Project: Multivariate Linear Regression and Logistic Regression Application

Jaehwan Han

We have a dataset which contain 400 patients and 11 variables. The purpose of this final project is we want to drive conclusion from Hypothesis 1, Gender has an impact on postemployment after controlling for the pre-employment and other covariant. and Hypothesis 2, Gender has impact on EQOL after controlling for pre-employment, post employment and other covariates. Thus, we want to decide which variable has significant effect on response variable (dependent variable) and try to find out final model by removing insignificant variable from the full model.

1. *Set the Female = 0 and Male = 1.*

project$Gender<- revalue(project$Gender, c("Female"=0)),
project$Gender<- revalue(project$Gender, c("Male"=1))
class(project$Gender) (Checking variable class. It should be factor variable, not character)

2. *Set the 0 = "not married" and 1 = "married"*

project$marital<-as.factor(project$marital) (marital variable also should be **factor** variable, not character)
project$marital<- revalue(project$marital, c("0"="not married", "1"="married"))

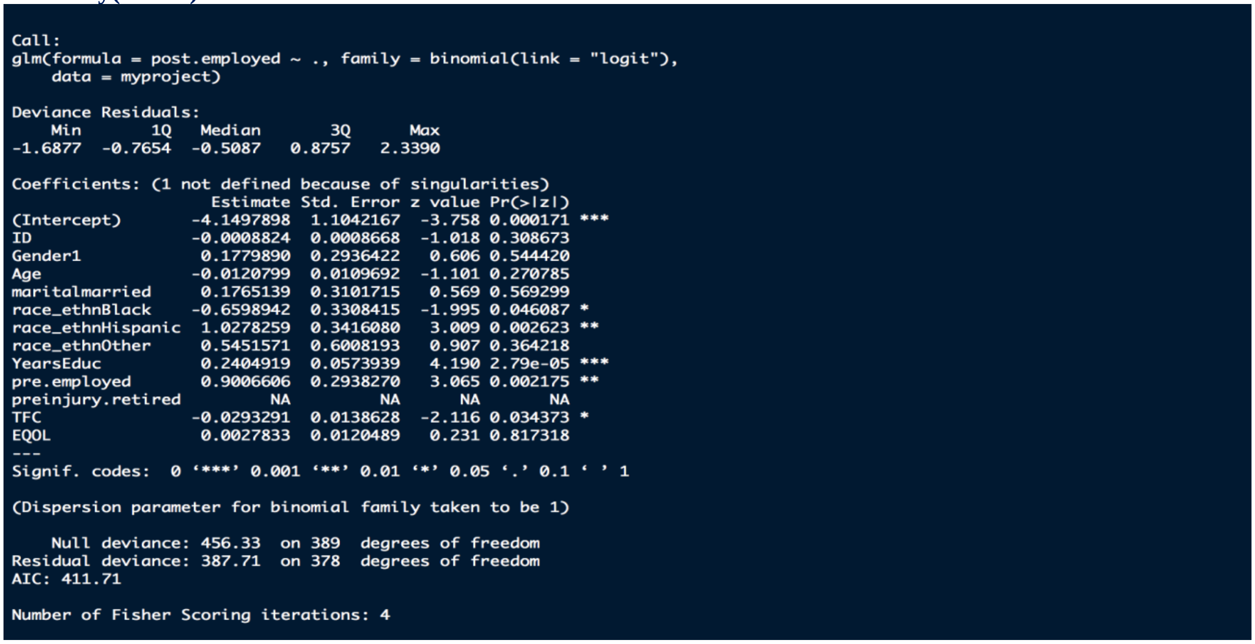
3. *Ethnicity. Use White as reference group.*

project$race_ethn <- relevel(project$race_ethn, ref="White")

4. ***sum(is.na(project))*** *(Use function to check whether dataset has NA value.)*

For Hypothesis 1, I am going to do first is removing the patients with preinjury.retired = 1
myproject <- project[project$preinjury.retired!=1,]
(Create new dataset name, myproject which only contains preinjury.retiered is not =1)

