Linear Model

Josep Franquet Fàbregas

26/2/2024

Linear Model

Simulation of data to model

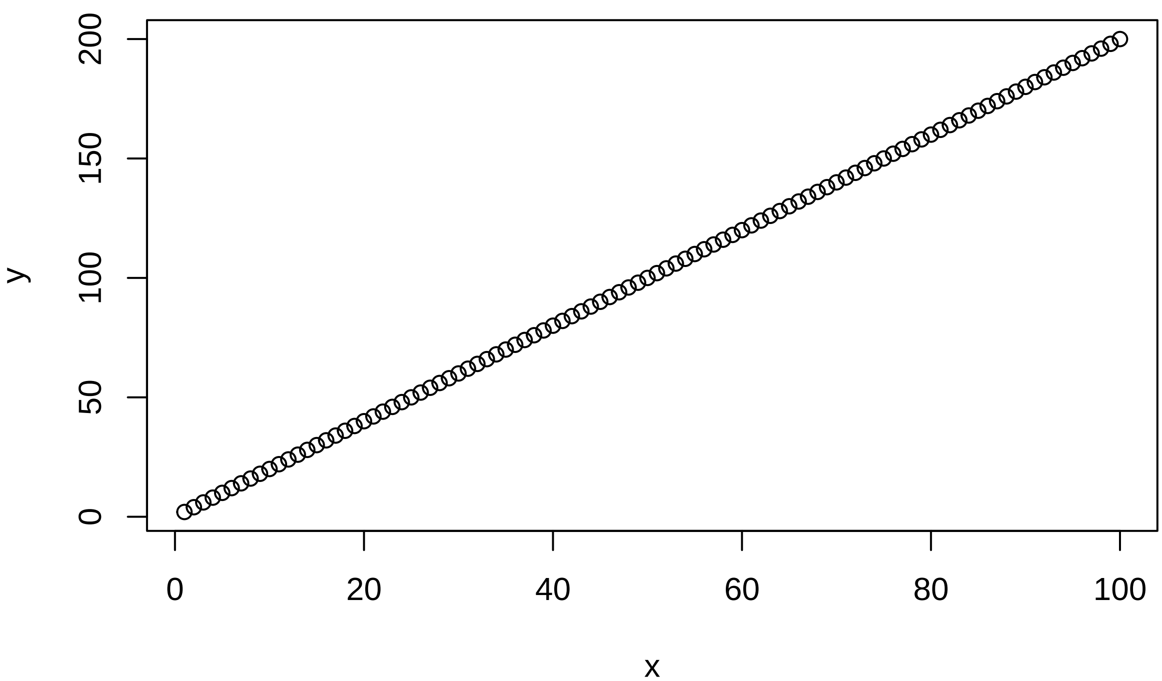
```
library(car)
```

```
## S'està carregant el paquet requerit: carData
```

To simulate data to be modelled with a linear model under the normal assumptions, we create a vectors related by a linear equation and we add some normally distributed noise.

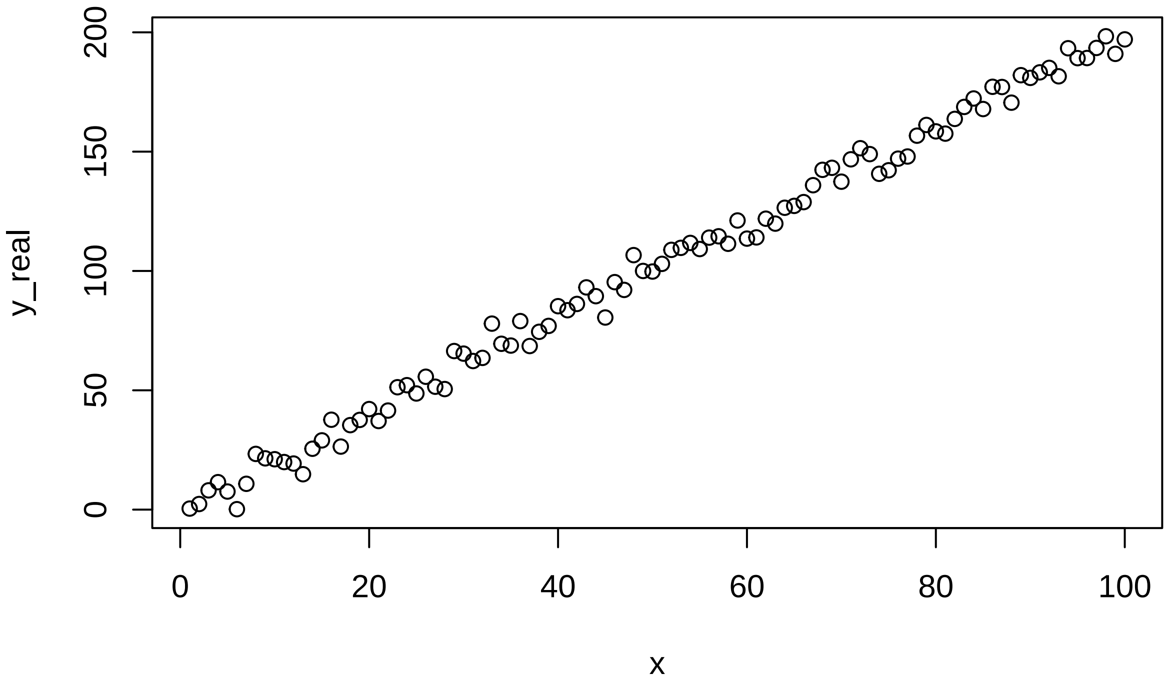
First, we simulate $y = 2x + \varepsilon$:

```
# Explanatory variable
x <- 1:100
# Target variable
y <- 2*x

plot(x, y)
```




```
# Normal distributed error
e <- rnorm(100, 0, 5)
```

```r
# Real data from a sample
y_real <- y + e

plot(x, y_real)
```



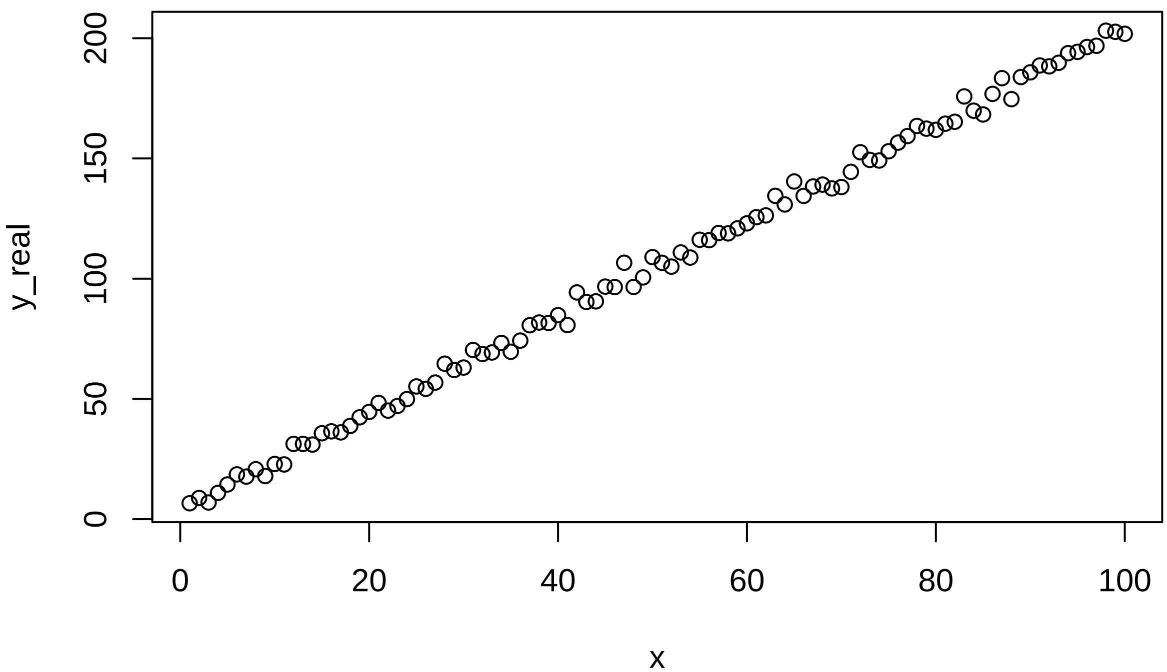

```r
df <- data.frame(x, y_real)

model_1 <- lm(y_real~x, data=df)

summary(model_1)
```

```
##
## Call:
## lm(formula = y_real ~ x, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8752  -2.6381  -0.2818   3.3213  11.8949
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.07454    0.93878   0.079    0.937
## x            1.99914    0.01614 123.868   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.659 on 98 degrees of freedom
## Multiple R-squared:  0.9937, Adjusted R-squared:  0.9936
## F-statistic: 1.534e+04 on 1 and 98 DF,  p-value: < 2.2e-16
```

