Deliverable 2

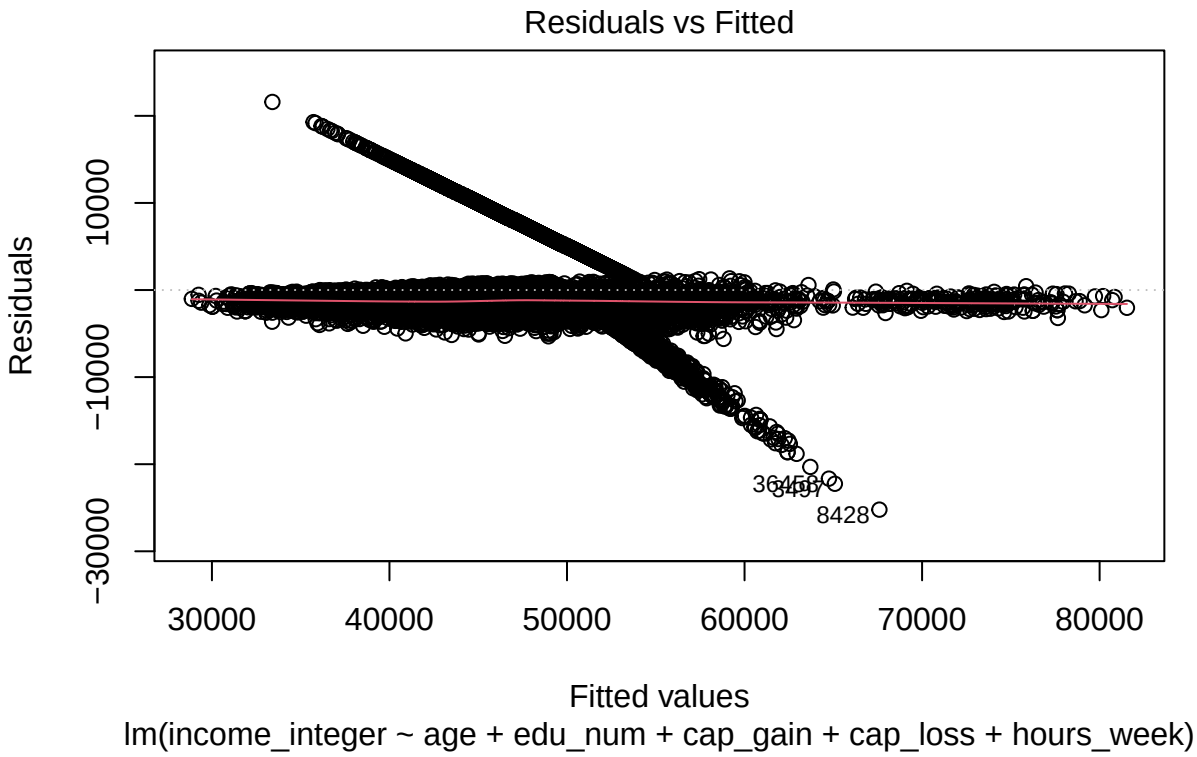
2025-04-28

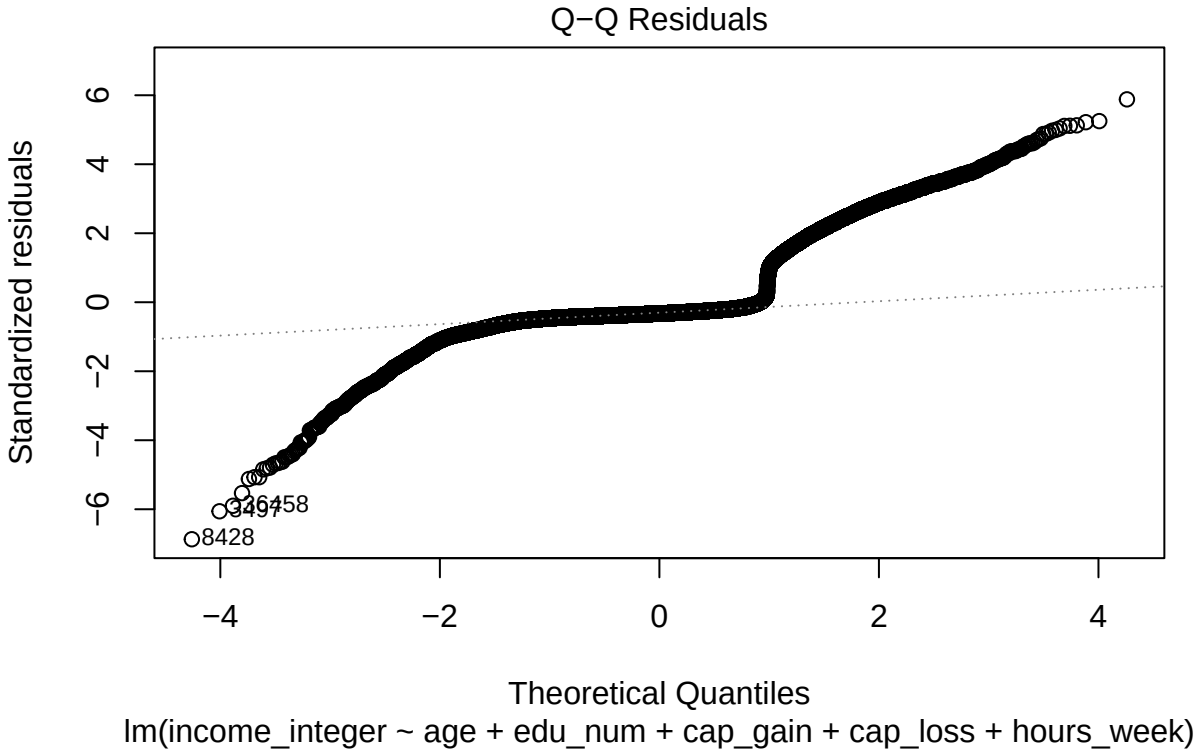
```r
setwd("~/Escritorio/ADEI/D2")
dd <- read.csv("adult_def.csv")
```

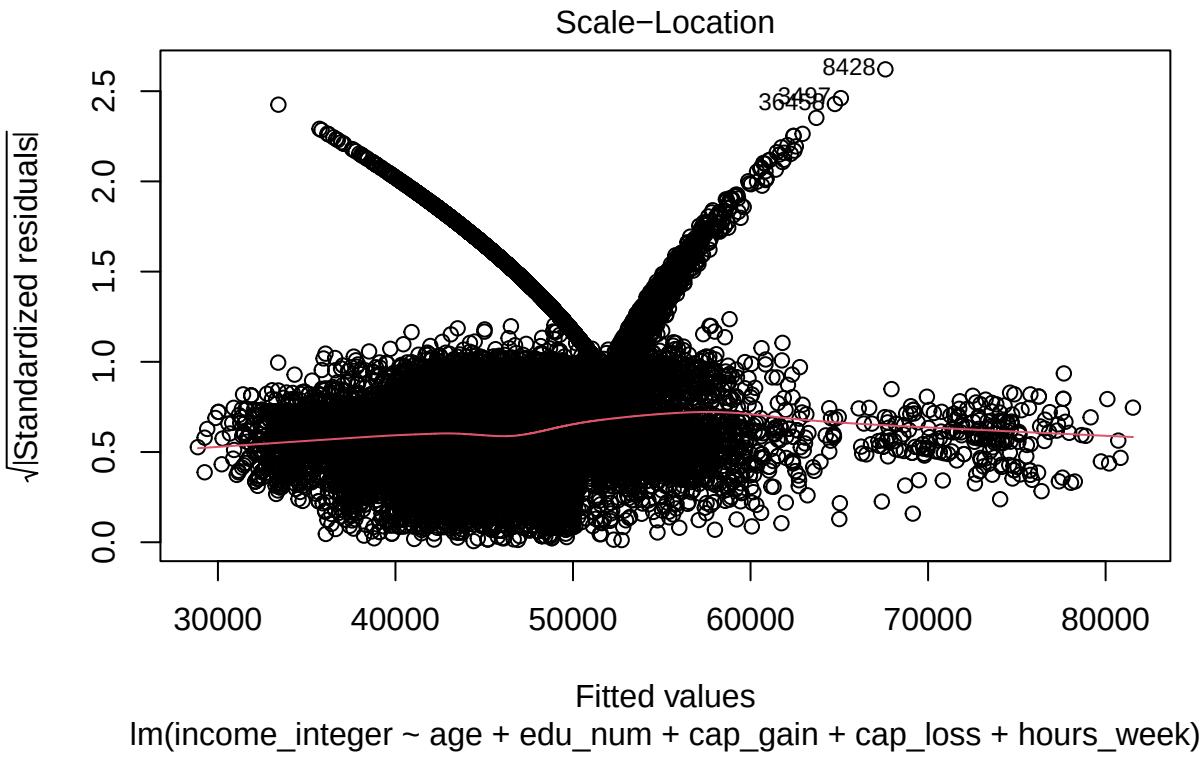
```r
initial_model <- lm(income_integer ~ age + edu_num + cap_gain + cap_loss + hours_week, data = dd)
summary(initial_model)
```

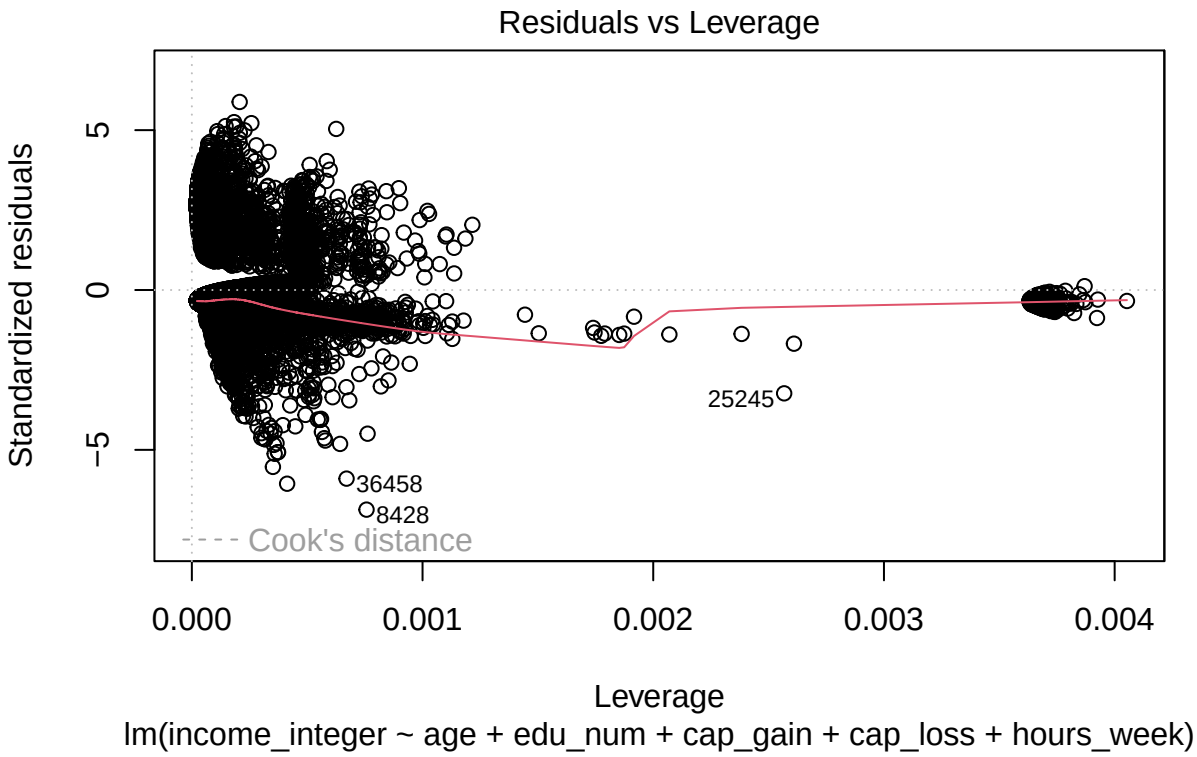
```
##
## Call:
## lm(formula = income_integer ~ age + edu_num + cap_gain + cap_loss +
##     hours_week, data = dd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25216.0  -1522.3  -1201.8   -699.4  21594.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)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

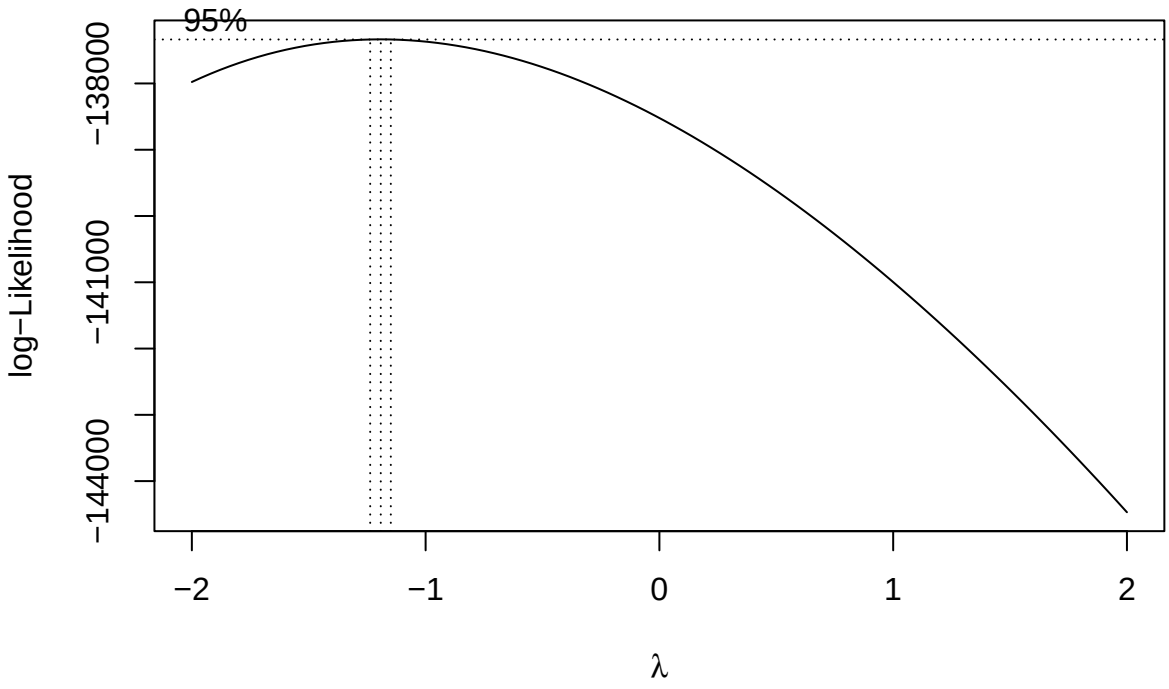

Residuals vs Fitted
Residuals
10000
−10000
−30000
30000
40000
50000
60000
70000
80000
8428
Fitted values
lm(income_integer ~ age + edu_num + cap_gain + cap_loss + hours_week)


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)
```

