Deliverable 2

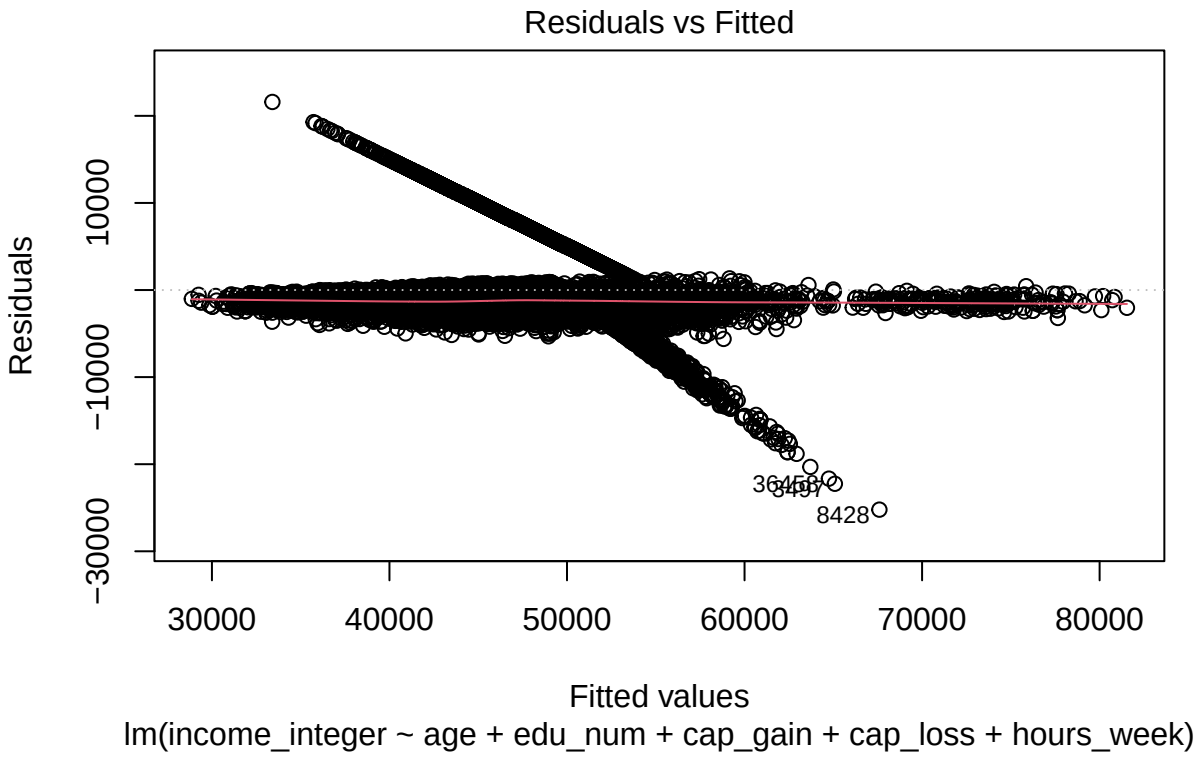
2025-04-23

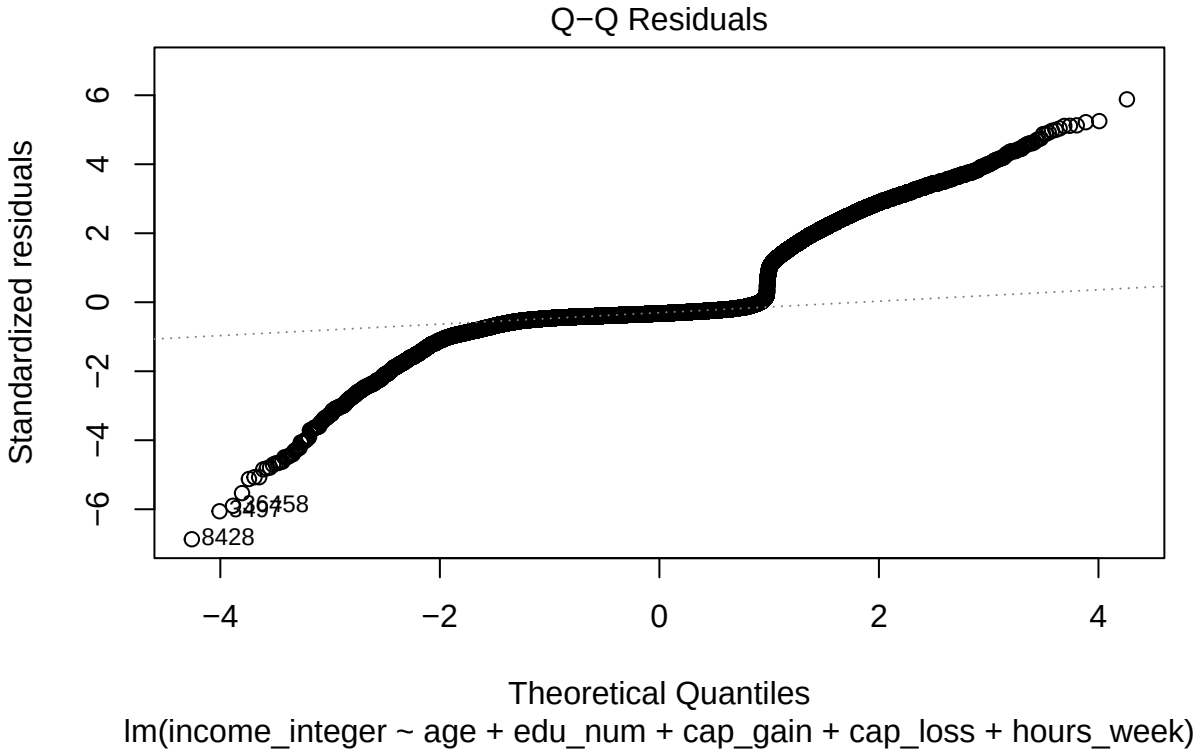
```r
setwd("~/Escritorio/ADEI/D2")
dd <- read.csv("adult_def.csv")
```

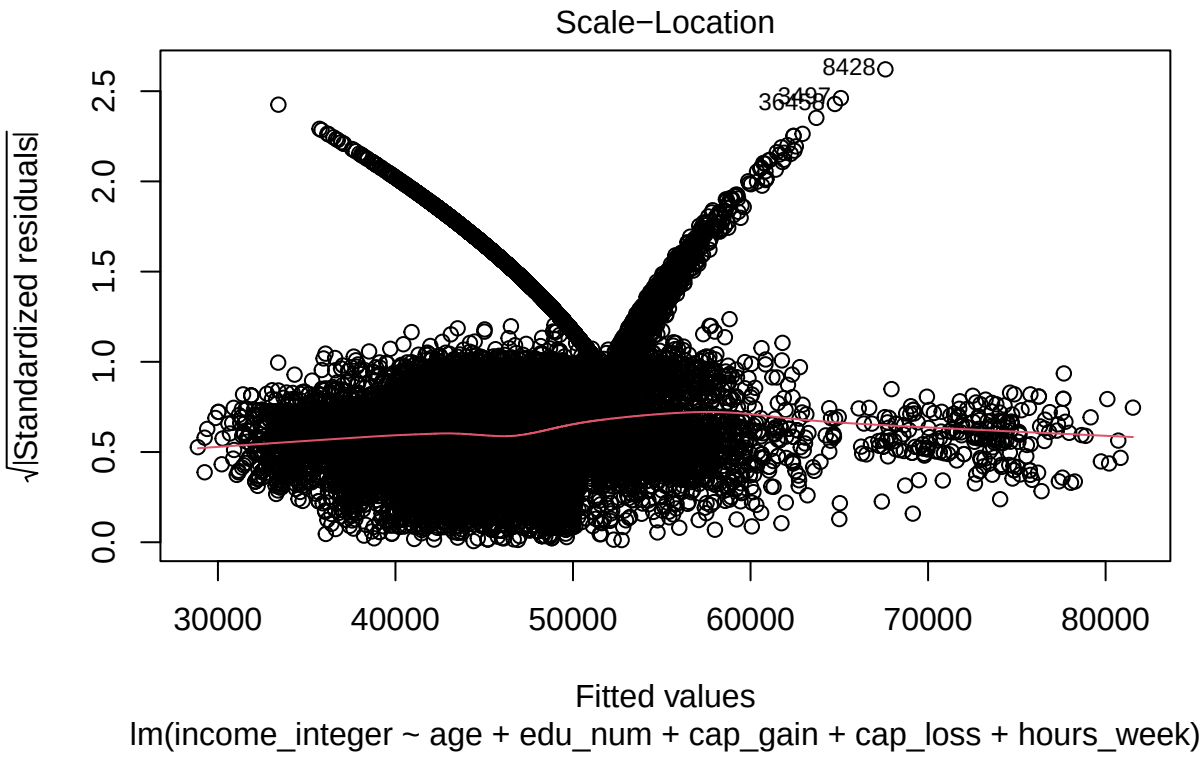
```r
initial_model <- lm(income_integer ~ age + edu_num + cap_gain + cap_loss + hours_week, data = dd)
summary(initial_model)
```

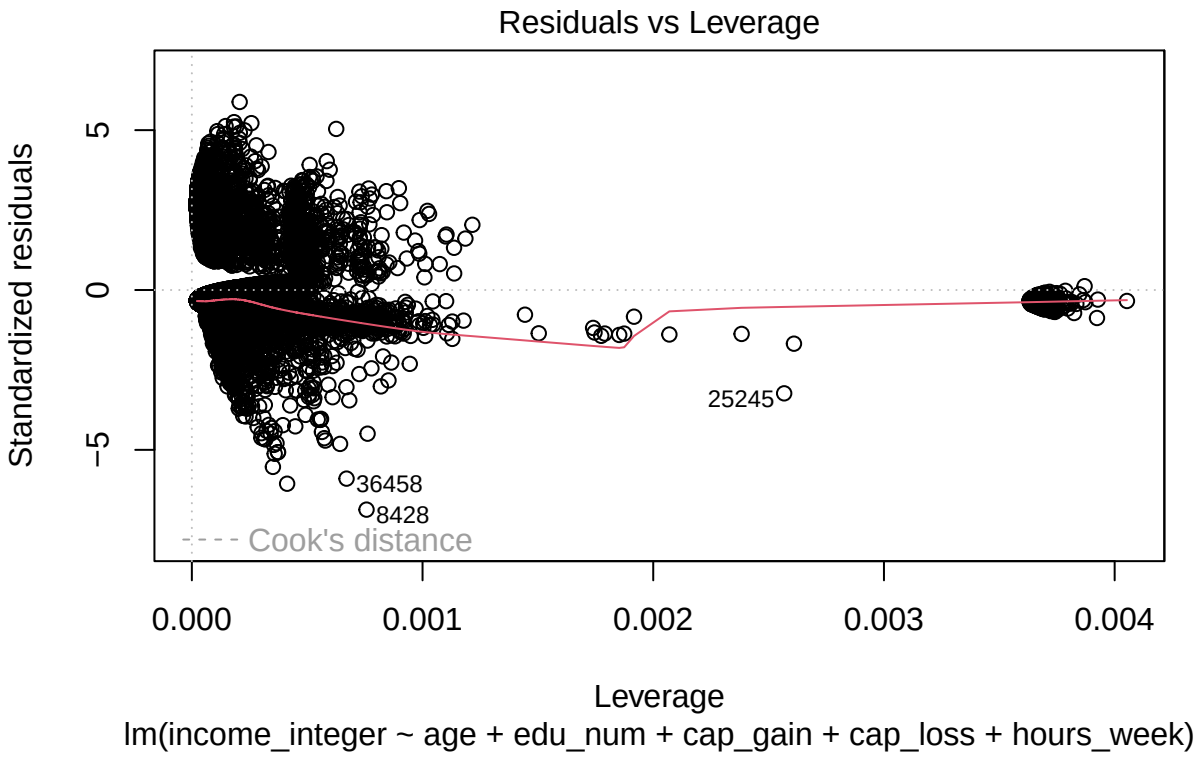
```
##
## Call:
## lm(formula = income_integer ~ age + edu_num + cap_gain + cap_loss +
##     hours_week, data = dd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25216.0  -1522.3  -1201.8   -699.4  21594.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (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
```

```r
plot(initial_model)
```