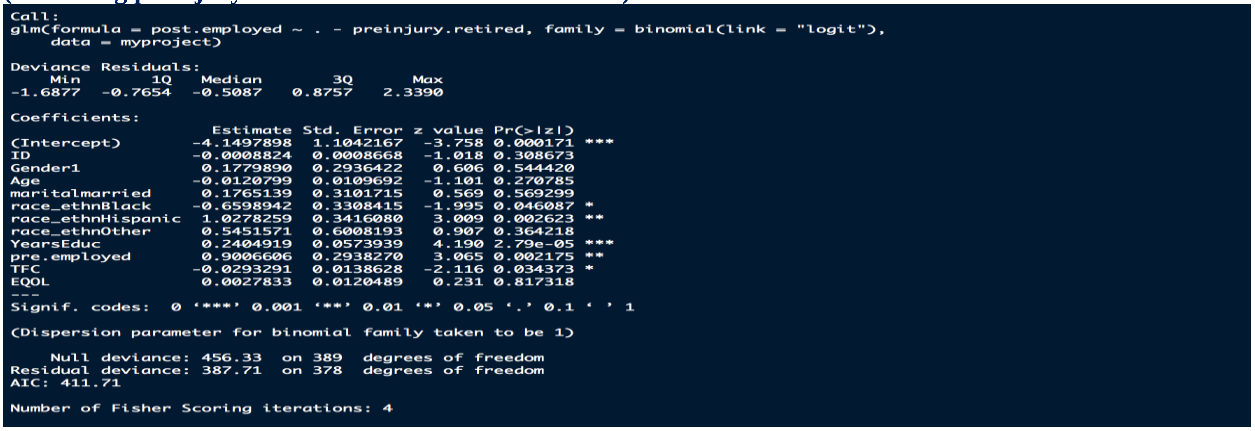
Then, I am going to use logistics model which is outcome (dependent variable) has to be two values (dichotomy). Postemployment is our dependent variable and rest of variables are independent variable.
model <- glm(post.employed ~ .,family=binomial(link='logit'),data=myproject)
summary(model)



```
Call:
glm(formula = post.employed ~ ., family = binomial(link = "logit"),
    data = myproject)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6877  -0.7654  -0.5087   0.8757   2.3390

Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -4.1497898  1.1042167  -3.758 0.000171 ***
ID                -0.0008824  0.0008668  -1.018 0.308673
Gender1            0.1779890  0.2936422   0.606 0.544420
Age               -0.0120799  0.0109692  -1.101 0.270785
maritalmarried     0.1765139  0.3101715   0.569 0.569299
race_ethnBlack    -0.6598942  0.3308415  -1.995 0.046087 *
race_ethnHispanic  1.0278259  0.3416080   3.009 0.002623 **
race_ethnOther     0.5451571  0.6008193   0.907 0.364218
YearsEduc          0.2404919  0.0573939   4.190 2.79e-05 ***
pre.employed       0.9006606  0.2938270   3.065 0.002175 **
preinjury.retired         NA         NA      NA       NA
TFC               -0.0293291  0.0138628  -2.116 0.034373 *
EQOL               0.0027833  0.0120489   0.231 0.817318
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 456.33  on 389  degrees of freedom
Residual deviance: 387.71  on 378  degrees of freedom
AIC: 411.71

Number of Fisher Scoring iterations: 4
```

This is the result of the code I just entered and there is **estimated coefficient, standard error and p-value for each of variable**. The simplest way to determine whether the variable is significant or insignificant is using P-value. If certain variable has p-value which is greater than 0.05, the variable need to be removed from our model. Thus, in order to find final best fit model, I need to remove the largest p-value and then, remove next largest p-value step by step until all variables in our model are significant. Since NA value in preinjury.retired, we should get rid of preinjury.retired variable first.

model_cor1 <- glm(post.employed ~ . - preinjury.retired, family=binomial(link='logit'),
data=myproject)
summary(model_cor1)
(Removing preinjury.retired variable from first model)



```
Call:
glm(formula = post.employed ~ . - preinjury.retired, family = binomial(link = "logit"),
    data = myproject)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6877  -0.7654  -0.5087   0.8757   2.3390

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -4.1497898  1.1042167  -3.758 0.000171 ***
ID                -0.0008824  0.0008668  -1.018 0.308673
Gender1            0.1779890  0.2936422   0.606 0.544420
Age               -0.0120799  0.0109692  -1.101 0.270785
maritalmarried     0.1765139  0.3101715   0.569 0.569299
race_ethnBlack    -0.6598942  0.3308415  -1.995 0.046087 *
race_ethnHispanic  1.0278259  0.3416080   3.009 0.002623 **
race_ethnOther     0.5451571  0.6008193   0.907 0.364218
YearsEduc          0.2404919  0.0573939   4.190 2.79e-05 ***
pre.employed       0.9006606  0.2938270   3.065 0.002175 **
TFC               -0.0293291  0.0138628  -2.116 0.034373 *
EQOL               0.0027833  0.0120489   0.231 0.817318
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 456.33  on 389  degrees of freedom
Residual deviance: 387.71  on 378  degrees of freedom
AIC: 411.71

Number of Fisher Scoring iterations: 4
```