Can you relate all the parameters with the R output? And calculate confidence intervals for the parameter estimations? Which are significative and why?

Let's add an intercept and a smaller variance for the error term.

```r
# Explanatory variable
x <- 1:100
# Target variable
y <- 4 + 2*x

# Normal distributed error
e <- rnorm(100, 0, 3)

# Real data from a sample
y_real <- y + e

plot(x, y_real)
```




```r
df <- data.frame(x, y_real)

model_2 <- lm(y_real~x, data=df)

summary(model_2)
```

```
##
## Call:
## lm(formula = y_real ~ x, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0005 -1.6873  0.0734  1.3094  8.5855
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.884268   0.556174   6.984 3.47e-10 ***
## x           2.003180   0.009562 209.504  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.76 on 98 degrees of freedom
## Multiple R-squared:  0.9978, Adjusted R-squared:  0.9977
## F-statistic: 4.389e+04 on 1 and 98 DF,  p-value: < 2.2e-16
```

Is the intercept now significative? How are the confidence intervals now for the parameter estimators, did they grow with the error term?

Remember that linear models are linear on the parameters, not on the variables. For example:

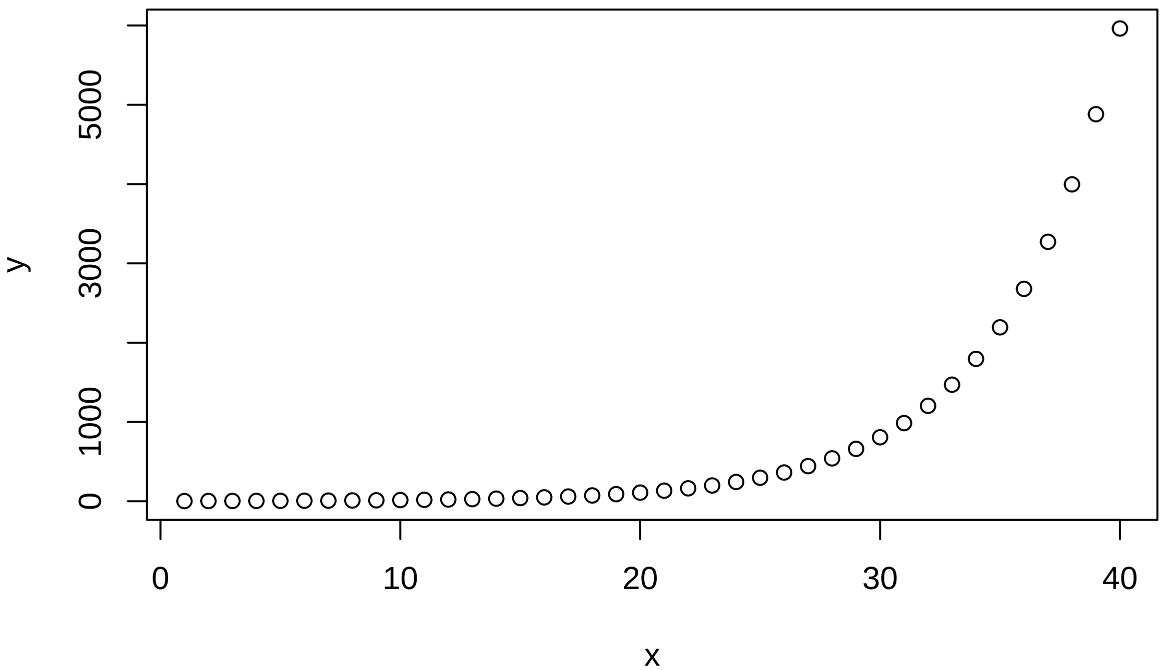
$$y = 2e^{0.2x} + \varepsilon$$

If we apply logarithms:

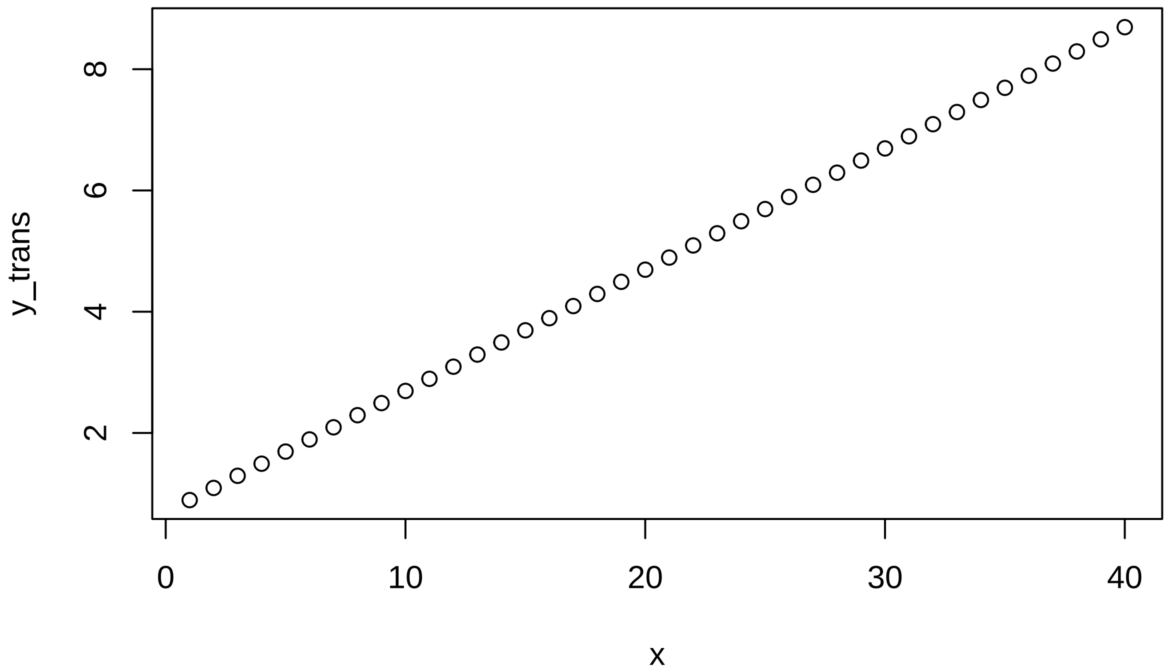
$$\log(y) = \log(2) + 0.2x + \epsilon$$

Let's simulate this data and then add an error term:

