

Deliverable 2

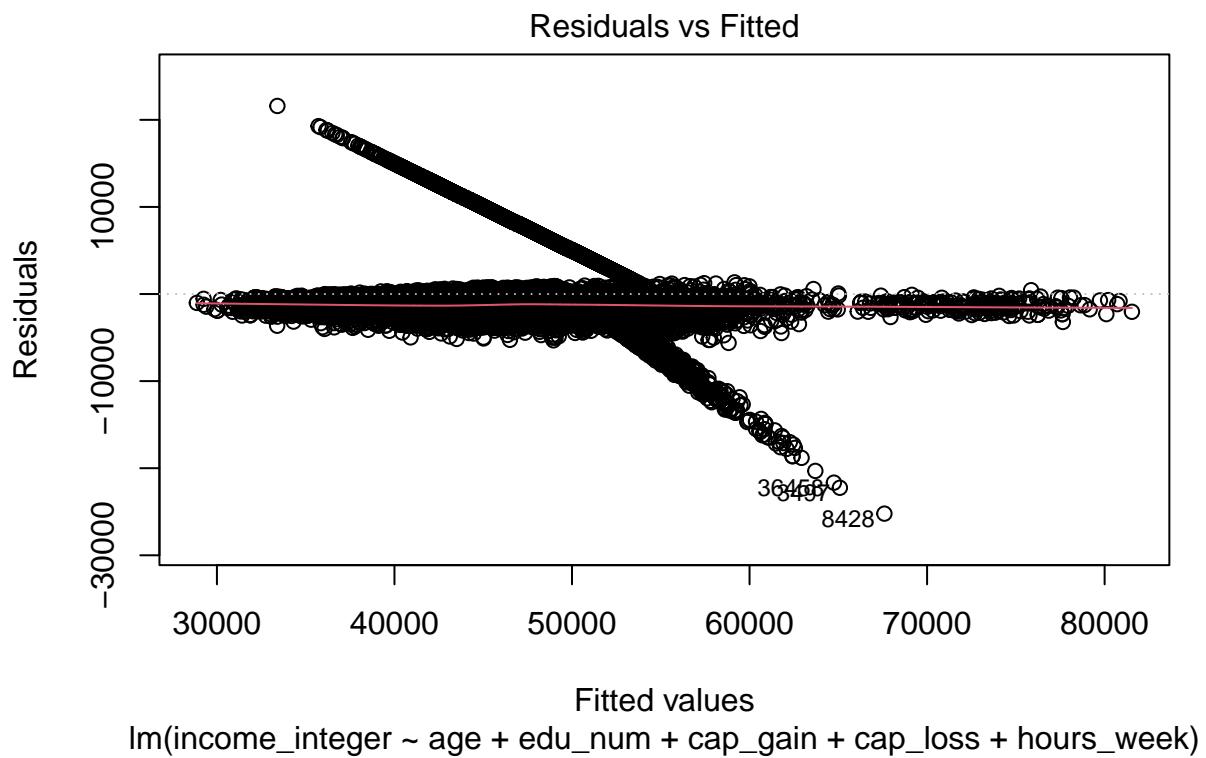
2025-05-01

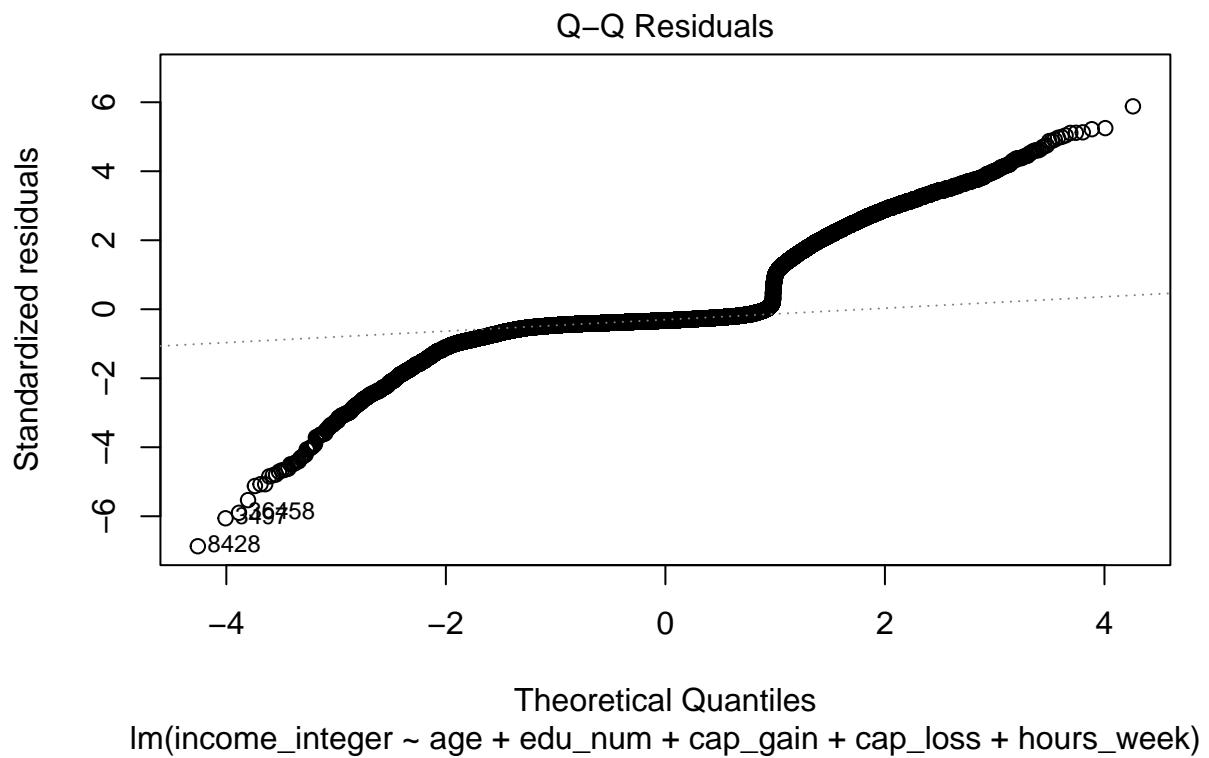
```
setwd("~/Escritorio/ADEI/D2")
dd <- read.csv("adult_def.csv")

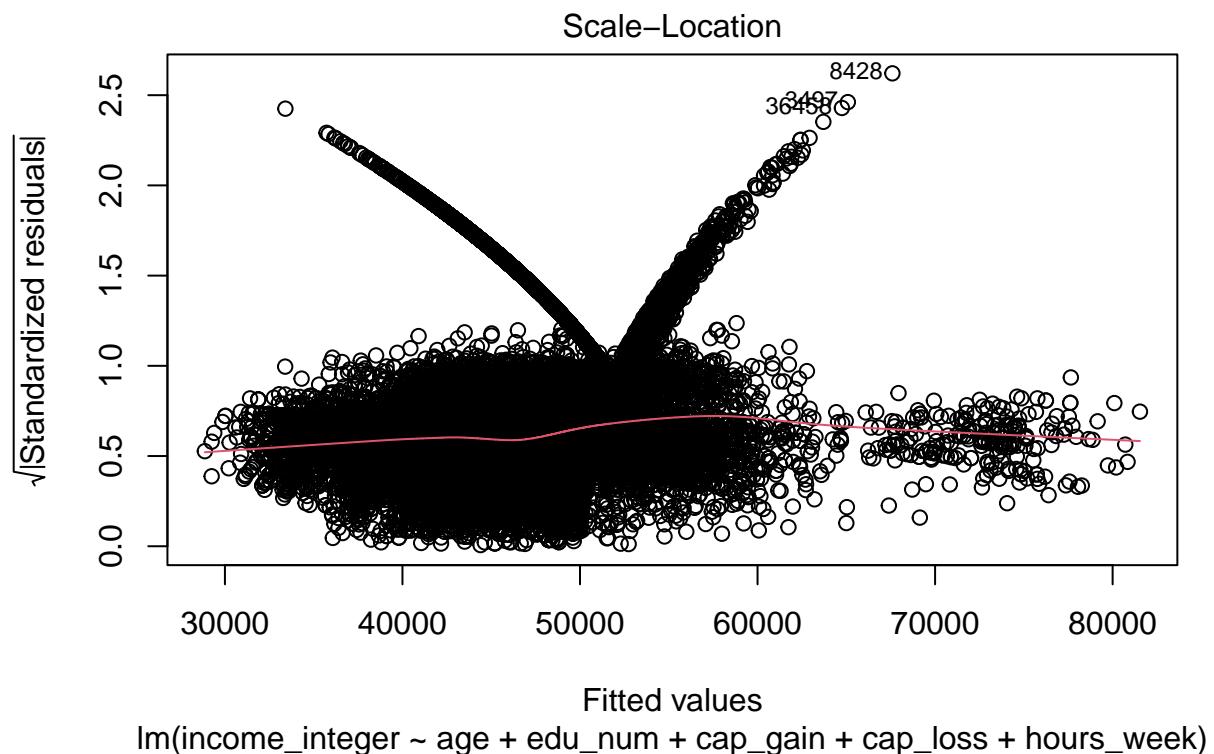
initial_model <- lm(income_integer ~ age + edu_num + cap_gain + cap_loss + hours_week, data = dd)
summary(initial_model)

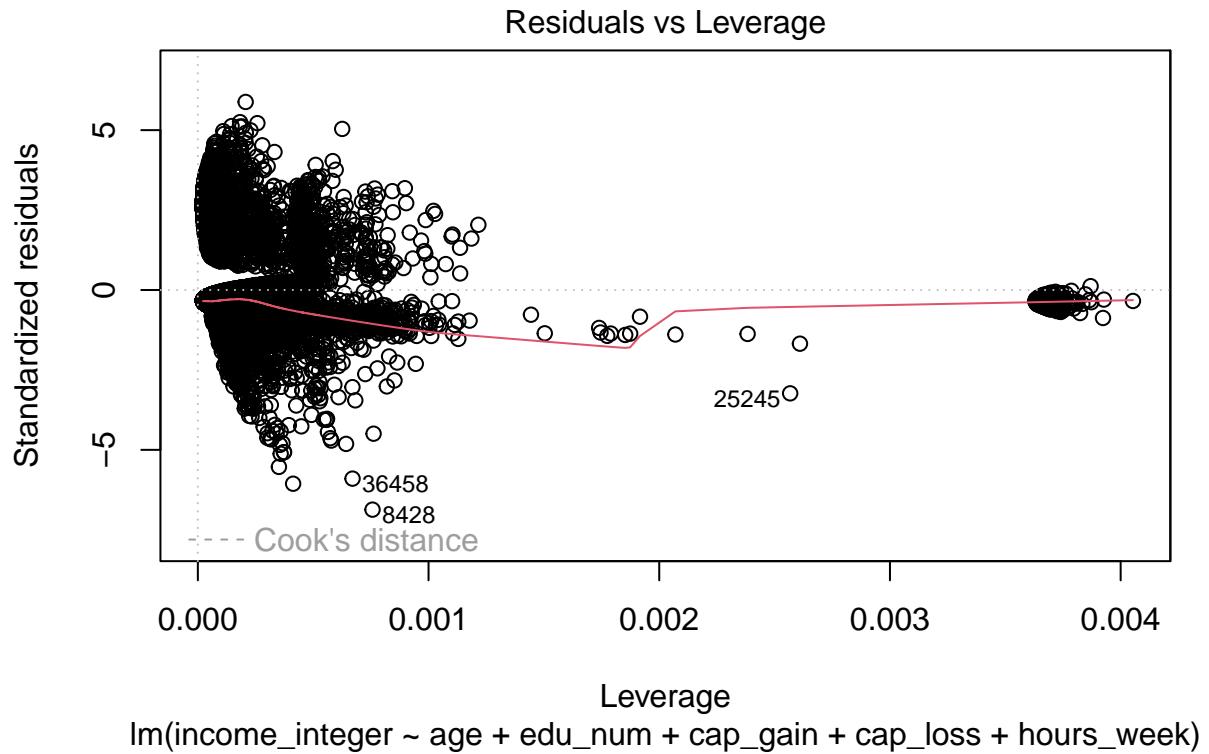
##
## Call:
## lm(formula = income_integer ~ age + edu_num + cap_gain + cap_loss +
##     hours_week, data = dd)
##
## Residuals:
##       Min      1Q      Median      3Q      Max
## -25216.0  -1522.3   -1201.8   -699.4  21594.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)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

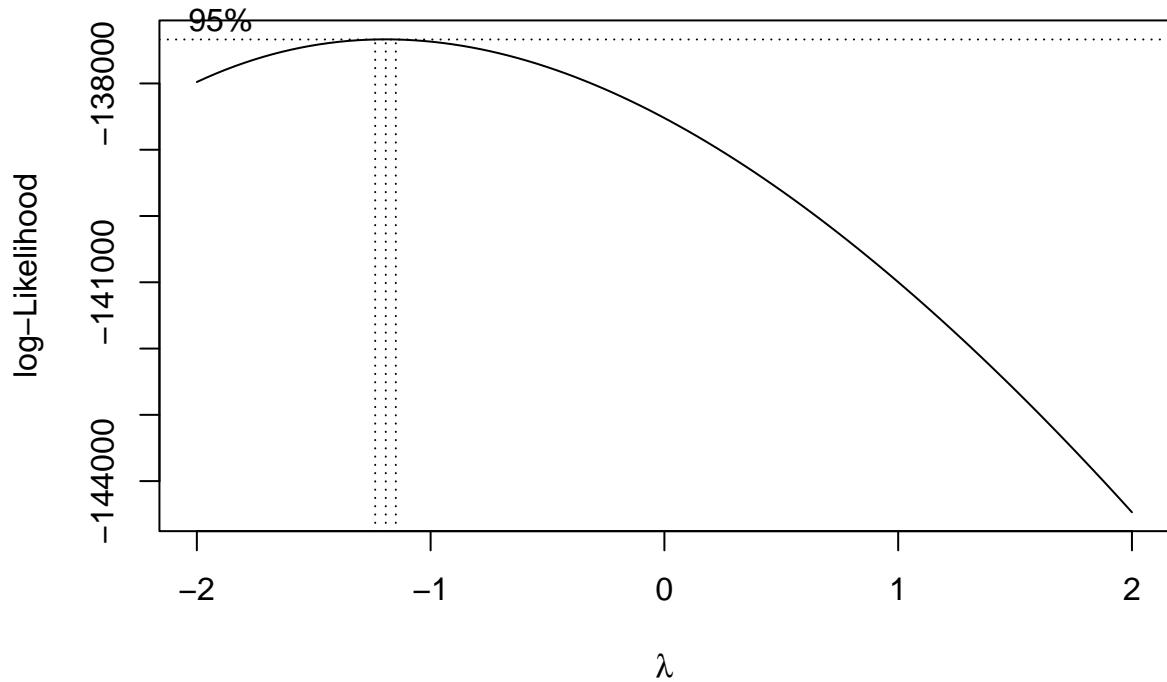








