36-617: Applied Linear Models

- Graphical Tools for Transformations (catching up!)
- Over- & Under-Specifying A Model

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

Announcements

- Quiz 02 – see in week 03 folder
  - 1 - learned.pdf
  - 1 - mystified.pdf
- Quiz 03 – Covers 6.4, 6.5, 6.6 (out at 5pm)
- HW04 – Out later today; due next Monday
- Reading
  - This week: Sheather 6.4, 6.5, 6.6, 7.1, 7.2
    - (supplemental: ISLR 3.3.3; G&H Ch 4)
  - Next week: Sheather, 7.3, 7.4, 8.1, 8.2
  - Supplementary: ISLR 3.3.3,& Ch 6; G&H Ch 4

Outline

- Graphical tools for Transformations (catching up!)
  - Added Variable Plots
  - Marginal Model Plots
  - Moral of the Story
- Over- and under-specifying a model
  - Too many predictors: Excess SE’s and Collinearity
  - Too few predictors: Omitted Variable Bias

Added-Variable Plots (add Z? or f(Z)?)

- Suppose the true model is

$$Y = X\beta + Z\gamma + \epsilon$$

- Let us fit the models

$$\begin{aligned} Y &= X\beta + \epsilon^{(1)} \text{ with residuals } \hat{e}^{(1)} = (I - H_X)Y \\ Z &= X\beta + \epsilon^{(2)} \text{ with residuals } \hat{e}^{(2)} = (I - H_X)Z \end{aligned}$$

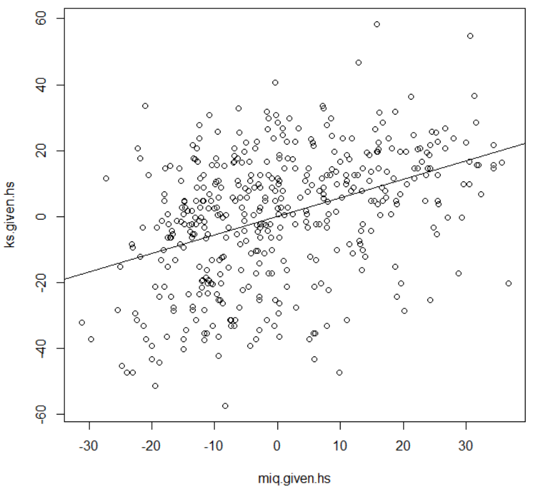
- If we multiply the true model by *(I-H_X)*, we get

$$\begin{aligned} (I - H_X)Y &= (I - H_X)X\beta + (I - H_X)Z\gamma + (I - H_X)\epsilon \\ \hat{e}^{(1)} &= 0 + \hat{e}^{(2)}\gamma + \epsilon^* \end{aligned}$$

so, plotting (or regressing) $\hat{e}^{(1)}$ on $\hat{e}^{(2)}$ will reveal γ !

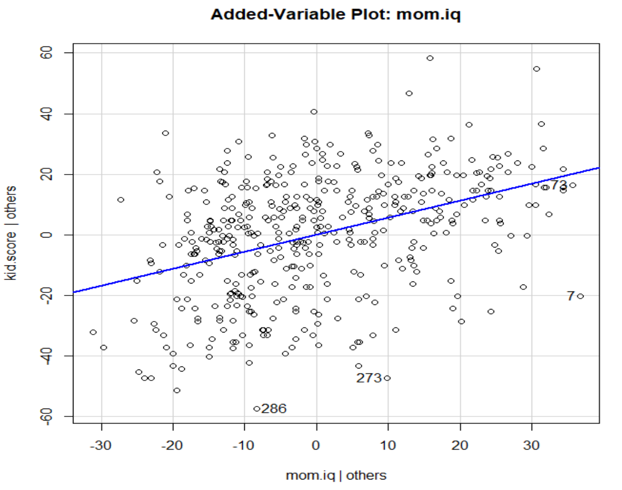
Added-variable plots: “graphical t-statistics”

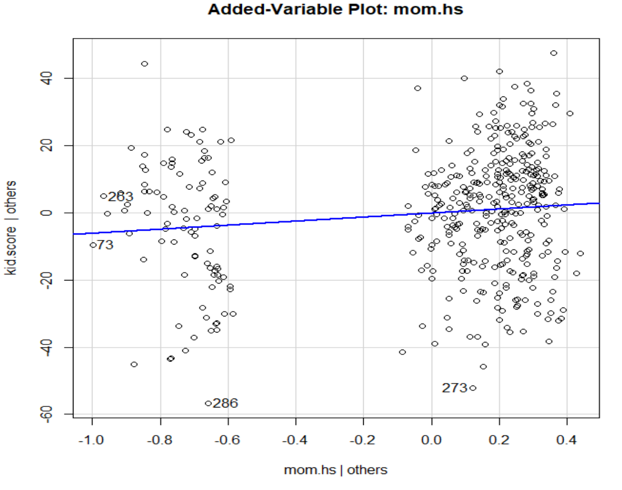
```
kidiq <- read.csv("kidiq.csv",header=TRUE)
round(summary(lm.3 <- lm(kid.score ~ mom.iq + mom.hs, data=kidiq))$coef,4)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   25.7315     5.8752  4.3797   0.0000
## mom.iq         0.5639     0.0606  9.3094   0.0000
## mom.hs         5.9501     2.2118  2.6902   0.0074
ks.given.hs <- residuals(lm(kid.score ~ mom.hs, data=kidiq))
miq.given.hs <- residuals(lm(mom.iq ~ mom.hs, data=kidiq))
plot(ks.given.hs ~ miq.given.hs)
abline(lm(ks.given.hs ~ miq.given.hs))
round(summary(lm.4 <- lm(ks.given.hs ~ miq.given.hs))$coef,4)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.0000     0.8695  0.0000        1
## miq.given.hs    0.5639     0.0605  9.3202        0
##
## The t-statistic in a multiple regression gives the same information
## as the t-statistic in an added-variable regression: it tests the
## significance of adding the variable *after* accounting for all
## other X's in the model
##        In this sense, the added variable plot is the graphical equivalent
##        of the t-statistic
```



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")
```



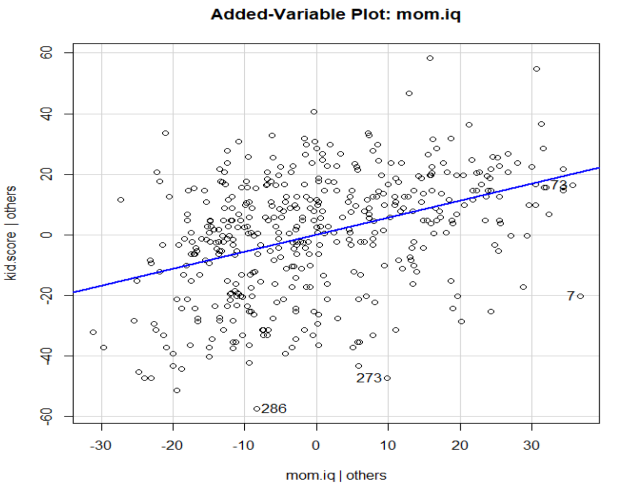


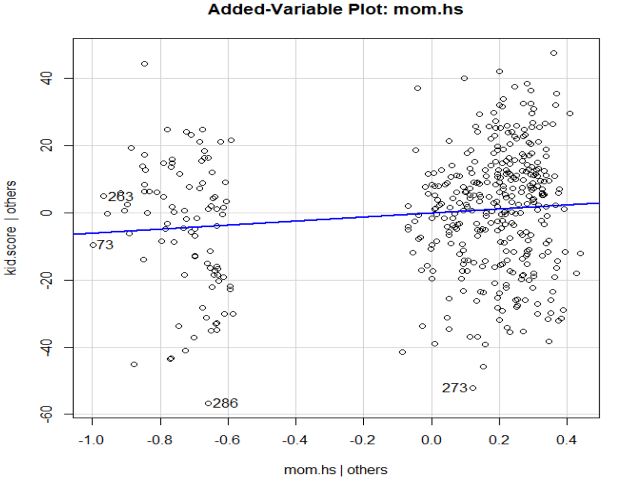
Added-Variable Plots – Interpretations

- Shows γ as the effect of *Z* after controlling for *X*, on *Y*, after controlling for *X*
- Allows you to visually assess the importance of γ , after controlling for all the other X's.
  - A visual form of the t-statistic!
- Also allows you to check for nonlinearity in predicting *Y* from *Z*, after controlling for *X*
- Another plot that allows us to assess nonlinearity is the "marginal model plot" – *later in this lecture*

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")
```





