Practice work 1

Maria Korzun

29 october 2016

Introduction

We use data which contains information about time when person relax (sleep, sleep & naps, relax all), time spent working in the market (first job, second job, total) and different economic variables.

Time is scarce; and we always have to choose between work and rest. In section I we establish the existence of a relationship between rest time and time spent working in the market and other socio-demographic and labor variables. In section II we try to understand how timing affects wage.

Section I

Desciptive statistics

A table of statistics for our sample:

```
##
## ================================================
## Statistic  N    Mean    St. Dev.   Min    Max
## ------------------------------------------------
## age       405  38.899    11.571     23     65
## black     405   0.049     0.217     0       1
## clerical  405   0.195     0.397     0       1
## construc  405   0.032     0.176     0       1
## educ      405  12.679     2.691     1      17
## earns74   405 9,688.889 8,553.928   0    42,500
## gdhlth    405   0.884     0.321     0       1
## inlf      405   1.000     0.000     1       1
## leis1     405 4,670.232  868.360  2,140  7,335
## leis2     405 4,548.817  868.517  2,140  7,297
## leis3     405 4,498.141  868.454  2,140  7,282
## smsa      405   0.393     0.489     0       1
## lhrwage   405   1.417     0.636   -0.673 3.570
## lothinc   405   6.458     4.034   0.000  10.657
## male      405   0.548     0.498     0       1
## marr      405   0.812     0.391     0       1
## prot      405   0.677     0.468     0       1
## rlxall    405 3,439.889  515.833  1,905  6,110
## selfe     405   0.077     0.266     0       1
## sleep     405 3,267.798  416.819  1,905  4,695
## slpnaps   405 3,389.212  493.727  1,905  6,110
## south     405   0.210     0.408     0       1
## spsepay   405 5,250.494 7,501.976   0    50,000
## spwrk75   405   0.523     0.500     0       1
## totwrk    405 2,141.970  907.177    0    5,020
## union     405   0.225     0.418     0       1
```

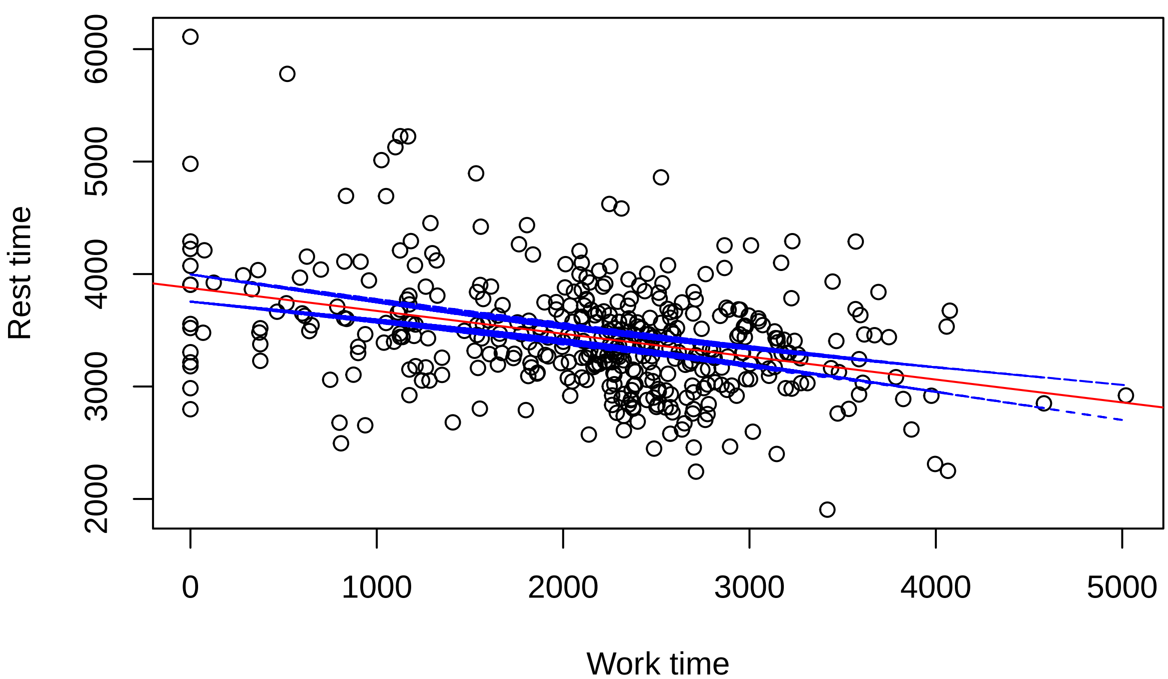
```
## worknrm   405 2,113.768  905.887    0    5,020
## workscnd  405   28.202   144.836    0    1,337
## exper     405  20.220    12.479     1      55
## yngkid    405   0.133     0.340     0       1
## yrsmarr   405  11.373    11.585     0      43
## hrwage    405   5.045     3.705   0.510  35.510
## agesq     405 1,646.672  970.308   529   4,225
## ----------------------------------------------
```

In column (3) of table above we present the means of the variables. Minimum and maximum meanings are presented in colums (5) and (6). As can be seen from the table, data contains information about 405 respondents. All of them work in the market. Their age ranges from 23 to 65 years. 54.8% of respondents are men, others - women. Respondents have various socio-demographic and labor characteristics.

Average time of total work is 2142 minutes per week. It is approximately 5 hours per day. Time of total work is very diversified: from 0 to 12 hours per day. The same situation with total time spent relaxing: it varies from 5 to 15 hours per day. The average rest duration is nearly 8 hours per day.

1. Firstly, we will check relationship between rest time (*rlxall*) and work time (*totwrk*).

Picture 1



As can be seen from picture 1, there are negative linear relationship between time spent relaxing and work time.

Summary shows that actually rest time and work time correlate with each other. The coefficient on the variable *totwrk* equals -0.20, what is more, it is high significant (***) relationship.

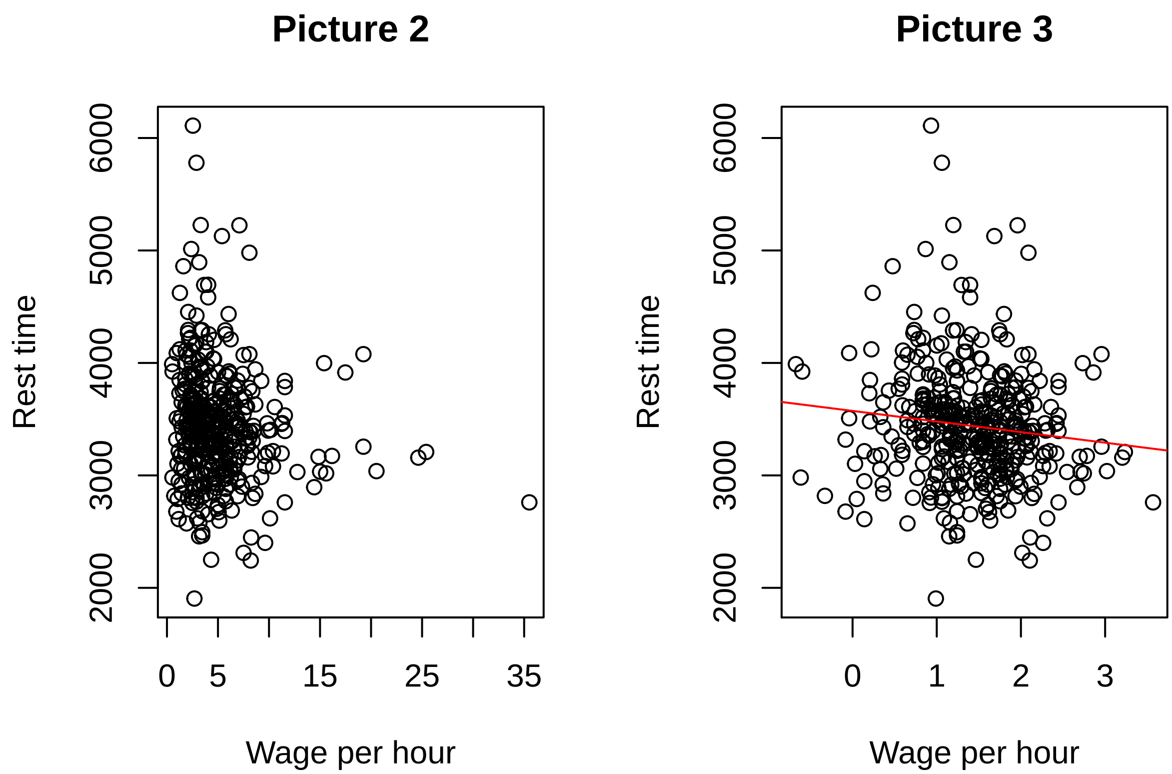
We reject hypothesis of equality of this coefficient to zero because p-value: 1.128e-13 (F-test).

Also we can check information about relationship using confidence interval (blue lines at picture 1).

```
##                     2.5 %       97.5 %
## (Intercept) 3754.7105858 3996.5706810
## totwrk        -0.2554318   -0.1514382
```

Confidence interval goes beyond zero, so we can surely say that our coefficient on variable *totwrk* is important and unequal to zero.

2. Then, we will check relationship between rest time (*rlxall*) and wage per hour (*hrwage*)(picture 2).

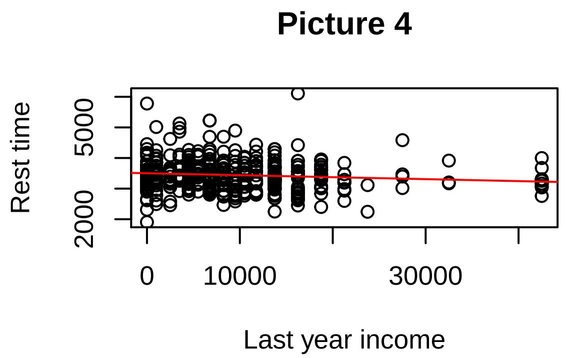


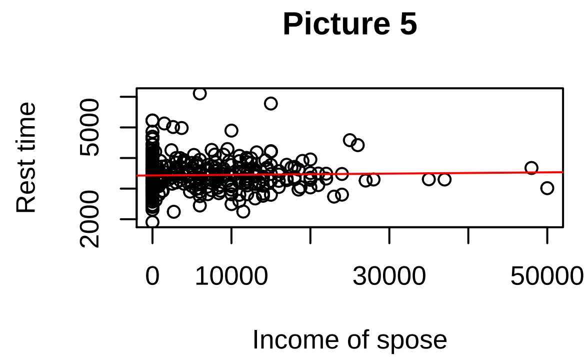
As can be seen from the graph there are nonlinear relationship between variables *rlxall* and *hrwage*. To make our future linear model more appropriate for the analysis we should use variable *lhrwage* instead *hrwage* (picture 3).

