Linear Final

Bowei Li

2023-05-02

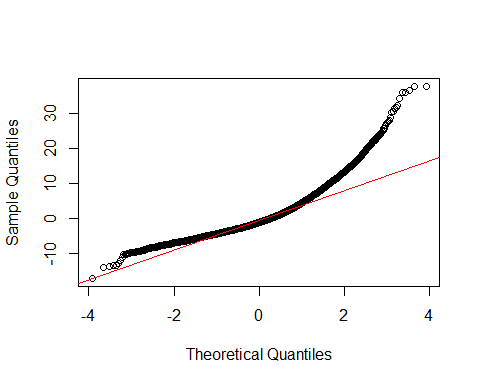
#import data and load packages  
library(broom)  
library(car)  
library(tidyverse)  
library(tidycomm)  
library(MASS)  
  
data <- read.csv("PATH\_W1.csv")  
  
# change type of data  
cate<-colnames(data)[-c(1,2,11,12,17)]  
for(x in cate){  
 data[,x]<-as.factor(data[,x])  
}

model 1

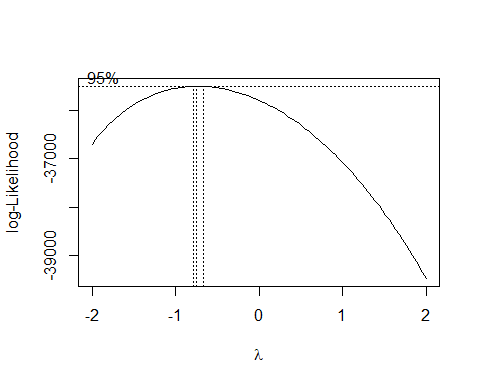
#run analysis with categorical variables, people who never used tobacco, Male,Non-Hispanic White and age 12-14 would be the references groups.  
model\_1<- lm(BMI~ Evertob+Old+relevel(Male,ref='1')+relevel(Race,ref='2'),data=data)  
summary(model\_1)

##   
## Call:  
## lm(formula = BMI ~ Evertob + Old + relevel(Male, ref = "1") +   
## relevel(Race, ref = "2"), data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.028 -3.402 -1.118 2.288 37.688   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 21.03558 0.09481 221.867 < 2e-16 \*\*\*  
## Evertob1 0.87100 0.13429 6.486 9.18e-11 \*\*\*  
## Old1 1.70269 0.09754 17.457 < 2e-16 \*\*\*  
## relevel(Male, ref = "1")0 0.03005 0.09532 0.315 0.75258   
## relevel(Race, ref = "2")1 1.26660 0.11309 11.200 < 2e-16 \*\*\*  
## relevel(Race, ref = "2")3 1.27647 0.14778 8.637 < 2e-16 \*\*\*  
## relevel(Race, ref = "2")4 0.47150 0.17152 2.749 0.00599 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.113 on 11522 degrees of freedom  
## Multiple R-squared: 0.04565, Adjusted R-squared: 0.04515   
## F-statistic: 91.86 on 6 and 11522 DF, p-value: < 2.2e-16

#see if residuals of the model follows normal distribution  
qqnorm(residuals(model\_1), main = "")  
qqline(residuals(model\_1), col = "red")



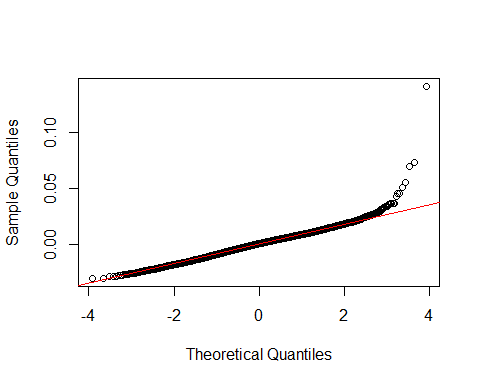
#using box-cox to transformation the model  
obj <- boxcox(model\_1, plotit = TRUE)