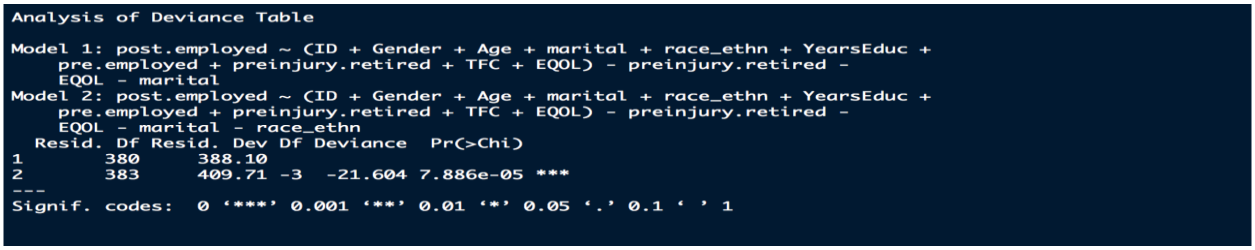
what we need to do next is removing highest p-value as I mentioned above. Variable with largest p-value is insignificant and has little or no effect on response variable, post.employment outcome. Even if the Gender variable has largest p-value, I cannot eliminate the Gender variable because I want to know the effect of gender on post.employment.

Step 1) Remove EQOL which is the highest p-value. Insignificant variable.
Step 2) Next largest insiginificant variable is Marital. The greatest p-value need to be removed.
Step 3) Except for Gender, race_ethnother has next largest p-value. I decided to eliminate race_ehtn variable from model.

model_cor4 <- glm(post.employed ~.-preinjury.retired-EQOL-marital-race_ethn, family=binomial(link='logit'),data=myproject), summary(model_cor4)

After removing race variable, **AIC value is increasing. It indicates selecting wrong variable.**
AIC is one of measure of relative goodness of fit and is useful for comparing model.

anova(model_cor3,model_cor4, test="LRT")
Using ANOVA test to check if RACE variable is useful or not by comparing two model. The likelihood ratio test is highly significant and we would conclude that the variable RACE should remain in the model. P-value is 7.888e-05

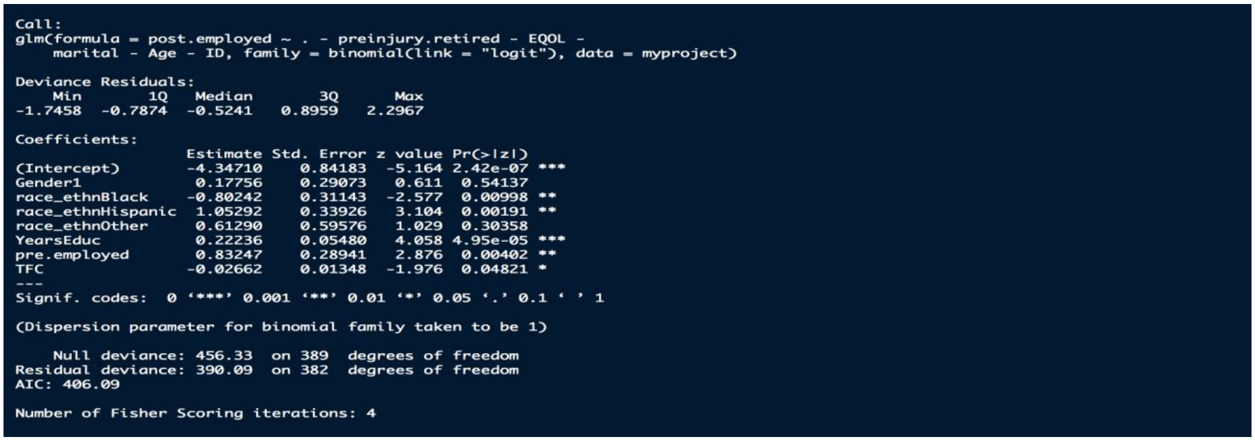


```
Analysis of Deviance Table

Model 1: post.employed ~ (ID + Gender + Age + marital + race_ethn + YearsEduc +
    pre.employed + preinjury.retired + TFC + EQOL) - preinjury.retired -
    EQOL - marital
Model 2: post.employed ~ (ID + Gender + Age + marital + race_ethn + YearsEduc +
    pre.employed + preinjury.retired + TFC + EQOL) - preinjury.retired -
    EQOL - marital - race_ethn
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       380     388.10
2       383     409.71 -3  -21.604 7.886e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Instead of removing race variable, next highest insignificant variable is Age. We decide to remove Age from model.

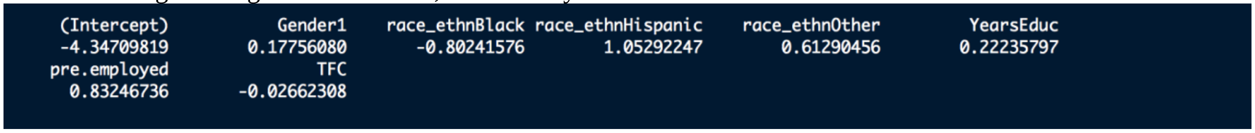
model_cor5 <- glm(post.employed ~.-preinjury.retired-EQOL-marital-Age, family=binomial(link='logit'),data=myproject), summary(model_cor5)
## AIC value is getting decreasing which means selecting right variable. Good decision.
But model_cor5 Still has insignificant variable which is ID, except for race_ehtnOther and Gender.

model_cor6 <- glm(post.employed ~.-preinjury.retired-EQOL-marital-Age-ID,
family=binomial(link='logit'),data=myproject), summary(model_cor6)