```
#target variable transformation, so the normality assumption is met
boxcox(income_integer ~ age + edu_num + cap_gain + cap_loss + hours_week, data = dd)
```

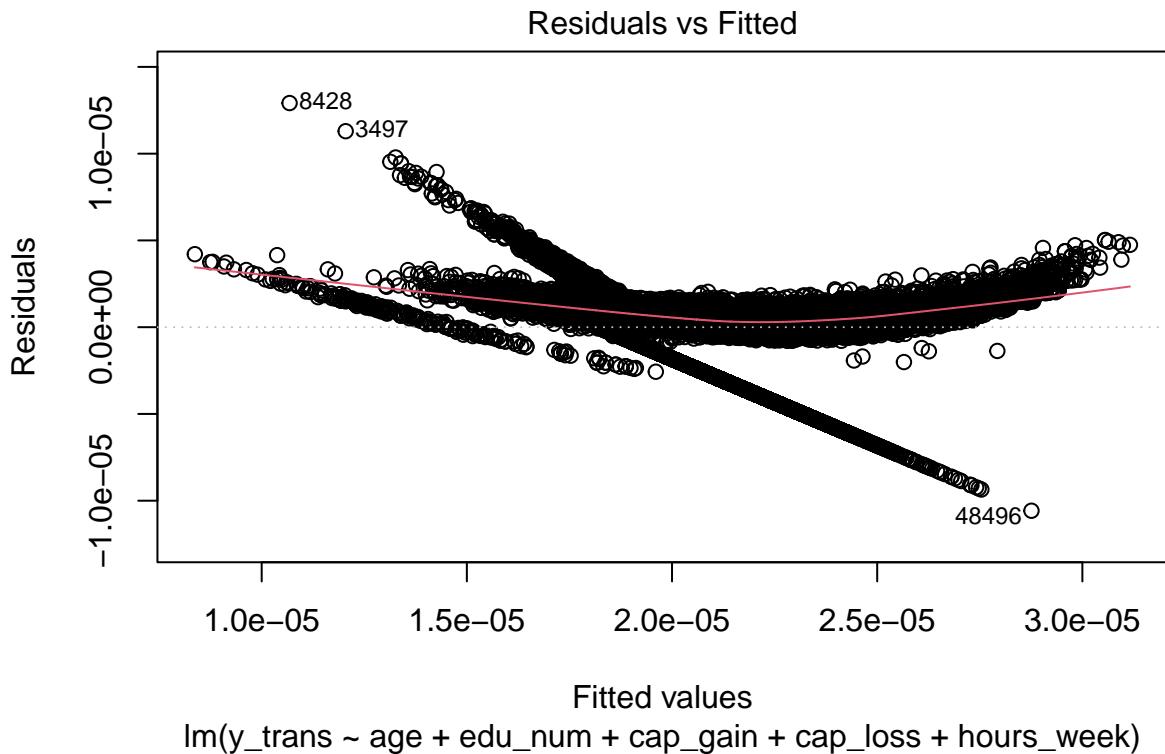


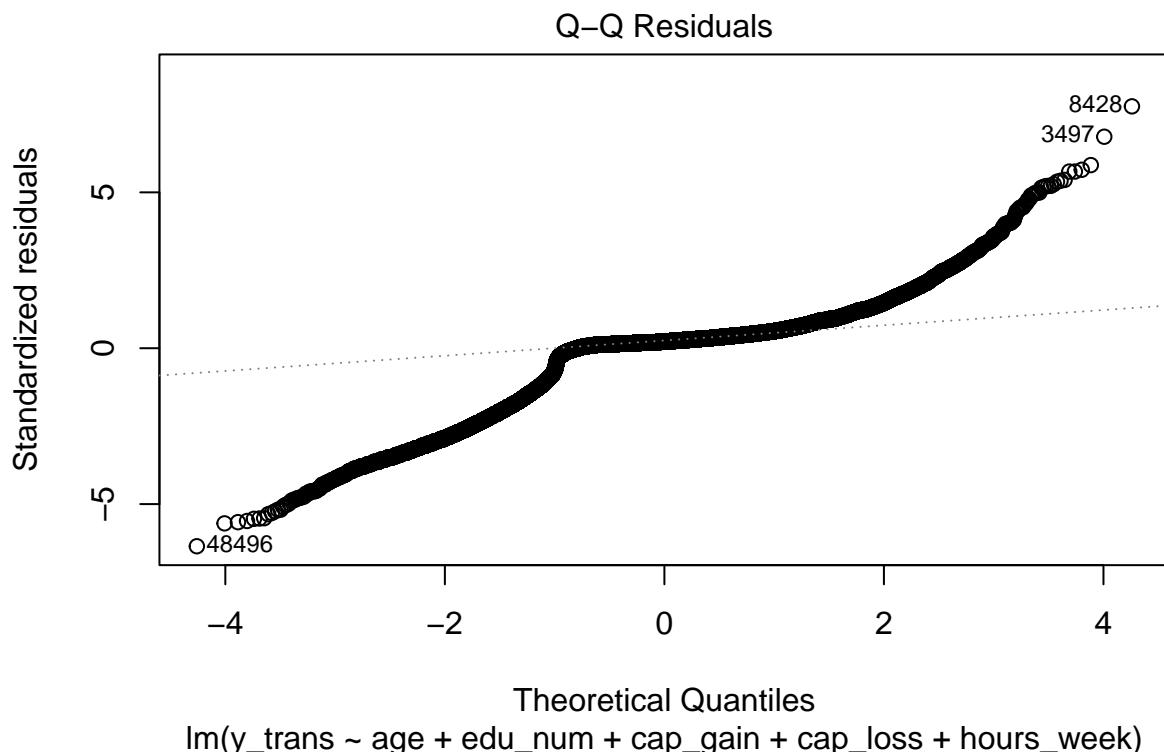
```
#given lambda is approximatedly -1 we do the inverse transformation
y_trans <- 1 / dd$income_integer
transformed_model <- lm(y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week, data = dd)
summary(transformed_model)

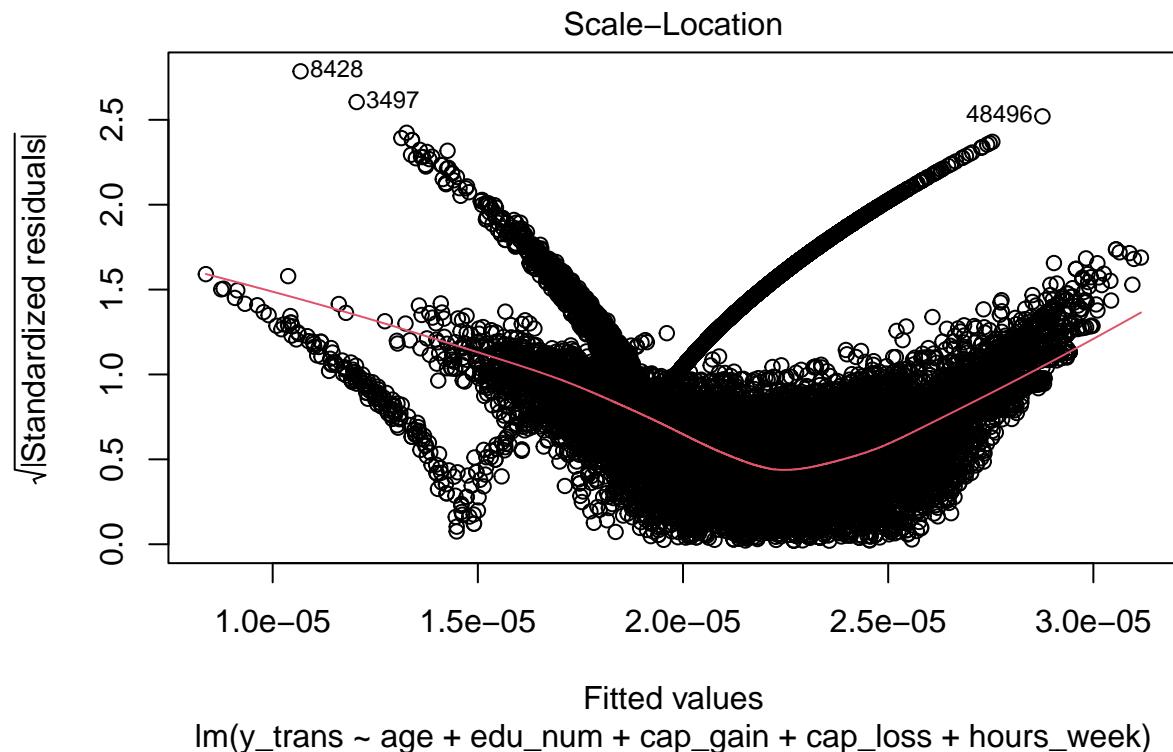
##
## Call:
## lm(formula = y_trans ~ age + edu_num + cap_gain + cap_loss +
##     hours_week, data = dd)
##
## Residuals:
##      Min        1Q    Median        3Q       Max
## -1.058e-05  1.373e-07  3.552e-07  6.859e-07  1.292e-05
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)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

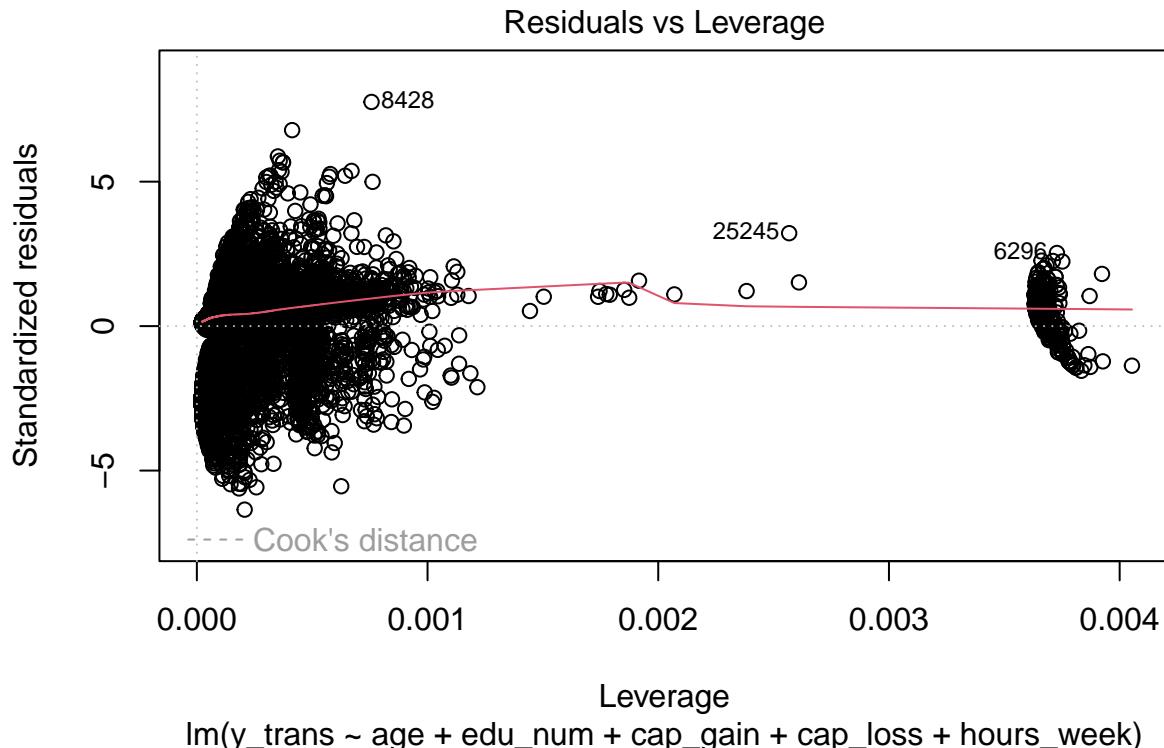
```
## Multiple R-squared:  0.7109, Adjusted R-squared:  0.7109  
## F-statistic: 2.402e+04 on 5 and 48836 DF,  p-value: < 2.2e-16
```

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

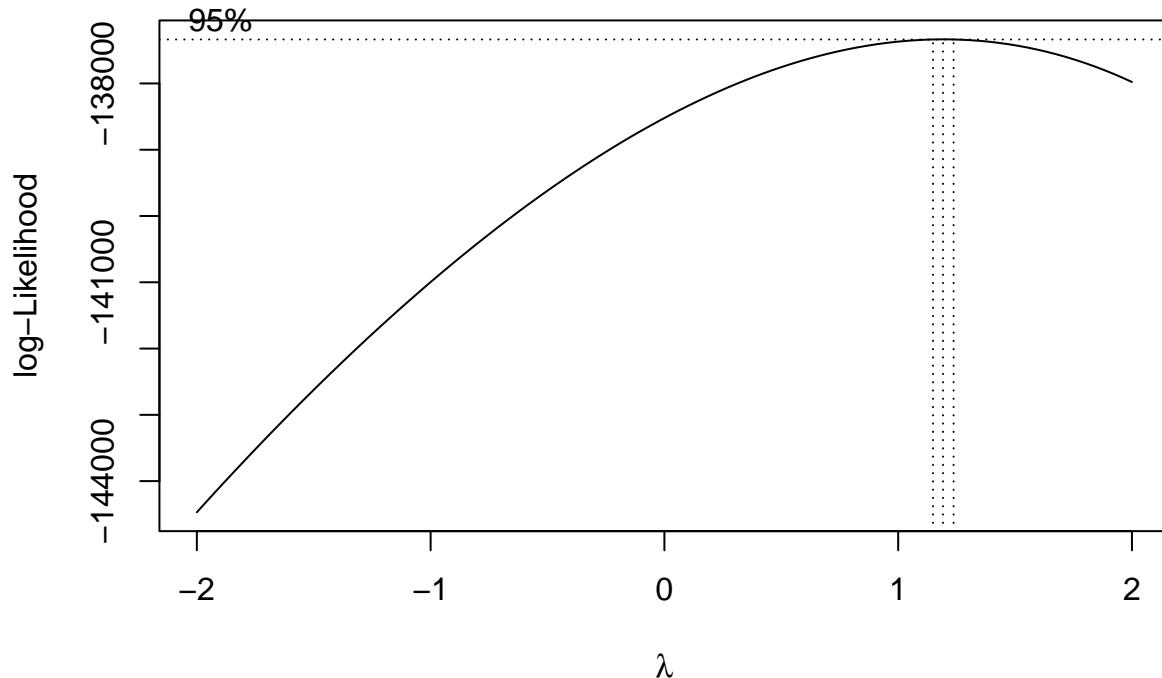








```
boxcox(y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week, data = dd) # lambda is now around 1
```



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

```
## MLE of lambda Score Statistic (t) Pr(>|t|)
## age          -0.69324      75.0355 < 2.2e-16 ***
## edu_num       0.38487      40.7224 < 2.2e-16 ***
## hours_week    0.85116      4.6881 2.764e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  4
##
## Score test for null hypothesis that all lambdas = 1:
## F = 2566.4, df = 3 and 48835, Pr(>F) = < 2.2e-16
```

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

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

```

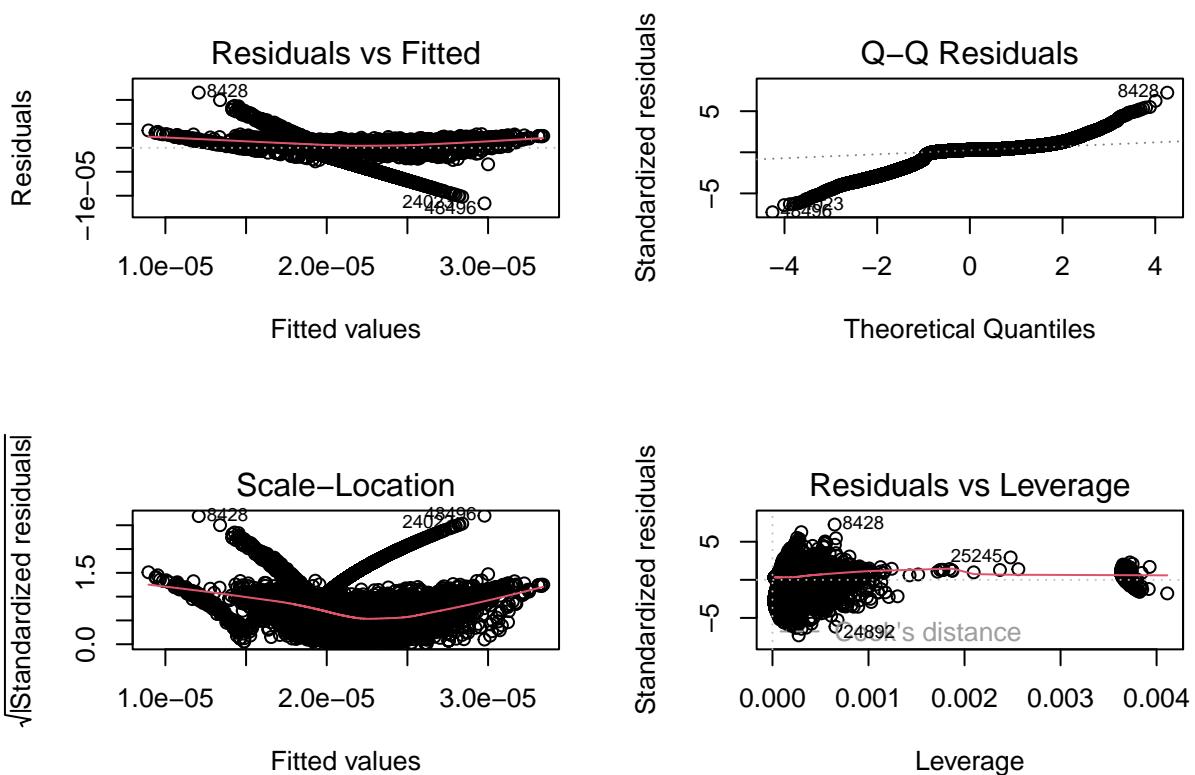
## Residuals:
##      Min       1Q    Median       3Q      Max
## -1.160e-05 1.249e-07 4.788e-07 6.404e-07 1.153e-05
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.542e-05 6.717e-08 676.15 <2e-16 ***
## agebt      -1.606e-06 6.642e-09 -241.80 <2e-16 ***
## edu_num_bt -3.585e-06 1.679e-08 -213.48 <2e-16 ***
## cap_gain   -5.684e-11 9.779e-13 -58.12 <2e-16 ***
## cap_loss   -2.708e-10 1.797e-11 -15.07 <2e-16 ***
## hours_week -4.752e-08 5.911e-10 -80.38 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.591e-06 on 48836 degrees of freedom
## Multiple R-squared:  0.7363, Adjusted R-squared:  0.7362
## F-statistic: 2.727e+04 on 5 and 48836 DF, p-value: < 2.2e-16

