Linear Regression - Example

SSSA - Applied Statistics - Chiara Seghieri and Costanza Tortù

2023-12-07

Preliminaries

Recall packages

Import Data

The data consists of a number of demographic variables (age, race, academic background, and previous real earnings), as well as a treatment indicator, and the real earnings in the year 1978 (the response).

Robert Lalonde, "Evaluating the Econometric Evaluations of Training Programs", American Economic Review, Vol. 76, pp. 604-620

```
rm(list=ls())
data("lalonde")
```

Have a first look at data

```
dim(lalonde)# units x variables
```

```
## [1] 614   9
```

```
head(lalonde)
```

```
##      treat age educ   race married nodegree re74 re75       re78
## NSW1     1  37   11  black       1        1    0    0  9930.0460
## NSW2     1  22    9 hispan       0        1    0    0  3595.8940
## NSW3     1  30   12  black       0        0    0    0 24909.4500
## NSW4     1  27   11  black       0        1    0    0  7506.1460
## NSW5     1  33    8  black       0        1    0    0   289.7899
## NSW6     1  22    9  black       0        1    0    0  4056.4940
```

Inspect variables

```
colnames(lalonde)
```

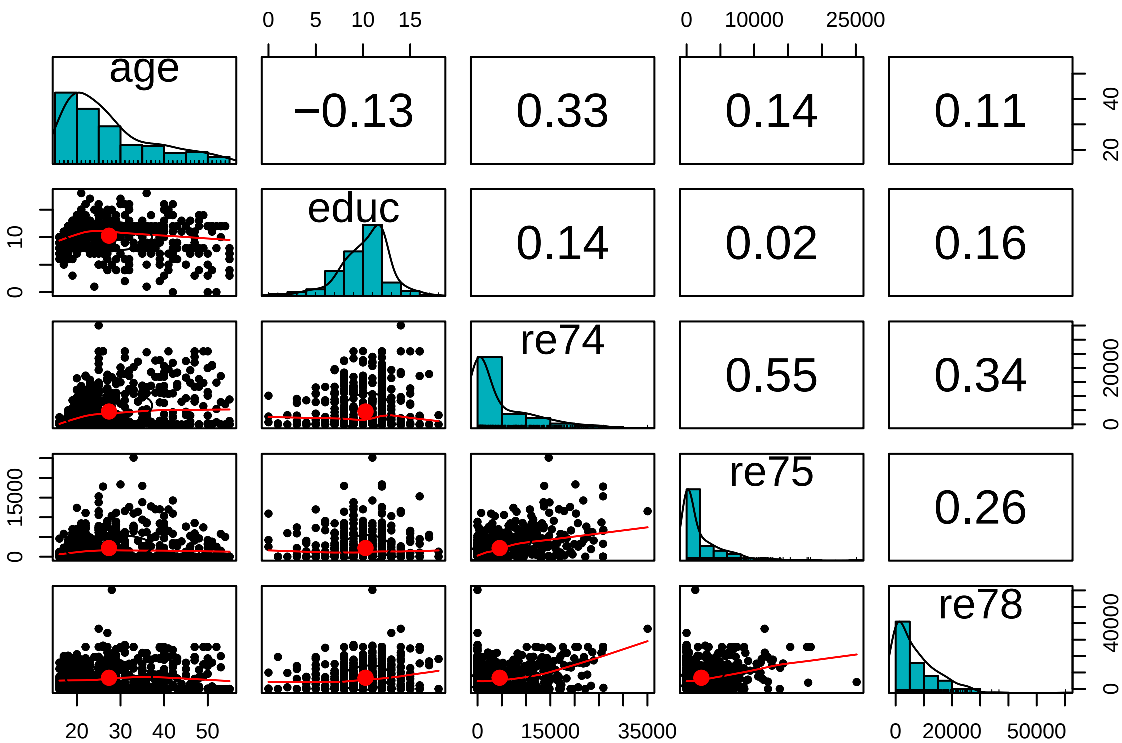
```
## [1] "treat"    "age"      "educ"     "race"     "married"  "nodegree" "re74"
## [8] "re75"     "re78"
```

```
quantitative_variables <- c("age", "educ", "re74", "re75", "re78")
qualitative_variables <- c("treat", "race", "married", "nodegree")
dummies <- c("treat", "married", "nodegree" )
```

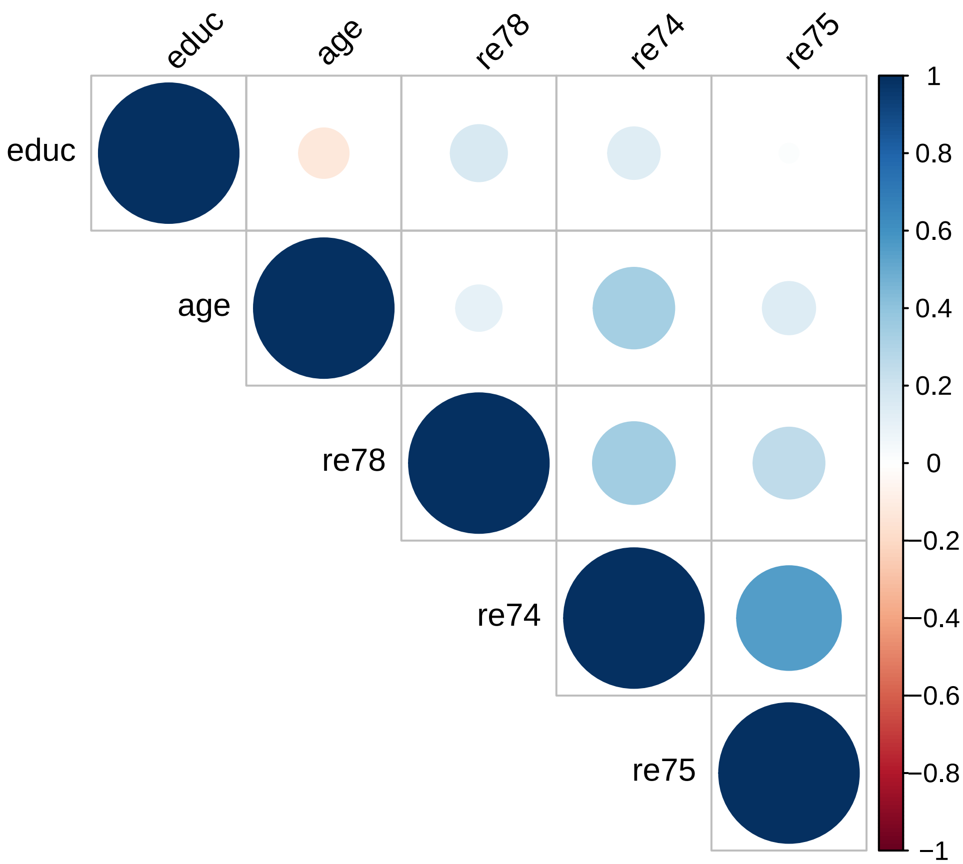
```r
lalonde$treat_factor <- as.factor(lalonde$treat)
lalonde$race_factor <- as.factor(lalonde$race)
lalonde$married_factor <- as.factor(lalonde$married)
lalonde$nodegree_factor <- as.factor(lalonde$nodegree)

qualitative_variables_factors <-  c("treat_factor", "race_factor",
                                    "married_factor", "nodegree_factor")

all_variables <- c(quantitative_variables, qualitative_variables_factors)
```