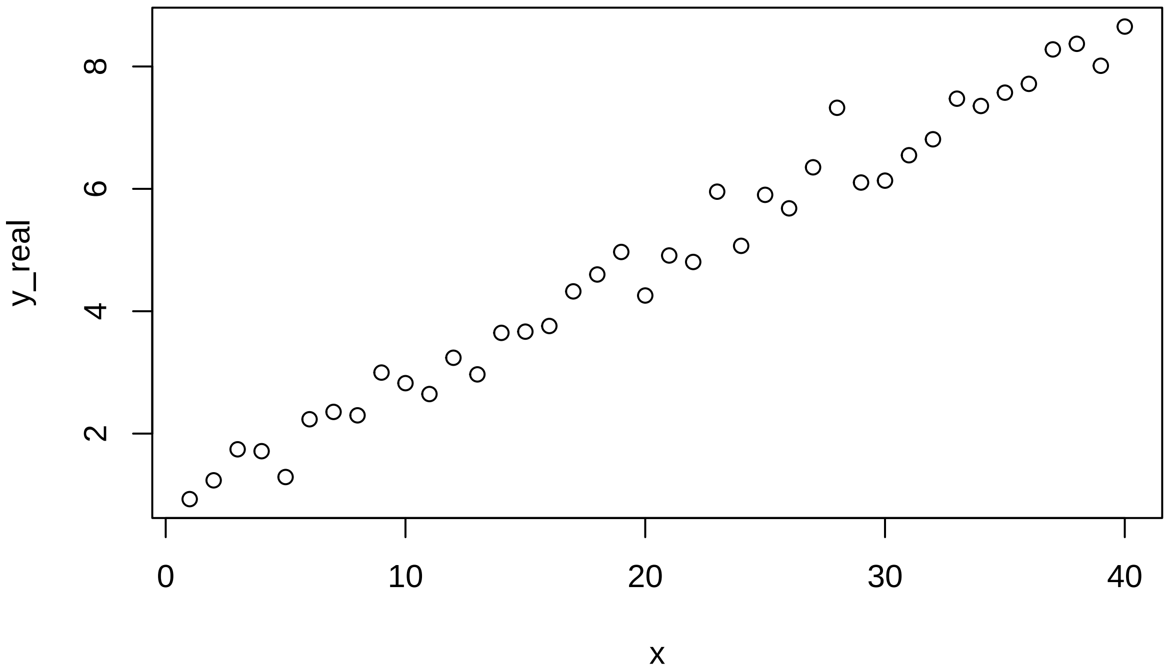
```
# Explanatory variable
x <- 1:40
# Target variable
y <- 2*exp(0.2*x)

plot(x, y)
```




```
y_trans <- log(2) + 0.2*x

plot(x, y_trans)
```




```
# Normal distributed error
e <- rnorm(40, 0, 0.4)

# Real data from a sample
y_real <- y_trans + e

plot(x, y_real)
```




```
df <- data.frame(x, y_real)

model_3 <- lm(y_real~x, data=df)
```

```r
summary(model_3)
```

```
##
## Call:
## lm(formula = y_real ~ x, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54131 -0.24091 -0.01033  0.21180  1.06162
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.868012   0.108768    7.98 1.21e-09 ***
## x           0.192682   0.004623   41.68  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3375 on 38 degrees of freedom
## Multiple R-squared:  0.9786, Adjusted R-squared:  0.978
## F-statistic:  1737 on 1 and 38 DF,  p-value: < 2.2e-16
```

Do you recognide the parameters? Are they well estimated? Do you find log(2)?

Prestige example

We load `Prestige` dataset:

```r
df <- Prestige
names(df)
```

```
## [1] "education" "income"    "women"     "prestige"  "census"    "type"
```

We consider prestige as target - as explanatory variables only numeric ones as education, income and women.

We first create the simplest model: no explanatory variables:

```r
m0<-lm(prestige~1, data = df)
summary(m0) # mean(prestige)
```

```
##
## Call:
## lm(formula = prestige ~ 1, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.033 -11.608  -3.233  12.442  40.367
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   46.833      1.703   27.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.2 on 101 degrees of freedom
```

Notice that the intercept is the mean of the variable and the residual standard error is the standard deviation of the variable:

```
mean(df$prestige)
```

```
## [1] 46.83333
```

```
sd(df$prestige)
```

```
## [1] 17.20449
```

Does it make sense?

Let's add one explanatory variable: education.

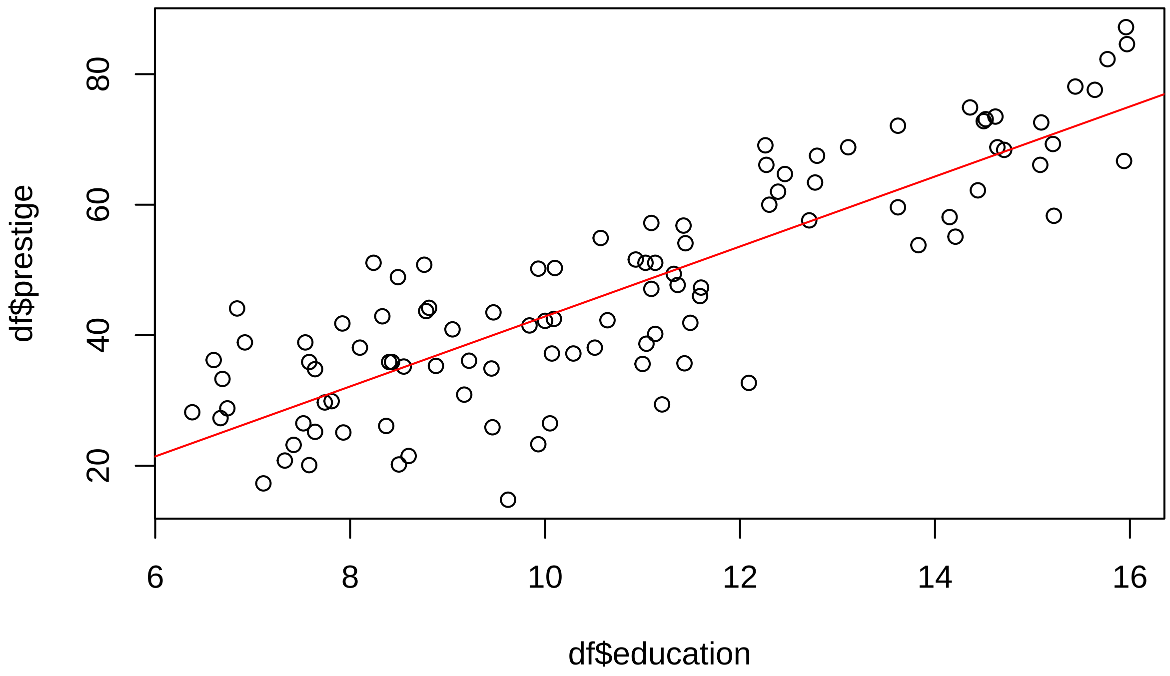
```
m1<-lm(prestige~education, data = df) # y = -11 + 5.4*educ
summary(m1)
```

```
##
## Call:
## lm(formula = prestige ~ education, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.0397  -6.5228   0.6611   6.7430  18.1636
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -10.732      3.677  -2.919  0.00434 **
## education      5.361      0.332  16.148  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.103 on 100 degrees of freedom
## Multiple R-squared:  0.7228, Adjusted R-squared:   0.72
## F-statistic: 260.8 on 1 and 100 DF,  p-value: < 2.2e-16
```

We have reduced quite considerable the residual standard error: adding education allows us to be more refined in the definition of the mean of the target variable. We improved the model: education explains up to 72% of the response variability. Adding this parameter, we improve the null model.