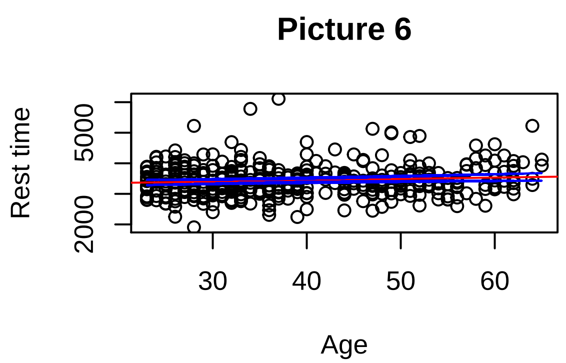
We can see negative linear relationship between wage per hour and time spent resting.

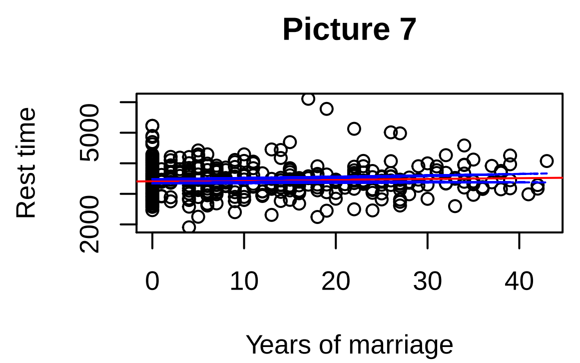
Summary confirms information on the picture 3 - the coefficient on the variable *lhrwage* is valuable at 5% significance level. P-value: $0.0195 < 0.05$ (F-test) so we should include this variable to the model.

3. What about last year income (picture 4)?









The coefficient is significant at 5% significance level. P-value: $0.03176 < 0.05$ (F-test) so we can reject the hypothesis of eqality of coefficient to zero on 5% significance level.

However, we can assume that problem of multicollinearity with variable *lhrwage* exist. Let's check it.

```
##   totwrk  lhrwage
## 1.061002 1.061002
```

Everything is OK we can include this variable to model.

4. The same algorithm with income of spose (picture 5):

This variable is unsignificant because p-value is too large: p-value: $0.5514 > 0.05$ (F-test). We can exclude this variable from model.

5. Next step is checking relationship between *rlxall* ang *age*.

We could see significant (*) positive linear correlation between age and total rest time (picture 6).

P-value: 0.04975 (F-test) which is less than 0.05 so we can reject the hypothesis of equality of this coefficient to zero. To sum up, the relationship between this two variables is significant and we should include this parameter from regression.

6. *Rlxall* vs *yrsmarr*:

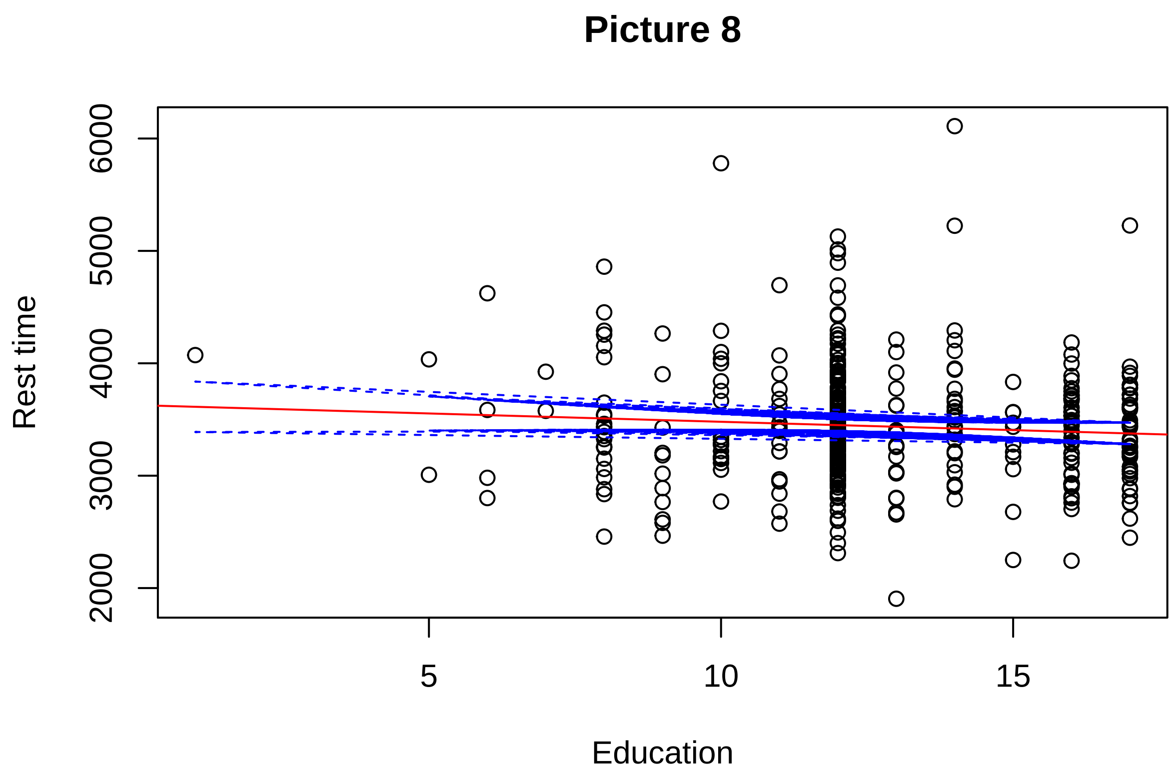
We could see small positive correlation between years of marriage and total rest time (picture 7).

P-value: 0.248 (F-test) which is more than 0.05 so we cannot reject the hypothesis of equality of this coefficient to zero.

Let's check multicolinearity *age* and *yrsmarr*:

```
##      age  yrsmarr
## 1.403781 1.403781
```

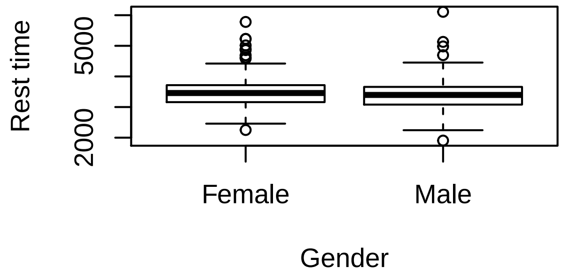
Everything is OK. Hovewer, the relationship between variables *rlxall* and *yrsmarr* is unsignificant and we can exclude this parameter from regression.



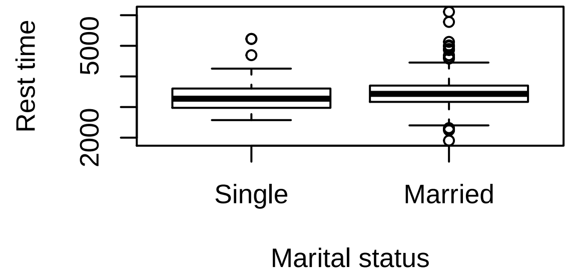
7. Variables *rlxall* and *educ* have significant (*) negative linear relationship (picture 8). P-value: 0.1204 > 0.05 (F-test) -> hypothesis of equality of this coefficient to zero is true. This coefficient should not be included to model.

At the next step of our work we will show the influence of dummy variables on relax time.

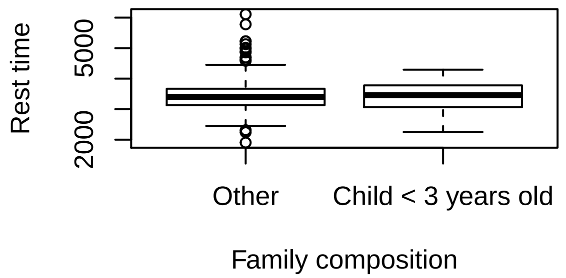
Picture 9



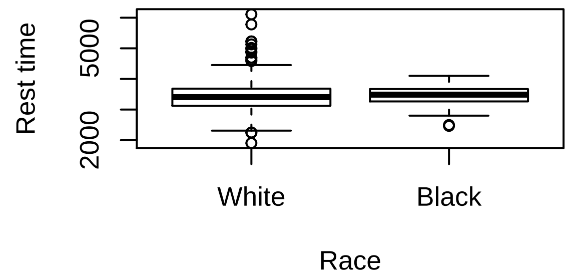
Picture 10



Picture 11



Picture 12



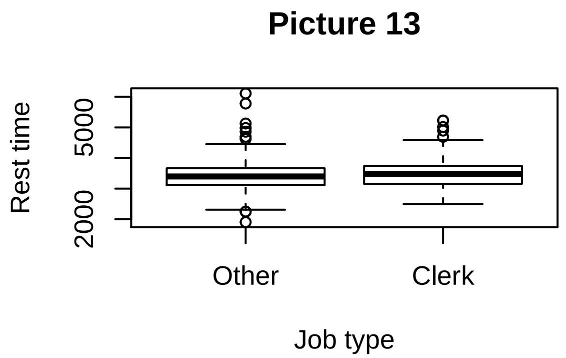
Picture 9 shows that average total rest time of women slightly exceeds total rest time of men most likely due to the fact that usually women spend more time on domestic affairs and the official length of their working day is less.

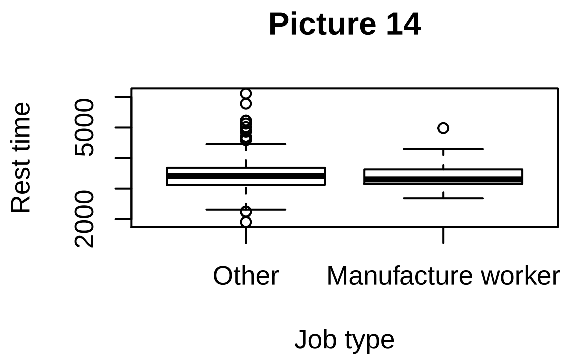
Married people on average devote more time to total relax. Maybe because they have to make a choice between work and family (picture 10).