Let's focus on quantitative variables

```r
pairs.panels(lalonde[, quantitative_variables],
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
             density = TRUE,  # show density plots
             ellipses = TRUE # show correlation ellipses
             )
```

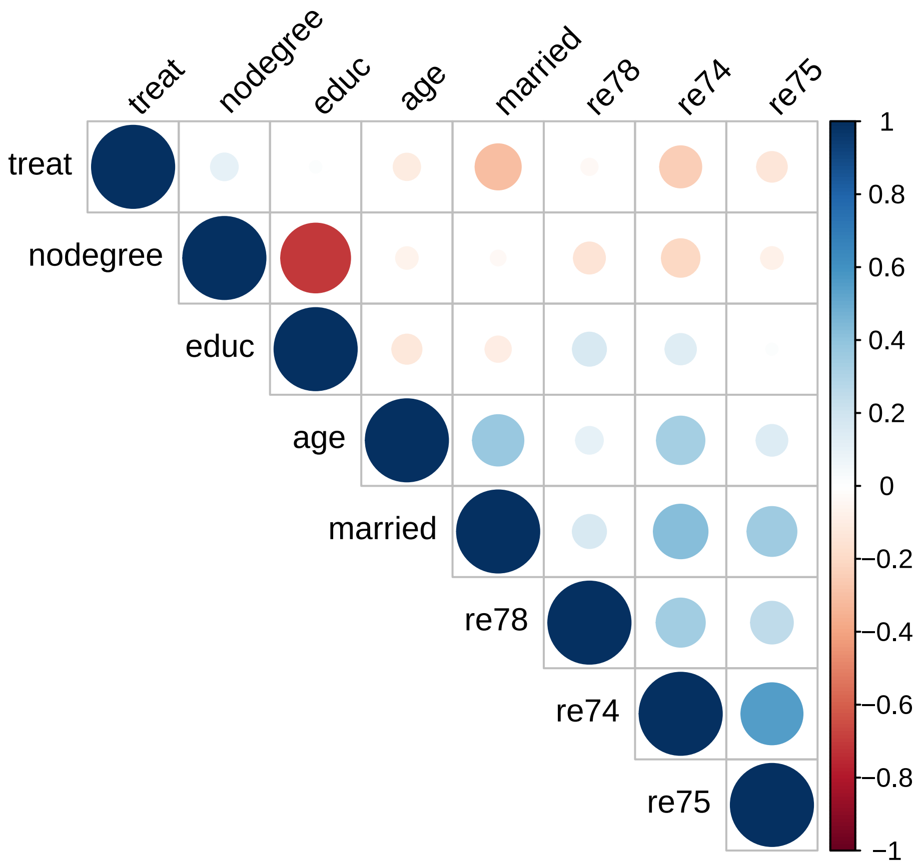



```r
correlation_matrix <- cor(lalonde[, quantitative_variables])

corrplot(correlation_matrix, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



You may also treat dummies as quantitative variables and compute correlation, but pay attention to the interpretation!!!!!!

```
correlation_matrix_withdummies <- cor(lalonde[, c(quantitative_variables,dummies)])

corrplot(correlation_matrix_withdummies, type = "upper",
         order = "hclust",
         tl.col = "black",
         tl.srt = 45)
```



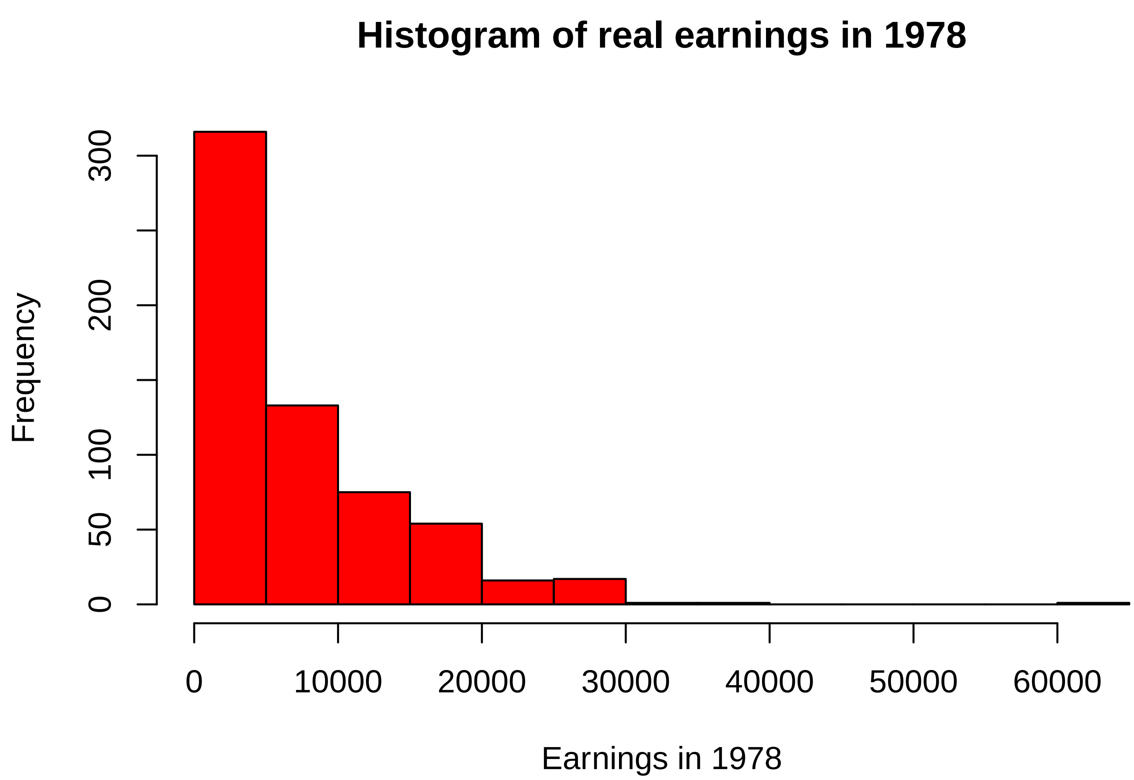
Run a regression model

The response variable measures earnings in 1978 while the marital status the age, the education, the race and the training program are independent variables.

Make sure your data meet the normality assumption

Let's have a look at the distribution of earnings in 1978