Given that we are using only one regressor variable, we can do a scatterplot and the regression line obtained:

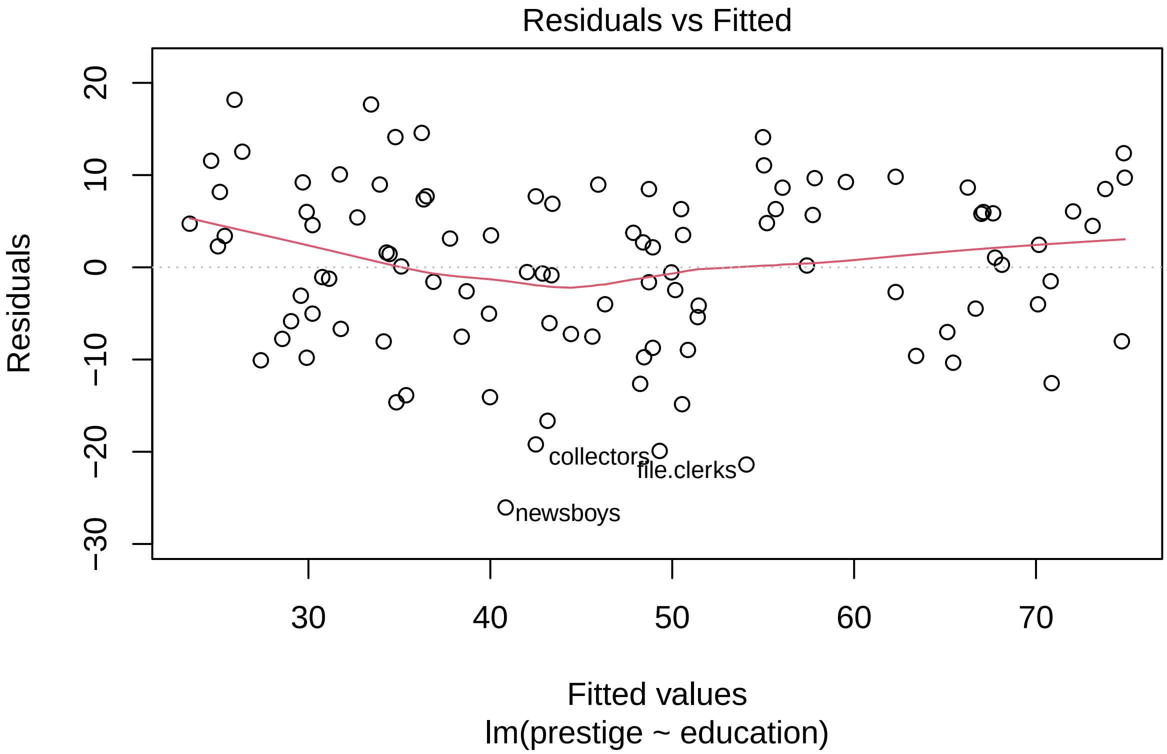
```
plot(df$education, df$prestige)
abline(a = m1$coefficients[1], b = m1$coefficients[2], col = "red")
```

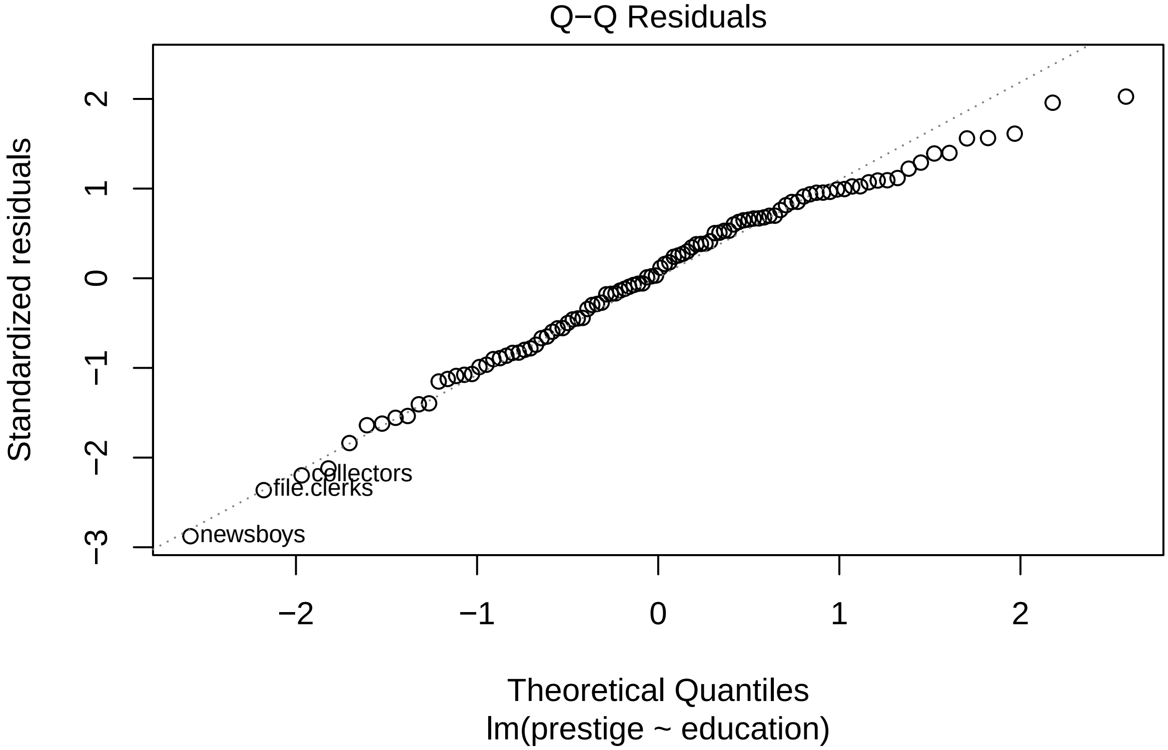


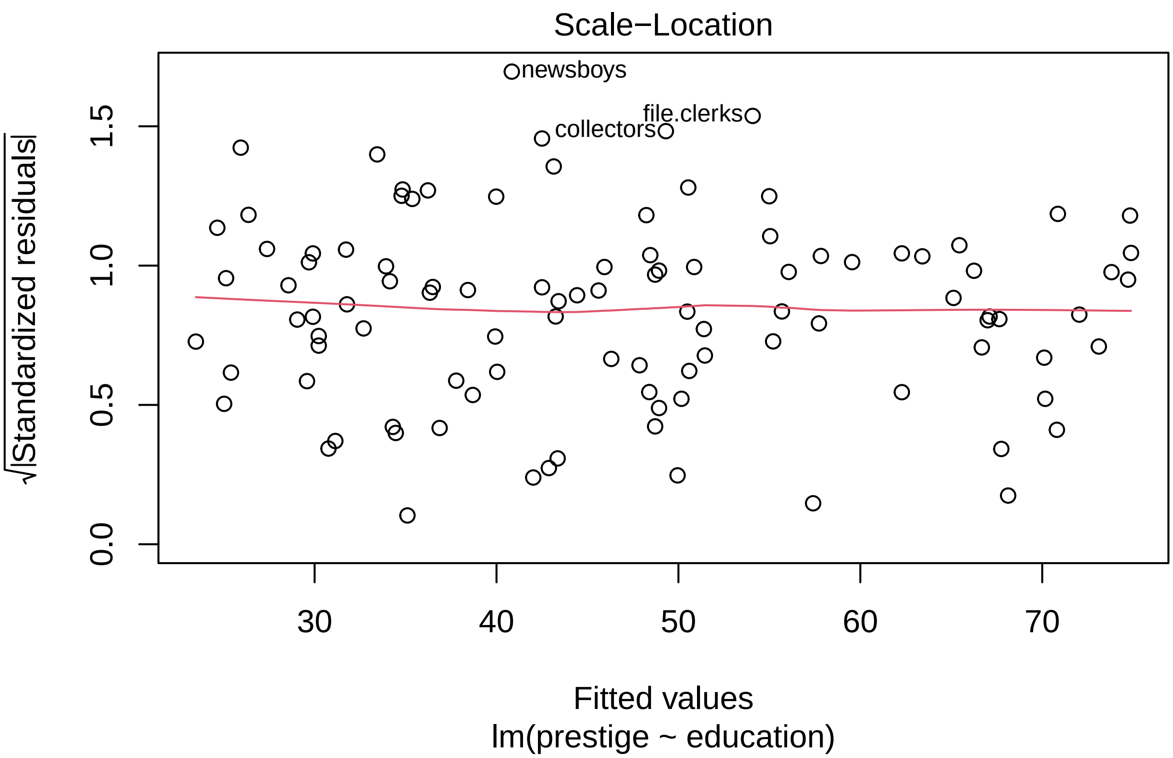
We **always** need to validate the model, doing residual analysis.