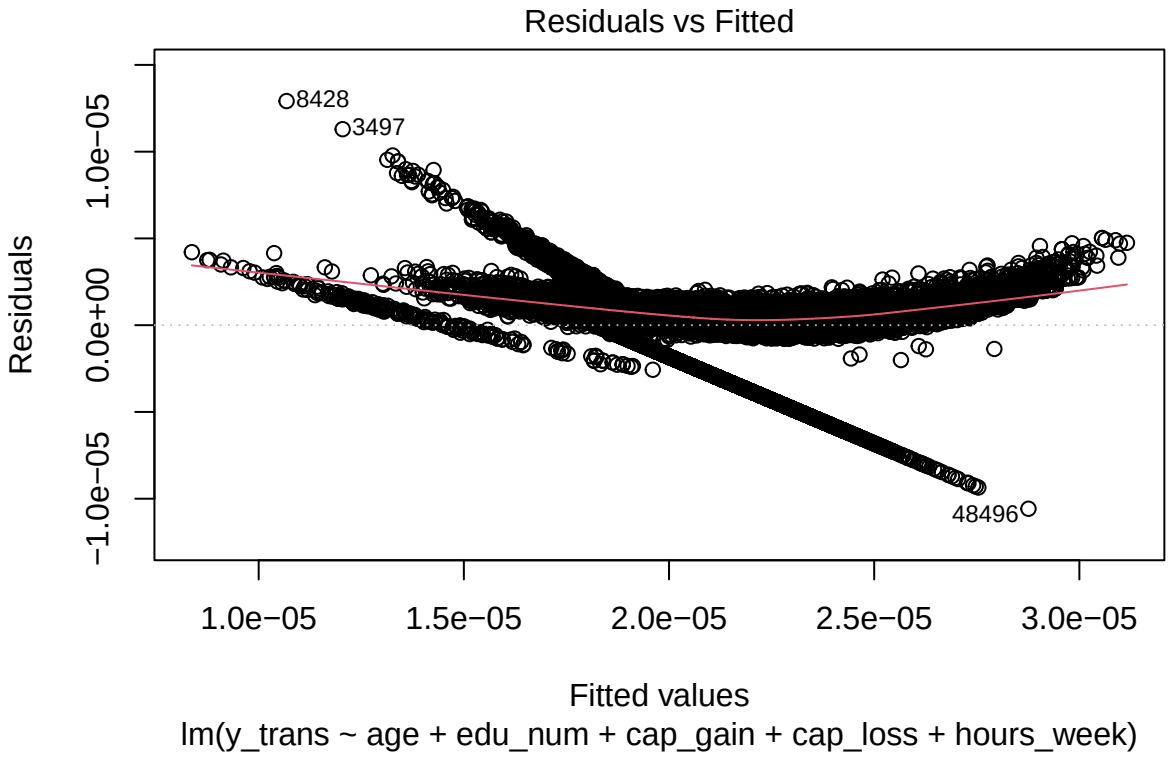


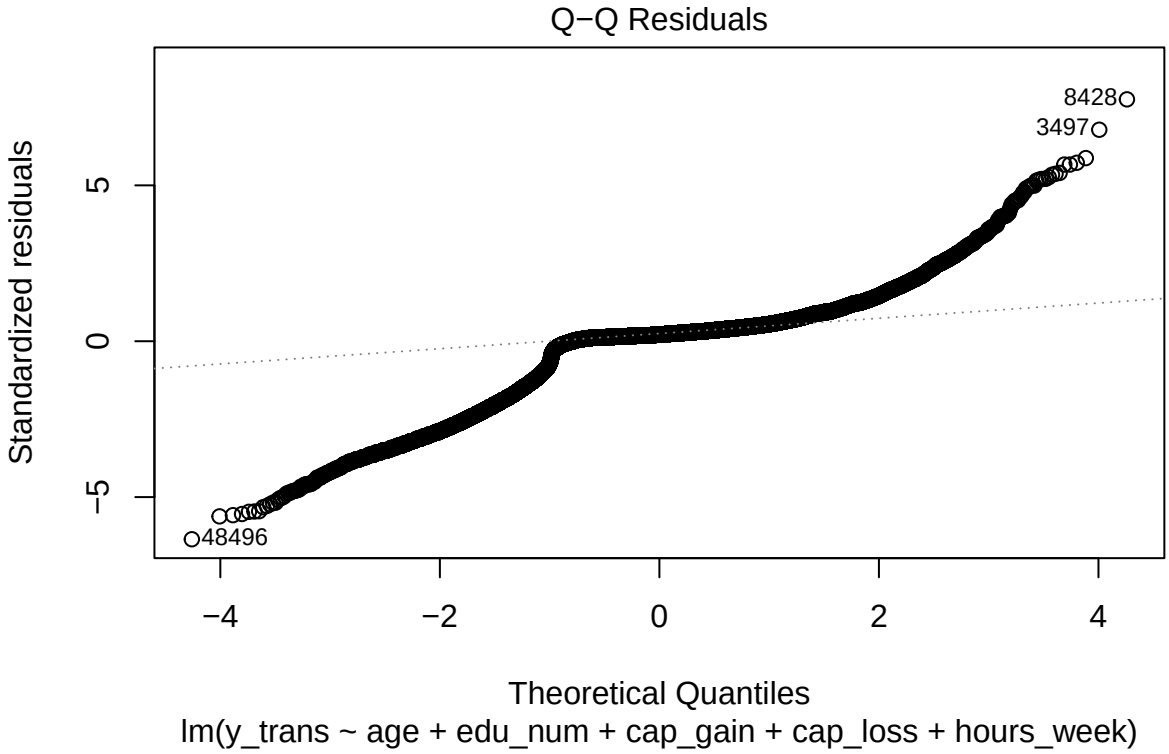

```
#given lambda is aproximatedly -1 we do the inverse transformation
y_trans <- 1 / dd$income_integer
transformed_model <- lm(y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week, data = dd)
summary(transformed_model)
```

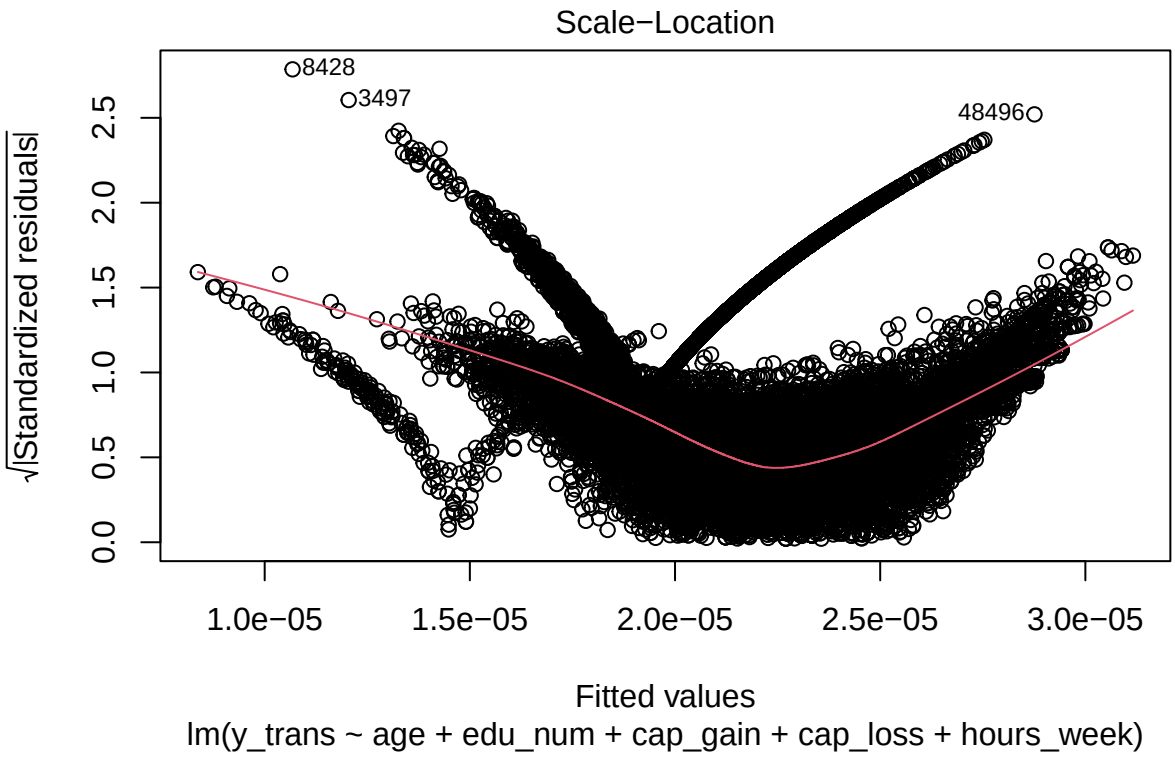
```
##
## Call:
## lm(formula = y_trans ~ age + edu_num + cap_gain + cap_loss +
##     hours_week, data = dd)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.058e-05  1.373e-07  3.552e-07  6.859e-07  1.292e-05
##
## Coefficients:
##               Estimate Std. Error t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

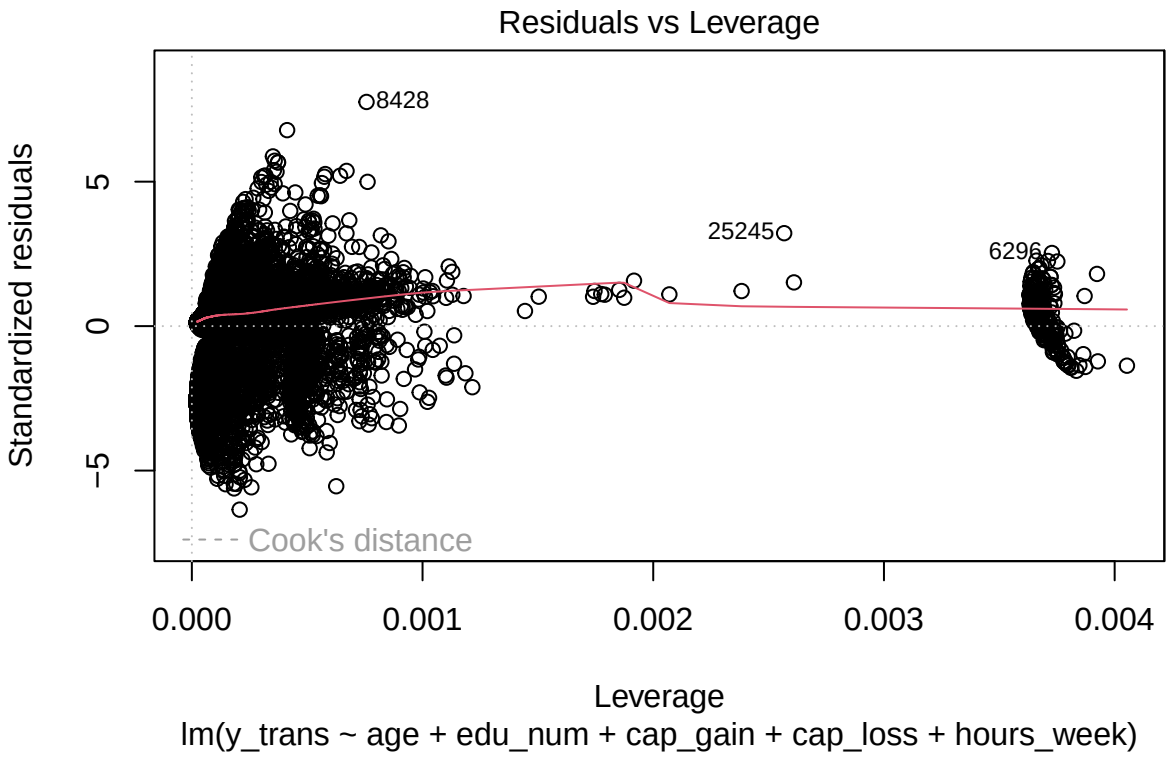
```
## Multiple R-squared:  0.7109, Adjusted R-squared:  0.7109
## F-statistic: 2.402e+04 on 5 and 48836 DF,  p-value: < 2.2e-16
```

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

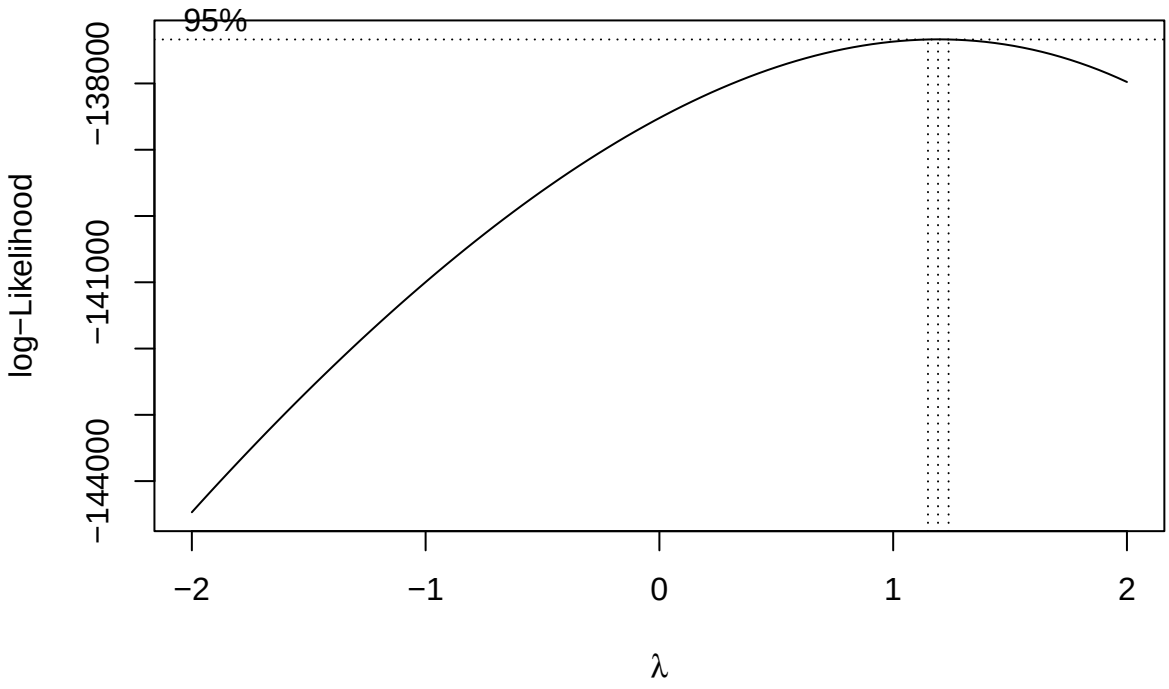



Q−Q Residuals
Standardized residuals
5
0
−5
8428
3497
48496
−4
−2
0
2
4
Theoretical Quantiles
lm(y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week)


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)
```