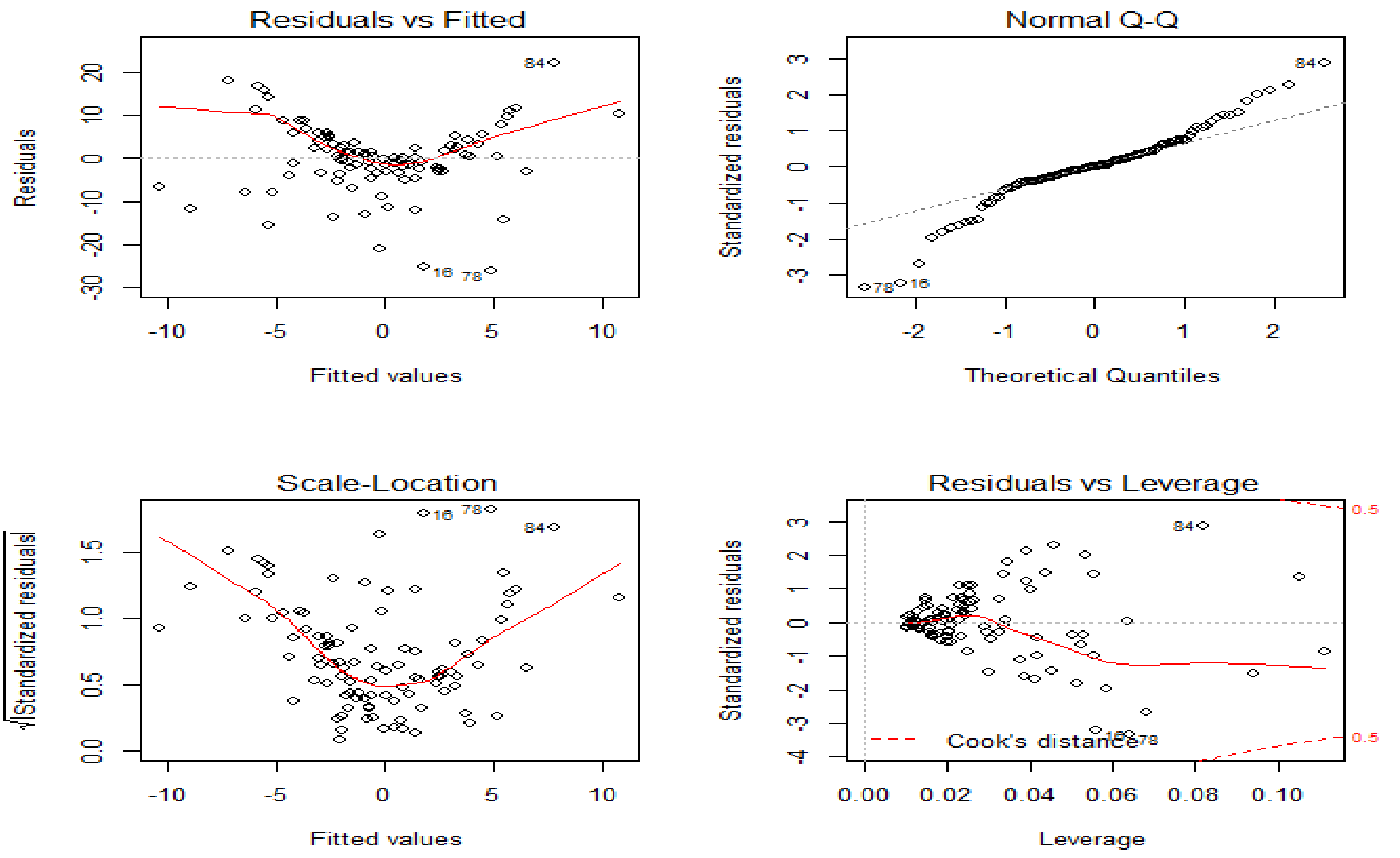
An example

```
> 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
```

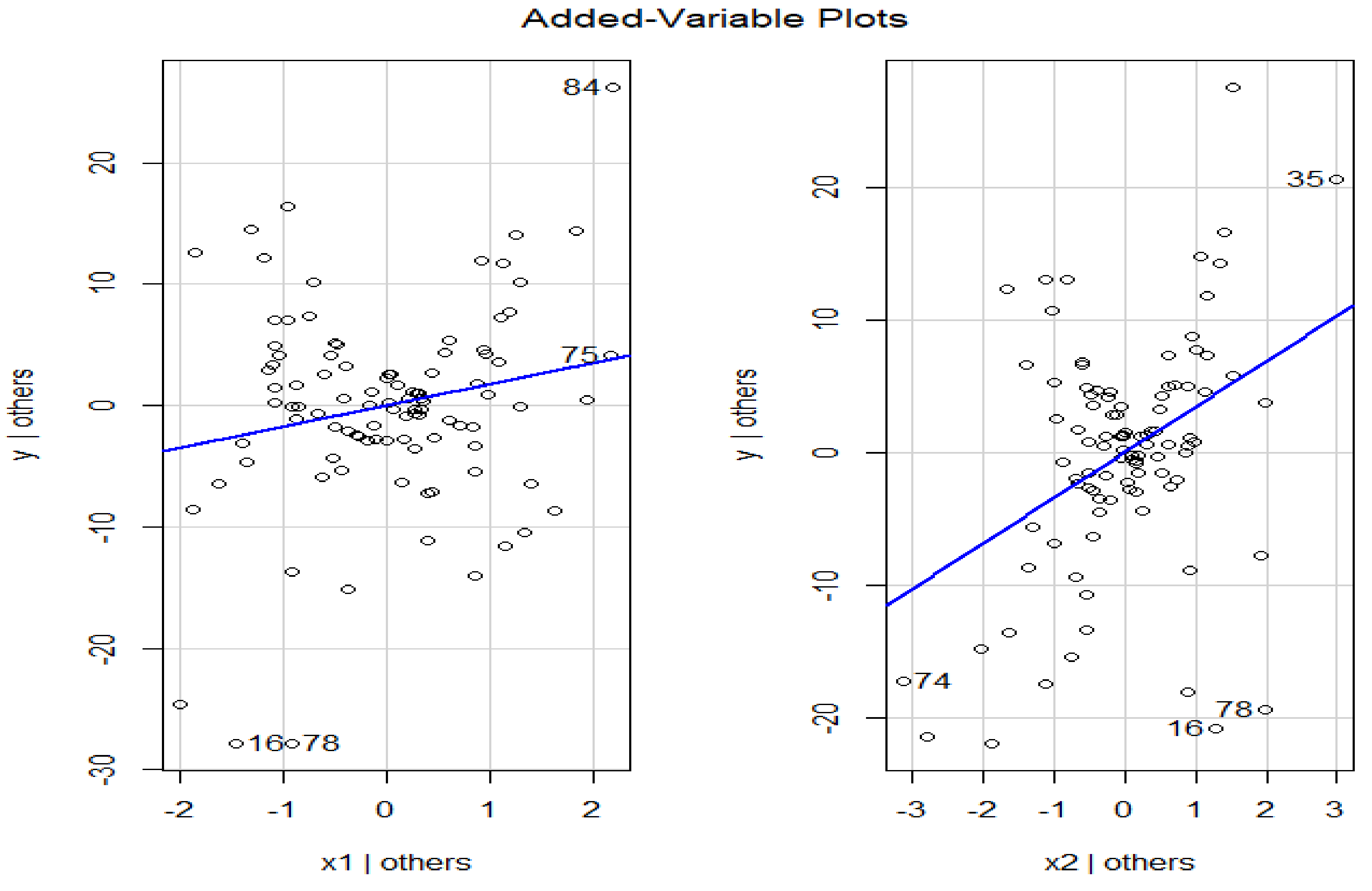
Casewise Diagnostic Plots

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



Added Variable Plots

> avPlots(lm.x1px2)

