ASM - Module 4 Solutions

Michal Siwek

Monday, November 16, 2015

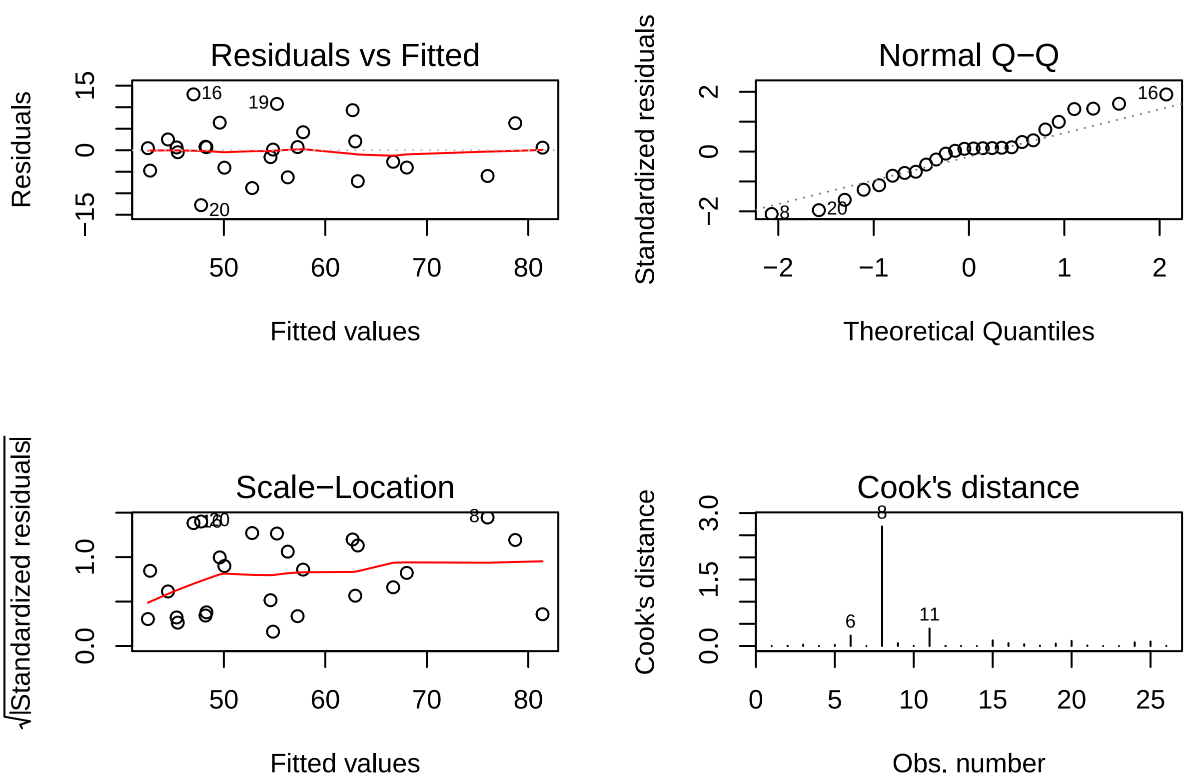
Task 3

Collecting data and fitting the model:

```
file <- "realest.txt"
data <- read.table(file, header = T)
fit <- lm(Price ~ ., data = data)
```

a) Diagnostic plots

```
par(mfrow = c(2, 2))
plot(fit, which = 1:4)
```



Residual plot shows heteroscedasticity - the variance is higher for lower fitted values, and decreases as the fitted values increase. **Normal QQ plot** indicates the distribution of residuals has lighter tails than the normal distribution. **Cook's distance plot** indicates some outlying influential observations, especially observation 8, but also 11 and 6.

b) Outliers

Based on the heuristic rule we conclude that there are three outliers:

```
res <- rstudent(fit)
res[abs(res) > 2]
```

```
##         8        16        20
## -2.352002  2.092962 -2.163371
```

c) Influential observations

Observations having highest Cook's distance:

```
cook <- cooks.distance(fit)
cook[order(cook, decreasing = T)][1:3]
```

```
##         8        11         6
## 2.6976331 0.3952608 0.2365181
```

Influential observation according to the heuristinc rule for the values of leverage:

```
hat <- hatvalues(fit)
threshold <- 2 * sum(hat) / nrow(data)
hat[hat > threshold]
```

```
##         8
## 0.8475338
```