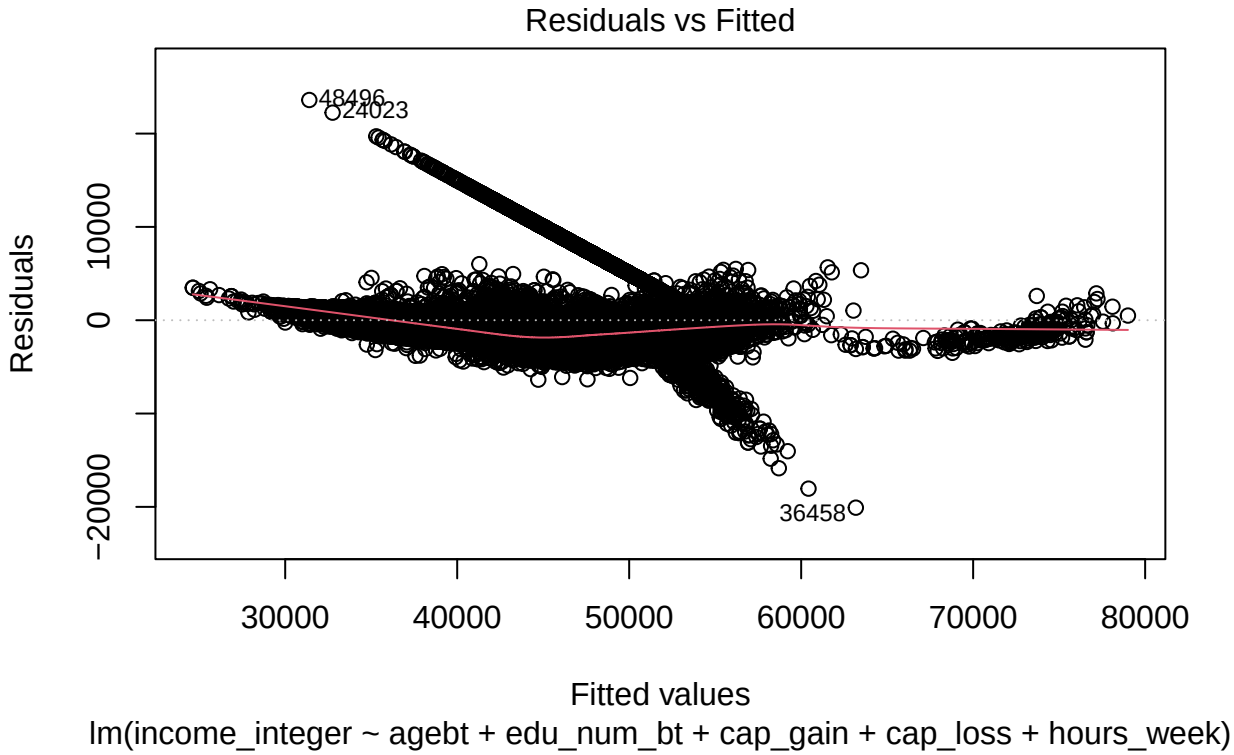
```
##             MLE of lambda Score Statistic (t) Pr(>|t|)
## age              -0.36944            -52.9070   <2e-16 ***
## edu_num           0.71757            -13.2951   <2e-16 ***
## hours_week        0.95940             -0.2224    0.824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05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

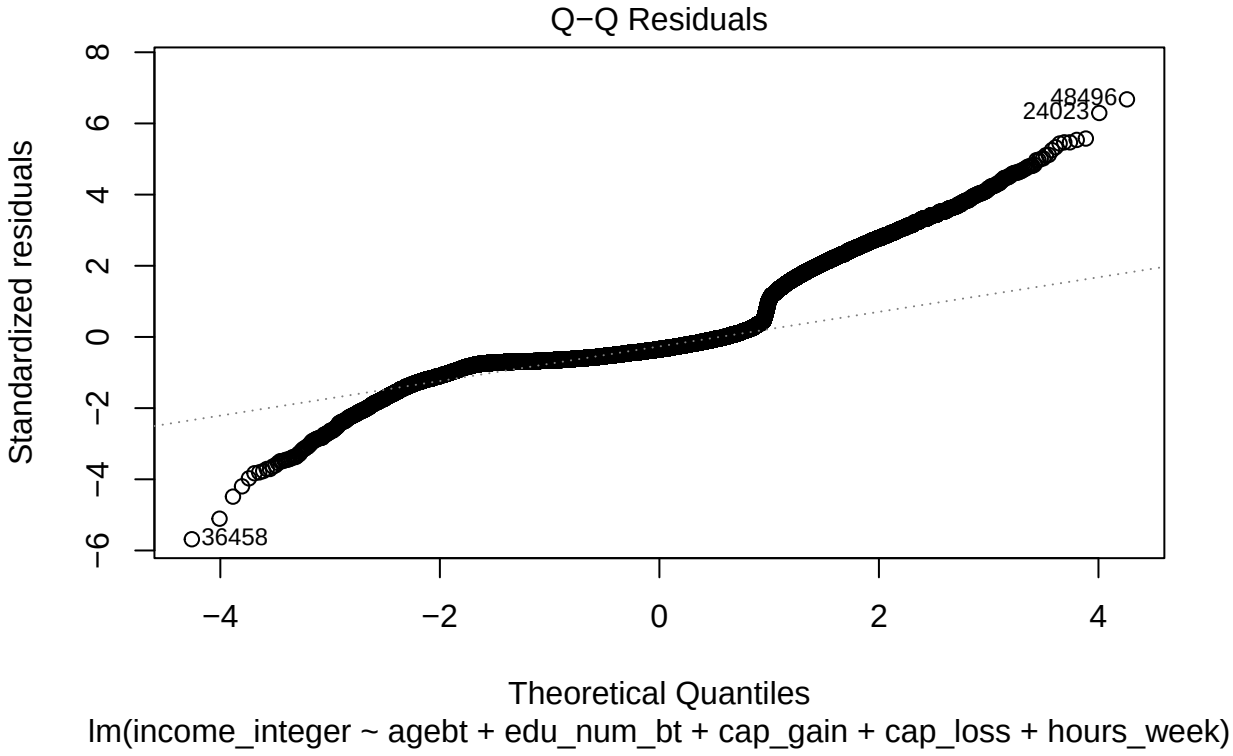
```
agebt <- 1 / sqrt(dd$age)
edu_num_bt <- sqrt(dd$edu_num)
btmodel <- lm(income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week, data = dd)
summary(btmodel)
```

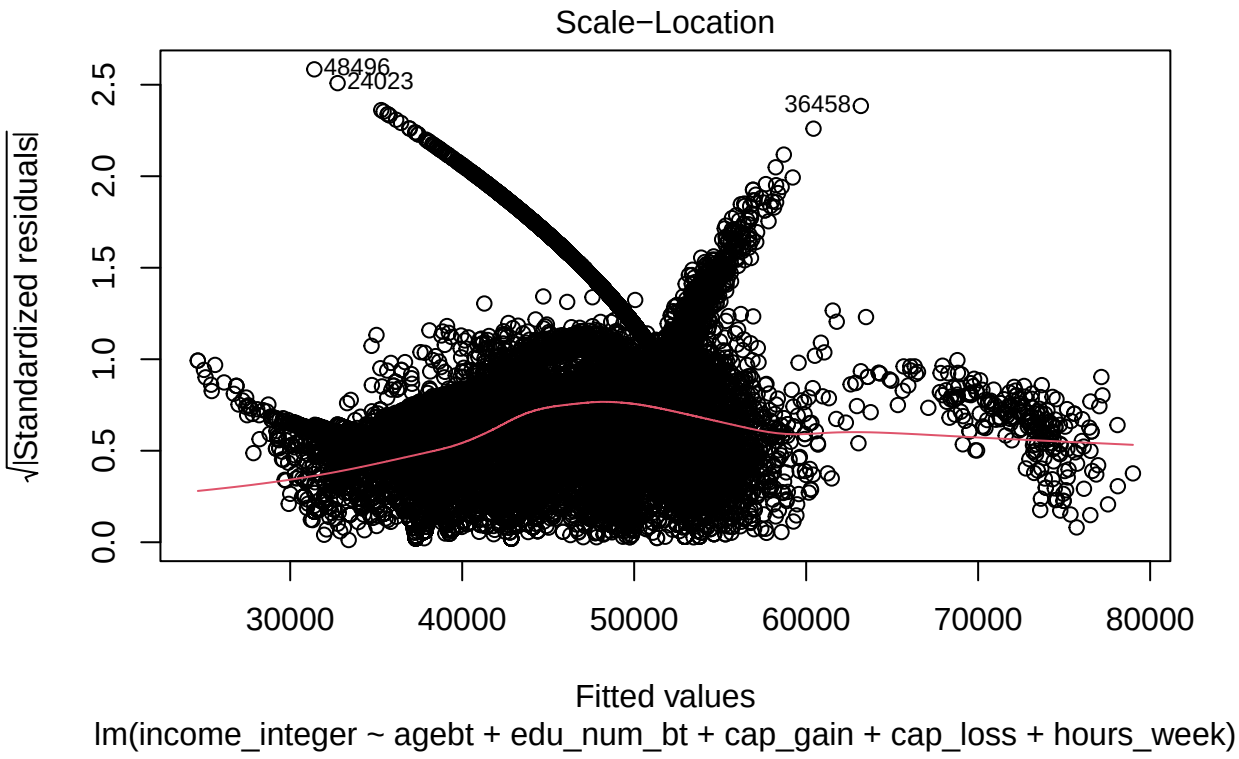
```
##
## Call:
## lm(formula = income_integer ~ agebt + edu_num_bt + cap_gain +
##     cap_loss + hours_week, data = dd)
##
```

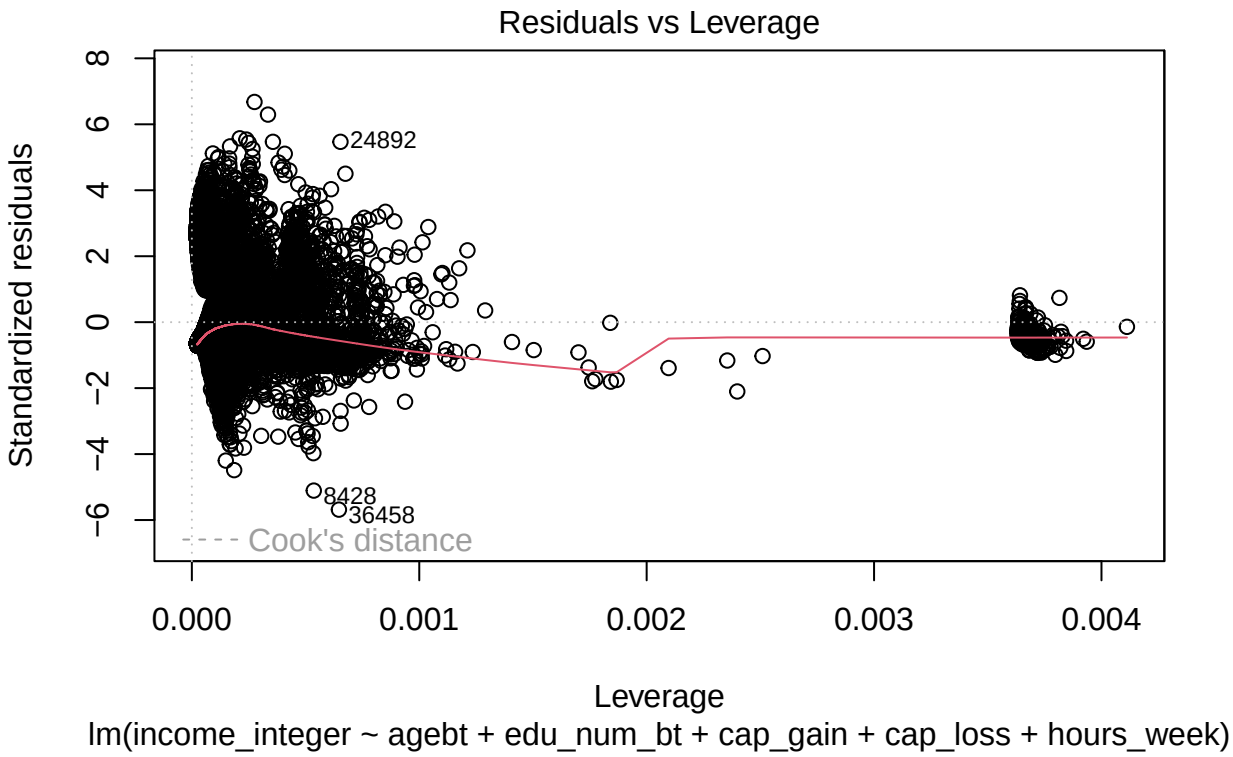
```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20085.5  -2100.2  -1218.1    216.9  23598.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.102e+04  1.601e+02  256.25   <2e-16 ***
## agebt       -1.106e+05  5.281e+02 -209.44   <2e-16 ***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