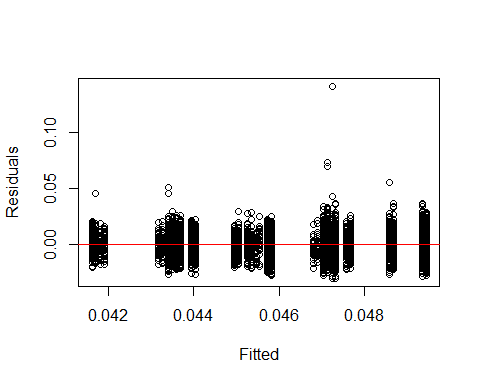
obj$x[which.max(obj$y)] # calculate lambda ~-1 -> transform x to 1/x

## [1] -0.7474747

data[,"new\_BMI"]<-1/data[,"BMI"]  
new\_model\_1<-lm(new\_BMI~ Evertob+Old+relevel(Male,ref='1')+relevel(Race,ref='2'),data=data)  
#see if now residuals follow normal distribution  
qqnorm(residuals(new\_model\_1), main = "")  
qqline(residuals(new\_model\_1), col = "red")



#see if there is equal variance  
plot(fitted(new\_model\_1), residuals(new\_model\_1), xlab = "Fitted",  
ylab = "Residuals")  
abline (h=0, col = "red")



summary(new\_model\_1)

##   
## Call:  
## lm(formula = new\_BMI ~ Evertob + Old + relevel(Male, ref = "1") +   
## relevel(Race, ref = "2"), data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.030662 -0.005975 0.000353 0.005915 0.140955   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.945e-02 1.712e-04 288.761 < 2e-16 \*\*\*  
## Evertob1 -1.782e-03 2.425e-04 -7.347 2.17e-13 \*\*\*  
## Old1 -3.634e-03 1.762e-04 -20.631 < 2e-16 \*\*\*  
## relevel(Male, ref = "1")0 -9.106e-05 1.722e-04 -0.529 0.5969   
## relevel(Race, ref = "2")1 -2.324e-03 2.043e-04 -11.376 < 2e-16 \*\*\*  
## relevel(Race, ref = "2")3 -2.117e-03 2.669e-04 -7.933 2.34e-15 \*\*\*  
## relevel(Race, ref = "2")4 -7.646e-04 3.098e-04 -2.468 0.0136 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.009234 on 11522 degrees of freedom  
## Multiple R-squared: 0.05727, Adjusted R-squared: 0.05678   
## F-statistic: 116.7 on 6 and 11522 DF, p-value: < 2.2e-16

model 2

#run analysis to see the interaction between ever tobacco use and gender  
model\_2a<- lm(new\_BMI~Evertob \* relevel(Male, ref = "1") +  
 relevel(Race, ref = "2")+Old,data=data)  
summary(model\_2a)

##   
## Call:  
## lm(formula = new\_BMI ~ Evertob \* relevel(Male, ref = "1") + relevel(Race,   
## ref = "2") + Old, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.030697 -0.005967 0.000346 0.005932 0.140985   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.0494791 0.0001752 282.390 < 2e-16 \*\*\*  
## Evertob1 -0.0019755 0.0003277 -6.029 1.70e-09 \*\*\*  
## relevel(Male, ref = "1")0 -0.0001560 0.0001874 -0.833 0.4050   
## relevel(Race, ref = "2")1 -0.0023248 0.0002043 -11.381 < 2e-16 \*\*\*  
## relevel(Race, ref = "2")3 -0.0021155 0.0002669 -7.926 2.48e-15 \*\*\*  
## relevel(Race, ref = "2")4 -0.0007665 0.0003098 -2.474 0.0134 \*   
## Old1 -0.0036349 0.0001762 -20.634 < 2e-16 \*\*\*  
## Evertob1:relevel(Male, ref = "1")0 0.0004171 0.0004745 0.879 0.3794   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.009234 on 11521 degrees of freedom  
## Multiple R-squared: 0.05733, Adjusted R-squared: 0.05676   
## F-statistic: 100.1 on 7 and 11521 DF, p-value: < 2.2e-16

#run analysis to see the interaction between ever tobacco use and age  
model\_2b<- lm(new\_BMI~Evertob\*Old + relevel(Male, ref = "1") +  
 relevel(Race, ref = "2"),data=data)  
summary(model\_2b)