Task 5

Collecting data:

```
file <- "strongx.txt"
data <- read.table(file, header = T)
```

a) LS fit

```
fitLS <- lm(crossx ~ energy, data = data)
```

b) WLS fit

```
fitWLS <- lm(crossx ~ energy, weights = sd^-2, data = data)
```

c) Models comparison

Models output:

```r
summary(fitLS)
```

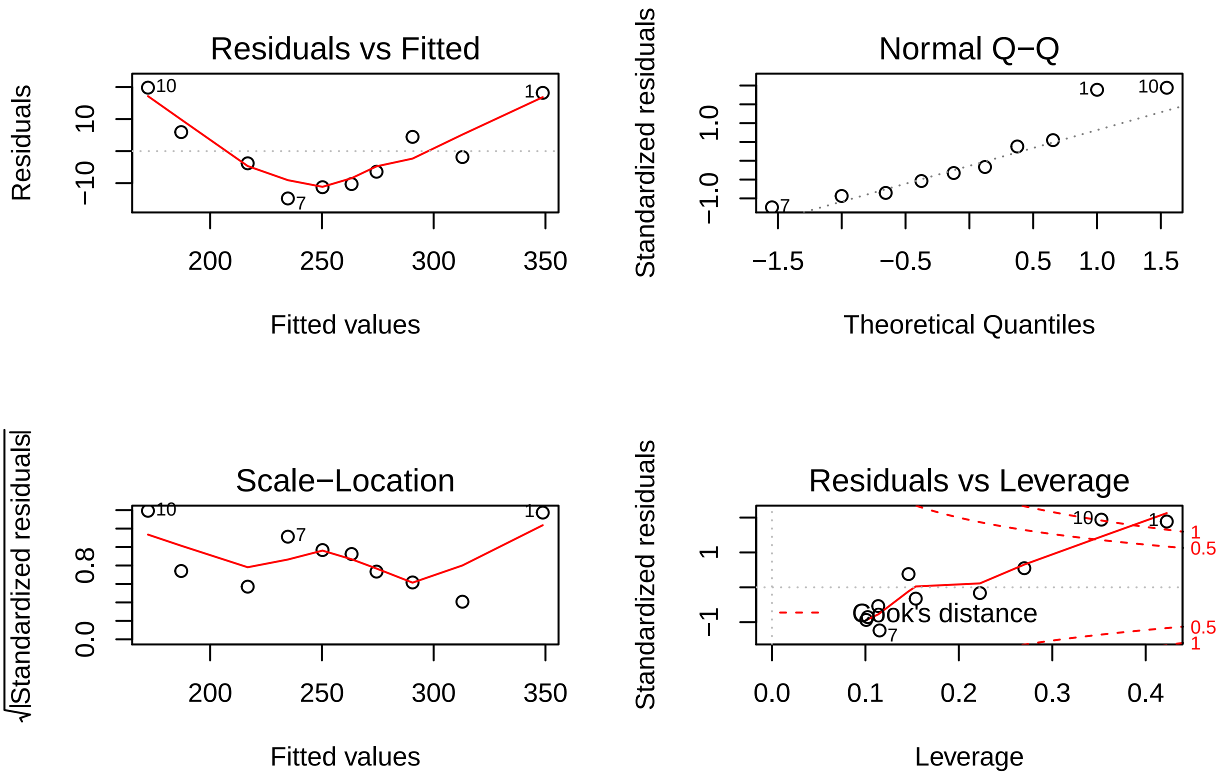
```
##
## Call:
## lm(formula = crossx ~ energy, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.773  -9.319  -2.829   5.571  19.817
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   135.00      10.08    13.4 9.21e-07 ***
## energy        619.71      47.68    13.0 1.16e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.69 on 8 degrees of freedom
## Multiple R-squared:  0.9548, Adjusted R-squared:  0.9491
## F-statistic: 168.9 on 1 and 8 DF,  p-value: 1.165e-06
```

```r
summary(fitWLS)
```

