Group_Project

Yueyang Zhang

2019??12??4??

```
# Loading packages needed in following steps
library("tidyverse")
library(haven)
library(dplyr)
library(tidyr)
library(ResourceSelection)
library(ggplot2)
library(foreign)#
library(nnet)#
library(ggplot2)
library(reshape2)
library(lmerTest)
library(car)
library(nlme)

multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
  ## A function used to plot several plots on the same page.
  ## found this func from internet
  ## input: ggplot item
  ## output: just plot

  require(grid)

  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)

  numPlots = length(plots)

  # If layout is NULL, then use 'cols' to determine layout
  if (is.null(layout)) {
    # Make the panel
    # ncol: Number of columns of plots
    # nrow: Number of rows needed, calculated from # of cols
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                    ncol = cols, nrow = ceiling(numPlots/cols))
  }

  if (numPlots==1) {
    print(plots[[1]])

  } else {
    # Set up the page
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

    # Make each plot, in the correct location
    for (i in 1:numPlots) {
      # Get the i,j matrix positions of the regions that contain this subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                      layout.pos.col = matchidx$col))
    }
  }
}
```

Here need to states how we deal with our data.(important)

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
detach("package:MASS", unload=TRUE)
```

```
## Warning: 'MASS' namespace cannot be unloaded:
##   namespace 'MASS' is imported by 'lmerTest', 'lme4' so cannot be unloaded
```

```
# Load data and select variables we need and drop NA
X<-read_xpt("https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/DEMO_D.XPT")
X_variable<-X%>%select("SEQN","RIAGENDR","RIDAGEYR","DMDEDUC2","RIDRETH1")%>%
  drop_na()%>%
  filter(RIDAGEYR>=20,DMDEDUC2!=7,DMDEDUC2!=9)%>%
  mutate(RIAGENDR=as.numeric(RIAGENDR==1))%>%
  transmute(SEQN,gender=RIAGENDR,age=RIDAGEYR,race=RIDRETH1,education=DMDEDUC2)

health_insurance<-read_xpt("https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/HIQ_D.XPT")
health_insurance<-health_insurance%>%select(SEQN,HIQ011)%>%
  drop_na()%>%
  filter(HIQ011!=7,HIQ011!=9)%>%
  mutate(insurance=as.numeric(HIQ011==1))%>%
  select(SEQN,insurance)

Smoking<-read_xpt("https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/SMQ_D.XPT")
Smoking<-Smoking%>%
  select(SEQN,SMQ020)%>%
  drop_na()%>%
  filter(SMQ020<7)%>%
  mutate(smoking=as.numeric(SMQ020!=1))%>%
  select(SEQN,smoking)

BMI<-read_xpt("https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/BMX_D.XPT")
BMI<-BMI%>%select(SEQN,BMXBMI)%>%
  drop_na()%>%
  mutate(BMI=as.numeric(BMXBMI>=18.5&BMXBMI<=24.9))%>%
  select(SEQN,BMI)

Blood_pressure<-read_xpt("https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/BPX_D.XPT")
Blood_pressure<-Blood_pressure%>%select(SEQN,BPXSY1,BPXSY2,BPXSY3,BPXDI1,BPXDI2,BPXDI3)%>%
  gather(condition, BPX, BPXSY1:BPXDI3)%>%
  mutate(condition=substring(condition,1,5))%>%
  group_by(SEQN,condition)%>%
  summarise(BPX=mean(BPX,na.rm=T))%>%
  ungroup()%>%
  spread(condition,BPX)%>%
  drop_na()%>%
  filter(BPXDI!=0,BPXSY!=0)%>%
  transmute(SEQN,Blood_pressure=as.numeric((BPXDI<80)&(BPXSY<120)))
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```r
Diet_raw<-read_xpt("https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/DBQ_D.XPT")
Diet<-Diet_raw%>%select(SEQN,DBQ700)%>%
  drop_na()%>%
  filter(DBQ700!=7,DBQ700!=9)%>%
  transmute(SEQN,Diet=as.numeric(DBQ700<=3))

