot(lm.3,"mom.iq")
> avPlot(lm.3,"mom.hs")
```





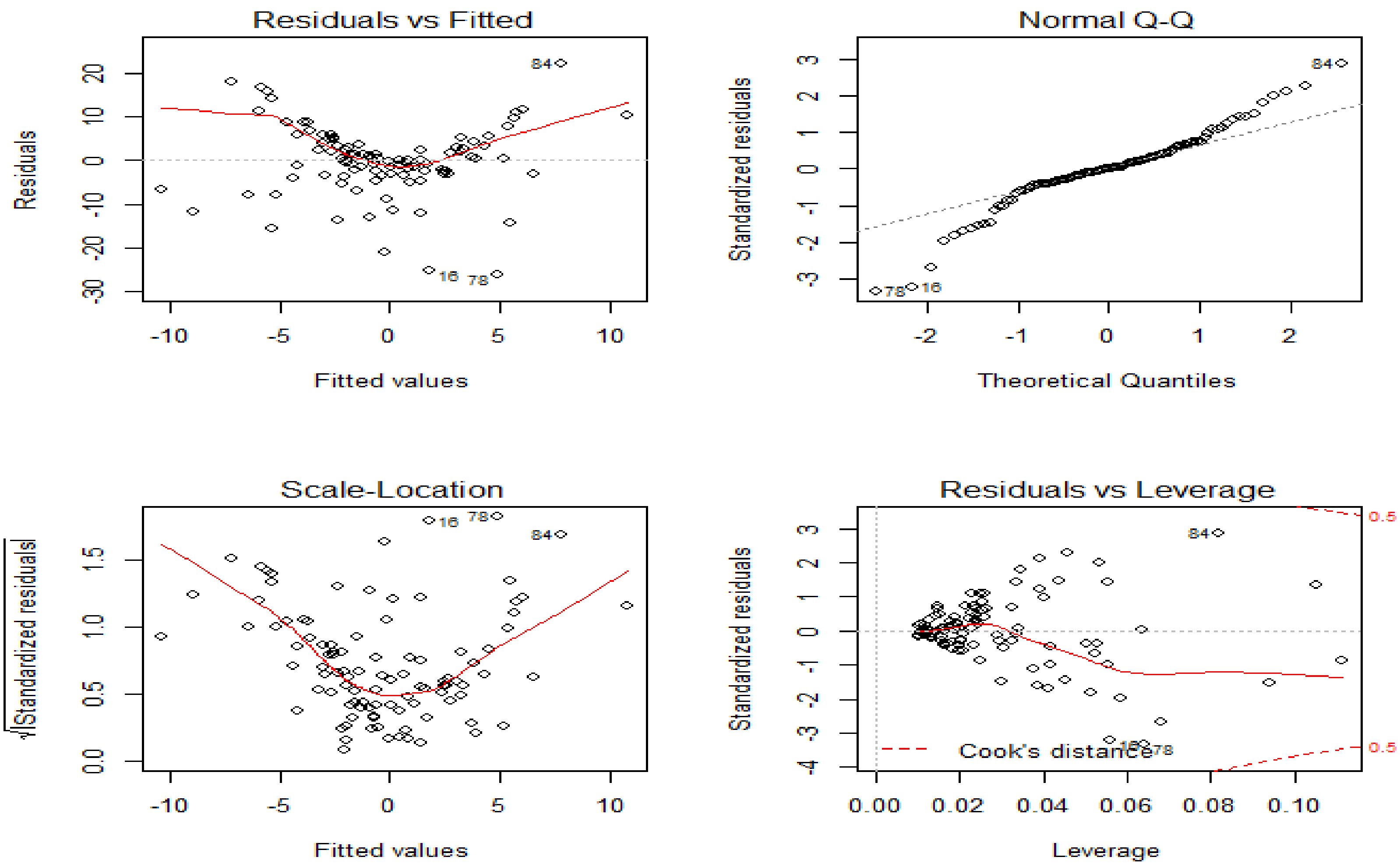
An example

```
> library(car)
> x1 <- rnorm(100)
> x2 <- rnorm(100)
> y <- 1 + x1 + 2*x2 +
+ 10*x1*x2 + rnorm(100)
>
> lm.x1px2 <- lm(y ~ x1 + x2)
> lm.x1mx2 <- lm(y ~ x1 * x2)
>
> summary(lm.x1px2)

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

```
Coefficients:
          Est   SE      t p
(Int) -0.05 0.82 -0.06 0.95
x1     1.77 0.87  2.03 0.04 *
x2     3.44 0.82  4.20 0.00 ***
---

Residual standard error: 8.13 on
97 degrees of freedom
Multiple R-squared:  0.1722,
Adjusted R-squared:  0.1551
F-statistic: 10.09 on 2 and 97 DF,
p-value: 0.0001045
```

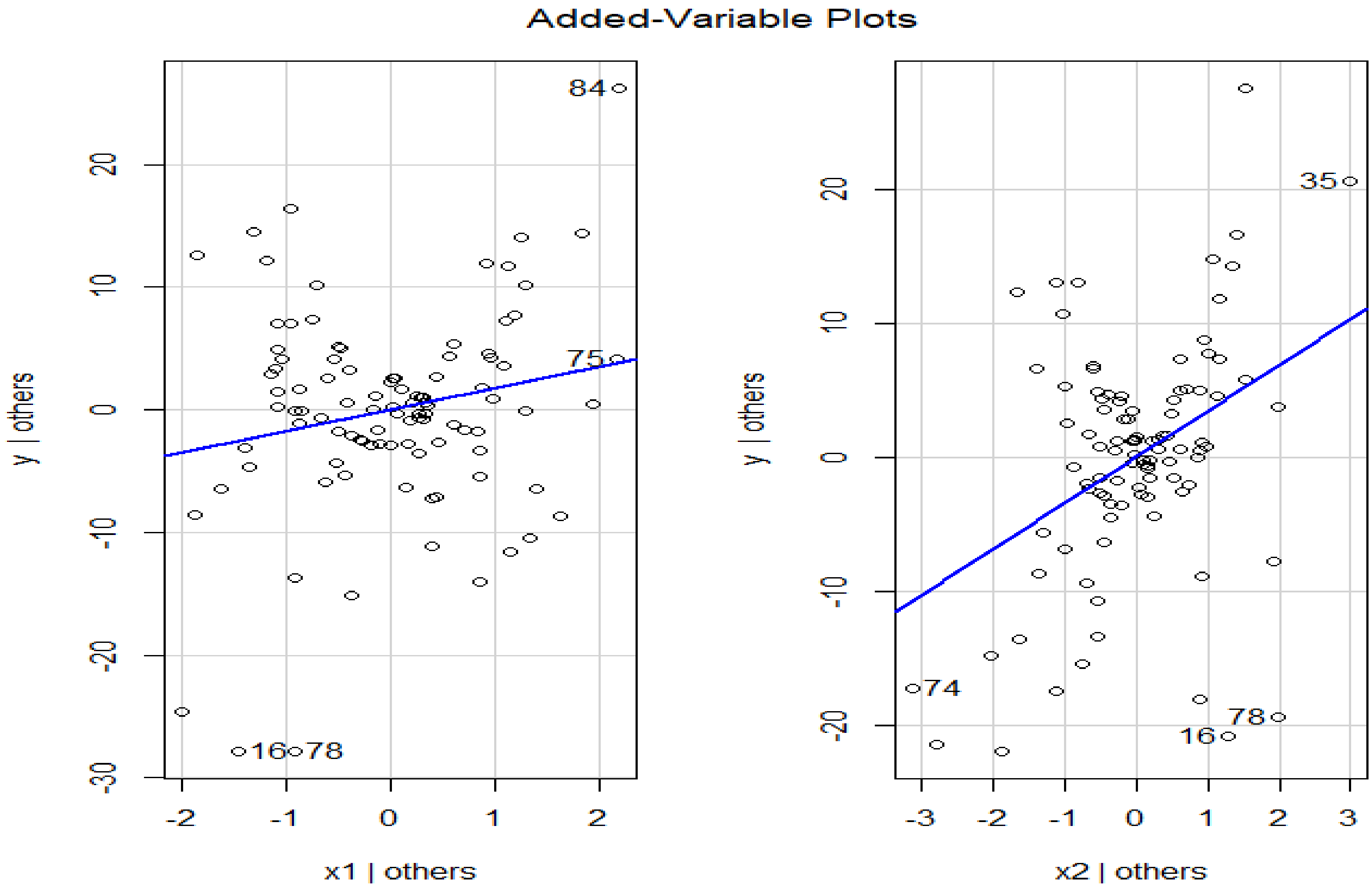
Casewise Diagnostic Plots

> par(mfrow=c(2,2))
> plot(lm.x1px2)



