Case Study 8: Diagnosing problems

Richard Wilkinson

Lets look at some data from a Canadian survey on wages in Ontario. The data contains information on

- wages: hourly rate
- education: number of years of schooling
- age in years
- sex
- language spoken

of 7425 individuals.

```
library(car)
data(SLID)
str(SLID)
```

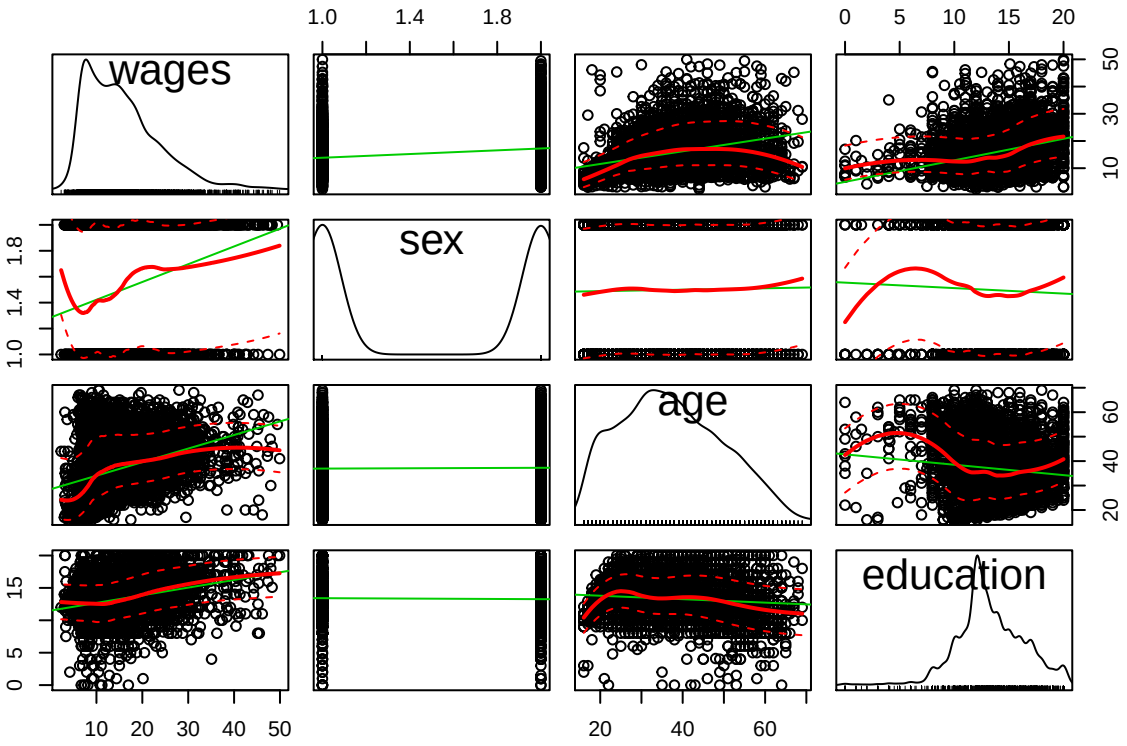
```
## 'data.frame':    7425 obs. of  5 variables:
##  $ wages    : num  10.6 11 NA 17.8 NA ...
##  $ education: num  15 13.2 16 14 8 16 12 14.5 15 10 ...
##  $ age      : int  40 19 49 46 71 50 70 42 31 56 ...
##  $ sex      : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 1 1 1 2 1 ...
##  $ language : Factor w/ 3 levels "English","French",..: 1 1 3 3 1 1 1 1 1 1 ...
```

```
SLID[1:10,]
```

```
##    wages education age    sex language
## 1  10.56      15.0  40   Male  English
## 2  11.00      13.2  19   Male  English
## 3     NA      16.0  49   Male    Other
## 4  17.76      14.0  46   Male    Other
## 5     NA       8.0  71   Male  English
## 6  14.00      16.0  50 Female  English
## 7     NA      12.0  70 Female  English
## 8     NA      14.5  42 Female  English
## 9   8.20      15.0  31   Male  English
## 10    NA      10.0  56 Female  English
```