##   
## Call:  
## lm(formula = new\_BMI ~ Evertob \* Old + relevel(Male, ref = "1") +   
## relevel(Race, ref = "2"), data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.030755 -0.005962 0.000336 0.005929 0.140864   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.953e-02 1.736e-04 285.245 < 2e-16 \*\*\*  
## Evertob1 -2.721e-03 4.225e-04 -6.441 1.23e-10 \*\*\*  
## Old1 -3.823e-03 1.894e-04 -20.190 < 2e-16 \*\*\*  
## relevel(Male, ref = "1")0 -9.345e-05 1.721e-04 -0.543 0.58716   
## relevel(Race, ref = "2")1 -2.323e-03 2.042e-04 -11.377 < 2e-16 \*\*\*  
## relevel(Race, ref = "2")3 -2.103e-03 2.669e-04 -7.881 3.54e-15 \*\*\*  
## relevel(Race, ref = "2")4 -7.534e-04 3.097e-04 -2.432 0.01501 \*   
## Evertob1:Old1 1.399e-03 5.151e-04 2.716 0.00663 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.009232 on 11521 degrees of freedom  
## Multiple R-squared: 0.05787, Adjusted R-squared: 0.0573   
## F-statistic: 101.1 on 7 and 11521 DF, p-value: < 2.2e-16

#run analysis to see the how other three variables Mental Health Internalizing symptom, Mental Health Externalizing symptom, and Grade would have impact on the model 2b which is the best model among the previous three models.  
model\_3<- lm(new\_BMI~Evertob\*Old + relevel(Male, ref = "1") +  
 relevel(Race, ref = "2")+ MHint + MHext+ Grade ,data=data)  
summary(model\_3)

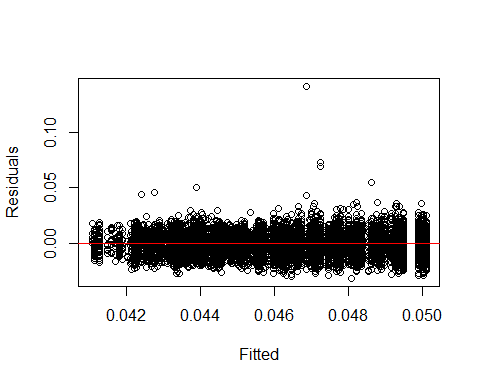
##   
## Call:  
## lm(formula = new\_BMI ~ Evertob \* Old + relevel(Male, ref = "1") +   
## relevel(Race, ref = "2") + MHint + MHext + Grade, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.031385 -0.006029 0.000343 0.005912 0.141319   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.890e-02 2.514e-04 194.473 < 2e-16 \*\*\*  
## Evertob1 -2.382e-03 4.257e-04 -5.595 2.26e-08 \*\*\*  
## Old1 -3.776e-03 1.893e-04 -19.944 < 2e-16 \*\*\*  
## relevel(Male, ref = "1")0 -1.250e-04 1.769e-04 -0.707 0.47975   
## relevel(Race, ref = "2")1 -2.235e-03 2.047e-04 -10.918 < 2e-16 \*\*\*  
## relevel(Race, ref = "2")3 -1.897e-03 2.692e-04 -7.047 1.94e-12 \*\*\*  
## relevel(Race, ref = "2")4 -7.320e-04 3.092e-04 -2.368 0.01791 \*   
## MHint1 -6.229e-04 1.955e-04 -3.187 0.00144 \*\*   
## MHext1 7.344e-05 1.994e-04 0.368 0.71270   
## Grade1 1.132e-03 1.941e-04 5.834 5.56e-09 \*\*\*  
## Evertob1:Old1 1.318e-03 5.148e-04 2.560 0.01048 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.009214 on 11518 degrees of freedom  
## Multiple R-squared: 0.06167, Adjusted R-squared: 0.06085   
## F-statistic: 75.7 on 10 and 11518 DF, p-value: < 2.2e-16

rs<-summary(model\_3)  
#check vif of model\_3  
vif(model\_3)