```
Call:
glm(formula = post.employed ~ . - preinjury.retired - EQOL -
    marital - Age - ID, family = binomial(link = "logit"), data = myproject)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7458  -0.7874  -0.5241   0.8959   2.2967

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -4.34710    0.84183  -5.164 2.42e-07 ***
Gender1            0.17756    0.29073   0.611  0.54137
race_ethnBlack    -0.80242    0.31143  -2.577  0.00998 **
race_ethnHispanic  1.05292    0.33926   3.104  0.00191 **
race_ethnOther     0.61290    0.59576   1.029  0.30358
YearsEduc          0.22236    0.05480   4.058 4.95e-05 ***
pre.employed       0.83247    0.28941   2.876  0.00402 **
TFC               -0.02662    0.01348  -1.976  0.04821 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 456.33  on 389  degrees of freedom
Residual deviance: 390.09  on 382  degrees of freedom
AIC: 406.09

Number of Fisher Scoring iterations: 4
```

After removing all insignificant variable, I can finally come to a final conclusion.



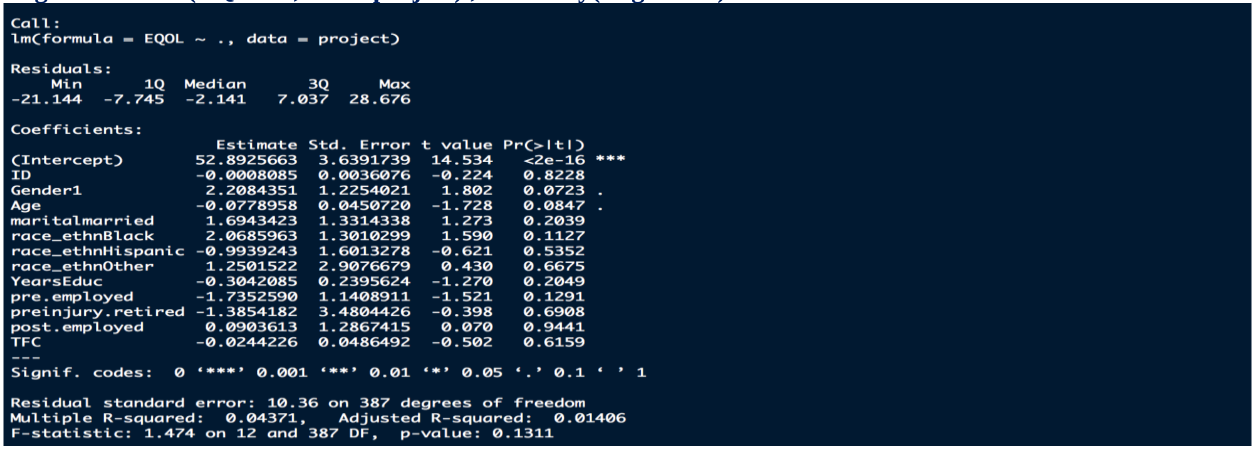
```
  (Intercept)         Gender1    race_ethnBlack race_ethnHispanic    race_ethnOther         YearsEduc
  -4.34709819      0.17756080       -0.80241576        1.05292247        0.61290456        0.22235797
 pre.employed             TFC
   0.83246736     -0.02662308
```

This is the estimated coefficient of each variable in Final Model.
Intercept is -4.347, Gender coefficient is 0.1775. Male= 1, Female=0 and White is reference race variable. Since the white is reference, black person has the least chance of being post-employed due to negative coefficient. Hispanic is the highest chance of being post employed.
Odds of male is e^(-4.347+(0.1775*1)), Odds of female is e^(-4.347+(0.1775*0)). Thus, Male Odds is 0.01546, Female Odds is 0.01295. Odds ratio = odds for male / odds for female, 0.01546/0.01295 = 1.1938.
The odds of post.employed is about 1.2 times greater for male than for female.

ll.null <- model_cor6$null.deviance/-2
ll.proposed <- model_cor6$deviance/-2
(ll.null - ll.proposed) / ll.null

R squared is low only 0.14514. It means model doesn't explain much of variable of the data, only 14.5% of y(post.employed) are explained by dependent variables.

For Hypothesis 2: I will be using all variable in this case. Since the model has 10 independent variables and 1 dependent variable, the regression model is called multivariable regression model.
Unlike Hypothesis 1, EQOL is our dependent variable and rest of variables are independent.
The purpose of this model is to find out effect of each variable on Post-Injury Economics Quality of Life(EQOL).
Here is our Multivariable Regression Model.
RegModel <- lm(EQOL~., data=project) , summary(RegModel)



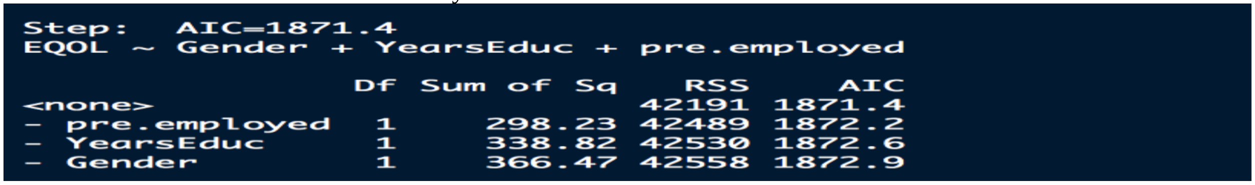
```
Call:
lm(formula = EQOL ~ ., data = project)