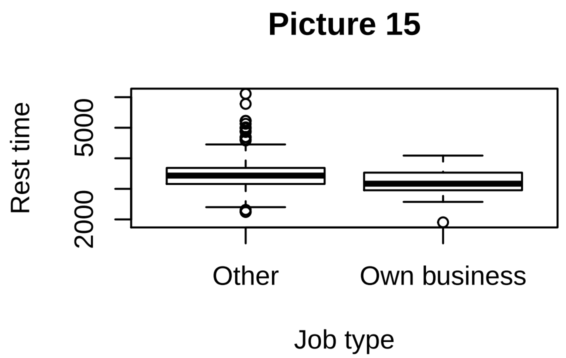
People with kids have a longer rest time on average which is also logical because children less 3 years old need more attention and care (picture 11).

Finally, African Americans on average have more rest than Europeans (picture 12).

However, all differences are small.





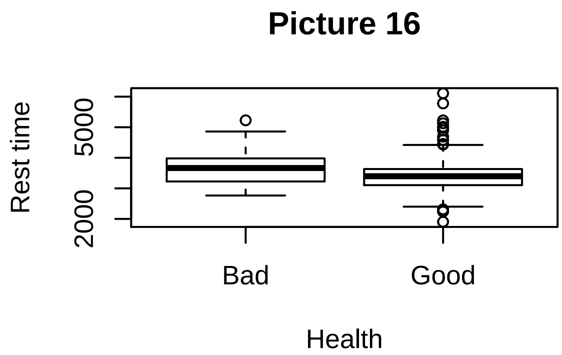


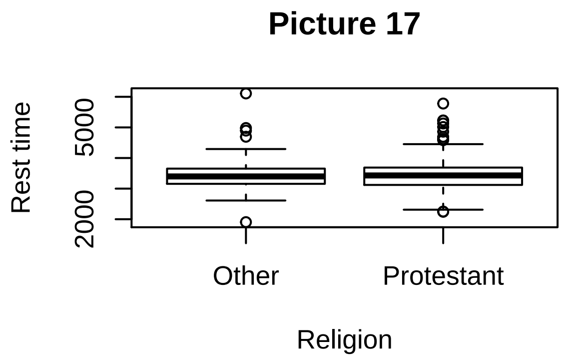
Average rest time of clerks is more than others' time for relax (picture 13).

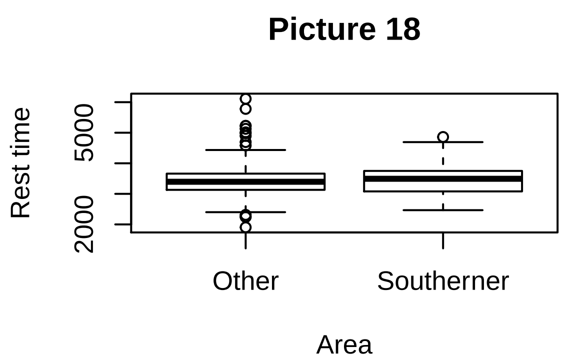
Picture 14 shows that manufacture workers on average rest less than other categories of people.

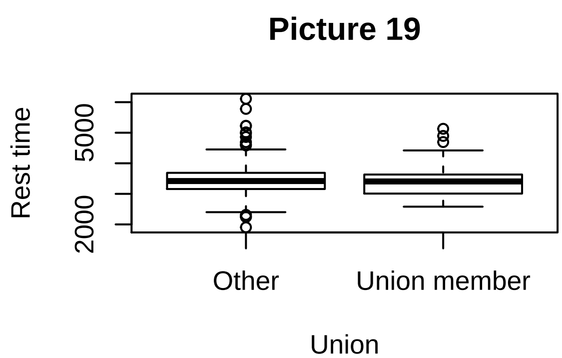
Businessmen on average relax less than others (picture 15).

All this results are easily explained with features of these professions.









Picture 16 shows that average total rest time of people with good health is less than time spent relaxing by people with bad health.

Protestants on average devote more time to total relax (picture 17).

Southerners have a longer rest time on average than others (picture 18).

Finally, union members on average relax less than other (picture 19).

Model

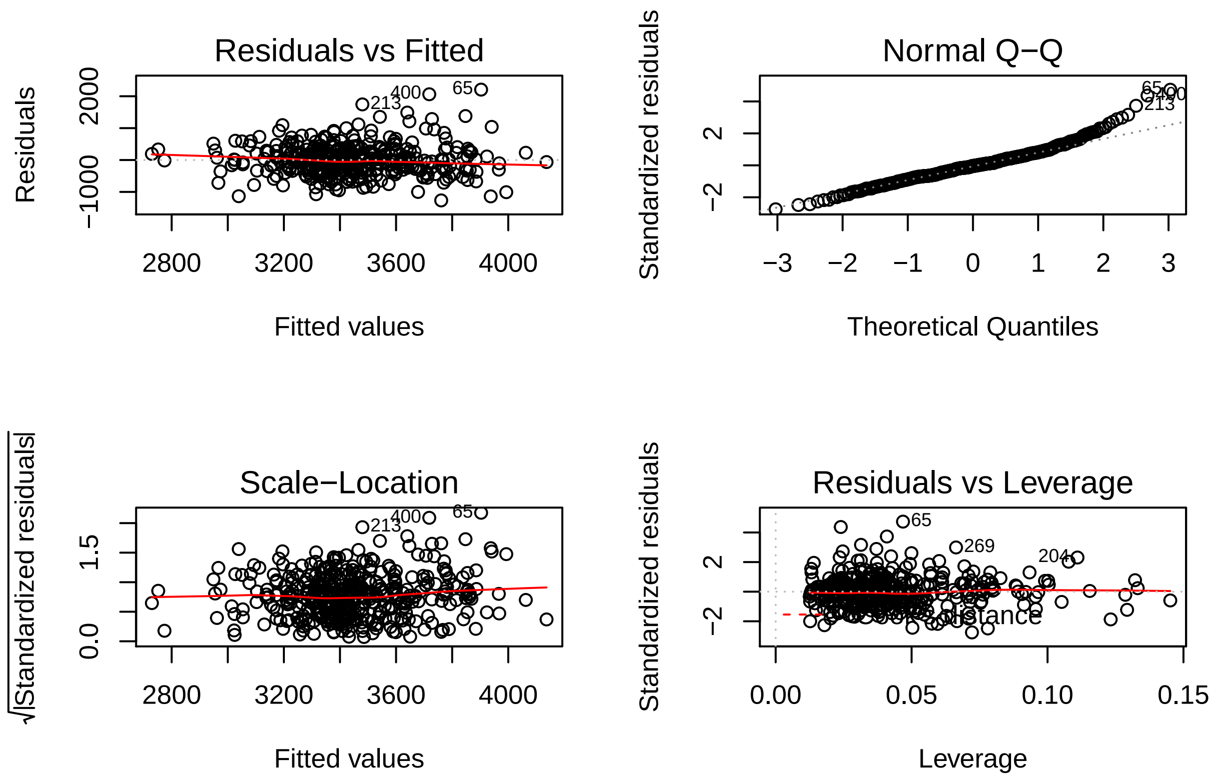
There are four principal assumptions which justify the use of linear regression models for purposes of inference or prediction:

1. Linearity and additivity of the relationship between dependent and independent variables ($residuals$ $vs\ fitted$).
2. Normality of the error distribution ($normality$).
3. Homoscedasticity (constant variance) of the errors ($scale - location$).
4. Statistical independence of the errors (in particular, no correlation between consecutive errors in the case of time series data) ($residuals\ vs\ leverage$).

Based on above data, we can assume that our basic model looks like:

$rlxall = totwrk + lhrwage + earns74 + age + agesq + male + marr + yngkid + black + clerical + construc + selfe + gdhlth + south + union + prot$

We can check conditions 1-4 of linear model using the graph.



We can see 3 blowouts: 65th, 213th, 400th observations. These respondents have much more time spent relaxing then others. We will not get rid from them because it can spoil owerall picture.

1. Residuals vs Fitted.

This plot shows if residuals have non-linear patterns. There could be a non-linear relationship between predictor variables and an outcome variable and the pattern could show up in this plot if the model does not capture the non-linear relationship. If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you do not have non-linear relationships.

We do not see any distinctive pattern in our case.

2. Normal Q-Q.

This plot shows if residuals are normally distributed. It is good if residuals are lined well on the straight dashed line.

Everything is OK.

3. Scale-Location.

This plot shows if residuals are spread equally along the ranges of predictors.It is good if you see a horizontal line with equally (randomly) spread points.

In our case the residuals appear randomly spread.

4. Residuals vs Leverage.

This plot helps us to find influential subjects if any. Not all outliers are influential in linear regression analysis (whatever outliers mean). Even though data have extreme values, they might not be influential to determine a regression line. That means, the results would not be much different if we either include or exclude them from analysis. They follow the trend in the majority of cases and they do not really matter; they are not influential. On the other hand, some cases could be very influential even if they look to be within a reasonable range of the values. They could be extreme cases against a regression line and can alter the results if we exclude them from analysis. Another way to put it is that they don't get along with the trend in the majority of the cases.

Our case is the typical look when there is no influential cases. All cases are well inside of the Cook's distance lines.

Summary shows that in such regression variables $totwrk$, $gdhth$ and $selfe$ are statistically significant.

Then we use Wald test and compare models from basic to the easiest.

```
## Wald test
##
## Model 1: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     yngkid + black + clerical + construc + selfe + gdhlth + south +
##     union
## Model 2: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     yngkid + black + clerical + construc + selfe + gdhlth + south +
##     union + prot
##   Res.Df Df      F Pr(>F)
## 1    389
## 2    388  1 0.0289  0.865
```

Pr(>F)=0.865 > 0.05 -> we should accept hypothesis Ho: true Model 1. However, F=0.0357 -> differences between the models are very small.

```
## Wald test
##
## Model 1: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     yngkid + black + clerical + construc + selfe + gdhlth + south
## Model 2: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     yngkid + black + clerical + construc + selfe + gdhlth + south +
##     union
##   Res.Df Df      F Pr(>F)
## 1    390
## 2    389  1 1.2863 0.2574
```

Pr(>F)=0.2574 > 0.05 -> we should accept hypothesis Ho: true Model 1.

```
## Wald test
##
## Model 1: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     yngkid + black + clerical + construc + selfe + gdhlth
## Model 2: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     yngkid + black + clerical + construc + selfe + gdhlth + south
##   Res.Df Df      F Pr(>F)
## 1    391
## 2    390  1 1.4538 0.2287
```

We accept hypothesis Ho: true Model 1.

```
## Wald test
##
## Model 1: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     yngkid + black + clerical + construc + selfe
## Model 2: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     yngkid + black + clerical + construc + selfe + gdhlth
##   Res.Df Df      F  Pr(>F)
## 1    392
## 2    391  1 4.0481 0.04491 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Please note that removing a significant variable we get opposite result of Wald test - we reject hypothesis Ho: true Model 1. We should try to reject other variables.