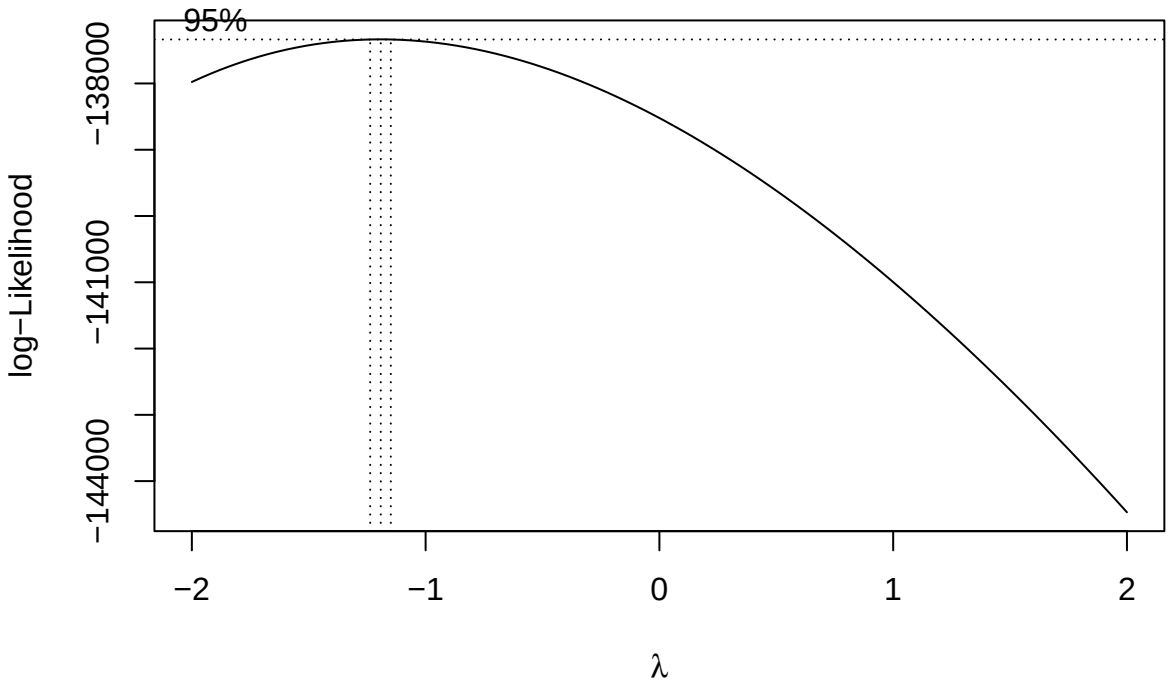

Residuals vs Fitted
Residuals
10000
−10000
−30000
30000
40000
50000
60000
70000
80000
8428
Fitted values
lm(income_integer ~ age + edu_num + cap_gain + cap_loss + hours_week)


Q−Q Residuals
Standardized residuals
6
4
2
0
−2
−4
−6
26458
3497
8428
−4
−2
0
2
4
Theoretical Quantiles
lm(income_integer ~ age + edu_num + cap_gain + cap_loss + hours_week)


Scale−Location
8428
3497
36458
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
30000
40000
50000
60000
70000
80000
Fitted values
lm(income_integer ~ age + edu_num + cap_gain + cap_loss + hours_week)




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

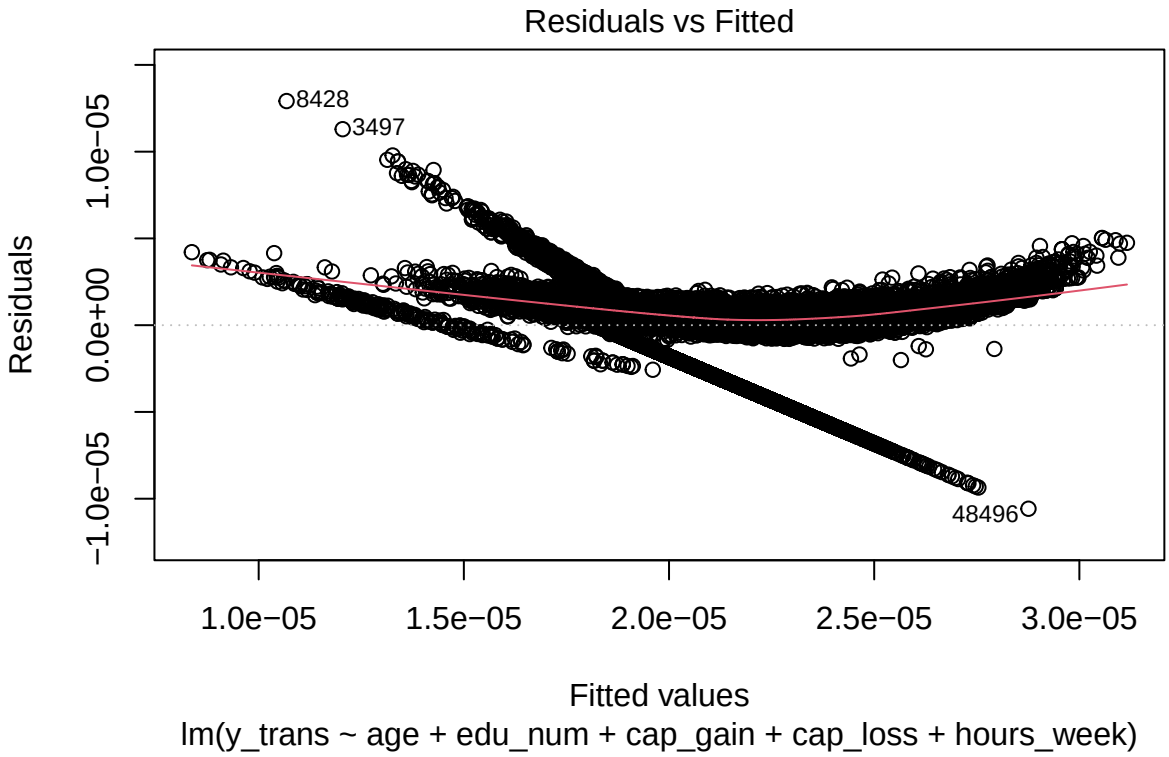


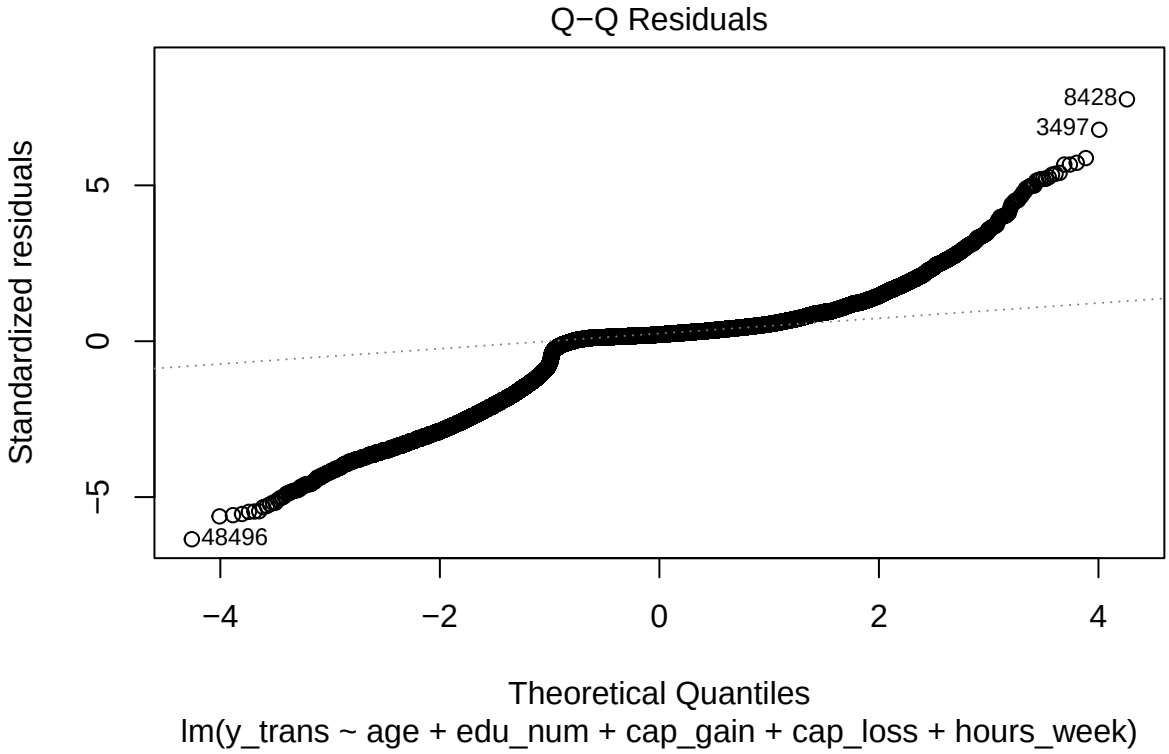

```
#given lambda is aproximatedly -1 we do the inverse transformation
y_trans <- 1 / dd$income_integer
transformed_model <- lm(y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week, data = dd)
summary(transformed_model)
```