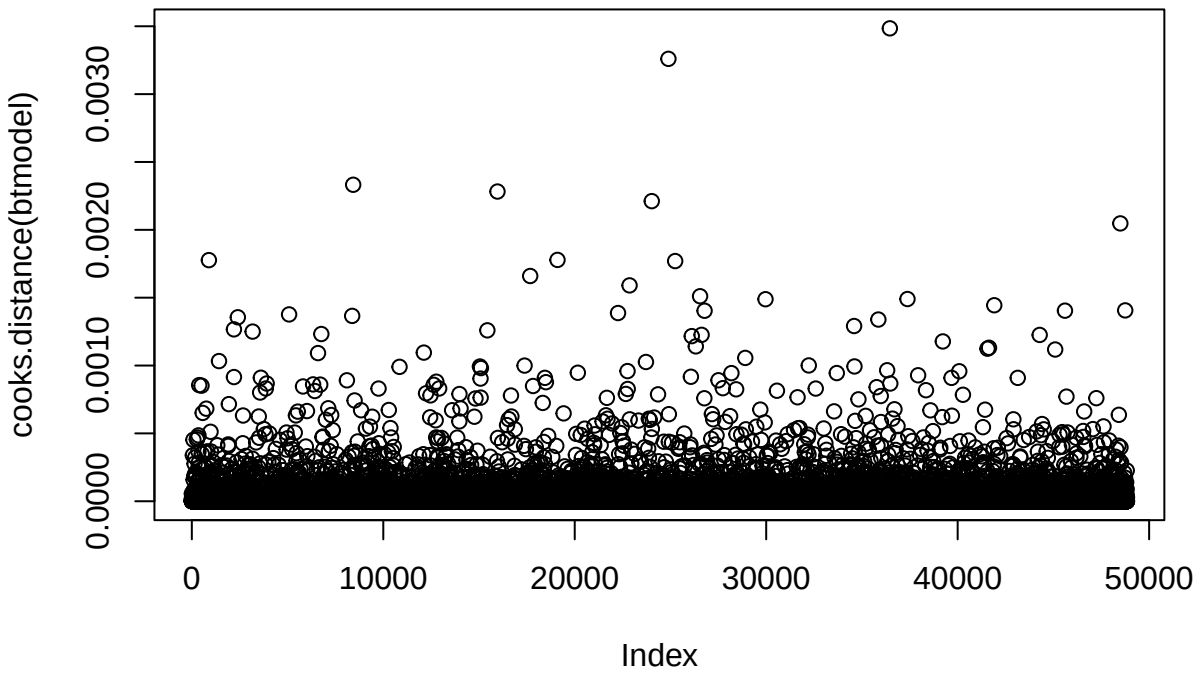



Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(income_integer ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week)
48496
24023
36458
−6
−4
−2
0
2
4
6
8
−4
−2
0
2
4


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)
```

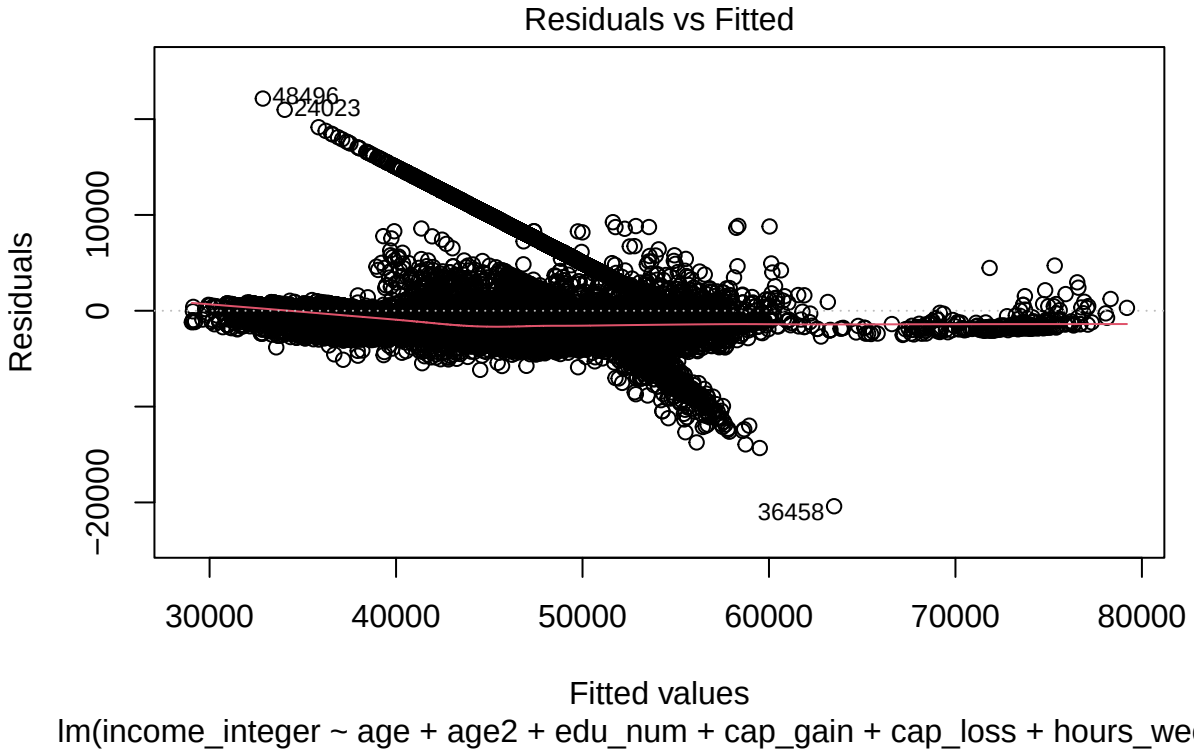
```
##
## Call:
## lm(formula = income_integer ~ age + age2 + edu_num + cap_gain +
##     cap_loss + hours_week + hours_week2, data = dd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20398.4  -1917.5  -1288.9    -20.5  22142.2
##
## Coefficients:
##               Estimate Std.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

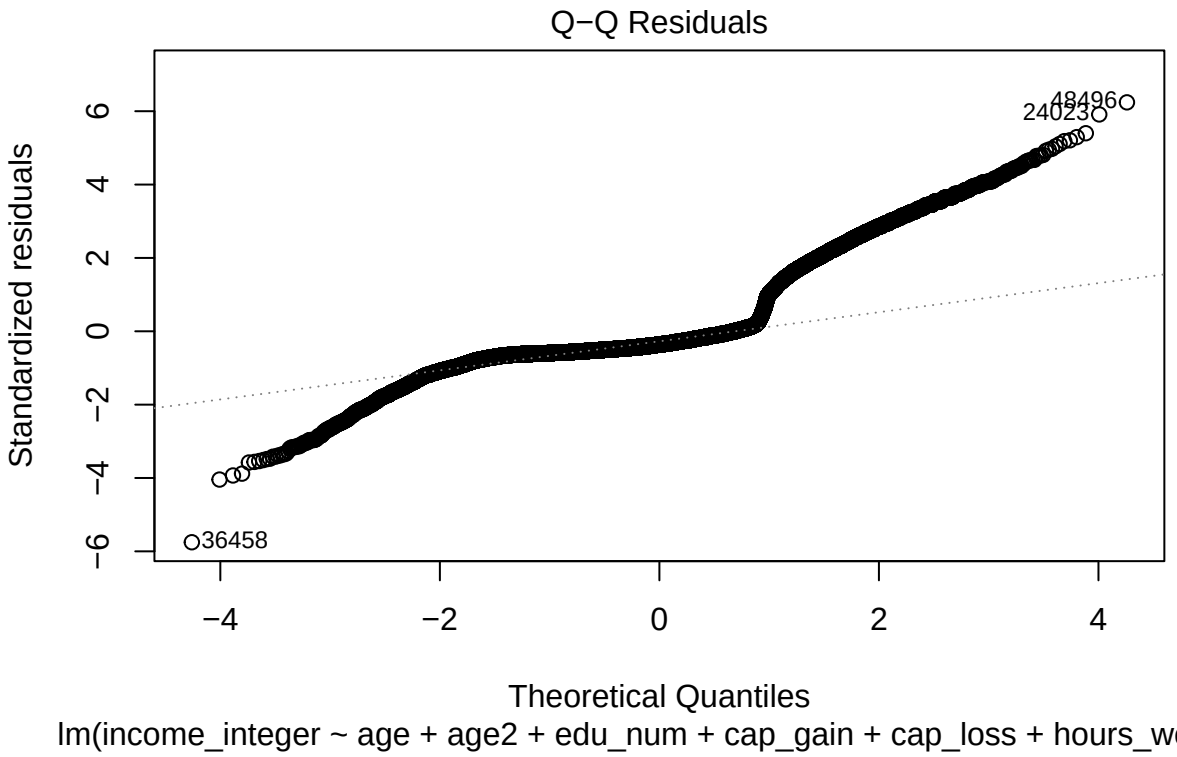
```
## hours_week2 -3.089e-01  4.847e-02  -6.374 1.85e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3548 on 48834 degrees of freedom
## Multiple R-squared:  0.6904, Adjusted R-squared:  0.6903
## F-statistic: 1.555e+04 on 7 and 48834 DF,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

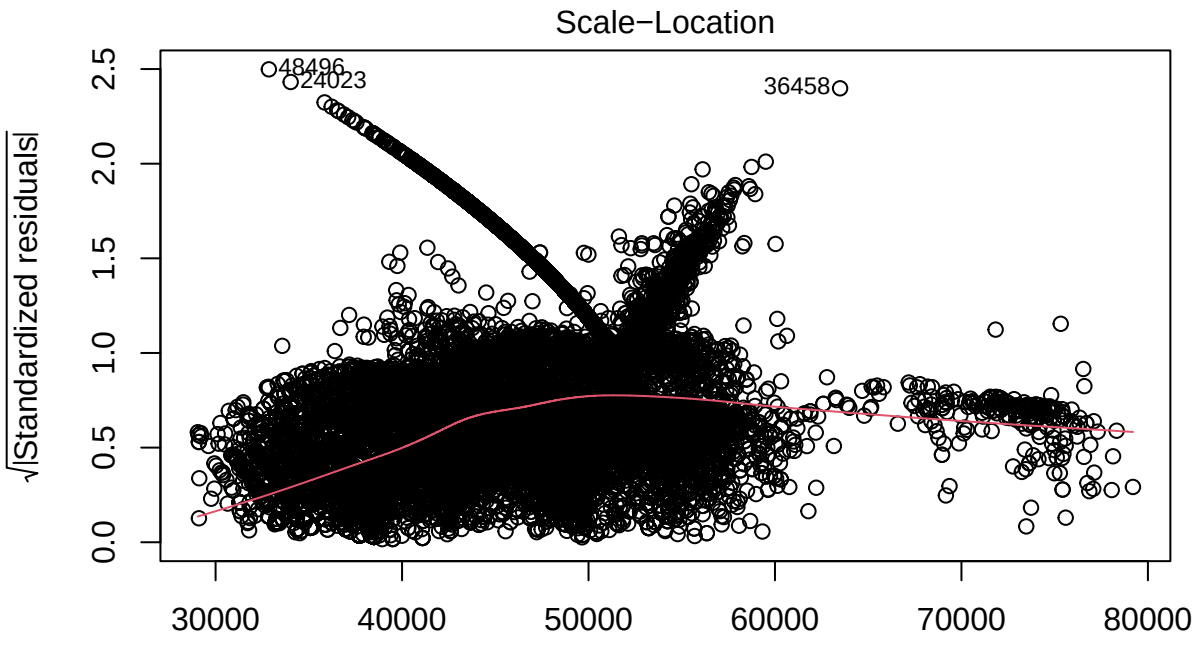
```
anova(initial_model, model_poly)
```

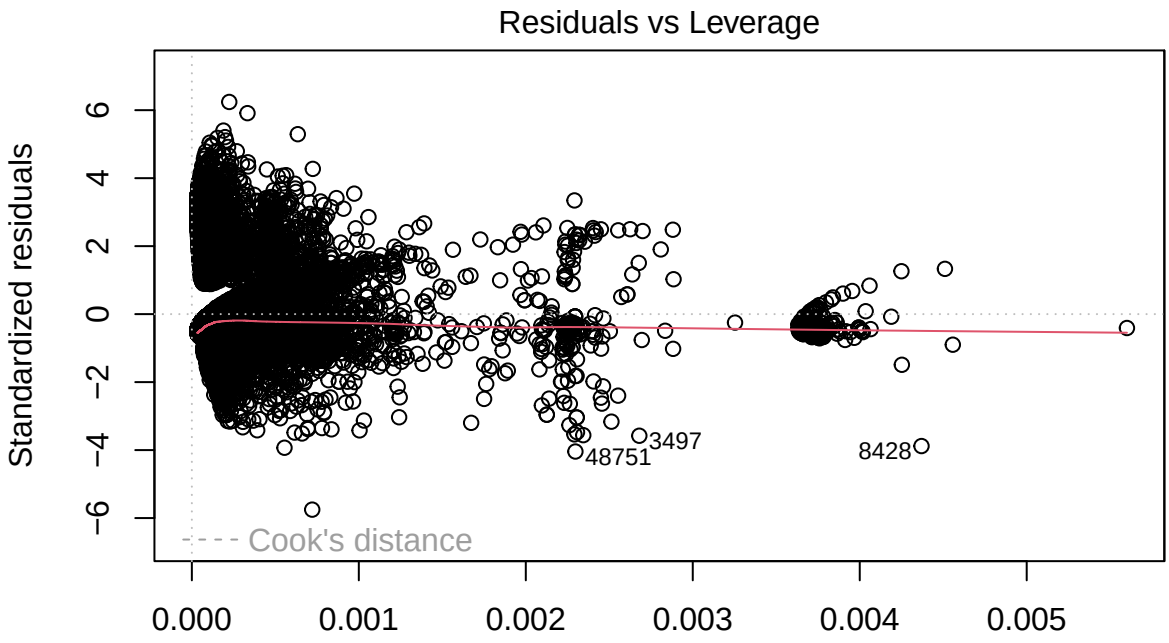
```
## Analysis of Variance Table
##
## Model 1: income_integer ~ age + edu_num + cap_gain + cap_loss + hours_week
## Model 2: income_integer ~ age + age2 + edu_num + cap_gain + cap_loss +
##     hours_week + hours_week2
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  48836 6.5856e+11
##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

```r
#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
```

```r
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
```

```r
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
```