Added Variable Plots

> avPlots(lm.x1px2)

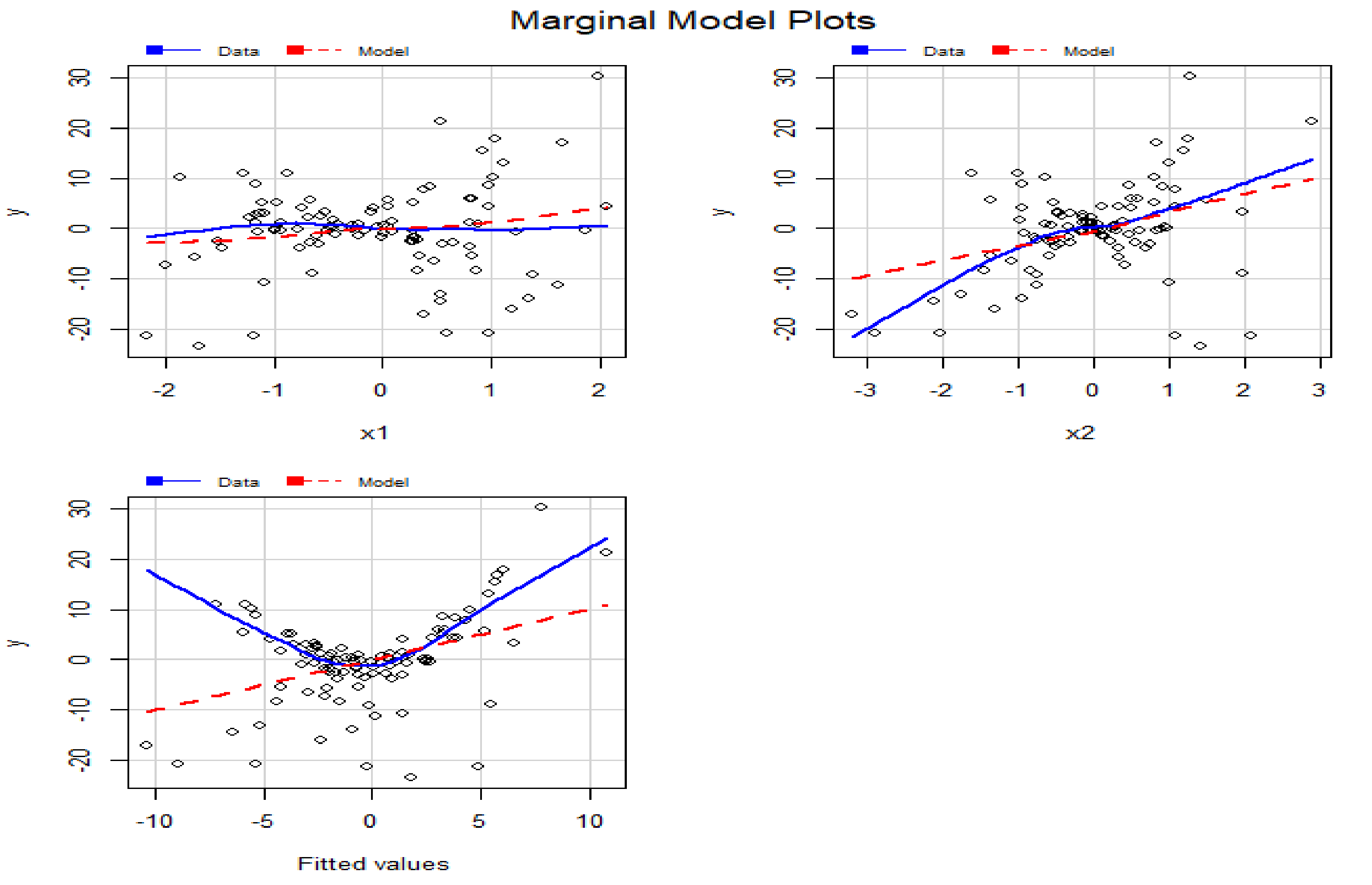


Marginal Model Plot

- The idea is very simple:
  - Plot y against a predictor (e.g. one of the x_j's or even $\hat{y}$); we'll call it x.
  - Use a nonparametric regression procedure (e.g. loess) to estimate E[y|x]
  - Use the fitted model to estimate E[y|x]
- The two should agree. If they do not,
  - x or y may need to be transformed
  - A term may be missing in the model
  - (or both!)

Marginal Model Plots

> mmps(lm.x1px2)



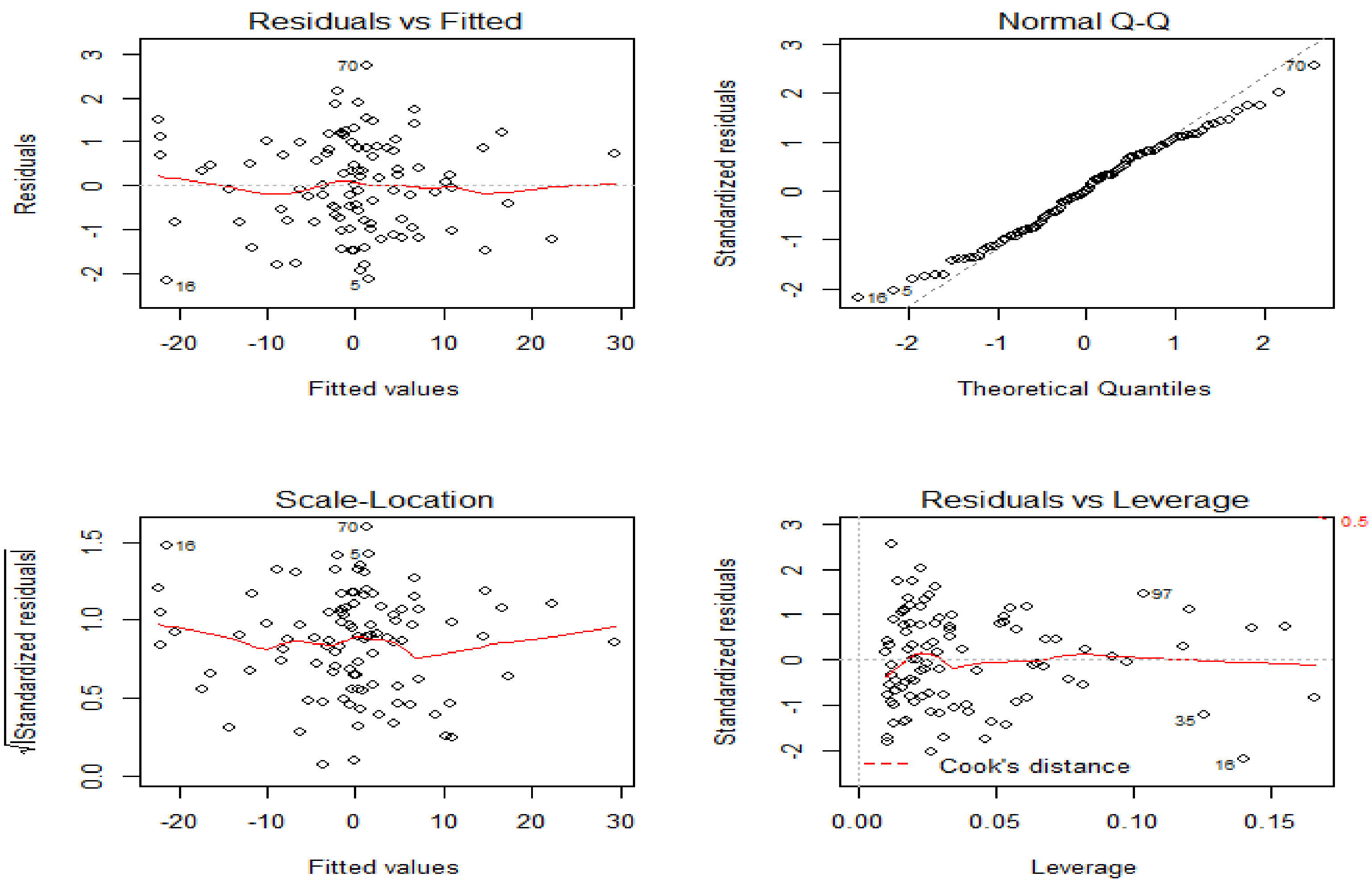
The “right” model (with interaction)

```
> summary(lm.x1mx2)

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


Coefficients:
          Est    SE      t     p
(Int)   0.77  0.11   7.06  0.00 ***
x1      0.69  0.12   5.93  0.00 ***
x2      2.03  0.11  18.33  0.00 ***
x1:x2   9.90  0.13  73.58  0.00 ***
---
```

```
Residual standard error:
1.079 on 96 degrees of
freedom
Multiple R-squared:  0.9856,
Adjusted R-squared:  0.9851
F-statistic:  2187 on 3 and
96 DF,  p-value: < 2.2e-16
```