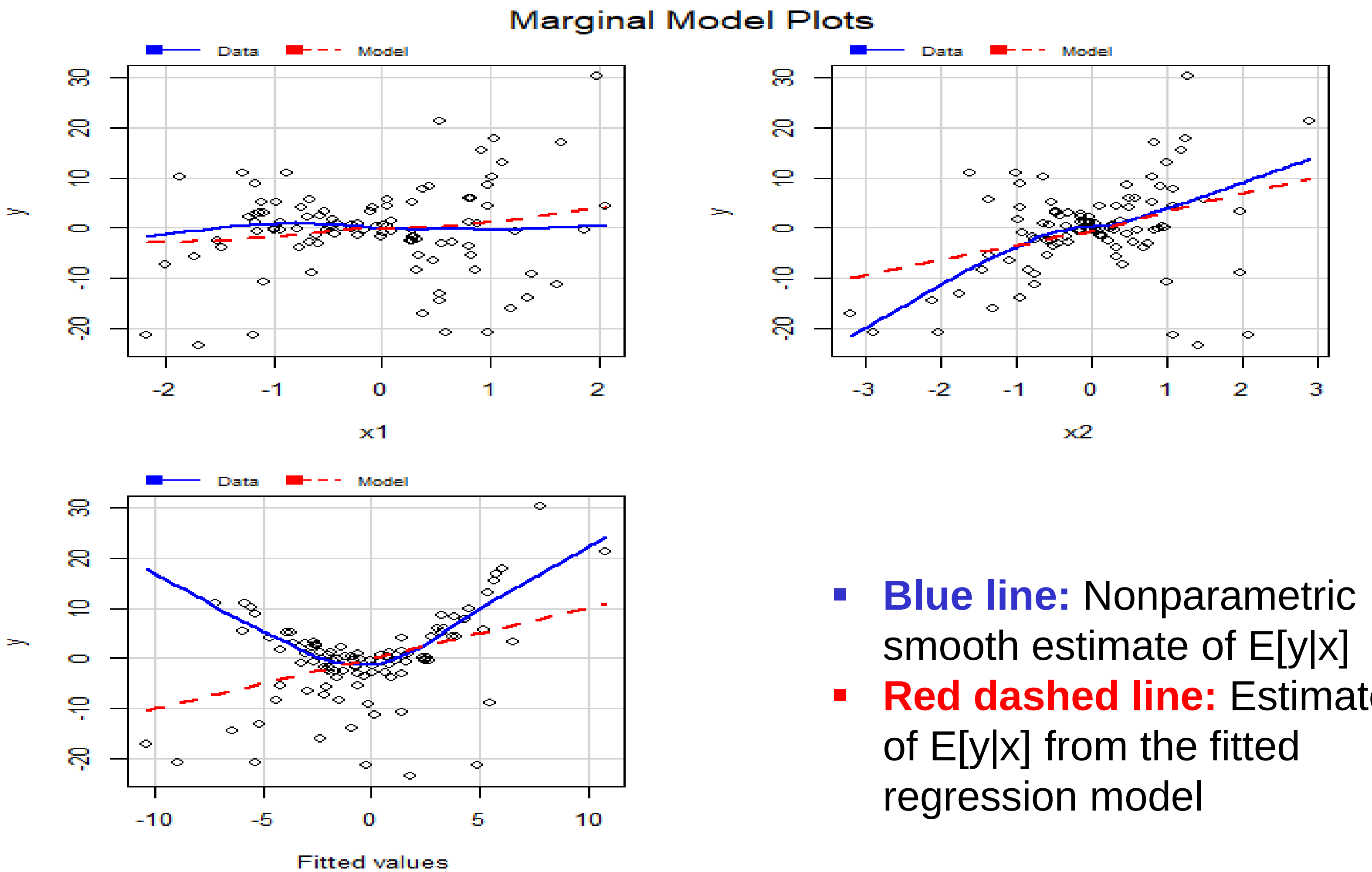


Marginal Model Plot

- 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)



- **Blue line:** Nonparametric smooth estimate of E[y|x]
- **Red dashed line:** Estimate of E[y|x] from the fitted regression model

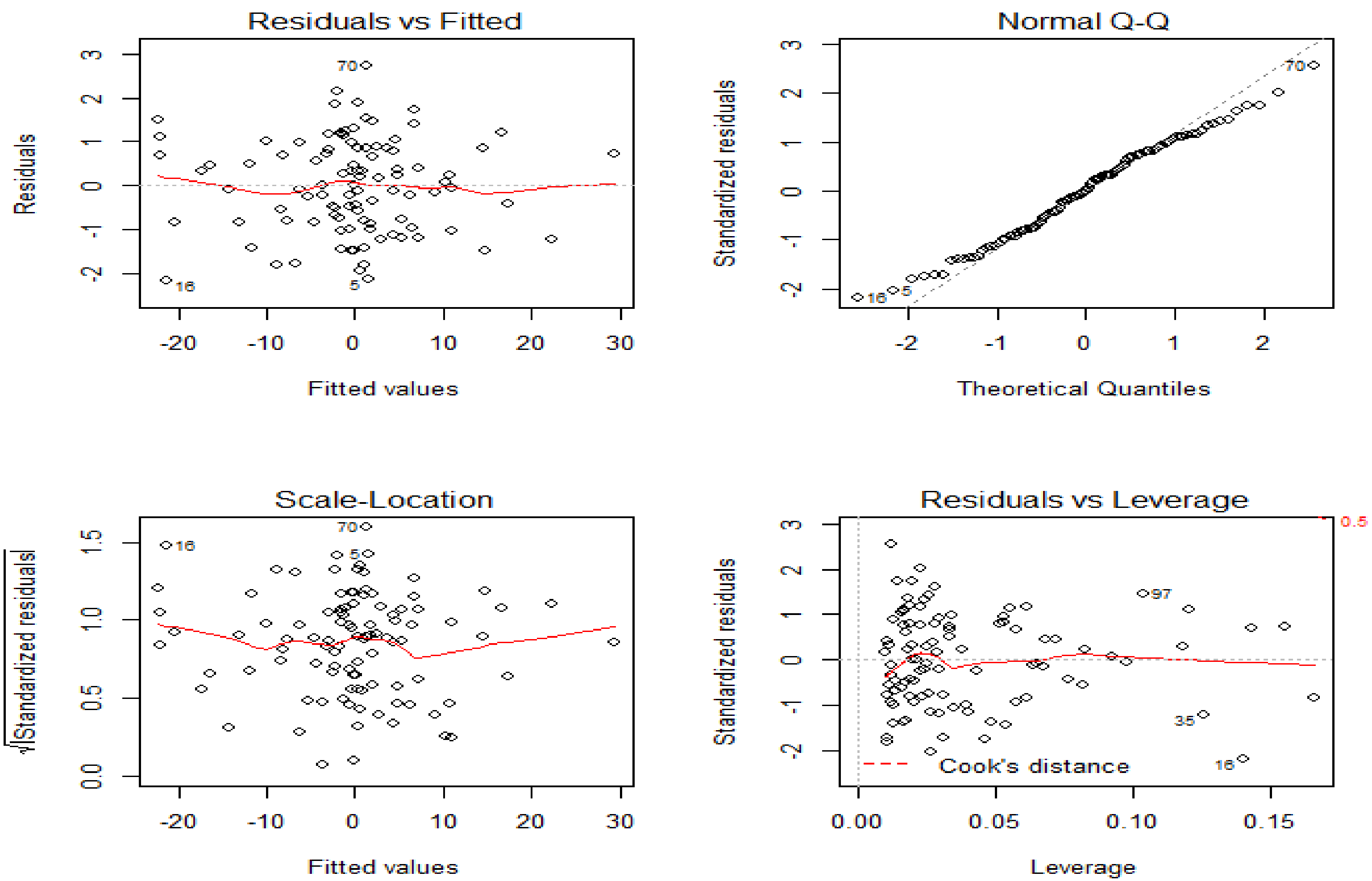
The “right” model (with interaction)

```
> 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
```