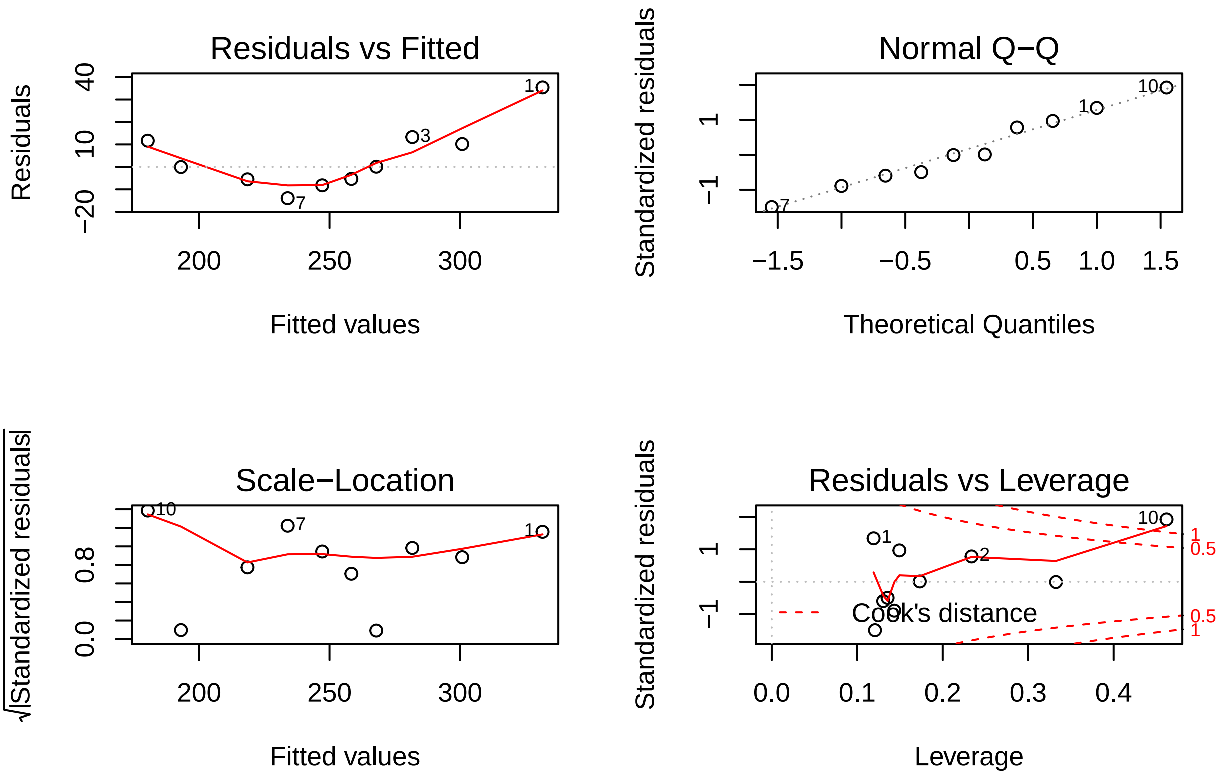
```
##
## Call:
## lm(formula = crossx ~ energy, data = data, weights = sd^-2)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3230 -0.8842  0.0000  1.3900  2.3353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  148.473      8.079   18.38 7.91e-08 ***
## energy       530.835     47.550   11.16 3.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.657 on 8 degrees of freedom
## Multiple R-squared:  0.9397, Adjusted R-squared:  0.9321
## F-statistic: 124.6 on 1 and 8 DF,  p-value: 3.71e-06
```