Residuals:
    Min      1Q  Median      3Q     Max
-21.144  -7.745  -2.141   7.037  28.676

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       52.8925663  3.6391739  14.534   <2e-16 ***
ID                -0.0008085  0.0036076  -0.224   0.8228
Gender1            2.2084351  1.2254021   1.802   0.0723 .
Age               -0.0778958  0.0450720  -1.728   0.0847 .
maritalmarried     1.6943423  1.3314338   1.273   0.2039
race_ethnBlack     2.0685963  1.3010299   1.590   0.1127
race_ethnHispanic -0.9939243  1.6013278  -0.621   0.5352
race_ethnOther     1.2501522  2.9076679   0.430   0.6675
YearsEduc         -0.3042085  0.2395624  -1.270   0.2049
pre.employed      -1.7352590  1.1408911  -1.521   0.1291
preinjury.retired -1.3854182  3.4804426  -0.398   0.6908
post.employed      0.0903613  1.2867415   0.070   0.9441
TFC               -0.0244226  0.0486492  -0.502   0.6159
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.36 on 387 degrees of freedom
Multiple R-squared:  0.04371,   Adjusted R-squared:  0.01406
F-statistic: 1.474 on 12 and 387 DF,  p-value: 0.1311
```

Look at the p-value of each variable, none of variable are significant and Multiple R Squared value which is statistical measure of how close the data are to fitted regression line is value low, 0.04371. Need to remove insignificant predictors with large P-value. Start removing from post.employed variable.

It is good to remove step by step like I did in first hypothesis test, but the fastest and efficient way is using model selection function. “Backward, Forward and Stepwise” The program automatically finds the best fit model by comparing AIC, R Squared and P-value. Thus I am going to use backward function to find final model.

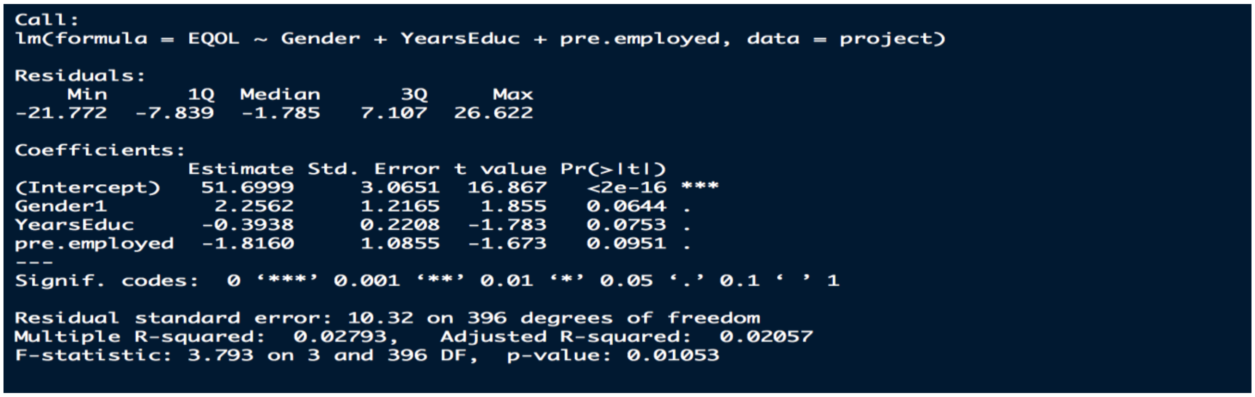
Reg_Final <- stepAIC(RegModel, direction="backward"), summary(Reg_Final)

After you enter the code, program automatically remove one significant term each time and you can see the AIC value is getting decrease. Because AIC is an estimator of the relative quality of statistical model for given set of data. The last model contain predictor variable (Gender, YearsEdu, pre.employed) and response variable(EQOL) is selected as our best fit model from the very first model.