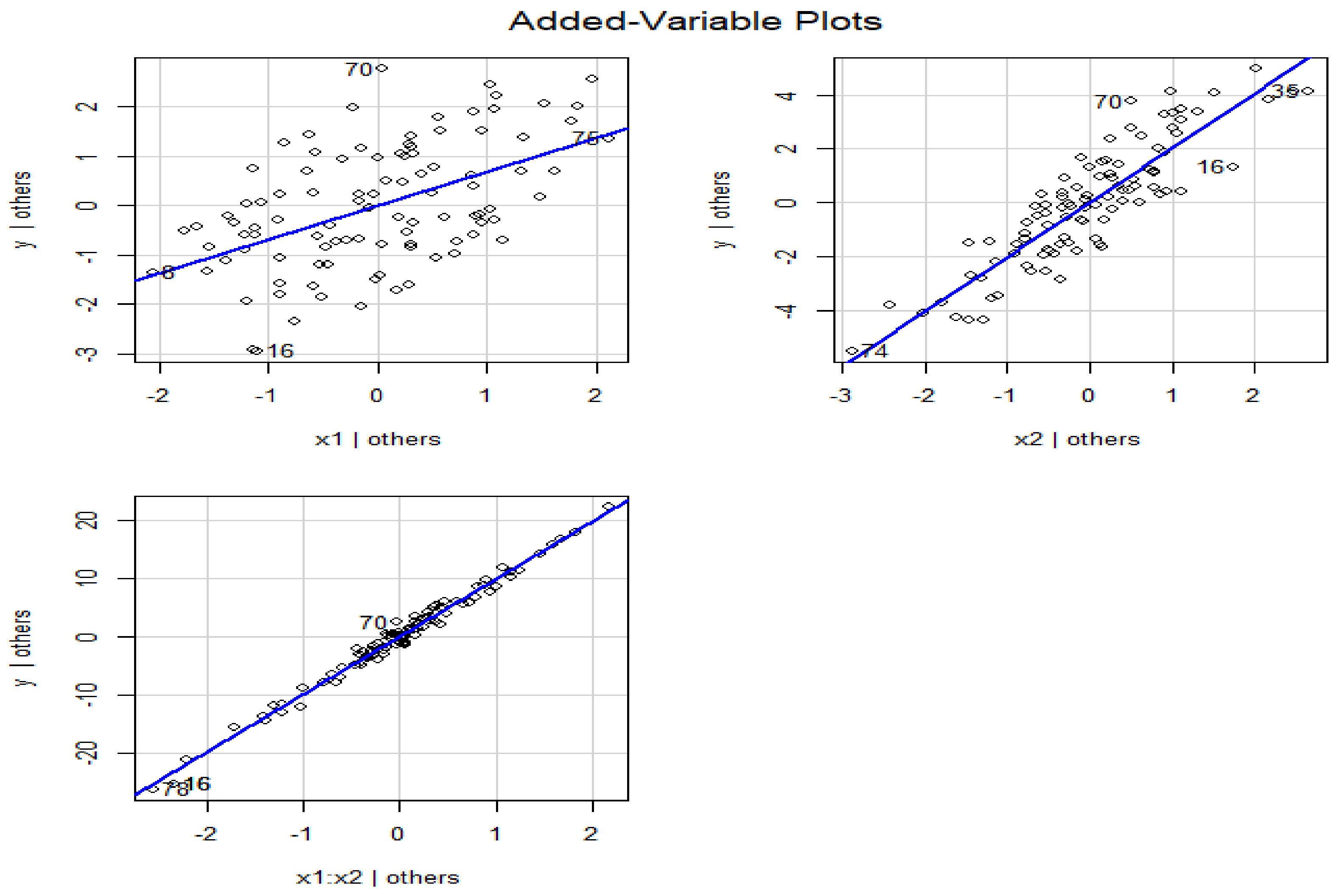
Casewise Diagnostic Plots

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



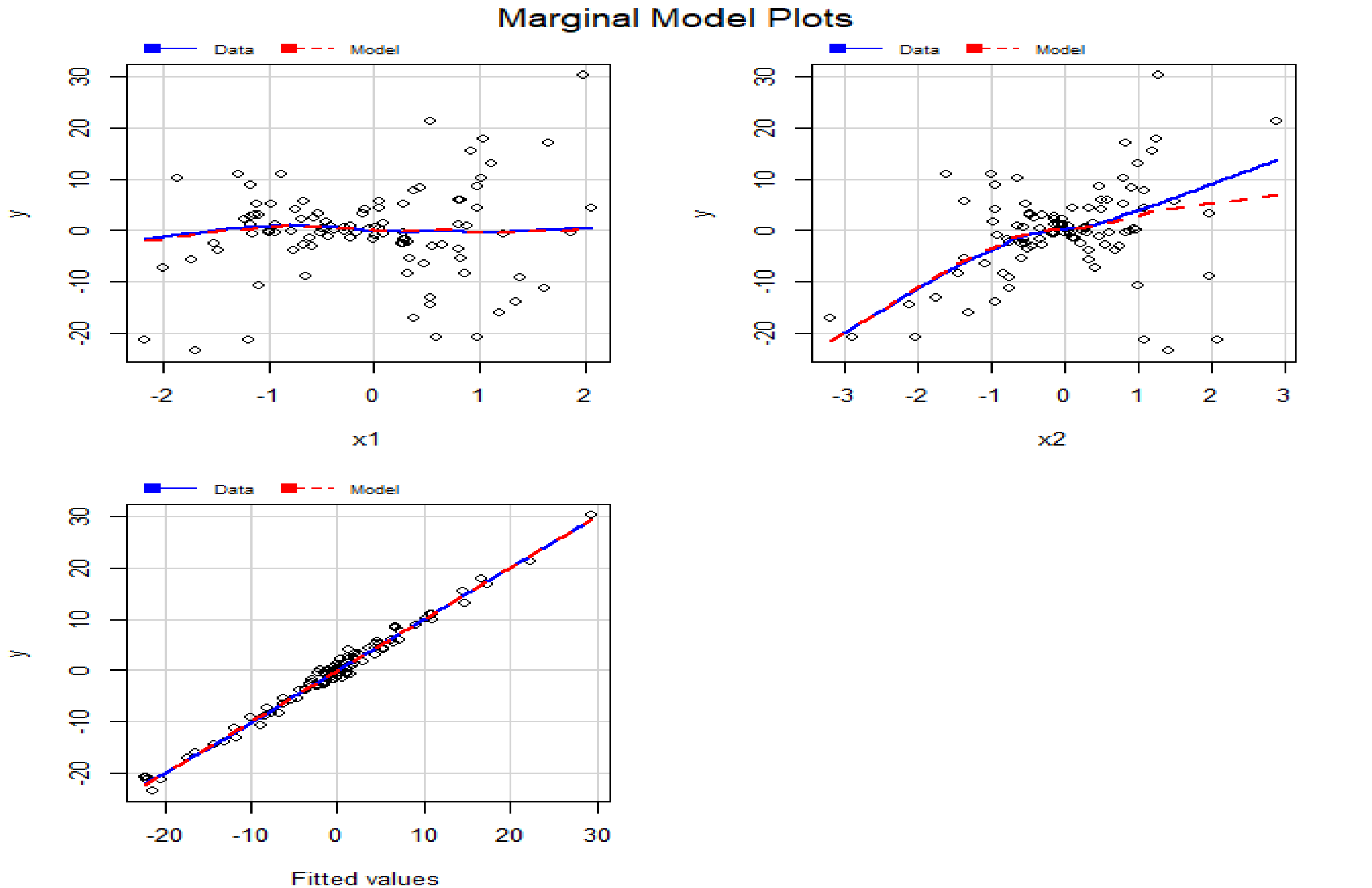
Added Variable Plots

> avPlots(lm.x1mx2)



Marginal Model Plots

> mmps(lm.x1mx2)



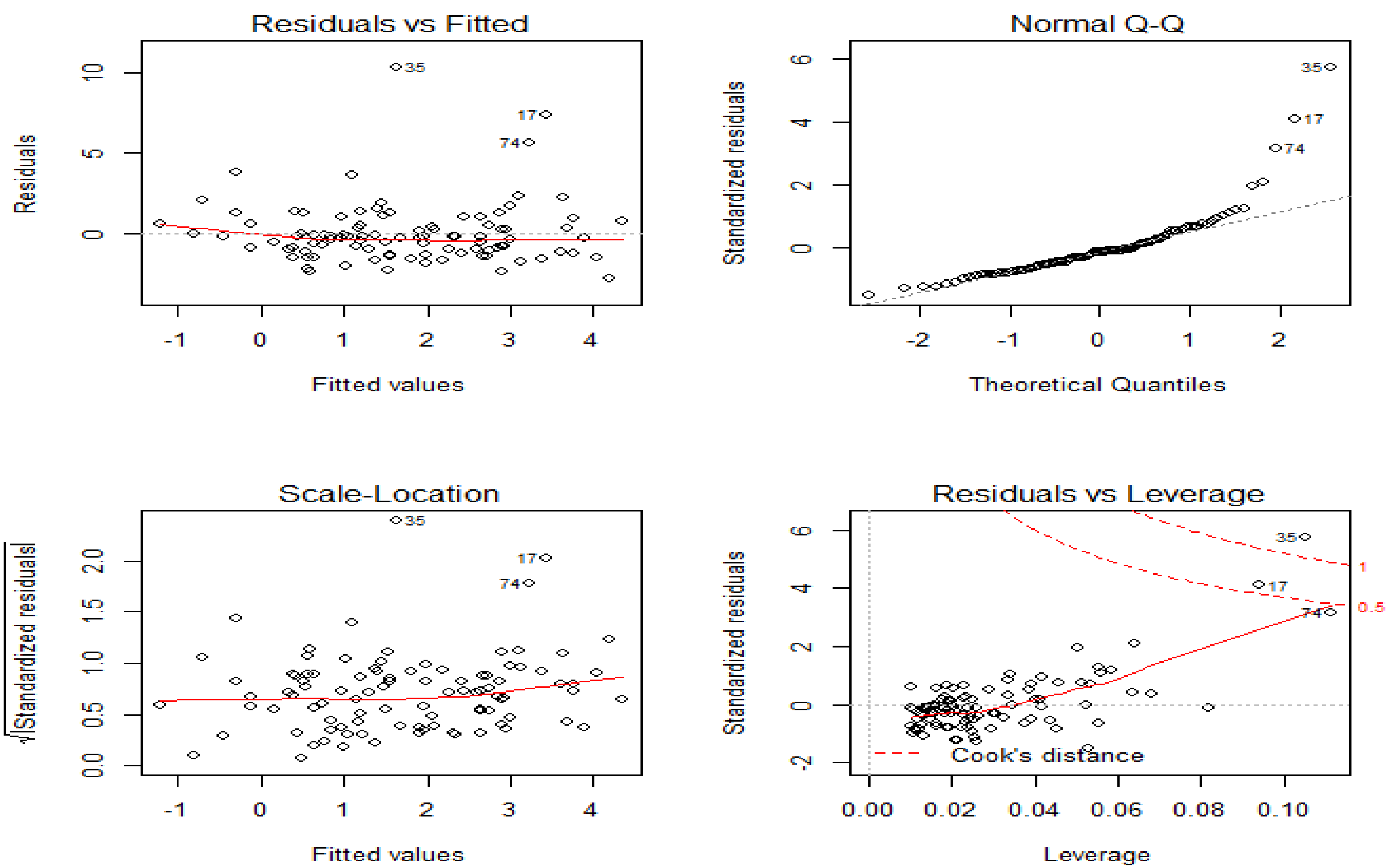
Another example

```
> 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
```