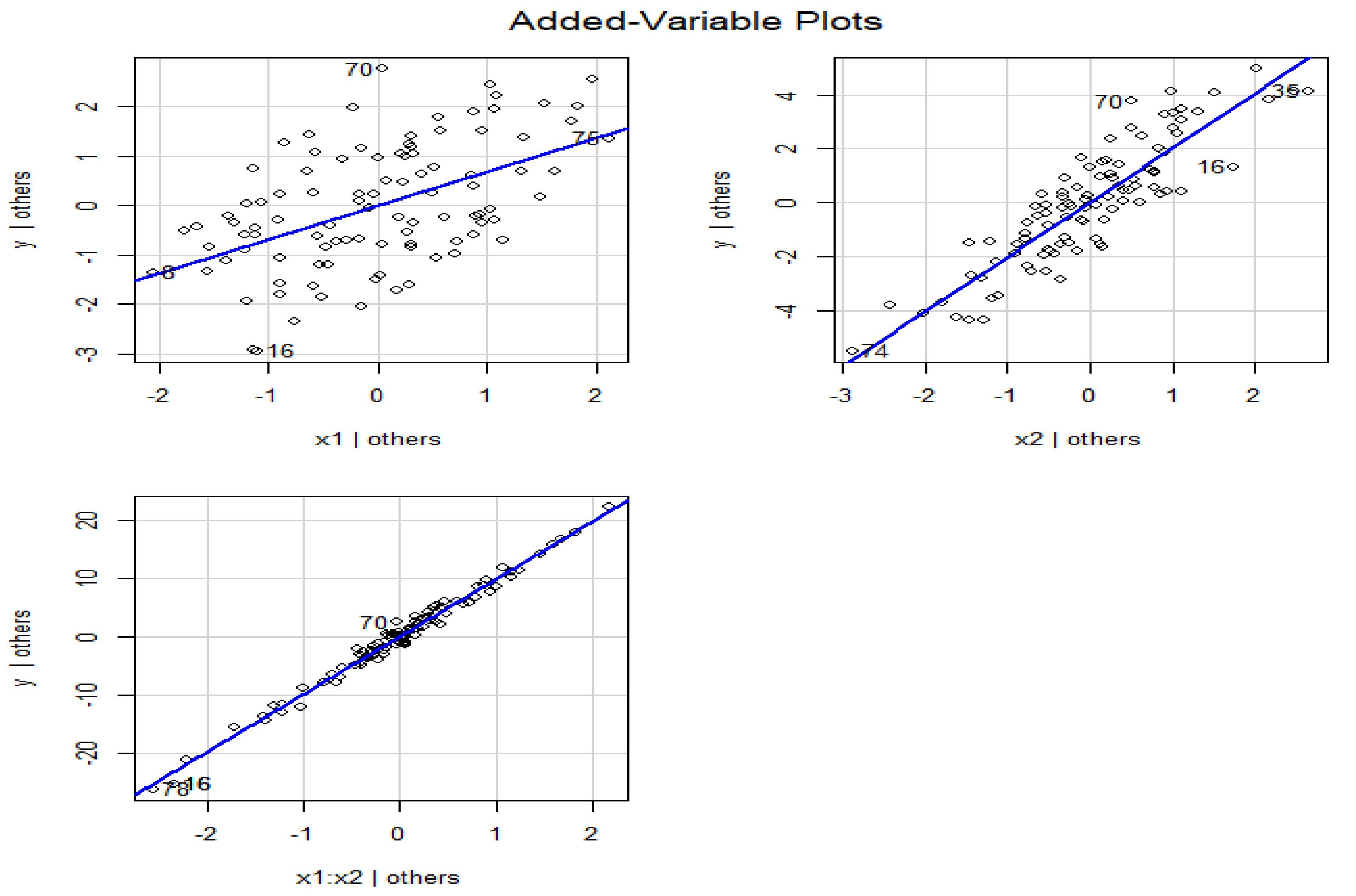
Casewise Diagnostic Plots

```
> par(mfrow=c(2,2))
> plot(lm.x1mx2)
```