```
Step:  AIC=1871.4
EQOL ~ Gender + YearsEduc + pre.employed

               Df Sum of Sq   RSS    AIC
<none>                      42191 1871.4
- pre.employed  1    298.23 42489 1872.2
- YearsEduc     1    338.82 42530 1872.6
- Gender        1    366.47 42558 1872.9
```

After finding final model, you definitely want to see how significant each of variable on the dependent variable(EQOL). Thus, I run the code to check p-value of each variable based on the result from backward model selection method.

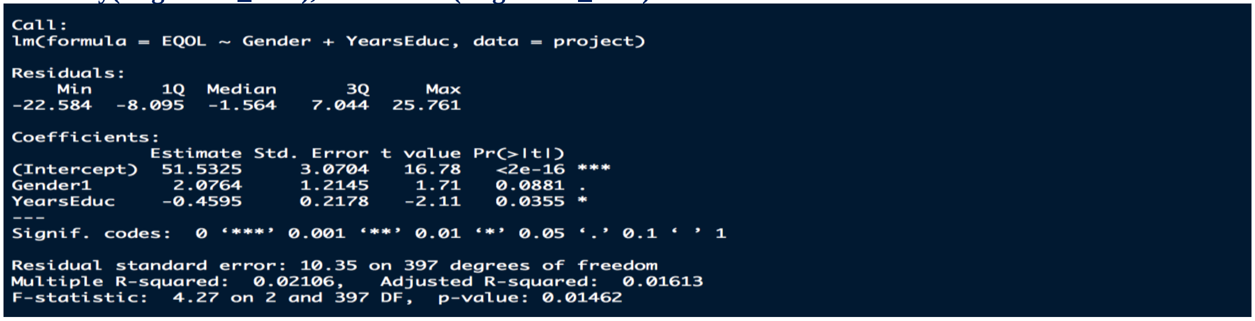


```
Call:
lm(formula = EQOL ~ Gender + YearsEduc + pre.employed, data = project)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.772  -7.839  -1.785   7.107  26.622 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   51.6999     3.0651  16.867   <2e-16 ***
Gender1        2.2562     1.2165   1.855   0.0644 .  
YearsEduc     -0.3938     0.2208  -1.783   0.0753 .  
pre.employed  -1.8160     1.0855  -1.673   0.0951 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.32 on 396 degrees of freedom
Multiple R-squared:  0.02793,	Adjusted R-squared:  0.02057 
F-statistic: 3.793 on 3 and 396 DF,  p-value: 0.01053
```

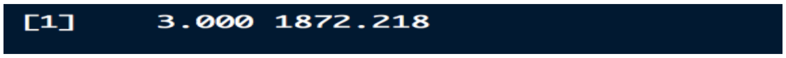
Post injury Economic Quality life is related to Gender and YearsEduc and pre.employed. In the case of male, EQOL increase 2.2562. And the more people are educated the less economics quality of life. Every one year of education increase, EQOL level decrease -0.3938. Person was pre-employed is -1.8160 less confident about their economic situation than the person who was not pre-employed.

Strangely, even if the model above is selected as final model, predictors of the model are not significant. P-value is greater than 0.05. Thus, I decided to remove pre.employed variable.

RegModel_cor8 <-lm(EQOL~Gender+YearsEduc, data=project)
summary(RegModel_cor8), extractAIC(RegModel_cor8)