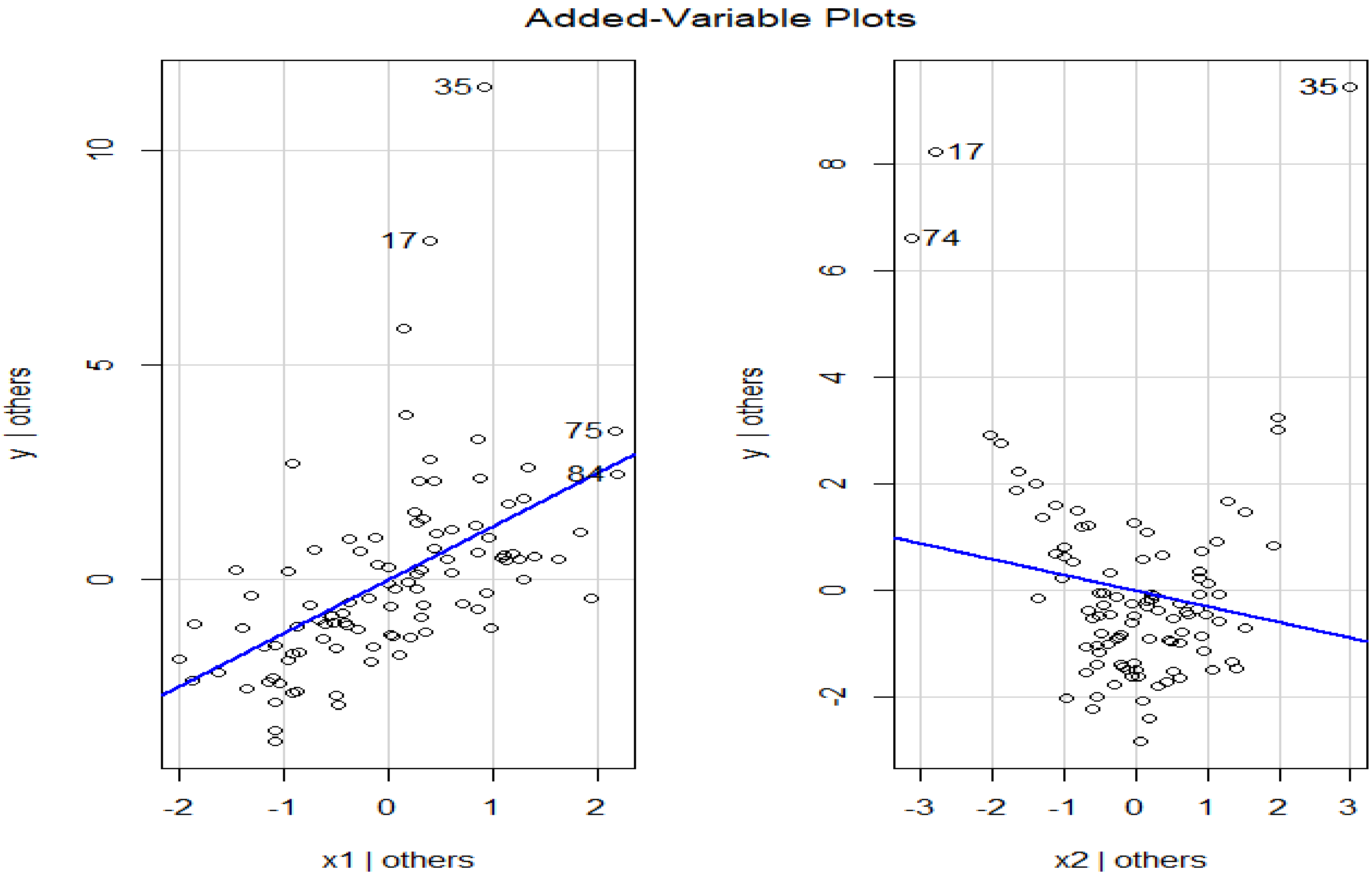
Casewise Diagnostic Plots

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



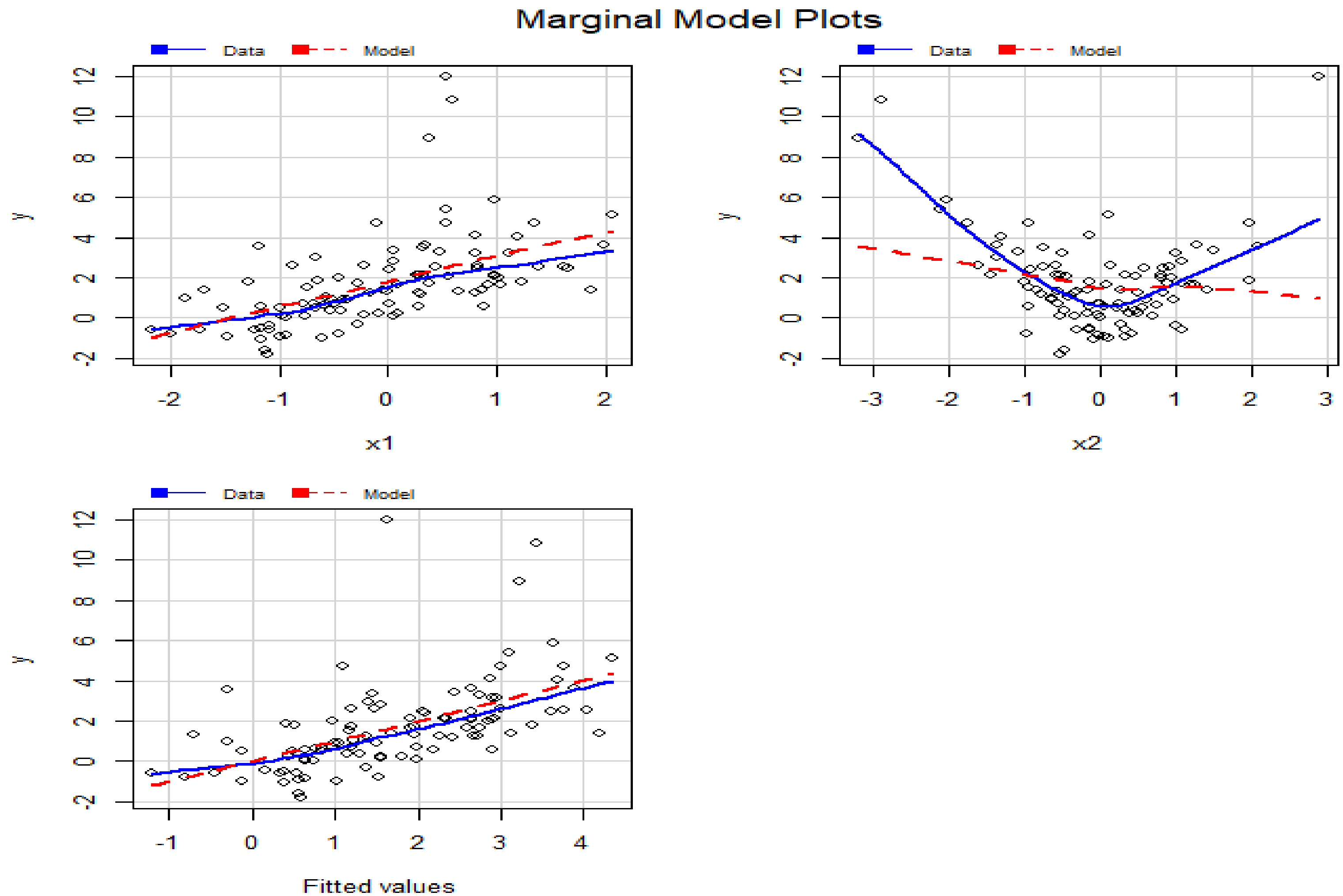
Added Variable Plots

> avPlots(lm.x1px2)



Marginal Model Plots

> mmps(lm.x1px2)

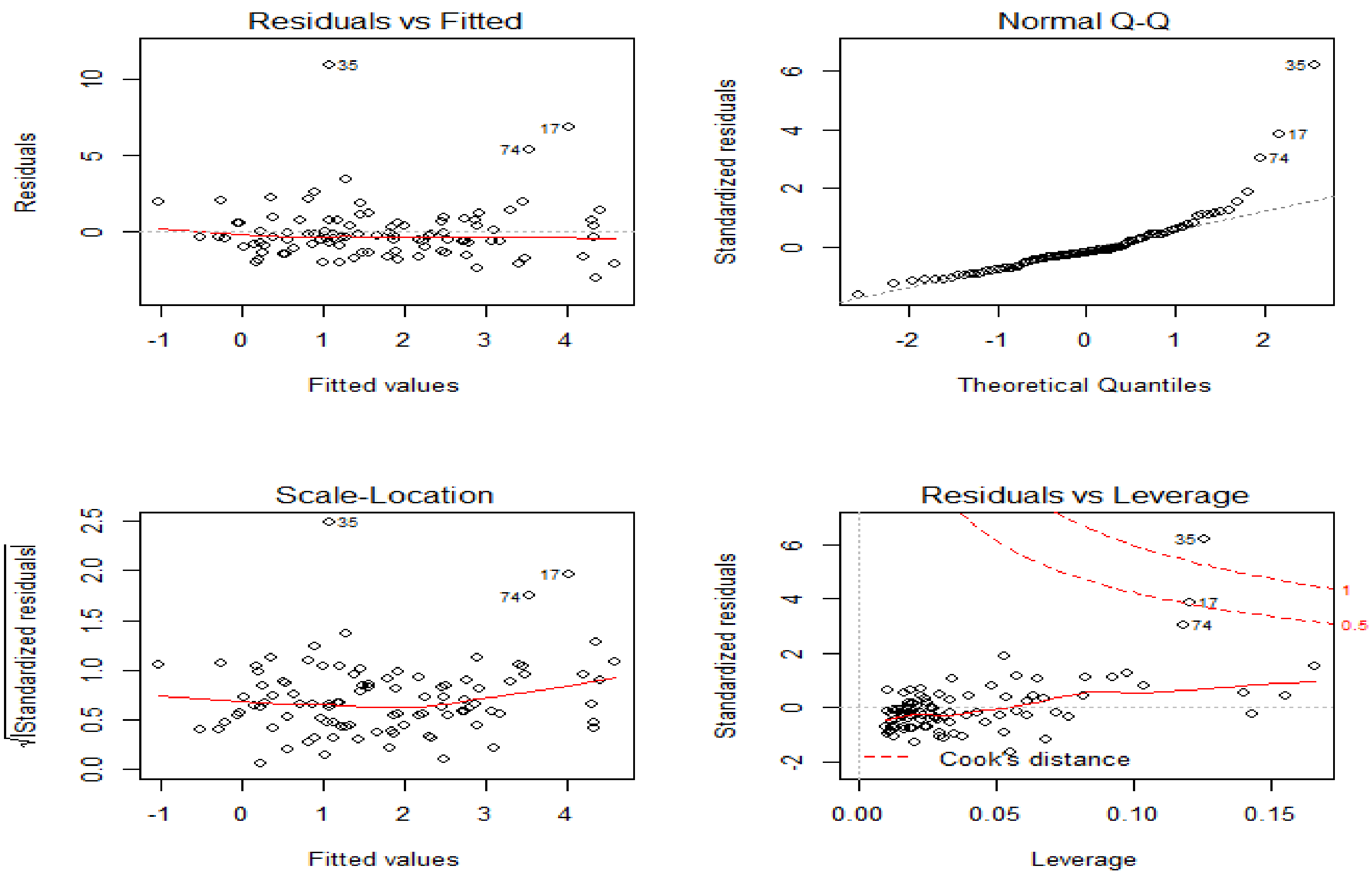


What if we think an interaction will fix it?

```
> 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
```