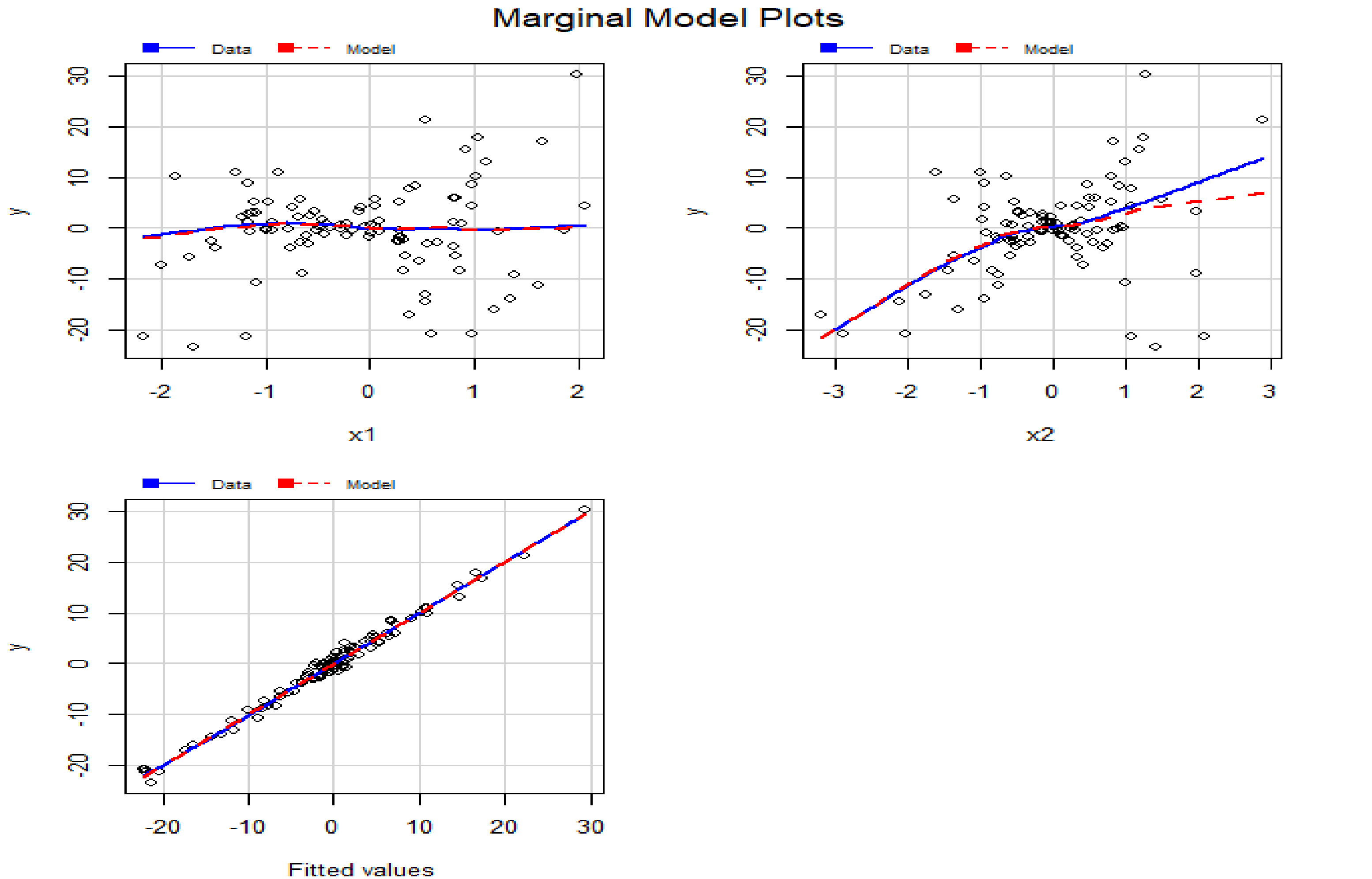
Added Variable Plots

> avPlots(lm.x1mx2)



Marginal Model Plots

> mmps(lm.x1mx2)



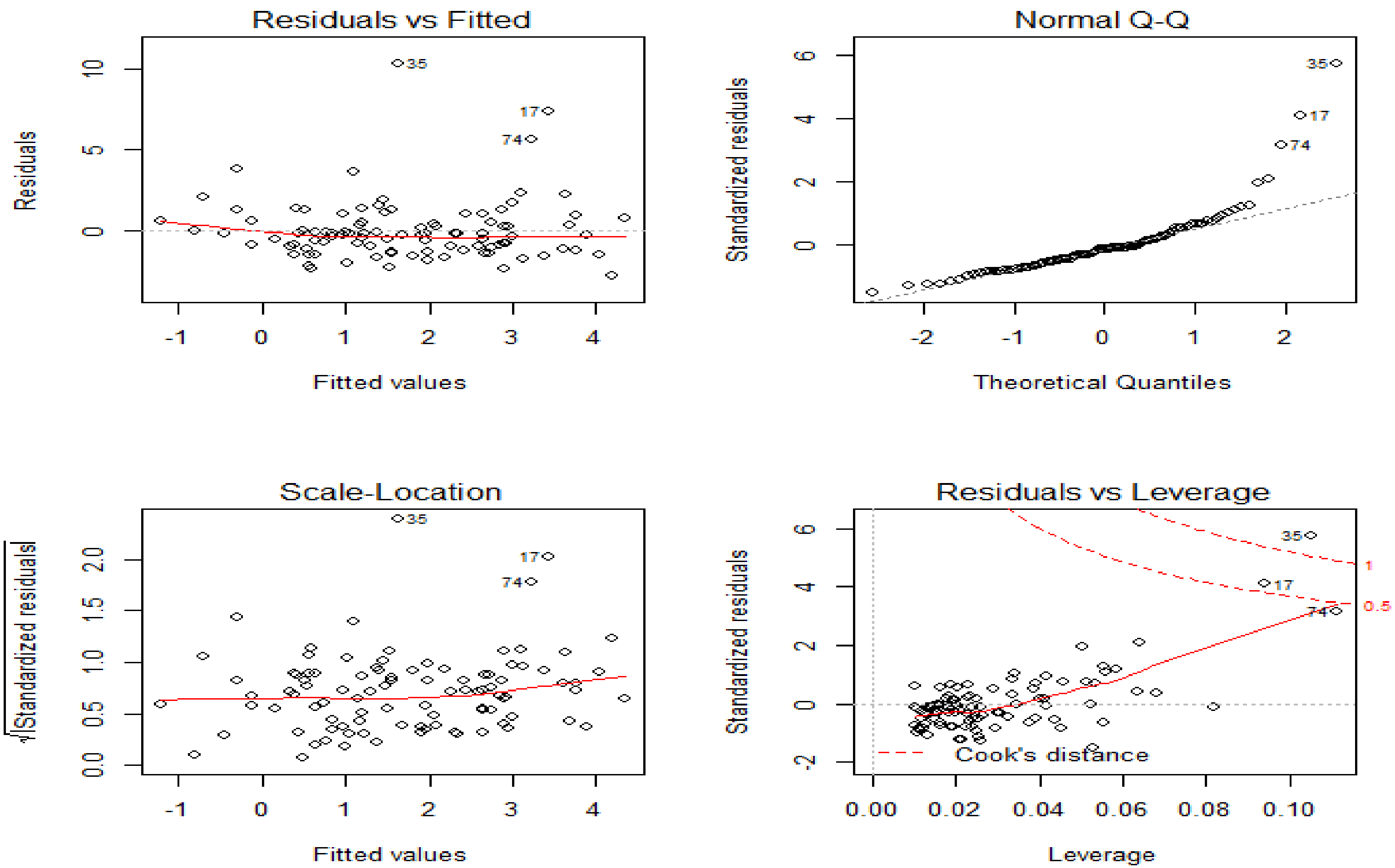
Another example

```
> y <- 1 + x1 + x2^2 +
+ rnorm(100)
>
> lm.x1px2 <- lm(y ~ x1 + x2)
> lm.x1mx2 <- lm(y ~ x1 * x2)
> lm.x1px2sq <- lm(y ~ x1 +
+ I(x2^2))
>
> summary(lm.x1px2)

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

```
Coefficients:
          Est   SE     t    p
(Int)   1.82 0.19  9.49 0.00 ***
x1      1.24 0.20  6.08 0.00 ***
x2     -0.30 0.19 -1.55 0.12
---
Residual standard error: 1.904
on 97 degrees of freedom
Multiple R-squared:  0.3014,
Adjusted R-squared:  0.287
F-statistic: 20.93 on 2 and 97
DF,  p-value: 2.779e-08
```

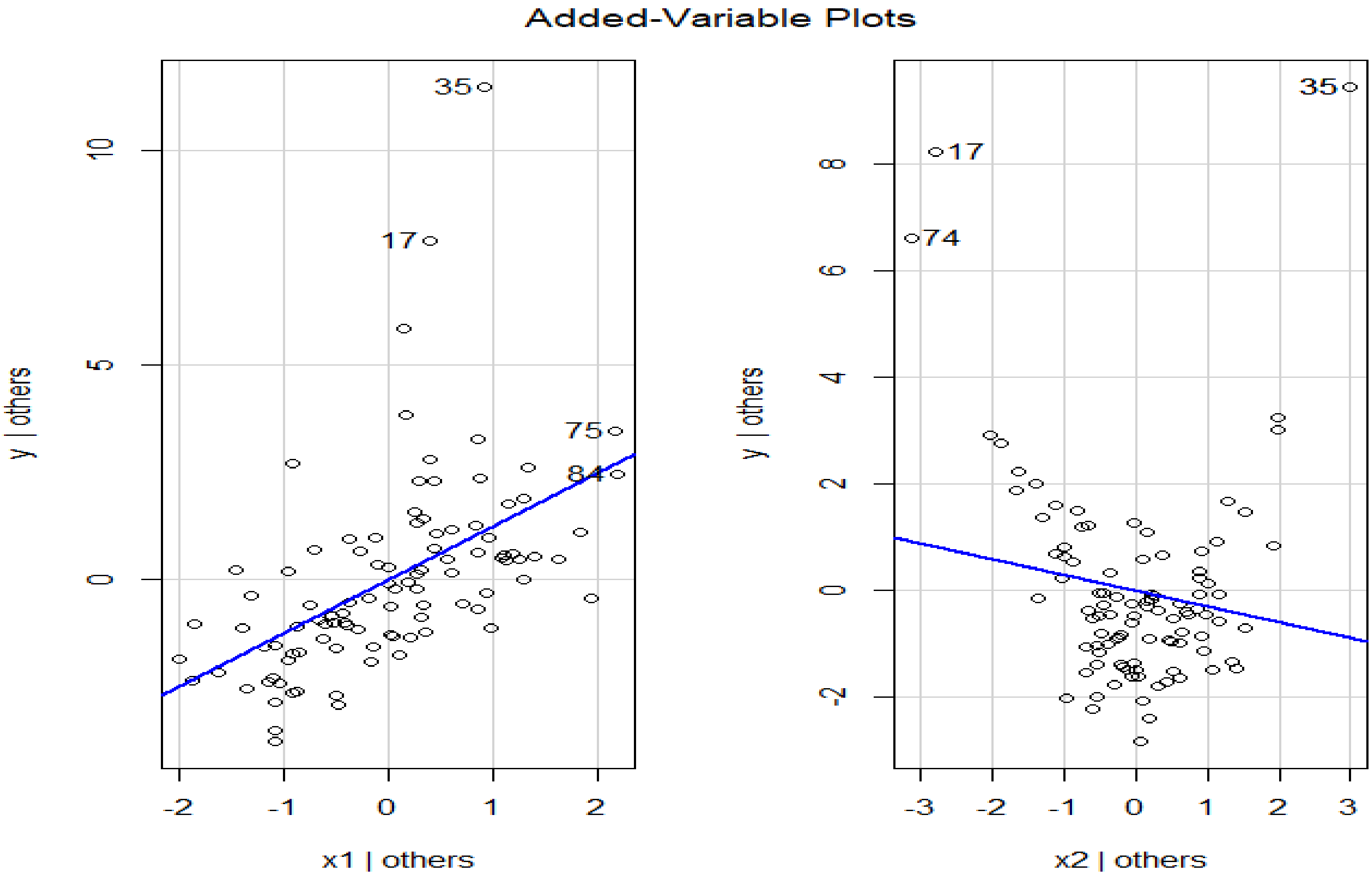
Casewise Diagnostic Plots