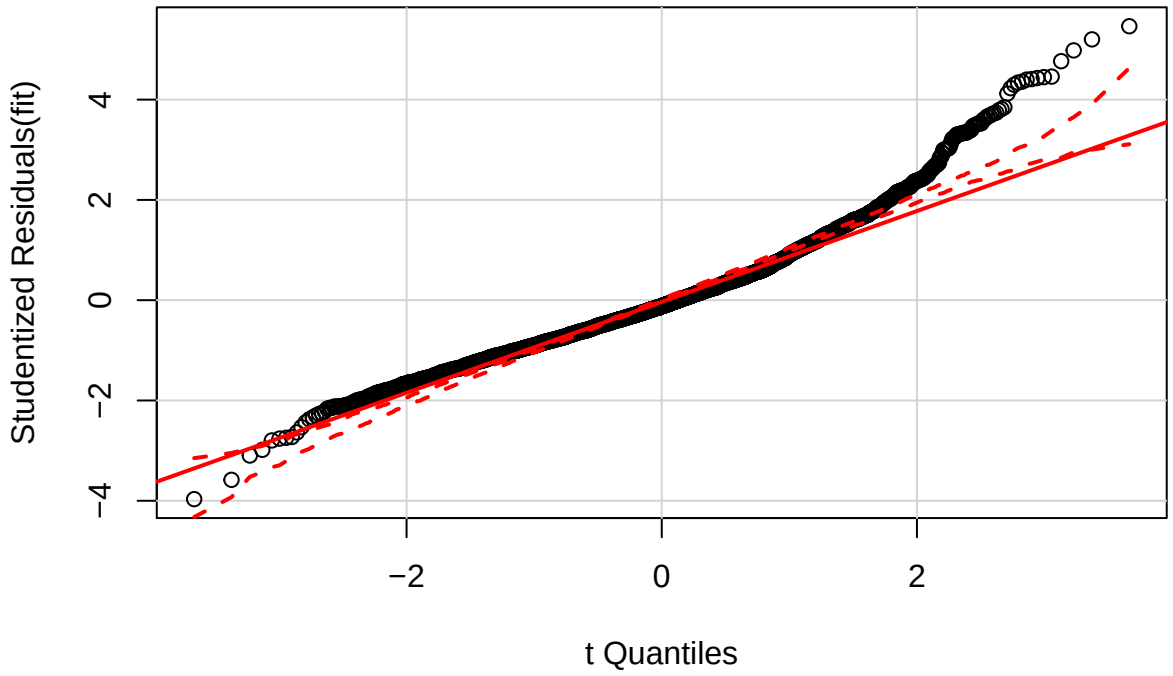
There are a number of missing data points. Lets remove these from the dataset and only work with individuals for whom we have complete information.

```
library(dplyr)
WageData <- filter(SLID, !is.na(wages), !is.na(education), !is.na(language))
scatterplotMatrix(WageData[,c('wages','sex', 'age', 'education')])
```