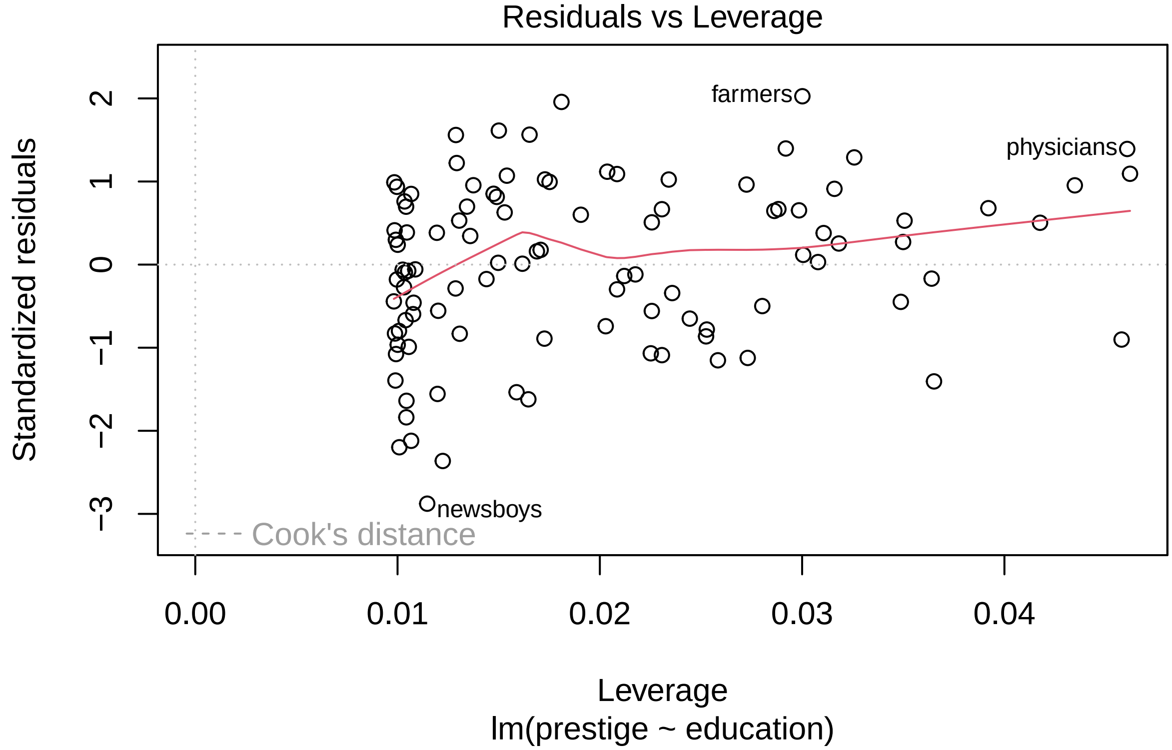
Standard model plots:

```
plot(m1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(prestige ~ education)
collectors
file.clerks
newsboys


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(prestige ~ education)
collectors
file.clerks
newsboys





We have moreless homocedasticity (some education levels have larger variance), higher values have higher residuals (breaking a bit normality). There are not many too influencial observations.

We can add now a second regressor: income.

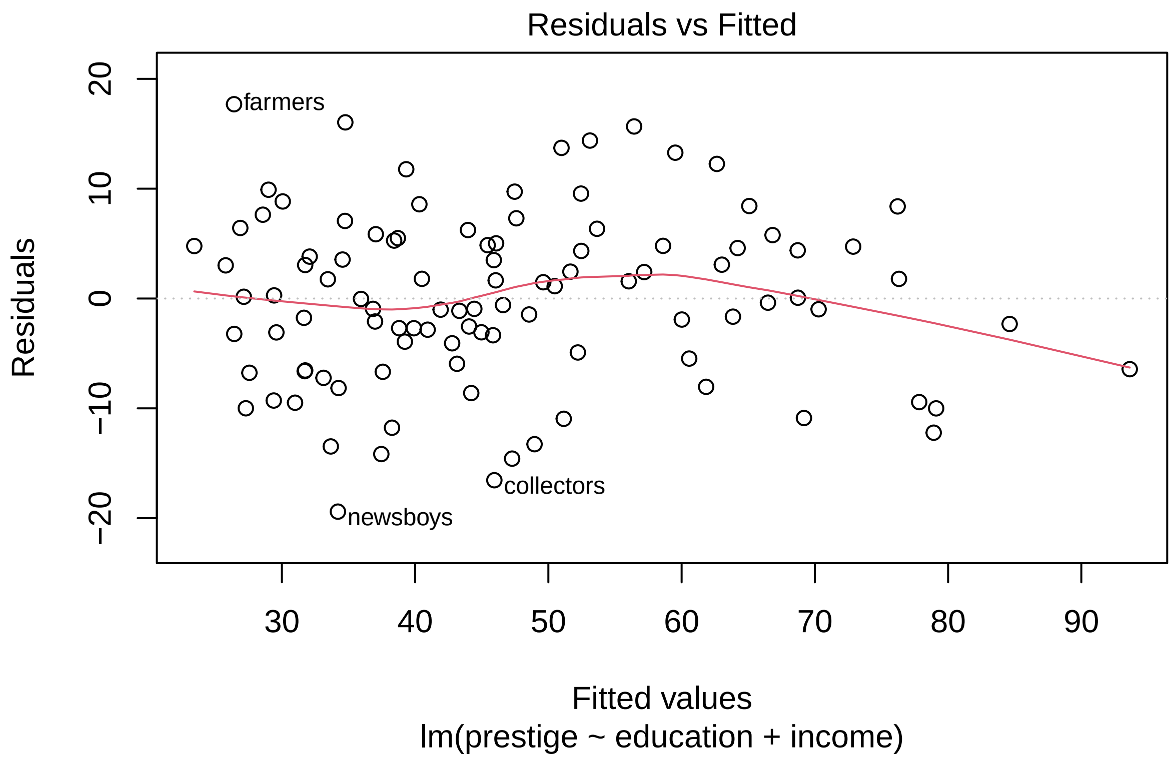
```
m2 <- lm(prestige~education+income, data=df)
# scatter3d(prestige~education+income, data=df)
```

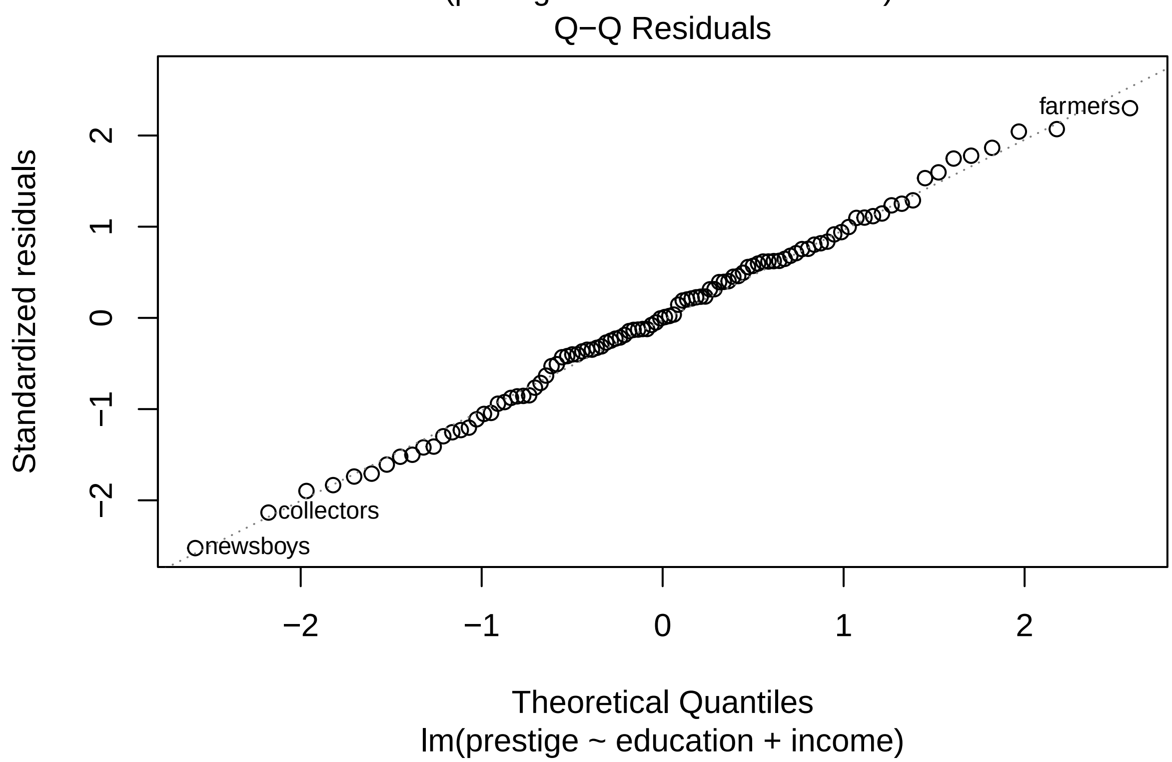
```r
summary(m2)
```

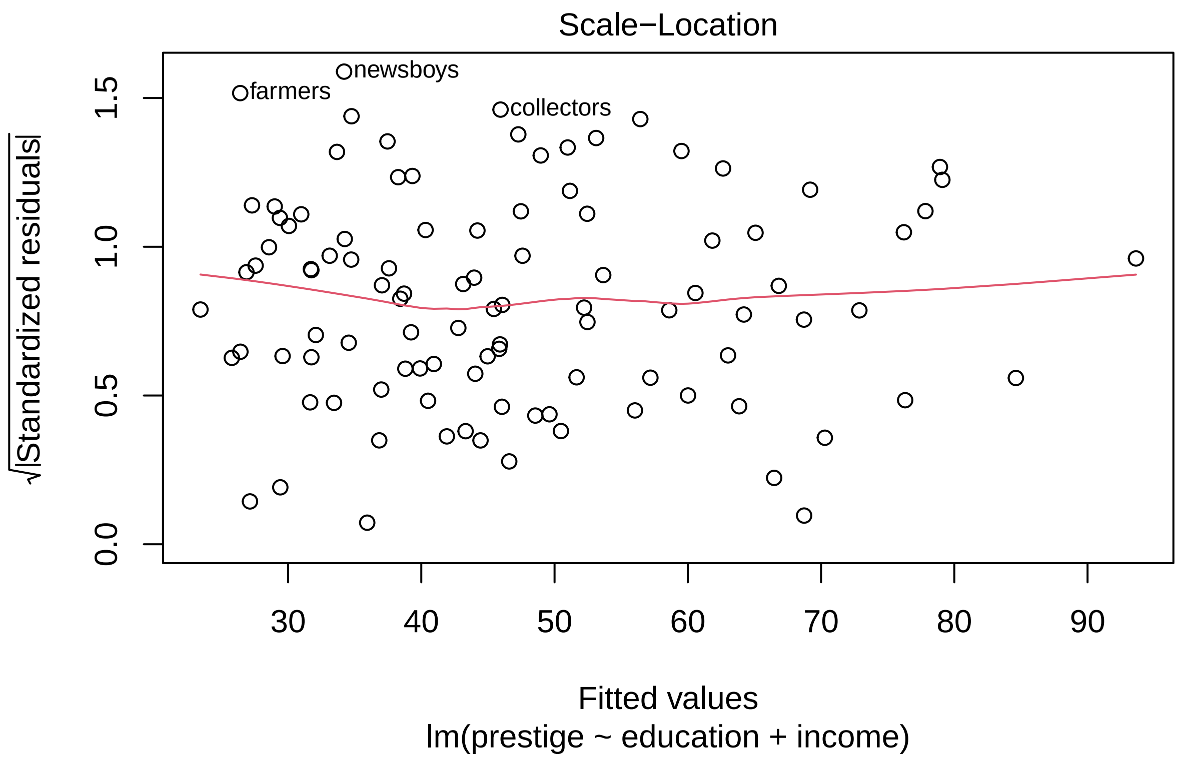
```
