

Deliverable 2

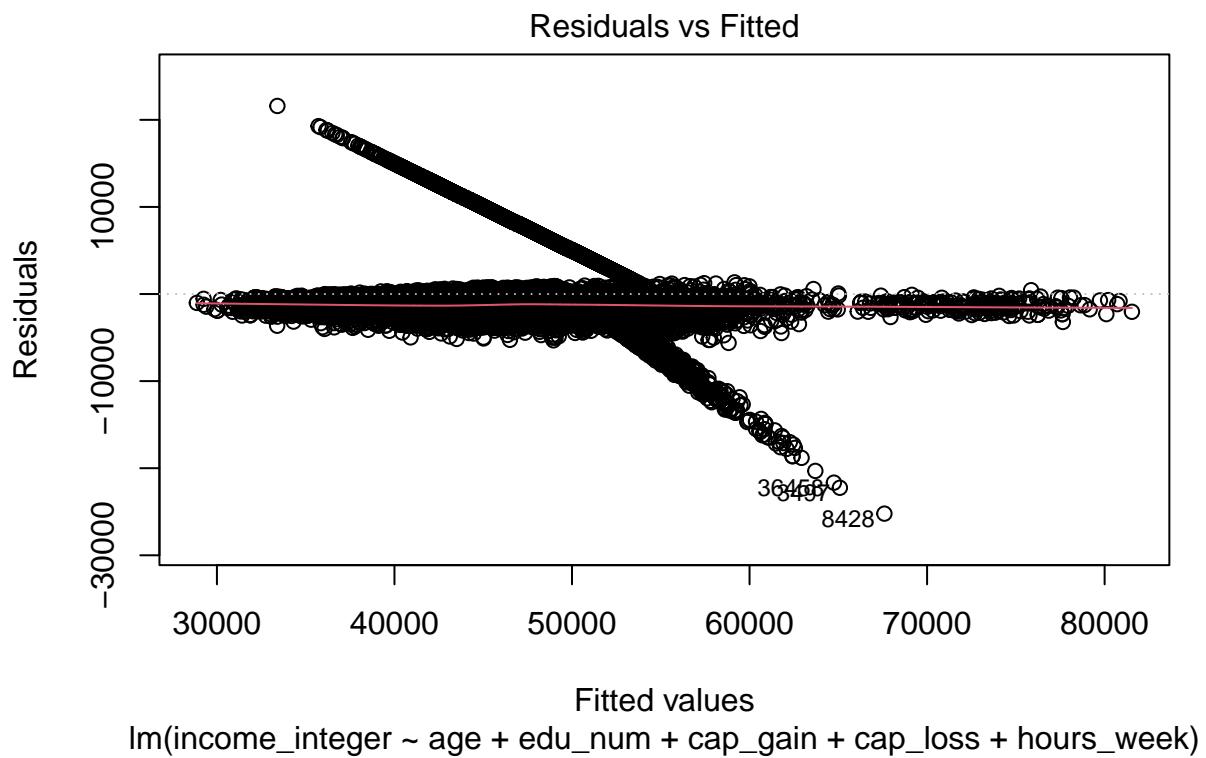
2025-04-27

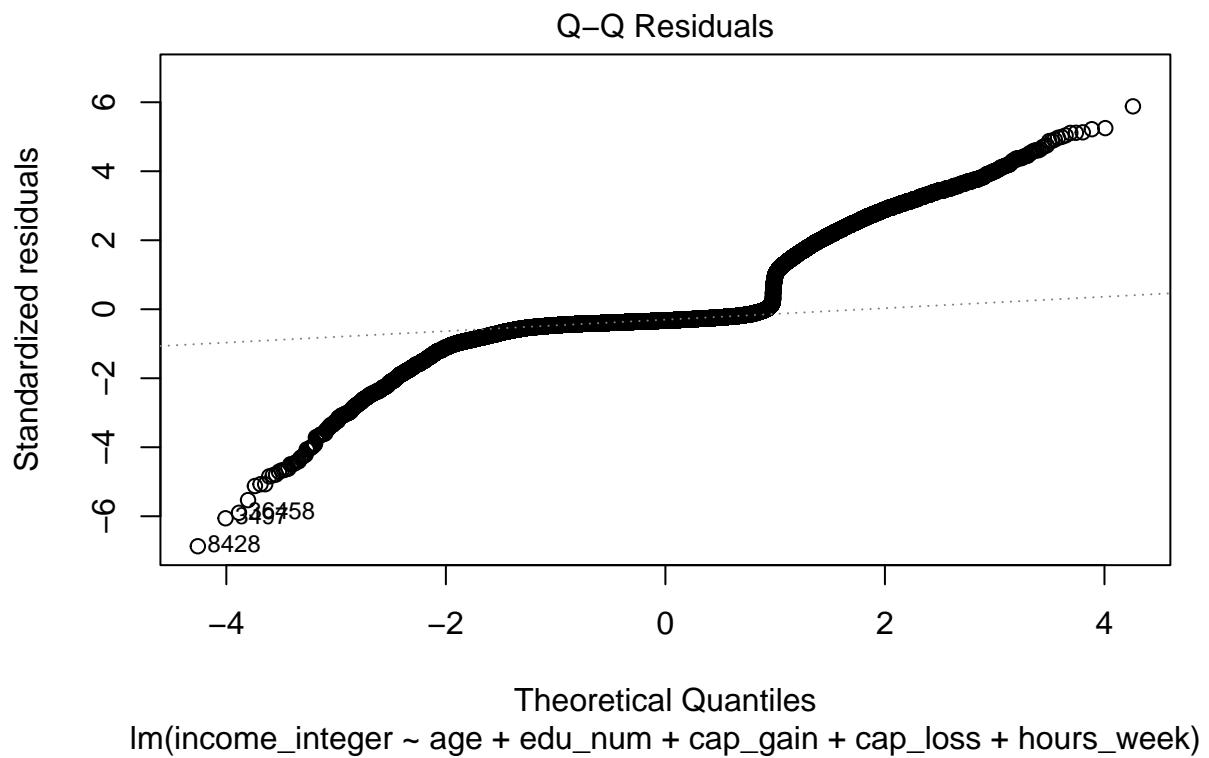
```
setwd("~/Escritorio/ADEI/D2")
dd <- read.csv("adult_def.csv")

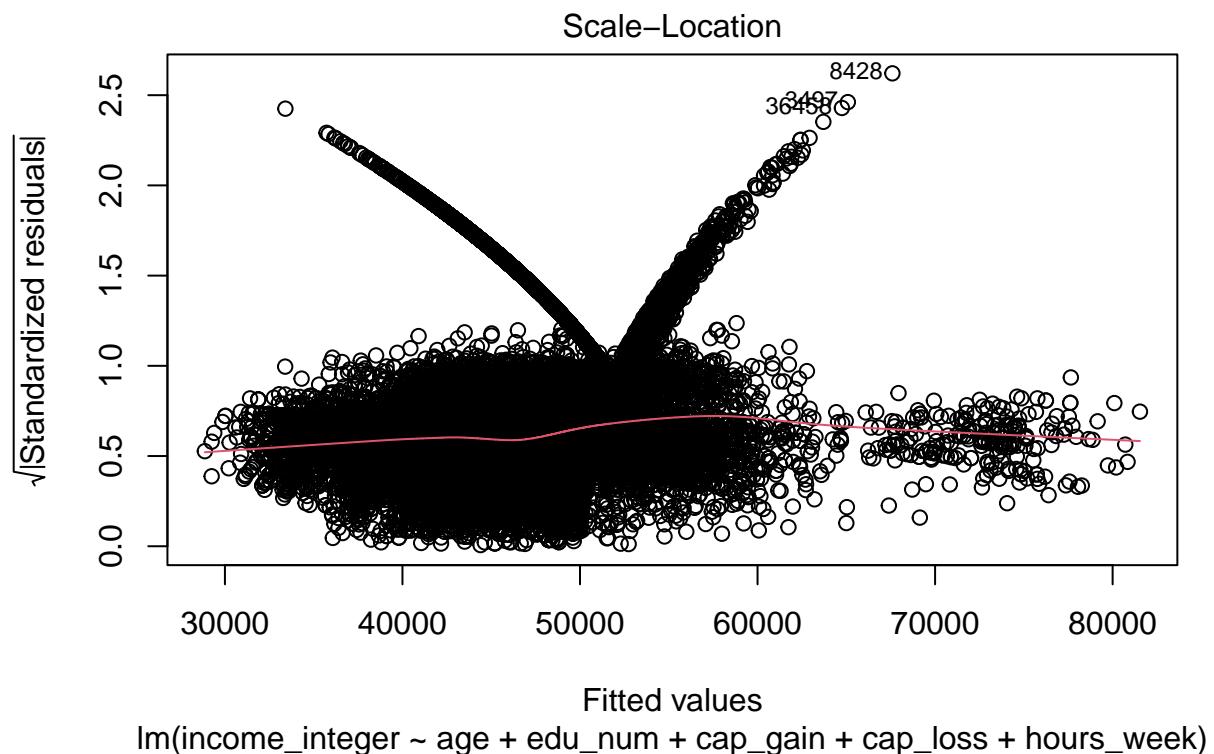
initial_model <- lm(income_integer ~ age + edu_num + cap_gain + cap_loss + hours_week, data = dd)
summary(initial_model)

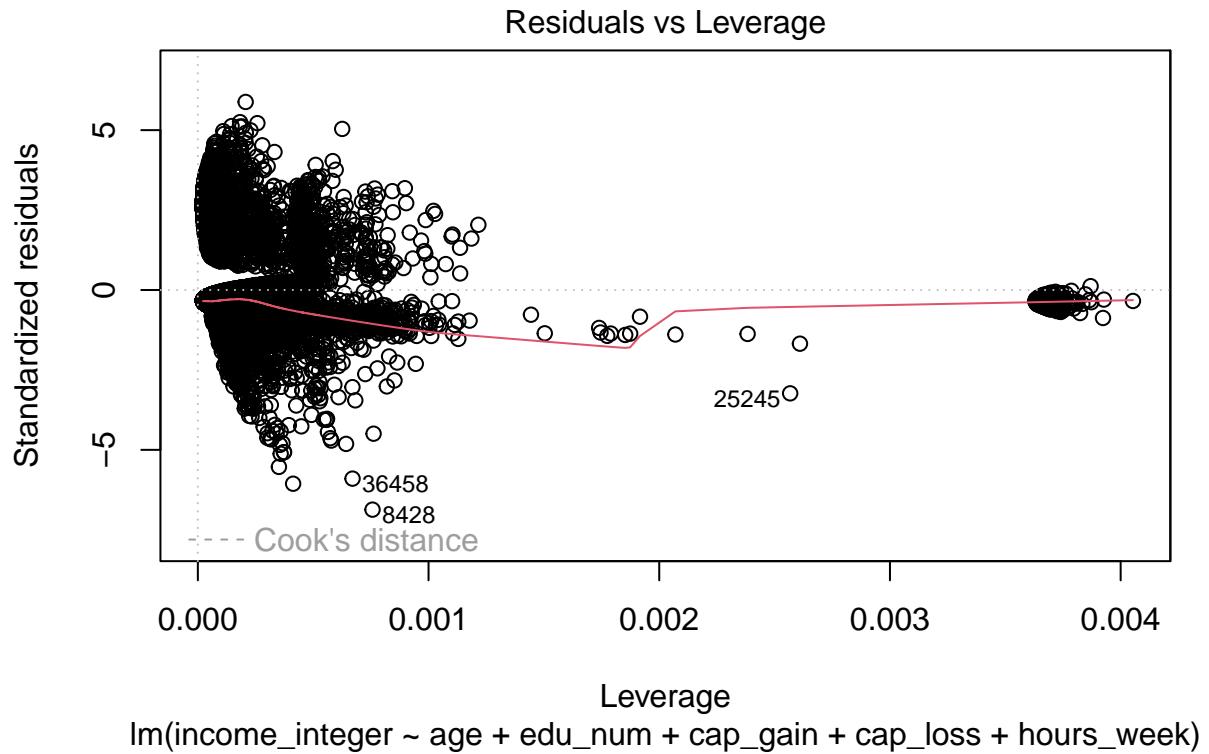
##
## Call:
## lm(formula = income_integer ~ age + edu_num + cap_gain + cap_loss +
##     hours_week, data = dd)
##
## Residuals:
##       Min      1Q      Median      3Q      Max
## -25216.0  -1522.3   -1201.8   -699.4  21594.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.091e+04  9.221e+01  226.74  <2e-16 ***
## age         2.317e+02  1.220e+00  189.87  <2e-16 ***
## edu_num     1.144e+03  6.595e+00  173.42  <2e-16 ***
## cap_gain    2.003e-01  2.260e-03   88.60  <2e-16 ***
## cap_loss    7.848e-01  4.151e-02   18.91  <2e-16 ***
## hours_week  9.924e+01  1.362e+00   72.87  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3672 on 48836 degrees of freedom
## Multiple R-squared:  0.6683, Adjusted R-squared:  0.6682
## F-statistic: 1.968e+04 on 5 and 48836 DF, p-value: < 2.2e-16

plot(initial_model)
```