```
## Wald test
##
## Model 1: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     yngkid + black + clerical + construc + gdhlth
## Model 2: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     yngkid + black + clerical + construc + selfe + gdhlth
##   Res.Df Df      F  Pr(>F)
## 1    392
## 2    391  1 4.3889 0.03682 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pr(>F) = 0.03682 < 0.05 -> true model 2. We cannot reject variable $selfe$.

```
## Wald test
##
## Model 1: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     yngkid + black + clerical + selfe + gdhlth
## Model 2: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     yngkid + black + clerical + construc + selfe + gdhlth
##   Res.Df Df      F Pr(>F)
## 1    392
## 2    391  1 0.1505 0.6983
```

Model 1 is true.

```
## Wald test
##
## Model 1: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     yngkid + black + selfe + gdhlth
## Model 2: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     yngkid + black + clerical + selfe + gdhlth
##   Res.Df Df      F Pr(>F)
## 1    393
## 2    392  1 0.7775 0.3784
```

Pr(>F)=0.3784 > 0.05 -> we should accept hypothesis Ho: true Model 1.

```
## Wald test
##
## Model 1: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     yngkid + selfe + gdhlth
## Model 2: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     yngkid + black + selfe + gdhlth
##   Res.Df Df      F Pr(>F)
## 1    394
## 2    393  1 0.0037 0.9517
```

We can reject variable *black*.

```
## Wald test
##
## Model 1: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     selfe + gdhlth
## Model 2: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     yngkid + selfe + gdhlth
##   Res.Df Df      F Pr(>F)
## 1    395
## 2    394  1 0.0547 0.8151
```

We can reject variable *yngkid*.

```
## Wald test
##
## Model 1: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + selfe +
##     gdhlth
## Model 2: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + marr +
##     selfe + gdhlth
##   Res.Df Df      F Pr(>F)
## 1    396
## 2    395  1 2.6726 0.1029
```

We can reject variable *marr*.

```
## Wald test
##
## Model 1: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + selfe + gdhlth
## Model 2: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + selfe +
##     gdhlth
##   Res.Df Df      F  Pr(>F)
## 1    397
## 2    396  1 3.3765 0.06688 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We will not reject variable *male* as at 10% significance level [H0: Model 1 is true] can be rejected.

```
## Wald test
##
## Model 1: rlxall ~ totwrk + lhrwage + earns74 + age + male + selfe + gdhlth
## Model 2: rlxall ~ totwrk + lhrwage + earns74 + age + agesq + male + selfe +
##     gdhlth
##   Res.Df Df     F Pr(>F)
## 1    397
## 2    396  1 0.202 0.6534
```

Variable *agesq* can be rejected.

```
## Wald test
##
## Model 1: rlxall ~ totwrk + lhrwage + earns74 + male + selfe + gdhlth
## Model 2: rlxall ~ totwrk + lhrwage + earns74 + age + male + selfe + gdhlth
##   Res.Df Df      F  Pr(>F)
## 1    398
## 2    397  1 2.9726 0.08547 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We cannot reject variable *age* at 10% significance level.

```
## Wald test
##
## Model 1: rlxall ~ totwrk + lhrwage + age + male + selfe + gdhlth
## Model 2: rlxall ~ totwrk + lhrwage + earns74 + age + male + selfe + gdhlth
##   Res.Df Df      F Pr(>F)
## 1    398
## 2    397  1 0.3805 0.5377
```

We can reject variable *earns*74.

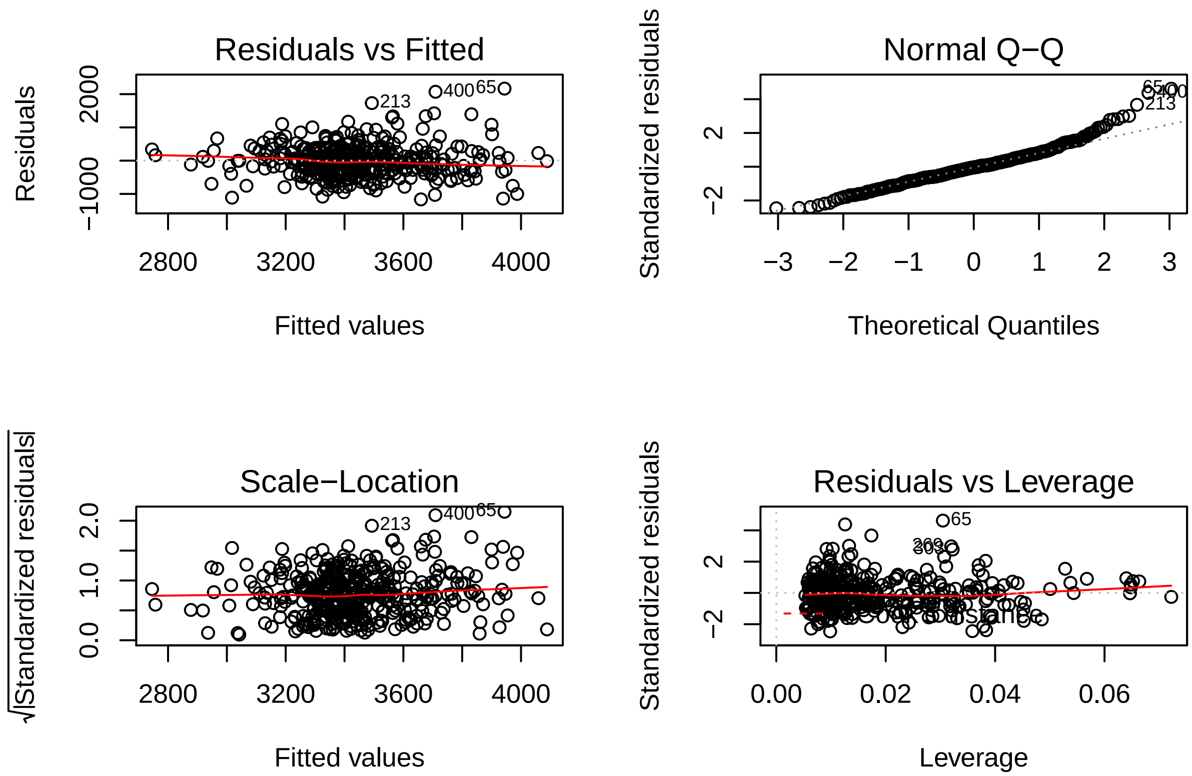
```
## Wald test
##
## Model 1: rlxall ~ lhrwage + age + male + selfe + gdhlth
## Model 2: rlxall ~ totwrk + lhrwage + age + male + selfe + gdhlth
##   Res.Df Df      F    Pr(>F)
## 1    399
## 2    398  1 48.129 1.625e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pr(>F)=1.625e-11 $<$ 0.05 -> we should reject hypothesis Ho: true Model 1 (reg15).

```
## Wald test
##
## Model 1: rlxall ~ totwrk + age + male + selfe + gdhlth
## Model 2: rlxall ~ totwrk + lhrwage + age + male + selfe + gdhlth
##   Res.Df Df      F  Pr(>F)
## 1    399
## 2    398  1 2.9696 0.08562 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Also we cannot reject variable $hrwage$ because at 10% significance level true model 2.

So the best model for us: $rlxall = totwrk + lhrwage + age + male + selfe + gdhlth$.



Graphs show that this model is also OK.

Ramsey test:

```
##
##  RESET test
##
## data:  reg14
## RESET = 0.34001, df1 = 2, df2 = 396, p-value = 0.712
```

P-value = 0.712 > 0.05 -> we can accept Ho: no omitted variables.

Chow test:

Ho: No difference between rest time of people with good and bad health.

H1: Long model is true -> there is difference between people with good and bad health.

```
##
## =====================================================================
##                                    Dependent variable:
##                      ------------------------------------------------
##                                          rlxall
##                                (1)                      (2)
## ---------------------------------------------------------------------
```

```
## totwrk                      -0.159**               -0.205***
##                              (0.081)                (0.029)
##
## lhrwage                   -450.158**              -77.168*
##                            (214.603)              (43.264)
##
## age                        18.066***               3.890*
##                              (5.614)                (2.075)
##
## male                       379.431**              111.515*
##                            (182.905)              (57.678)
##
## selfe                      -290.883              -218.648**
##                            (388.614)              (90.747)
##
## gdhlth                      242.361
##                            (357.705)
##
## I(totwrk * gdhlth)          -0.046
##                              (0.086)
##
## I(lhrwage * gdhlth)        392.862*
##                            (219.039)
##
## I(age * gdhlth)           -16.877***
##                              (6.047)
##
## I(male * gdhlth)           -275.127
##                            (192.799)
##
## I(selfe * gdhlth)           67.026
##                            (399.520)
##
## Constant                 3,612.450***           3,792.603***
##                            (337.989)              (109.597)
##
## ------------------------------------------------------------------
## Observations                  405                    405
## R2                           0.183                  0.153
## Adjusted R2                  0.160                  0.143
## Residual Std. Error   472.757 (df = 393)      477.574 (df = 399)
## F Statistic         7.998*** (df = 11; 393) 14.464*** (df = 5; 399)
## ==================================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

```
## $`F-stat`
## [1] 2.362251
##
## $`P-value`
## [1] 0.02969833
```

Coefficients of *totwrk*, *lrwage*, *male* changed a lot. P-value (F-test) = 0.02969833 < 0.05 -> we can reject Ho.

Ho: No difference between rest time of people with good and bad health. H1: Long model is true -> there is difference between people with good and bad health.