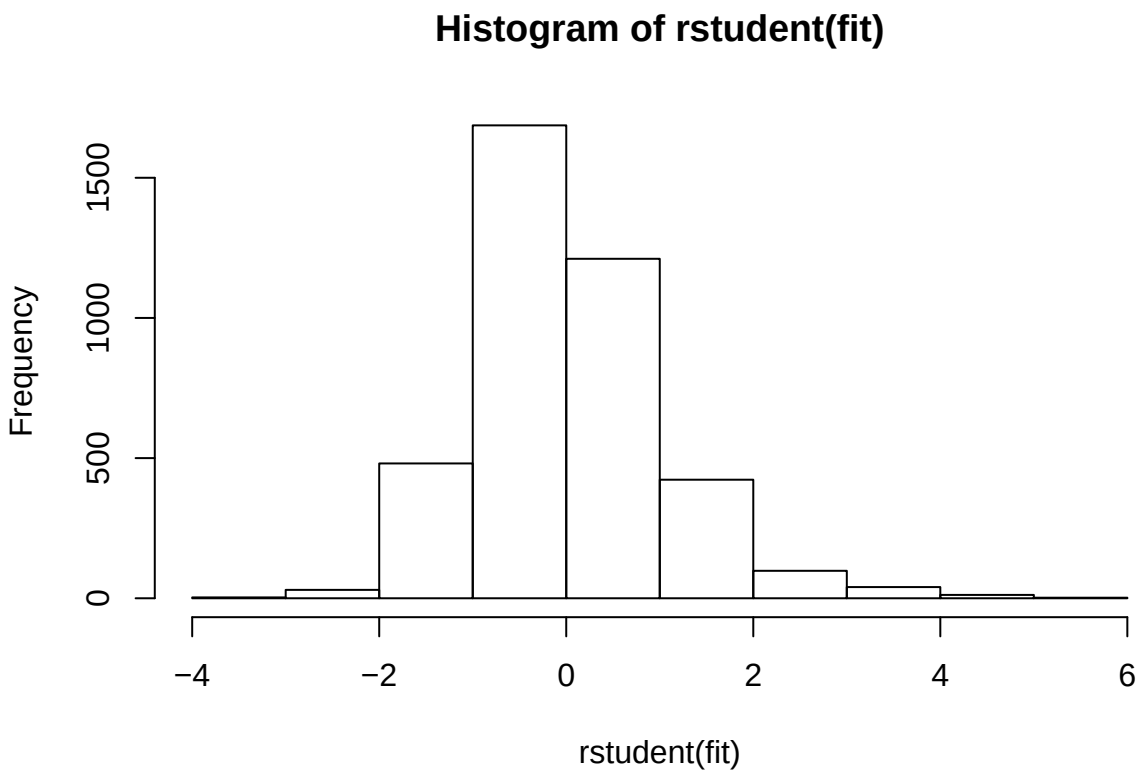


Lets start by fitting a basic model and examining some diagnostic plots.

```
fit <- lm(wages~ sex + age + education, data = WageData)
qqPlot(fit)
```

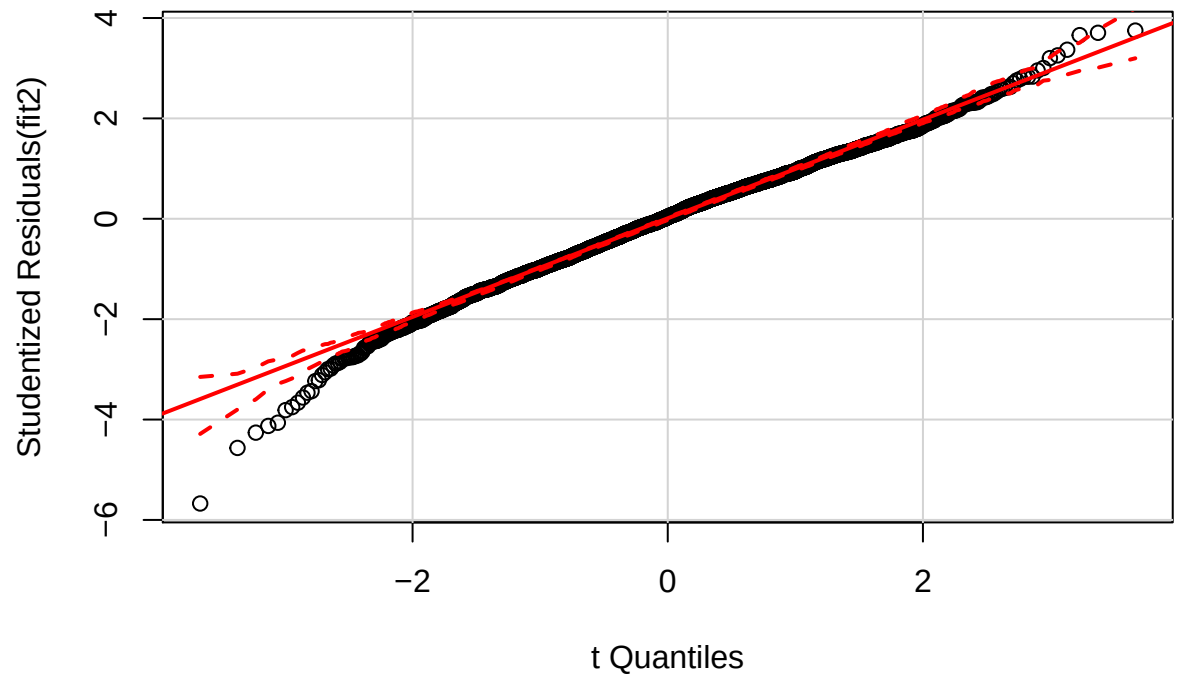



```
hist(rstudent(fit))
```



The QQ-plot and the histogram of the residuals suggests a positive skew (as did the original histogram of the wage data in the scatterplot matrix). We can reduce positive skew by moving the response variable down the ladder of powers.

```
fit2 <- lm(log(wages) ~ sex + age + education, data = WageData)
qqPlot(fit2)
```