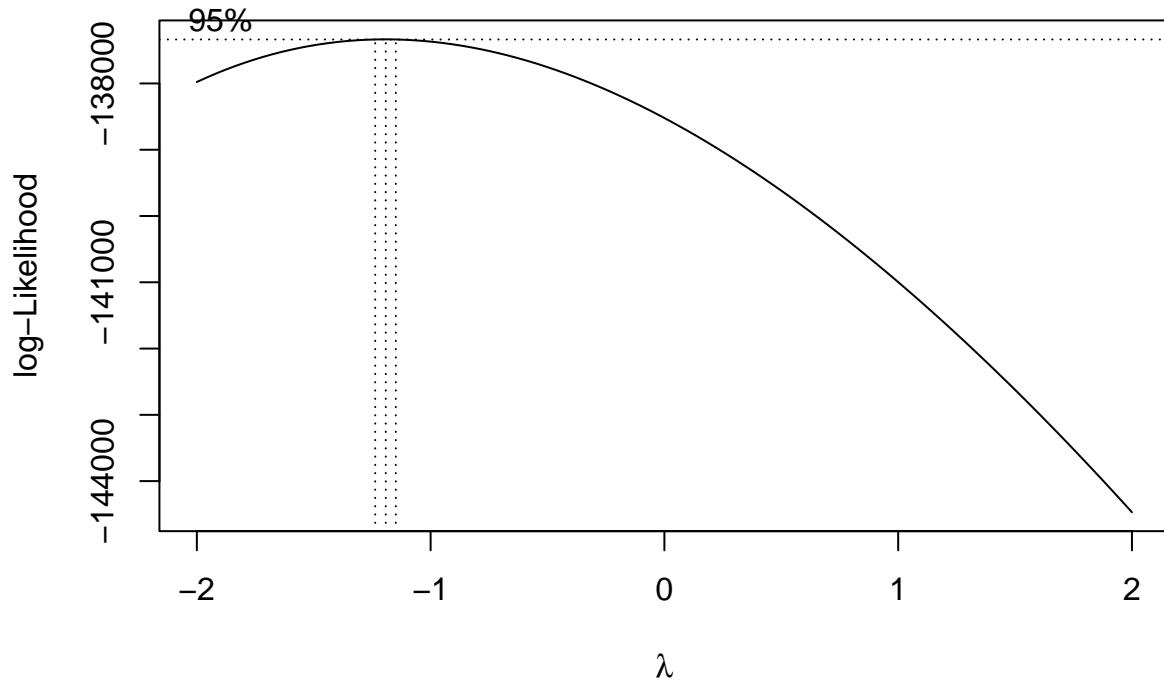








```
#target variable transformation, so the normality assumption is met
boxcox(income_integer ~ age + edu_num + cap_gain + cap_loss + hours_week, data = dd)
```

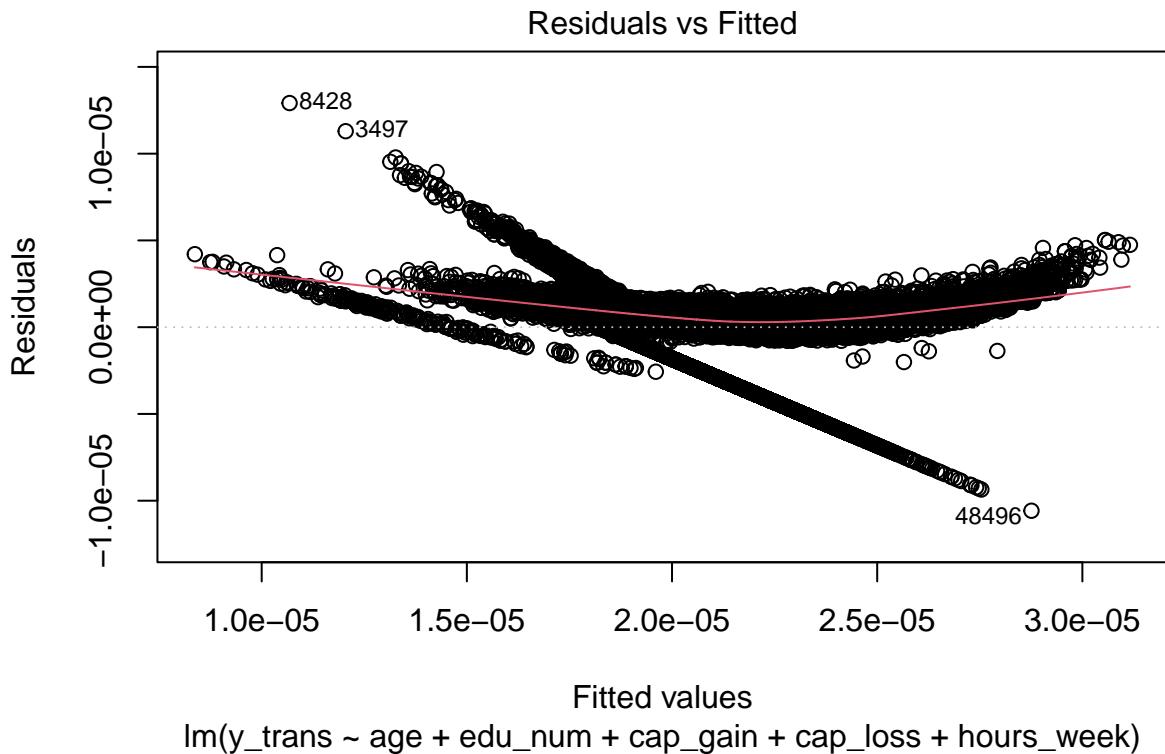


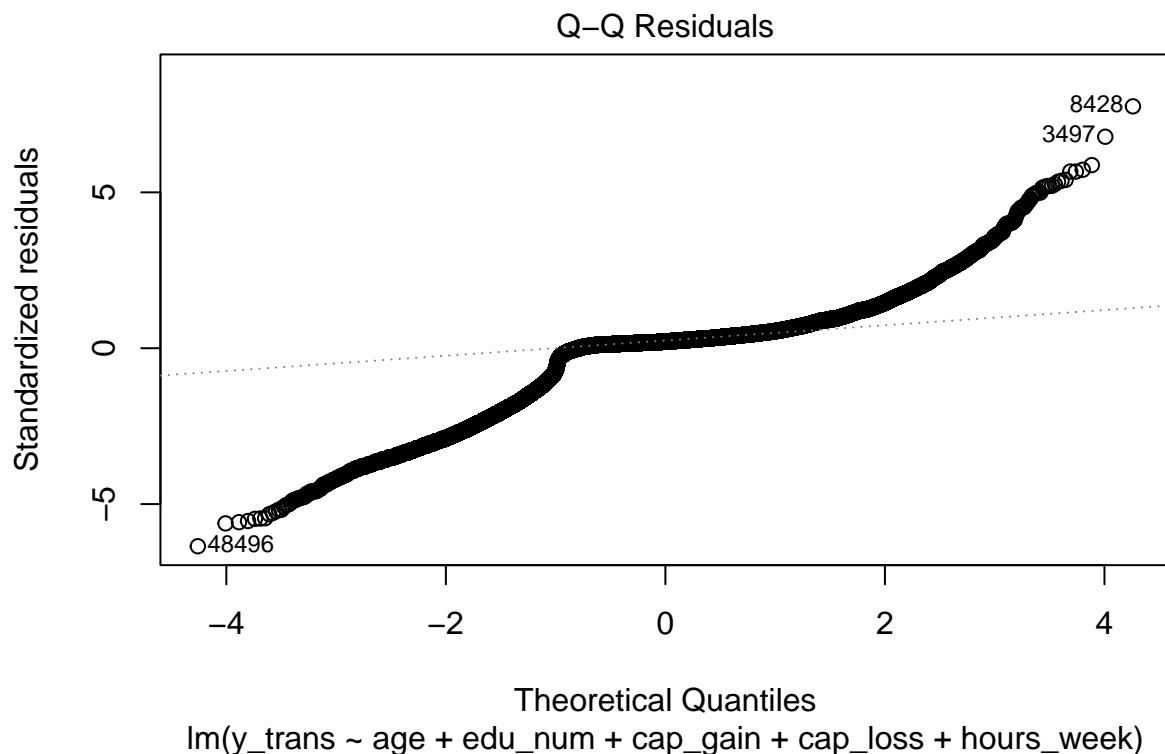
```
#given lambda is approximatedly -1 we do the inverse transformation
y_trans <- 1 / dd$income_integer
transformed_model <- lm(y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week, data = dd)
summary(transformed_model)

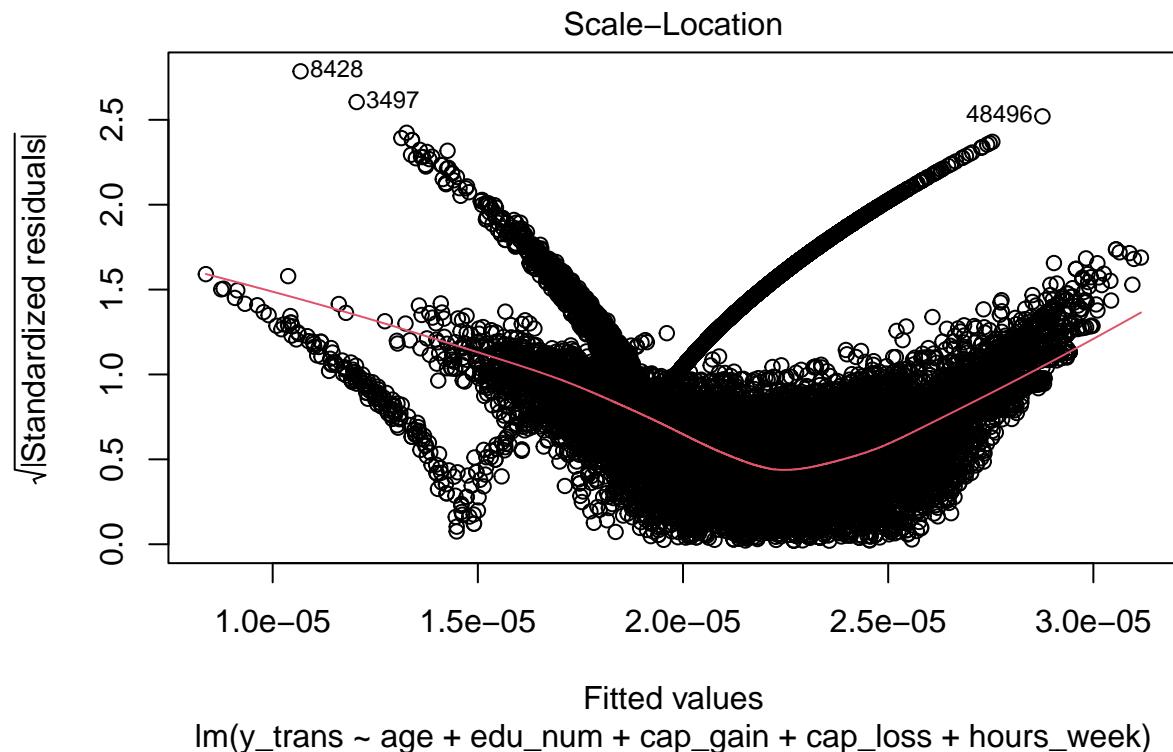
##
## Call:
## lm(formula = y_trans ~ age + edu_num + cap_gain + cap_loss +
##     hours_week, data = dd)
##
## Residuals:
##      Min        1Q    Median        3Q       Max
## -1.058e-05  1.373e-07  3.552e-07  6.859e-07  1.292e-05
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.536e-05  4.182e-08  845.57   <2e-16 ***
## age         -1.221e-07  5.533e-10 -220.74   <2e-16 ***
## edu_num     -6.086e-07  2.991e-09 -203.48   <2e-16 ***
## cap_gain    -5.503e-11  1.025e-12  -53.68   <2e-16 ***
## cap_loss    -2.612e-10  1.883e-11  -13.88   <2e-16 ***
## hours_week  -5.219e-08  6.176e-10  -84.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.665e-06 on 48836 degrees of freedom
```