```
##
## ===================================================================
##                                  Dependent variable:
##                     -----------------------------------------------
##                                        rlxall
##                               (1)                     (2)
## -------------------------------------------------------------------
## totwrk                     -0.189***               -0.182***
##                             (0.041)                 (0.027)
##
## lhrwage                     -16.960                 -37.577
##                            (68.814)                (38.861)
##
## age                          3.781                   3.051
##                             (2.997)                 (2.071)
##
## male                       434.984*
##                            (248.826)
##
## selfe                      -175.498               -194.024**
##                            (162.918)               (90.306)
##
## I(totwrk * male)            -0.039
##                             (0.059)
##
## I(lhrwage * male)          -108.385
##                            (90.582)
##
## I(age * male)                0.515
##                             (4.219)
##
## I(gdhlth * male)           -121.818
##                            (111.912)
##
## I(selfe * male)             -39.366
##                            (197.044)
##
## gdhlth                                             -147.866*
##                                                    (75.261)
##
## Constant                  3,701.184***            3,910.182***
##                            (151.347)               (127.326)
##
## -------------------------------------------------------------------
## Observations                  405                     405
## R2                           0.161                   0.154
## Adjusted R2                  0.140                   0.143
## Residual Std. Error   478.358 (df = 394)      477.502 (df = 399)
## F Statistic         7.578*** (df = 10; 394) 14.493*** (df = 5; 399)
## ===================================================================
## Note:                                 *p<0.1; **p<0.05; ***p<0.01
```

```
## $`F-stat`
## [1] 0.7148177
##
## $`P-value`
## [1] 0.6125989
```

Coefficients almost did not change. P-value (F-test) = 0.6125989 > 0.05 -> we can accept Ho -> short model is better.

```
##
## ===================================================================
##                                   Dependent variable:
##                     -----------------------------------------------
##                                         rlxall
##                               (1)                     (2)
## -------------------------------------------------------------------
## totwrk                     -0.215***               -0.204***
##                             (0.030)                 (0.029)
##
## lhrwage                    -86.103*                 -59.092
##                            (48.394)                (42.881)
##
## age                          3.388                   3.469*
##                             (2.153)                 (2.088)
##
## male                       122.180**                97.268*
##                            (60.117)                (57.484)
##
## selfe                      -256.973
##                            (586.614)
##
## I(totwrk * selfe)            0.083
##                             (0.096)
##
## I(lhrwage * selfe)          29.481
##                            (121.769)
##
## I(age * selfe)               5.941
##                             (8.569)
##
## I(gdhlth * selfe)          -355.392
##                            (368.906)
##
## I(male * selfe)            -138.041
##                            (248.088)
##
## gdhlth                                             -153.906**
##                                                     (75.375)
##
## Constant                 3,840.742***            3,907.357***
##                            (117.881)               (127.591)
##
## -------------------------------------------------------------------
## Observations                  405                     405
```

```
## R2                           0.159                   0.150
## Adjusted R2                  0.138                   0.139
## Residual Std. Error   478.895 (df = 394)      478.542 (df = 399)
## F Statistic         7.473*** (df = 10; 394) 14.083*** (df = 5; 399)
## ======================================================================
## Note:                                      *p<0.1; **p<0.05; ***p<0.01
```

```
## $`F-stat`
## [1] 0.8826427
##
## $`P-value`
## [1] 0.4926465
```

Coefficients almost did not change.

P-value (F-test) = 0.4926465 > 0.05 -> we can accept Ho -> short model is better.

So we can modify our model $rlxall = totwrk + lhrwage + age + male + selfe + gdhlth$.

Now it is $rlxall = totwrk + lhrwage + age + male + selfe + gdhlth + I(age * gdhlth) + I(lhrwage * gdhlth)$.

```
##
## Call:
## lm(formula = rlxall ~ totwrk + lhrwage + age + male + selfe +
##     gdhlth + I(age * gdhlth) + I(lhrwage * gdhlth))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1152.65  -287.22   -20.51   265.98  2153.06
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3712.5599   329.7391  11.259  < 2e-16 ***
## totwrk                -0.2005     0.0285  -7.035 8.82e-12 ***
## lhrwage             -280.5881   176.3843  -1.591  0.11246
## age                   14.7368     5.1295   2.873  0.00429 **
## male                 130.2196    57.5084   2.264  0.02409 *
## selfe               -217.3216    89.9076  -2.417  0.01609 *
## gdhlth               129.6020   345.8764   0.375  0.70808
## I(age * gdhlth)      -13.4483     5.5824  -2.409  0.01645 *
## I(lhrwage * gdhlth)  212.8817   177.8711   1.197  0.23209
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 472.6 on 396 degrees of freedom
## Multiple R-squared:  0.1772, Adjusted R-squared:  0.1605
## F-statistic: 10.66 on 8 and 396 DF,  p-value: 1.357e-13
```

Summary tells that modified model is good -> p-value: 1.357e-13 < 0.05.

Again *reset.test* for new variables $I(age * gdhlth)$ and $I(lhrwage * gdhlth)$.

```
## Wald test
##
## Model 1: rlxall ~ totwrk + lhrwage + age + male + selfe + gdhlth
```

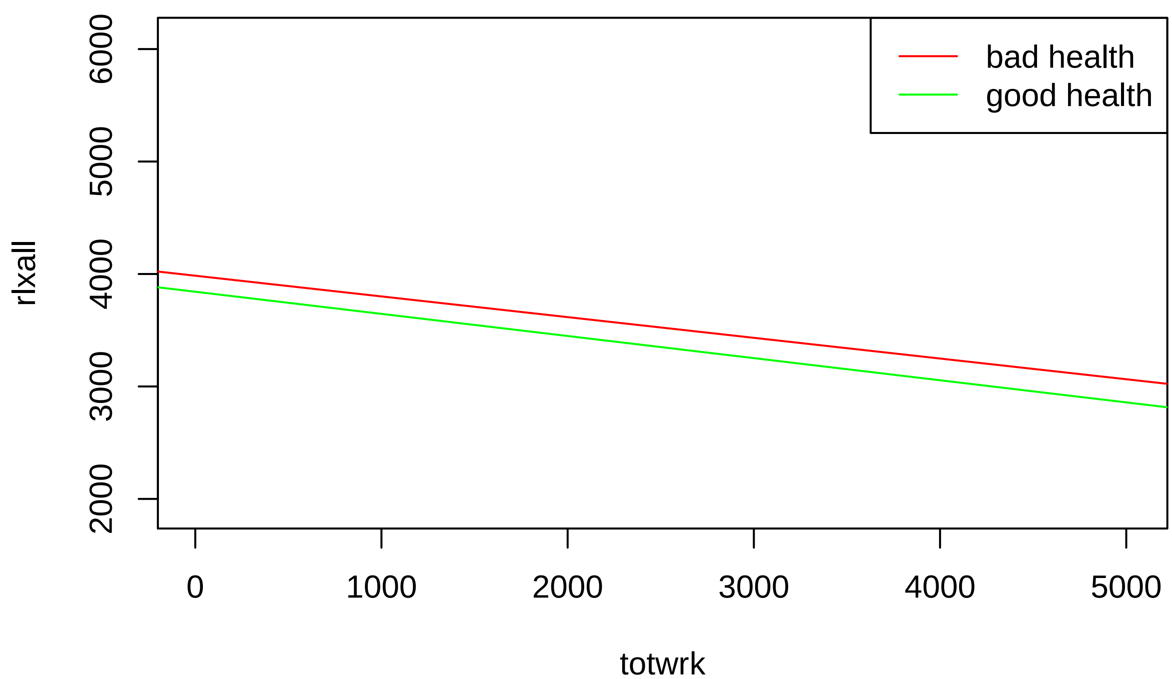
```
## Model 2: rlxall ~ totwrk + lhrwage + age + male + selfe + gdhlth + I(age *
##     gdhlth)
##   Res.Df Df      F  Pr(>F)
## 1    398
## 2    397  1 6.0586 0.01426 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value = 0.01426 -> true model 2.

```
## Wald test
##
## Model 1: rlxall ~ totwrk + lhrwage + age + male + selfe + gdhlth + I(age *
##     gdhlth)
## Model 2: rlxall ~ totwrk + lhrwage + age + male + selfe + gdhlth + I(age *
##     gdhlth) + I(lhrwage * gdhlth)
##   Res.Df Df      F Pr(>F)
## 1    397
## 2    396  1 1.4324 0.2321
```

True model $rlxall = totwrk + lhrwage + age + male + selfe + gdhlth + I(age * gdhlth)$.

We also can build a graph, to see whether the effect of $totwrk$ on $rlxall$ is different for people with good and bad health:



People with good health relax less than people with bad health.

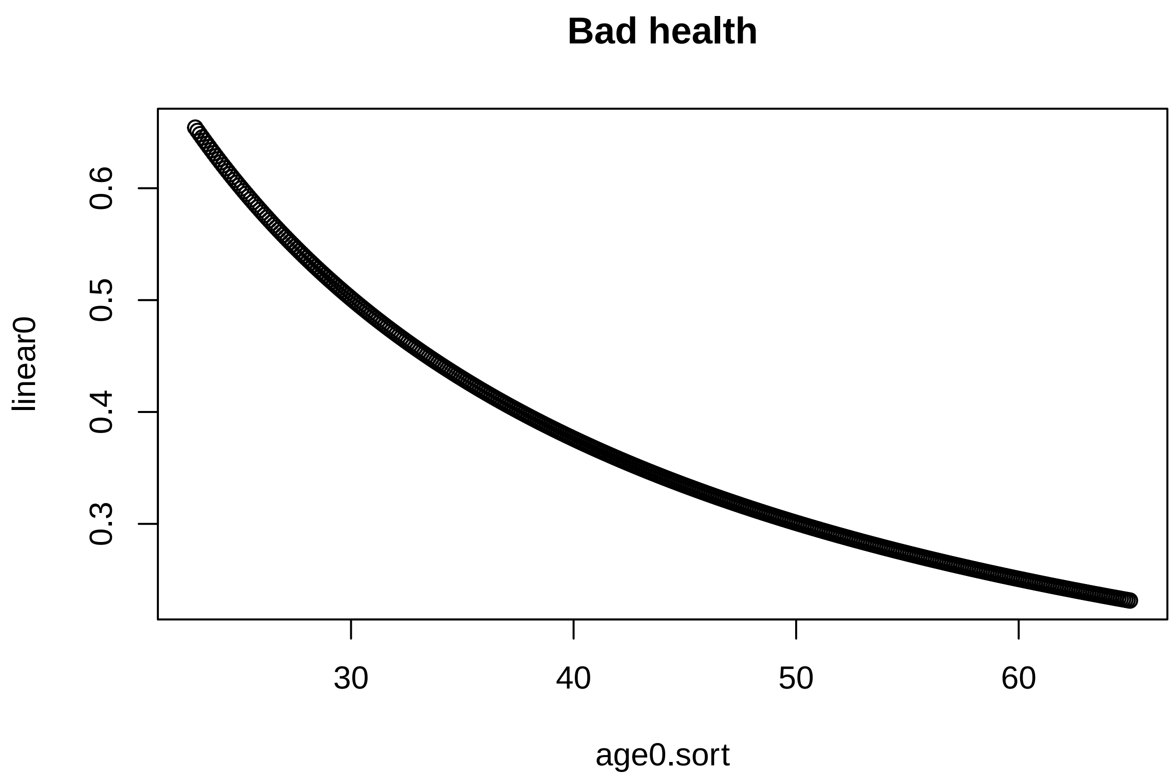
RE-test.