##
## Call:
## lm(formula = prestige ~ education + income, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.4040  -5.3308   0.0154   4.9803  17.6889
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.8477787  3.2189771  -2.127   0.0359 *
## education    4.1374444  0.3489120  11.858  < 2e-16 ***
## income       0.0013612  0.0002242   6.071 2.36e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.81 on 99 degrees of freedom
## Multiple R-squared:  0.798,  Adjusted R-squared:  0.7939
## F-statistic: 195.6 on 2 and 99 DF,  p-value: < 2.2e-16
```

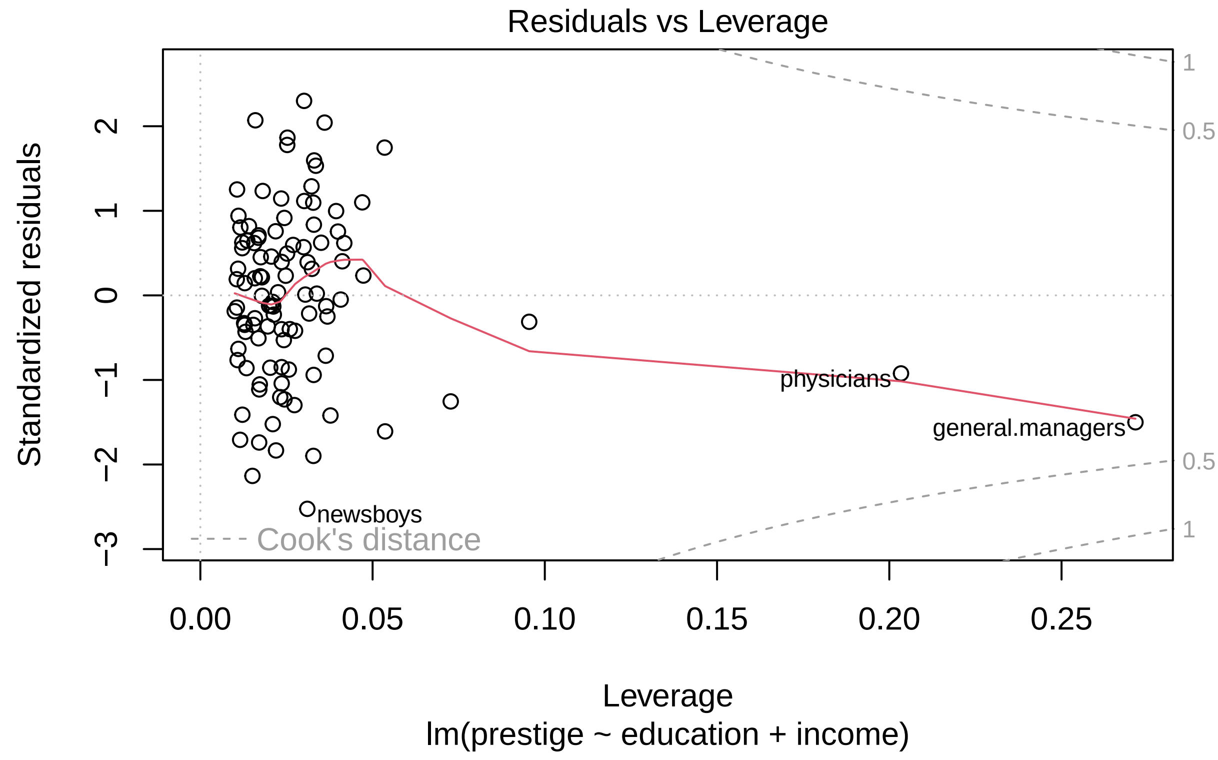
All parameters seem to be significative. On top of it, adding the new variable reduces standard error of the predictions and improves explained variability (80%). However, visual exploration now is much more complex: we need to find alternatives to validate model assumptions such as residual analysis. Are they normally distributed? Do we have homocedasticity?