```

```

par(mfrow=c(2,2))
plot(btmodel)

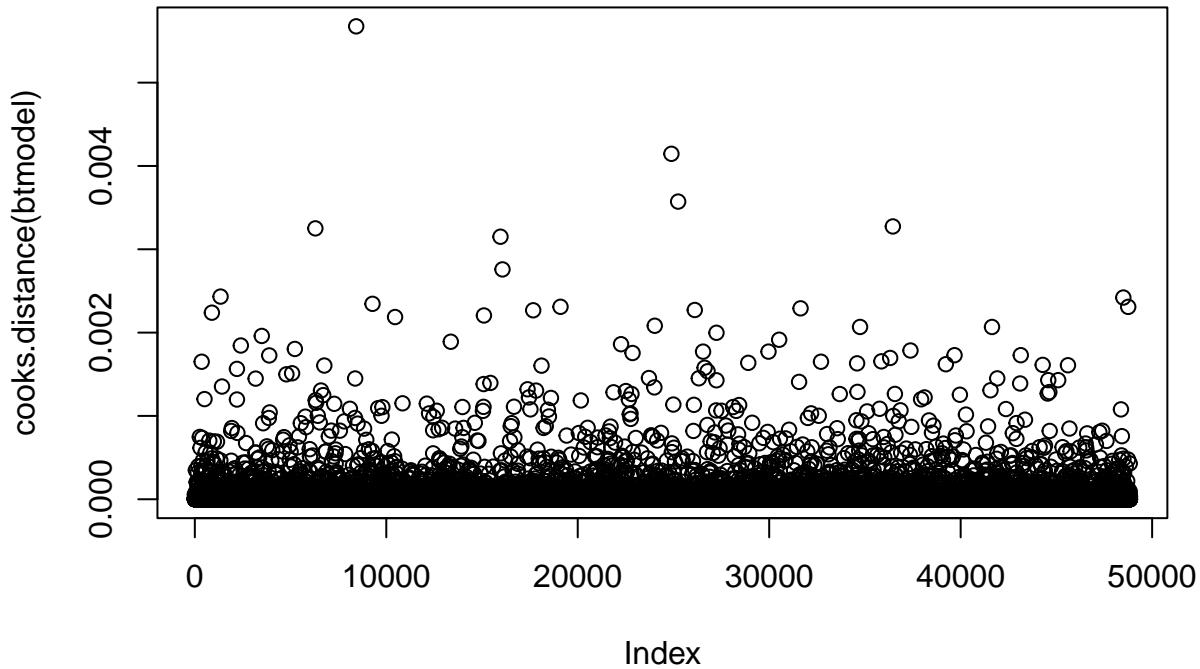
```



```

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

```



```

#Try adding polynomial terms
age2 <- dd$agebt^2
hours_week2 <- dd$hours_week^2
model_poly <- lm(y_trans ~ agebt + age2 + edu_num + cap_gain + cap_loss + hours_week + hours_week2 , data = dd)

#comparing model performance
summary(model_poly)

## 
## Call:
## lm(formula = y_trans ~ agebt + age2 + edu_num + cap_gain + cap_loss +
##     hours_week + hours_week2, data = dd)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.090e-05 -1.907e-07  5.084e-07  8.102e-07  8.100e-06 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 5.161e-05 2.063e-07 250.184 <2e-16 ***
## agebt      -5.548e-06 7.237e-08 -76.663 <2e-16 ***
## age2        3.172e-07 5.759e-09  55.070 <2e-16 ***
## edu_num    -5.729e-07 2.841e-09 -201.635 <2e-16 ***
## cap_gain   -5.647e-11 9.619e-13 -58.706 <2e-16 ***
## cap_loss   -2.568e-10 1.766e-11 -14.540 <2e-16 ***
## hours_week -5.224e-08 1.946e-09 -26.843 <2e-16 ***

```

```

## hours_week2  2.073e-10  2.140e-11     9.687   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.562e-06 on 48834 degrees of freedom
## Multiple R-squared:  0.7456, Adjusted R-squared:  0.7456
## F-statistic: 2.045e+04 on 7 and 48834 DF,  p-value: < 2.2e-16

anova(initial_model, model_poly)

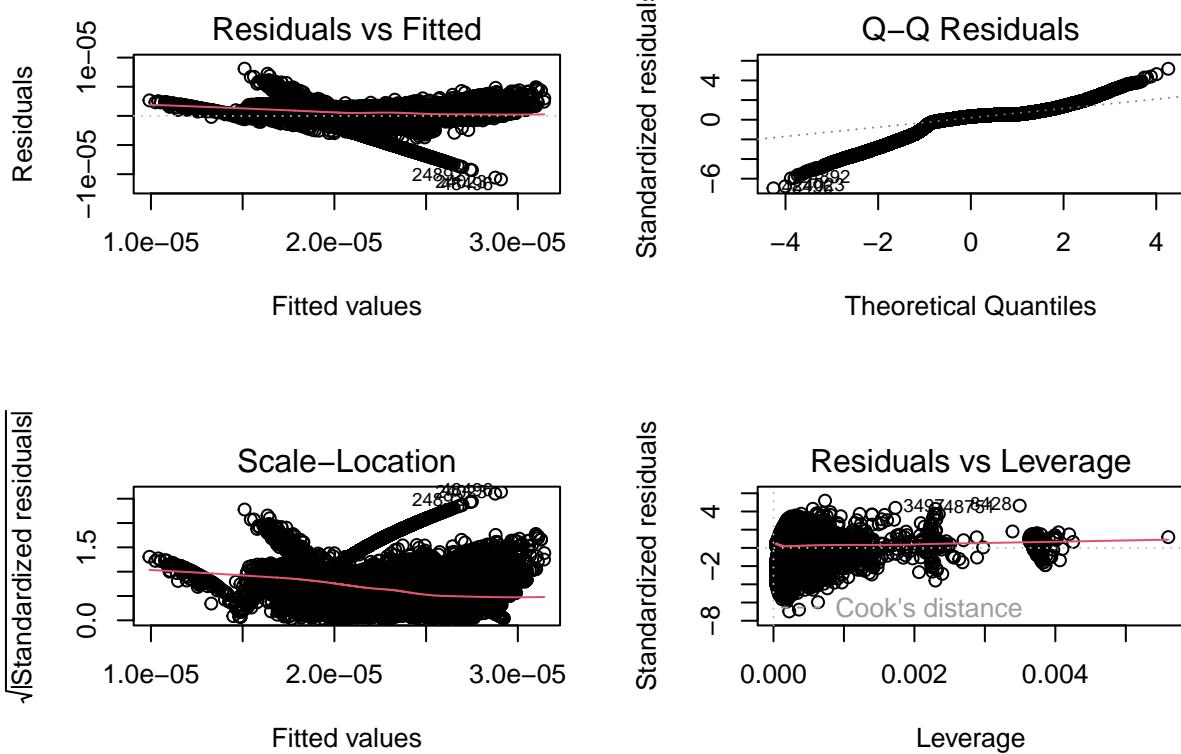
## Warning in anova.lmlist(object, ...): models with response '"y_trans"' removed
## because response differs from model 1

## Analysis of Variance Table

## Response: income_integer
##             Df    Sum Sq    Mean Sq  F value    Pr(>F)
## age          1 5.9443e+11 5.9443e+11 44080.01 < 2.2e-16 ***
## edu_num      1 5.3884e+11 5.3884e+11 39958.37 < 2.2e-16 ***
## cap_gain     1 1.1520e+11 1.1520e+11  8542.91 < 2.2e-16 ***
## cap_loss     1 6.5529e+09 6.5529e+09   485.93 < 2.2e-16 ***
## hours_week    1 7.1603e+10 7.1603e+10  5309.76 < 2.2e-16 ***
## Residuals  48836 6.5856e+11 1.3485e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

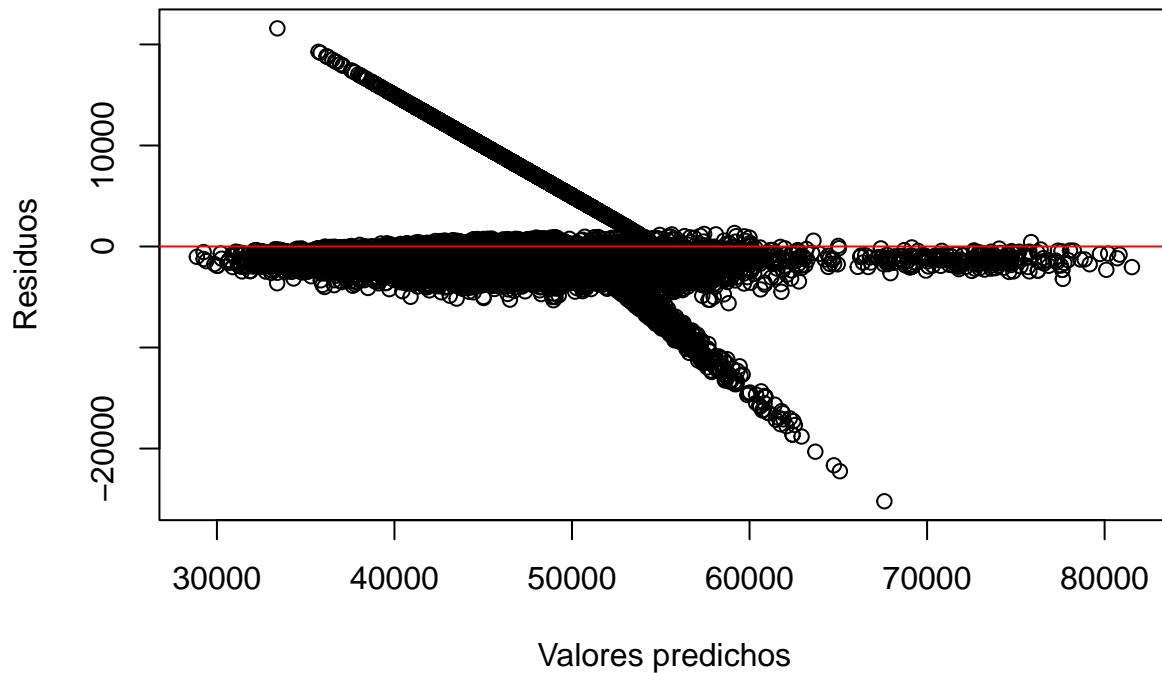
par(mfrow=c(2,2))
plot(model_poly)

```



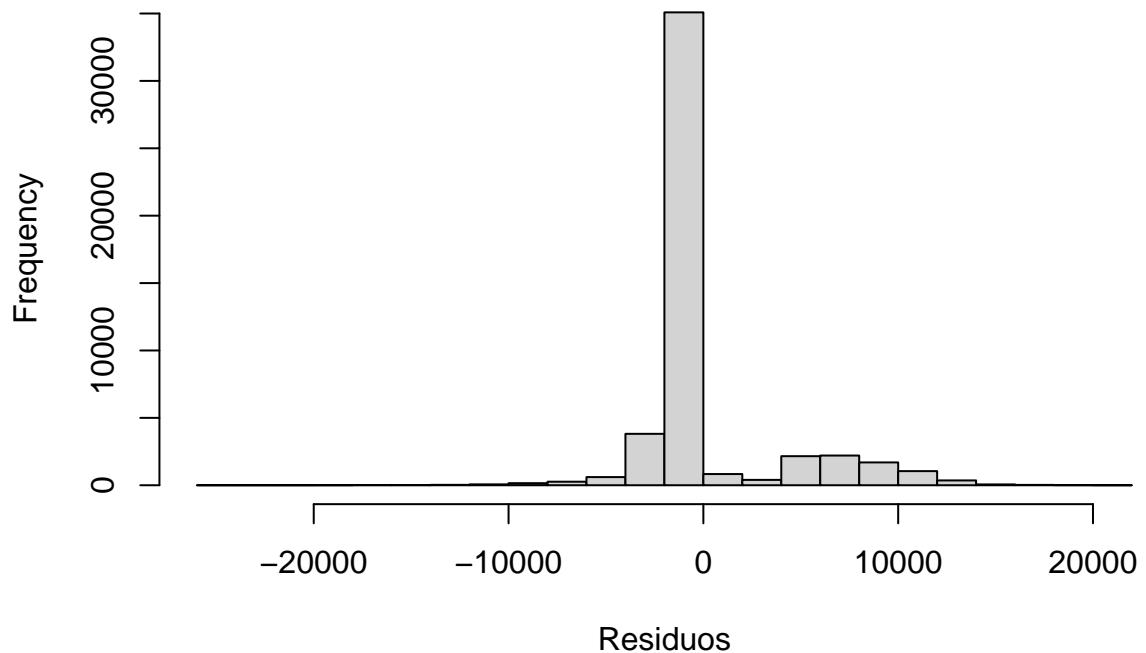
```
# Análisis de residuos (modelo original)
plot(initial_model$fitted.values, resid(initial_model),
     xlab = "Valores predichos", ylab = "Residuos",
     main = "Residuos vs Valores Predichos")
abline(h = 0, col = "red")
```

Residuos vs Valores Predichos



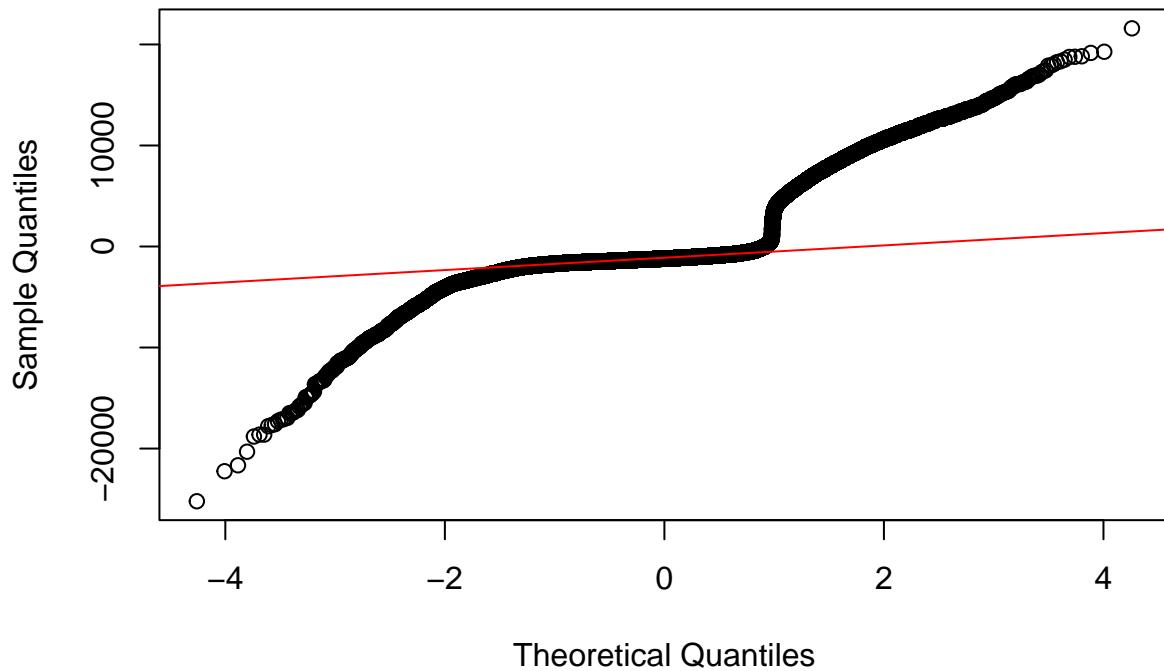
```
hist(resid(initial_model), main = "Histograma de residuos", xlab = "Residuos")
```