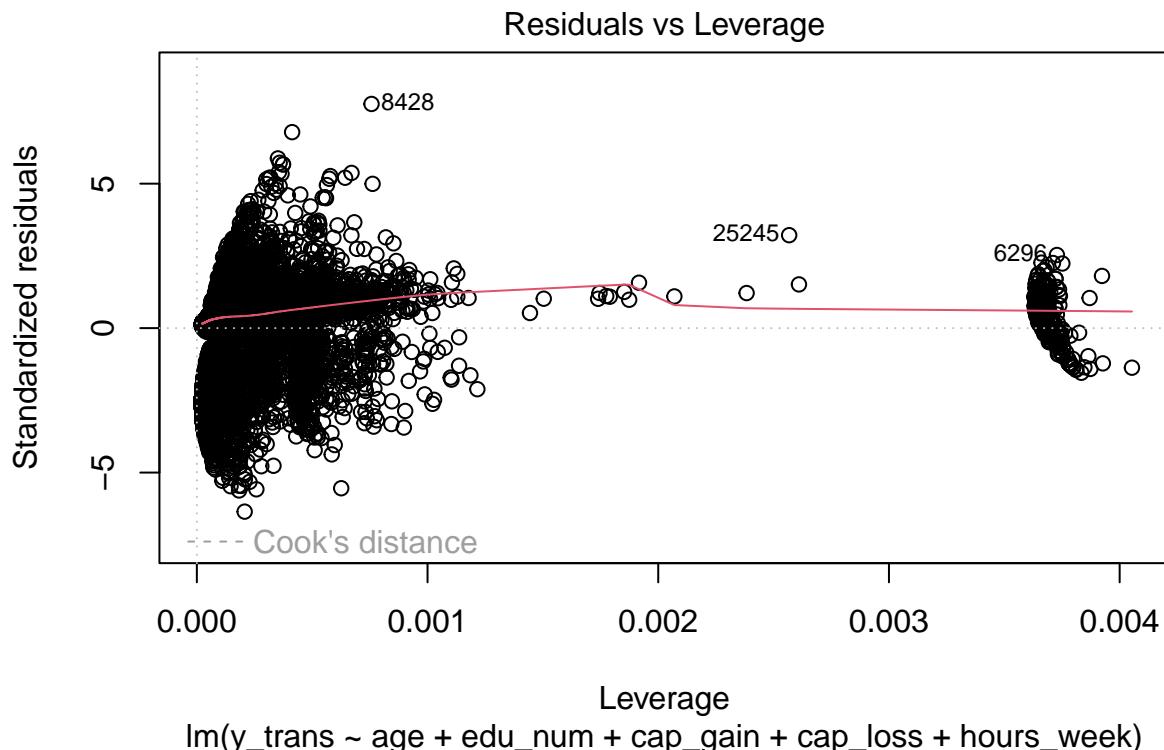
```
## Multiple R-squared:  0.7109, Adjusted R-squared:  0.7109  
## F-statistic: 2.402e+04 on 5 and 48836 DF,  p-value: < 2.2e-16
```

```
plot(transformed_model) #we cannot accept the basic hypothesis yet
```

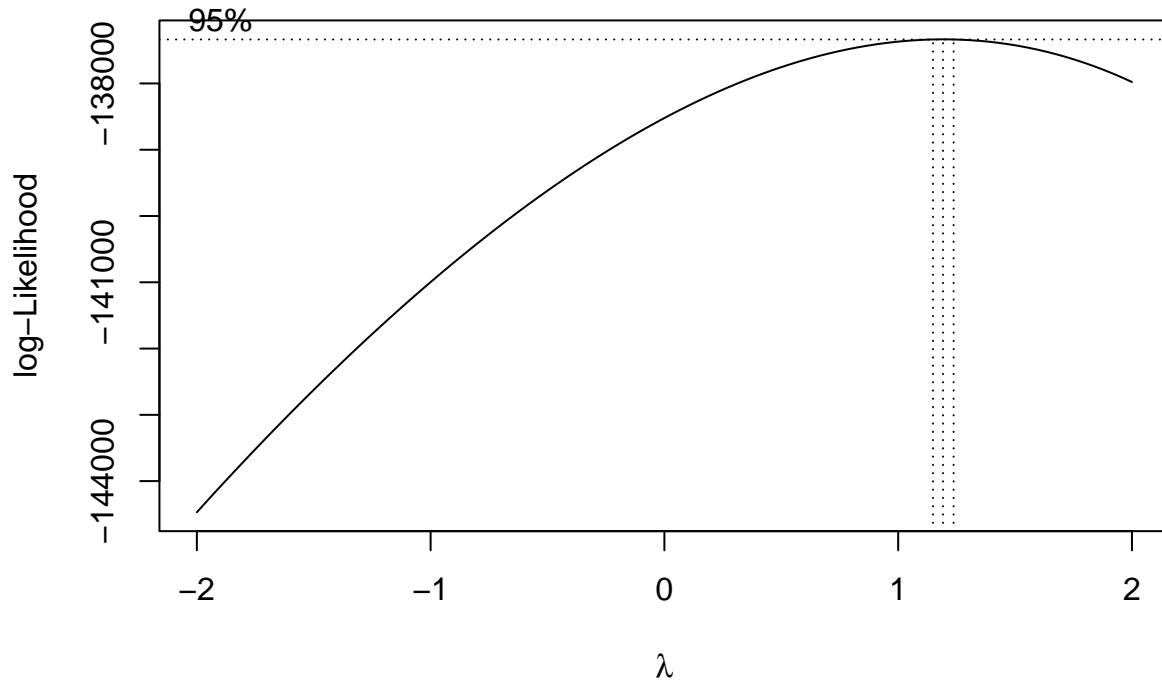








```
boxcox(y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week, data = dd) # lambda is now around 1
```



```
#As seen before, the basic hypothesis cannot be accepted, we need to perform transformation on the regr
boxTidwell(income_integer ~ age + edu_num + hours_week, data = dd)
```

```
## MLE of lambda Score Statistic (t) Pr(>|t|)
## age          -0.36944      -52.9070    <2e-16 ***
## edu_num       0.71757      -13.2951    <2e-16 ***
## hours_week   0.95940      -0.2224     0.824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  6
##
## Score test for null hypothesis that all lambdas = 1:
## F = 1034.7, df = 3 and 48835, Pr(>F) = < 2.2e-16

agebt <- 1 / sqrt(dd$age)
edu_num_bt <- sqrt(dd$edu_num)
btmodel <- lm(income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week, data = dd)
summary(btmodel)
```