Diagnostic plots:

```r
par(mfrow = c(2, 2))
plot(fitLS)
```

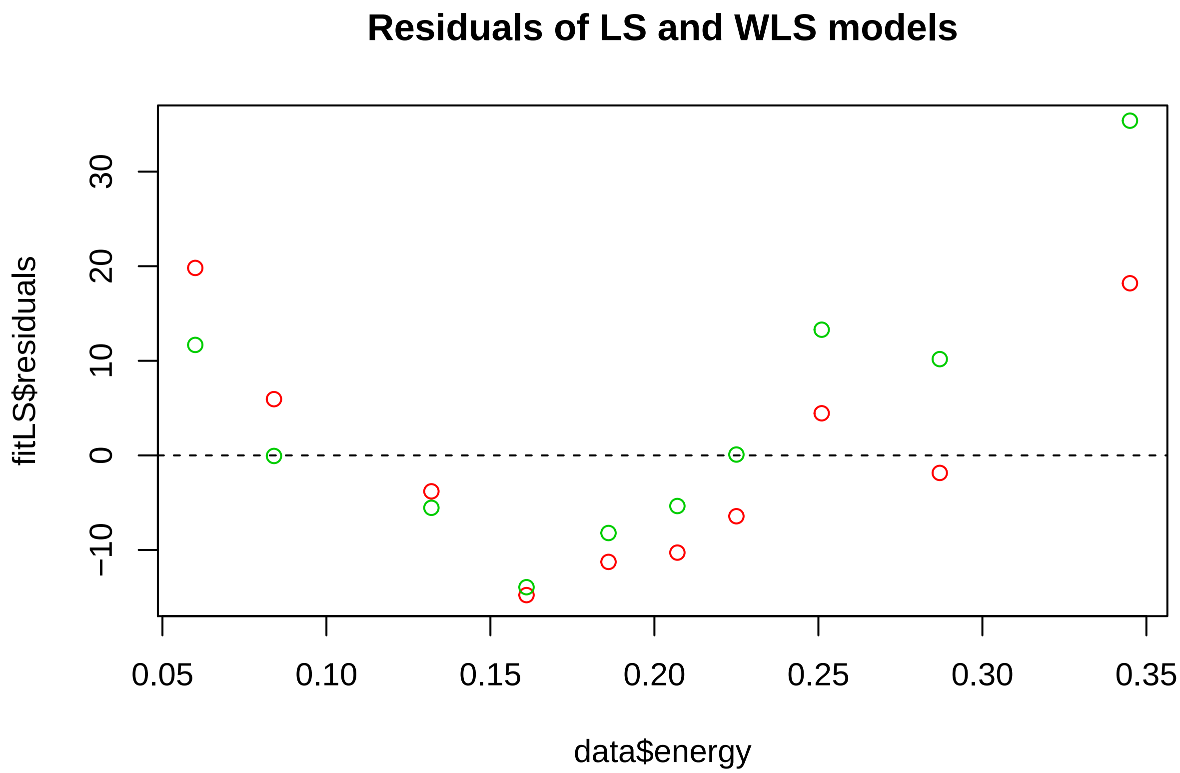



```r
plot(fitWLS)
```



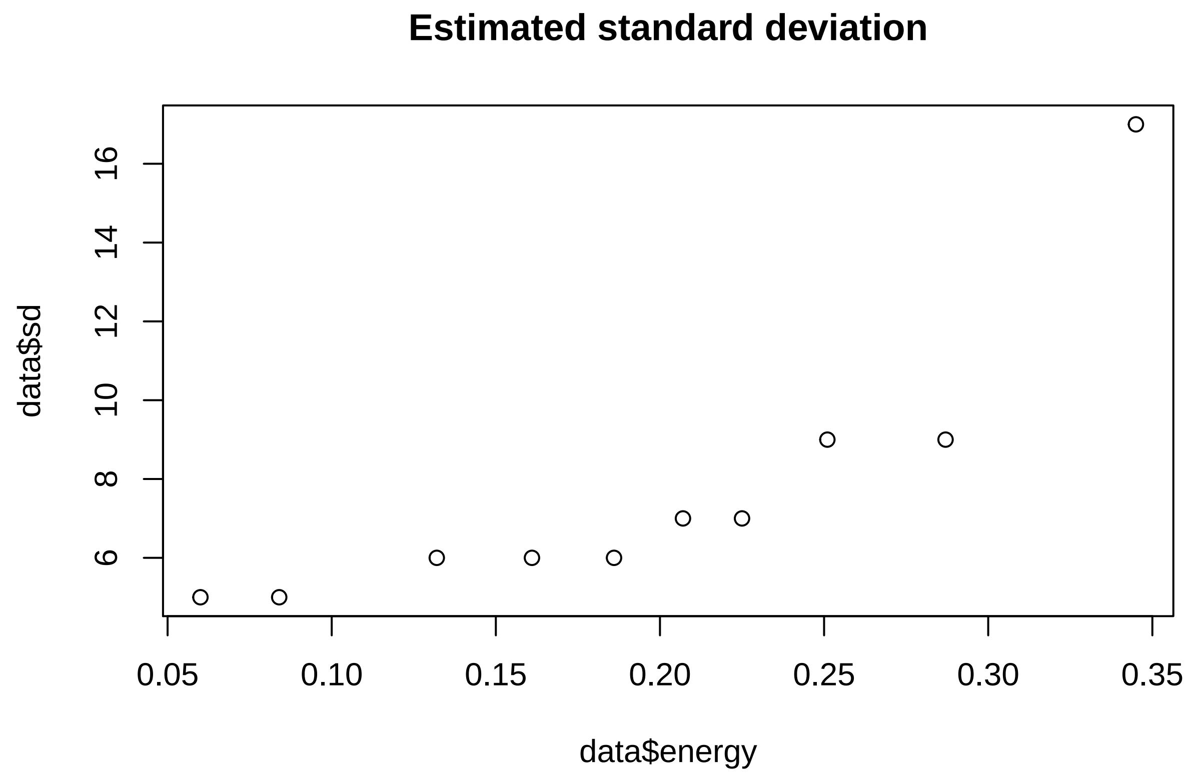
We can see that the **WLS model (green in the plot below)** has less scattered residuals for small `energy` values (these are actualy small values of the inverse of the energy) than the **LS model (red)**.

```
plot(data$energy, fitLS$residuals, ylim = c(-15, 35), col = 2,
     main = "Residuals of LS and WLS models")
points(data$energy, fitWLS$residuals, col = 3)
abline(h = 0, lty = 2)
```