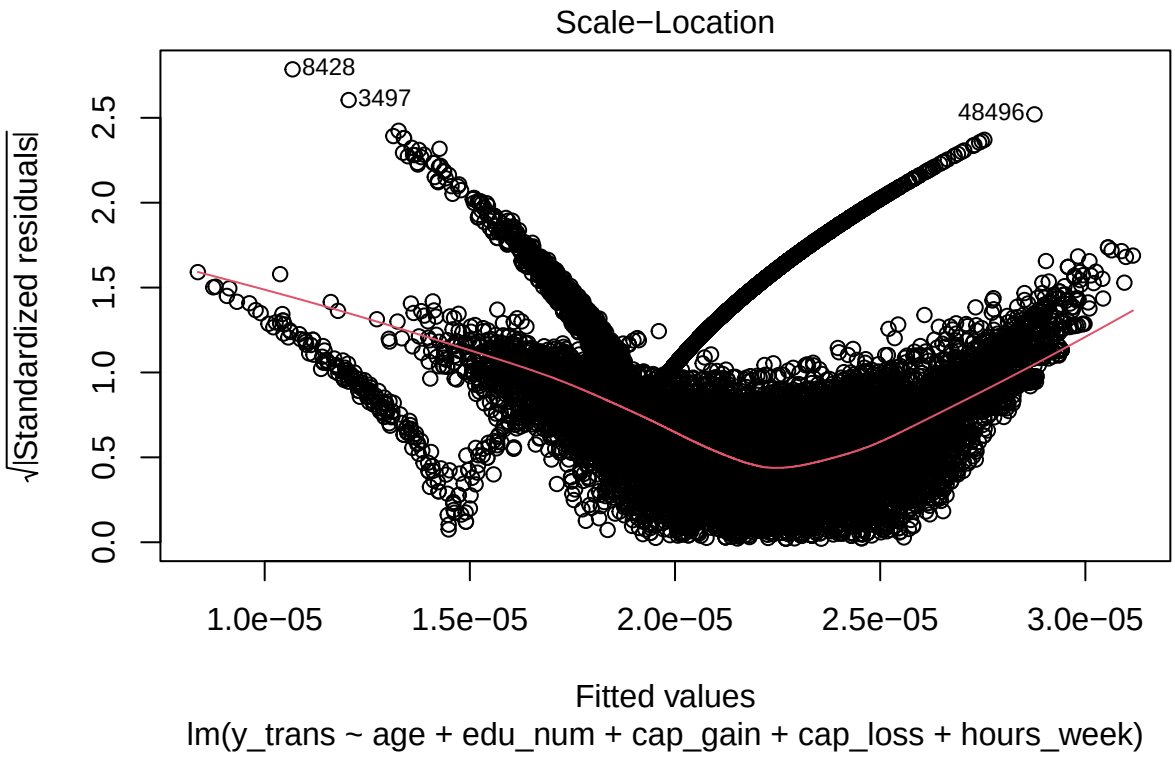
```
##
## Call:
## lm(formula = y_trans ~ age + edu_num + cap_gain + cap_loss +
##     hours_week, data = dd)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.058e-05  1.373e-07  3.552e-07  6.859e-07  1.292e-05
##
## Coefficients:
##               Estimate Std. Error t 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
```

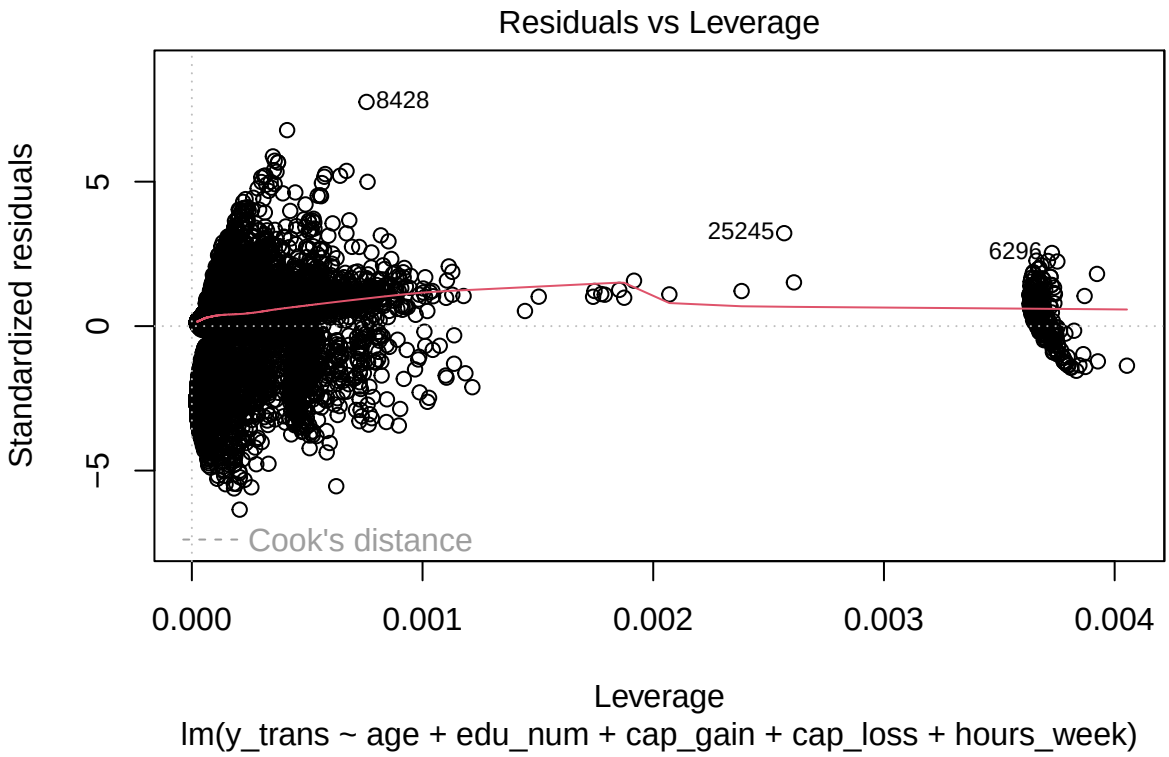
```
## Multiple R-squared:  0.7109, Adjusted R-squared:  0.7109
## F-statistic: 2.402e+04 on 5 and 48836 DF,  p-value: < 2.2e-16
```

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

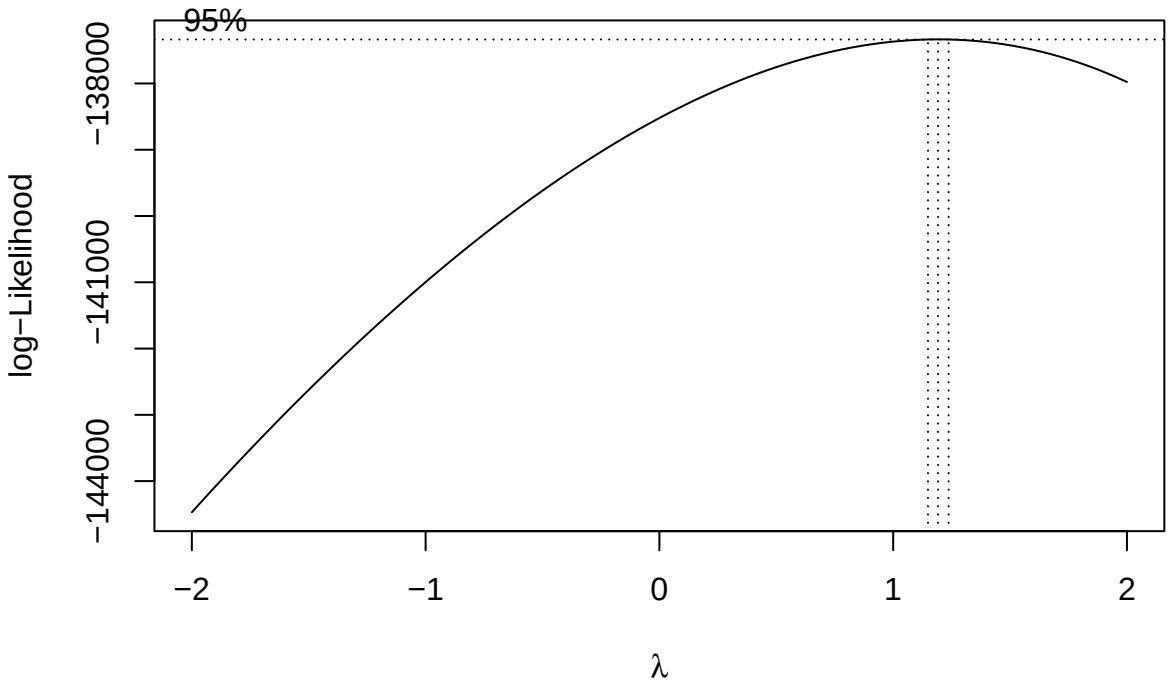



Q−Q Residuals
Standardized residuals
5
0
−5
8428
3497
48496
−4
−2
0
2
4
Theoretical Quantiles
lm(y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week)


Scale−Location
8428
3497
48496
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
1.0e−05
1.5e−05
2.0e−05
2.5e−05
3.0e−05
Fitted values
lm(y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week)




```
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)
```

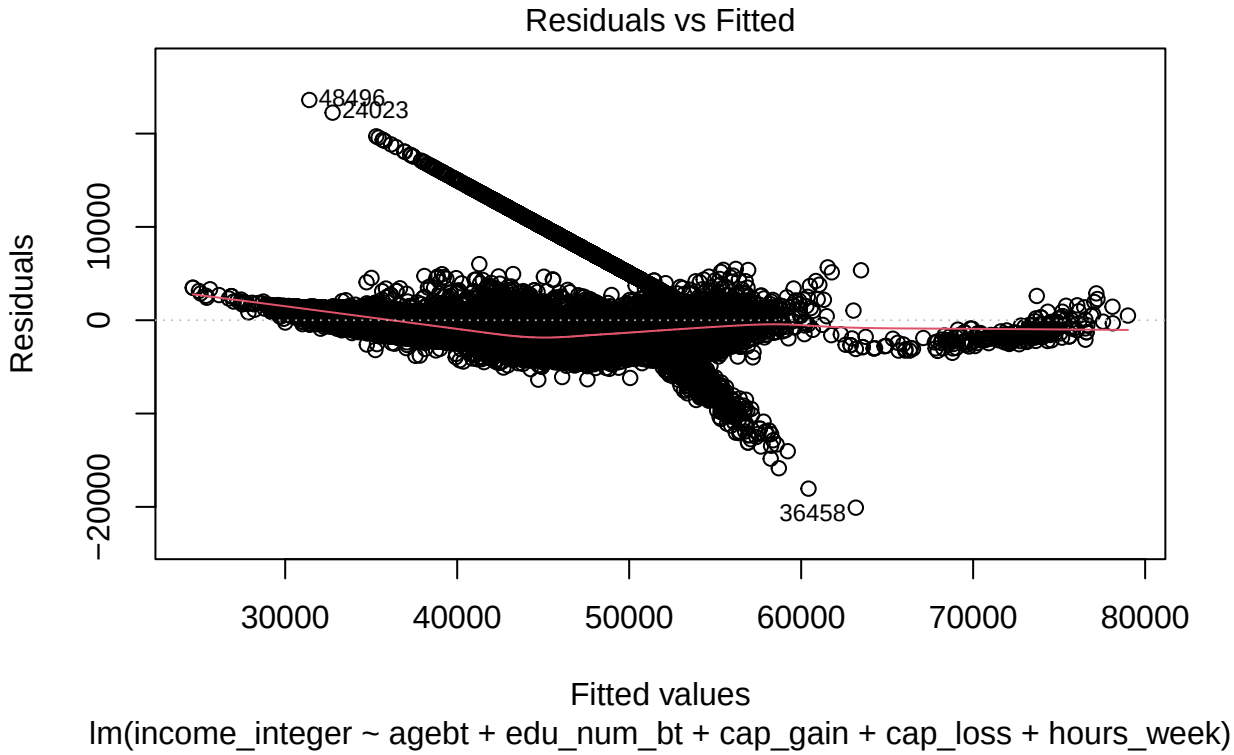
```
##             MLE of lambda Score Statistic (t) Pr(>|t|)
## age              -0.36944            -52.9070   <2e-16 ***
## edu_num           0.71757            -13.2951   <2e-16 ***
## hours_week        0.95940             -0.2224    0.824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 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
```