Diet_alt<-Diet_raw%>%
  select(SEQN,DBQ780)%>%
  drop_na()%>%
  filter(DBQ780!=77,DBQ780!=99)%>%
  transmute(SEQN,Diet=as.numeric(DBQ780<=4))

Physical_Activity<-read_xpt("https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/PAQIAF_D.XPT")
Physical_Activity<-Physical_Activity%>%
  select(SEQN,PADLEVEL,PADTIMES,PADDURAT)%>%
  drop_na()%>%
  mutate(times=PADTIMES*PADDURAT*PADLEVEL)%>%
  group_by(SEQN)%>%
  summarise(phy_act=as.numeric(sum(times)>=600))%>%
  select(SEQN,phy_act)

Blood_Cholesterol<-read_xpt("https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/TCHOL_D.XPT")
Blood_Cholesterol<-Blood_Cholesterol%>%
  select(SEQN,LBXTC)%>%
  drop_na()%>%
  transmute(SEQN,blood_cho=as.numeric(LBXTC<200))

Blood_Glucose<-read_xpt("https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/GLU_D.XPT")
Blood_Glucose<-Blood_Glucose%>%
  select(SEQN,LBXGLU)%>%
  drop_na()%>%
  transmute(SEQN,blood_glu=as.numeric(LBXGLU<=100))

# merge all seperate datasets together by SEQN
raw_data<-X_variable%>%inner_join(health_insurance, by = "SEQN")%>%
  inner_join(Smoking, by = "SEQN")%>%
  inner_join(BMI, by = "SEQN")%>%
  inner_join(Blood_pressure, by = "SEQN")%>%
  inner_join(Diet, by = "SEQN")%>%
  inner_join(Physical_Activity, by = "SEQN")%>%
  inner_join(Blood_Cholesterol, by = "SEQN")%>%
  inner_join(Blood_Glucose, by = "SEQN")

data<-raw_data%>%transmute(SEQN,CVH=smoking+Blood_pressure+phy_act+blood_cho+blood_glu+BMI+Diet,smoking,Blood_pressure,phy_act,blood_ch
o,blood_glu,BMI,Diet,gender,age,race,education,insurance)

# Then we get our final version dataset
data
```

```
## # A tibble: 1,255 x 14
##     SEQN   CVH smoking Blood_pressure phy_act blood_cho blood_glu   BMI  Diet
##    <dbl> <dbl>   <dbl>          <dbl>   <dbl>     <dbl>     <dbl> <dbl> <dbl>
##  1 31132     5       1              0       1         1         0     1     1
##  2 31134     4       1              0       0         1         1     0     1
##  3 31150     3       0              0       1         1         1     0     0
##  4 31153     4       0              0       1         1         1     0     1
##  5 31155     5       1              0       1         1         1     0     1
##  6 31158     4       0              0       1         1         0     1     1
##  7 31162     3       1              1       0         0         0     0     1
##  8 31167     3       0              0       1         0         1     0     1
##  9 31183     6       1              1       1         1         1     0     1
## 10 31187     5       1              1       1         1         1     0     0
## # ... with 1,245 more rows, and 5 more variables: gender <dbl>, age <dbl>,
## #   race <dbl>, education <dbl>, insurance <dbl>
```

```
# First we analyze the relationship between gender and each facor of CVH score using logistic model
gender_smoking <- summary(glm(smoking~gender+education+age+insurance+race,data=data, family = "binomial"))
gender_BP <-summary(glm(Blood_pressure~gender+education+age+insurance+race,data, family = "binomial"))
gender_phy <- summary(glm(phy_act~gender+education+age+insurance+race,data, family = "binomial"))
gender_BC <- summary(glm(blood_cho~gender+education+age+insurance+race,data, family = "binomial"))
gender_BG <- summary(glm(blood_glu~gender+education+age+insurance+race,data, family = "binomial"))
gender_BMI <- summary(glm(BMI~gender+education+age+insurance+race,data, family = "binomial"))
gender_Diet <- summary(glm(Diet~gender+education+age+insurance+race,data, family = "binomial"))