Histograma de residuos



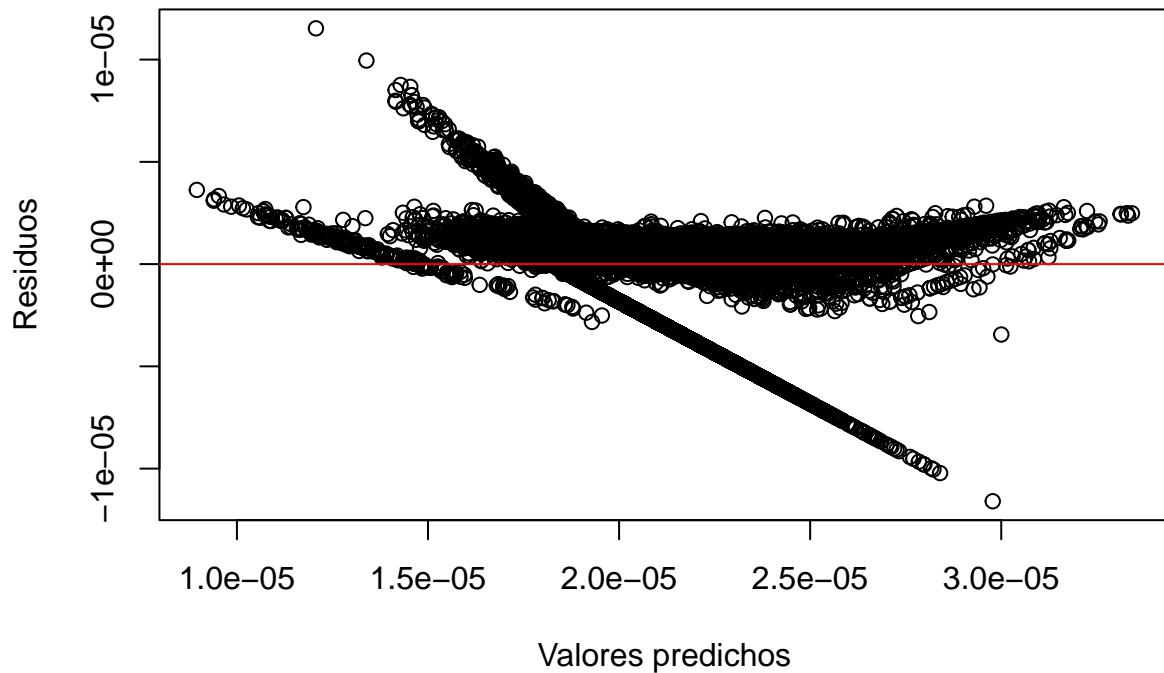
```
qqnorm(resid(initial_model))
qqline(resid(initial_model), col = "red")
```

Normal Q-Q Plot



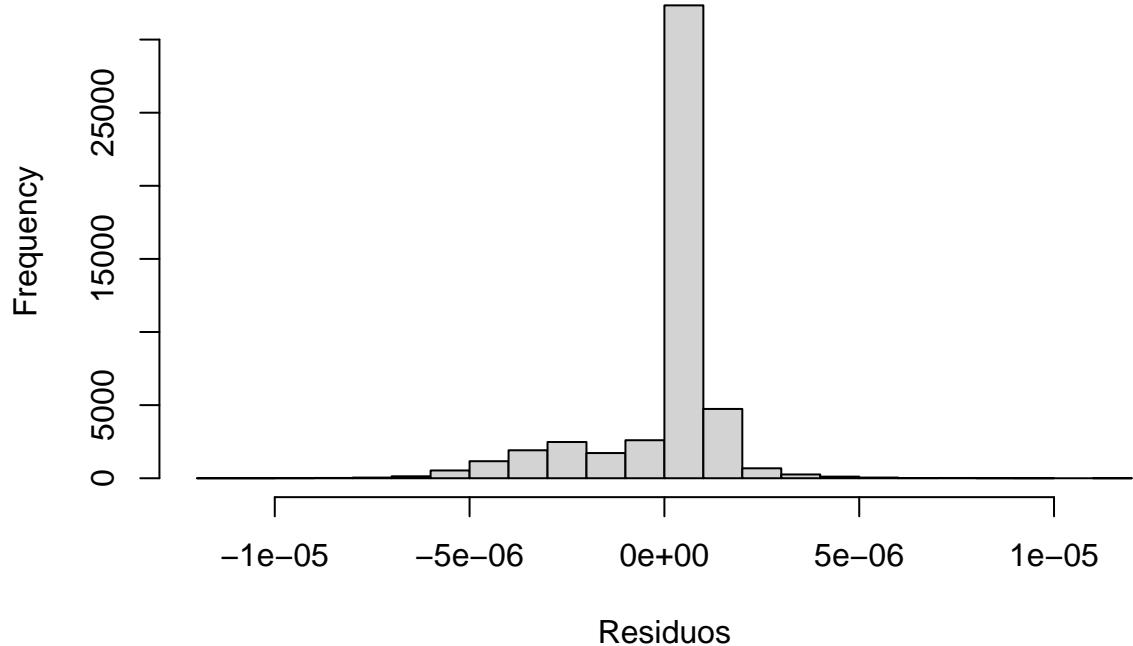
```
# Análisis de residuos (bt_model transformado)
plot(btmodel$fitted.values, resid(btmodel),
      xlab = "Valores predichos", ylab = "Residuos",
      main = "Residuos vs Valores Predichos")
abline(h = 0, col = "red")
```

Residuos vs Valores Predichos



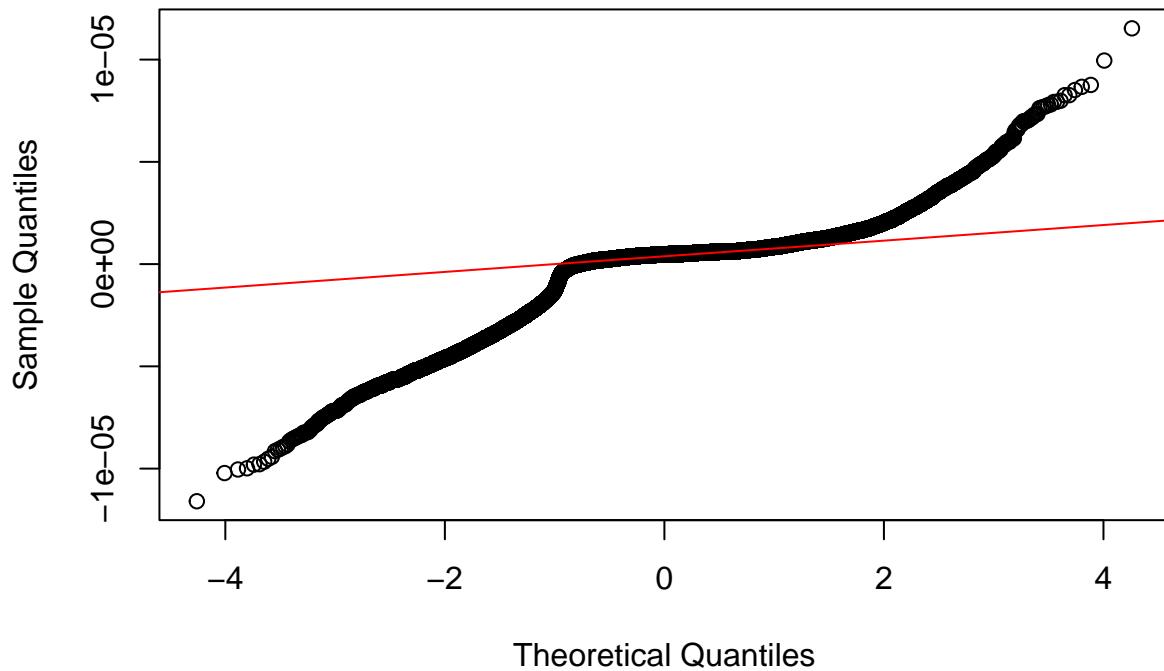
```
hist(resid(btmodel), main = "Histograma de residuos", xlab = "Residuos")
```

Histograma de residuos



```
qqnorm(resid(btmodel))
qqline(resid(btmodel), col = "red")
```

Normal Q-Q Plot



```
# Multicolinealidad (initial_model original)
vif(initial_model)
```

```
##      age     edu_num   cap_gain   cap_loss hours_week
## 1.013606 1.041277 1.027722 1.013638 1.031481
```

```
# Multicolinealidad (bt_model transformado)
vif(btmodel)
```

```
##      agebt edu_num_bt   cap_gain   cap_loss hours_week
## 1.020432 1.034825 1.024972 1.012790 1.035586
```

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

```
## Analysis of Variance Table
##
## Model 1: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week
## Model 2: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##           occupation
## Res.Df          RSS Df  Sum of Sq      F    Pr(>F)
## 1  48836 1.2357e-07
```

```

## 2 48823 1.2040e-07 13 3.1647e-09 98.712 < 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 ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
## occupation
## Model 2: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
## occupation + marital
## Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1 48823 1.2040e-07
## 2 48819 1.0817e-07 4 1.2237e-08 1380.7 < 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 = 0.009058 ***

## Analysis of Variance Table
##
## Model 1: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
## occupation + marital
## Model 2: y_trans ~ 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 1.0817e-07
## 2 48818 1.0815e-07 1 1.5091e-11 6.8119 0.009058 **
## ---
## 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 ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
## occupation + marital + sex
## Model 2: y_trans ~ 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 1.0815e-07
## 2 48812 1.0787e-07 6 2.8323e-10 21.36 < 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 ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##           occupation + marital + sex + workclass
## Model 2: y_trans ~ 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 1.0787e-07
## 2 48807 1.0664e-07  5 1.2266e-09 112.28 < 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 = 0.0008009

## Analysis of Variance Table
##
## Model 1: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##           occupation + marital + sex + workclass + relationship
## Model 2: y_trans ~ 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 1.0664e-07
## 2 48803 1.0660e-07  4 4.1418e-11 4.7404 0.0008009 ***
## ---
## 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 ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##           occupation + marital + sex + workclass + relationship + race
## Model 2: y_trans ~ 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 1.0660e-07
## 2 48802 1.0644e-07  1 1.5805e-10 72.462 < 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 ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##           occupation + marital + sex + workclass + relationship + race +
##           native_country
## Model 2: y_trans ~ 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 1.0644e-07
## 2 48801 4.9257e-08  1 5.7187e-08 56657 < 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 ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week + occupation + marital +
stepmodel <- stepAIC(catmodel, direction = "back")

```

```

## Start:  AIC=-1349059
## y_trans ~ 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>                 4.9257e-08 -1349059
## - sex                  1 4.8000e-11 4.9305e-08 -1349013
## - cap_loss              1 8.4000e-11 4.9341e-08 -1348978
## - workclass             6 1.0700e-10 4.9364e-08 -1348965
## - race                  4 1.0400e-10 4.9361e-08 -1348964
## - native_country         1 1.6400e-10 4.9420e-08 -1348899
## - marital                4 2.9400e-10 4.9551e-08 -1348777
## - relationship           5 9.7500e-10 5.0232e-08 -1348112
## - occupation             13 1.2400e-09 5.0497e-08 -1347871
## - cap_gain               1 1.9790e-09 5.1236e-08 -1347137
## - hours_week              1 4.8230e-09 5.4080e-08 -1344499
## - agebt                  1 5.0523e-08 9.9780e-08 -1314583
## - edu_num_bt              1 5.1493e-08 1.0075e-07 -1314110
## - income                  1 5.7187e-08 1.0644e-07 -1311425

```

```
vif(catmodel)
```

	GVIF	Df	GVIF^(1/(2*Df))
## agebt	1.781237	1	1.334630
## edu_num_bt	1.429725	1	1.195711
## cap_gain	1.071123	1	1.034951
## cap_loss	1.030597	1	1.015183
## hours_week	1.221805	1	1.105353
## occupation	2.257130	13	1.031807
## marital	60.315146	4	1.669373
## sex	1.990125	1	1.410718
## workclass	1.441251	6	1.030928
## relationship	75.507075	5	1.540986
## race	1.283651	4	1.031706

```

## native_country 1.239624 1      1.113384
## income         1.549349 1      1.244728

summary(catmodel)

##
## Call:
## lm(formula = y_trans ~ 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
## -8.257e-06 -5.550e-07 -8.660e-08  3.838e-07  7.935e-06
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 4.073e-05 9.922e-08 410.476 < 2e-16 ***
## agebt       -1.240e-06 5.542e-09 -223.731 < 2e-16 ***
## edu_num_bt  -2.816e-06 1.247e-08 -225.869 < 2e-16 ***
## cap_gain    -2.796e-11 6.314e-13 -44.280 < 2e-16 ***
## cap_loss     1.046e-10 1.145e-11   9.131 < 2e-16 ***
## hours_week  -2.803e-08 4.055e-10 -69.124 < 2e-16 ***
## occupationArmy 2.978e-07 2.611e-07   1.141 0.254025
## occupationCraftRep -7.949e-08 2.017e-08  -3.940 8.14e-05 ***
## occupationExecMan 2.275e-07 1.960e-08  11.607 < 2e-16 ***
## occupationFarmFish 1.664e-07 3.137e-08   5.304 1.14e-07 ***
## occupationHandlCl 2.901e-07 2.697e-08  10.756 < 2e-16 ***
## occupationHouse 3.830e-07 6.655e-08   5.755 8.70e-09 ***
## occupationMachOp 8.582e-08 2.361e-08   3.634 0.000279 ***
## occupationOther 2.784e-07 2.008e-08  13.860 < 2e-16 ***
## occupationProf 3.612e-07 1.776e-08  20.340 < 2e-16 ***
## occupationProtServ -1.413e-07 3.627e-08  -3.897 9.77e-05 ***
## occupationSales 1.443e-07 1.977e-08   7.297 2.99e-13 ***
## occupationTech -1.148e-07 2.987e-08  -3.843 0.000122 ***
## occupationTrans -8.449e-08 2.598e-08  -3.253 0.001145 ** 
## maritalMarried -1.581e-07 5.642e-08  -2.803 0.005072 ** 
## maritalNevMarr 2.225e-07 1.719e-08  12.945 < 2e-16 ***
## maritalSep     8.720e-08 2.516e-08   3.466 0.000529 ***
## maritalWidow  -1.863e-07 2.954e-08  -6.307 2.87e-10 ***
## sexMale        9.425e-08 1.362e-08   6.919 4.61e-12 ***
## workclassLoc  -3.535e-08 3.269e-08  -1.081 0.279562
## workclassNoPay 2.314e-07 1.828e-07   1.266 0.205479
## workclassPriv -3.407e-08 2.776e-08  -1.227 0.219771
## workclassSelfI 2.224e-07 3.698e-08   6.015 1.81e-09 ***
## workclassSelfN 1.276e-09 3.221e-08   0.040 0.968406
## workclassState 2.610e-08 3.522e-08   0.741 0.458720
## relationshipNot-in-family -2.334e-07 5.621e-08  -4.153 3.29e-05 ***
## relationshipOther-relative 1.129e-07 5.518e-08   2.046 0.040760 *
## relationshipOwn-child    2.358e-07 5.614e-08   4.200 2.67e-05 ***
## relationshipUnmarried   -9.660e-08 5.829e-08  -1.657 0.097445 .
## relationshipWife       -2.354e-07 2.564e-08  -9.182 < 2e-16 ***
## raceAsian-Pac-Islander -2.456e-07 5.447e-08  -4.509 6.53e-06 ***
## raceBlack          -7.987e-08 4.871e-08  -1.640 0.101042