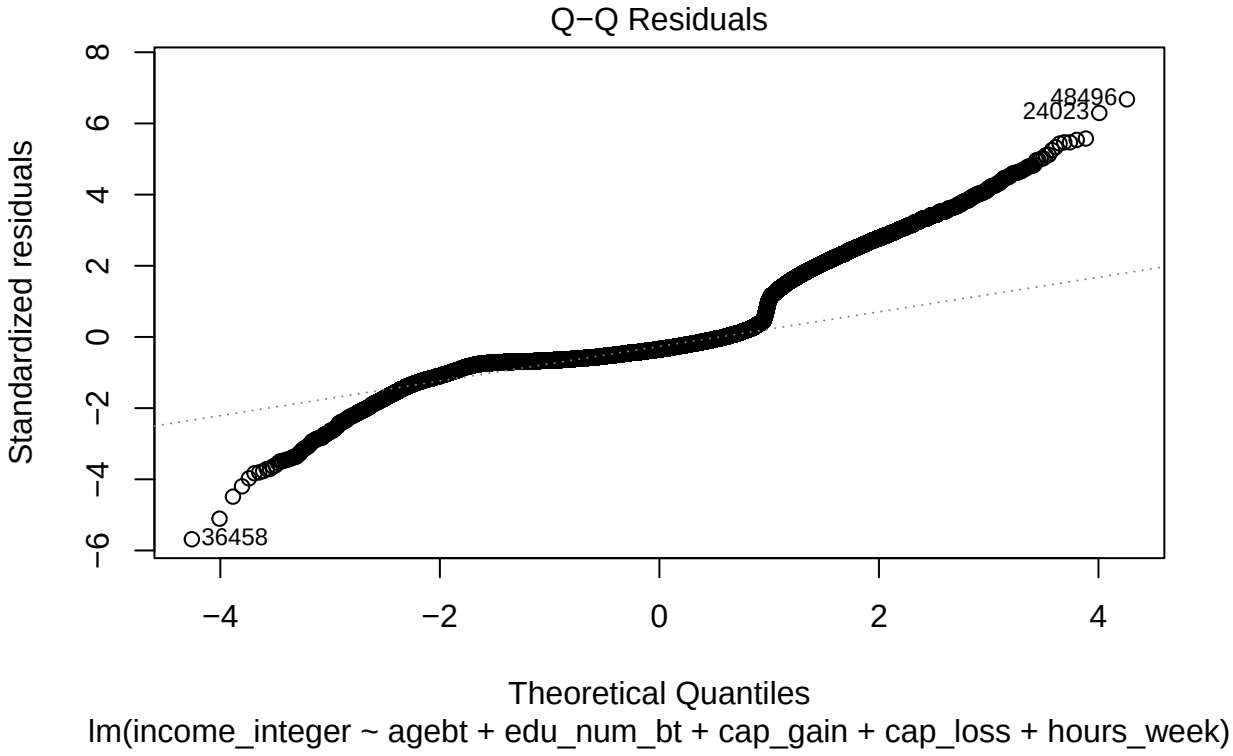
```
agebt <- 1 / sqrt(dd$age)
edu_num_bt <- sqrt(dd$edu_num)
btmodel <- lm(income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week, data = dd)
summary(btmodel)
```

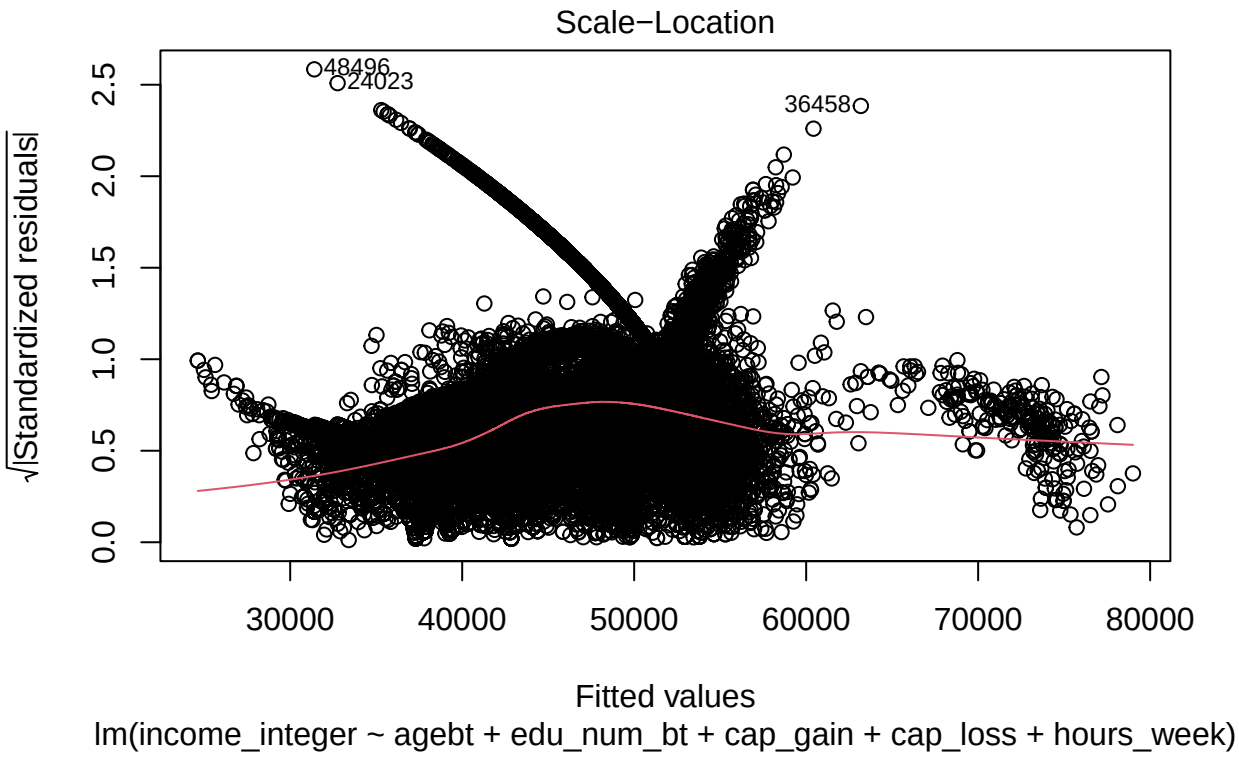
```
##
## Call:
## lm(formula = income_integer ~ agebt + edu_num_bt + cap_gain +
##     cap_loss + hours_week, data = dd)
##
```

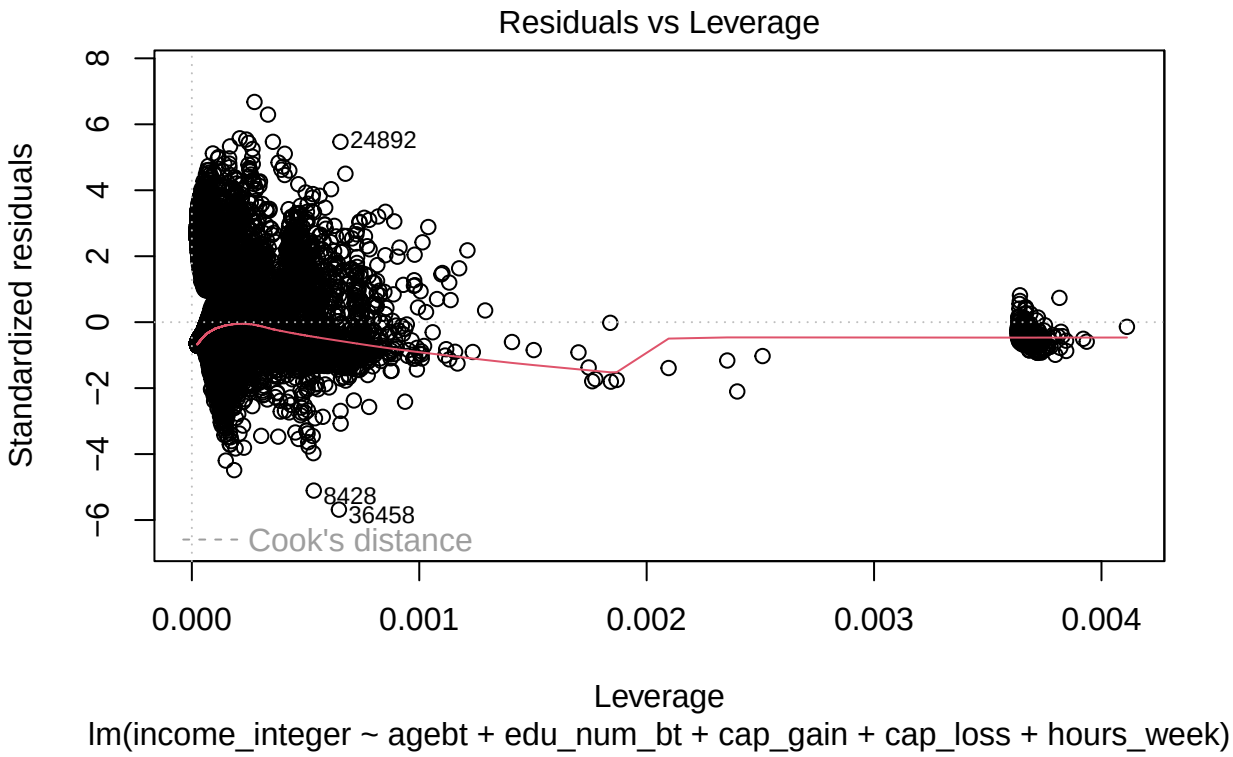
```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20085.5  -2100.2  -1218.1    216.9  23598.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.102e+04  1.601e+02  256.25   <2e-16 ***
## agebt       -1.106e+05  5.281e+02 -209.44   <2e-16 ***
## 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)
```