```
hist(lalonde$re78,
     main ="Histogram of real earnings in 1978",
     col = "red",
     xlab = "Earnings in 1978")
```

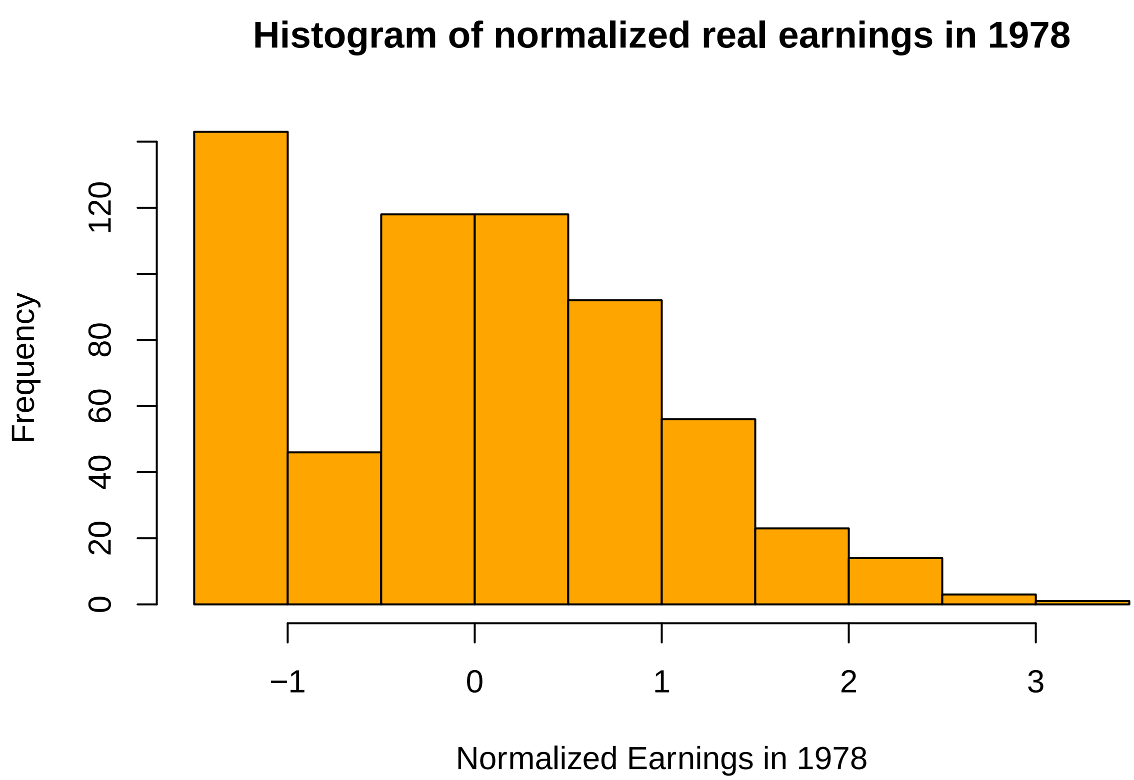


This is far from normality, let's apply a normalization trasformation

```
re78_BN <- bestNormalize(lalonde$re78)
re78_BN
```

```
## Best Normalizing transformation with 614 Observations
##  Estimated Normality Statistics (Pearson P / df, lower => more normal):
##  - arcsinh(x): 12.5241
##  - Center+scale: 7.1465
##  - Double Reversed Log_b(x+a): 8.3531
##  - Log_b(x+a): 17.6774
##  - orderNorm (ORQ): 3.7705
##  - sqrt(x + a): 3.7965
##  - Yeo-Johnson: 5.2671
## Estimation method: Out-of-sample via CV with 10 folds and 5 repeats
##
## Based off these, bestNormalize chose:
## orderNorm Transformation with 614 nonmissing obs and ties
##  - 457 unique values
##  - Original quantiles:
##        0%       25%       50%       75%      100%
##     0.000   238.283  4759.018 10893.592 60307.930
```