seperate<-data.frame(factor=c("smoking","Blood_pressure","phy_act","blood_cho","blood_glu","BMI","Diet"),gender_effect=rep(0,7),p_value=
rep(0,7),significance=rep("*",7),stringsAsFactors = FALSE)
j=1
for (i in list(gender_smoking,gender_BP,gender_phy,gender_BC,gender_BG,gender_BMI,gender_Diet)){
  seperate$gender_effect[j]=i$coefficients[2,1]
  seperate$p_value[j]=i$coefficients[2,4]
  p=rank(c(i$coefficients[2,4],0.001,0.01,0.05,0.1))[1]
  seperate$significance[j]=switch(p,
                                 "***",
                                 "**",
                                 "*",
                                 ".",
                                 " ")
  j=j+1
}

formattable::formattable(seperate)
```

factor	gender_effect	p_value	significance
smoking	-0.66312766	1.358578e-08	***
Blood_pressure	-0.90073688	3.975678e-13	***
phy_act	0.43817923	2.664810e-04	***
blood_cho	0.22166593	5.469620e-02	.
blood_glu	-0.72252699	1.961419e-08	***
BMI	-0.13881502	2.688388e-01	
Diet	0.03604901	7.928781e-01	

Then we will conduct OLS analysis

```
# We begin first with OLS regression and some diagnostics to view the general relationship between our data.
OLS_full<-lm(CVH~gender+race+education+insurance+age,data)
summary(OLS_full)
```

```
##
## Call:
## lm(formula = CVH ~ gender + race + education + insurance + age,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6916 -0.8478  0.0298  0.9311  3.7477
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.347783   0.187468  23.192  < 2e-16 ***
## gender      -0.371149   0.076324  -4.863 1.30e-06 ***
## race        -0.005439   0.037801  -0.144   0.8856
## education    0.165708   0.033903   4.888 1.15e-06 ***
## insurance    0.226256   0.103003   2.197   0.0282 *
## age         -0.023155   0.002223 -10.415  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.34 on 1249 degrees of freedom
## Multiple R-squared:  0.1355, Adjusted R-squared:  0.1321
## F-statistic: 39.17 on 5 and 1249 DF,  p-value: < 2.2e-16
```