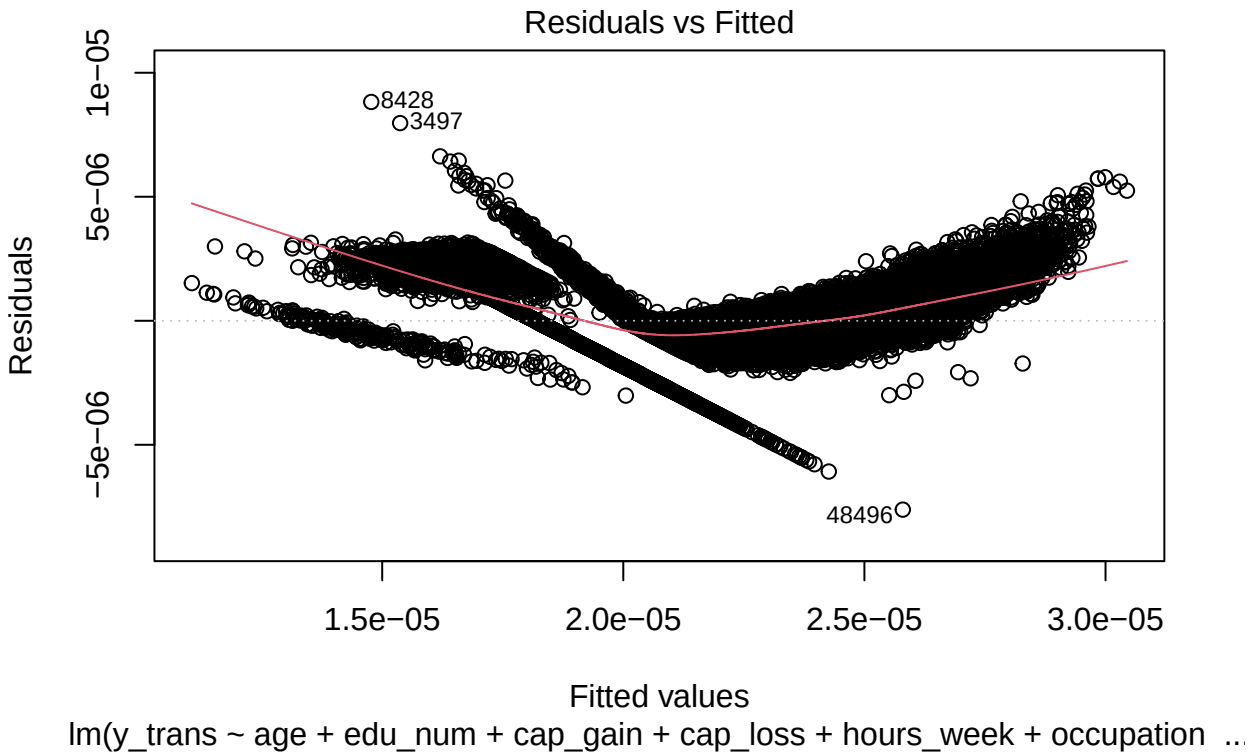
```
anova(catmodel)
```

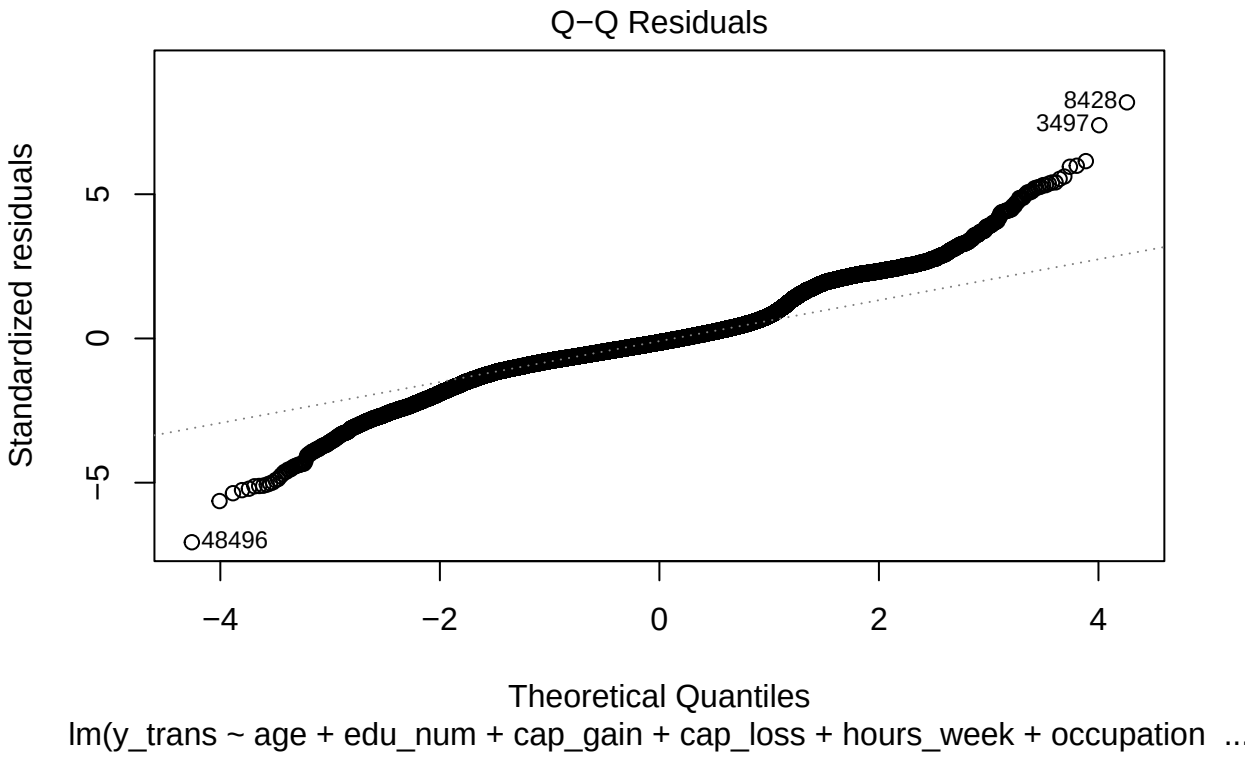
```
## Analysis of Variance Table
##
## Response: y_trans
##                   Df     Sum Sq    Mean Sq    F value    Pr(>F)
## age                1 1.5949e-07 1.5949e-07 137000.941 < 2.2e-16 ***
## edu_num            1 1.4352e-07 1.4352e-07 123282.092 < 2.2e-16 ***
## cap_gain           1 9.4400e-09 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
```

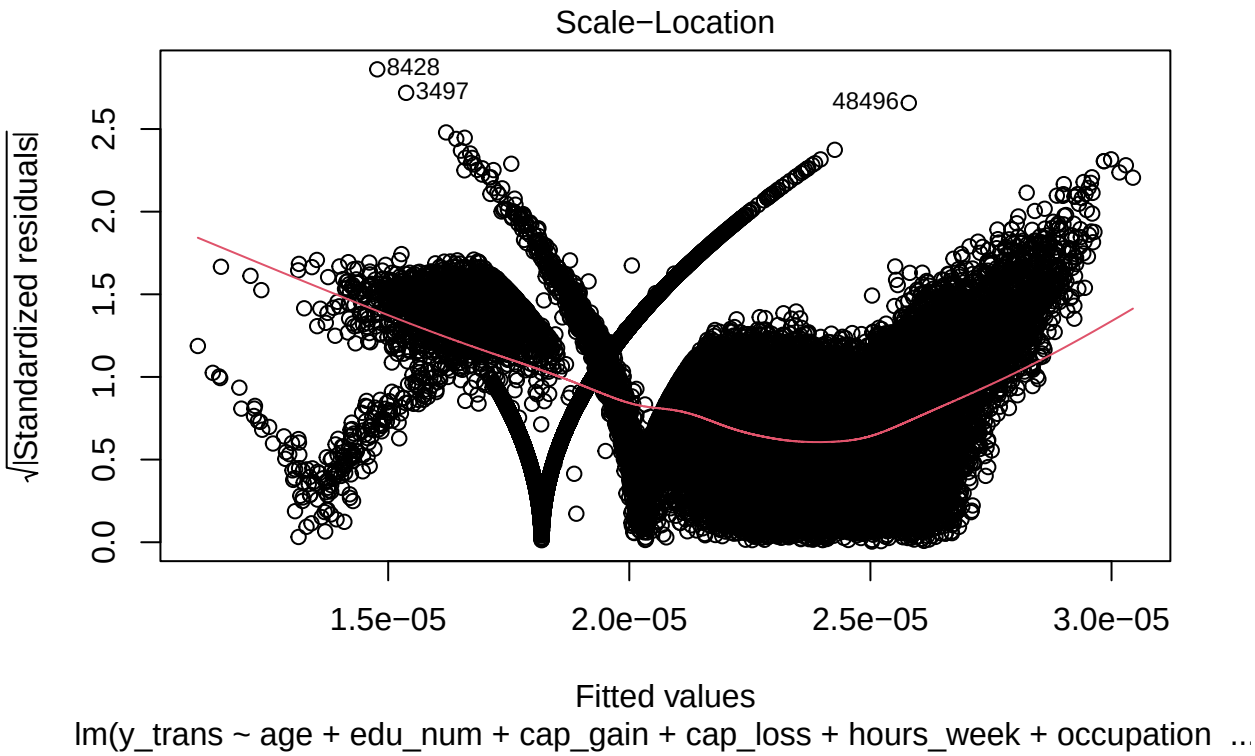
```
anova(transformed_model, catmodel)
```