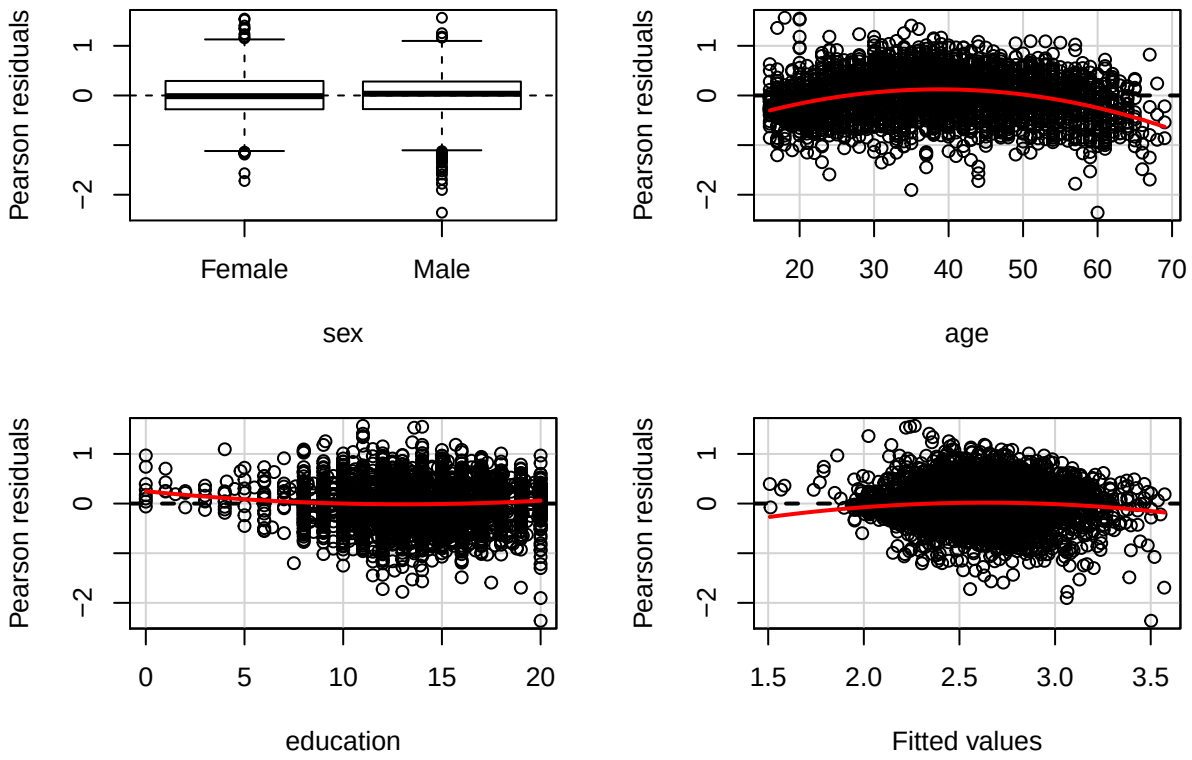
You could try other powers such as sqrt(wage), but log(wage) seems to have produced reasonable results.

Note that you can highlight the 5 most outlying data points in QQ-plots by using the command