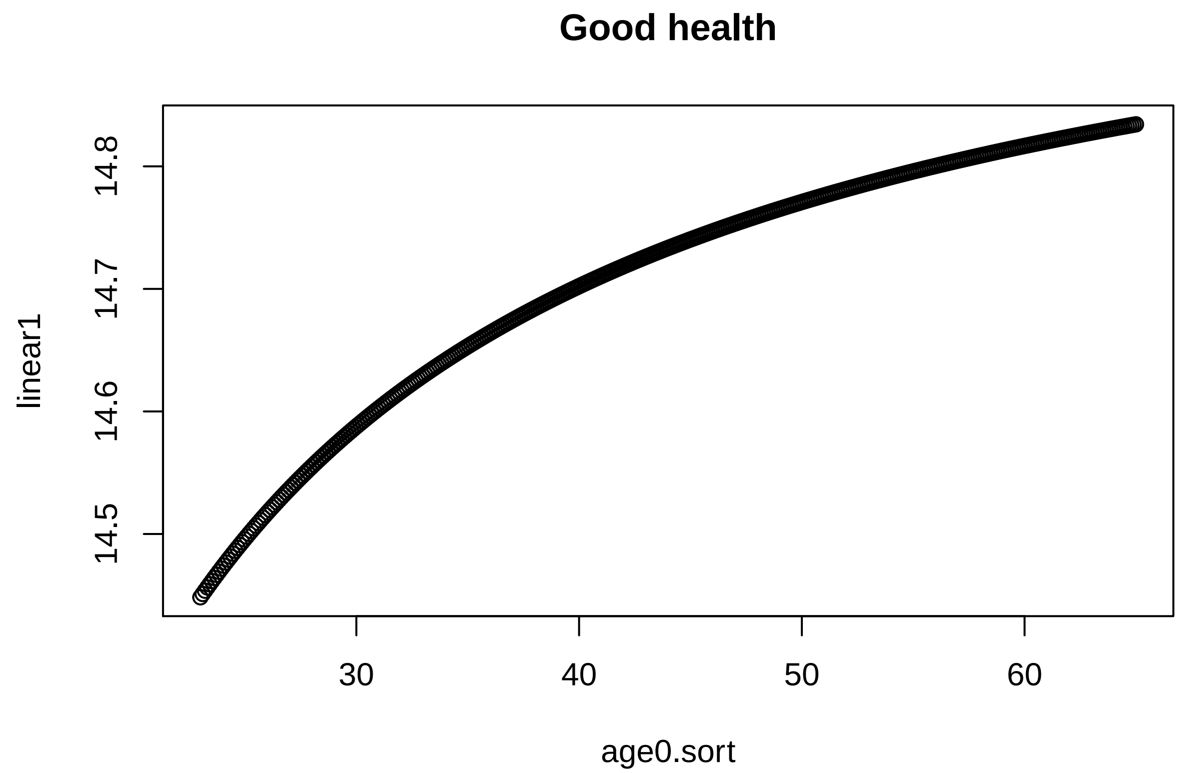
Ho: lin-lin model is true. Ho: log-lin model is true.

```
## PE test
##
## Model 1: log(rlxall) ~ totwrk + lhrwage + age + male + selfe + gdhlth +
##     I(age * gdhlth)
## Model 2: rlxall ~ totwrk + lhrwage + age + male + selfe + gdhlth + I(age *
##     gdhlth)
##                            Estimate Std. Error t value Pr(>|t|)
## M1 + fit(M2)-exp(fit(M1))       0.0        0.0 -0.4233   0.6723
## M2 + log(fit(M2))-fit(M1)  -8165.5     7780.3 -1.0495   0.2946
```

Both models are true. P-value $(lin - lin)>$ P-value $(log - lin)$ so $lin - lin$ model is better.

In addition, we can draw the graph directly to marginal effect of any variable.





With age ME for people with good health is growing, wor people with bad health - opposite situation.

Interpretation

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                               rlxall
## -----------------------------------------------
## totwrk                       -0.202***
##                               (0.028)
##
## lhrwage                      -75.808*
##                              (42.867)
##
## age                          15.046***
##                               (5.126)
##
## male                         127.689**
##                              (57.501)
##
## selfe                       -214.133**
##                              (89.917)
##
```

```
## gdhlth                         423.547*
##                                (243.669)
##
## I(age * gdhlth)                -13.735**
##                                 (5.580)
##
## Constant                      3,434.552***
##                                (234.159)
##
## -----------------------------------------------
## Observations                      405
## R2                               0.174
## Adjusted R2                      0.160
## Residual Std. Error     472.873 (df = 397)
## F Statistic           11.963*** (df = 7; 397)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

It's time to interpret our results about relationship between rest time and different variables for people in labor force.

Growth of *totwrk* by one minute per week gives drop of *rlxall* by 0.2 minutes per week.

Growth of *lhrwage* by one percent gives drop of *rlxall* by 0.76 minutes per week.

Growth of *age* by one year gives growth of *rlxall* by 15 minutes per week.

Men on average relax more then women by 128 minutes per week.

Businessmen relax less then others on average by 214 minutes per week.

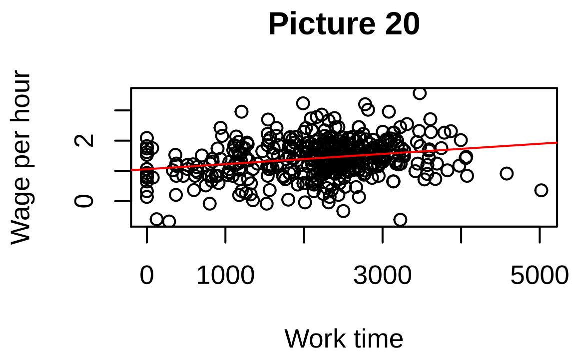
People with good health on average relax more then others by 423 minutes per week. This result is strange but OK.

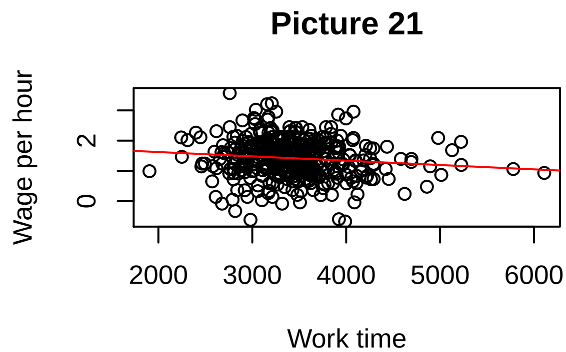
And each year of life of people with good health reduces their rest time by 13 minutes per week.

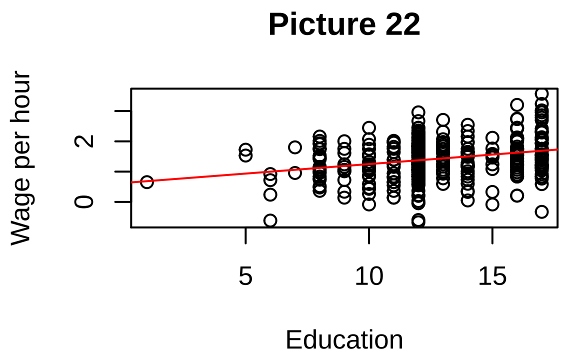
Section II

Desciptive statistics

Impact of timing on wage.







As can be seen from picture 20, there are positive linear relationship between time spent relaxing and work time.

Summary shows that actually work time and wage correlate with each other. The coefficient on the variable *totwrk* equals 1.680e-04 or 0.000168, what is more, it is high significant (***) relationship.

We reject hypothesis of equality of this coefficient to zero because p-value: 1.05e-06 (F-test).

As can be seen from picture 21, there are positive linear relationship between time spent relaxing and work time.

Summary shows that actually work time and wage correlate with each other. The coefficient on the variable *rlxall* equals -1.430e-04 or 0.000143, what is more, it is significant (*) relationship.

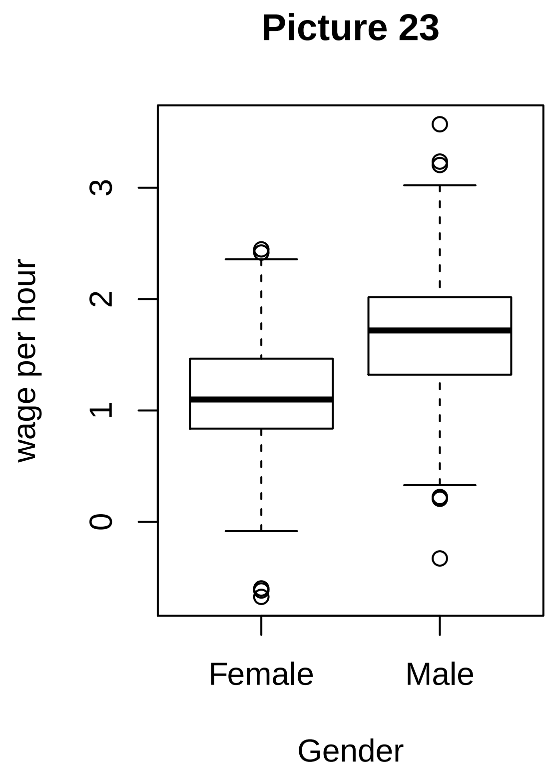
We reject hypothesis of equality of this coefficient to zero because p-value: 0.0195 (F-test).