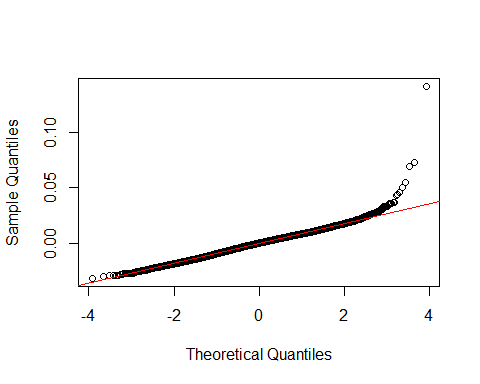
## there are higher-order terms (interactions) in this model  
## consider setting type = 'predictor'; see ?vif

## GVIF Df GVIF^(1/(2\*Df))  
## Evertob 3.252938 1 1.803590  
## Old 1.212839 1 1.101290  
## relevel(Male, ref = "1") 1.061623 1 1.030351  
## relevel(Race, ref = "2") 1.035157 3 1.005775  
## MHint 1.296822 1 1.138781  
## MHext 1.264377 1 1.124445  
## Grade 1.070862 1 1.034825  
## Evertob:Old 3.561977 1 1.887320

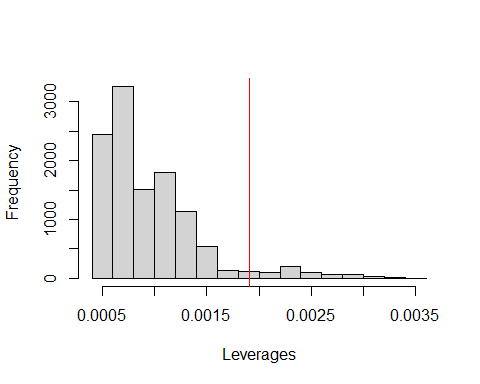
#use residual-fitted plot to check if the model's residual follows normal distribution  
plot(fitted(model\_3), residuals(model\_3), xlab = "Fitted",  
ylab = "Residuals")  
abline (h=0, col = "red")



qqnorm(residuals(model\_3), main = "")  
qqline(residuals(model\_3), col = "red")



#check leverage points in model\_3  
hatv<-hatvalues(model\_3)  
hist (hatv, xlab = "Leverages", main = "")  
abline(v = 2 \* mean(hatv), col="red") #found leverage points to the right of the red line



#data table for each variable  
data\_table<- data[,c("BMI","Evertob","Old","Male","MHint","MHext","Race","Grade")]  
summary(data\_table)

## BMI Evertob Old Male MHint MHext Race   
## Min. : 5.314 0:9722 0:6097 0:5640 0:5695 0:4311 1:3172   
## 1st Qu.:19.020 1:1807 1:5432 1:5889 1:5834 1:7218 2:5788   
## Median :21.454 3:1518   
## Mean :22.549 4:1051   
## 3rd Qu.:24.800   
## Max. :60.000   
## Grade   
## 0:3440   
## 1:8089   
##   
##   
##   
##

# continuous dep variabe with dichotomous ind dep, table 1  
data\_table %>% t\_test(Evertob)

## # A tibble: 1 × 10  
## Variable M\_0 SD\_0 M\_1 SD\_1 Delta\_M t df p d  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 BMI 22.4 5.21 23.6 5.25 -1.24 -9.27 11527 2.16e-20 -0.238

# continuous dep variabe with polytomous ind dep, one way anova post hoc test to see differences in BMI between ethnic groups.Table 3  
data\_table %>%   
 unianova(Race,post\_hoc=TRUE)%>%  
 dplyr::select(Var, post\_hoc) %>%   
 tidyr::unnest(post\_hoc)

## # A tibble: 6 × 8  
## Var term contrast null.value estimate conf.low conf.high adj.p.value  
## <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 BMI Race 2-1 0 -1.18 -1.48 -0.887 0   
## 2 BMI Race 3-1 0 0.00805 -0.409 0.425 1.00   
## 3 BMI Race 4-1 0 -0.786 -1.26 -0.310 0.000129  
## 4 BMI Race 3-2 0 1.19 0.805 1.58 0   
## 5 BMI Race 4-2 0 0.397 -0.0517 0.845 0.104   
## 6 BMI Race 4-3 0 -0.794 -1.33 -0.258 0.000822