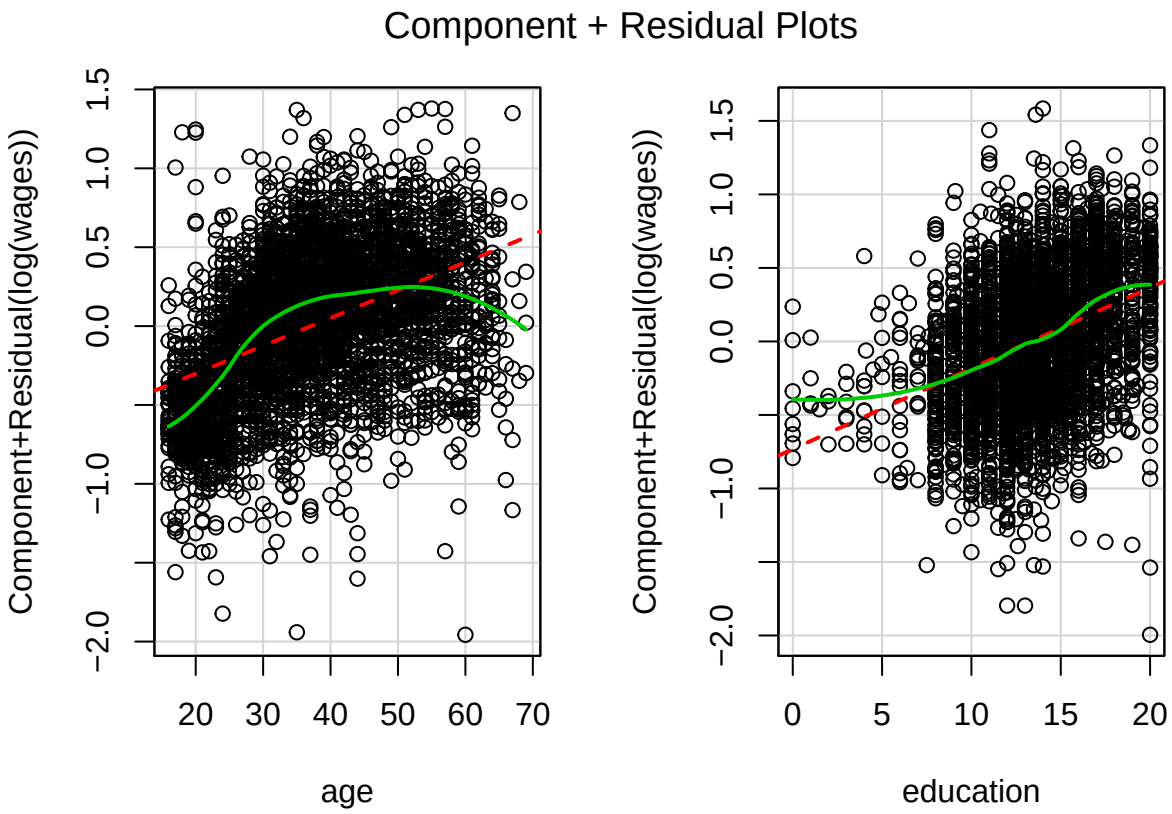
```
qqPlot(fit2, id.n=5)
```

Lets now check the residual plots:

```
residualPlots(fit2, tests=FALSE)
```




```
crPlots(fit2, terms=~age+education)
```