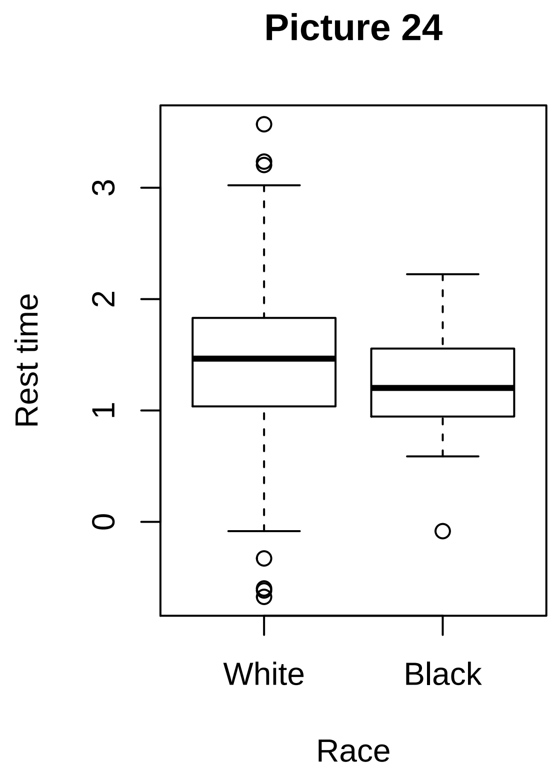
```
##   totwrk   rlxall
## 1.146792 1.146792
```

No problems with multicollinearity.

P-value: 5.829e-08 < 0.05 (F-test). *Educ* is high significant variable (***) (picture 22) -> accept in the model.

P-value: 0.00432 < 0.05 (F-test). *Age* and *agesq* - significant relationship (**) -> accept in the model.





Lhrwage of men on average is bigger than *lhrwage* of women. *Lhrwage* of European people is bigger than *lhrwage* of Arican Americans.

Model

Model: $lhrwage\ rlxall + totwrk + age + agesq + educ + male + black$.

Chow test.

```
##
## ===================================================================
##                                    Dependent variable:
##                      ----------------------------------------------
##                                          lhrwage
##                                (1)                      (2)
## -------------------------------------------------------------------
## totwrk                      0.0002***                0.0002***
##                             (0.00005)                (0.00003)
##
## rlxall                       0.00003                 -0.00002
##                              (0.0001)                (0.0001)
##
## educ                         0.061***                0.067***
##                              (0.015)                  (0.011)
##
## age                           0.041                  0.066***
```

```
##                            (0.027)                 (0.020)
##
## agesq                      -0.0005                 -0.001***
##                            (0.0003)                (0.0002)
##
## black                       -0.031                  -0.139
##                            (0.158)                 (0.136)
##
## male                        0.681
##                            (0.903)
##
## I(totwrk * male)          -0.0002***
##                            (0.0001)
##
## I(rlxall * male)           -0.0002*
##                            (0.0001)
##
## I(educ * male)              -0.006
##                            (0.020)
##
## I(age * male)               0.048
##                            (0.036)
##
## I(agesq * male)            -0.0005
##                            (0.0004)
##
## I(black * male)             -0.100
##                            (0.246)
##
## Constant                    -0.866                 -1.096**
##                            (0.670)                 (0.499)
##
## --------------------------------------------------------------------
## Observations                 405                     405
## R2                          0.354                   0.161
## Adjusted R2                 0.332                   0.149
## Residual Std. Error     0.520 (df = 391)        0.587 (df = 398)
## F Statistic         16.450*** (df = 13; 391) 12.773*** (df = 6; 398)
## ====================================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

```
## $`F-stat`
## [1] 16.59841
##
## $`P-value`
## [1] 0
```

There is difference between results for men and women.

```
##
## ====================================================================
##                                     Dependent variable:
##                       ----------------------------------------------
```

```
##                                    lhrwage
##                          (1)                      (2)
## -----------------------------------------------------------------
## totwrk                 0.00004                  0.00003
##                       (0.00003)                (0.00003)
##
## rlxall                 -0.0001                  -0.0001
##                        (0.0001)                 (0.0001)
##
## educ                   0.059***                 0.062***
##                        (0.010)                  (0.010)
##
## age                    0.071***                 0.068***
##                        (0.019)                  (0.018)
##
## agesq                 -0.001***                -0.001***
##                        (0.0002)                 (0.0002)
##
## black                   1.850
##                        (2.917)
##
## male                   0.566***                 0.564***
##                        (0.059)                  (0.057)
##
## I(totwrk * black)      -0.0002
##                        (0.0003)
##
## I(rlxall * black)      -0.0002
##                        (0.0003)
##
## I(educ * black)         0.041
##                        (0.048)
##
## I(age * black)          -0.078
##                        (0.133)
##
## I(agesq * black)        0.001
##                        (0.002)
##
## I(male * black)         -0.011
##                        (0.458)
##
## Constant               -1.055**                 -1.034**
##                        (0.459)                  (0.449)
##
## -----------------------------------------------------------------
## Observations             405                      405
## R2                      0.328                    0.324
## Adjusted R2             0.306                    0.314
## Residual Std. Error     0.530 (df = 391)        0.527 (df = 398)
## F Statistic         14.713*** (df = 13; 391) 31.780*** (df = 6; 398)
## =================================================================
## Note:                                 *p<0.1; **p<0.05; ***p<0.01
```

```
## $`F-stat`
## [1] 0.380995
##
## $`P-value`
## [1] 0.9133646
```

No difference between African American and European people. Short model is true.

New model: $lhrwage\ totwrk + rlxall + educ + age + agesq + black + male + I(totwrk * male) + I(rlxall * male)$

Waldtest:

```
## Wald test
##
## Model 1: lhrwage ~ totwrk + rlxall + educ + age + agesq + male + black +
##     I(totwrk * male)
## Model 2: lhrwage ~ totwrk + rlxall + educ + age + agesq + male + black +
##     I(totwrk * male) + I(rlxall * male)
##   Res.Df Df      F Pr(>F)
## 1    396
## 2    395  1 2.6282 0.1058
```

True short model because p-value = 0.1058.

```
## Wald test
##
## Model 1: lhrwage ~ totwrk + rlxall + educ + age + agesq + male + black
## Model 2: lhrwage ~ totwrk + rlxall + educ + age + agesq + male + black +
##     I(totwrk * male)
##   Res.Df Df      F  Pr(>F)
## 1    397
## 2    396  1 9.1428 0.00266 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

True long model. We cannot exclude $I(totwrk * male)$.

```
## Wald test
##
## Model 1: lhrwage ~ totwrk + rlxall + educ + age + agesq + male + I(totwrk *
##     male)
## Model 2: lhrwage ~ totwrk + rlxall + educ + age + agesq + male + black +
##     I(totwrk * male)
##   Res.Df Df      F Pr(>F)
## 1    397
## 2    396  1 0.2072 0.6492
```

Pr(>F)=0.6492 $>$ 0.05 -> we should accept hypothesis Ho: true Model 1 (reject *black*).

```
## Wald test
##
## Model 1: lhrwage ~ totwrk + rlxall + educ + age + agesq + I(totwrk * male)
## Model 2: lhrwage ~ totwrk + rlxall + educ + age + agesq + male + I(totwrk *
```

```
##     male)
##   Res.Df Df      F    Pr(>F)
## 1    398
## 2    397  1 44.663 7.944e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pr(>F)=7.944e-11 < 0.05 -> we should accept hypothesis Ho: true Model 2 (include *male*).

```
## Wald test
##
## Model 1: lhrwage ~ totwrk + rlxall + educ + age + male + I(totwrk * male)
## Model 2: lhrwage ~ totwrk + rlxall + educ + age + agesq + male + I(totwrk *
##     male)
##   Res.Df Df      F  Pr(>F)
## 1    398
## 2    397  1 10.086 0.00161 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

True $Model2$.

```
## Wald test
##
## Model 1: lhrwage ~ totwrk + rlxall + educ + agesq + male + I(totwrk *
##     male)
## Model 2: lhrwage ~ totwrk + rlxall + educ + age + agesq + male + I(totwrk *
##     male)
##   Res.Df Df      F    Pr(>F)
## 1    398
## 2    397  1 12.759 0.0003978 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

True $Model2$.

```
## Wald test
##
## Model 1: lhrwage ~ totwrk + rlxall + age + agesq + male + I(totwrk * male)
## Model 2: lhrwage ~ totwrk + rlxall + educ + age + agesq + male + I(totwrk *
##     male)
##   Res.Df Df      F    Pr(>F)
## 1    398
## 2    397  1 36.778 3.085e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

True $Model2$.

```
## Wald test
##
## Model 1: lhrwage ~ totwrk + educ + age + agesq + male + I(totwrk * male)
```

```
## Model 2: lhrwage ~ totwrk + rlxall + educ + age + agesq + male + I(totwrk *
##     male)
##   Res.Df Df      F Pr(>F)
## 1    398
## 2    397  1 1.2613 0.2621
```