```
lalonde$re78_normalized <- re78_BN$x.t

hist(lalonde$re78_normalized,
     main ="Histogram of normalized real earnings in 1978",
     col = "orange",
     xlab = "Normalized Earnings in 1978")
```



Simple Linear Regression

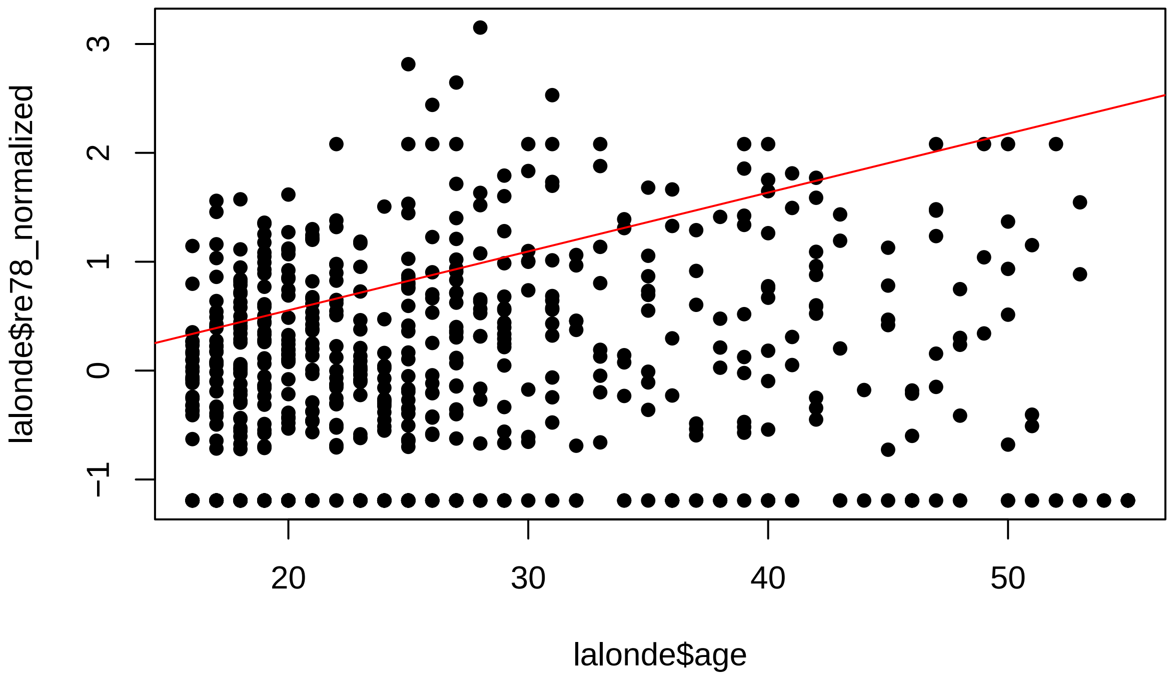
We investigate the relationship between the education and the real earnings in the year 1978

educ: years of education re78: earnings in 1978

```
simple_model_normalized <- lm(re78_normalized ~ educ,
                   data = lalonde)