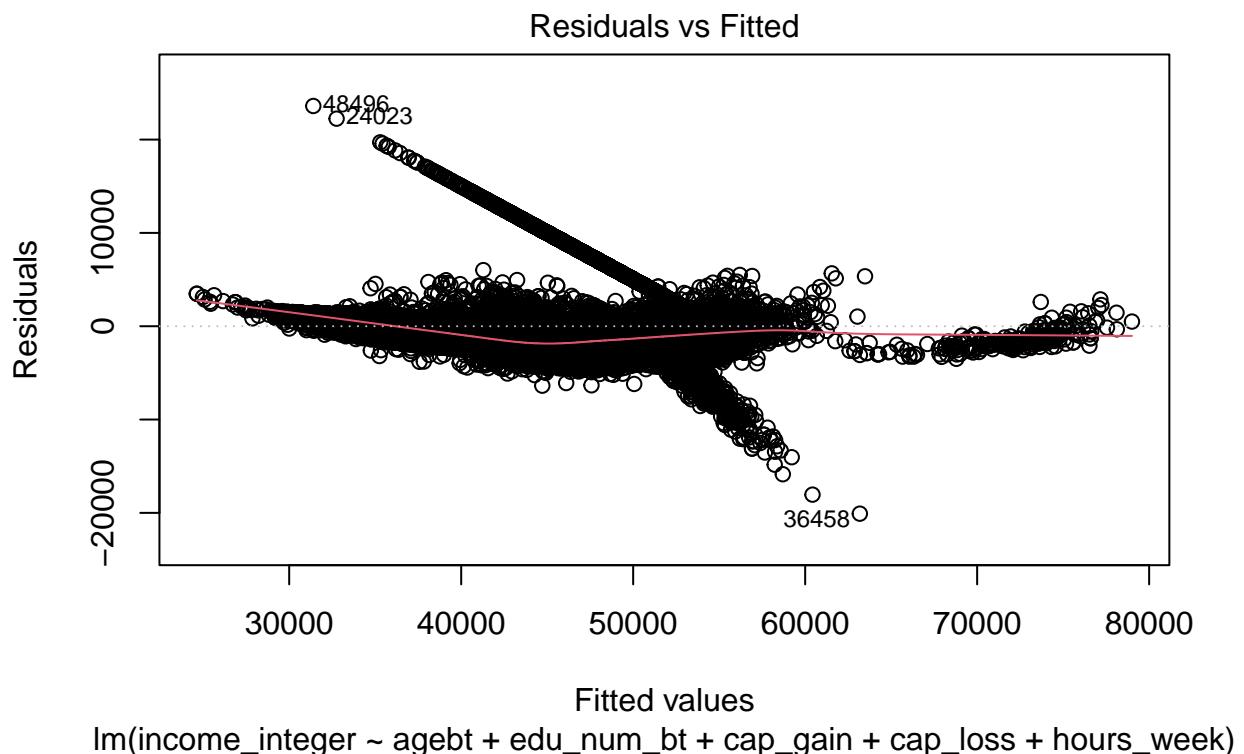
```
##
## Call:
## lm(formula = income_integer ~ agebt + edu_num_bt + cap_gain +
##     cap_loss + hours_week, data = dd)
##
```

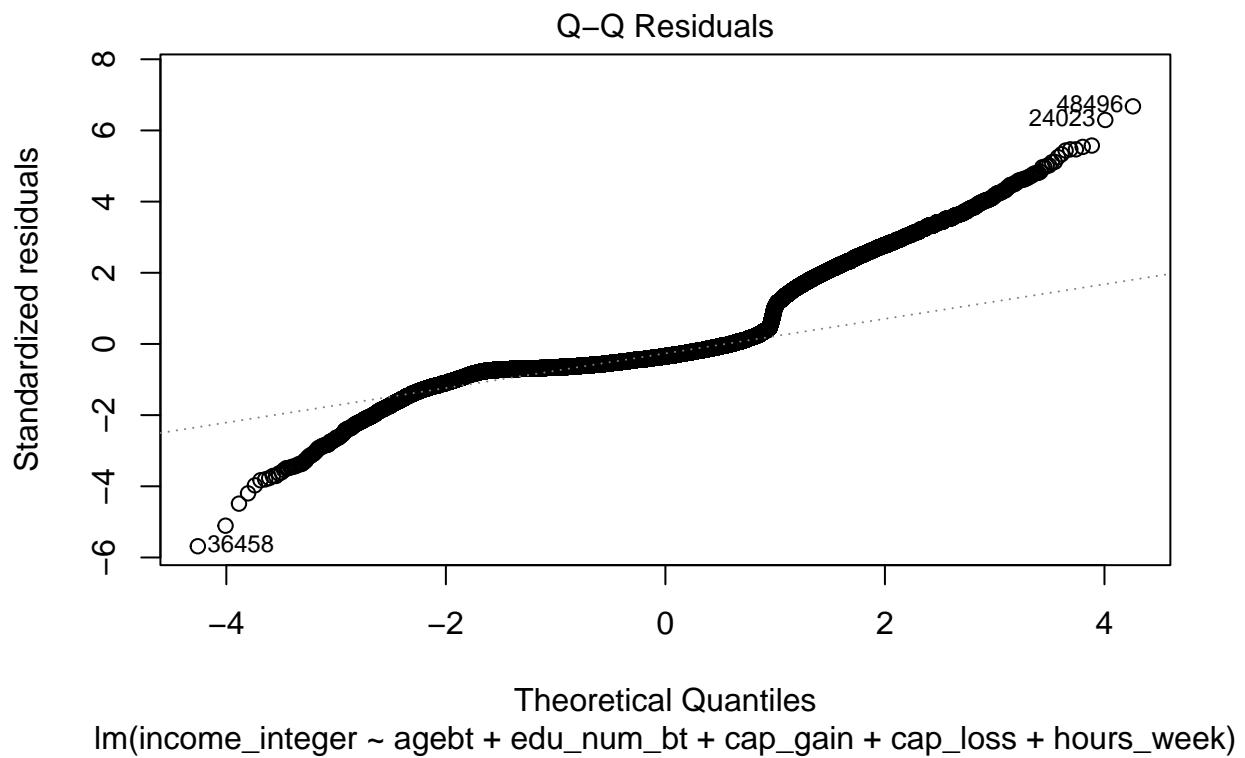
```

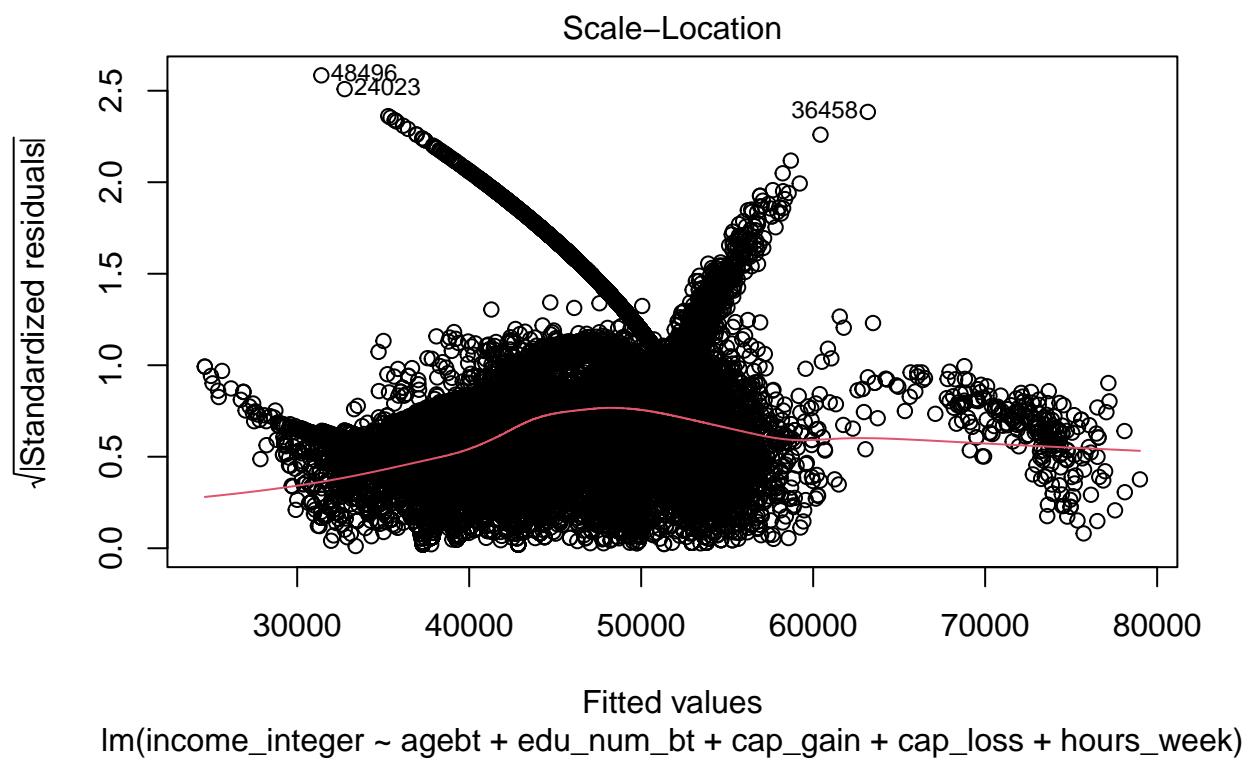
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20085.5 -2100.2 -1218.1   216.9 23598.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.102e+04 1.601e+02 256.25 <2e-16 ***
## agebt      -1.106e+05 5.281e+02 -209.44 <2e-16 ***
## edu_num_bt  6.408e+03 3.733e+01 171.65 <2e-16 ***
## cap_gain    2.067e-01 2.172e-03  95.15 <2e-16 ***
## cap_loss    8.164e-01 3.994e-02  20.44 <2e-16 ***
## hours_week   7.161e+01 1.325e+00  54.03 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3535 on 48836 degrees of freedom
## Multiple R-squared:  0.6926, Adjusted R-squared:  0.6926
## F-statistic: 2.201e+04 on 5 and 48836 DF, p-value: < 2.2e-16

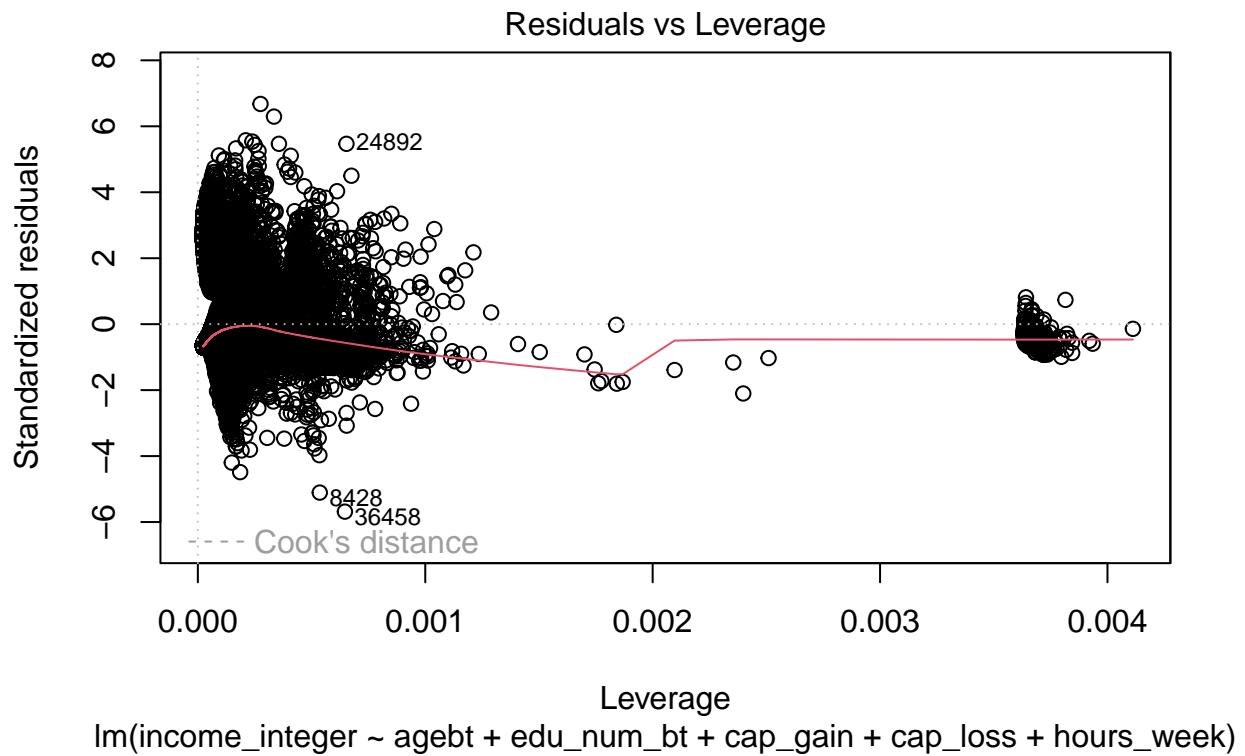
```