```
Call:
lm(formula = EQOL ~ Gender + YearsEduc, data = project)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.584  -8.095  -1.564   7.044  25.761 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  51.5325     3.0704   16.78   <2e-16 ***
Gender1       2.0764     1.2145    1.71   0.0881 .  
YearsEduc    -0.4595     0.2178   -2.11   0.0355 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.35 on 397 degrees of freedom
Multiple R-squared:  0.02106,	Adjusted R-squared:  0.01613 
F-statistic:  4.27 on 2 and 397 DF,  p-value: 0.01462
```



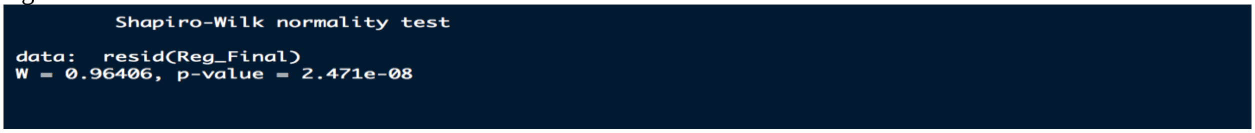
```
[1]      3.000 1872.218
```

<- AIC value of above model (1872.218)

After removing pre.employed variable, YearEduc become significant. But we need to consider find better model, not just making all variable significant. As I mentioned before, Multiple R squared which is statistical measure of how close the data are to fitted regression line and AIC are both effective way to decide better model between two models. Thus, even if YearEduc become significant, R squared value decreasing and AIC value are increasing. It means model contains YearEduc variable is better than not having YearEduc variable model.

Obviously, it would be perfect and reasonable result of best fit model when the model has all significant variables and large R squared value. My final regression model's R squared value is only 0.02793. It means 2.7% variance in dependent variable Y are explained by independent variables present in the model.

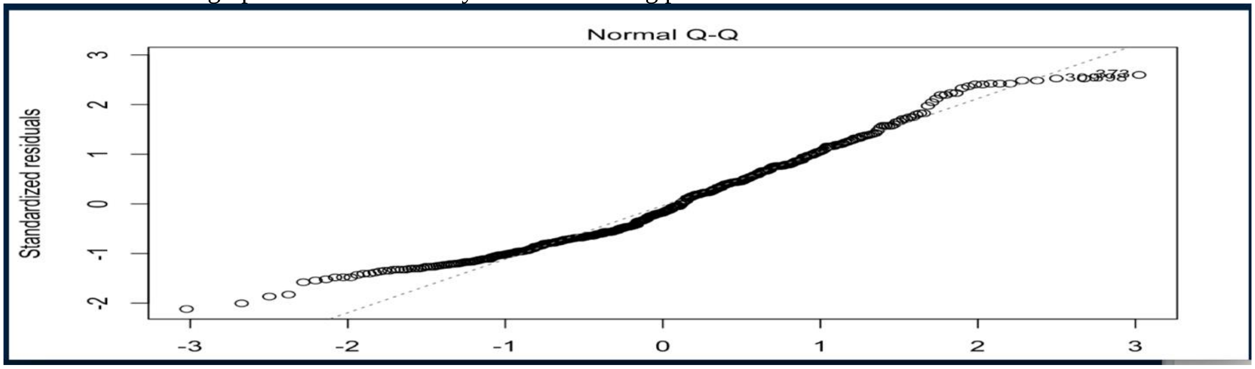
By using Sharpiro.test, I could check whether the model is normal or not since R squared is extremely low and not significant variables.



```
        Shapiro-Wilk normality test

data:  resid(Reg_Final)
W = 0.96406, p-value = 2.471e-08
```

If p-value is greater than 0.05, the data is normal. If it is below 0.05, the data significantly deviate from a normal distribution. Since P-value is 2.471e-0.5, the data is significantly deviate from normal distribution.

I could also utilize graph to check normality of the data using plot function.

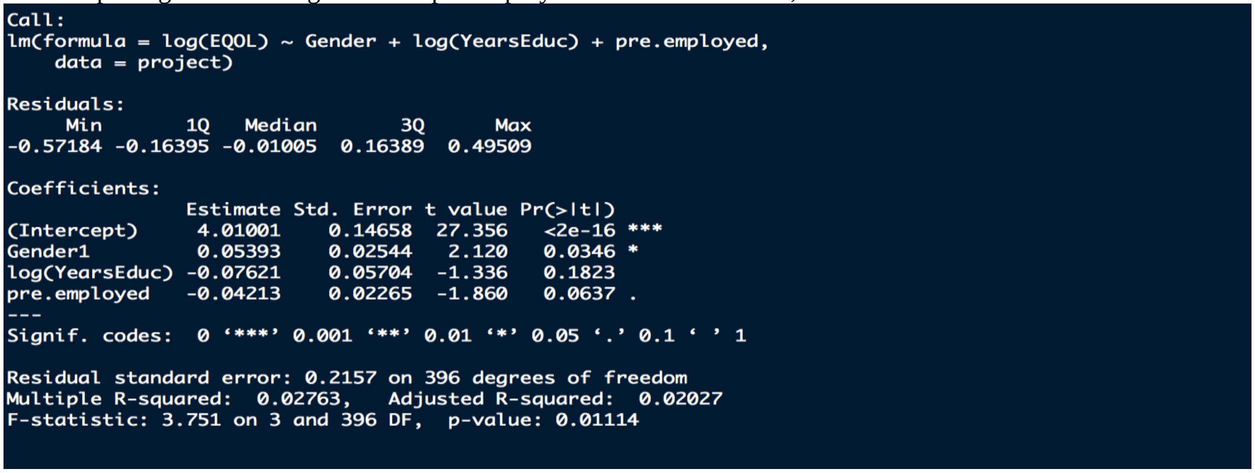


Values are supposed to be on the line, but some of values which is called "outlier" are extremely far from the line. I could say it fail to meet normality assumption because its diverge at the tails.

One of the way to deal with non normal distribution is log transformation.

RegModel_cor10 <-lm(log(EQOL)~Gender+log(YearsEduc)+pre.employed, data=project)

I cannot put log function on gender and pre.employed because those are 0,1 factor variable.