```

```

## raceOther           1.078e-07 6.862e-08   1.571 0.116123
## raceWhite          -5.742e-10 4.674e-08  -0.012 0.990198
## native_countryUSA -2.123e-07 1.668e-08  -12.728 < 2e-16 ***
## income>50K         -3.157e-06 1.326e-08 -238.028 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.005e-06 on 48801 degrees of freedom
## Multiple R-squared:  0.8949, Adjusted R-squared:  0.8948
## F-statistic: 1.039e+04 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)
## agebt        1 1.7759e-07 1.7759e-07 175949.879 < 2.2e-16 ***
## edu_num_bt   1 1.4030e-07 1.4030e-07 138997.998 < 2.2e-16 ***
## cap_gain     1 9.8790e-09 9.8790e-09  9787.355 < 2.2e-16 ***
## cap_loss     1 8.6200e-10 8.6200e-10   853.794 < 2.2e-16 ***
## hours_week   1 1.6350e-08 1.6350e-08 16198.266 < 2.2e-16 ***
## occupation   13 3.1650e-09 2.4300e-10   241.185 < 2.2e-16 ***
## marital      4 1.2237e-08 3.0590e-09  3030.888 < 2.2e-16 ***
## sex          1 1.5000e-11 1.5000e-11   14.951 0.0001105 ***
## workclass    6 2.8300e-10 4.7000e-11   46.767 < 2.2e-16 ***
## relationship 5 1.2270e-09 2.4500e-10  243.055 < 2.2e-16 ***
## race         4 4.1000e-11 1.0000e-11   10.259 2.664e-08 ***
## native_country 1 1.5800e-10 1.5800e-10   156.585 < 2.2e-16 ***
## income       1 5.7187e-08 5.7187e-08  56657.237 < 2.2e-16 ***
## Residuals    48801 4.9257e-08 1.0000e-12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

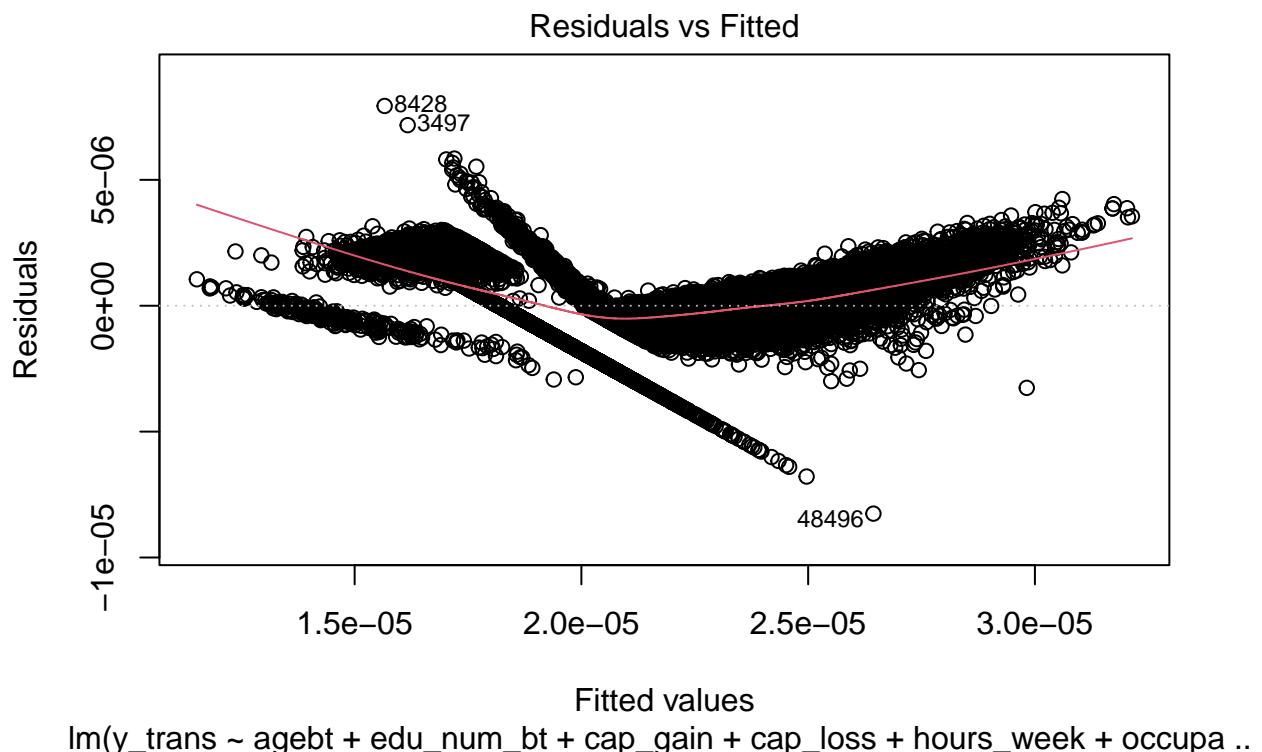
```
anova(transformed_model, catmodel)
```

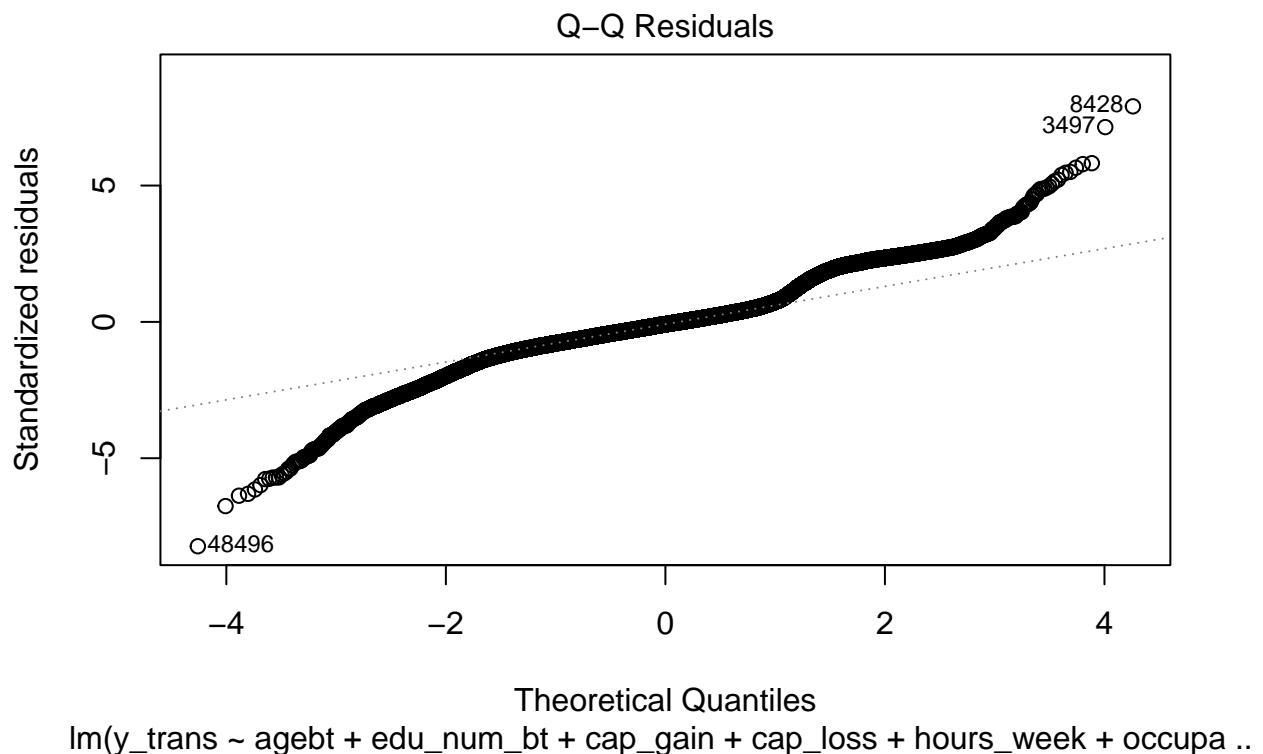
```

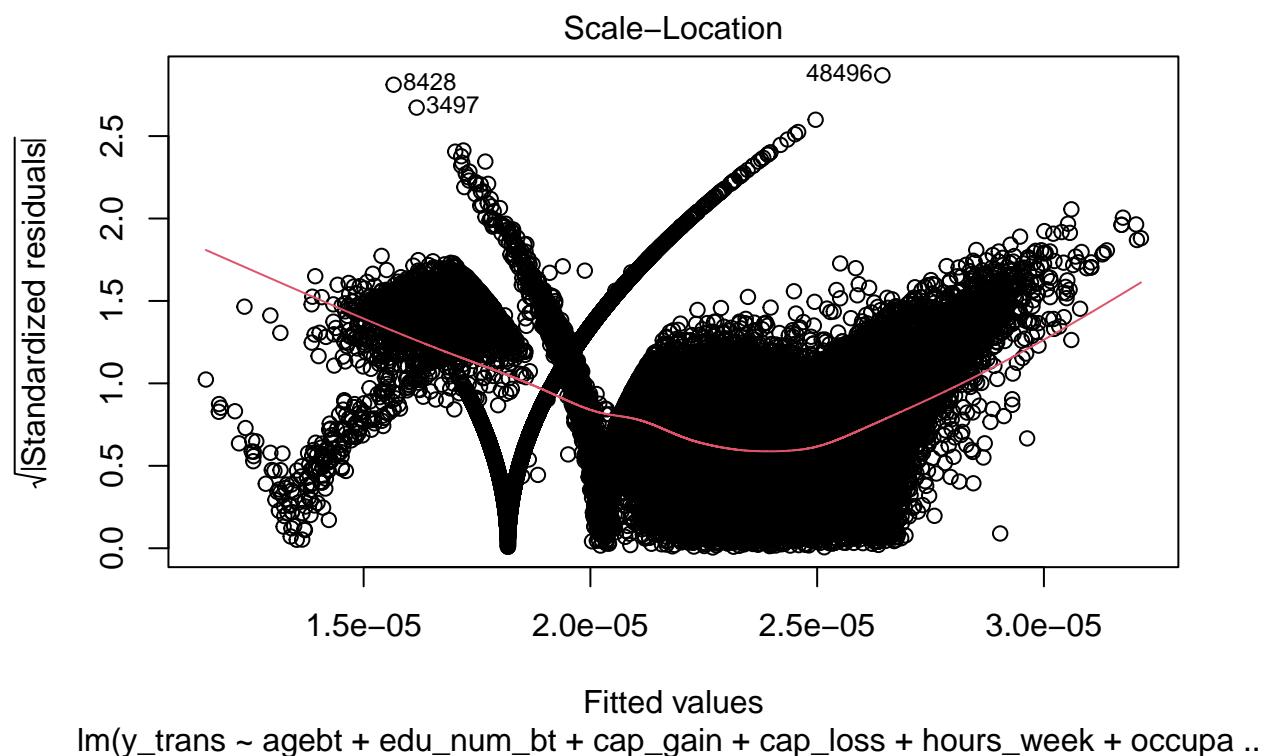
## Analysis of Variance Table
##
## Model 1: y_trans ~ age + edu_num + cap_gain + cap_loss + hours_week
## Model 2: y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##             occupation + marital + sex + workclass + relationship + race +
##             native_country + income
##   Res.Df    RSS Df  Sum of Sq   F   Pr(>F)
## 1  48836 1.3544e-07
## 2  48801 4.9257e-08 35 8.6187e-08 2439.7 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