We can explain this by pointing to the fact that observations with high `energy` have also high values of estimated standard deviation (see the plot below). It means that these observations are multiplied by relativaly smaller weights and *so the observations with low* `energy` *values have stronger impact on the regression line leading to a better fit for them.*

```
plot(data$energy, data$sd,
     main = "Estimated standard deviation")
```



d) Modification - WLS2 model

The residual plot shows nonlinearity: a convex depenence of actual `crossx` values not caught by the regression on `energy`. Thus we could **include square values of `energy`** to account for this dependence:

```
fitWLS2 <- lm(crossx ~ energy + I(energy^2), weights = sd^-2, data = data)
summary(fitWLS2)
```

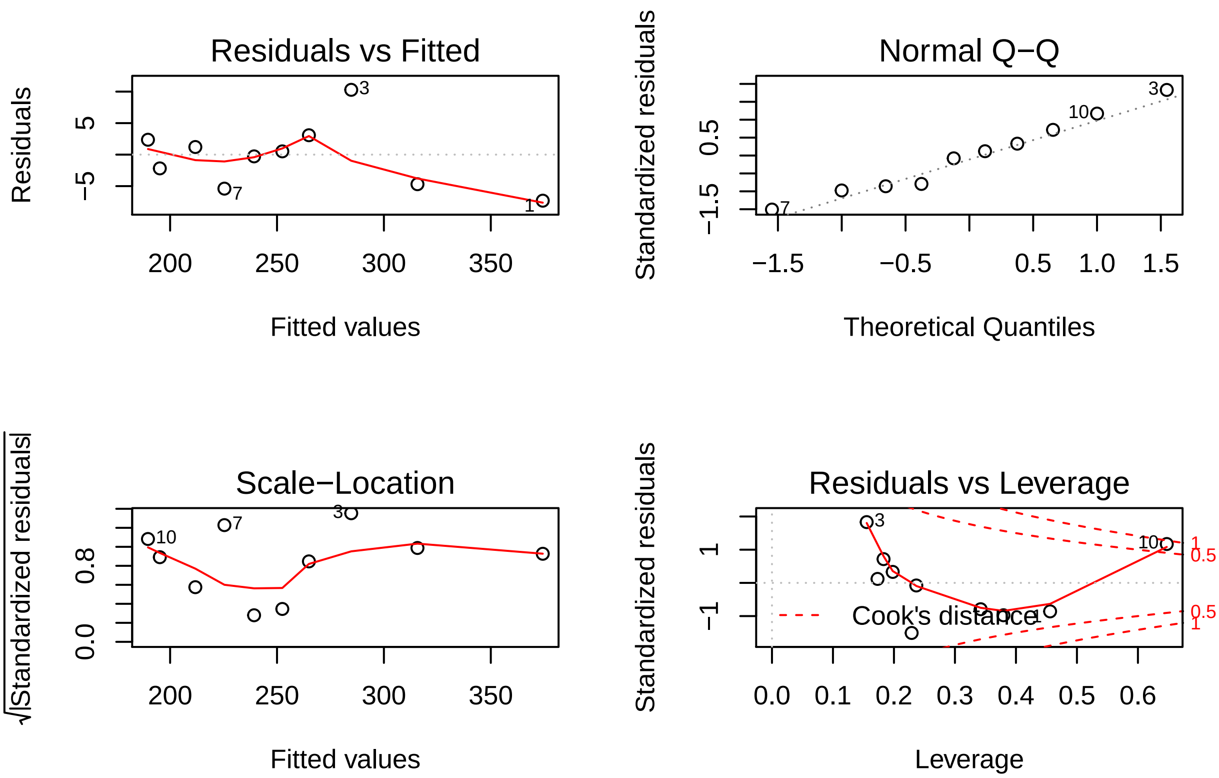
```
##
## Call:
## lm(formula = crossx ~ energy + I(energy^2), data = data, weights = sd^-2)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89928 -0.43508  0.01374  0.37999  1.14238
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  183.8305     6.4591  28.461  1.7e-08 ***
## energy         0.9709    85.3688   0.011 0.991243
## I(energy^2) 1597.5047   250.5869   6.375 0.000376 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6788 on 7 degrees of freedom
```

```
## Multiple R-squared:  0.9911, Adjusted R-squared:  0.9886
## F-statistic: 391.4 on 2 and 7 DF,  p-value: 6.554e-08
```

We see the squared term is significant, the adjusted R^2 is 99%, so much better than the previous WLS model (93%).