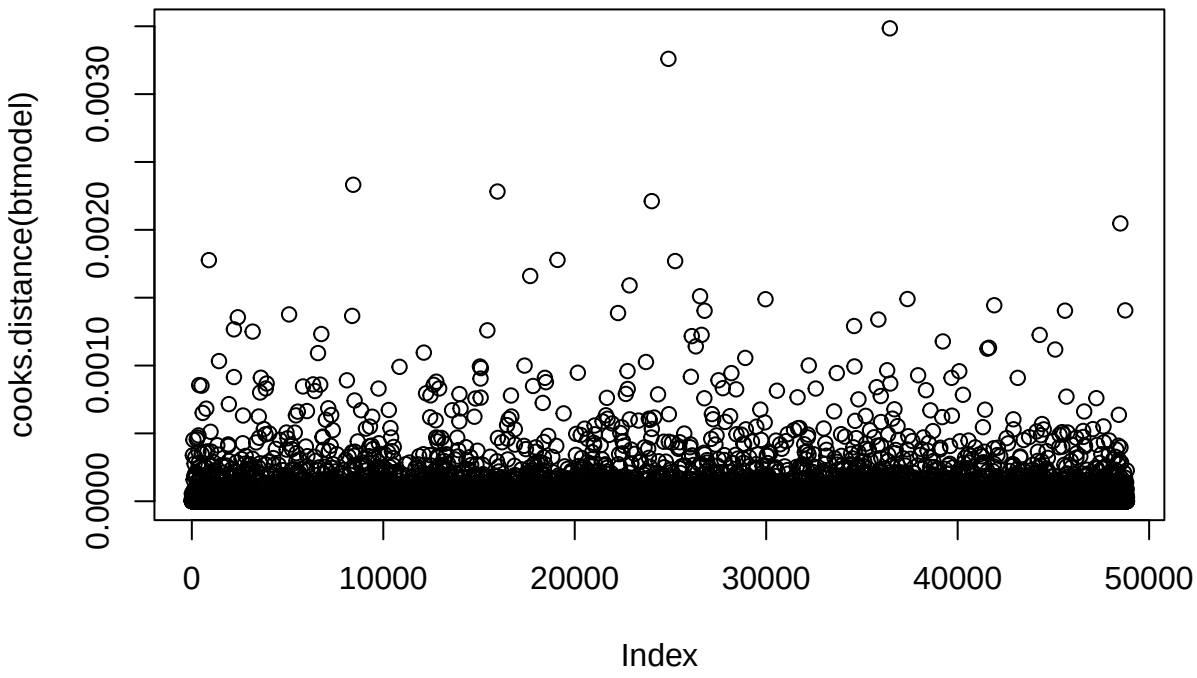



Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week)
48496
24023
36458
−6
−4
−2
0
2
4
6
8
−4
−2
0
2
4


Scale−Location
48496
24023
36458
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
30000
40000
50000
60000
70000
80000
Fitted values
lm(income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week)




```
#Check cook's distance
plot(cooks.distance(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)
```

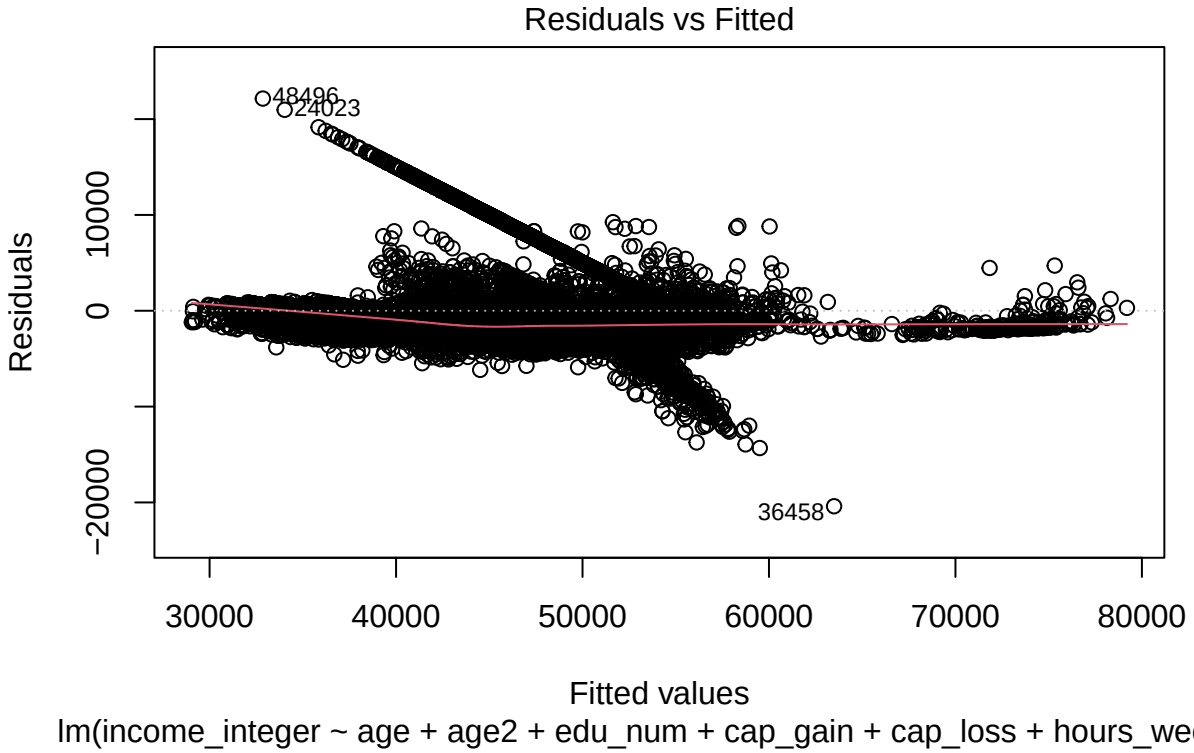
```
##
## Call:
## lm(formula = income_integer ~ age + age2 + edu_num + cap_gain +
##     cap_loss + hours_week + hours_week2, data = dd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20398.4  -1917.5  -1288.9    -20.5  22142.2
##
## Coefficients:
##               Estimate Std. 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 ***
```

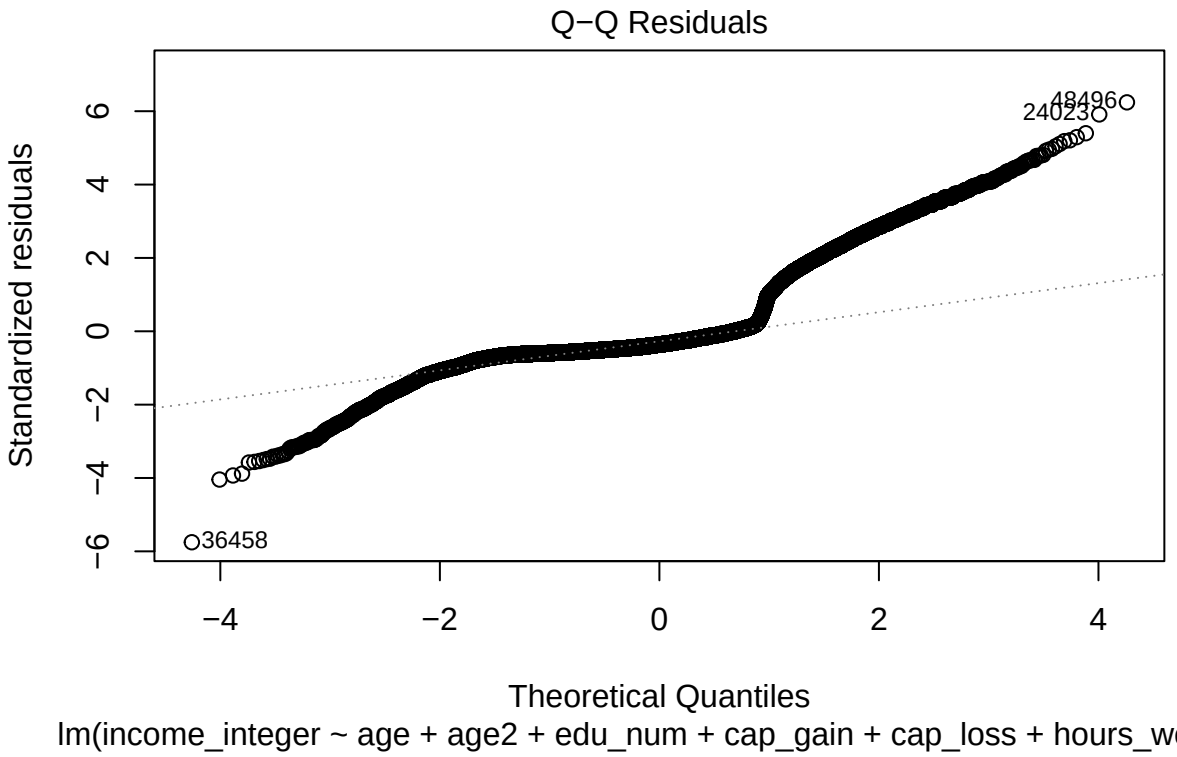
```
## hours_week2 -3.089e-01  4.847e-02  -6.374 1.85e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3548 on 48834 degrees of freedom
## Multiple R-squared:  0.6904, Adjusted R-squared:  0.6903
## F-statistic: 1.555e+04 on 7 and 48834 DF,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

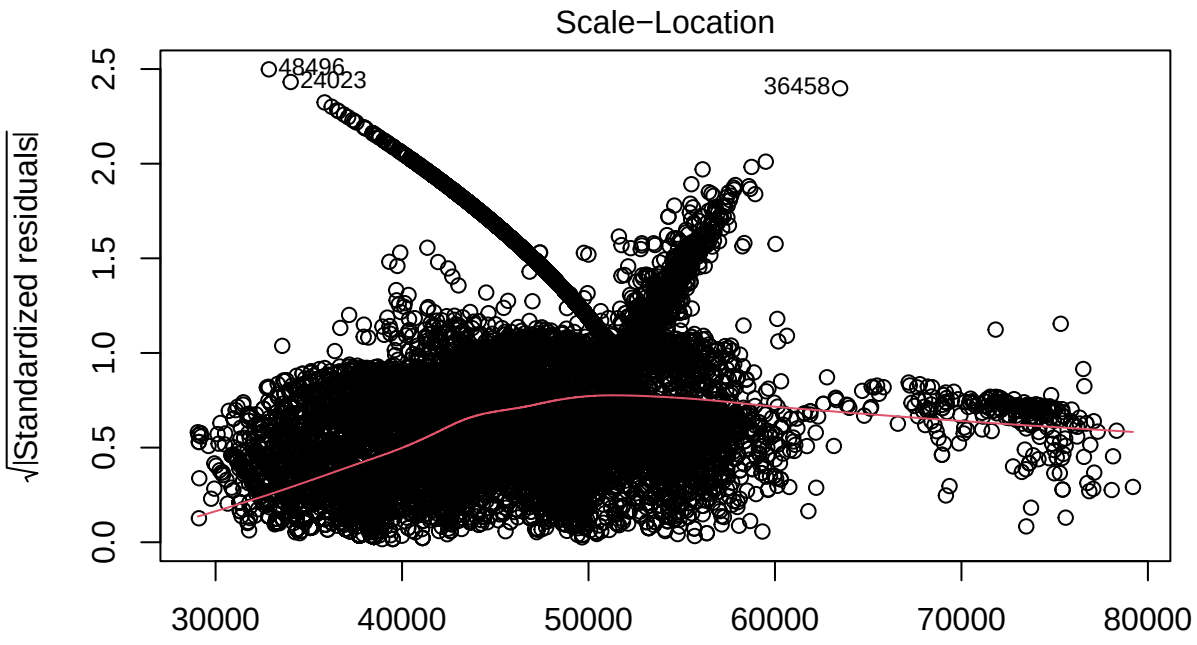
```
anova(initial_model, model_poly)
```

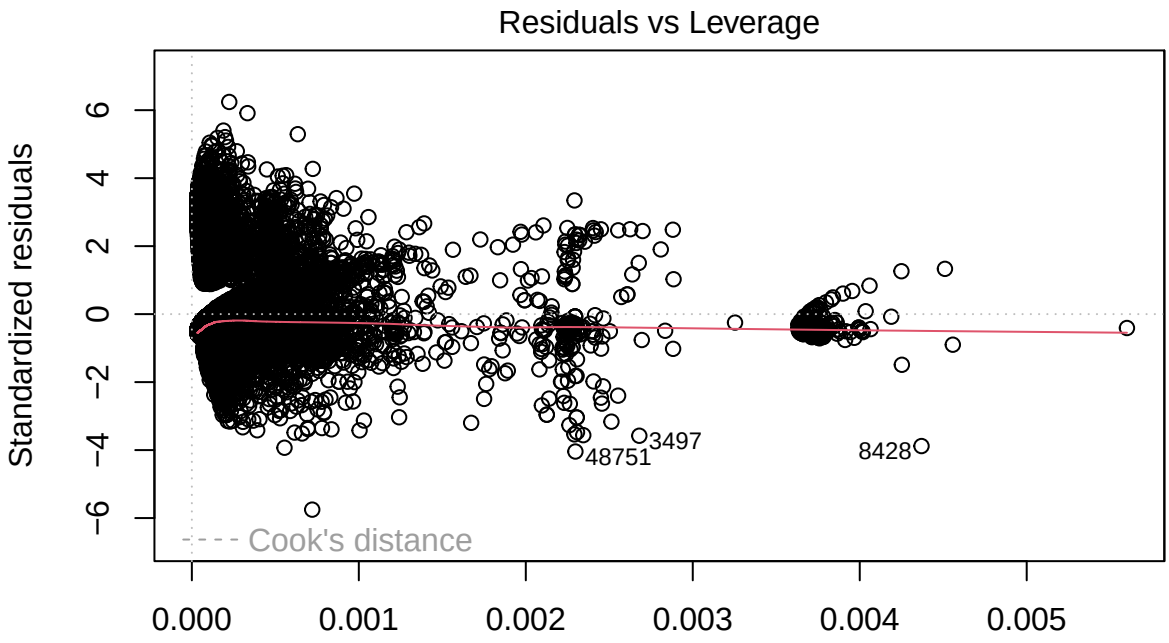
```
## Analysis of Variance Table
##
## Model 1: income_integer ~ age + edu_num + cap_gain + cap_loss + hours_week
## Model 2: income_integer ~ age + age2 + edu_num + cap_gain + cap_loss +
##     hours_week + hours_week2
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  48836 6.5856e+11
## 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)
```




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(income_integer ~ age + age2 + edu_num + cap_gain + cap_loss + hours_week .
48496
24023
36458
−6
−4
−2
0
2
4
6


Scale−Location
48496
24023
36458
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
30000
40000
50000
60000
70000
80000
Fitted values
lm(income_integer ~ age + age2 + edu_num + cap_gain + cap_loss + hours_week .




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