```
# we delete race variable and get a seemly good model.
OLS_opt<-lm(CVH~gender+education+insurance+age,data)
summary(OLS_opt)
```

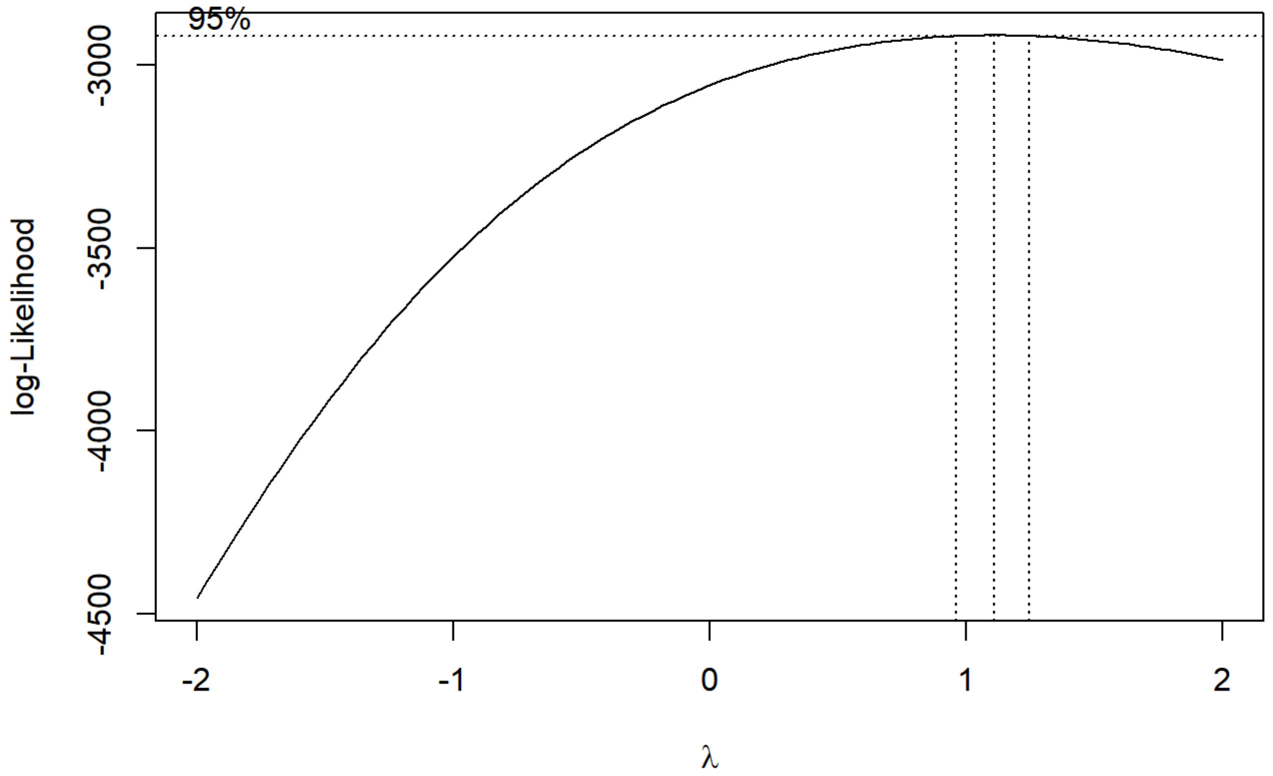
```
##
## Call:
## lm(formula = CVH ~ gender + education + insurance + age, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6907 -0.8440  0.0295  0.9355  3.7487
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.336613   0.170574  25.424  < 2e-16 ***
## gender      -0.371167   0.076294  -4.865 1.29e-06 ***
## education    0.164599   0.033002   4.988 6.97e-07 ***
## insurance    0.226468   0.102952   2.200    0.028 *
## age         -0.023178   0.002217 -10.455  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.339 on 1250 degrees of freedom
## Multiple R-squared:  0.1355, Adjusted R-squared:  0.1328
## F-statistic: 48.99 on 4 and 1250 DF,  p-value: < 2.2e-16
```

```
OLS2<-lm(CVH+1~gender+education+insurance+age,data)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