summary(simple_model_normalized)
```

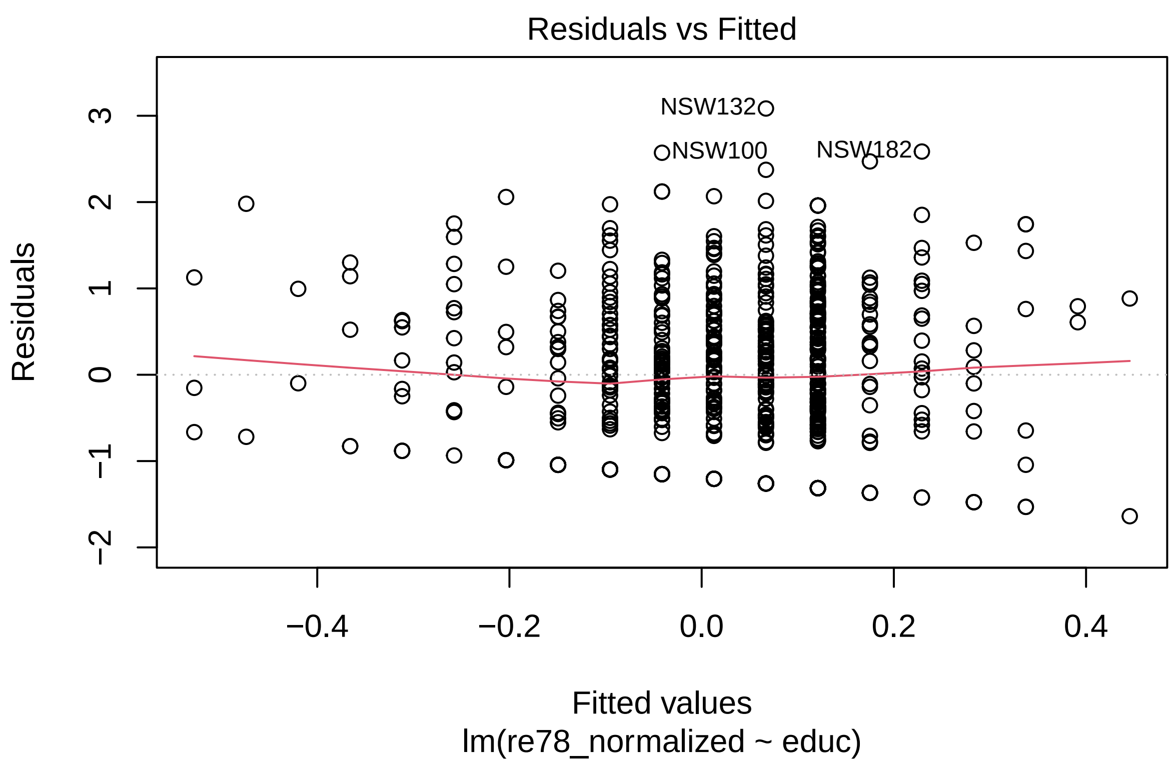
```
##
## Call:
## lm(formula = re78_normalized ~ educ, data = lalonde)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63840 -0.69846 -0.02051  0.64832  3.08381
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.52797    0.15053  -3.507 0.000486 ***
## educ         0.05408    0.01420   3.808 0.000154 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9242 on 612 degrees of freedom
## Multiple R-squared:  0.02315,    Adjusted R-squared:  0.02155
## F-statistic:  14.5 on 1 and 612 DF,  p-value: 0.0001542
```

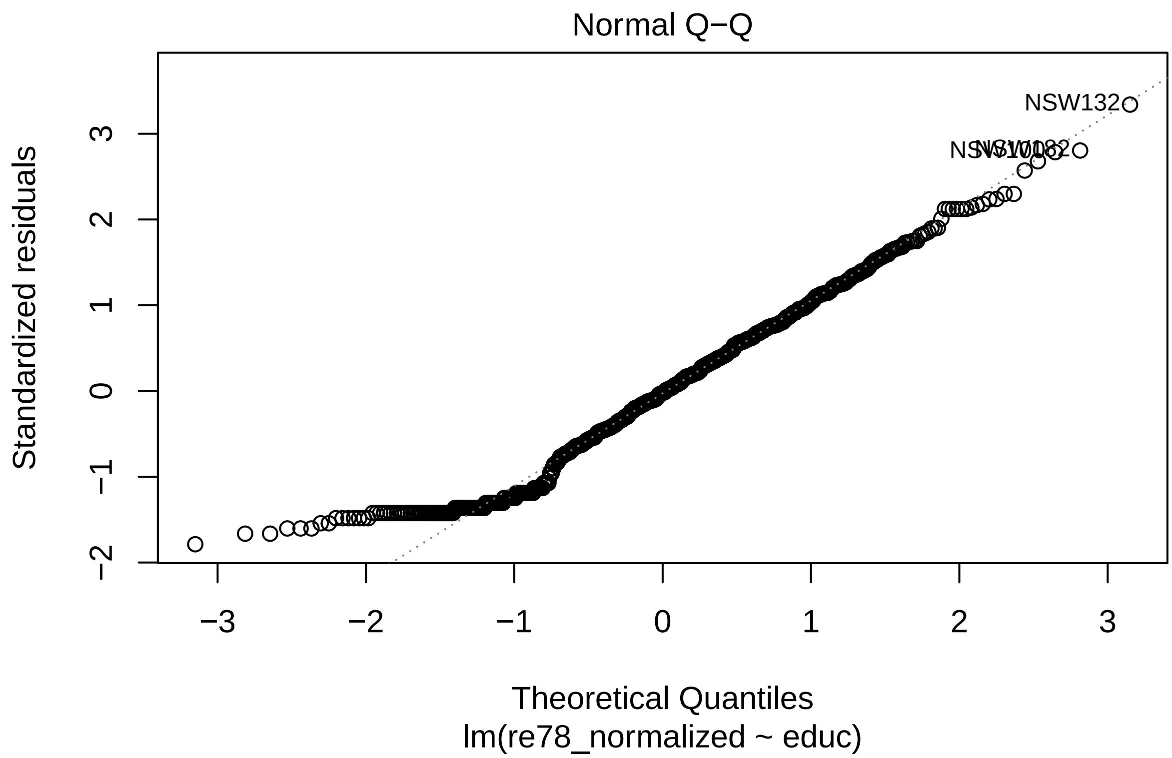
```r
plot(lalonde$age, lalonde$re78_normalized, pch=16)
abline(simple_model_normalized, col="red" )
```

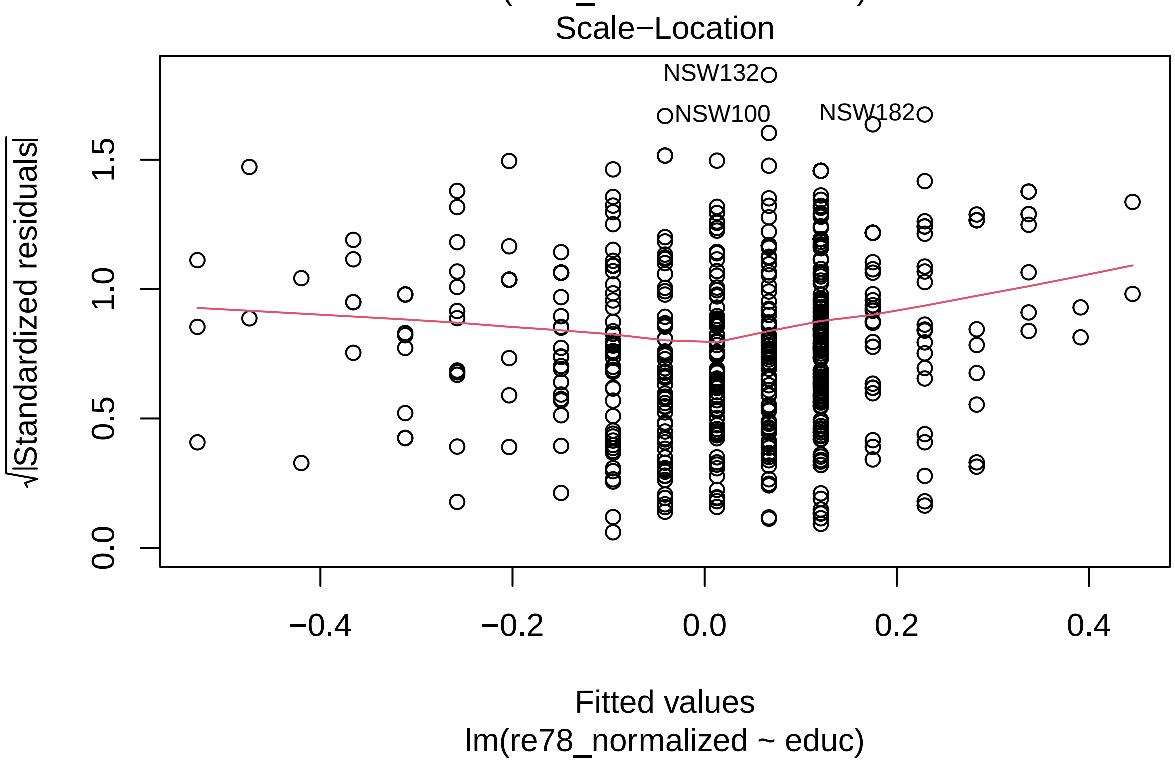


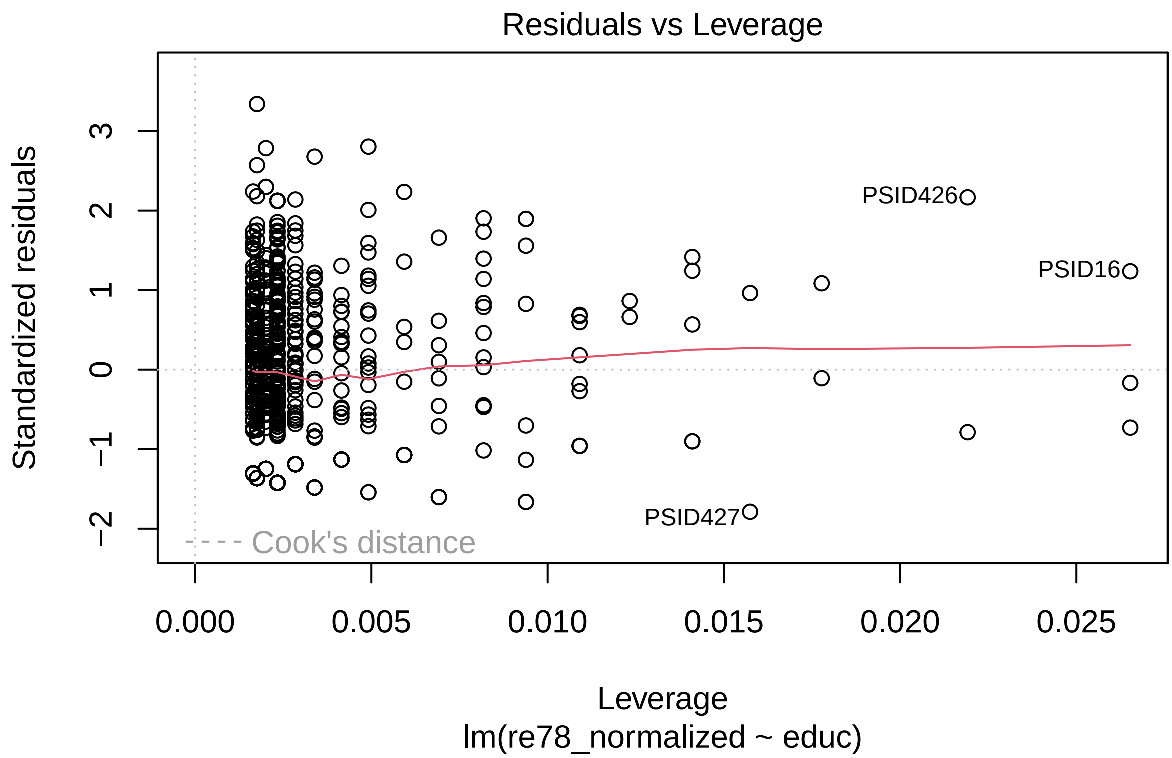
Further inspect your model