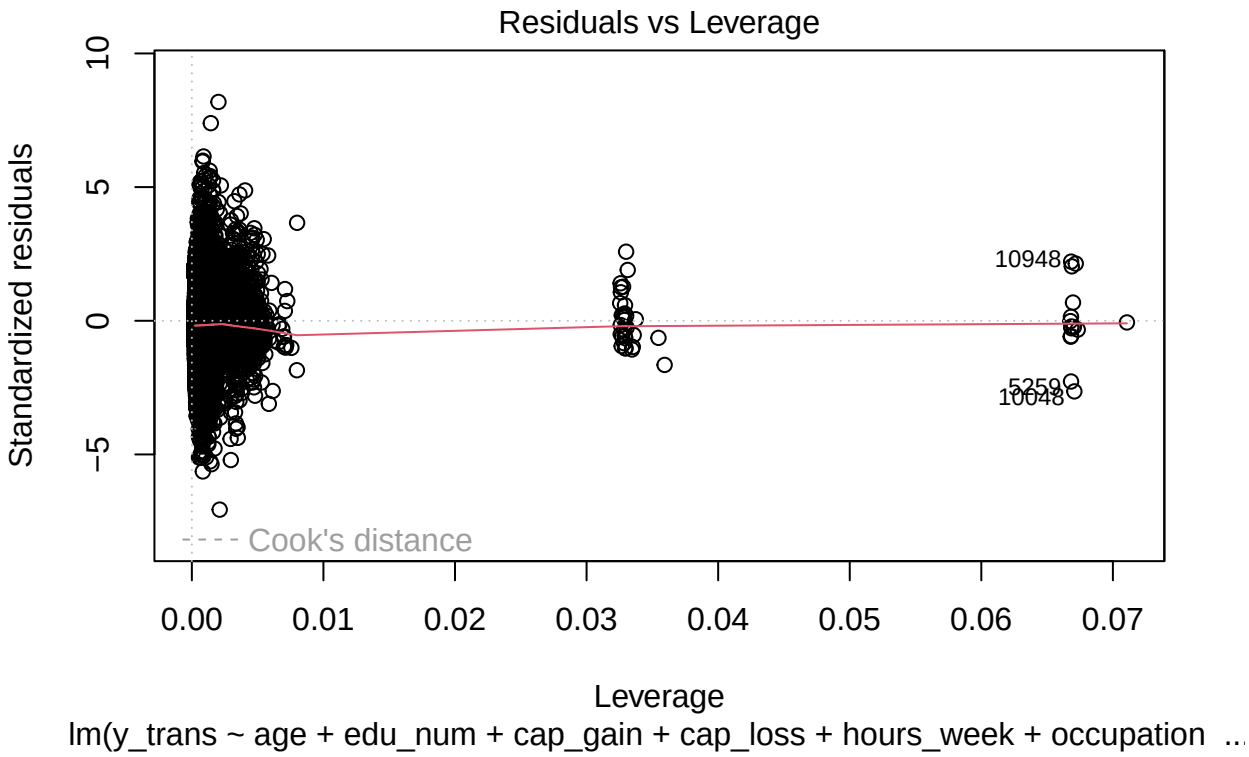
```
## Analysis of Variance Table
##
## Model 1: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week
## Model 2: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     occupation + marital + sex + workclass + relationship + race +
##     native_country + income
##   Res.Df        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)
```

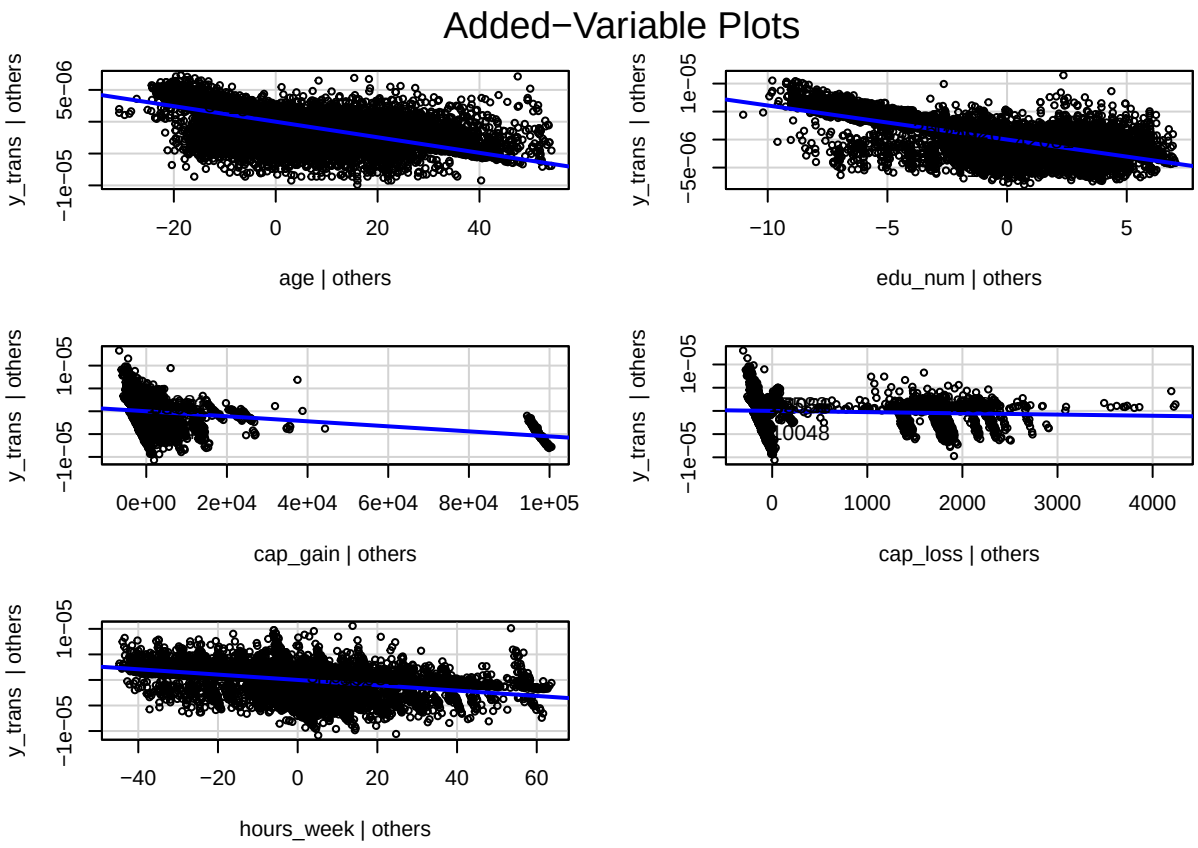

Residuals vs Fitted
Residuals
1e−05
5e−06
−5e−06
8428
3497
48496
1.5e−05
2.0e−05
2.5e−05
3.0e−05
Fitted values
lm(y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week + occupation ...


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 + occupation ...

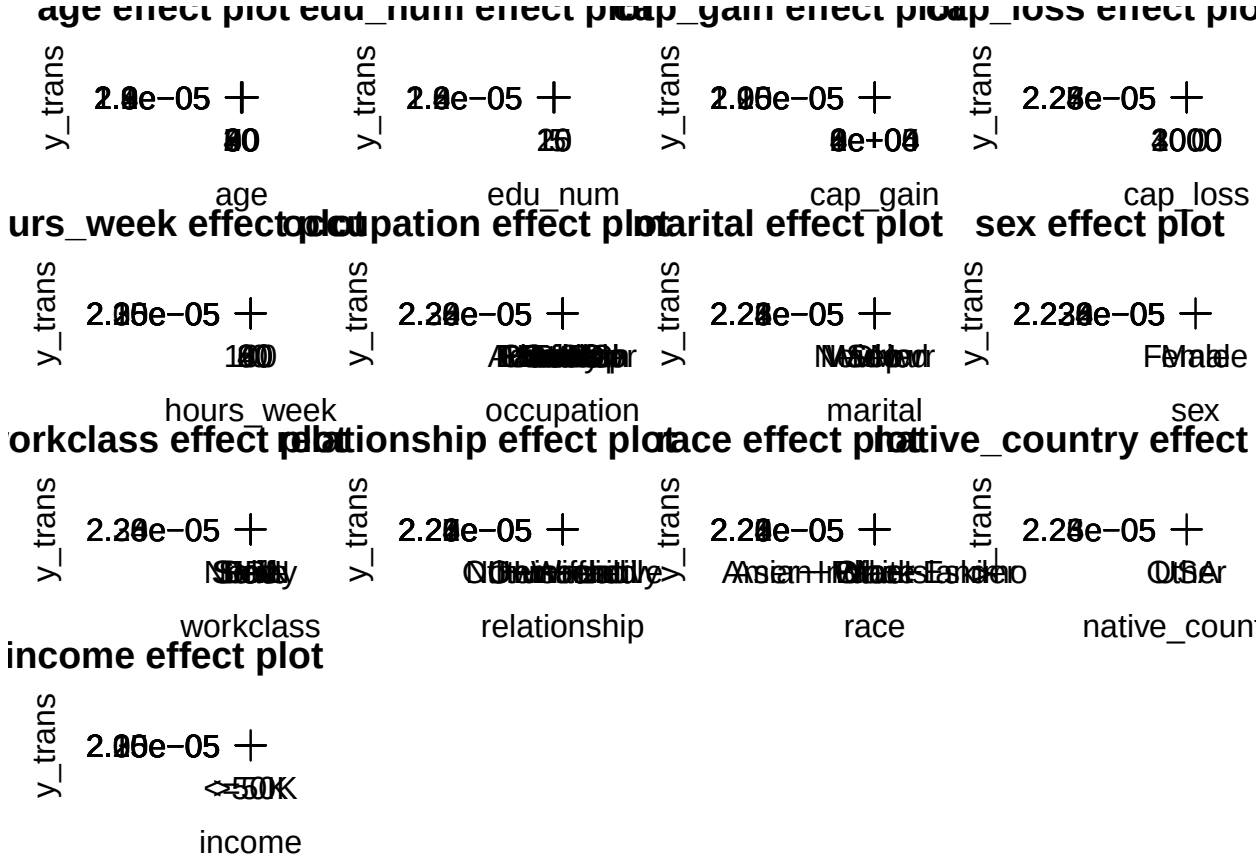

Scale−Location
8428
3497
48496
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
1.5e−05
2.0e−05
2.5e−05
3.0e−05
Fitted values
lm(y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week + occupation ...




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




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




```
#using this line will transform our income variable into a binary response variable we can use for our
dd$income_bin <- ifelse(dd$income == ">50K", 1, 0)
#transforming variable into factor
dd$income_bin <- as.factor(dd$income_bin)
```