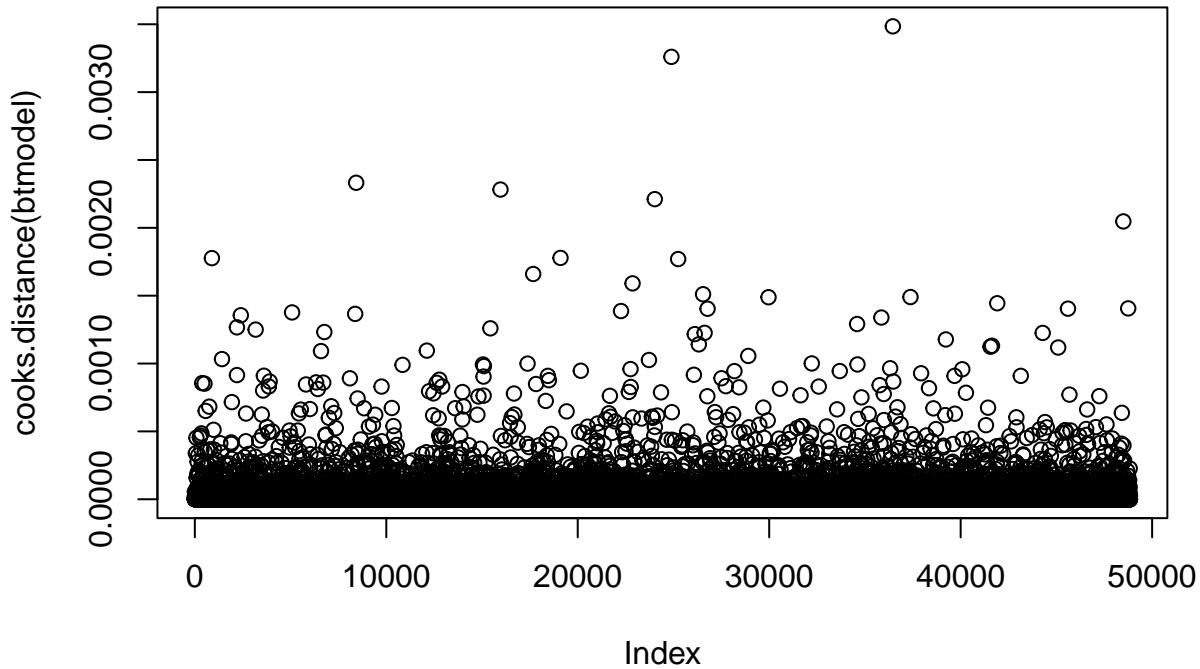
```
plot(btmodel)
```











```

#Try adding polynomial terms
age2 <- dd$age^2
hours_week2 <- dd$hours_week^2
model_poly <- lm(income_integer ~ age + age2 + edu_num + cap_gain + cap_loss + hours_week + hours_week2

#comparing model performance
summary(model_poly)

## 
## Call:
## lm(formula = income_integer ~ age + age2 + edu_num + cap_gain +
##     cap_loss + hours_week + hours_week2, data = dd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20398.4 -1917.5 -1288.9   -20.5 22142.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.504e+04 1.411e+02 106.580 < 2e-16 ***
## age         5.994e+02 6.672e+00  89.841 < 2e-16 ***
## age2        -4.304e+00 7.698e-02 -55.901 < 2e-16 ***
## edu_num     1.093e+03 6.432e+00 169.877 < 2e-16 ***
## cap_gain    2.018e-01 2.184e-03  92.369 < 2e-16 ***
## cap_loss    7.740e-01 4.011e-02 19.300 < 2e-16 ***
## hours_week  9.819e+01 4.395e+00 22.343 < 2e-16 ***

```

```

## hours_week2 -3.089e-01  4.847e-02  -6.374 1.85e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3548 on 48834 degrees of freedom
## Multiple R-squared:  0.6904, Adjusted R-squared:  0.6903
## F-statistic: 1.555e+04 on 7 and 48834 DF,  p-value: < 2.2e-16

```

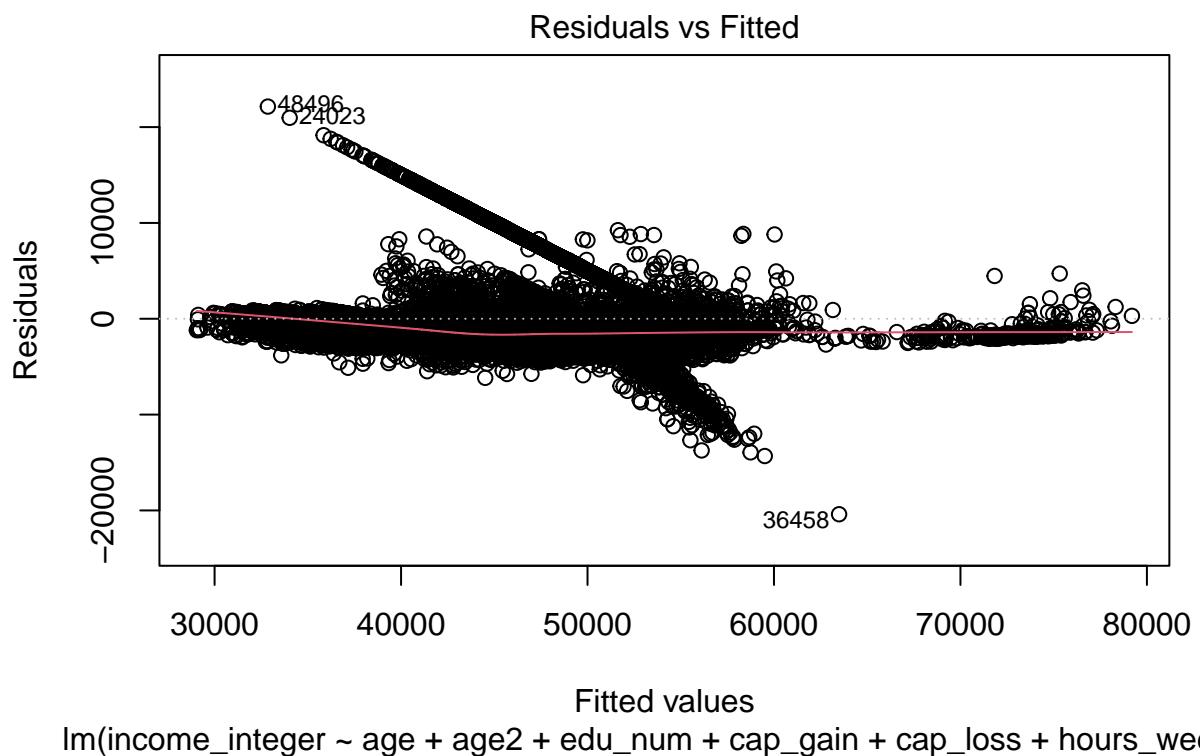
```
anova(initial_model, model_poly)
```