```
> par(mfrow=c(2,2))
> plot(lm.x1px2)
```



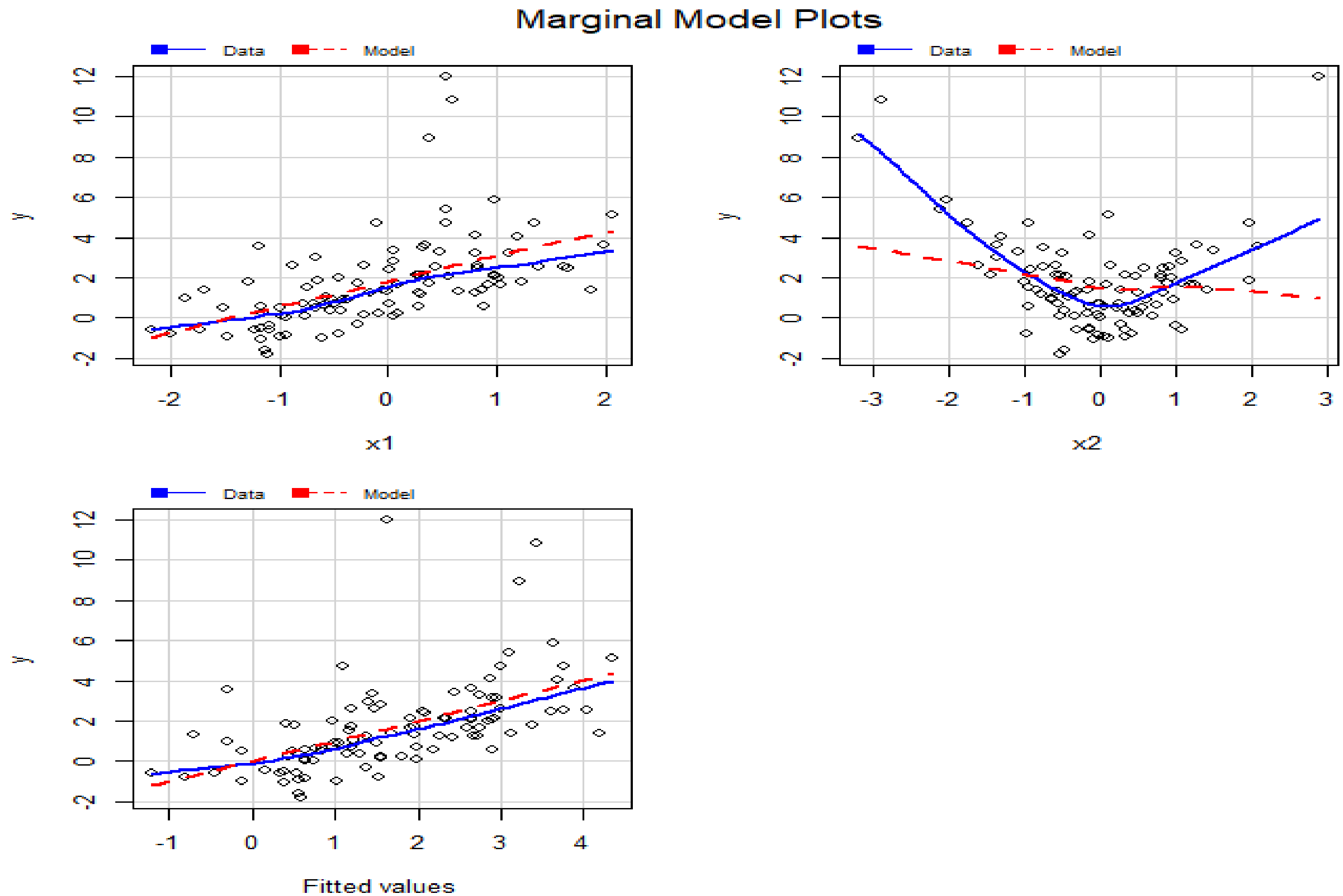
Added Variable Plots

> avPlots(lm.x1px2)



Marginal Model Plots

> mmps(lm.x1px2)

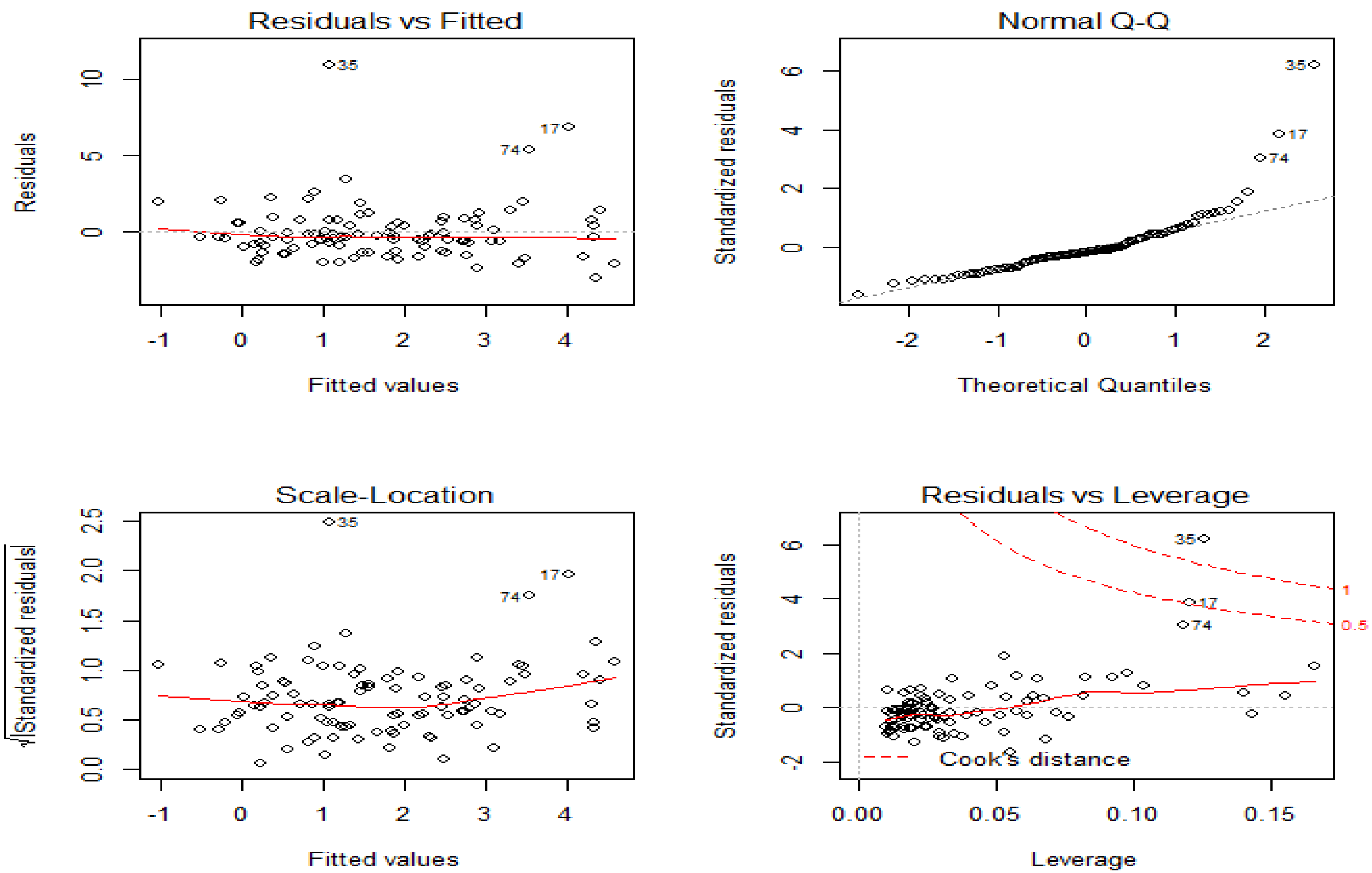


What if we think an interaction will fix it?