# perform t-tests for each specific variables  
t.test(data$BMI ~ data$Old)

##   
## Welch Two Sample t-test  
##   
## data: data$BMI by data$Old  
## t = -18.75, df = 11182, p-value < 2.2e-16  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -1.998427 -1.620126  
## sample estimates:  
## mean in group 0 mean in group 1   
## 21.69611 23.50538

t.test(data$BMI ~ data$MHint)

##   
## Welch Two Sample t-test  
##   
## data: data$BMI by data$MHint  
## t = -4.9803, df = 11513, p-value = 6.442e-07  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -0.6752953 -0.2938511  
## sample estimates:  
## mean in group 0 mean in group 1   
## 22.30336 22.78793

t.test(data$BMI ~ data$MHext)

##   
## Welch Two Sample t-test  
##   
## data: data$BMI by data$MHext  
## t = -1.8366, df = 9004.2, p-value = 0.0663  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -0.38317485 0.01247234  
## sample estimates:  
## mean in group 0 mean in group 1   
## 22.43252 22.61787

t.test(data$BMI ~ data$Male)

##   
## Welch Two Sample t-test  
##   
## data: data$BMI by data$Male  
## t = 0.30239, df = 11490, p-value = 0.7624  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -0.1616606 0.2206354  
## sample estimates:  
## mean in group 0 mean in group 1   
## 22.56363 22.53414

t.test(data$BMI ~ data$Evertob)

##   
## Welch Two Sample t-test  
##   
## data: data$BMI by data$Evertob  
## t = -9.2212, df = 2511.6, p-value < 2.2e-16  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -1.5015245 -0.9749056  
## sample estimates:  
## mean in group 0 mean in group 1   
## 22.35449 23.59271

t.test(data$BMI ~ data$Grade)

##   
## Welch Two Sample t-test  
##   
## data: data$BMI by data$Grade  
## t = 9.3108, df = 5809.9, p-value < 2.2e-16  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## 0.8206415 1.2583767  
## sample estimates:  
## mean in group 0 mean in group 1   
## 23.27791 22.23840

#calculate ANOVA for the four models  
anova(new\_model\_1,model\_2a,model\_2b,model\_3)

## Analysis of Variance Table  
##   
## Model 1: new\_BMI ~ Evertob + Old + relevel(Male, ref = "1") + relevel(Race,   
## ref = "2")  
## Model 2: new\_BMI ~ Evertob \* relevel(Male, ref = "1") + relevel(Race,   
## ref = "2") + Old  
## Model 3: new\_BMI ~ Evertob \* Old + relevel(Male, ref = "1") + relevel(Race,   
## ref = "2")  
## Model 4: new\_BMI ~ Evertob \* Old + relevel(Male, ref = "1") + relevel(Race,   
## ref = "2") + MHint + MHext + Grade  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 11522 0.98248   
## 2 11521 0.98242 1 0.0000659 0.776 0.3784   
## 3 11521 0.98185 0 0.0005626   
## 4 11518 0.97790 3 0.0039541 15.524 4.483e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# calculate AIC and BIC for each model  
model\_1\_AIC <- AIC(new\_model\_1)  
model\_1\_BIC <- BIC(new\_model\_1)  
  
model\_2a\_AIC <- AIC(model\_2a)  
model\_2a\_BIC <- BIC(model\_2a)  
  
model\_2b\_AIC <- AIC(model\_2b)  
model\_2b\_BIC <- BIC(model\_2b)  
  
model\_3\_AIC <- AIC(model\_3)  
model\_3\_BIC <- BIC(model\_3)  
  
#Present the result in a data frame  
results\_df <- data.frame(  
 Model = c("Model 1", "Model 2a", "Model 2b", "Model 3"),  
 AIC = c(model\_1\_AIC, model\_2a\_AIC, model\_2b\_AIC, model\_3\_AIC),  
 BIC = c(model\_1\_BIC, model\_2b\_BIC, model\_2b\_BIC, model\_3\_BIC)  
)  
results\_df

## Model AIC BIC  
## 1 Model 1 -75296.24 -75237.42  
## 2 Model 2a -75295.01 -75235.44  
## 3 Model 2b -75301.62 -75235.44  
## 4 Model 3 -75342.14 -75253.91