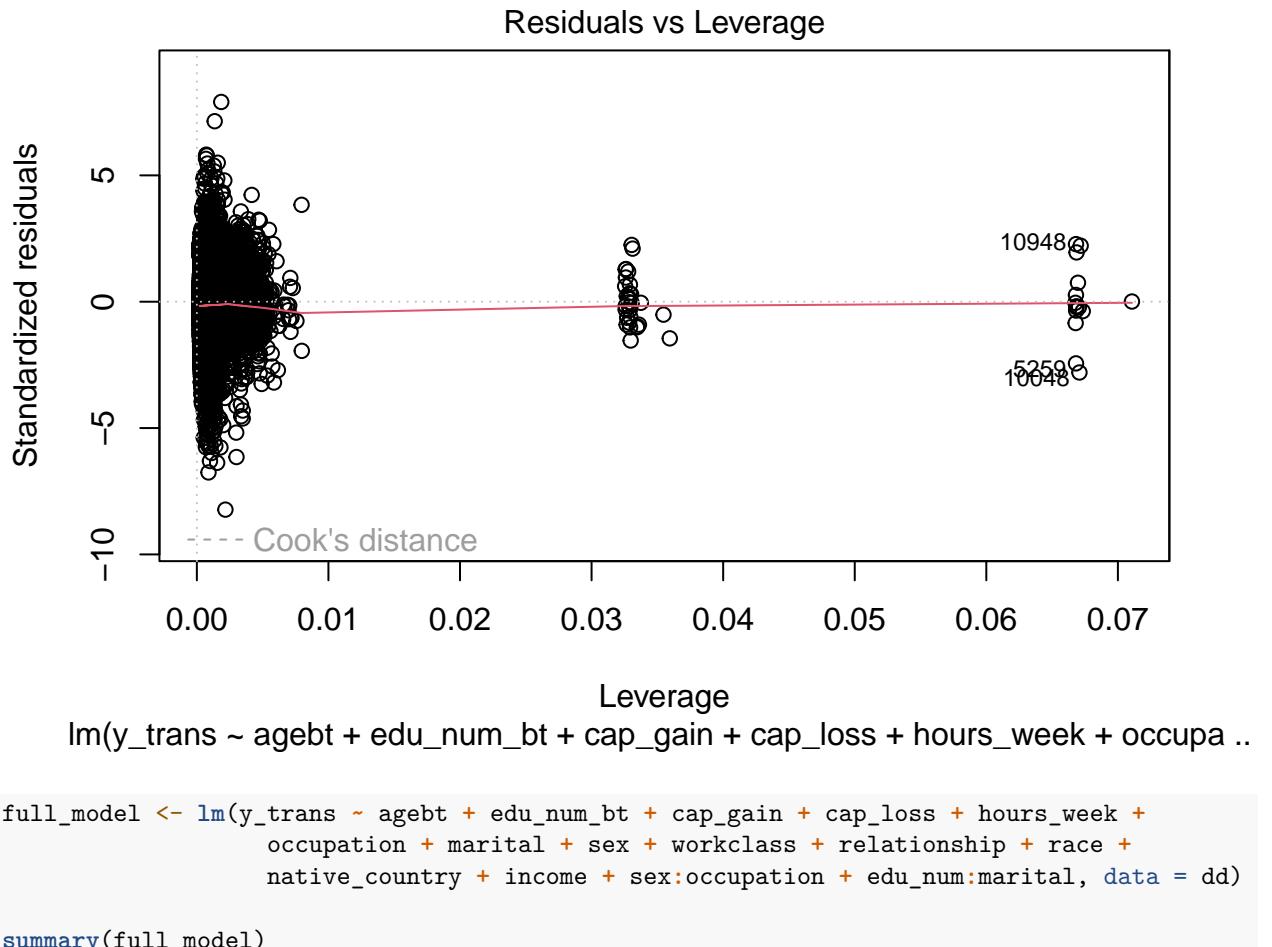
```

```
plot(catmodel)
```









```

##
## Call:
## lm(formula = y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss +
##     hours_week + occupation + marital + sex + workclass + relationship +
##     race + native_country + income + sex:occupation + edu_num:marital,
##     data = dd)
##
## Residuals:
##      Min        1Q    Median        3Q       Max
## -1.002e-05 -4.469e-07 -4.050e-08  3.931e-07  7.453e-06
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.568e-05 1.470e-07 310.629 < 2e-16 ***
## agebt       -1.214e-06 4.992e-09 -243.222 < 2e-16 ***
## edu_num_bt  -6.780e-06 7.403e-08 -91.586 < 2e-16 ***
## cap_gain    -3.115e-11 5.636e-13 -55.274 < 2e-16 ***
## cap_loss    7.709e-11 1.020e-11   7.554 4.29e-14 ***
## hours_week  -2.671e-08 3.629e-10 -73.619 < 2e-16 ***

```

## occupationArmy	2.571e-07	2.330e-07	1.103	0.269883
## occupationCraftRep	2.561e-08	5.199e-08	0.493	0.622354
## occupationExecMan	-4.778e-10	2.620e-08	-0.018	0.985451
## occupationFarmFish	2.250e-07	9.332e-08	2.411	0.015901 *
## occupationHandlCl	1.423e-07	5.814e-08	2.447	0.014429 *
## occupationHouse	5.819e-09	6.167e-08	0.094	0.924830
## occupationMachOp	7.918e-08	3.514e-08	2.253	0.024253 *
## occupationOther	1.853e-07	2.285e-08	8.110	5.18e-16 ***
## occupationProf	1.060e-07	2.147e-08	4.935	8.04e-07 ***
## occupationProtServ	5.344e-08	8.266e-08	0.647	0.517935
## occupationSales	1.806e-07	2.520e-08	7.165	7.87e-13 ***
## occupationTech	-1.163e-07	4.053e-08	-2.870	0.004110 **
## occupationTrans	6.822e-09	8.084e-08	0.084	0.932739
## maritalMarried	-1.080e-06	7.278e-08	-14.837	< 2e-16 ***
## maritalNevMarr	2.752e-06	6.000e-08	45.858	< 2e-16 ***
## maritalSep	4.225e-07	8.545e-08	4.944	7.67e-07 ***
## maritalWidow	-2.038e-06	9.467e-08	-21.525	< 2e-16 ***
## sexMale	-6.636e-08	2.650e-08	-2.504	0.012280 *
## workclassLoc	-6.722e-08	2.925e-08	-2.298	0.021577 *
## workclassNoPay	2.392e-07	1.630e-07	1.467	0.142282
## workclassPriv	-2.480e-08	2.483e-08	-0.999	0.317789
## workclassSelfI	1.610e-07	3.311e-08	4.862	1.17e-06 ***
## workclassSelfN	-1.652e-08	2.882e-08	-0.573	0.566637
## workclassState	-2.388e-08	3.146e-08	-0.759	0.447916
## relationshipNot-in-family	5.333e-08	5.024e-08	1.062	0.288382
## relationshipOther-relative	1.832e-07	4.919e-08	3.724	0.000196 ***
## relationshipOwn-child	4.002e-07	5.007e-08	7.993	1.34e-15 ***
## relationshipUnmarried	7.771e-08	5.200e-08	1.494	0.135114
## relationshipWife	-2.564e-07	2.302e-08	-11.141	< 2e-16 ***
## raceAsian-Pac-Islander	-1.748e-07	4.853e-08	-3.602	0.000316 ***
## raceBlack	-6.529e-08	4.340e-08	-1.504	0.132472
## raceOther	9.531e-08	6.113e-08	1.559	0.118978
## raceWhite	4.619e-09	4.164e-08	0.111	0.911681
## native_countryUSA	-1.757e-08	1.522e-08	-1.155	0.248276
## income>50K	-3.407e-06	1.207e-08	-282.260	< 2e-16 ***
## occupationArmy:sexMale	NA	NA	NA	NA
## occupationCraftRep:sexMale	2.416e-08	5.729e-08	0.422	0.673259
## occupationExecMan:sexMale	2.295e-07	3.630e-08	6.323	2.60e-10 ***
## occupationFarmFish:sexMale	-9.963e-08	9.850e-08	-1.012	0.311780
## occupationHandlCl:sexMale	1.408e-07	6.524e-08	2.158	0.030904 *
## occupationHouse:sexMale	4.951e-07	2.479e-07	1.997	0.045800 *
## occupationMachOp:sexMale	4.718e-08	4.499e-08	1.049	0.294285
## occupationOther:sexMale	2.715e-08	3.633e-08	0.747	0.454873
## occupationProf:sexMale	1.366e-07	3.228e-08	4.230	2.34e-05 ***
## occupationProtServ:sexMale	-1.057e-07	9.032e-08	-1.170	0.242040
## occupationSales:sexMale	-8.991e-08	3.615e-08	-2.487	0.012889 *
## occupationTech:sexMale	4.481e-08	5.463e-08	0.820	0.412014
## occupationTrans:sexMale	6.095e-08	8.567e-08	0.711	0.476815
## maritalDiv:edu_num	6.989e-07	1.332e-08	52.464	< 2e-16 ***
## maritalMarried:edu_num	8.181e-07	1.285e-08	63.649	< 2e-16 ***
## maritalNevMarr:edu_num	4.515e-07	1.302e-08	34.670	< 2e-16 ***
## maritalSep:edu_num	6.647e-07	1.509e-08	44.037	< 2e-16 ***
## maritalWidow:edu_num	8.992e-07	1.584e-08	56.777	< 2e-16 ***
## ---				

```

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.948e-07 on 48784 degrees of freedom
## Multiple R-squared:  0.9166, Adjusted R-squared:  0.9165
## F-statistic:  9410 on 57 and 48784 DF,  p-value: < 2.2e-16

reduced_model <- step(full_model)

## Start:  AIC=-1360353
## y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##          occupation + marital + sex + workclass + relationship + race +
##          native_country + income + sex:occupation + edu_num:marital
##
##              Df  Sum of Sq      RSS      AIC
## - native_country   1 1.0000e-12 3.9062e-08 -1360353
## <none>                  3.9061e-08 -1360353
## - cap_loss         1 4.6000e-11 3.9107e-08 -1360298
## - workclass        6 6.1000e-11 3.9122e-08 -1360288
## - race             4 6.2000e-11 3.9123e-08 -1360283
## - occupation:sex  12 9.0000e-11 3.9151e-08 -1360265
## - relationship     5 5.9700e-10 3.9658e-08 -1359622
## - cap_gain          1 2.4460e-09 4.1507e-08 -1357388
## - hours_week        1 4.3400e-09 4.3401e-08 -1355209
## - edu_num_bt        1 6.7160e-09 4.5777e-08 -1352605
## - marital:edu_num  5 9.8820e-09 4.8943e-08 -1349347
## - agebt             1 4.7367e-08 8.6428e-08 -1321565
## - income            1 6.3792e-08 1.0285e-07 -1313067
##
## Step:  AIC=-1360353
## y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss + hours_week +
##          occupation + marital + sex + workclass + relationship + race +
##          income + occupation:sex + marital:edu_num
##
##              Df  Sum of Sq      RSS      AIC
## <none>                  3.9062e-08 -1360353
## - cap_loss         1 4.6000e-11 3.9108e-08 -1360298
## - workclass        6 6.1000e-11 3.9124e-08 -1360289
## - race             4 6.2000e-11 3.9124e-08 -1360284
## - occupation:sex  12 9.0000e-11 3.9152e-08 -1360265
## - relationship     5 5.9600e-10 3.9658e-08 -1359624
## - cap_gain          1 2.4470e-09 4.1510e-08 -1357387
## - hours_week        1 4.3390e-09 4.3401e-08 -1355211
## - edu_num_bt        1 7.1180e-09 4.6180e-08 -1352180
## - marital:edu_num  5 1.0042e-08 4.9104e-08 -1349189
## - agebt             1 4.7414e-08 8.6476e-08 -1321540
## - income            1 6.3832e-08 1.0289e-07 -1313050

summary(reduced_model)

##
## Call:
## lm(formula = y_trans ~ agebt + edu_num_bt + cap_gain + cap_loss +
##     hours_week + occupation + marital + sex + workclass + relationship +

```

```

##      race + income + occupation:sex + marital:edu_num, data = dd)
##
## Residuals:
##      Min       1Q    Median     3Q    Max
## -1.001e-05 -4.472e-07 -3.970e-08  3.933e-07  7.451e-06
##
## Coefficients: (1 not defined because of singularities)
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.569e-05  1.467e-07 311.344 < 2e-16 ***
## agebt                      -1.214e-06  4.990e-09 -243.342 < 2e-16 ***
## edu_num_bt                  -6.800e-06  7.212e-08 -94.282 < 2e-16 ***
## cap_gain                     -3.116e-11  5.636e-13 -55.285 < 2e-16 ***
## cap_loss                      7.701e-11  1.020e-11   7.546 4.56e-14 ***
## hours_week                   -2.671e-08  3.629e-10 -73.612 < 2e-16 ***
## occupationArmy                2.569e-07  2.330e-07   1.103 0.270244
## occupationCraftRep            2.589e-08  5.199e-08   0.498 0.618574
## occupationExecMan              -6.786e-10  2.620e-08  -0.026 0.979337
## occupationFarmFish             2.252e-07  9.332e-08   2.413 0.015819 *
## occupationHandlC1               1.426e-07  5.814e-08   2.453 0.014181 *
## occupationHouse                 9.274e-09  6.160e-08   0.151 0.880324
## occupationMachOp                8.009e-08  3.513e-08   2.279 0.022646 *
## occupationOther                  1.856e-07  2.285e-08   8.124 4.63e-16 ***
## occupationProf                  1.058e-07  2.147e-08   4.929 8.31e-07 ***
## occupationProtServ              5.319e-08  8.266e-08   0.644 0.519880
## occupationSales                  1.804e-07  2.520e-08   7.159 8.26e-13 ***
## occupationTech                  -1.166e-07  4.053e-08  -2.876 0.004032 **
## occupationTrans                  7.127e-09  8.084e-08   0.088 0.929743
## maritalMarried                 -1.077e-06  7.275e-08  -14.809 < 2e-16 ***
## maritalNevMarr                  2.753e-06  5.998e-08   45.909 < 2e-16 ***
## maritalSep                      4.263e-07  8.539e-08   4.992 5.99e-07 ***
## maritalWidow                     -2.039e-06  9.467e-08  -21.535 < 2e-16 ***
## sexMale                         -6.591e-08  2.650e-08  -2.488 0.012862 *
## workclassLoc                     -6.740e-08  2.925e-08  -2.304 0.021222 *
## workclassNoPay                   2.383e-07  1.630e-07   1.462 0.143858
## workclassPriv                   -2.444e-08  2.483e-08  -0.984 0.324941
## workclassSelfI                  1.615e-07  3.310e-08   4.879 1.07e-06 ***
## workclassSelfN                  -1.633e-08  2.882e-08  -0.567 0.571056
## workclassState                   -2.421e-08  3.146e-08  -0.770 0.441579
## relationshipNot-in-family        5.415e-08  5.023e-08   1.078 0.280998
## relationshipOther-relative       1.858e-07  4.913e-08   3.782 0.000156 ***
## relationshipOwn-child             4.005e-07  5.007e-08   7.998 1.29e-15 ***
## relationshipUnmarried             7.898e-08  5.199e-08   1.519 0.128758
## relationshipWife                  -2.559e-07  2.301e-08  -11.122 < 2e-16 ***
## raceAsian-Pac-Islander           -1.632e-07  4.747e-08  -3.438 0.000588 ***
## raceBlack                        -6.454e-08  4.339e-08  -1.487 0.136905
## raceOther                         1.029e-07  6.078e-08   1.693 0.090555 .
## raceWhite                        5.380e-09  4.164e-08   0.129 0.897189
## income>50K                       -3.407e-06  1.207e-08 -282.346 < 2e-16 ***
## occupationArmy:sexMale             NA          NA          NA          NA
## occupationCraftRep:sexMale         2.372e-08  5.729e-08   0.414 0.678841
## occupationExecMan:sexMale          2.293e-07  3.629e-08   6.317 2.70e-10 ***
## occupationFarmFish:sexMale         -1.002e-07  9.850e-08  -1.017 0.309164
## occupationHandlC1:sexMale          1.404e-07  6.524e-08   2.152 0.031371 *
## occupationHouse:sexMale            4.917e-07  2.479e-07   1.984 0.047292 *

```

```

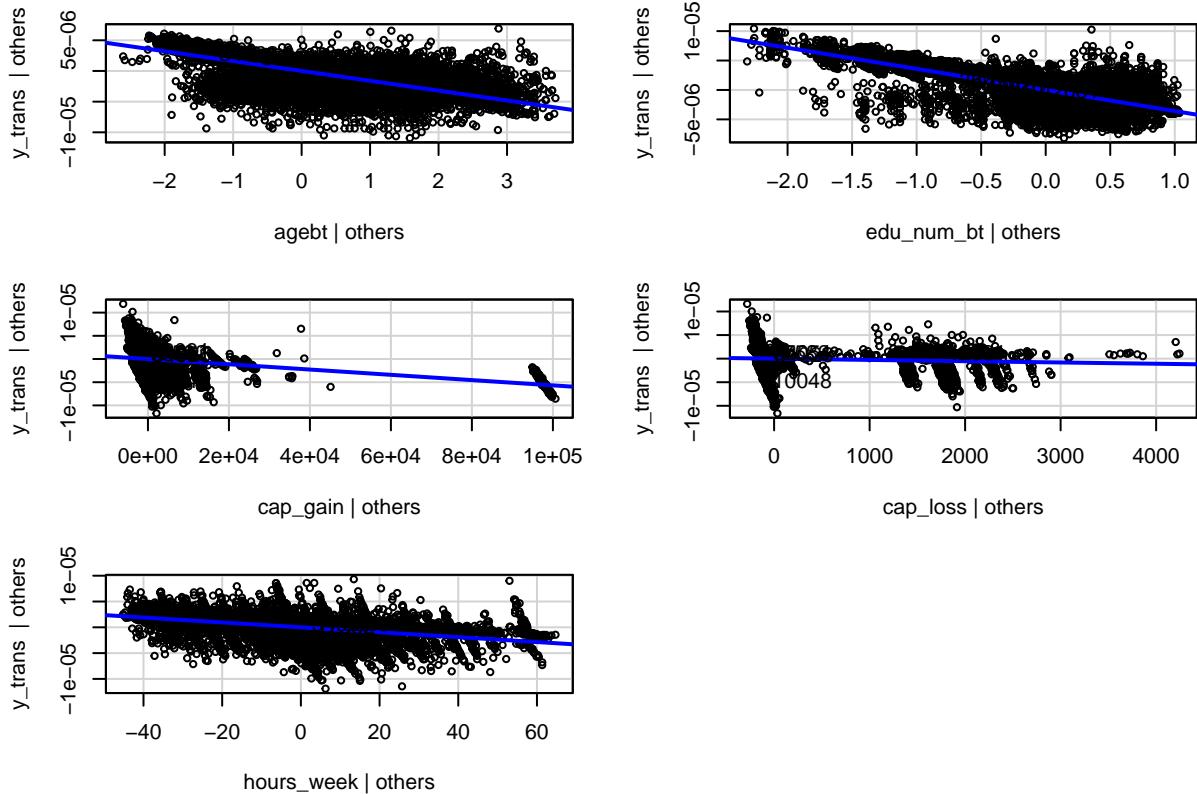
## occupationMachOp:sexMale      4.639e-08  4.498e-08   1.031  0.302404
## occupationOther:sexMale       2.757e-08  3.633e-08   0.759  0.447899
## occupationProf:sexMale        1.363e-07  3.228e-08   4.222  2.42e-05 ***
## occupationProtServ:sexMale    -1.056e-07 9.032e-08  -1.169  0.242421
## occupationSales:sexMale       -9.026e-08 3.615e-08  -2.497  0.012542 *
## occupationTech:sexMale        4.475e-08  5.463e-08   0.819  0.412660
## occupationTrans:sexMale       5.997e-08  8.567e-08   0.700  0.483941
## maritalDiv:edu_num            7.021e-07  1.303e-08  53.867 < 2e-16 ***
## maritalMarried:edu_num        8.212e-07  1.258e-08  65.291 < 2e-16 ***
## maritalNevMarr:edu_num        4.546e-07  1.276e-08  35.628 < 2e-16 ***
## maritalSep:edu_num            6.677e-07  1.488e-08  44.875 < 2e-16 ***
## maritalWidow:edu_num          9.025e-07  1.558e-08  57.937 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.948e-07 on 48785 degrees of freedom
## Multiple R-squared:  0.9166, Adjusted R-squared:  0.9165
## F-statistic:  9578 on 56 and 48785 DF,  p-value: < 2.2e-16

```

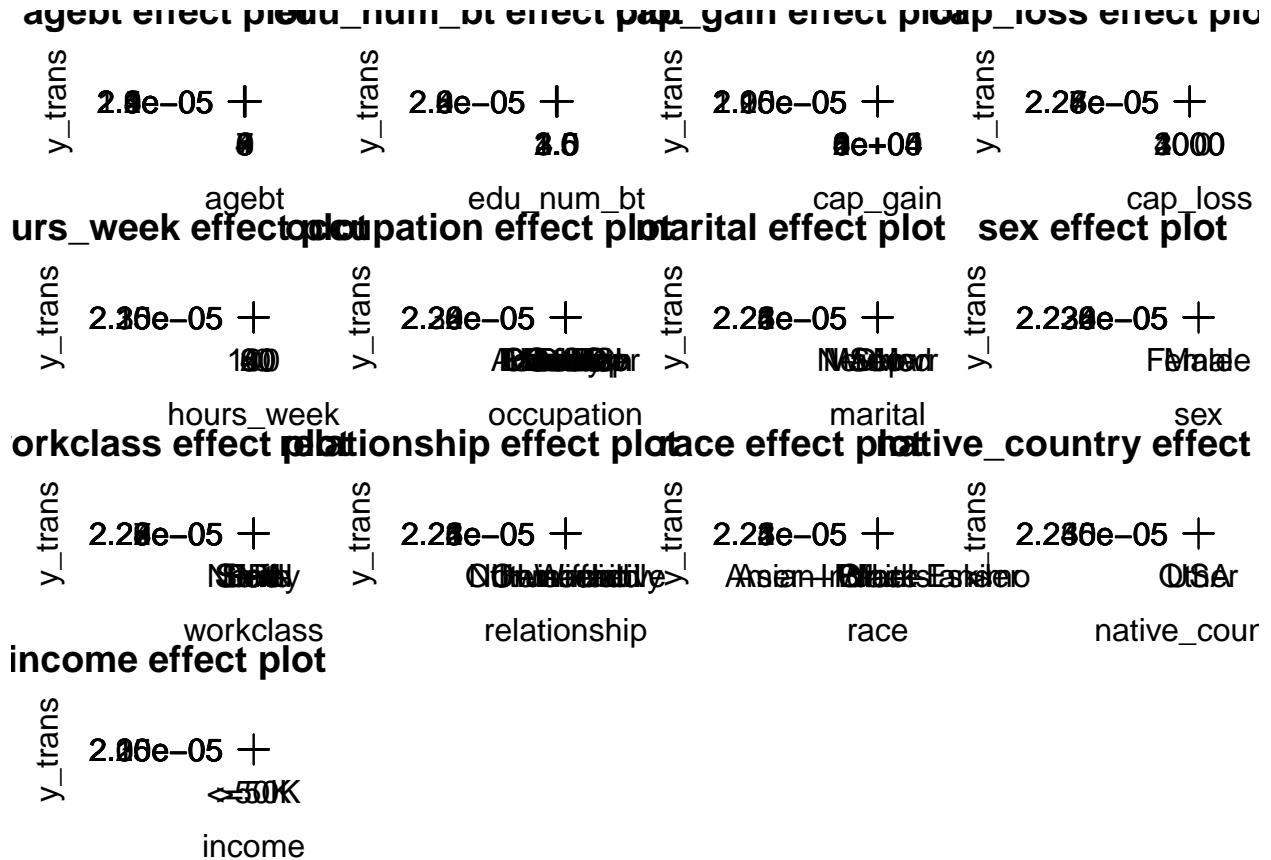
#Possible Conclusions

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

Added-Variable Plots



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