```
> summary(lm.x1mx2)

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

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7774     0.1895   9.380 3.19e-15 ***
x1            1.2891     0.2024   6.370 6.52e-09 ***
x2           -0.2321     0.1922  -1.208   0.2301
x1:x2        -0.4607     0.2340  -1.969   0.0518 .
---
Residual standard error: 1.877 on 96 degrees of freedom
Multiple R-squared:  0.3286,    Adjusted R-squared:  0.3076
F-statistic: 15.66 on 3 and 96 DF,  p-value: 2.29e-08
```

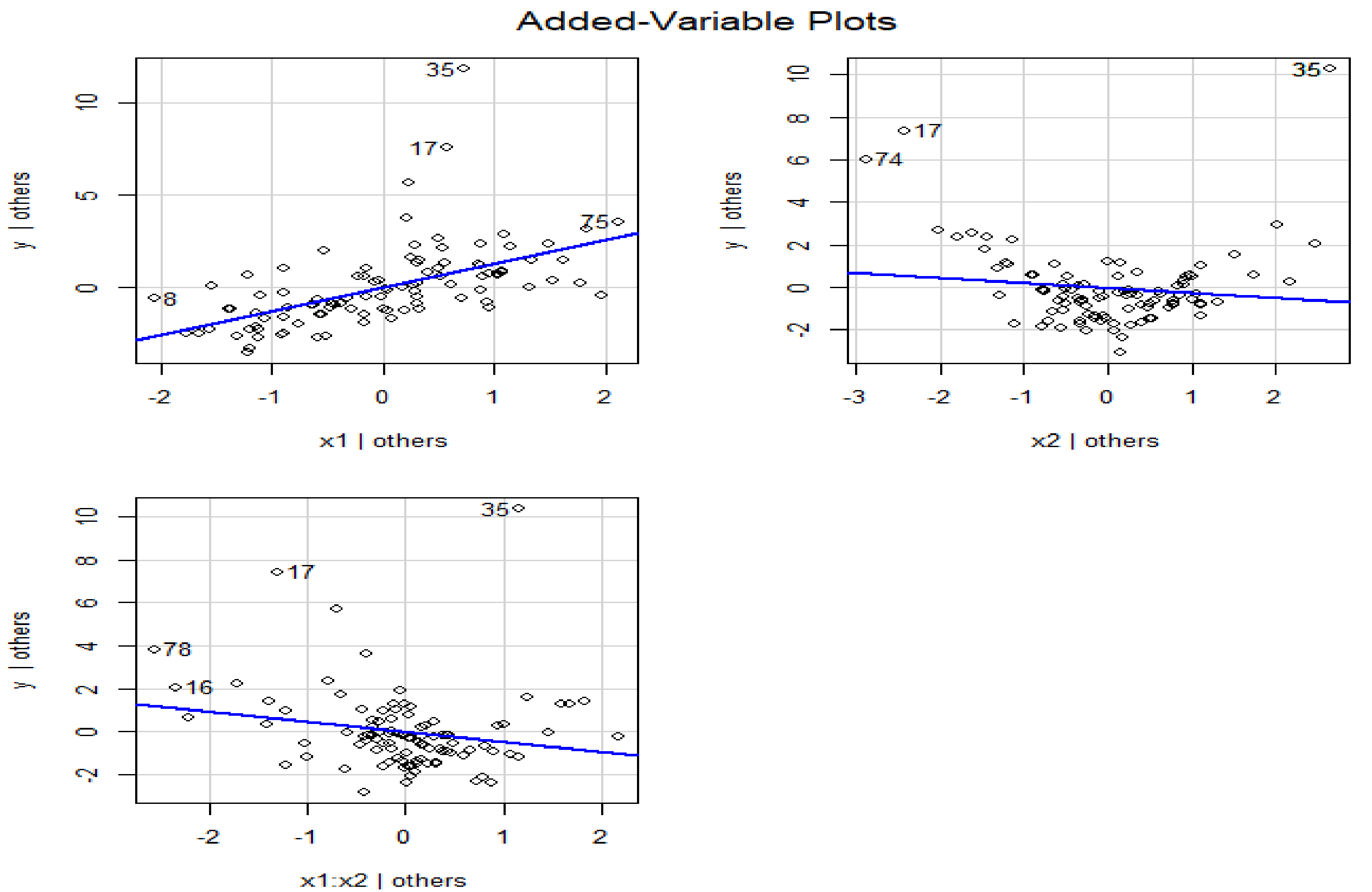
Casewise Diagnostic Plots