The diagnostic plots:

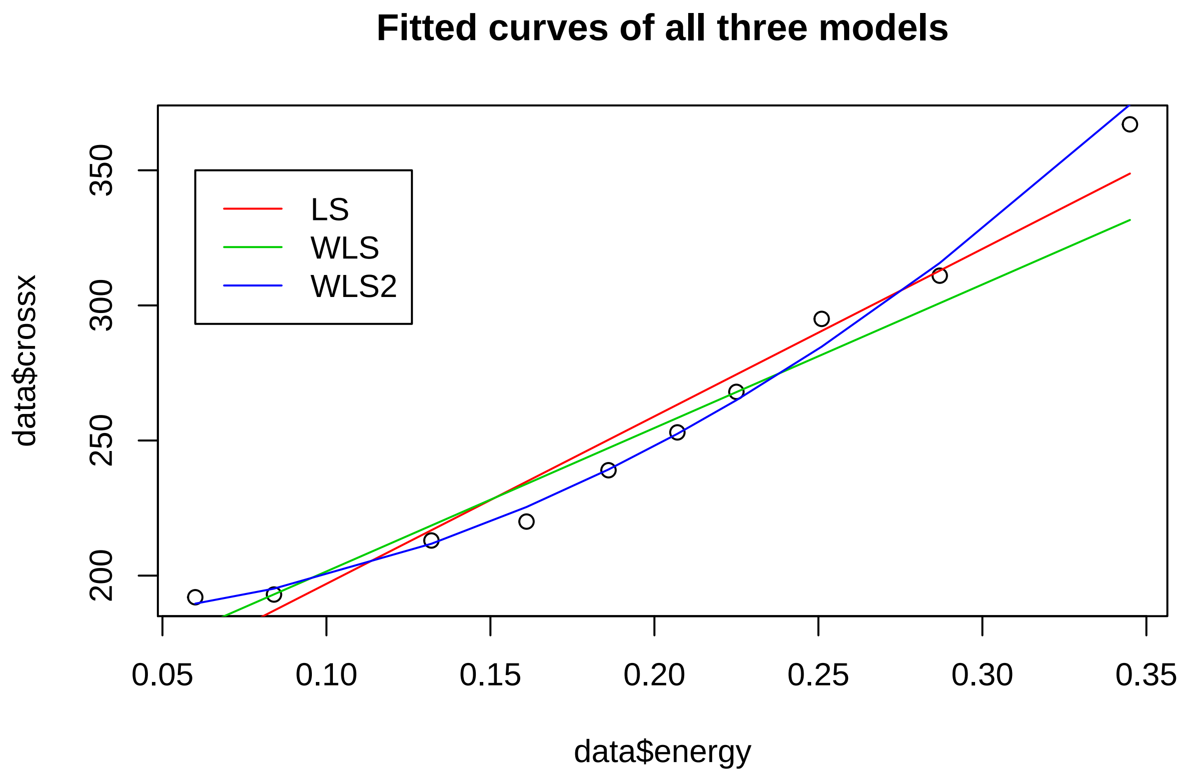
```r
par(mfrow = c(2, 2))
plot(fitWLS2)
```



The plots show that there are still many problems with the fit.

e) Drawing fitted curves

```r
par(mfrow = c (1, 1))
plot(data$energy, data$crossx,
     main = "Fitted curves of all three models")
lines(data$energy, fitLS$fitted.values, col = 2)
lines(data$energy, fitWLS$fitted.values, col = 3)
lines(data$energy, fitWLS2$fitted.values, col = 4)
legend(x = 0.06, y = 350,
       legend=c("LS", "WLS", "WLS2"),
       lty = 1, col = 2:4)
```


Fitted curves of all three models
LS
WLS
WLS2
data$crossx
200
250
300
350
0.05
0.10
0.15
0.20
0.25
0.30
0.35
data$energy