```r
plot(simple_model_normalized)
```




Normal Q−Q
NSW132
NSW100
NSW182
Standardized residuals
Theoretical Quantiles
lm(re78_normalized ~ educ)


Scale−Location
NSW132
NSW100
NSW182
√|Standardized residuals|
Fitted values
lm(re78_normalized ~ educ)



Multiple Regression

We investigate the determinants of real earnings in the year 1978

age: gae (years), numeric treat: attendence of a training program educ: years of education married: marital status re78: earnings in 1978

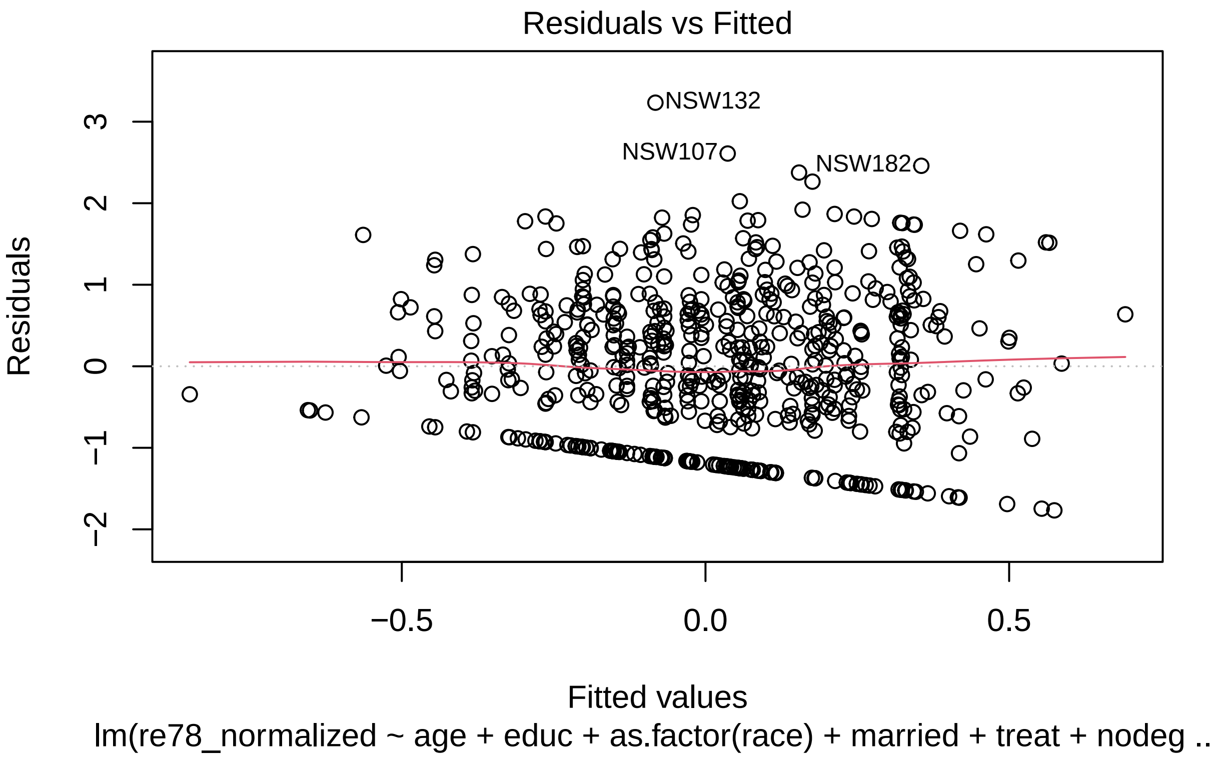
```
multiple_model_normalized <-lm(re78_normalized ~ age + educ + as.factor(race)
                              + married + treat + nodegree,
                                data = lalonde)