```
> par(mfrow=c(2,2))
> plot(lm.x1mx2)
```



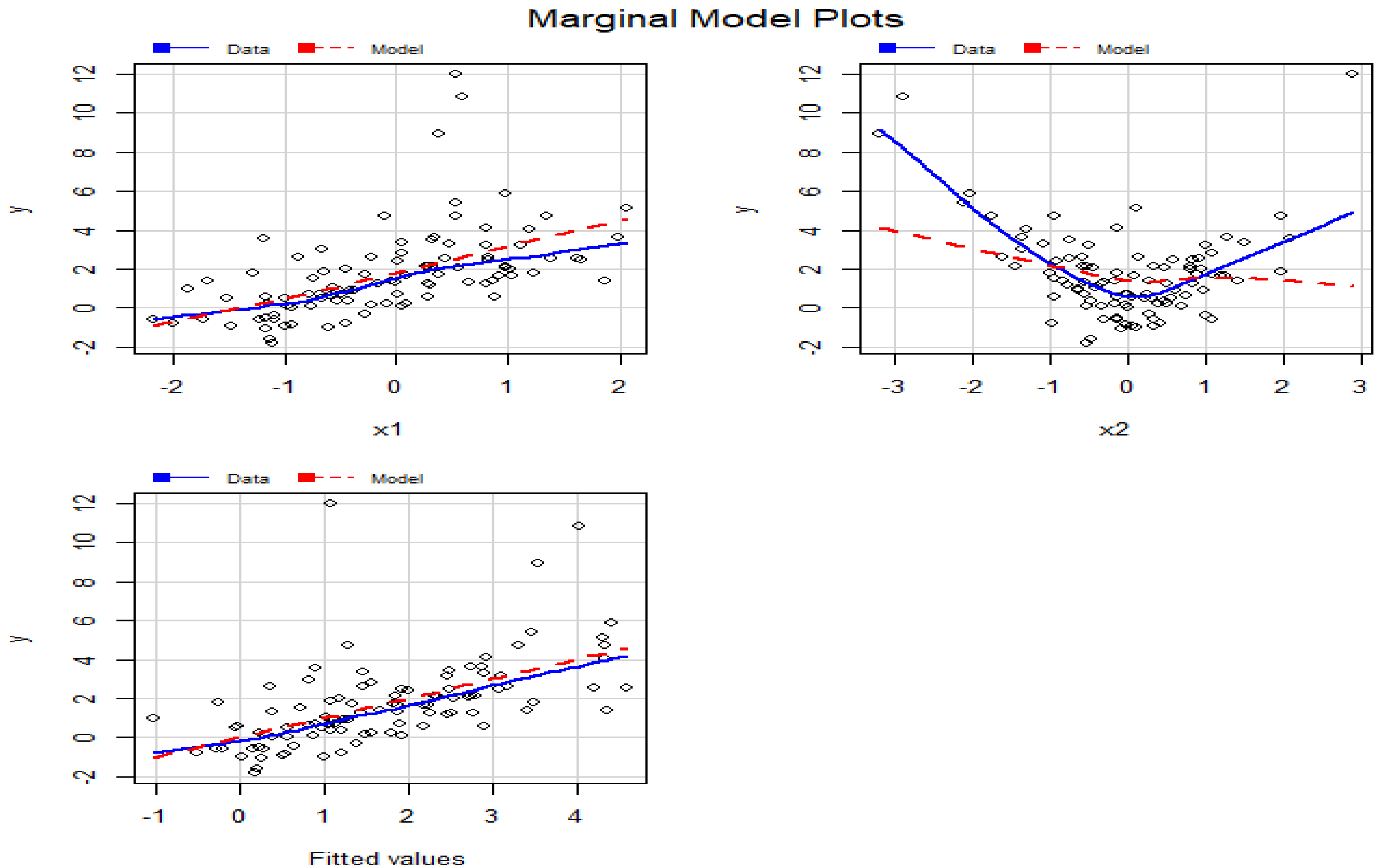
Added Variable Plots

> avPlots(lm.x1mx2)



Marginal Model Plots

> mmps(lm.x1mx2)

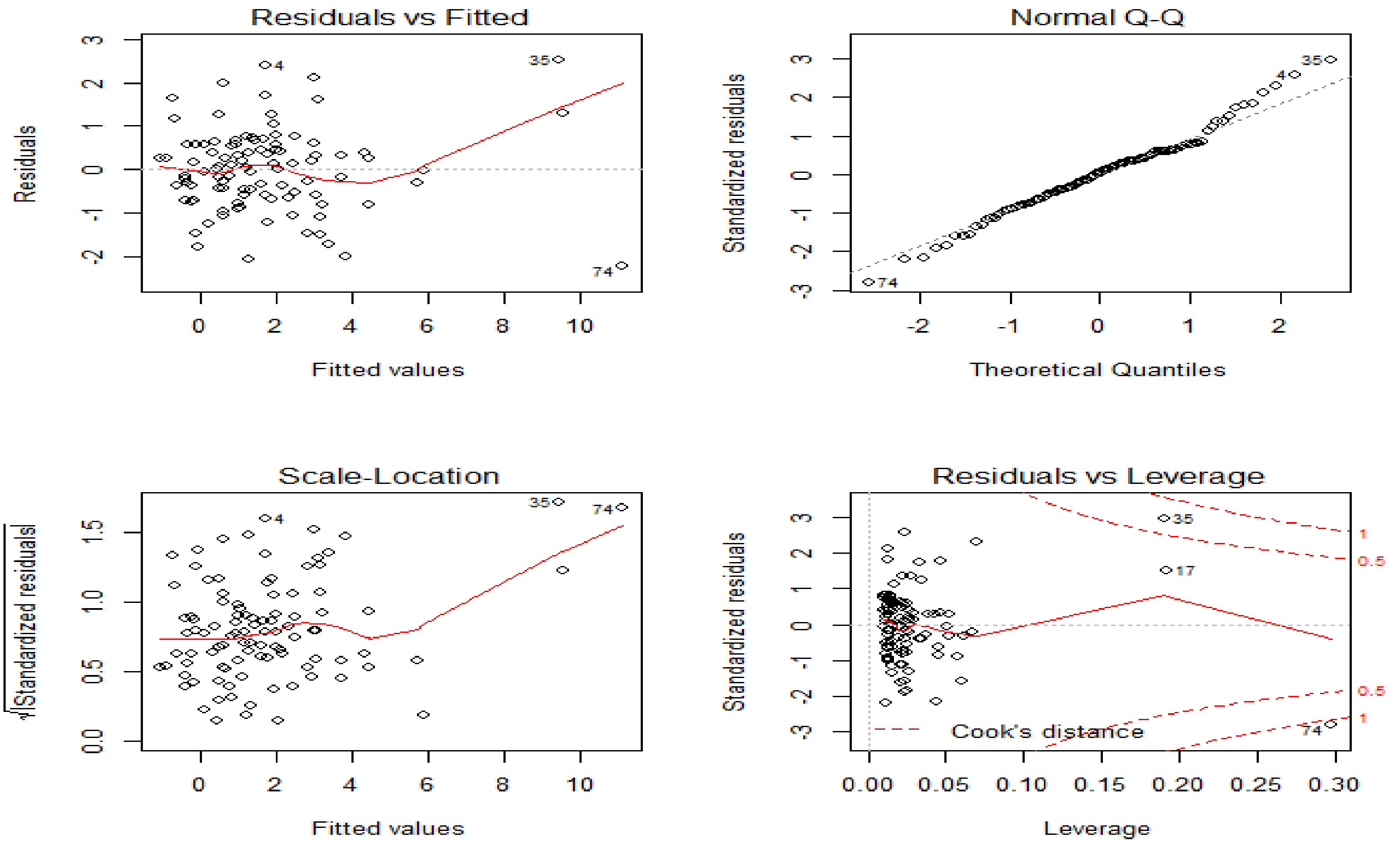


And now the correct model (with x2 squared term)..