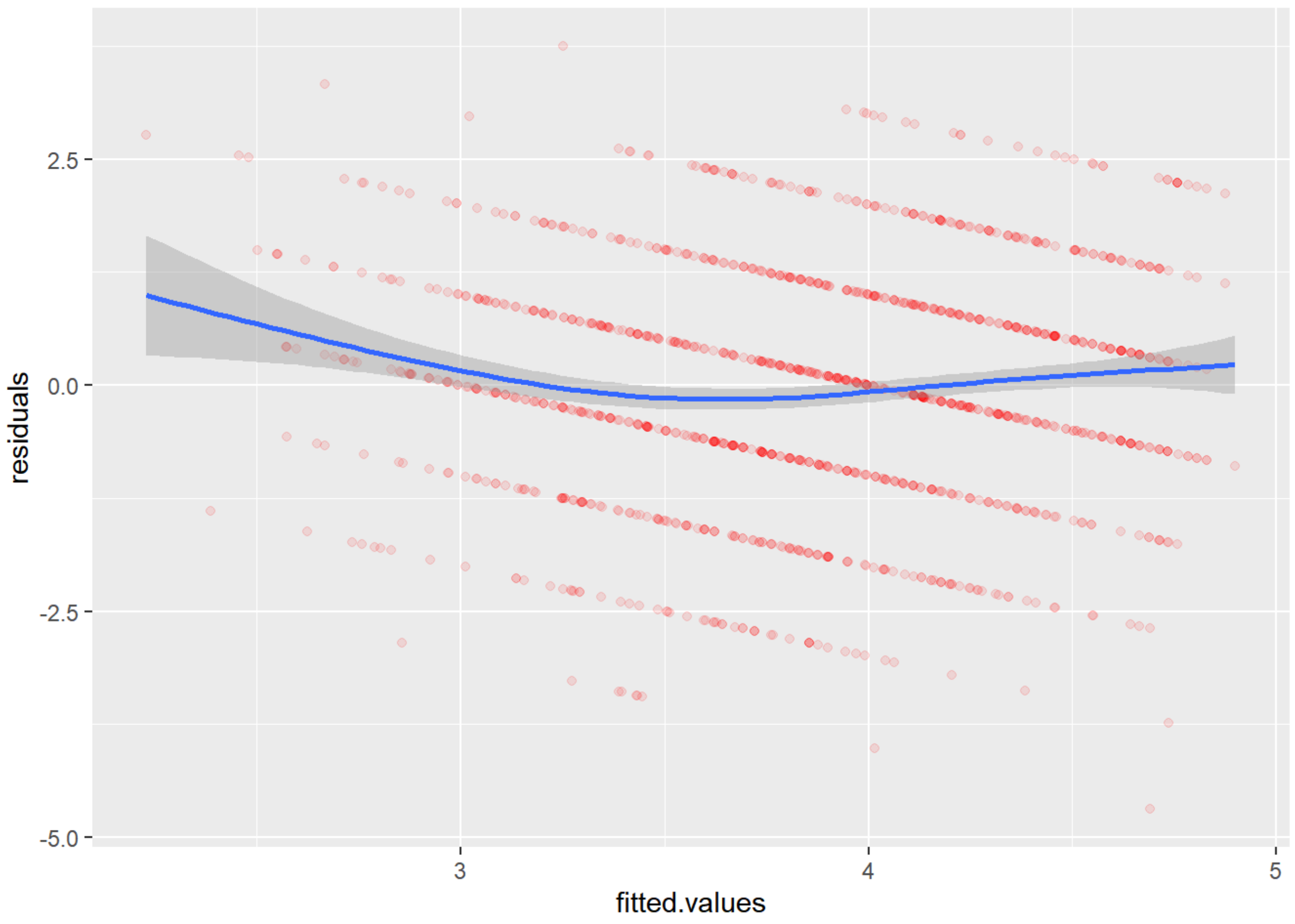
```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
boxcox(OLS2,plotit=T)# 1 is in the confidence interval so no need to do transformation
```



```
dat=data.frame(fitted.values=as.vector(OLS_opt$fitted),residuals=as.vector(OLS_opt$residuals))
ggplot(data=dat,aes(x=fitted.values,y=residuals))+geom_point(color="red",alpha=0.1)+geom_smooth(se=T)