```

income_bin <- ifelse(dd$income == ">50K", 1, 0)
income_bin <- as.factor(income_bin)
set.seed(123)
total_n   <- nrow(dd)
train_index <- sample(seq_len(total_n), size = floor(0.8 * total_n))
train_data <- dd[train_index, ]
test_data  <- dd[-train_index, ]
test_data$income_bin <- factor(ifelse(test_data$income == ">50K", "1", "0"), levels = c("0","1"))
train_data$income_bin <- factor(ifelse(train_data$income == ">50K", "1", "0"), levels = c("0","1"))

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

## 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|)

```

```

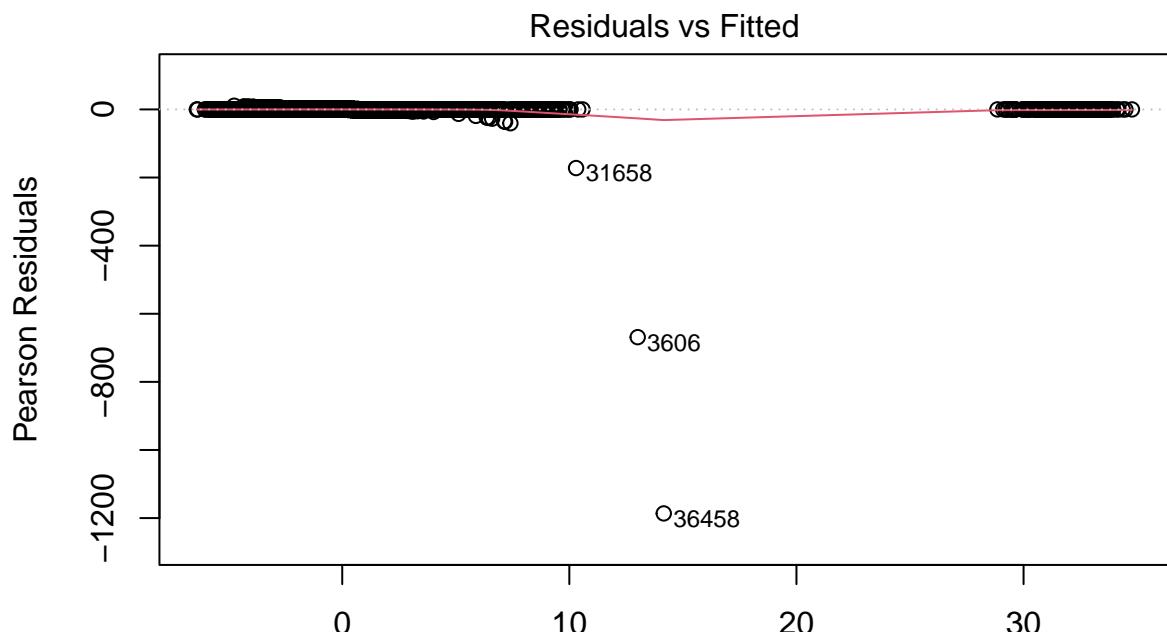
## (Intercept) -8.260e+00 9.371e-02 -88.14 <2e-16 ***
## age         4.220e-02 9.915e-04  42.57 <2e-16 ***
## edu_num     3.223e-01 5.556e-03  58.01 <2e-16 ***
## cap_gain    3.205e-04 7.985e-06  40.14 <2e-16 ***
## cap_loss    6.799e-04 2.634e-05  25.81 <2e-16 ***
## hours_week  4.012e-02 1.070e-03  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

```

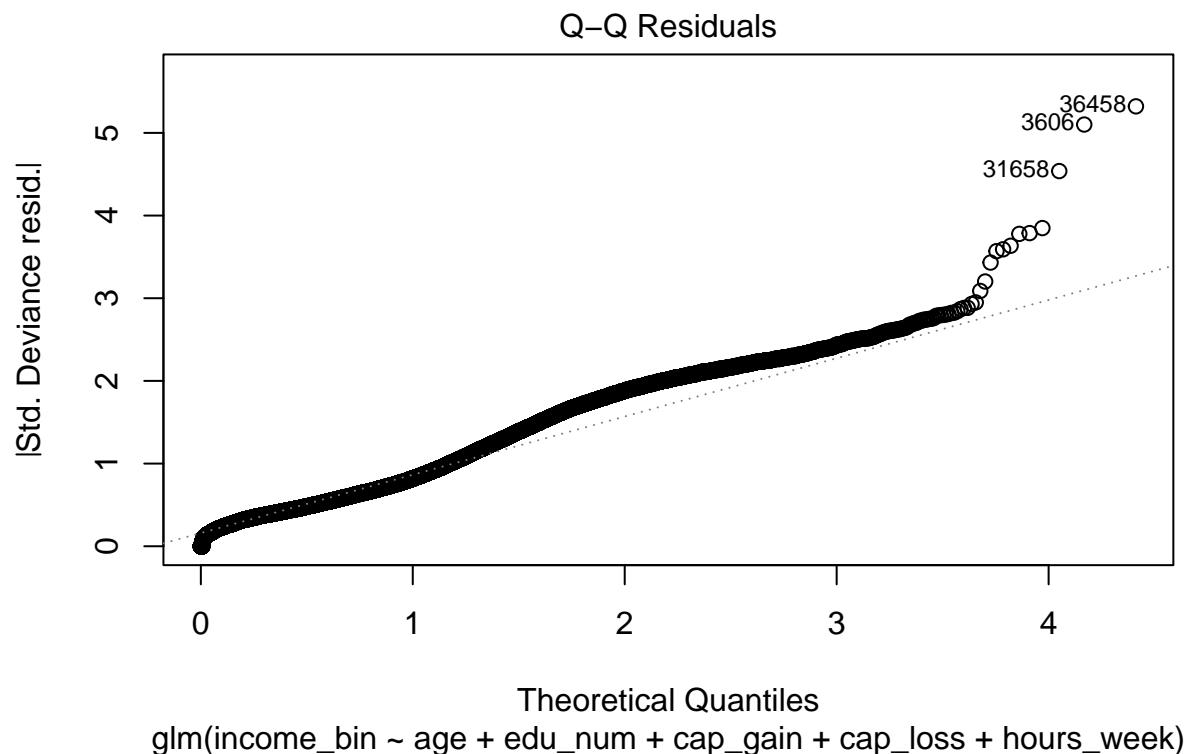
```

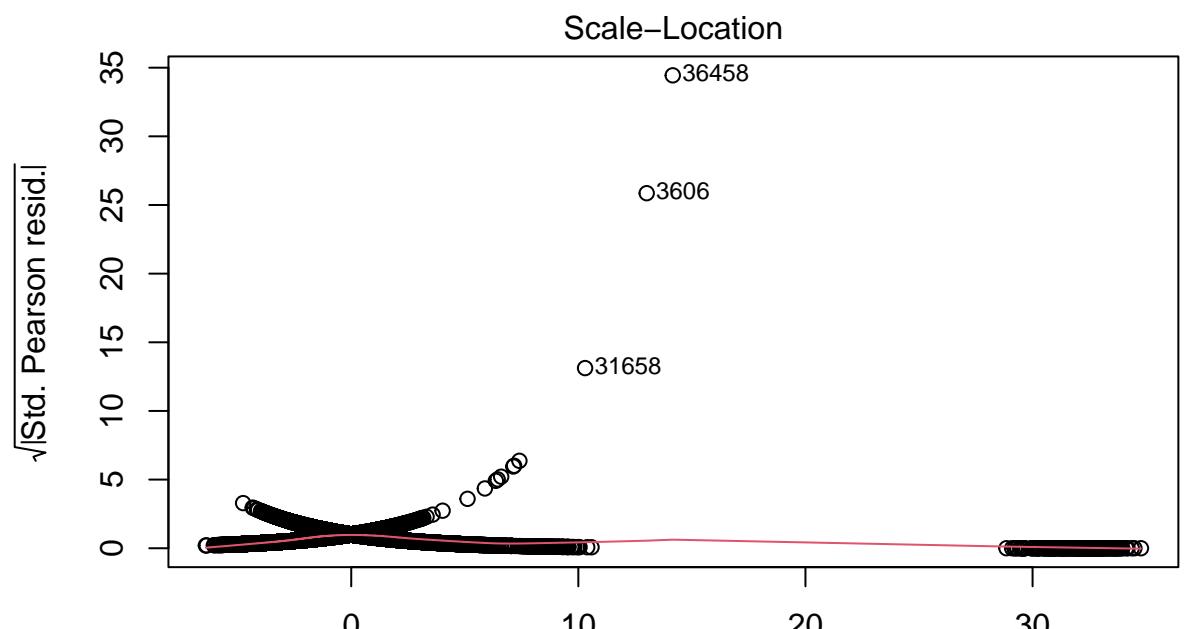
## Check Model Diagnostics
plot(initial_model_b) #the basic hypothesis are met, beware of Homoscedasticity

```



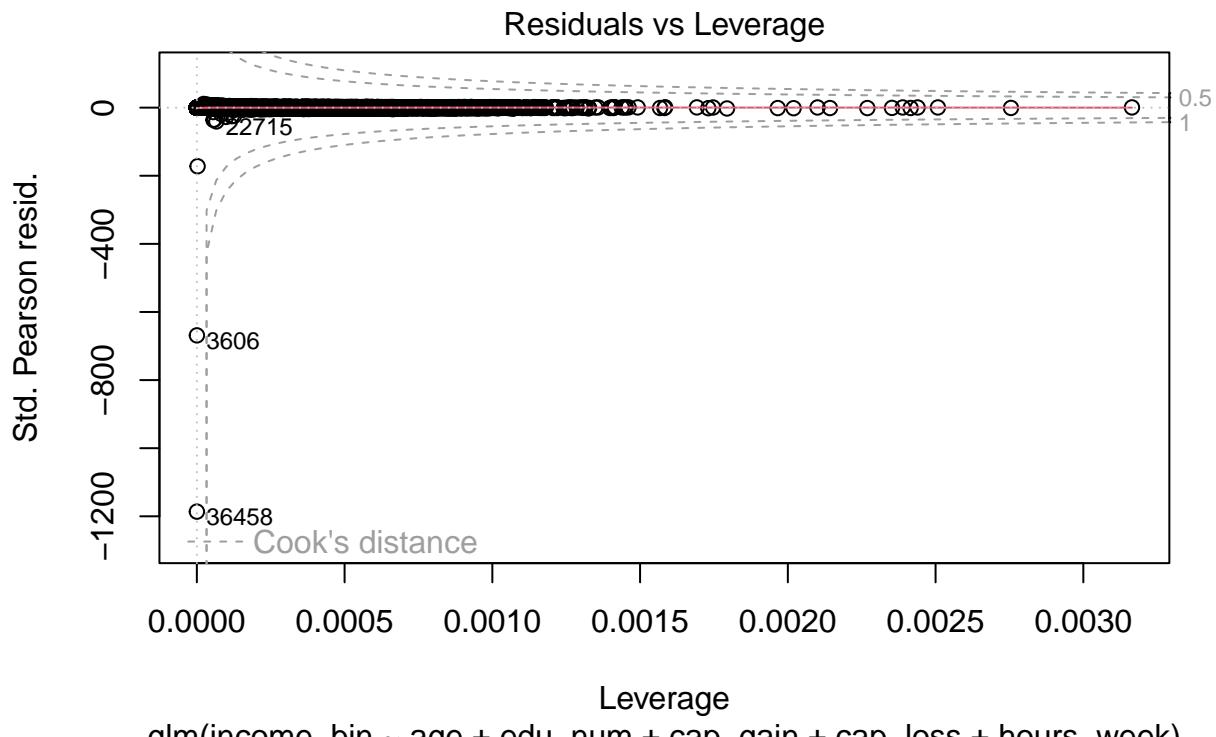
Predicted values
glm(income_bin ~ age + edu_num + cap_gain + cap_loss + hours_week)



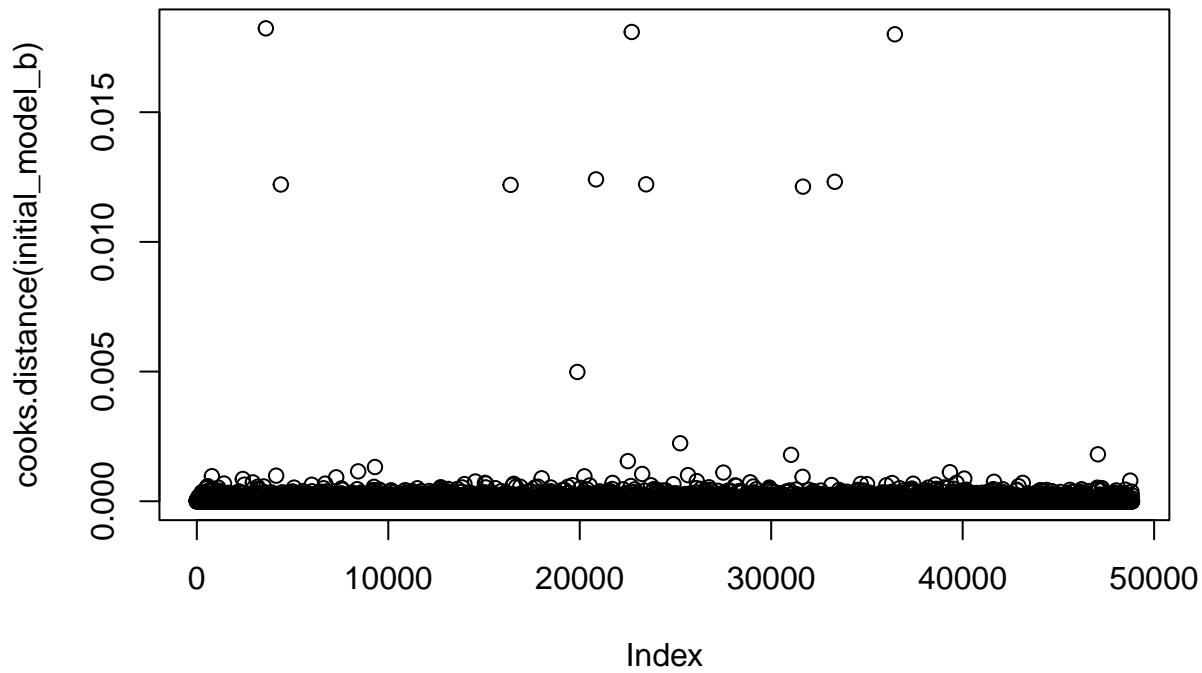


Predicted values

`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_Chisquared
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

##Add_marital_and_test_with_Chisquared
model_marital <- update(model_workclass, . ~ . + marital)