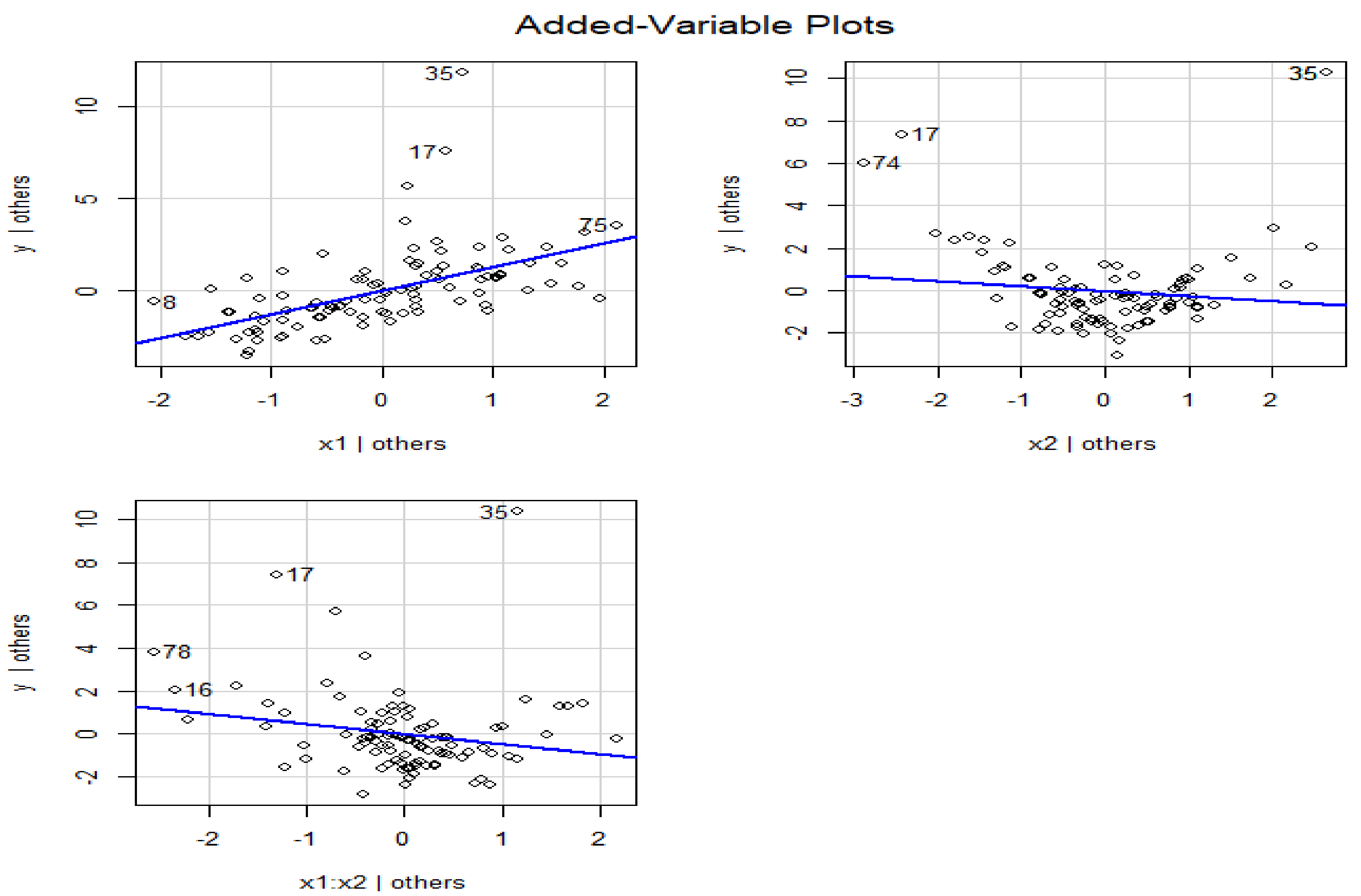
Casewise Diagnostic Plots

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



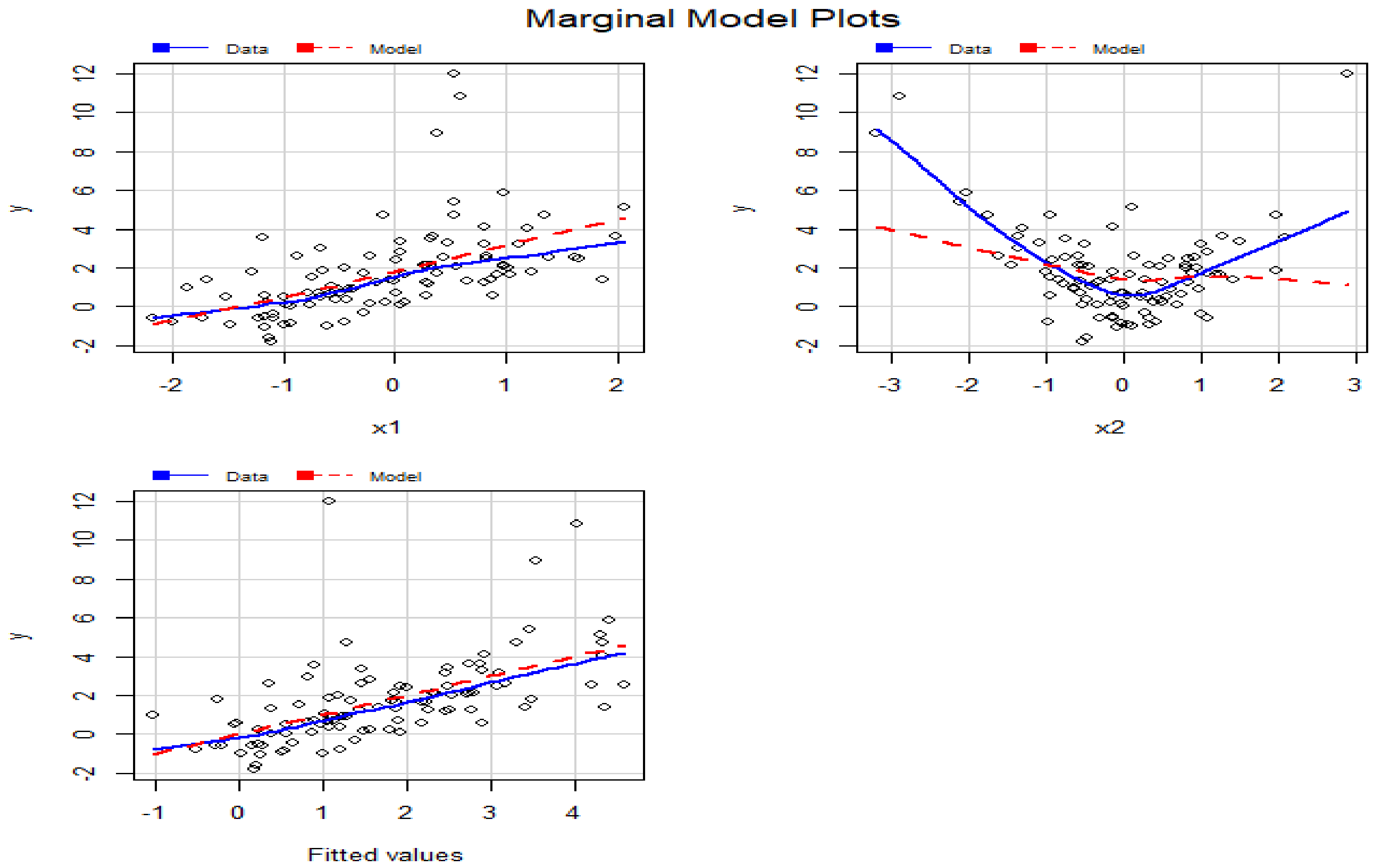
Added Variable Plots

> avPlots(lm.x1mx2)



Marginal Model Plots

> mmps(lm.x1mx2)