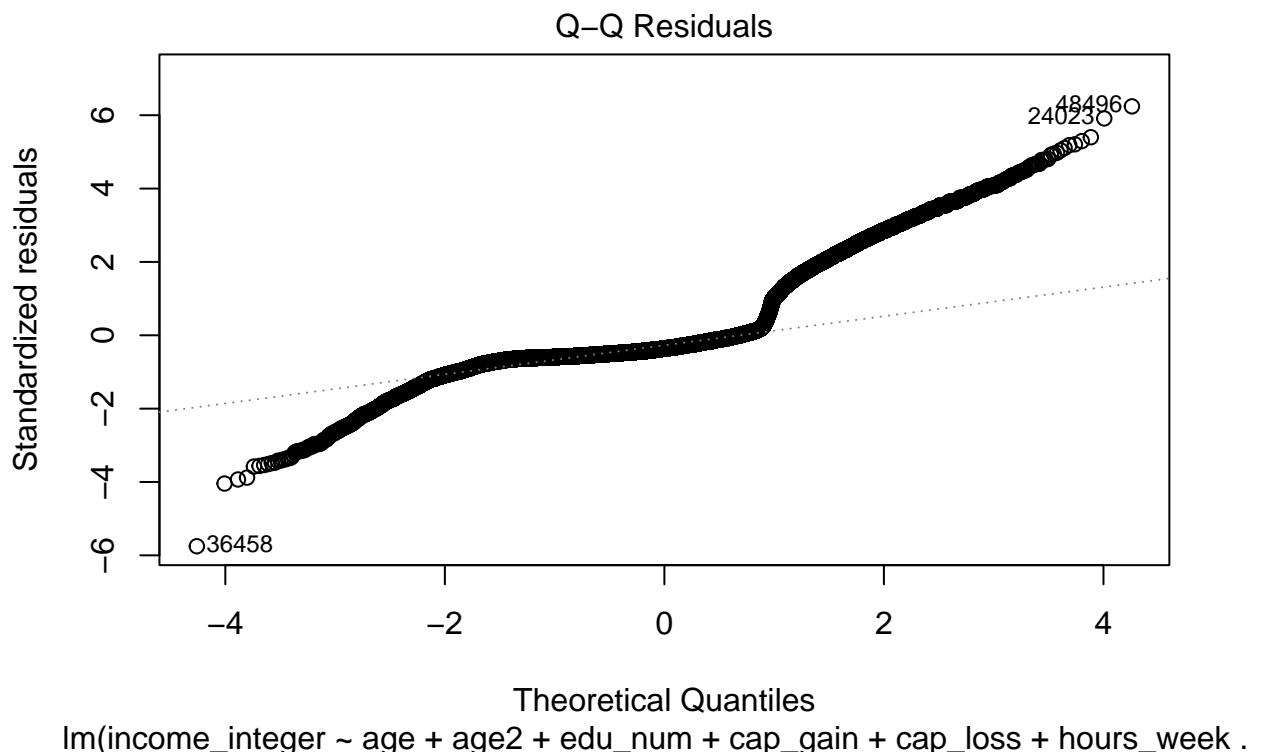
```

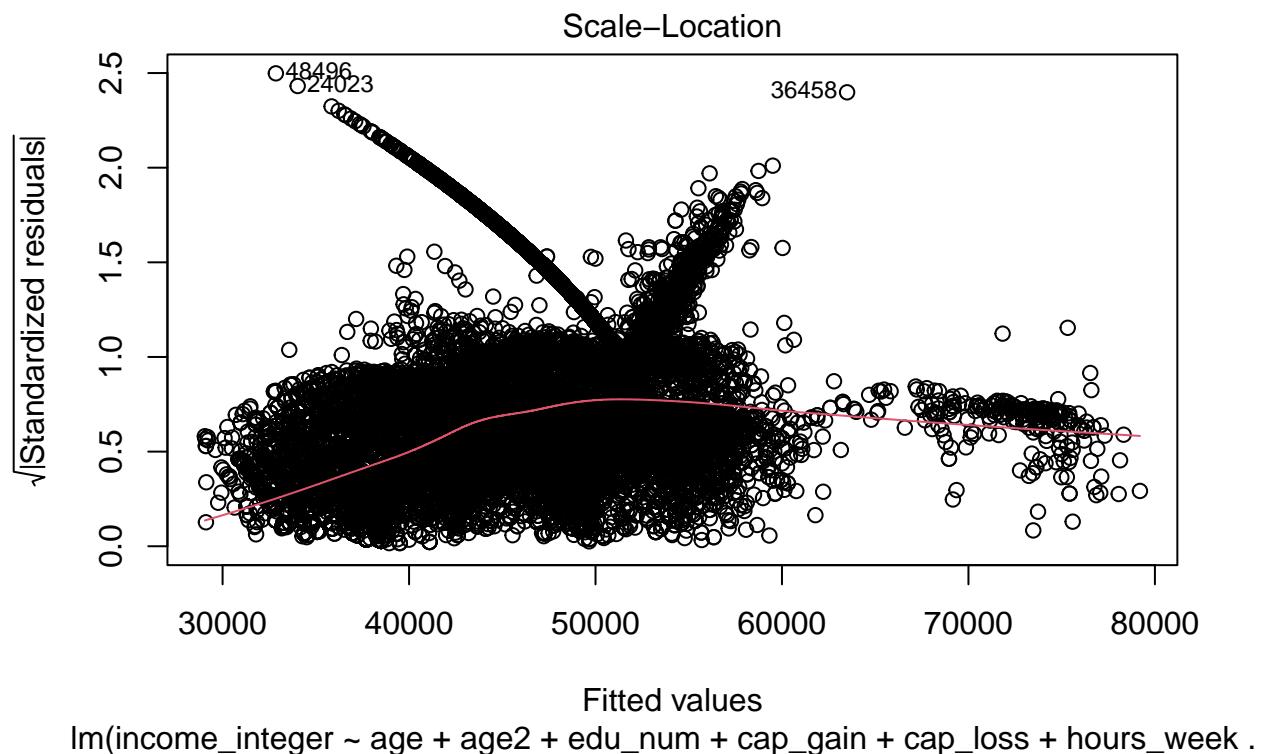
## Analysis of Variance Table
##
## Model 1: income_integer ~ age + edu_num + cap_gain + cap_loss + hours_week
## Model 2: income_integer ~ age + age2 + edu_num + cap_gain + cap_loss +
##           hours_week + hours_week2
##   Res.Df      RSS Df  Sum of Sq      F    Pr(>F)
## 1 48836 6.5856e+11
## 2 48834 6.1467e+11  2 4.3888e+10 1743.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

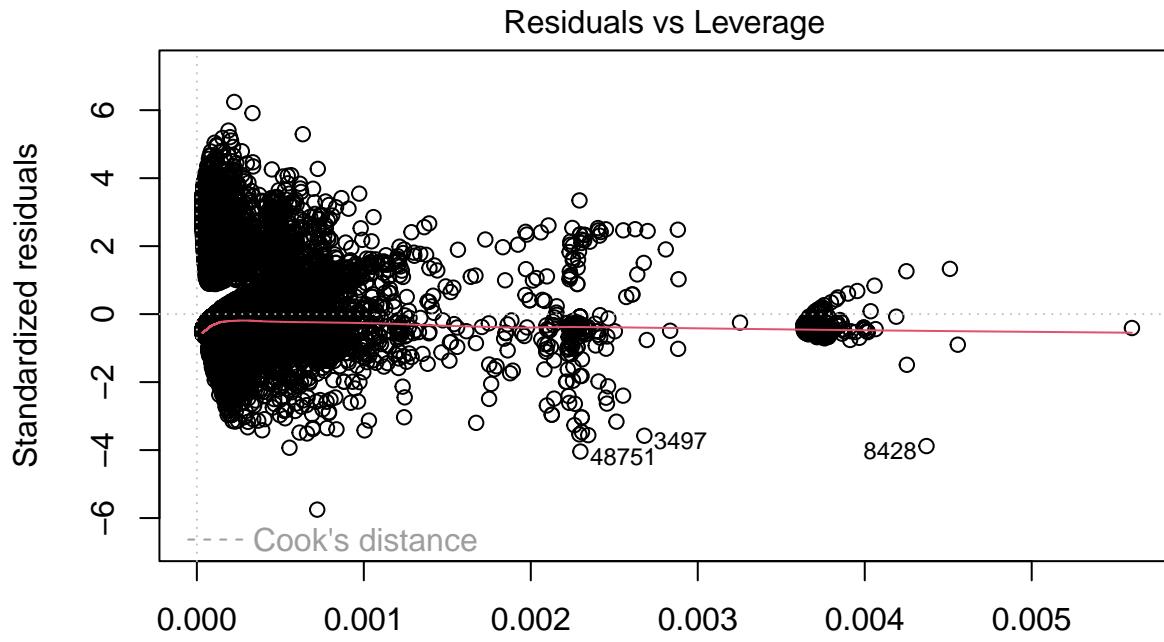
```

```
plot(model_poly)
```









Leverage

lm(income_integer ~ age + age2 + edu_num + cap_gain + cap_loss + hours_week .

```
#Incorporating Factors
#Add Occupation
modelo_occ <- update(transformed_model, . ~ . + occupation)
anova(transformed_model, modelo_occ) # p < 2.2e-16 ***
```

```
## Analysis of Variance Table
##
## Model 1: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week
## Model 2: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
##          occupation
##   Res.Df      RSS Df  Sum of Sq      F    Pr(>F)
## 1  48836 1.3544e-07
## 2  48823 1.3105e-07 13 4.3978e-09 126.03 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1
```

```
# Add estado civil (7 categories)
modelo_marital <- update(modelo_occ, . ~ . + marital)
anova(modelo_occ, modelo_marital) # p = < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
##          occupation
## Model 2: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
```

```

##      occupation + marital
##  Res.Df      RSS Df  Sum of Sq      F    Pr(>F)
## 1 48823 1.3105e-07
## 2 48819 1.1666e-07  4 1.4389e-08 1505.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Add género (2 categories)
modelo_gender <- update(modelo_marital, . ~ . + sex)
anova(modelo_marital, modelo_gender) # p = 4.18e-05 ***

## Analysis of Variance Table
##
## Model 1: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
##          occupation + marital
## Model 2: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
##          occupation + marital + sex
##  Res.Df      RSS Df  Sum of Sq      F    Pr(>F)
## 1 48819 1.1666e-07
## 2 48818 1.1662e-07  1 4.0111e-11 16.791 4.18e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Add clase trabajadora (9 categories)
modelo_workclass <- update(modelo_gender, . ~ . + workclass)
anova(modelo_gender, modelo_workclass) # p = < 2.2e-16

## Analysis of Variance Table
##
## Model 1: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
##          occupation + marital + sex
## Model 2: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
##          occupation + marital + sex + workclass
##  Res.Df      RSS Df  Sum of Sq      F    Pr(>F)
## 1 48818 1.1662e-07
## 2 48812 1.1630e-07  6 3.1315e-10 21.904 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Add relación familiar (6 categories)
modelo_relat <- update(modelo_workclass, . ~ . + relationship)
anova(modelo_workclass, modelo_relat) # p = < 2.2e-16

## Analysis of Variance Table
##
## Model 1: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
##          occupation + marital + sex + workclass
## Model 2: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
##          occupation + marital + sex + workclass + relationship
##  Res.Df      RSS Df  Sum of Sq      F    Pr(>F)
## 1 48812 1.1630e-07
## 2 48807 1.1467e-07  5 1.6292e-09 138.69 < 2.2e-16 ***

```

```

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Add raza (5 categories)
modelo_race <- update(modelo_relat, . ~ . + race)
anova(modelo_relat, modelo_race) # p = 3.471e-06

## Analysis of Variance Table
##
## Model 1: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
##           occupation + marital + sex + workclass + relationship
## Model 2: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
##           occupation + marital + sex + workclass + relationship + race
##   Res.Df      RSS Df  Sum of Sq      F    Pr(>F)
## 1  48807 1.1467e-07
## 2  48803 1.1460e-07  4 7.2189e-11 7.6854 3.471e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Add país origen (42 categorías)
modelo_country <- update(modelo_race, . ~ . + native_country)
anova(modelo_race, modelo_country) # p = < 2.2e-16

## Analysis of Variance Table
##
## Model 1: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
##           occupation + marital + sex + workclass + relationship + race
## Model 2: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
##           occupation + marital + sex + workclass + relationship + race +
##           native_country
##   Res.Df      RSS Df  Sum of Sq      F    Pr(>F)
## 1  48803 1.1460e-07
## 2  48802 1.1413e-07  1 4.6909e-10 200.58 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Add income (2 categorias)
modelo_income <- update(modelo_country,. ~ . + income)
anova(modelo_country,modelo_income)

## Analysis of Variance Table
##
## Model 1: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
##           occupation + marital + sex + workclass + relationship + race +
##           native_country
## Model 2: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
##           occupation + marital + sex + workclass + relationship + race +
##           native_country + income
##   Res.Df      RSS Df  Sum of Sq      F    Pr(>F)
## 1  48802 1.1413e-07
## 2  48801 5.6812e-08  1 5.7321e-08 49239 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

#Model with all significant variables including categorical variables
catmodel <- lm(y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week + occupation + marital + sex +
stepmodel <- stepAIC(catmodel, direction = "back")

## Start: AIC=-1342090
## y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
##   occupation + marital + sex + workclass + relationship + race +
##   native_country + income
##
##          Df  Sum of Sq      RSS      AIC
## <none>            5.6812e-08 -1342090
## - sex             1 5.9000e-11 5.6871e-08 -1342041
## - cap_loss        1 9.6000e-11 5.6908e-08 -1342009
## - workclass       6 1.0800e-10 5.6920e-08 -1342009
## - race            4 1.8300e-10 5.6995e-08 -1341941
## - native_country  1 4.2800e-10 5.7239e-08 -1341725
## - marital          4 6.2600e-10 5.7438e-08 -1341563
## - relationship    5 1.3970e-09 5.8209e-08 -1340913
## - cap_gain         1 1.8280e-09 5.8640e-08 -1340545
## - occupation       13 1.9100e-09 5.8722e-08 -1340501
## - hours_week       1 5.5490e-09 6.2361e-08 -1337540
## - age              1 4.5742e-08 1.0255e-07 -1313244
## - edu_num          1 4.8472e-08 1.0528e-07 -1311960
## - income           1 5.7321e-08 1.1413e-07 -1308018

vif(catmodel)

##          GVIF Df GVIF^(1/(2*Df))
## age        1.692382  1     1.300916
## edu_num    1.487519  1     1.219639
## cap_gain   1.071958  1     1.035354
## cap_loss   1.030813  1     1.015290
## hours_week 1.228567  1     1.108407
## occupation 2.319271  13    1.032885
## marital    59.855229  4     1.667776
## sex         1.990143  1     1.410724
## workclass  1.444560  6     1.031125
## relationship 74.715581  5     1.539363
## race        1.282085  4     1.031548
## native_country 1.229145  1     1.108668
## income      1.556683  1     1.247671

summary(catmodel)

##
## Call:
## lm(formula = y_trans ~ age + edu_num + cap_gain + cap_loss +
##   hours_week + occupation + marital + sex + workclass + relationship +
##   race + native_country + income, data = dd)
##
## Residuals:
##       Min        1Q      Median        3Q       Max