```
## Analysis of Variance Table
##
## Model 1: income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week
## Model 2: income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  48836 6.1019e+11
## 2  48823 5.9657e+11 13 1.3613e+10 85.698 < 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: income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation
## Model 2: income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
```

```
##     occupation + marital
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1  48823 5.9657e+11
## 2  48819 5.3488e+11  4 6.169e+10 1407.6 < 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 = 0.008045 ***
```

```
## Analysis of Variance Table
##
## Model 1: income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital
## Model 2: income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex
##   Res.Df        RSS Df Sum of Sq     F   Pr(>F)
## 1  48819 5.3488e+11
## 2  48818 5.3481e+11  1  76947979 7.024 0.008045 **
## ---
## 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: income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex
## Model 2: income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  48818 5.3481e+11
## 2  48812 5.3341e+11  6 1393156238 21.248 < 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: income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass
## Model 2: income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass + relationship
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  48812 5.3341e+11
## 2  48807 5.3054e+11  5 2875841238 52.913 < 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 = 1.172e-08
```

```
## Analysis of Variance Table
##
## Model 1: income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass + relationship
## Model 2: income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass + relationship + race
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1  48807 5.3054e+11
## 2  48803 5.3007e+11  4 464401605 10.689 1.172e-08 ***
## ---
## 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 = 5.832e-09
```

```
## Analysis of Variance Table
##
## Model 1: income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass + relationship + race
## Model 2: income_integer ~ agebt + edu_num_bt + 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 5.3007e+11
## 2  48802 5.2970e+11  1 367977806 33.902 5.832e-09 ***
## ---
## 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: income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass + relationship + race +
##     native_country
## Model 2: income_integer ~ agebt + edu_num_bt + 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 5.2970e+11
## 2  48801 1.7078e+11  1 3.5892e+11 102564 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Model with all significant variables including categorical variables
catmodel <- lm(income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week + occupation + ma
stepmodel <- stepAIC(catmodel, direction = "back")
```

```
## Start:  AIC=735998.8
## income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass + relationship + race +
##     native_country + income
##
##                   Df  Sum of Sq        RSS    AIC
## <none>                          1.7078e+11 735999
## - workclass        6 1.5589e+08 1.7094e+11 736031
## - cap_loss         1 1.3337e+08 1.7091e+11 736035
## - race             4 1.6108e+08 1.7094e+11 736037
## - sex              1 1.6094e+08 1.7094e+11 736043
## - native_country   1 3.6216e+08 1.7114e+11 736100
## - relationship     5 8.7310e+08 1.7165e+11 736238
## - marital          4 1.2041e+09 1.7198e+11 736334
## - occupation      13 1.7580e+09 1.7254e+11 736473
## - hours_week       1 7.5736e+09 1.7835e+11 738116
## - cap_gain         1 4.3699e+10 2.1448e+11 747125
## - edu_num_bt       1 1.2909e+11 2.9987e+11 763493
## - agebt            1 1.6731e+11 3.3809e+11 769353
## - income           1 3.5892e+11 5.2970e+11 791283
```

```r
vif(catmodel)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## agebt            1.953376  1        1.397632
## edu_num_bt       1.424807  1        1.193653
## cap_gain         1.070758  1        1.034774
## cap_loss         1.030525  1        1.015148
## hours_week       1.216447  1        1.102927
## occupation       2.250465 13        1.031689
## marital         60.754482  4        1.670888
## sex              1.990188  1        1.410740
## workclass        1.439311  6        1.030812
## relationship    77.803628  5        1.545610
## race             1.285289  4        1.031870
## native_country   1.239526  1        1.113340
## income           1.550355  1        1.245133
```

```r
summary(catmodel)
```

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