```r
plot(m2)
```


Residuals vs Fitted
Residuals
farmers
newsboys
collectors
20
10
0
−10
−20
30
40
50
60
70
80
90
Fitted values
lm(prestige ~ education + income)


Q−Q Residuals
Standardized residuals
farmers
collectors
newsboys
2
1
0
−1
−2
−2
−1
0
1
2
Theoretical Quantiles
lm(prestige ~ education + income)





We have homocedasticity up to a certain range, where we find as well more infuential observations (physicians, general managers).

We can compute confidence intervals for the parameters of the model given that we have a probability distribution associated to them:

```r
confint(m2, level = 0.95)
```

```
##                     2.5 %       97.5 %
## (Intercept) -1.323493e+01 -0.460629799
## education    3.445127e+00  4.829761535
## income       9.162805e-04  0.001806051
```

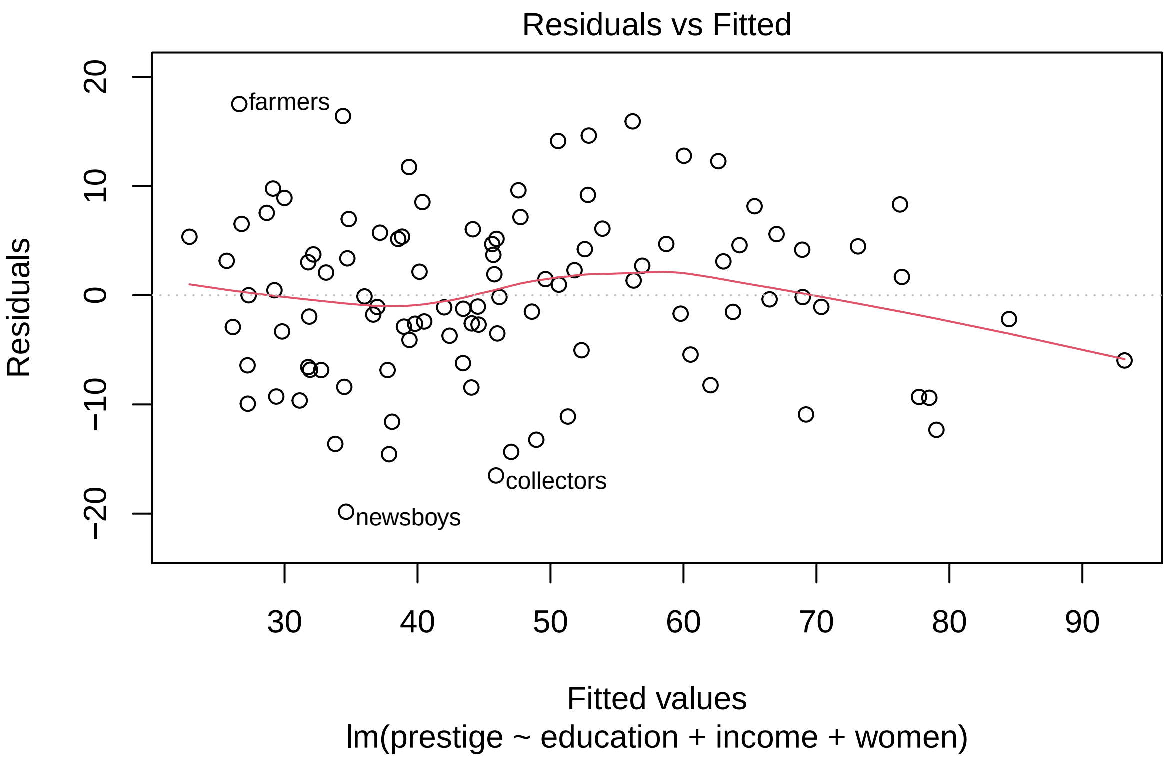
We add now a third regressor: women.

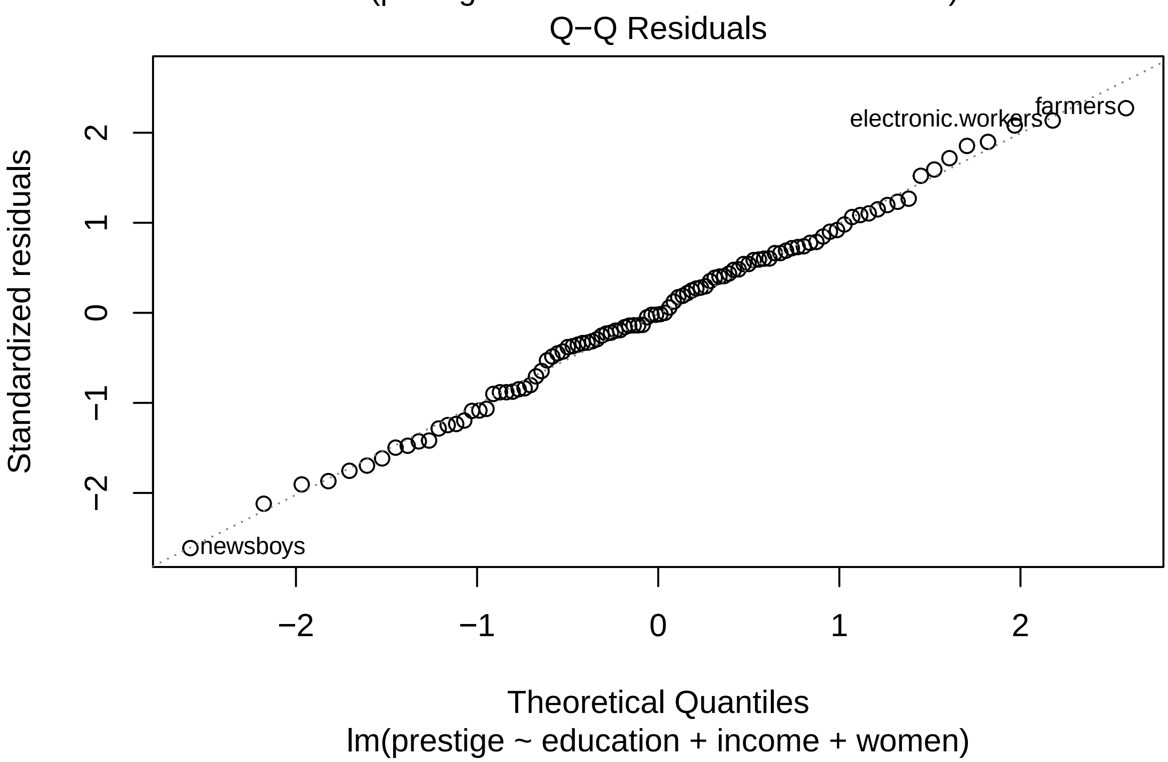
```r
m3 <- lm(prestige~education+income+women, data=df)
summary(m3)
```

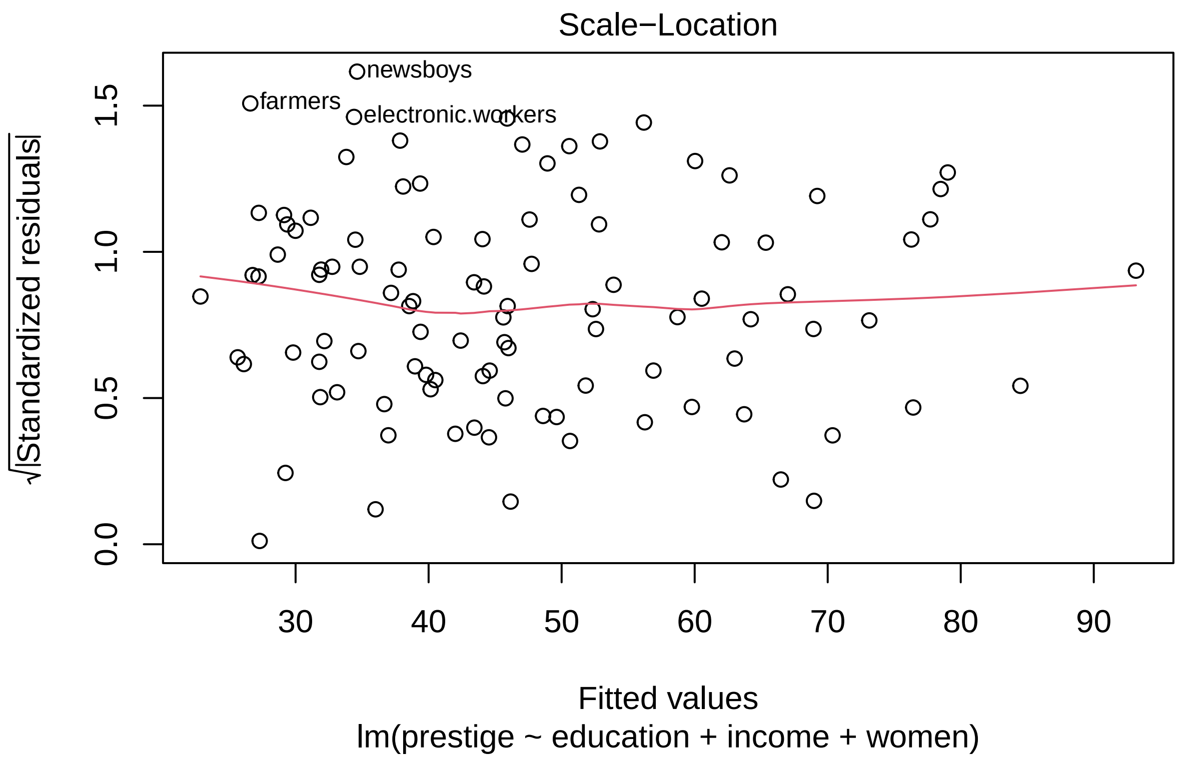
```
##
## Call:
## lm(formula = prestige ~ education + income + women, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.8246  -5.3332  -0.1364   5.1587  17.5045
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.7943342  3.2390886  -2.098   0.0385 *
## education    4.1866373  0.3887013  10.771  < 2e-16 ***
## income       0.0013136  0.0002778   4.729 7.58e-06 ***
## women       -0.0089052  0.0304071  -0.293   0.7702
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.846 on 98 degrees of freedom