```
Call:
lm(formula = log(EQOL) ~ Gender + log(YearsEduc) + pre.employed,
    data = project)

Residuals:
     Min       1Q   Median       3Q      Max
-0.57184 -0.16395 -0.01005  0.16389  0.49509

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.01001    0.14658  27.356   <2e-16 ***
Gender1         0.05393    0.02544   2.120   0.0346 *
log(YearsEduc) -0.07621    0.05704  -1.336   0.1823
pre.employed   -0.04213    0.02265  -1.860   0.0637 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2157 on 396 degrees of freedom
Multiple R-squared:  0.02763,	Adjusted R-squared:  0.02027
F-statistic: 3.751 on 3 and 396 DF,  p-value: 0.01114
```

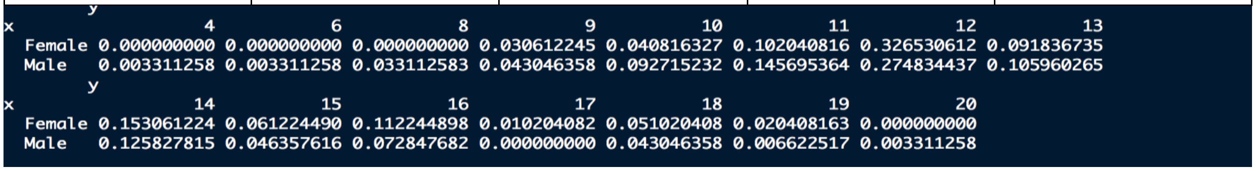
Even if I use log transformation, it looks like this transformation remedy does not work. R squared value is still low. Thus, based on the result I find out, I could conclude that dataset is not normal distributed and the final model seems to be not good enough to explain Y(EQOL) variable well.

Summary Table

Continuous Variable	Female (n=127)	Male (n=377)	Test Used	P-value
Age	38.71429	38.59272	Two sample or wilcoxon	0.9376(Compute independent t-test)
TFC	7.784138	8,478787		0.5839
EQOL	45.44214	47.87464		0.0447
Discrete Variable	Female (Percentage)	Male(Percentage)	Test Used	P-value

Race / Ethnicity White	0.4693	0.4238	Chi square test or Fisher exact	0.544, df=3
				df <-data.frame (x=project$Gender, y=project$race_ethn)
				mytab <-with(df,table(x,y))
				chisq.test(mytab)

Black	0.3163	0.3973		
Hispanic	0.1734	0.1456		
Other	0.04081	0.03311		
Year Educ	4 year (0.0000) 6 year (0.00000) 8 year(0.00000)	4 year(0.003311) 6 year (0.003311) 8 year (0.03311)	Chi square test or Fisher exact	0.3386



```
          y
x                  4           6           8           9          10          11          12          13
  Female 0.000000000 0.000000000 0.000000000 0.030612245 0.040816327 0.102040816 0.326530612 0.091836735
  Male   0.003311258 0.003311258 0.033112583 0.043046358 0.092715232 0.145695364 0.274834437 0.105960265
          y
x                 14          15          16          17          18          19          20
  Female 0.153061224 0.061224490 0.112244898 0.010204082 0.051020408 0.020408163 0.000000000
  Male   0.125827815 0.046357616 0.072847682 0.000000000 0.043046358 0.006622517 0.003311258
```