Build the Initial Logistic Regression Model

```
## Build the Initial Logistic Regression Model
initial_model_b <- glm(income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week, data = dd, family
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

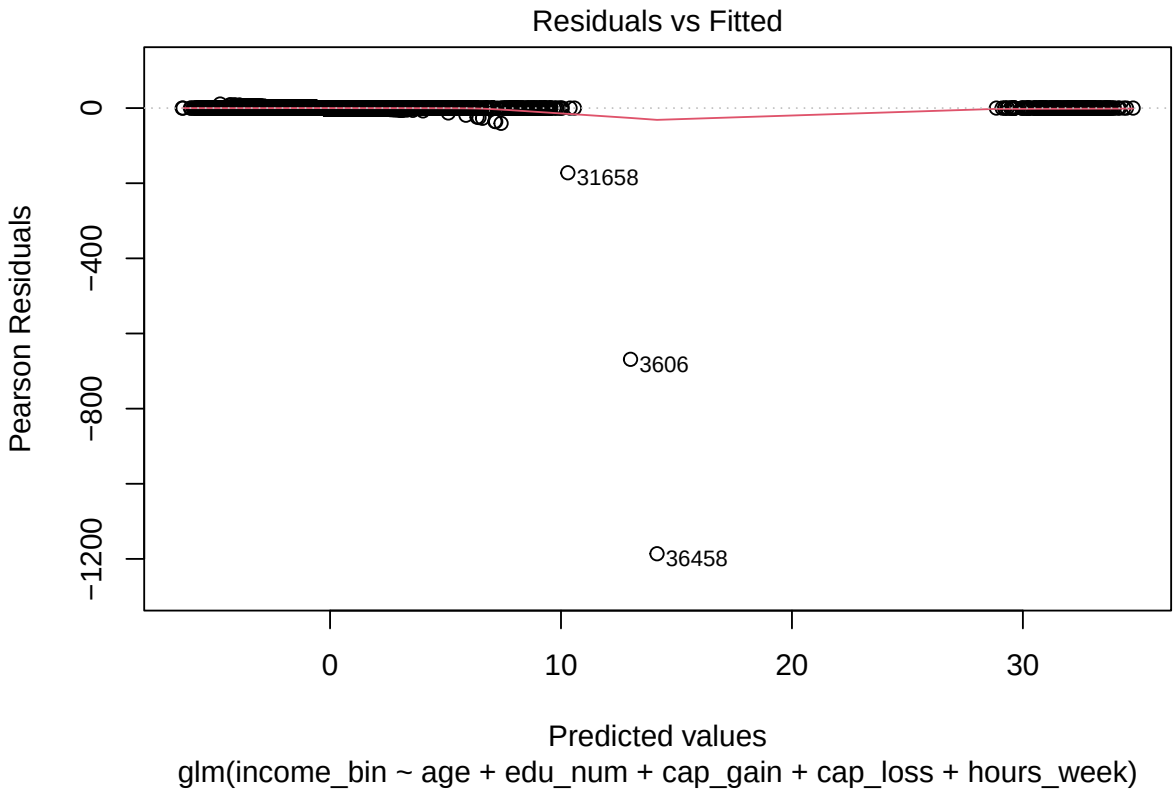
```
summary(initial_model_b)
```

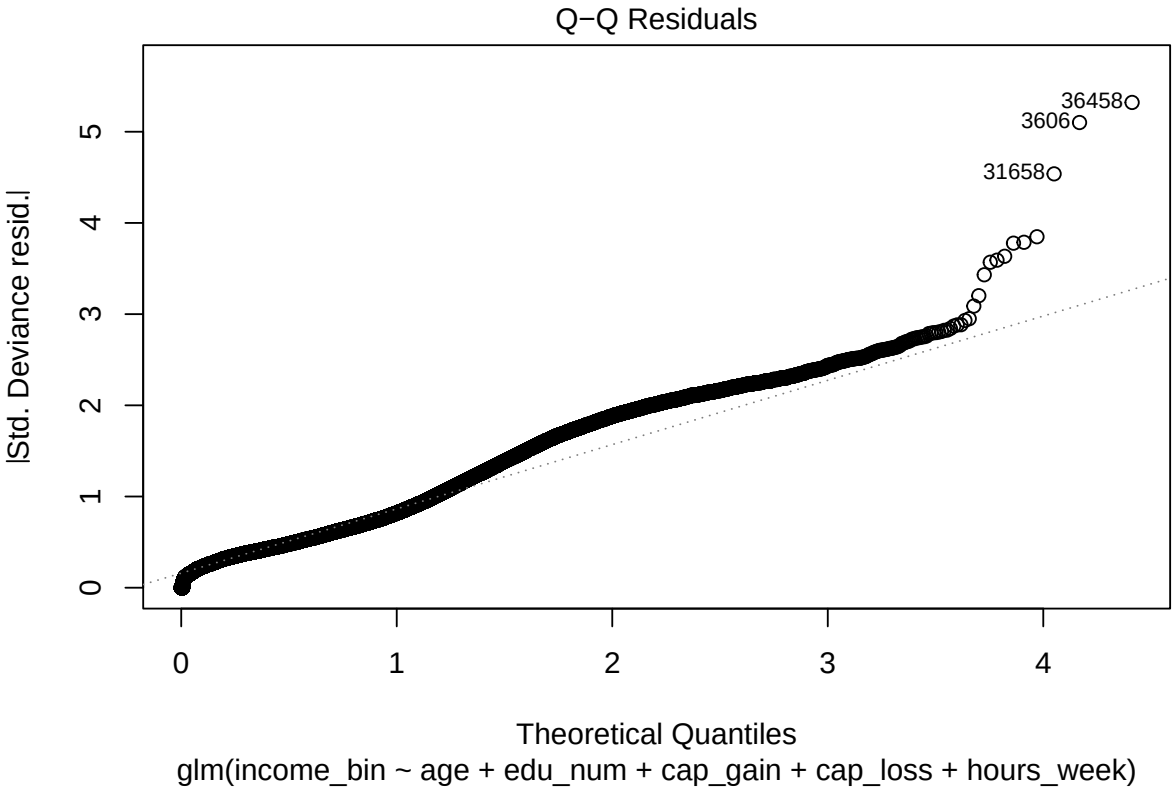
```
##
## Call:
## glm(formula = income_bin ~ age + edu_num + cap_gain + cap_loss +
##     hours_week, family = binomial, data = dd)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.260e+00  9.371e-02  -88.14   <2e-16 ***
```

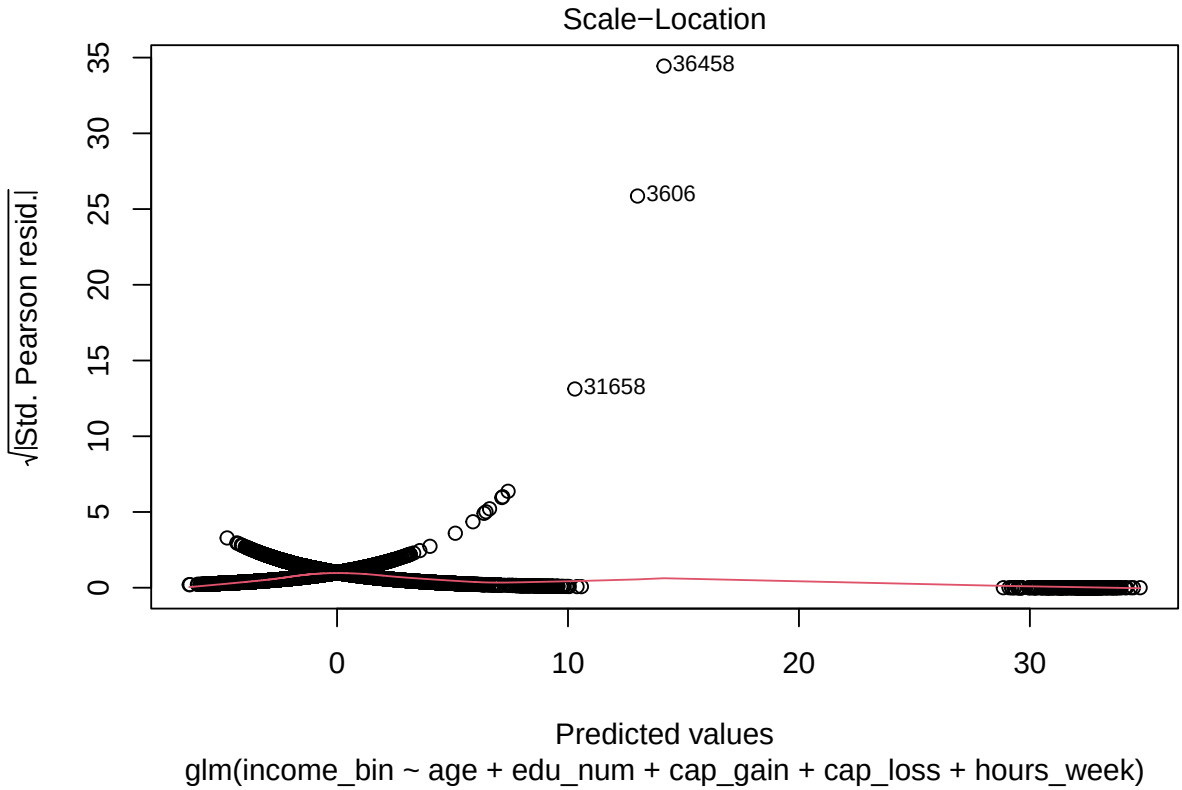
```
## age          4.220e-02  9.915e-04   42.57   <2e-16 ***
## edu_num      3.223e-01  5.556e-03   58.01   <2e-16 ***
## cap_gain     3.205e-04  7.985e-06   40.14   <2e-16 ***
## cap_loss     6.799e-04  2.634e-05   25.81   <2e-16 ***
## hours_week   4.012e-02  1.070e-03   37.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 53751  on 48841  degrees of freedom
## Residual deviance: 39775  on 48836  degrees of freedom
## AIC: 39787
##
## Number of Fisher Scoring iterations: 7
```

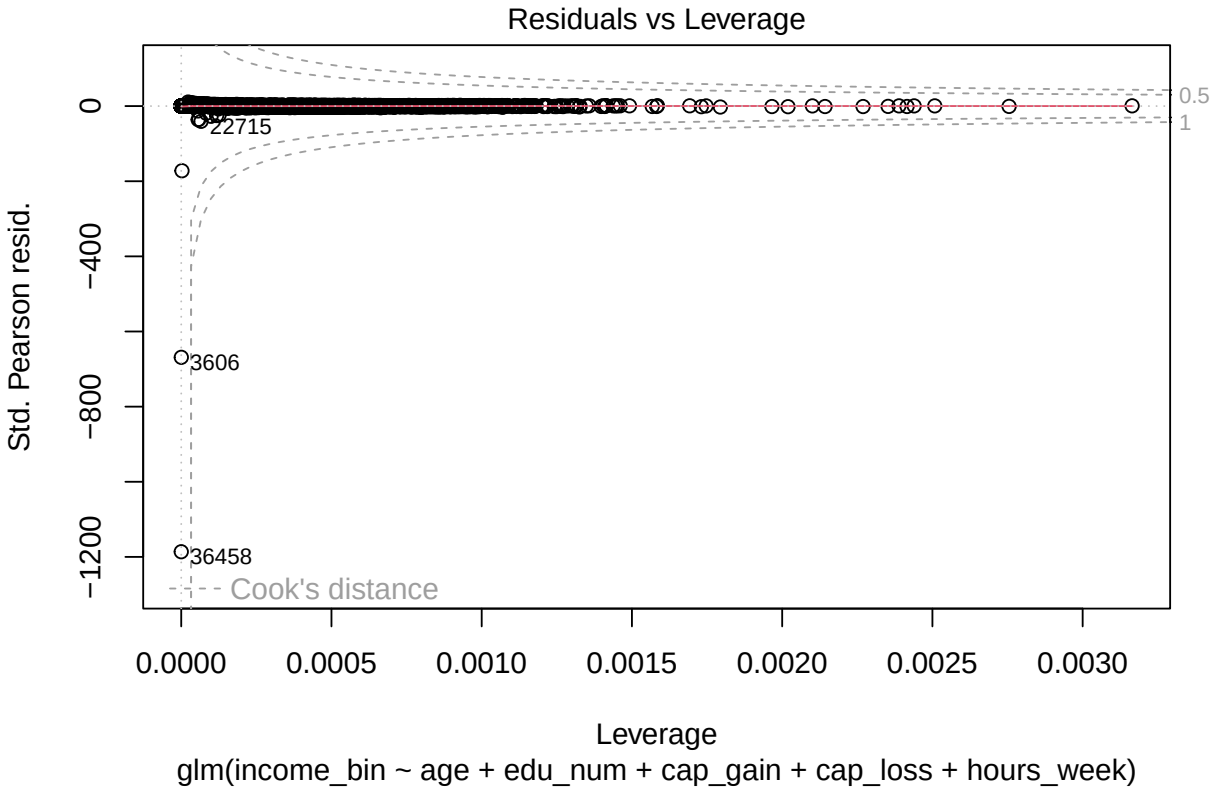
Check Model Diagnostics

```
## Check Model Diagnostics
plot(initial_model_b) #the basic hypothesis are met, beware of Homoscedasticity
```

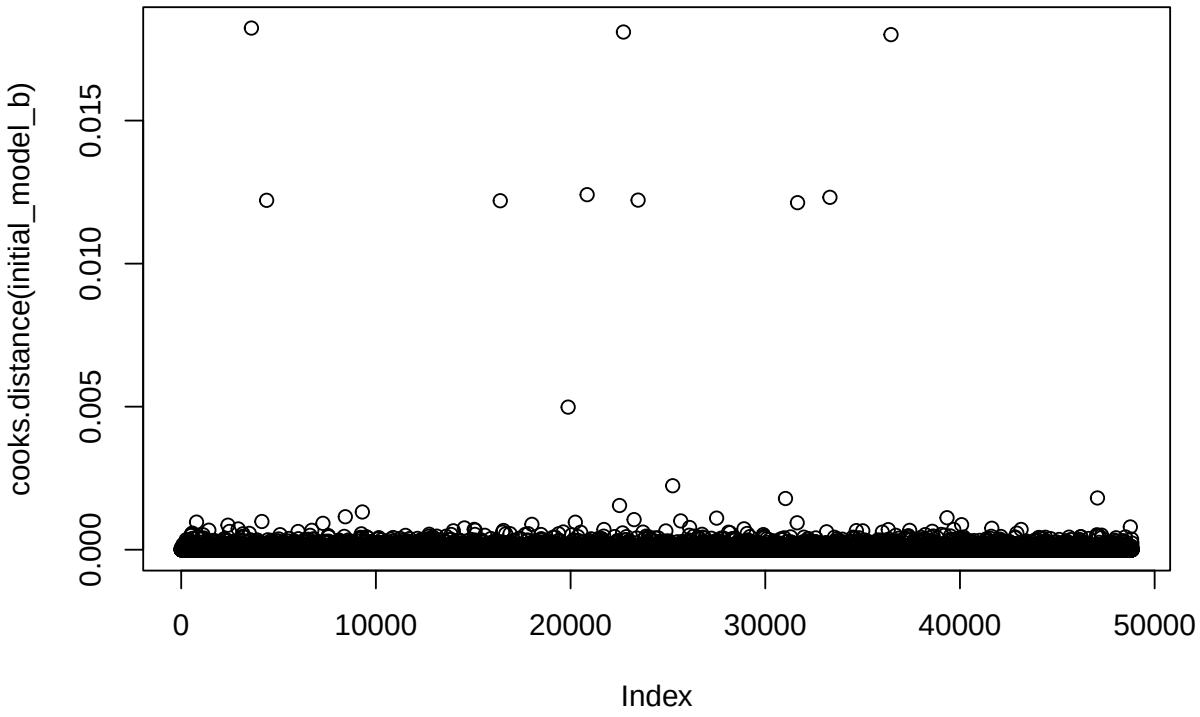



Q−Q Residuals
|Std. Deviance resid.|
0
1
2
3
4
5
36458
3606
31658
0
1
2
3
4
Theoretical Quantiles
glm(income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week)


Scale−Location
√|Std. Pearson resid.|
0
5
10
15
20
25
30
35
36458
3606
31658
0
10
20
30
Predicted values
glm(income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week)