Model 1 is better than Model 2.

```
## Wald test
##
## Model 1: lhrwage ~ educ + age + agesq + male + I(totwrk * male)
## Model 2: lhrwage ~ totwrk + educ + age + agesq + male + I(totwrk * male)
##   Res.Df Df     F   Pr(>F)
## 1    399
## 2    398  1 9.653 0.002026 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

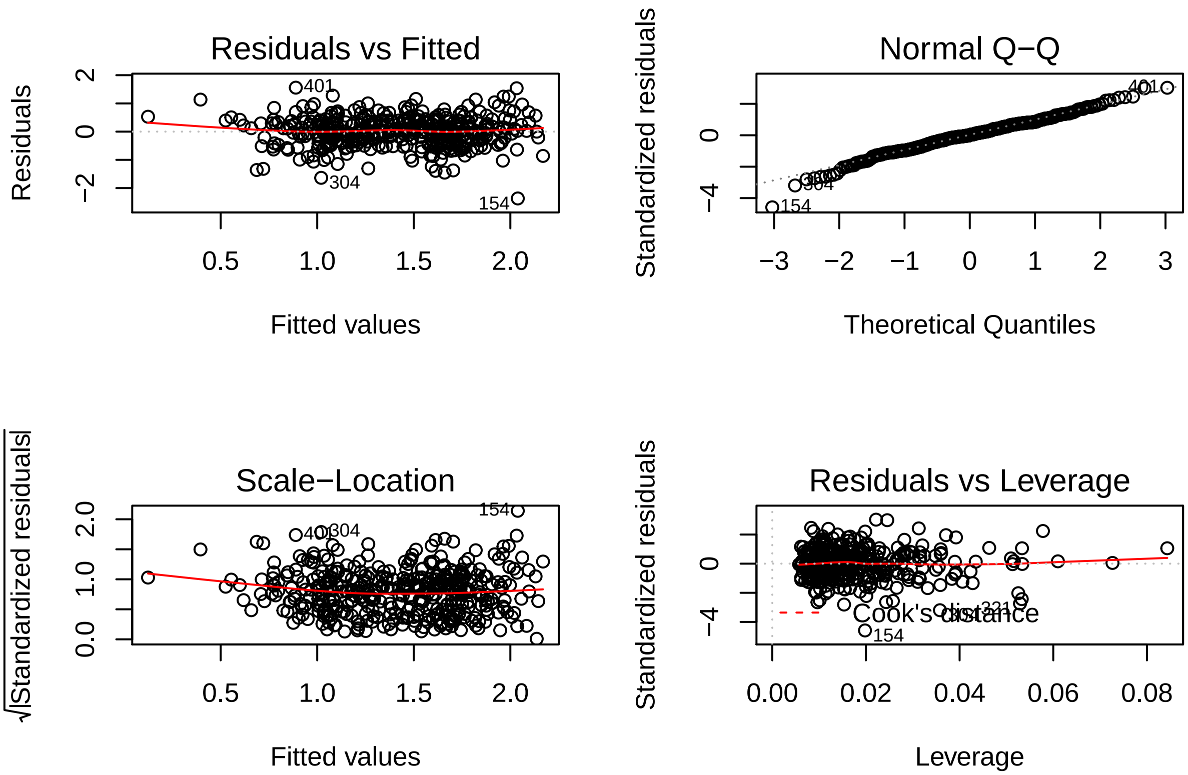
We will not exclude variable *totwrk* because it is important.

Our model will be:

$lhrwage = totwrk + educ + age + agesq + male + I(totwrk * male)$

```
##
## Call:
## lm(formula = lhrwage ~ totwrk + educ + age + agesq + male + I(totwrk *
##     male))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36701 -0.32222  0.00876  0.35237  1.55759
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.312e+00  3.819e-01  -3.436 0.000653 ***
## totwrk            1.344e-04  4.325e-05   3.107 0.002026 **
## educ              6.111e-02  9.945e-03   6.144 1.95e-09 ***
## age               6.528e-02  1.802e-02   3.622 0.000330 ***
## agesq            -6.953e-04  2.148e-04  -3.237 0.001311 **
## male              9.541e-01  1.441e-01   6.620 1.16e-10 ***
## I(totwrk * male) -1.862e-04  6.268e-05  -2.970 0.003154 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5216 on 398 degrees of freedom
## Multiple R-squared:  0.337,  Adjusted R-squared:  0.327
## F-statistic: 33.72 on 6 and 398 DF,  p-value: < 2.2e-16
```

We reject hypothesis of equality of all coefficients to zero because p-value: 2.2e-16 (F-test).



Graphs are good -> model is OK.

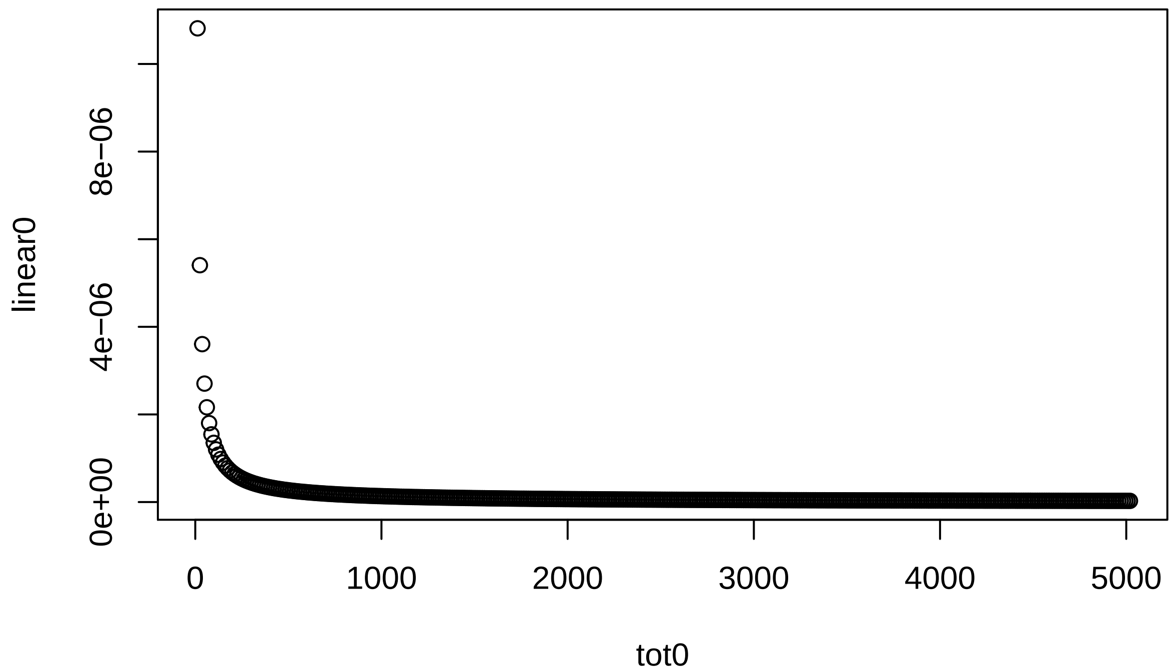
Ramsey test:

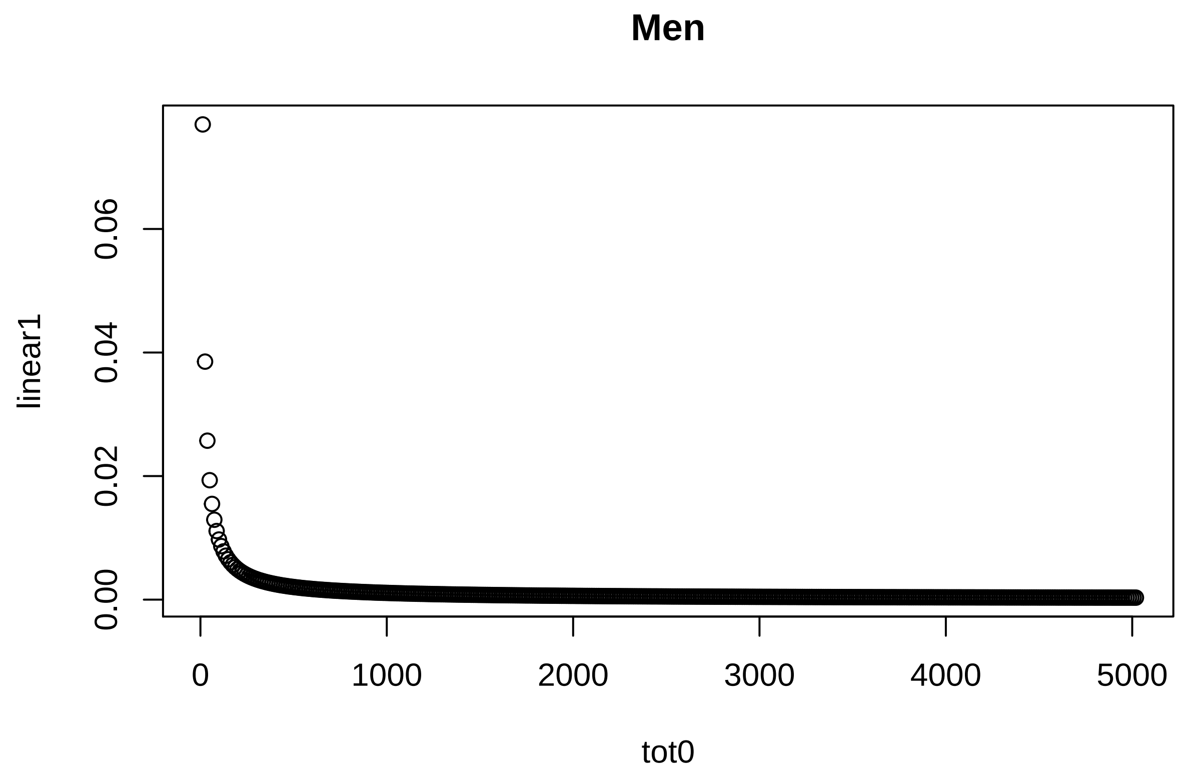
```
##
##  RESET test
##
## data:  reg8
## RESET = 1.6707, df1 = 2, df2 = 396, p-value = 0.1894
```

P-value = 0.1894 > 0.05 -> we can accept Ho: no omitted variables.

ME for male:

Women





ME is bigger for men.

Interpretation

```
##
## ===============================================
##                          Dependent variable:
##                      ---------------------------
##                                lhrwage
## -----------------------------------------------
## totwrk                        0.0001***
##                               (0.00004)
##
## educ                          0.061***
##                                (0.010)
##
## age                           0.065***
##                                (0.018)
##
## agesq                         -0.001***
##                               (0.0002)
##
## male                          0.954***
##                                (0.144)
##
```

```
## I(totwrk * male)             -0.0002***
##                               (0.0001)
##
## Constant                      -1.312***
##                                (0.382)
##
## -----------------------------------------------
## Observations                     405
## R2                              0.337
## Adjusted R2                     0.327
## Residual Std. Error      0.522 (df = 398)
## F Statistic           33.718*** (df = 6; 398)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Growth of *totwrk* by one minute per week gives growth of *lhrwage* by 0.01%.

Growth of *educ* by one year per week gives growth of *lhrwage* by 6.1%.

Men' salary is bigger by 95.4% then women' salary.

Salary increases with age up to a certain value and the falls.

Each minute of *totwork* gives for men grop of *hrwage* by 0.02%.

Conclusion.

In this work we made two models and interpreted them. We used descriptive statistics, different tests, made marginal effects.We verified that there are a lot of different models but it is important to choose the best one using different methods. Well-written model is irreplaceable for writing dofferent scientific papers.