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



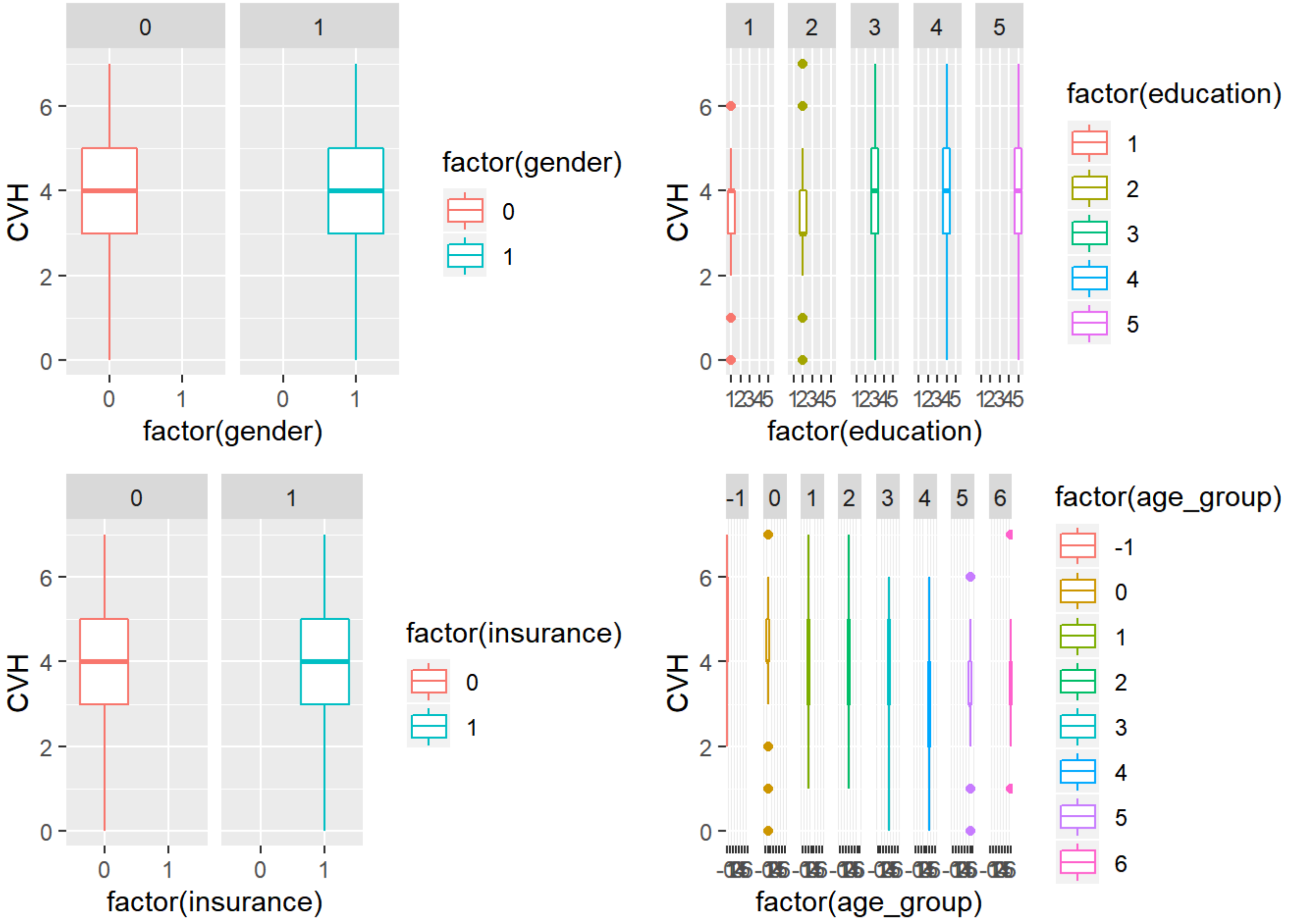
```
# It is obvious that OLS model doesn’t fit well with our dependent variables discontinuous. But we can still obtain the information that
 among the 5 predictors, gender, education level, insurance status, and age are more significant that race.
```