```

```

## -7.614e-06 -6.165e-07 -1.486e-07  4.157e-07  8.822e-06
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.260e-05 9.431e-08 345.621 < 2e-16 ***
## age                      -9.182e-08 4.632e-10 -198.221 < 2e-16 ***
## edu_num                  -4.726e-07 2.316e-09 -204.053 < 2e-16 ***
## cap_gain                 -2.688e-11 6.783e-13 -39.624 < 2e-16 ***
## cap_loss                  1.120e-10 1.230e-11   9.103 < 2e-16 ***
## hours_week                -3.015e-08 4.367e-10 -69.042 < 2e-16 ***
## occupationArmy             3.609e-07 2.804e-07   1.287 0.198063
## occupationCraftRep        -9.955e-08 2.168e-08 -4.592 4.41e-06 ***
## occupationExecMan          2.678e-07 2.107e-08  12.712 < 2e-16 ***
## occupationFarmFish         2.623e-07 3.366e-08   7.793 6.68e-15 ***
## occupationHandlCl           3.421e-07 2.896e-08  11.813 < 2e-16 ***
## occupationHouse            5.351e-07 7.145e-08   7.489 7.08e-14 ***
## occupationMachOp           1.040e-07 2.536e-08   4.103 4.08e-05 ***
## occupationOther             3.311e-07 2.156e-08  15.360 < 2e-16 ***
## occupationProf              4.773e-07 1.916e-08  24.913 < 2e-16 ***
## occupationProtServ         -1.347e-07 3.895e-08 -3.458 0.000545 ***
## occupationSales             1.841e-07 2.124e-08   8.669 < 2e-16 ***
## occupationTech              1.057e-07 3.208e-08 -3.294 0.000989 ***
## occupationTrans             -8.798e-08 2.791e-08 -3.152 0.001621 **
## maritalMarried             -9.439e-08 6.059e-08 -1.558 0.119255
## maritalNevMarr             3.784e-07 1.832e-08  20.655 < 2e-16 ***
## maritalSep                  1.301e-07 2.702e-08   4.814 1.49e-06 ***
## maritalWidow                -1.007e-07 3.185e-08 -3.162 0.001567 **
## sexMale                     1.045e-07 1.463e-08   7.145 9.11e-13 ***
## workclassLoc                -1.122e-08 3.511e-08 -0.320 0.749349
## workclassNoPay              3.057e-07 1.963e-07   1.557 0.119437
## workclassPriv               1.727e-09 2.981e-08   0.058 0.953805
## workclassSelfI              2.568e-07 3.972e-08   6.465 1.02e-10 ***
## workclassSelfN              2.503e-08 3.459e-08   0.723 0.469400
## workclassState              6.483e-08 3.783e-08   1.714 0.086581 .
## relationshipNot-in-family -2.170e-07 6.038e-08 -3.594 0.000326 ***
## relationshipOther-relative  1.780e-07 5.925e-08   3.005 0.002661 **
## relationshipOwn-child       3.485e-07 6.027e-08   5.782 7.41e-09 ***
## relationshipUnmarried       -1.088e-07 6.261e-08 -1.738 0.082252 .
## relationshipWife            -2.545e-07 2.754e-08 -9.241 < 2e-16 ***
## raceAsian-Pac-Islander     -2.893e-07 5.850e-08 -4.946 7.61e-07 ***
## raceBlack                   -1.001e-07 5.231e-08 -1.914 0.055601 .
## raceOther                   1.668e-07 7.369e-08   2.264 0.023598 *
## raceWhite                   1.981e-08 5.020e-08   0.395 0.693175
## native_countryUSA           -3.419e-07 1.784e-08 -19.166 < 2e-16 ***
## income>50K                  -3.168e-06 1.428e-08 -221.898 < 2e-16 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1
##
## Residual standard error: 1.079e-06 on 48801 degrees of freedom
## Multiple R-squared:  0.8787, Adjusted R-squared:  0.8787
## F-statistic:  8842 on 40 and 48801 DF, p-value: < 2.2e-16

```

```
anova(catmodel)
```

```

## Analysis of Variance Table
##
## Response: y_trans
##                               Df      Sum Sq   Mean Sq   F value   Pr(>F)
## age                      1 1.5949e-07 1.5949e-07 137000.941 < 2.2e-16 ***
## edu_num                  1 1.4352e-07 1.4352e-07 123282.092 < 2.2e-16 ***
## cap_gain                 1 9.4400e-09 9.4400e-09  8109.099 < 2.2e-16 ***
## cap_loss                 1 8.5000e-10 8.5000e-10   730.241 < 2.2e-16 ***
## hours_week                1 1.9806e-08 1.9806e-08 17013.078 < 2.2e-16 ***
## occupation                13 4.3980e-09 3.3800e-10   290.591 < 2.2e-16 ***
## marital                   4 1.4389e-08 3.5970e-09  3090.113 < 2.2e-16 ***
## sex                       1 4.0000e-11 4.0000e-11   34.455 4.391e-09 ***
## workclass                 6 3.1300e-10 5.2000e-11   44.832 < 2.2e-16 ***
## relationship               5 1.6290e-09 3.2600e-10  279.902 < 2.2e-16 ***
## race                      4 7.2000e-11 1.8000e-11   15.502 1.117e-12 ***
## native_country              1 4.6900e-10 4.6900e-10  402.948 < 2.2e-16 ***
## income                     1 5.7321e-08 5.7321e-08 49238.802 < 2.2e-16 ***
## Residuals                 48801 5.6812e-08 1.0000e-12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

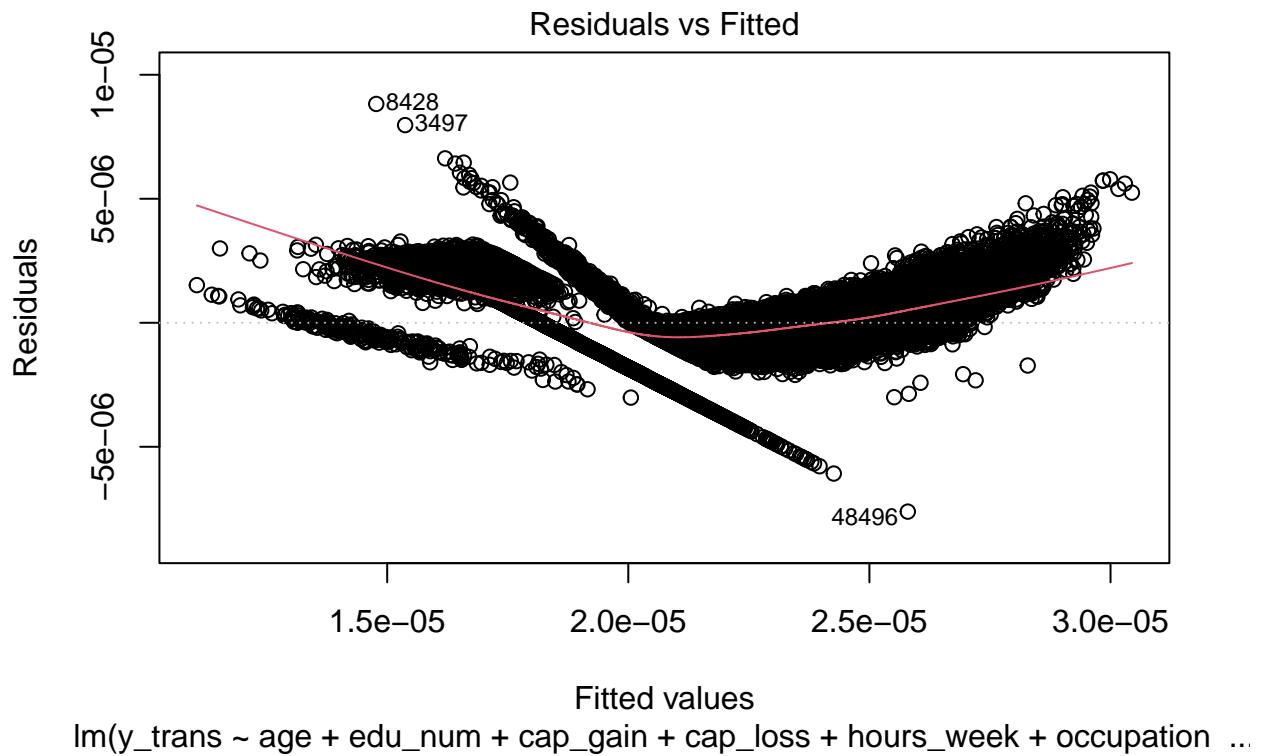
```
anova(transformed_model, catmodel)
```

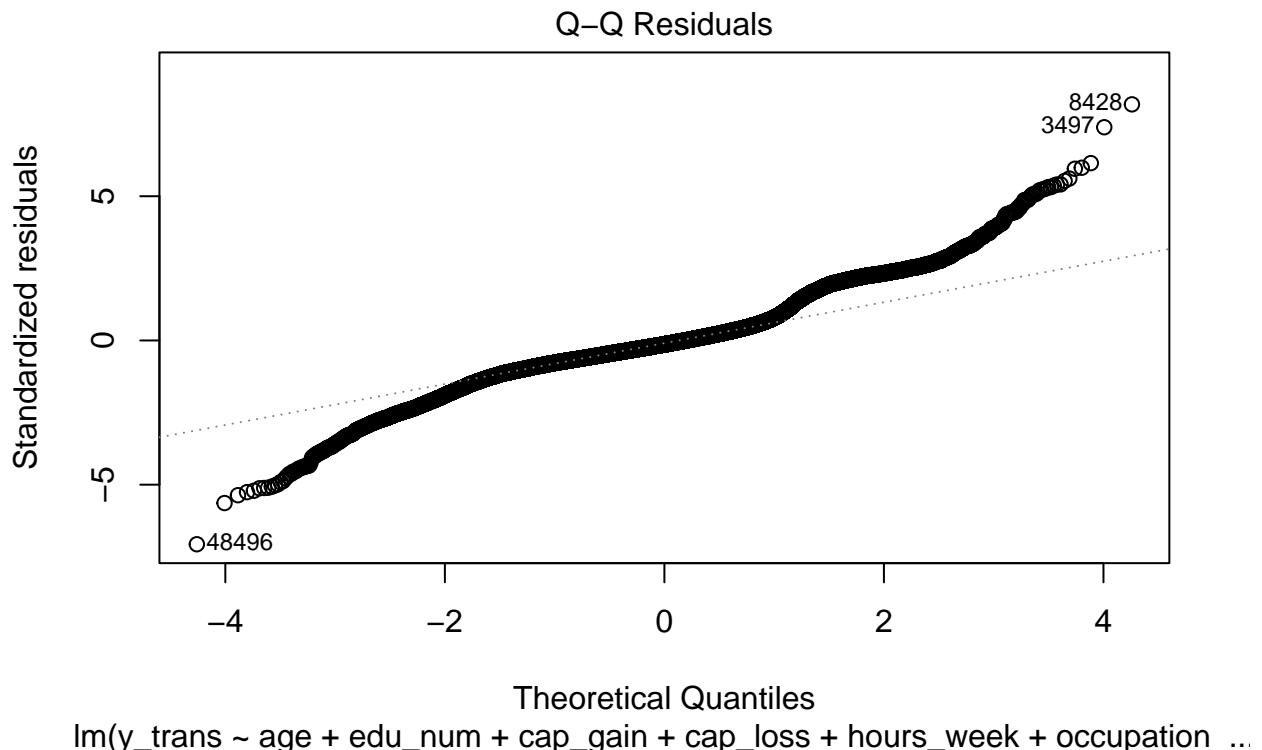
```