```
## -12724.0   -861.8    -19.9   1116.1  14390.3
##
## Coefficients:
##                               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                  4.215e+04  1.754e+02  240.260  < 2e-16 ***
## agebt                       -8.366e+04  3.826e+02 -218.655  < 2e-16 ***
## edu_num_bt                   4.450e+03  2.317e+01  192.061  < 2e-16 ***
## cap_gain                     1.313e-01  1.175e-03  111.746  < 2e-16 ***
## cap_loss                    -1.316e-01  2.132e-02   -6.173 6.73e-10 ***
## hours_week                   3.505e+01  7.534e-01   46.521  < 2e-16 ***
## occupationArmy              -7.698e+02  4.862e+02   -1.583  0.11334
## occupationCraftRep          -1.496e-01  3.754e+01   -0.004  0.99682
## occupationExecMan           -3.301e+02  3.649e+01   -9.046  < 2e-16 ***
## occupationFarmFish          -1.257e+02  5.841e+01   -2.153  0.03136 *
## occupationHandlCl           -4.723e+02  5.022e+01   -9.403  < 2e-16 ***
## occupationHouse             -3.851e+02  1.239e+02   -3.108  0.00188 **
## occupationMachOp            -3.081e+02  4.396e+01   -7.009 2.43e-12 ***
## occupationOther             -3.892e+02  3.740e+01  -10.406  < 2e-16 ***
## occupationProf              -3.559e+02  3.306e+01  -10.767  < 2e-16 ***
## occupationProtServ           2.877e+02  6.754e+01    4.259 2.06e-05 ***
## occupationSales             -9.442e+01  3.683e+01   -2.564  0.01036 *
## occupationTech               1.668e+02  5.561e+01    2.999  0.00271 **
## occupationTrans              6.186e+01  4.837e+01    1.279  0.20093
## maritalMarried               1.252e+02  1.050e+02    1.192  0.23309
## maritalNevMarr              -8.543e+01  3.236e+01   -2.640  0.00829 **
## maritalSep                  -1.982e+02  4.685e+01   -4.231 2.34e-05 ***
## maritalWidow                 8.888e+02  5.462e+01   16.272  < 2e-16 ***
## sexMale                     -1.720e+02  2.537e+01   -6.782 1.20e-11 ***
## workclassLoc                 5.614e+01  6.088e+01    0.922  0.35641
## workclassNoPay               1.687e+02  3.404e+02    0.496  0.62021
## workclassPriv                7.450e+01  5.170e+01    1.441  0.14964
## workclassSelfI              -2.118e+02  6.884e+01   -3.076  0.00210 **
## workclassSelfN               8.091e+01  5.996e+01    1.349  0.17723
## workclassState              -5.482e+01  6.558e+01   -0.836  0.40322
## relationshipNot-in-family    1.319e+02  1.046e+02    1.261  0.20740
## relationshipOther-relative -2.061e+02  1.027e+02   -2.006  0.04484 *
## relationshipOwn-child      -1.867e+02  1.046e+02   -1.784  0.07435 .
## relationshipUnmarried      -2.483e+02  1.085e+02   -2.290  0.02205 *
## relationshipWife            2.477e+02  4.769e+01    5.193 2.07e-07 ***
## raceAsian-Pac-Islander      4.592e+02  1.014e+02    4.527 5.99e-06 ***
## raceBlack                   9.471e+01  9.070e+01    1.044  0.29636
## raceOther                  -1.795e+01  1.278e+02   -0.140  0.88827
## raceWhite                   1.511e+02  8.703e+01    1.736  0.08260 .
## native_countryUSA           3.160e+02  3.106e+01   10.173  < 2e-16 ***
## income>50K                  7.911e+03  2.470e+01  320.256  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1871 on 48801 degrees of freedom
## Multiple R-squared:  0.914,  Adjusted R-squared:  0.9139
## F-statistic: 1.296e+04 on 40 and 48801 DF,  p-value: < 2.2e-16
```

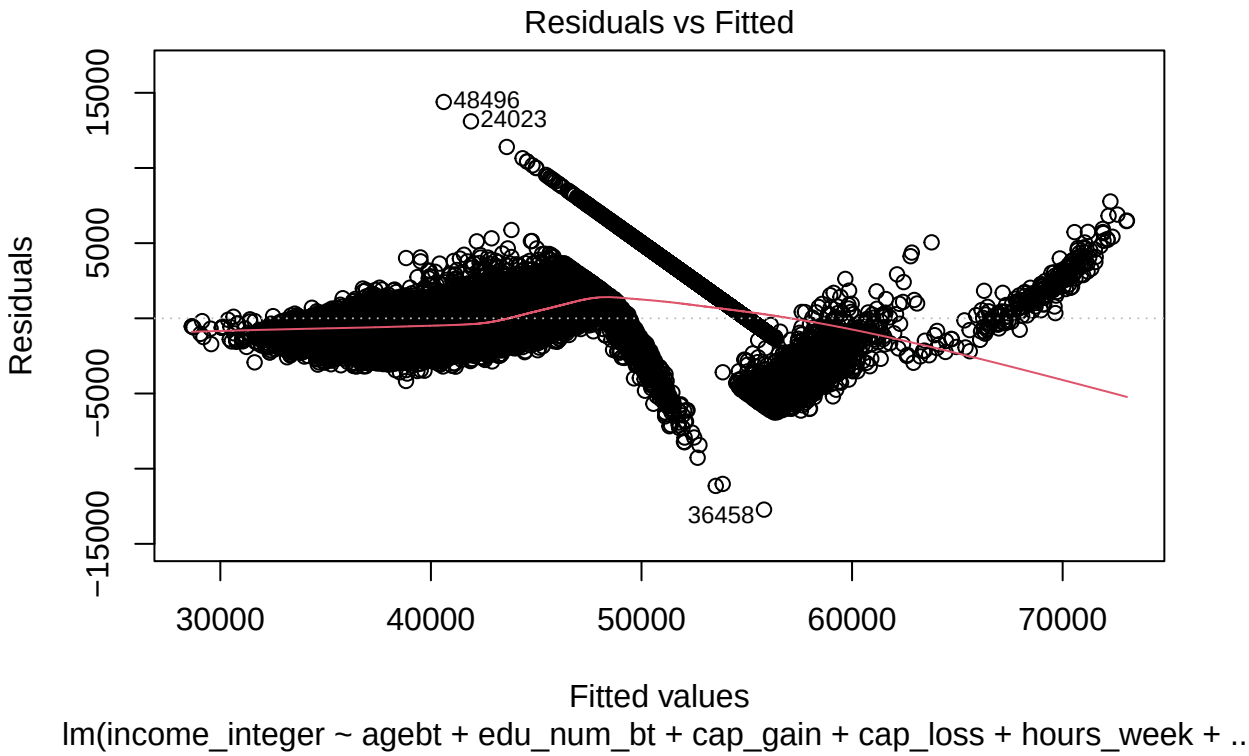
```
anova(catmodel)
```

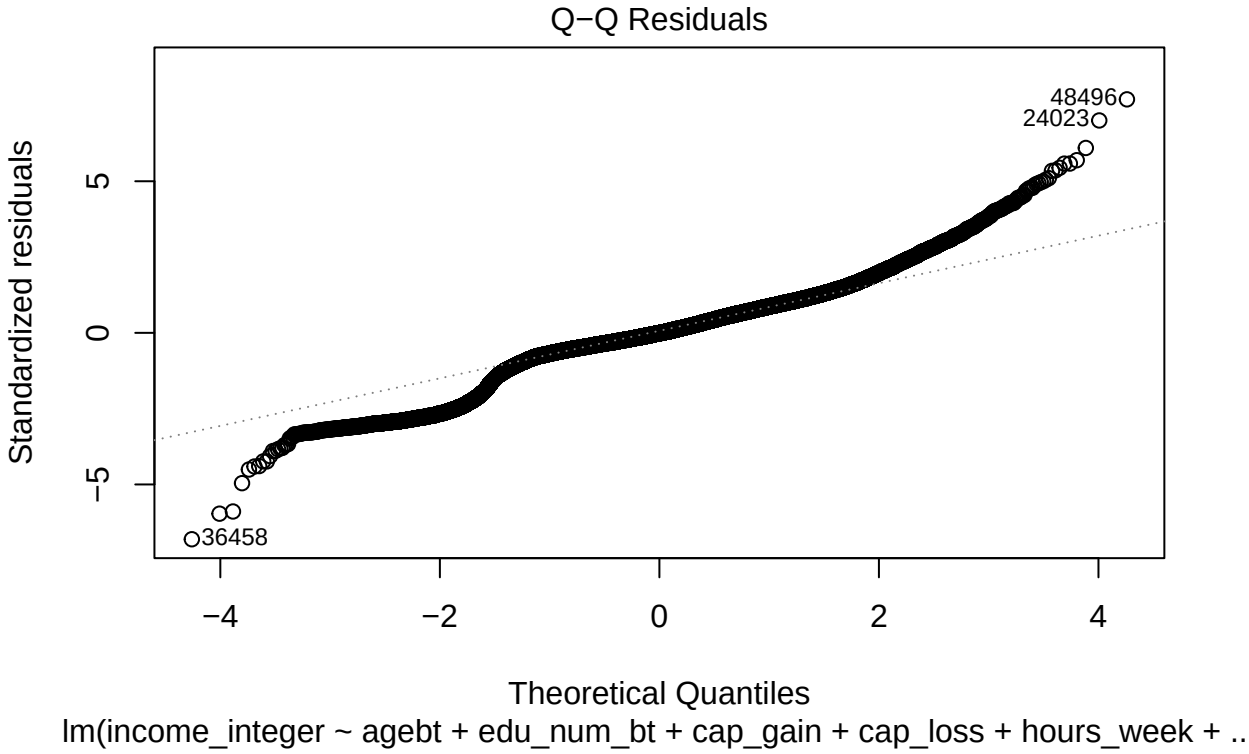
```
## Analysis of Variance Table
##
## Response: income_integer
##                   Df     Sum Sq    Mean Sq    F value    Pr(>F)
## agebt              1 7.4820e+11 7.4820e+11 213799.669 < 2.2e-16 ***
## edu_num_bt         1 4.6514e+11 4.6514e+11 132915.257 < 2.2e-16 ***
## cap_gain           1 1.1886e+11 1.1886e+11  33964.116 < 2.2e-16 ***
## cap_loss           1 6.3326e+09 6.3326e+09   1809.554 < 2.2e-16 ***
## hours_week         1 3.6479e+10 3.6479e+10  10424.102 < 2.2e-16 ***
## occupation        13 1.3613e+10 1.0472e+09    299.228 < 2.2e-16 ***
## marital            4 6.1690e+10 1.5423e+10   4407.065 < 2.2e-16 ***
## sex                1 7.6948e+07 7.6948e+07     21.988 2.751e-06 ***
## workclass          6 1.3932e+09 2.3219e+08     66.350 < 2.2e-16 ***
## relationship       5 2.8758e+09 5.7517e+08    164.357 < 2.2e-16 ***
## race               4 4.6440e+08 1.1610e+08     33.176 < 2.2e-16 ***
## native_country     1 3.6798e+08 3.6798e+08    105.151 < 2.2e-16 ***
## income             1 3.5892e+11 3.5892e+11 102563.829 < 2.2e-16 ***
## Residuals      48801 1.7078e+11 3.4995e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