```

```

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

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

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

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

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

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
final_model <- model_country

## 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
## 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
## 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

```

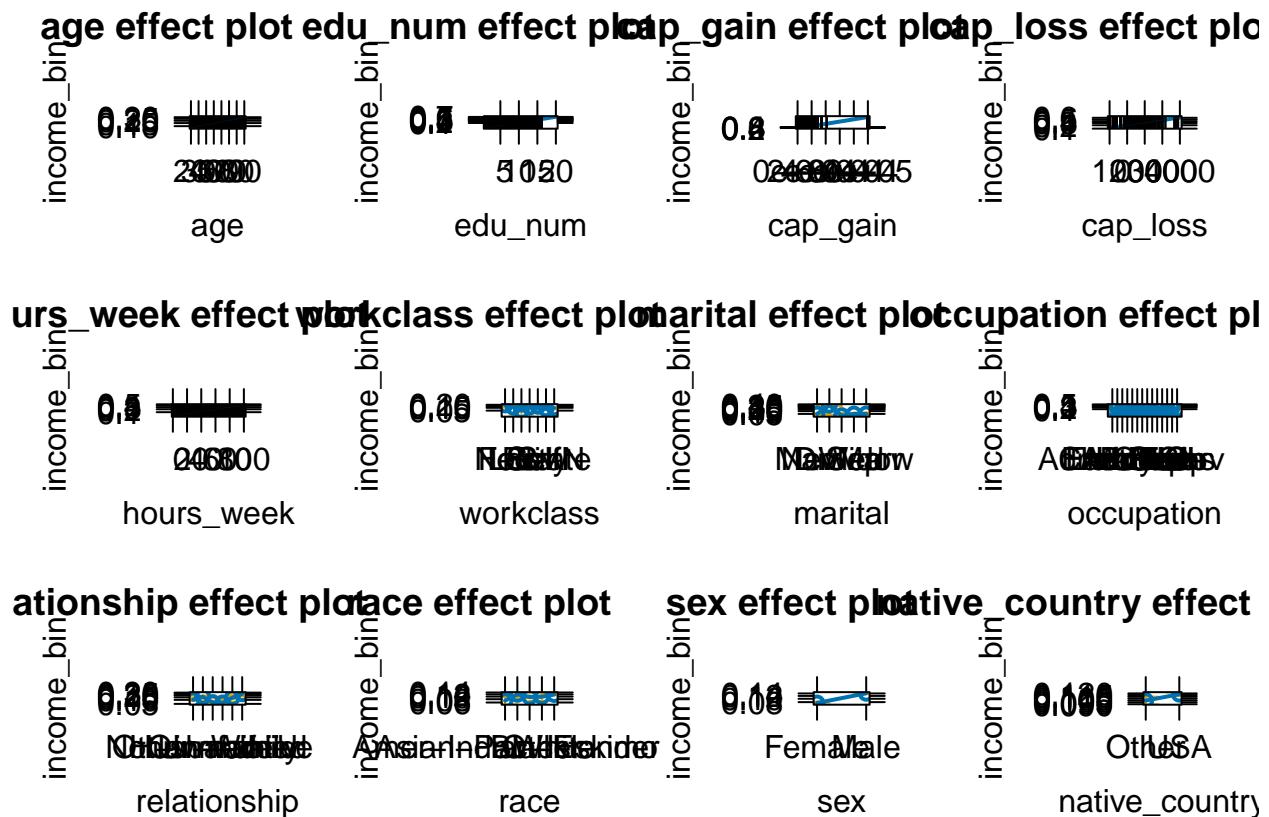
```
AIC(initial_model_b, final_model)
```

```

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

```

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



```

##Final Model Diagnostics and Test-Sample Evaluation
##Predictions on test set (probabilities)
probs_test <- predict(final_model, newdata = test_data, type = "response")

##Convert probabilities to class labels (threshold = 0.5)
preds_test <- factor(ifelse(probs_test > 0.5, "1", "0"), levels=c("0","1"))

##Confusion matrix and derived metrics
conf_tab <- table(Pred=preds_test, True=test_data$income_bin)
TP <- conf_tab["1","1"]; TN <- conf_tab["0","0"]
FP <- conf_tab["1","0"]; FN <- conf_tab["0","1"]
accuracy <- (TP + TN) / sum(conf_tab)
sensitivity <- TP / (TP + FN)
specificity <- TN / (TN + FP)
cat("Accuracy=", round(accuracy,3), " Sensitivity=", round(sensitivity,3),
    " Specificity=", round(specificity,3), "\n")

## Accuracy= 0.855  Sensitivity= 0.597  Specificity= 0.934

##ROC curve and AUC
test_roc <- roc(test_data$income_bin, probs_test)

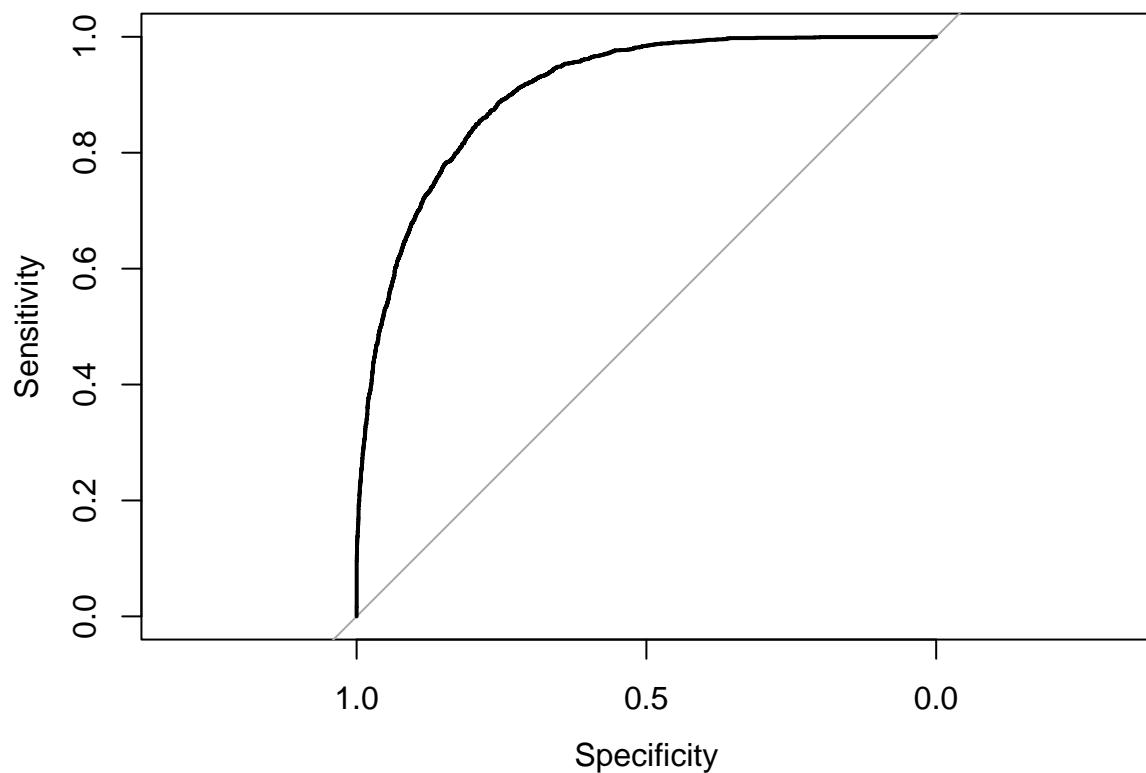
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

plot(test_roc, main="ROC Curve - Test set")

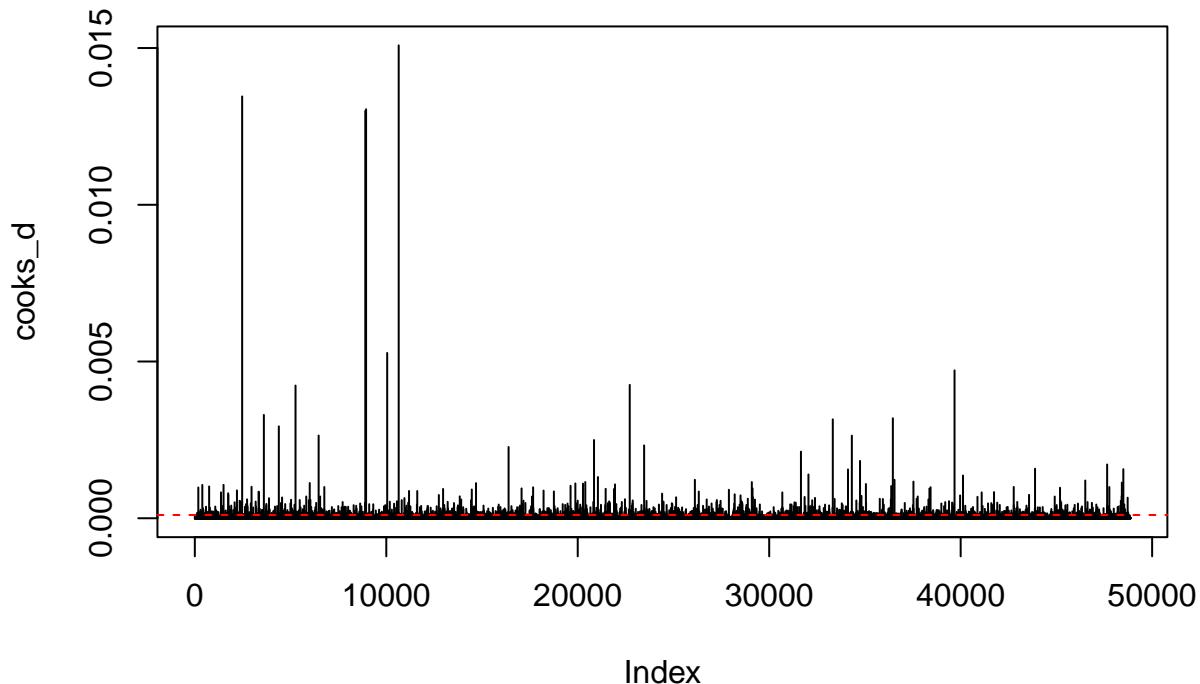
```

ROC Curve – Test set



```
cat("AUC=", round(auc(test_roc),3), "\n")  
  
## AUC= 0.906  
  
##Influential observations (Cook's distance on training data)  
cooks_d <- cooks.distance(final_model)  
p <- length(coef(final_model))  
threshold <- 4 / (nrow(train_data) - p)  
inf_index <- which(cooks_d > threshold)  
cat("Cook's D threshold=", round(threshold,4), " Influential obs indices:", inf_index, "\n")  
  
## Cook's D threshold= 1e-04 Influential obs indices: 89 120 129 183 192 200 237 257 266 282 297 300 311  
  
plot(cooks_d, type="h", main="Cook's Distance for Final Model")  
abline(h=threshold, col="red", lty=2)
```

Cook's Distance for Final Model



```
## Odds ratios and 95% confidence intervals
or_vals <- exp(coef(final_model))
ci_vals <- exp(confint(final_model))

## Waiting for profiling to be done...

## 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
## 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
## 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
```



```

## cap_gain           cap_gain 1.000316e+00 1.000300e+00
## cap_loss           cap_loss 1.000646e+00 1.000587e+00
## hours_week         hours_week 1.031460e+00 1.028872e+00
## workclassLoc       workclassLoc 5.558142e-01 4.651980e-01
## workclassNoPay     workclassNoPay 2.227845e-01 3.367114e-02
## workclassPriv      workclassPriv 5.689082e-01 4.904874e-01
## workclassSelfI     workclassSelfI 6.974690e-01 5.740282e-01
## workclassSelfN     workclassSelfN 3.596217e-01 3.022322e-01
## workclassState      workclassState 4.651813e-01 3.822748e-01
## maritalMarried     maritalMarried 1.012625e+01 6.629108e+00
## maritalNevMarr     maritalNevMarr 6.526277e-01 5.677316e-01
## maritalSep          maritalSep 1.004049e+00 8.051710e-01
## maritalWidow        maritalWidow 1.040055e+00 8.111525e-01
## occupationArmy       occupationArmy 1.667736e+00 3.256233e-01
## occupationCraftRep   occupationCraftRep 1.087513e+00 9.584799e-01
## occupationExecMan    occupationExecMan 2.089855e+00 1.851066e+00
## occupationFarmFish   occupationFarmFish 3.616225e-01 2.879027e-01
## occupationHandlC1    occupationHandlC1 5.219654e-01 4.165146e-01
## occupationHouse      occupationHouse 1.393042e-01 2.286890e-02
## occupationMachOp     occupationMachOp 7.678019e-01 6.530106e-01
## occupationOther       occupationOther 4.194998e-01 3.469070e-01
## occupationProf       occupationProf 1.309883e+00 1.162893e+00
## occupationProtServ   occupationProtServ 1.528116e+00 1.249553e+00
## occupationSales       occupationSales 1.279129e+00 1.122684e+00
## occupationTech        occupationTech 1.703469e+00 1.429878e+00
## occupationTrans       occupationTrans 9.260154e-01 7.911969e-01
## relationshipNot-in-family relationshipNot-in-family 1.779173e+00 1.169519e+00
## relationshipOther-relative relationshipOther-relative 5.883787e-01 3.955321e-01
## relationshipOwn-child  relationshipOwn-child 5.527625e-01 3.645845e-01
## relationshipUnmarried  relationshipUnmarried 1.464372e+00 9.375227e-01
## relationshipWife       relationshipWife 2.989404e+00 2.545543e+00
## raceAsian-Pac-Islander raceAsian-Pac-Islander 1.808387e+00 1.234210e+00
## raceBlack            raceBlack 1.416335e+00 9.918690e-01
## raceOther             raceOther 1.453177e+00 8.623964e-01
## raceWhite            raceWhite 1.774620e+00 1.265170e+00
## sexMale              sexMale 2.004777e+00 1.771307e+00
## native_countryUSA    native_countryUSA 1.205882e+00 1.079277e+00
## CI_high
## (Intercept)          1.777925e-04
## age                  1.025551e+00
## edu_num              1.370759e+00
## cap_gain              1.000333e+00
## cap_loss              1.000705e+00
## hours_week            1.034063e+00
## workclassLoc          6.640413e-01
## workclassNoPay         8.612850e-01
## workclassPriv          6.600606e-01
## workclassSelfI         8.476099e-01
## workclassSelfN          4.278719e-01
## workclassState          5.658140e-01
## maritalMarried         1.535589e+01
## maritalNevMarr          7.503575e-01
## maritalSep              1.244593e+00
## maritalWidow            1.323117e+00

```

```
## occupationArmy          7.149958e+00
## occupationCraftRep      1.234579e+00
## occupationExecMan       2.361061e+00
## occupationFarmFish      4.521330e-01
## occupationHandlCl       6.503350e-01
## occupationHouse         4.865854e-01
## occupationMachOp        9.018544e-01
## occupationOther          5.054139e-01
## occupationProf          1.476340e+00
## occupationProtServ       1.867142e+00
## occupationSales          1.458004e+00
## occupationTech           2.028561e+00
## occupationTrans          1.083284e+00
## relationshipNot-in-family 2.685854e+00
## relationshipOther-relative 8.576652e-01
## relationshipOwn-child    8.251197e-01
## relationshipUnmarried     2.271785e+00
## relationshipWife          3.512334e+00
## raceAsian-Pac-Islander   2.687690e+00
## raceBlack                 2.055852e+00
## raceOther                  2.441012e+00
## raceWhite                  2.534800e+00
## sexMale                    2.270644e+00
## native_countryUSA          1.348540e+00
```