## Multiple R-squared:  0.7982, Adjusted R-squared:  0.792
## F-statistic: 129.2 on 3 and 98 DF,  p-value: < 2.2e-16
```

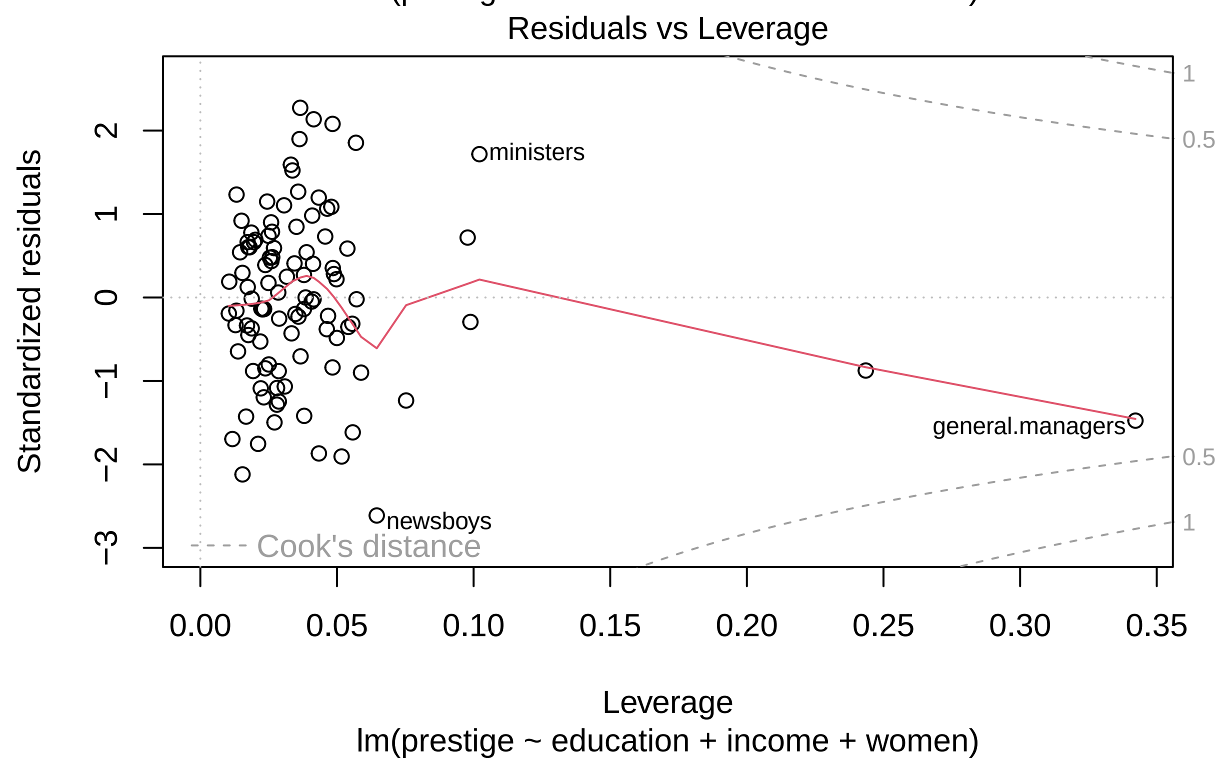
We see now that women is not significative, that is, it does not explain the variable Prestige. We see that with the addition of this variable we do not improve the residual standard error. The explained variability (R2) always increases with the addition of variables. To solve this, we check R2 adj, which takes into consideration the number of parameters and penalises the addition. We see that the addition of variable women worsens R2 adj.

```r
plot(m3)
```


Residuals vs Fitted
Residuals
20
10
0
−10
−20
farmers
collectors
newsboys
30
40
50
60
70
80
90
Fitted values
lm(prestige ~ education + income + women)


Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
electronic.workers
farmers
newsboys
−2
−1
0
1
2
Theoretical Quantiles
lm(prestige ~ education + income + women)


Scale−Location
newsboys
farmers
electronic.workers
√|Standardized residuals|
0.0
0.5
1.0
1.5
30
40
50
60
70
80
90
Fitted values
lm(prestige ~ education + income + women)


Residuals vs Leverage
ministers
general.managers
newsboys
Cook's distance
1
0.5
0.5
1
Standardized residuals
−3
−2
−1
0
1
2
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Leverage
lm(prestige ~ education + income + women)