summary(multiple_model_normalized)
```

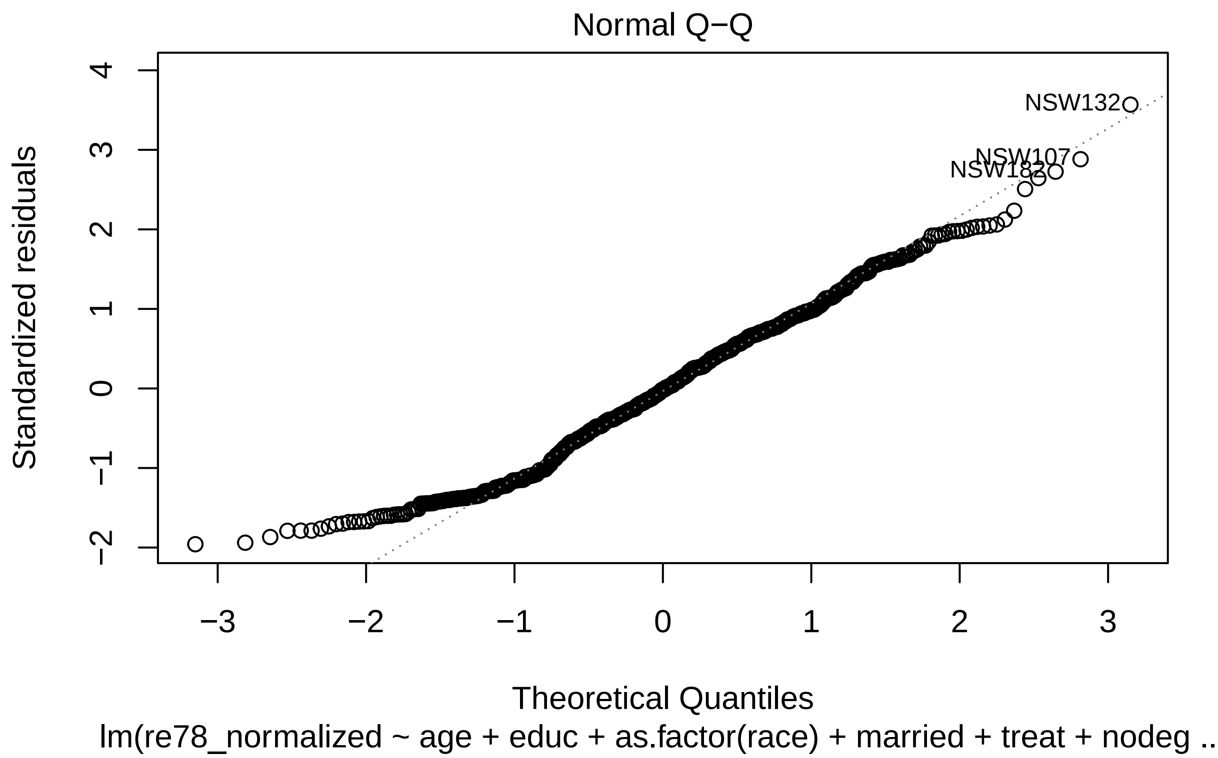
```
##
## Call:
## lm(formula = re78_normalized ~ age + educ + as.factor(race) +
##     married + treat + nodegree, data = lalonde)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7674 -0.6991 -0.0141  0.6426  3.2332
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -0.9429339  0.3191322  -2.955  0.00325 **
## age                     0.0009189  0.0041241   0.223  0.82377
## educ                    0.0602261  0.0206487   2.917  0.00367 **
## as.factor(race)hispan   0.3542669  0.1331197   2.661  0.00799 **
## as.factor(race)white    0.2565782  0.1003492   2.557  0.01080 *
## married                 0.2604112  0.0853849   3.050  0.00239 **
```

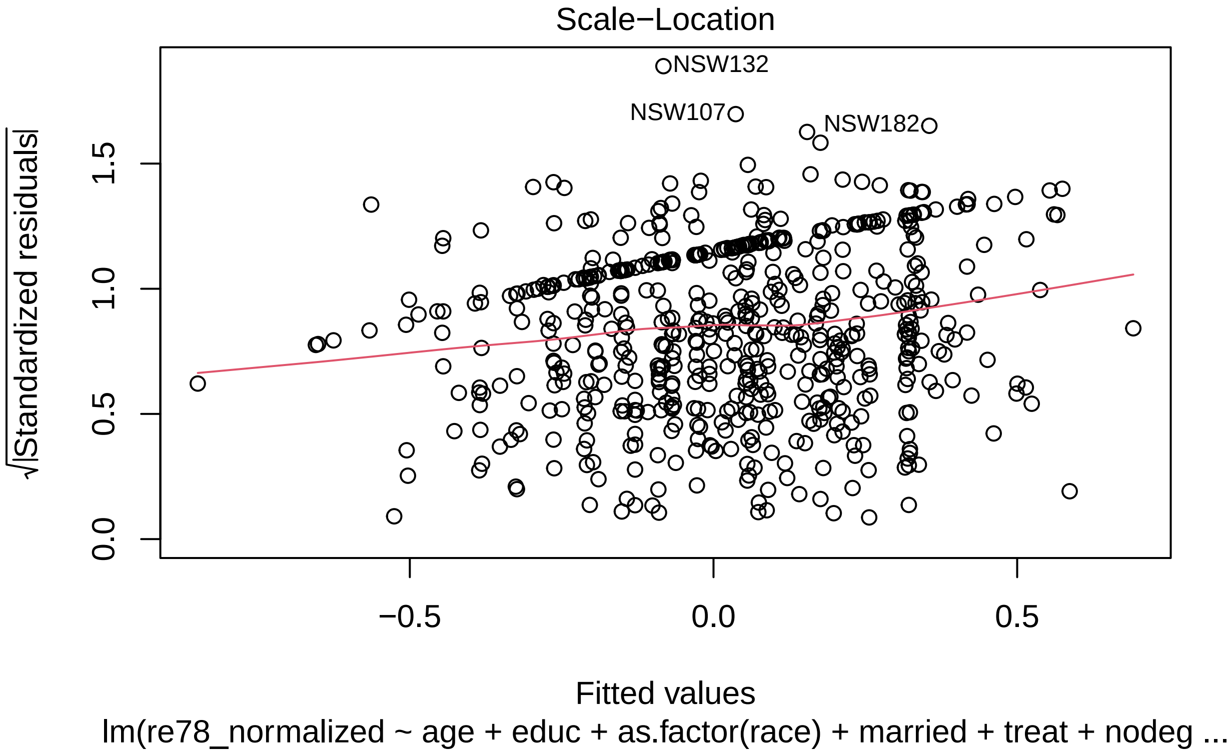
```
## treat                     0.1717658  0.1019511   1.685  0.09254 .
## nodegree                  0.0005064  0.1108397   0.005  0.99636
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9098 on 606 degrees of freedom
## Multiple R-squared:  0.06253,    Adjusted R-squared:  0.0517
## F-statistic: 5.775 on 7 and 606 DF,  p-value: 1.719e-06
```

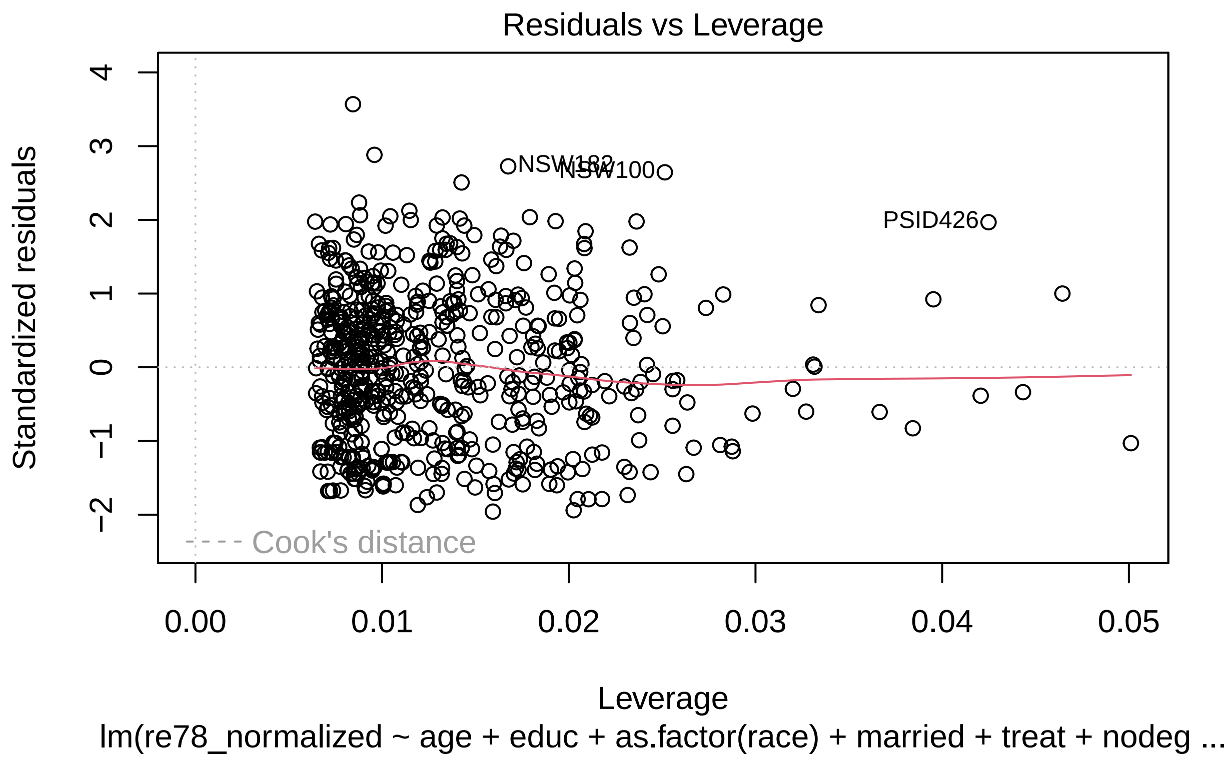
Furtehr inspect your model