```
#Check cook's distance
plot(cooks.distance(initial_model_b)) #there are some influential observations that skew the data a lit
```



Add Categorical Variables Step by Step

```
##Add_workclass_and_test_with_Chi_squared
model_workclass <- update(initial_model_b, . ~ . + workclass)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
anova(initial_model_b, model_workclass, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week
## Model 2: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     48836      39775
## 2     48830      39546  6   229.93 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
##Add_marital_and_test_with_Chi_squared
model_marital <- update(model_workclass, . ~ . + marital)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
anova(model_workclass, model_marital, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass
## Model 2: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     48830      39546
## 2     48826      32504  4   7041.9 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
##Add_occupation_and_test_with_Chi_squared
model_occupation <- update(model_marital, . ~ . + occupation)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
anova(model_marital, model_occupation, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital
## Model 2: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     48826      32504
## 2     48813      31676 13   827.94 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
##Add_relationship_and_test_with_Chi_squared
model_relationship <- update(model_occupation, . ~ . + relationship)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
anova(model_occupation, model_relationship, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation
```

```
## Model 2: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation + relationship
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     48813      31676
## 2     48808      31438  5    237.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##Add_race_and_test_with_Chi_squared
model_race <- update(model_relationship, . ~ . + race)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
anova(model_relationship, model_race, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation + relationship
## Model 2: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation + relationship + race
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     48808      31438
## 2     48804      31409  4   29.342 6.661e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##Add_sex_and_test_with_Chi_squared
model_sex <- update(model_race, . ~ . + sex)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
anova(model_race, model_sex, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation + relationship + race
## Model 2: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation + relationship + race +
##     sex
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     48804      31409
## 2     48803      31286  1   122.48 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##Add_native_country_and_test_with_Chi_squared
model_country <- update(model_sex, . ~ . + native_country)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
anova(model_sex, model_country, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation + relationship + race +
##     sex
## Model 2: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation + relationship + race +
##     sex + native_country
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     48803      31286
## 2     48802      31275  1       11 0.000911 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Define the Final Model

```r
## Define the Final Model
final_model <- model_country
```

Perform Stepwise Selection and Final Diagnostics

```r
## Perform Stepwise Selection and Final Diagnostics
stepmodel <- stepAIC(final_model, direction = "back")
```

```
## Start:  AIC=31355.09
## income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation + relationship + race +
##     sex + native_country
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##                   Df Deviance   AIC
## <none>                  31275 31355
## - native_country  1     31286 31364
## - race            4     31301 31373
## - sex             1     31399 31477
```

```
## - workclass       6    31447 31515
## - marital         4    31447 31519
## - age             1    31584 31662
## - relationship    5    31628 31698
## - cap_loss        1    31749 31827
## - hours_week      1    31875 31953
## - occupation     13    32064 32118
## - edu_num         1    33242 33320
## - cap_gain        1    34086 34164
```

```
vif(final_model)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## age              1.223938  1        1.106317
## edu_num          1.393982  1        1.180670
## cap_gain         1.024071  1        1.011964
## cap_loss         1.010195  1        1.005085
## hours_week       1.144362  1        1.069748
## workclass        1.492505  6        1.033934
## marital         47.015300  4        1.618191
## occupation       2.171232 13        1.030268
## relationship   109.766183  5        1.599731
## race             1.308231  4        1.034155
## sex              2.942333  1        1.715323
## native_country   1.257900  1        1.121561
```

```
summary(final_model)
```

```
##
## Call:
## glm(formula = income_bin ~ age + edu_num + cap_gain + cap_loss +
##     hours_week + workclass + marital + occupation + relationship +
##     race + sex + native_country, family = binomial, data = dd)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -9.271e+00  3.257e-01 -28.470  < 2e-16 ***
## age                         2.269e-02  1.295e-03  17.524  < 2e-16 ***
## edu_num                     3.011e-01  7.273e-03  41.395  < 2e-16 ***
## cap_gain                    3.160e-04  8.376e-06  37.733  < 2e-16 ***
## cap_loss                    6.460e-04  3.002e-05  21.519  < 2e-16 ***
## hours_week                  3.097e-02  1.284e-03  24.123  < 2e-16 ***
## workclassLoc               -5.873e-01  9.078e-02  -6.470 9.82e-11 ***
## workclassNoPay             -1.502e+00  7.861e-01  -1.910 0.056107 .
## workclassPriv              -5.640e-01  7.574e-02  -7.447 9.54e-14 ***
## workclassSelfI             -3.603e-01  9.942e-02  -3.624 0.000290 ***
## workclassSelfN             -1.023e+00  8.867e-02 -11.533  < 2e-16 ***
## workclassState             -7.653e-01  1.000e-01  -7.651 1.99e-14 ***
## maritalMarried              2.315e+00  2.141e-01  10.815  < 2e-16 ***
## maritalNevMarr             -4.267e-01  7.114e-02  -5.999 1.99e-09 ***
## maritalSep                  4.041e-03  1.110e-01   0.036 0.970972
## maritalWidow                3.927e-02  1.247e-01   0.315 0.752873
## occupationArmy              5.115e-01  7.782e-01   0.657 0.511025
```

```
## occupationCraftRep          8.389e-02  6.457e-02   1.299 0.193863
## occupationExecMan           7.371e-01  6.208e-02  11.874  < 2e-16 ***
## occupationFarmFish         -1.017e+00  1.151e-01  -8.837  < 2e-16 ***
## occupationHandlCl          -6.502e-01  1.136e-01  -5.723 1.05e-08 ***
## occupationHouse            -1.971e+00  7.554e-01  -2.609 0.009069 **
## occupationMachOp           -2.642e-01  8.235e-02  -3.209 0.001334 **
## occupationOther            -8.687e-01  9.597e-02  -9.052  < 2e-16 ***
## occupationProf              2.699e-01  6.088e-02   4.434 9.24e-06 ***
## occupationProtServ          4.240e-01  1.024e-01   4.139 3.48e-05 ***
## occupationSales             2.462e-01  6.667e-02   3.693 0.000222 ***
## occupationTech              5.327e-01  8.921e-02   5.971 2.36e-09 ***
## occupationTrans            -7.686e-02  8.015e-02  -0.959 0.337532
## relationshipNot-in-family   5.761e-01  2.119e-01   2.719 0.006541 **
## relationshipOther-relative -5.304e-01  1.971e-01  -2.690 0.007136 **
## relationshipOwn-child      -5.928e-01  2.082e-01  -2.848 0.004406 **
## relationshipUnmarried       3.814e-01  2.256e-01   1.691 0.090884 .
## relationshipWife            1.095e+00  8.212e-02  13.335  < 2e-16 ***
## raceAsian-Pac-Islander      5.924e-01  1.983e-01   2.987 0.002813 **
## raceBlack                   3.481e-01  1.857e-01   1.875 0.060856 .
## raceOther                   3.738e-01  2.651e-01   1.410 0.158508
## raceWhite                   5.736e-01  1.770e-01   3.240 0.001194 **
## sexMale                     6.955e-01  6.335e-02  10.980  < 2e-16 ***
## native_countryUSA           1.872e-01  5.682e-02   3.295 0.000984 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 53751  on 48841  degrees of freedom
## Residual deviance: 31275  on 48802  degrees of freedom
## AIC: 31355
##
## Number of Fisher Scoring iterations: 7
```

```
anova(final_model, test="LR")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: income_bin
##
## Terms added sequentially (first to last)
```

```
##
##
##                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                            48841      53751
## age             1   2537.0     48840      51214 < 2.2e-16 ***
## edu_num         1   5891.7     48839      45322 < 2.2e-16 ***
## cap_gain        1   3291.4     48838      42031 < 2.2e-16 ***
## cap_loss        1    753.5     48837      41277 < 2.2e-16 ***
## hours_week      1   1501.8     48836      39775 < 2.2e-16 ***
## workclass       6    229.9     48830      39546 < 2.2e-16 ***
## marital         4   7041.9     48826      32504 < 2.2e-16 ***
## occupation     13    827.9     48813      31676 < 2.2e-16 ***
## relationship    5    237.8     48808      31438 < 2.2e-16 ***
## race            4     29.3     48804      31409 6.661e-06 ***
## sex             1    122.5     48803      31286 < 2.2e-16 ***
## native_country  1     11.0     48802      31275  0.000911 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

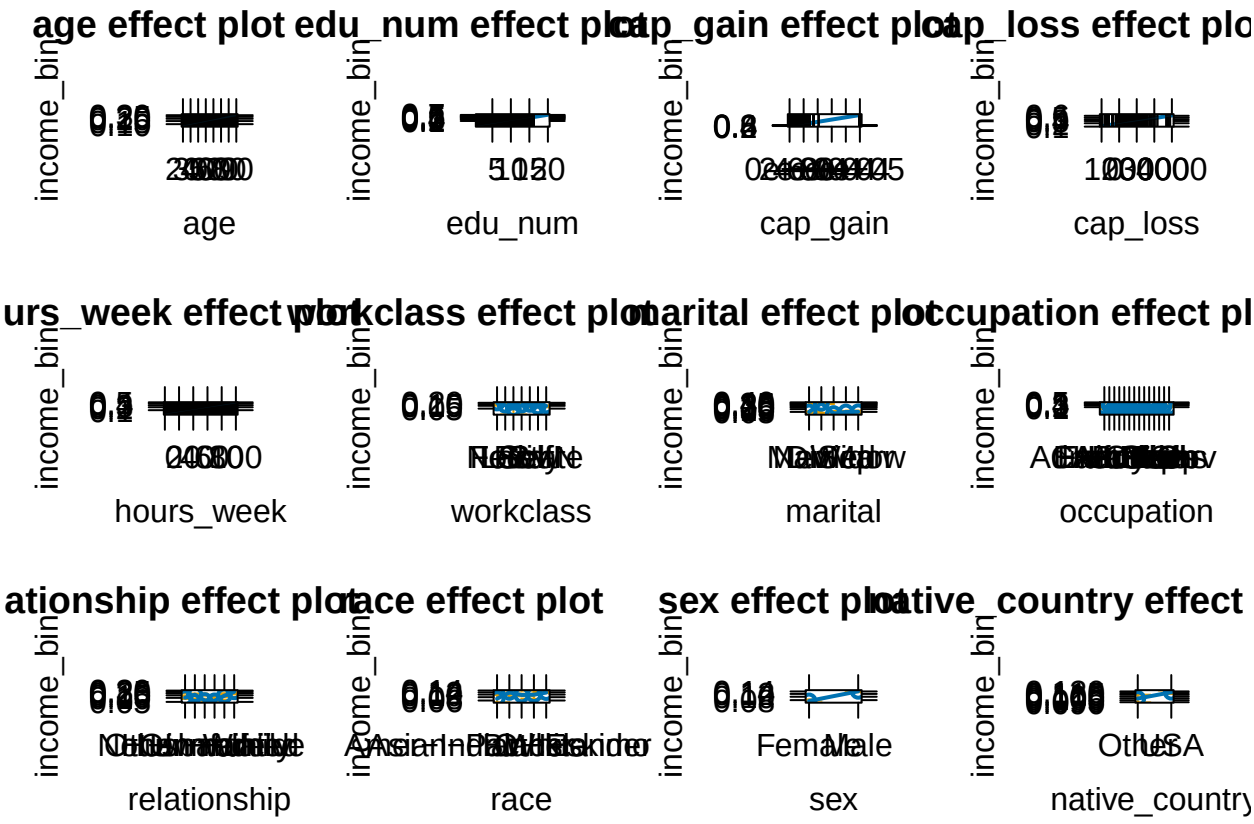
```
anova(initial_model_b, final_model)
```

```
## Analysis of Deviance Table
##
## Model 1: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week
## Model 2: income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week +
##     workclass + marital + occupation + relationship + race +
##     sex + native_country
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     48836      39775
## 2     48802      31275 34   8500.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(initial_model_b, final_model)
```

```
##                 df      AIC
## initial_model_b  6 39787.46
## final_model     40 31355.09
```

```
plot(allEffects(final_model))
```


age effect plot
edu_num effect plot
cap_gain effect plot
cap_loss effect plo
income_bin
age
edu_num
cap_gain
cap_loss
urs_week effect plot
workclass effect plot
marital effect plot
occupation effect pl
income_bin
hours_week
workclass
marital
occupation
ationship effect plot
race effect plot
sex effect plot
native_country effect
income_bin
Female
Male
Other
USA
relationship
race
sex
native_country