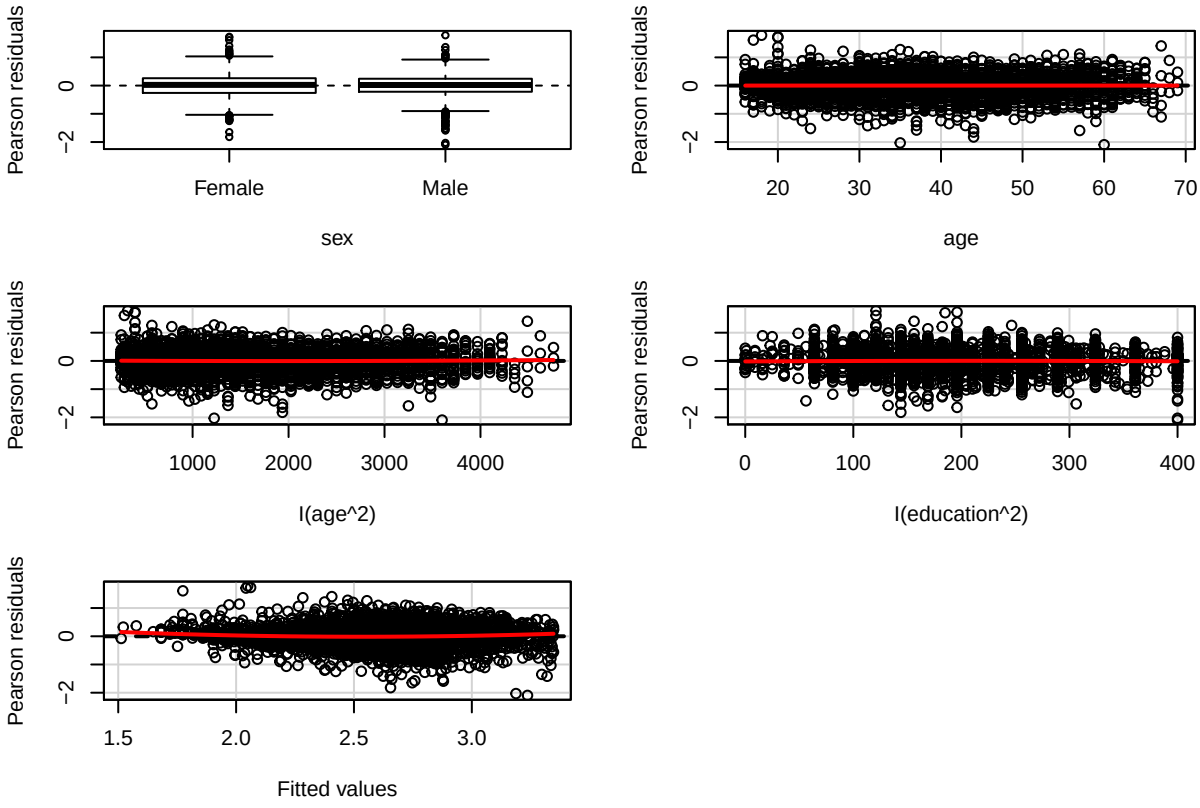
The command

```
crPlots(fit2)
```

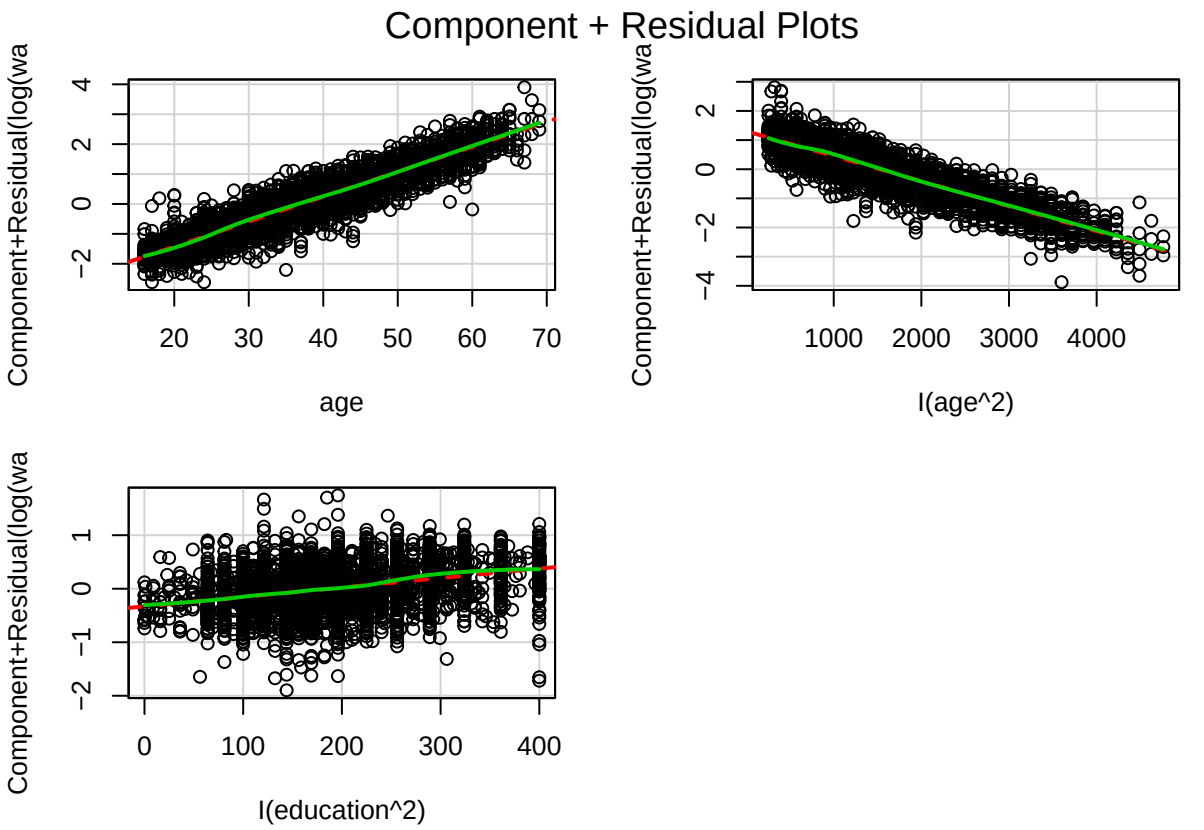
will also work, but gives a third plot for the categorical variable sex which is not useful. So I've removed it by specifying the terms to plot with the command crPlots(fit2, terms=~age+education).

This shows some non-linearity, particularly in how the response depends upon age. We could try transforming age down the ladder of powers, but the relationship looks like it could be non-monotone, in which case we will need to fit a quadratic model. Similarly, the dependence upon education looks non-linear. We can try transforming this up the ladder of powers (see Tukey and Mosteller's bulging rule).

```
fit3 <- lm(log(wages) ~ sex + age + I(age^2) + I(education^2), WageData)
residualPlots(fit3, tests=FALSE)
```




```
crPlots(fit3, terms=~.-sex)
```



These now looks pretty much perfect.