```
plot(multiple_model_normalized)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(re78_normalized ~ age + educ + as.factor(race) + married + treat + nodeg ...
NSW132
NSW107
NSW182
−3
−2
−1
0
1
2
3
4


Scale−Location
√|Standardized residuals|
Fitted values
lm(re78_normalized ~ age + educ + as.factor(race) + married + treat + nodeg ...
NSW132
NSW107
NSW182
0.0
0.5
1.0
1.5
−0.5



Add valuable interactions

```
multiple_model_normalized_winteractions <-lm(re78_normalized ~ age + educ + as.factor(race) +
                                             married + treat + nodegree +
                                             as.factor(race)*educ + treat*nodegree,
                                             data = lalonde)

summary(multiple_model_normalized_winteractions)
```

```
##
## Call:
## lm(formula = re78_normalized ~ age + educ + as.factor(race) +
##     married + treat + nodegree + as.factor(race) * educ + treat *
##     nodegree, data = lalonde)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8607 -0.6772  0.0010  0.6355  3.2442
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -1.0857162  0.3831280  -2.834  0.00475 **
## age                         0.0009009  0.0041194   0.219  0.82695
## educ                        0.0689941  0.0295796   2.332  0.02000 *
## as.factor(race)hispan       0.9586613  0.4438597   2.160  0.03118 *
## as.factor(race)white        0.0731624  0.3724158   0.196  0.84432
## married                     0.2573807  0.0857845   3.000  0.00281 **
## treat                       0.2720856  0.1625284   1.674  0.09463 .
## nodegree                    0.0862717  0.1290781   0.668  0.50415
## educ:as.factor(race)hispan -0.0666634  0.0445307  -1.497  0.13491
```

```
## educ:as.factor(race)white   0.0176651  0.0345201   0.512  0.60902
## treat:nodegree             -0.1502407  0.1878521  -0.800  0.42415
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9085 on 603 degrees of freedom
## Multiple R-squared:  0.06989,    Adjusted R-squared:  0.05447
## F-statistic: 4.531 on 10 and 603 DF,  p-value: 3.321e-06
```