```
> summary(lm.x1px2sq)
Call:
lm(formula = y ~ x1 + I(x2^2))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.85241    0.11038   7.722 1.04e-11 ***
x1           1.04216    0.10171  10.247  < 2e-16 ***
I(x2^2)      0.96300    0.05522  17.438  < 2e-16 ***
---
Residual standard error: 0.9481 on 97 degrees of freedom
Multiple R-squared:  0.8269,    Adjusted R-squared:
0.8233
F-statistic: 231.6 on 2 and 97 DF,  p-value: < 2.2e-16
```

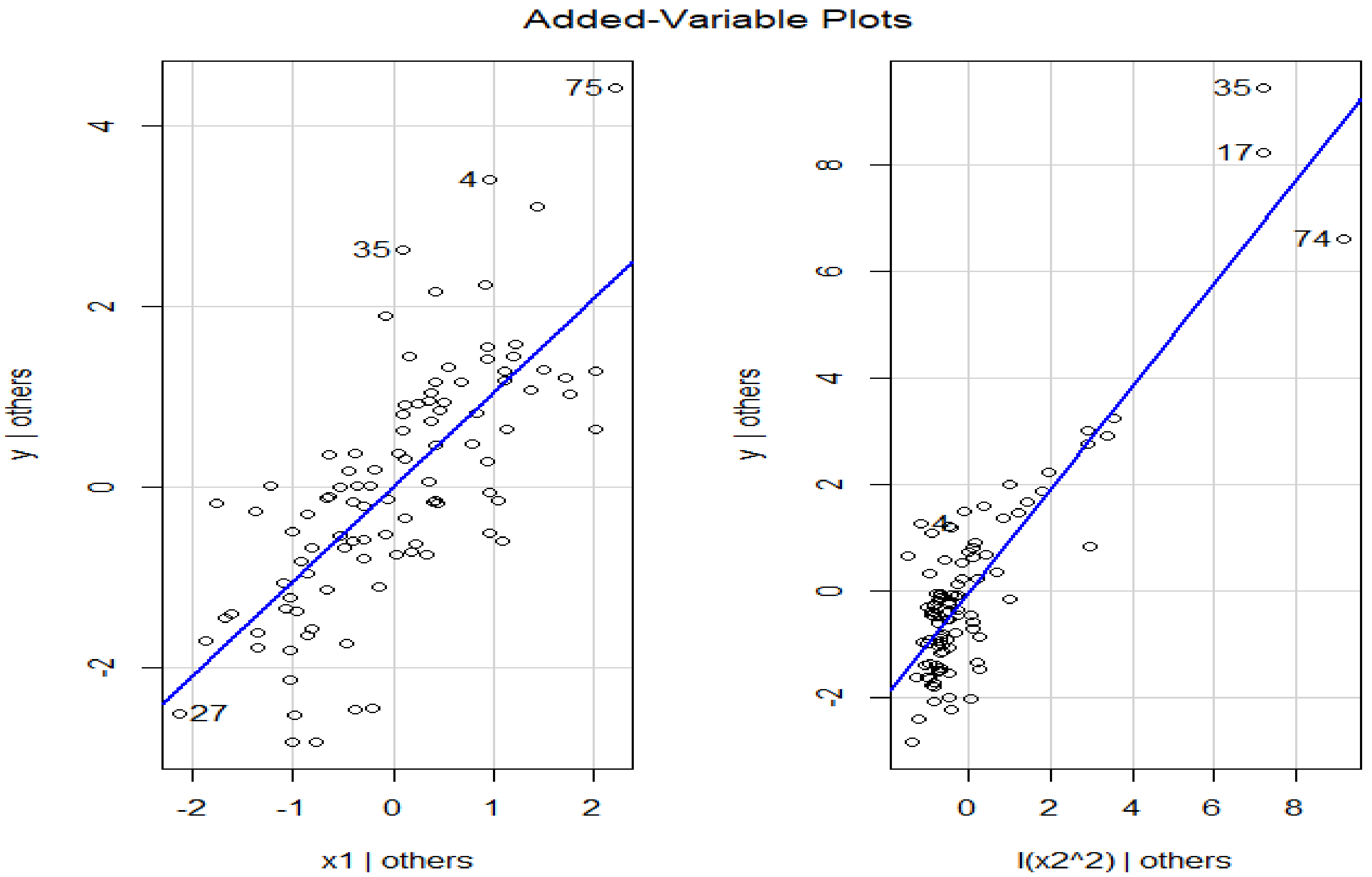
Casewise Diagnostic Plots

> par(mfrow=c(2,2))
> plot(lm.x1px2sq)



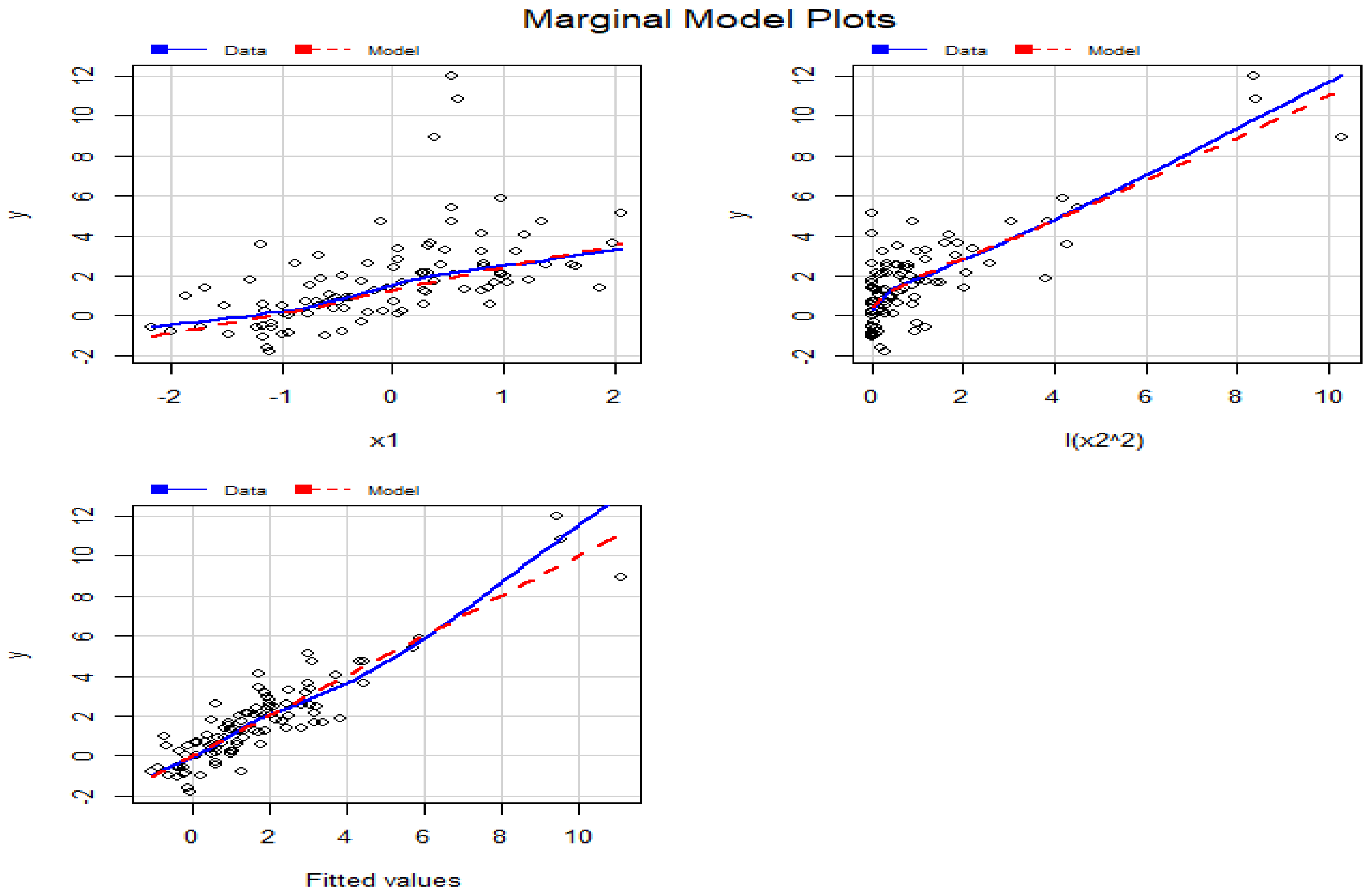
Added Variable Plots

> avPlots(lm.x1px2sq)



Marginal Model Plots

> mmps(lm.x1px2sq)



Moral of the Story

- Nonlinearity can show up in lots of ways, in lots of graphs
  - In casewise diagnostic plots
    - As nonlinearity
    - As nonconstant variance
    - As Non-normality (!!!)
  - In added-variable and marginal model plots
    - Nonlinearity shows up more clearly
    - Not always obvious what the right transformation would be.

Transformations

- As before, transformations useful for overcoming
  - Non-linearity
  - Non-normality
  - Non-constant variance
  - Poor interpretability

- Same transformation toolkit as before:
  - Inverse response plots – nonlinearity
  - Box-Cox – Non-normality
  - Variance-Stabilizing Transformations
  - Logs, scale-location transformations, etc.

- Use visual checks (scatter plots, histograms, normal qq plots, etc.), not just automatic methods!
- Don't forget that substantive knowledge and interpretability should predominate!

Summary

- Graphical methods for detecting non-linearity
  - Casewise diagnostic plots
  - Added variable plots
  - Marginal model plots
- Example(s)
- Transformations – a brief summary