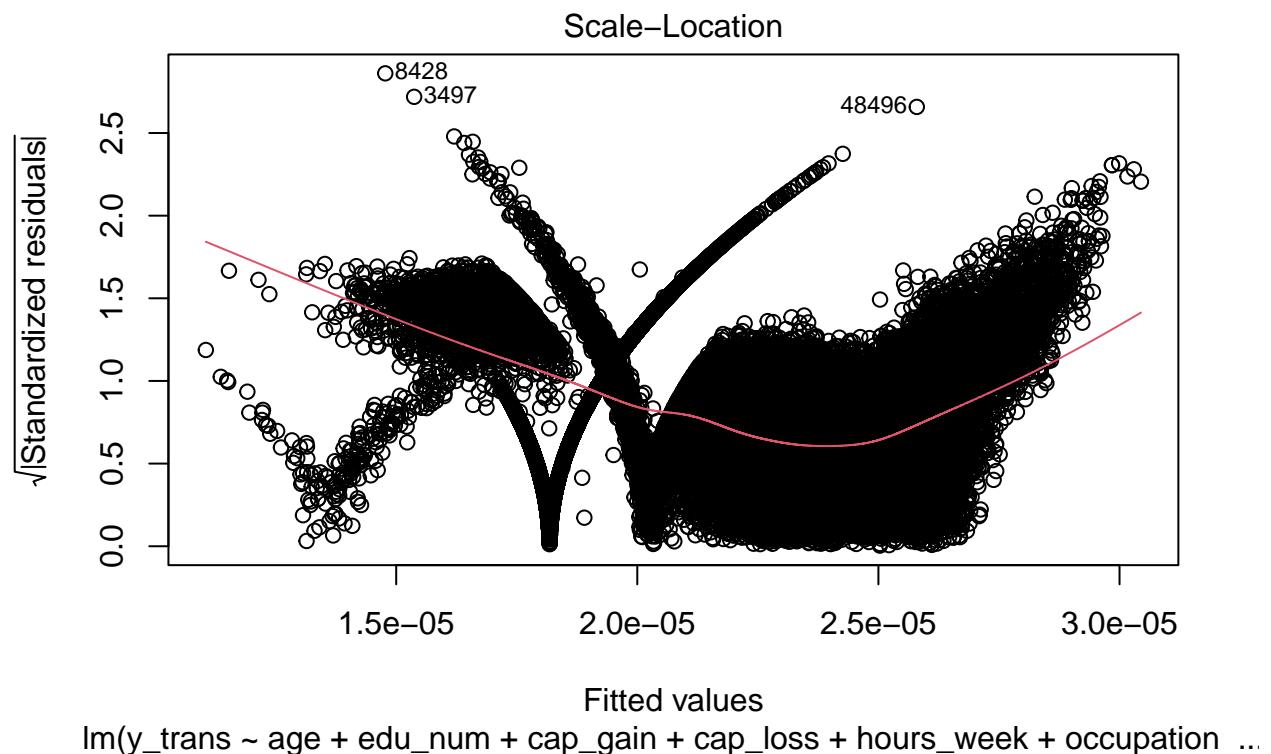
## Analysis of Variance Table
##
## Model 1: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week
## Model 2: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
##           occupation + marital + sex + workclass + relationship + race +
##           native_country + income
##   Res.Df      RSS Df  Sum of Sq      F   Pr(>F)
## 1  48836 1.3544e-07
## 2  48801 5.6812e-08 35 7.8633e-08 1929.9 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

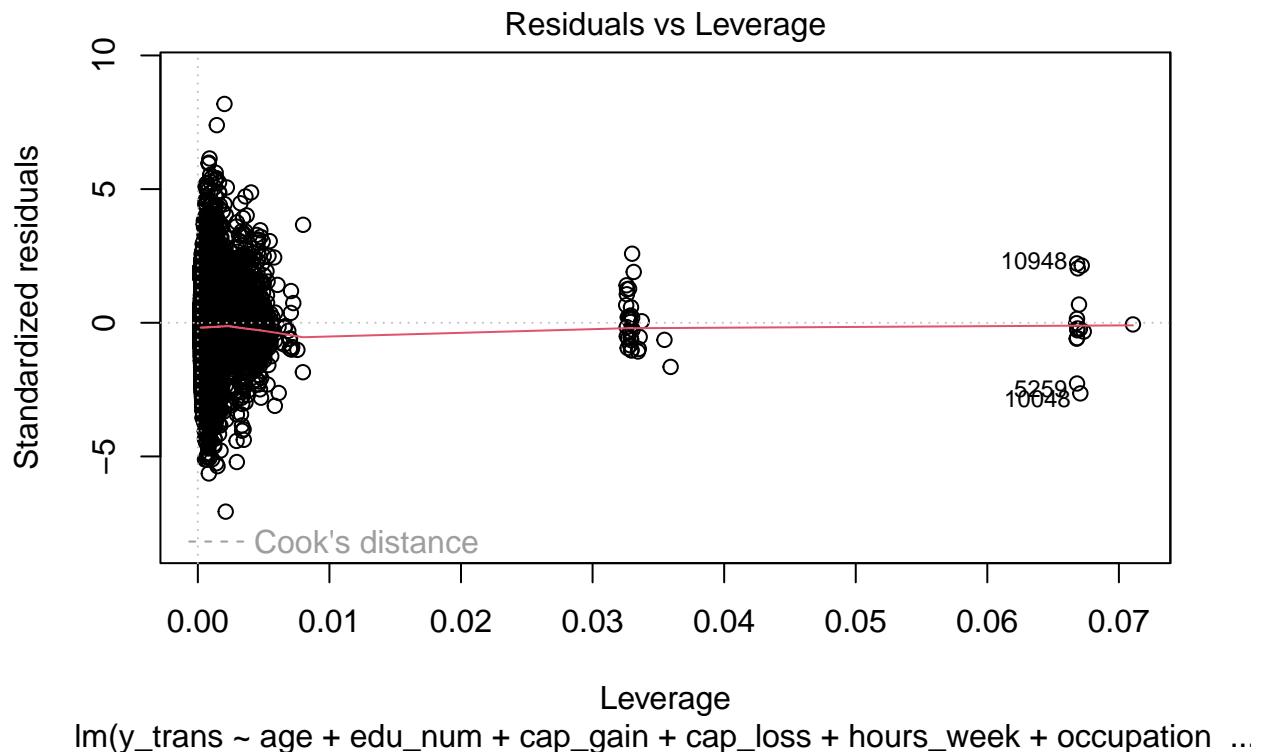
```

```
plot(catmodel)
```

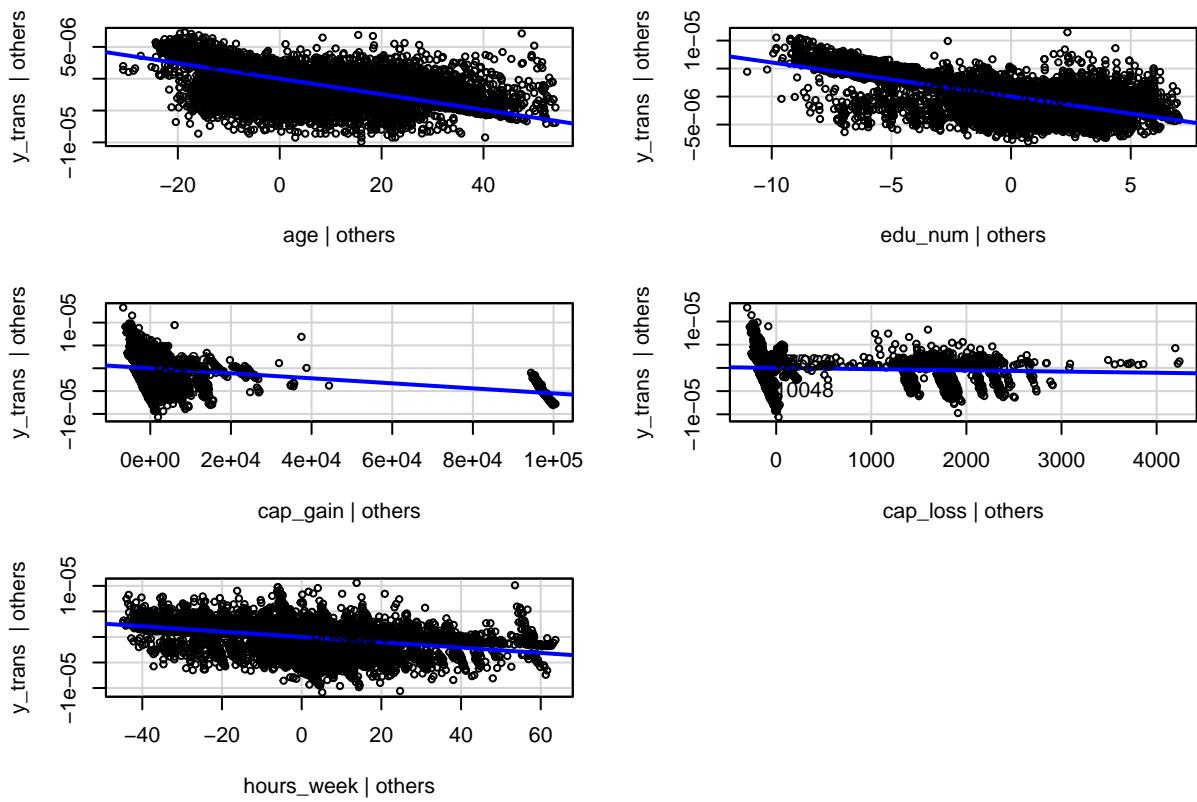








Added-Variable Plots



```
plot(allEffects(catmodel))
```