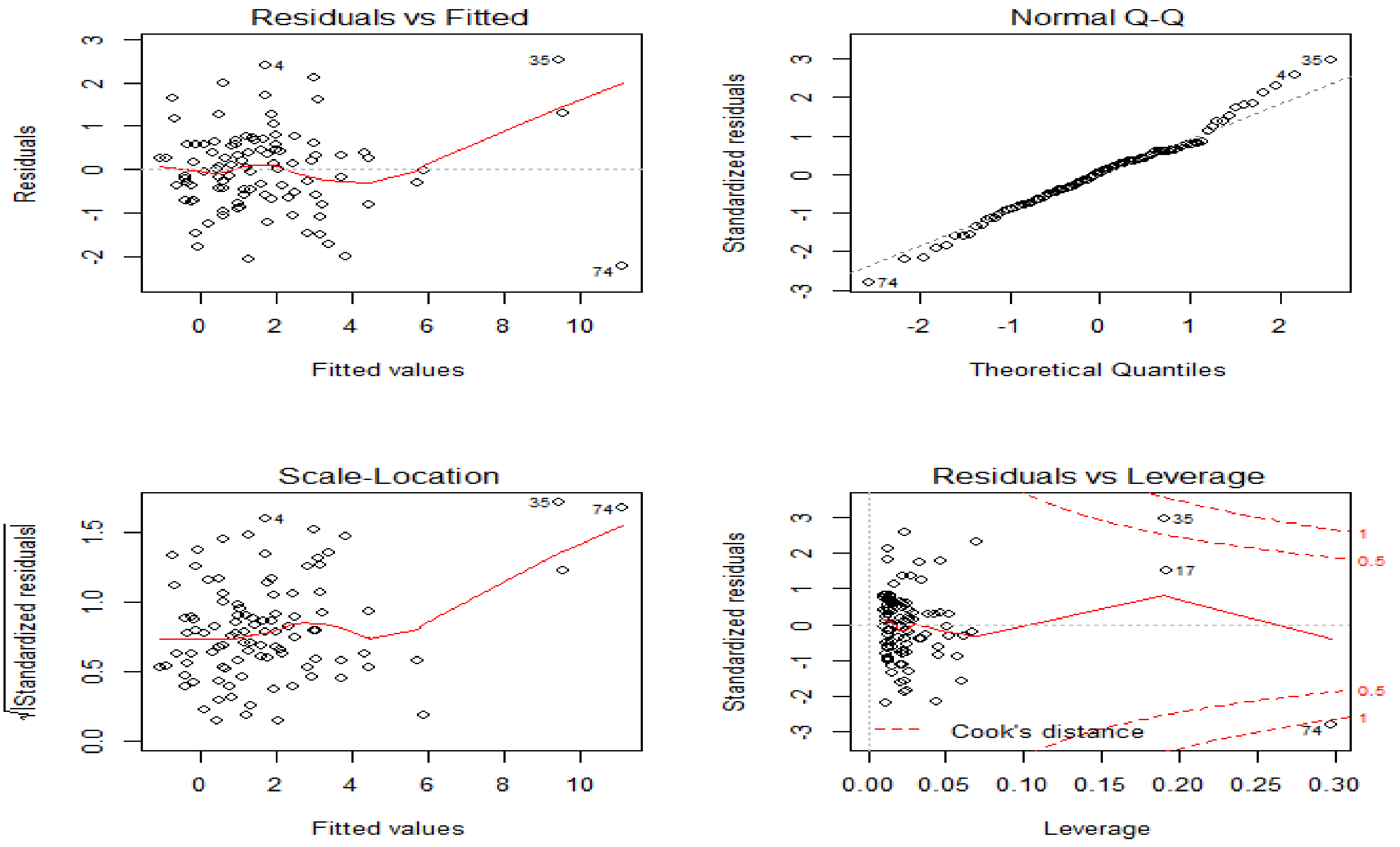


And now the correct model (with x2 squared term)..

```
> 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
```