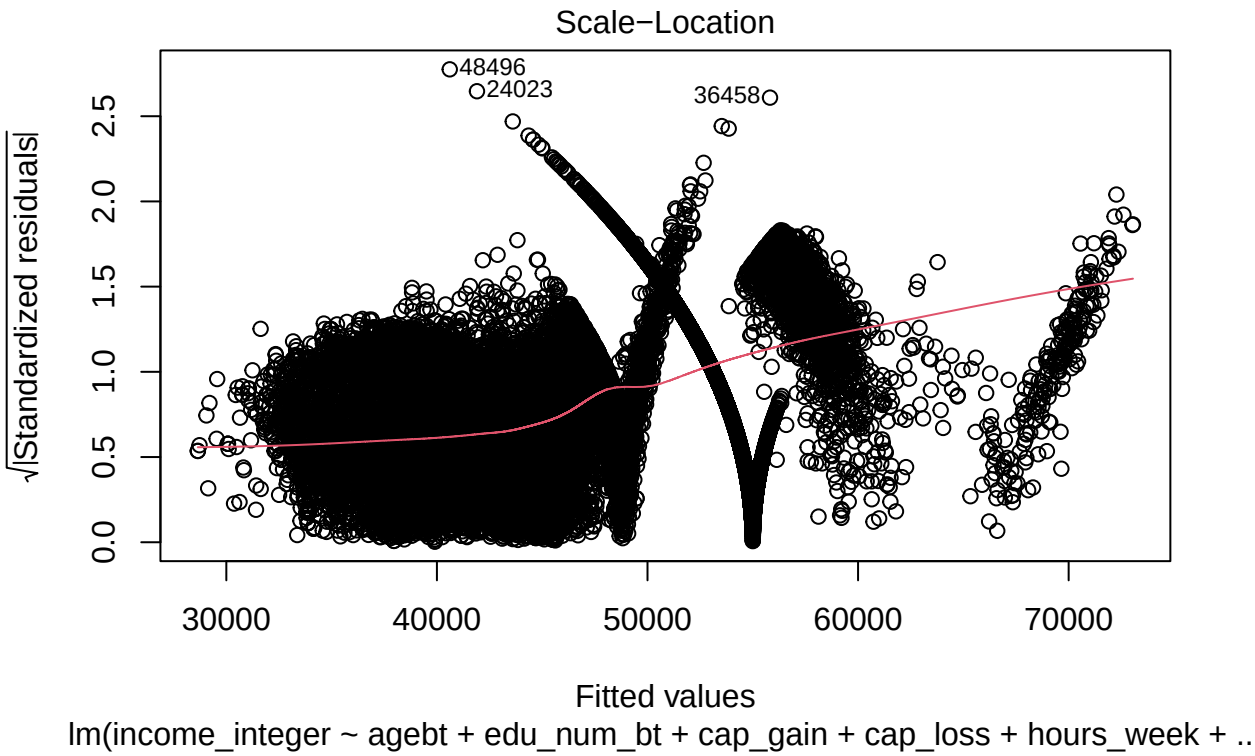
```
anova(btmodel, catmodel)
```

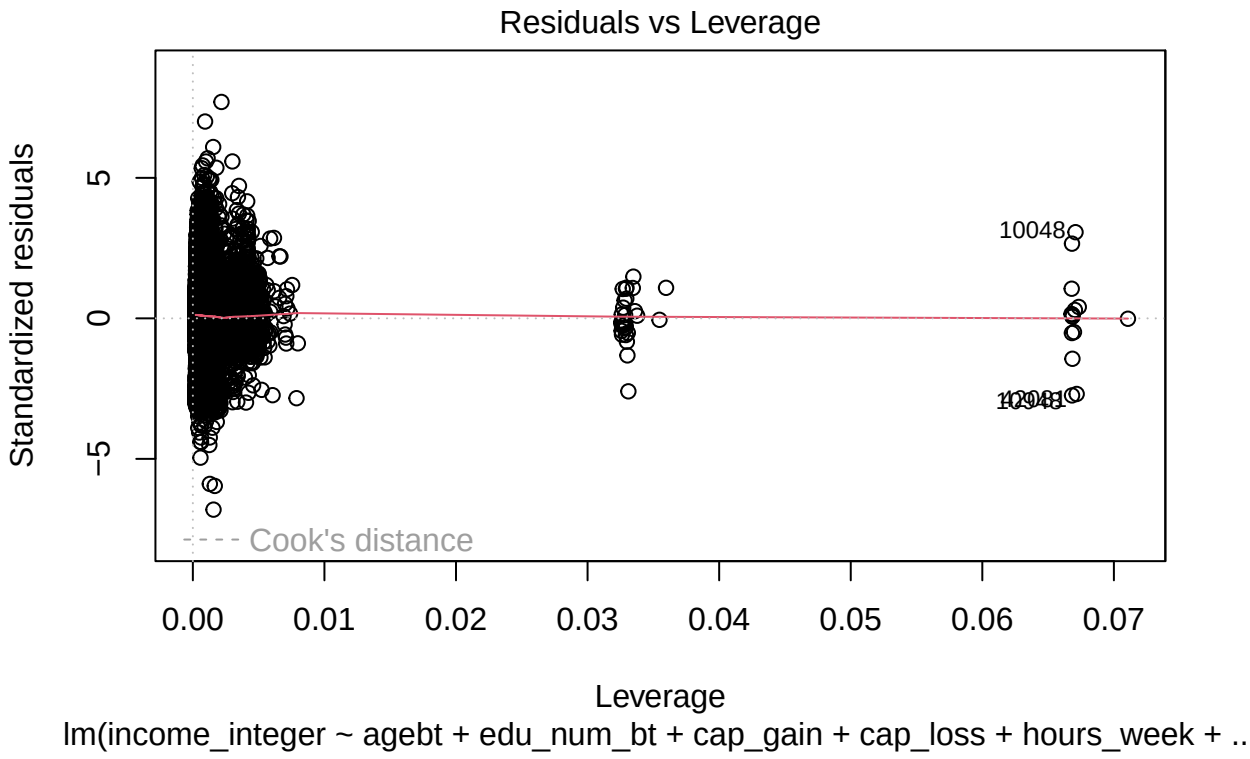
```
## Analysis of Variance Table
##
## Model 1: income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week
## Model 2: income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass + relationship + race +
##     native_country + income
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  48836 6.1019e+11
## 2  48801 1.7078e+11 35 4.3941e+11 3587.5 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(catmodel)
```

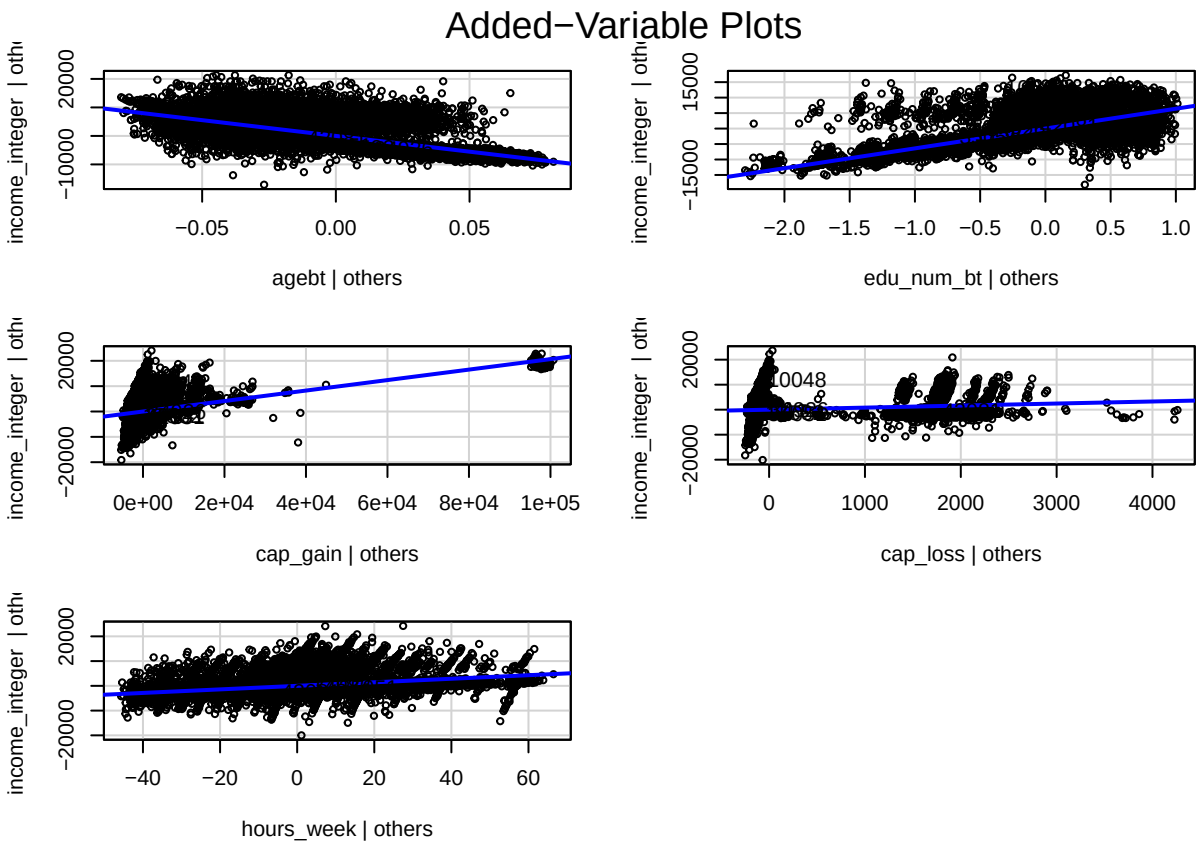

Residuals vs Fitted
Residuals
15000
5000
−5000
−15000
48496
24023
36458
30000
40000
50000
60000
70000
Fitted values
lm(income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week + ..


Q−Q Residuals
Standardized residuals
5
0
−5
48496
24023
36458
−4
−2
0
2
4
Theoretical Quantiles
lm(income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week + ..

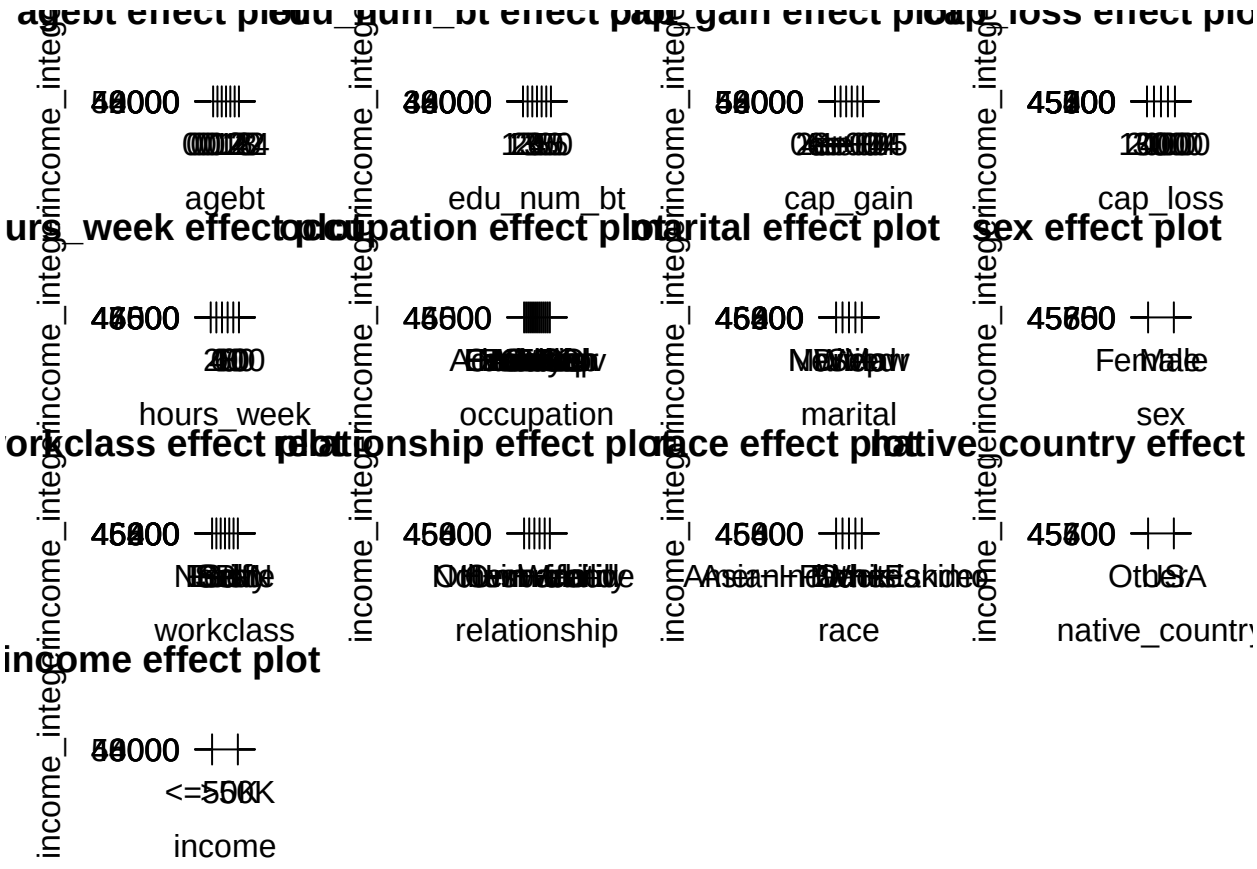

Scale−Location
48496
24023
36458
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
30000
40000
50000
60000
70000
Fitted values
lm(income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week + ..




```
#Possible Conclusio
avPlots(btmodel,id=list(method=hatvalues(catmodel),n=5))
```




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


agebt effect plot
edu_num_bt effect plot
cap_gain effect plot
cap_loss effect plot
hours_week effect plot
occupation effect plot
marital effect plot
sex effect plot
workclass effect plot
relationship effect plot
race effect plot
native_country effect
income effect plot
income_integer
agebt
edu_num_bt
cap_gain
cap_loss
hours_week
occupation
marital
sex
Female
Male
workclass
relationship
race
native_country
OtherUSA
income
<=50K