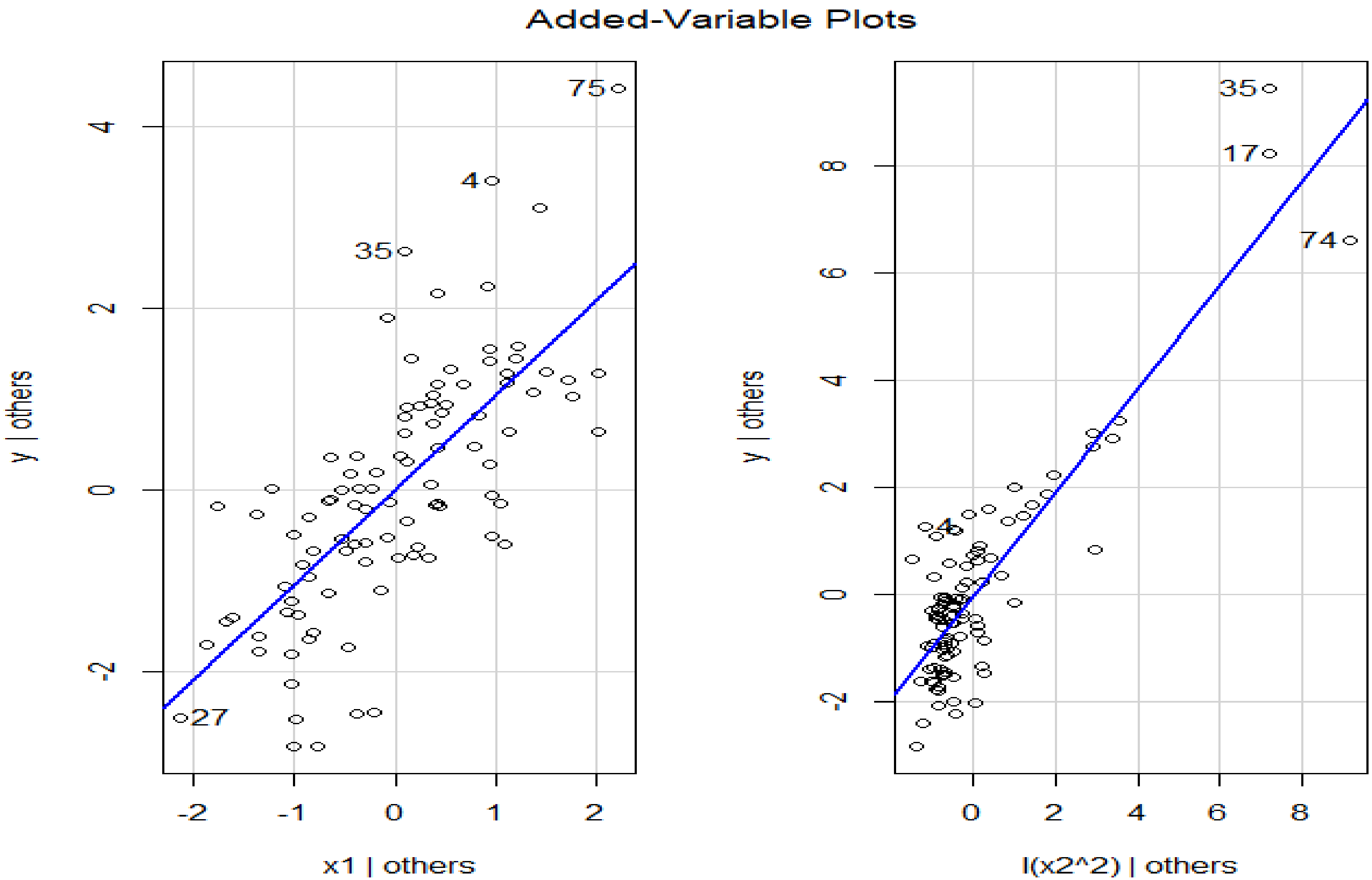
Casewise Diagnostic Plots

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



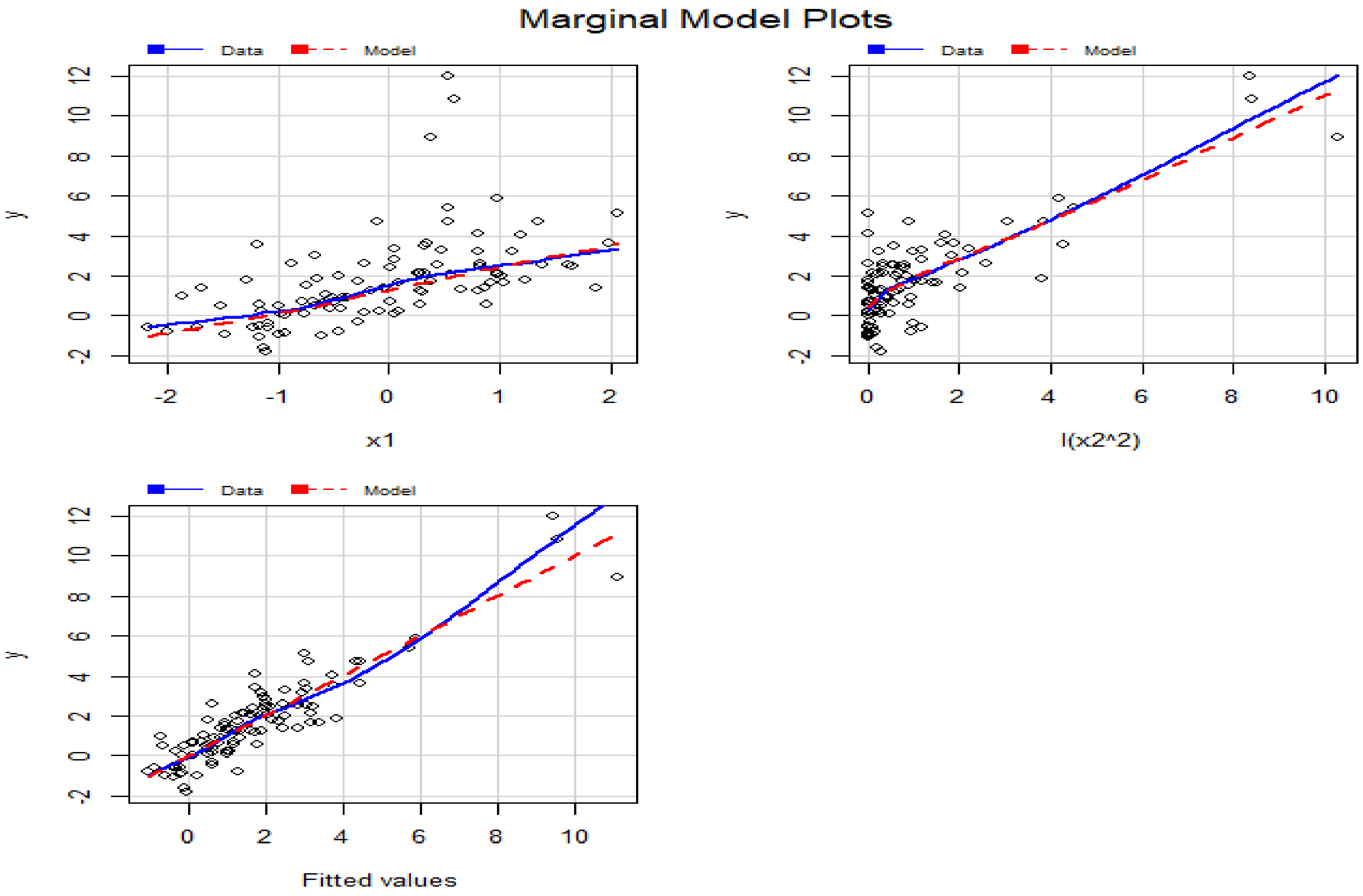
Added Variable Plots

> 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.

Too many predictors: (Multi)Collinearity

- Recall that

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$
$$\mathrm{Var}(\hat{\beta}) = (X^TX)^{-1}\sigma^2$$

- What can cause this to blow up?
  - If (X^TX) is full-rank (rank = dimension of $(X^TX) = p+1$), then $(X^TX)^{-1}$ exists.
  - If (X^TX) has rank less than $p+1$
    - At least one column of X is a linear combination of the others*
    - <u>*Perfect collinearity*</u>

  Then $(X^TX)^{-1}$ doesn't exist

    - Can fix by deleting columns of X until X^TX has full rank again.

*Fact: $rank(X^TX) = rank(X)$

Collinearity …

- If $X^T X$ is full-rank, but there is a column of X that is *nearly* a linear combination of the others…
  - $(X^T X)^{-1}$ will exist but will contain some wild values
  - $\mathrm{Var}\,(\hat{\beta}_j)$ can be wildly inflated
- How could we measure the amount of "almost collinearity"?
  - Regress X_j on the other X's; compute R_j^2 from this regression…
    - $R_j^2 = 1$ for perfect collinearity;
    - $R_j^2 \approx 1$ for near-collinearity.

Using R_j^2 as a Collinearity measure

- $1 - R_j^2$ is called the *tolerance* of $\hat{\beta}_j$ to collinearity.
- One can calculate* that for $y = X\beta + \varepsilon$,

$$\mathrm{Var}\,(\hat{\beta}_j) = \frac{1}{1-R_j^2}\frac{\sigma^2}{SX_jX_j}$$

and we know for simple regression $y = \beta_0 + \beta_j X_j + \varepsilon$,

$$\mathrm{Var}\,(\hat{\beta}_j) = \frac{\sigma^2}{SX_jX_j}$$

- Thus, $1/(1-R_j^2)$ is the ratio of $\mathrm{Var}\,(\hat{\beta}_j)$ under the full model to $\mathrm{Var}\,(\hat{\beta}_j)$ under simple regression on X_j alone.
- $VIF_j = \frac{1}{1-R_j^2}$ is the *variance inflation factor!*

* E.g. O'Brien, R. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity, 41,* 673—690.

Using VIF…

- No "significance tests" for VIF_j, since $X_1\ X_2 \ldots X_p$ are usually considered nonrandom.
- Some common rule-of-thumb cutoffs are
  - $VIF_j > 4$ or 5 (VIF=4 → $SE_{multiple} = 2 \times SE_{simple}$)
  - $VIF_j > 10$ (VIF=9 → $SE_{multiple} = 3 \times SE_{simple}$)
- What to do when VIF_j is "large"?
  - Eliminate columns of *X* until the *VIF*'s settle down?
  - Combine highly correlated columns of *X?*
    - *Principal Components?*
  - Try alternative models such as *ridge regression*?
    - Less sensitive to collinearity

Do the usual "fixes" make sense?

- Eliminate columns of *X* until the *VIF*'s settle down?
  - Unlike perfect collinearity, we are throwing away some information – collaborator may not agree!
  - If we do it, which columns to eliminate?
- Combine highly correlated columns of *X*?
  - This is a less obvious form of throwing away data…
- Use alternative models such as *ridge regression?*
  - $\hat{\beta}_j$ and $\mathrm{Var}\left(\hat{\beta}_j\right)$ biased; OLS estimates are not...
  - Is the changed model meaningful to collaborator?
- *What consideration is missing from these "fixes"?*

What inferences are we trying to make?

- *Is the goal accurate prediction?* We may not care if the $\hat{\beta}_j$ individually have high SE's as long as adding *X*'s to the model improves $\hat{y}$.
- *Is the goal selecting a "best model"?*
  - *"Best" does not only mean best statistical measures*
  - We may wish to include high-VIF *X*'s because they comport with substantive theory
- *Is the goal inference on individual β's?* High *VIFs* can be bad. This is often where the "fixes" seem to make some sense…

Aside on “generalized VIF”…

- GVIF is a more general form of VIF that applies to groups of variables and reduces to regular VIF for a single variable.
- The usual cutoffs of 5 and 10 work for a transformation of GVIF,

$$GVIF^{(2/(2*df))}$$

where “df” is the number of free coefficients in the for the group of variables.

  - http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/4-5-Multiple-collinearity.html
  - Fox, John, and Georges Monette. (1992). “Generalized Collinearity Diagnostics.” *Journal of the American Statistical Association* 87 (417): 178–83. http://www.jstor.org/stable/2290467.

Example…

- heights.dta…
- heights - VIFs, avplots, mmplots, etc.r

Too few predictors: Omitted variable bias

- Suppose

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

and $$x_2 = \alpha_0 + \alpha_1 x_1 + \varepsilon'$$

then $$y = (\beta_0 + \beta_2\alpha_0) + (\beta_1 + \beta_2\alpha_1)\, x_1 + (\varepsilon + \beta_2\varepsilon')$$

- So if we fit, $y = \gamma_0 + \gamma_1 x_1 + \varepsilon''$, we get $\gamma_1 = \beta_1 + \beta_2\alpha_1$
  - If $\beta_2 = 0$ or $\alpha_1 = 0$, then $\hat{\gamma}_1$ will be unbiased for β_1
  - If both are nonzero, then $\hat{\gamma}_1$ will be biased for β_1
  - Even if $\beta_1 = 0$, it can appear that y and x are correlated ("*spurious correlation*" – "*lurking variable correlation*")

Example (simulated)

```
> x1 <- rnorm(100)
> x2 <- 3*x1 + rnorm(100)
> y <- 2 + 4*x2 + rnorm(100)
> lm.y <- lm(y ~ x1)
> summary(lm.y)

Call:
lm(formula = y ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max
-13.2235  -2.7728   0.4441   2.3951   9.0288

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.5535     0.4054   6.299 8.55e-09 ***
x1           11.9012     0.3623  32.853  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.043 on 98 degrees of freedom
Multiple R-squared:  0.9168,    Adjusted R-squared:  0.9159
F-statistic:  1079 on 1 and 98 DF,  p-value: < 2.2e-16
```

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \\
  &= 2 + 0x_1 + 4x_2 + \epsilon \\
x_2 &= \alpha_0 + \alpha_1 x_1 + \epsilon' \\
  &= 0 + 3x_1 + \epsilon'
\end{aligned}
$$

In the fitted model, omitting x_2, it appears that the coeffient on x_1 should be around 12. In fact, we know it should be (estimating) zero.

Example (simulated)

```
> lm.y2 <- lm(y ~ x1 + x2)
> summary(lm.y2)

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

Residuals:
     Min       1Q   Median       3Q      Max
-2.61952 -0.71282 -0.04126  0.63074  2.51451

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.960843   0.100888   19.44   <2e-16 ***
x1          0.009569   0.317556    0.03    0.976
x2          3.996395   0.102431   39.02   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9947 on 97 degrees of freedom
Multiple R-squared:  0.995,     Adjusted R-squared:  0.9949
F-statistic:  9678 on 2 and 97 DF,  p-value: < 2.2e-16
```

Of course when we fit the right model, we get reasonable estimates of the β's and of the SE's.

Discussion of removal of one variable from this model inevitably starts with vifs. But,

$$VIF(x_1) = 12.696$$
$$VIF(x_2) = 12.696$$

In this case, both x's have vifs of 12.70!

Summary

- Graphical tools for Transformations (catching up!)
  - Added Variable Plots
  - Marginal Model Plots
  - Moral of the Story
- Over- and under-specifying a model
  - Too many predictors: Excess SE's and Collinearity
  - Too few predictors: Omitted Variable Bias