```
# We polt boxplots for response variable CVH grouped by different predictors. (Here we group age variable)
data$age_group=ceiling(data$age/10)-3
data$age_group[data$age==-1]=0

p1<-qplot(factor(gender), CVH, facets = . ~ factor(gender),
      colour = factor(gender), geom = "boxplot", data = data)
p2<-qplot(factor(insurance), CVH, facets = . ~ factor(insurance),
      colour = factor(insurance), geom = "boxplot", data = data)
p3<-qplot(factor(education), CVH, facets = . ~ factor(education),
      colour = factor(education), geom = "boxplot", data = data)
p4<-qplot(factor(age_group), CVH, facets = . ~ factor(age_group),
      colour = factor(age_group), geom = "boxplot", data = data)

multiplot(p1, p2, p3, p4,cols=2)
```

```
## Loading required package: grid
```



From the plots We can see that the CVH shows difference in different groups. It is resasonable to establish the following mixed effect model

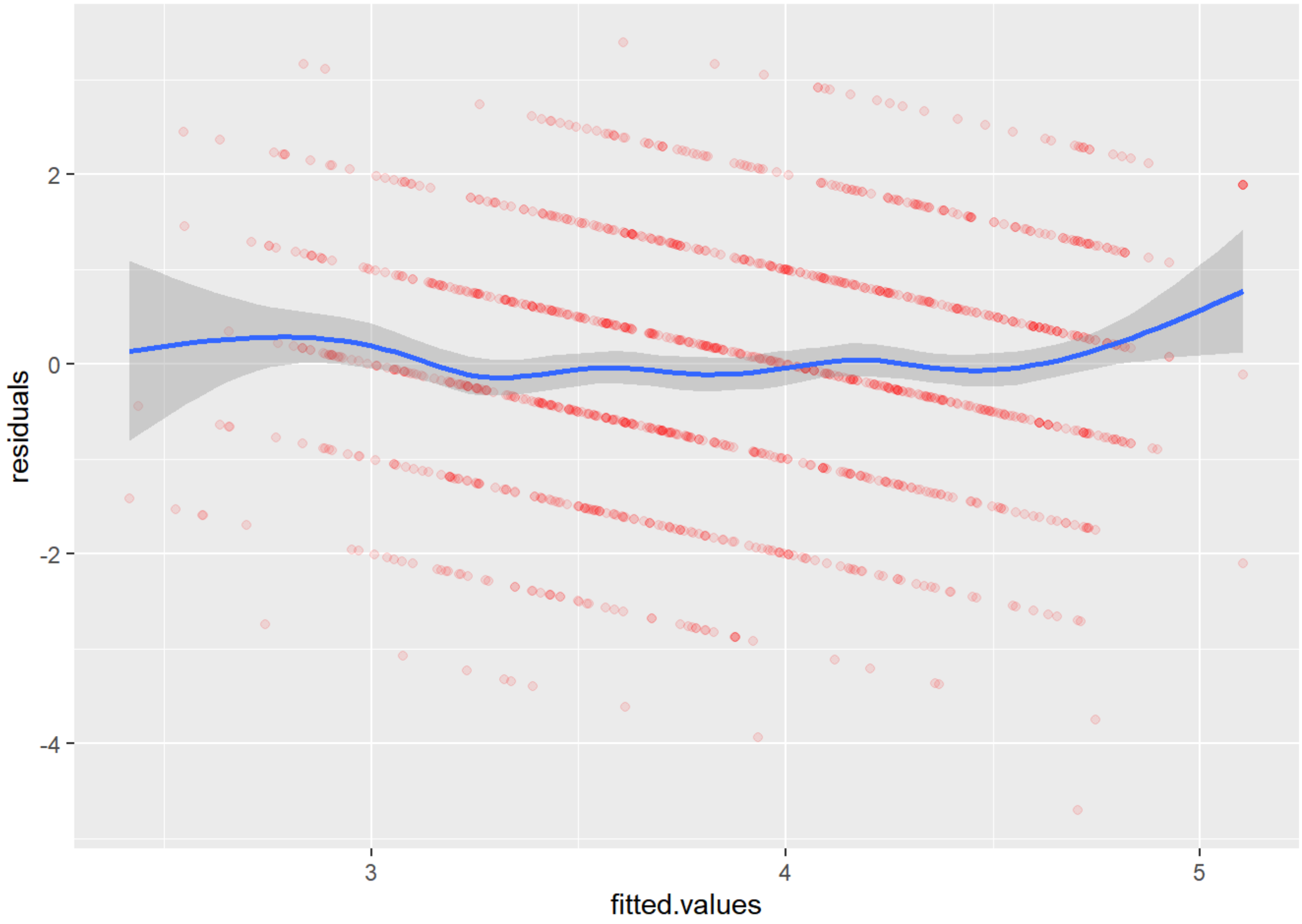
```
mixed=lme(CVH~gender+insurance+age+education, random=~1|age_group,
          method = ’ML’, data = data)

# Conduct Analysis of Variance and find this model dignificant. (?) and draw residuals_fitted plot
Anova(mixed)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: CVH
##              Chisq Df Pr(>Chisq)
## gender     24.3674  1  7.960e-07 ***
## insurance   4.5444  1    0.03303 *
## age        23.4255  1  1.298e-06 ***
## education  29.1570  1  6.674e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
dat=data.frame(fitted.values=as.vector(fitted(mixed)),residuals=as.vector(residuals(mixed)))
ggplot(data=dat,aes(x=fitted.values,y=residuals))+geom_point(color="red",alpha=0.1)+geom_smooth(se=T)
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Next we will test whether random effects are warranted

```
# lm.test
dev1 = -2*logLik(mixed);dev0 = -2*logLik(OLS_opt)
devdiff = as.numeric(dev0-dev1)
dfdiff <- attr(dev1,"df")-attr(dev0,"df");
cat('Chi-square =', devdiff, '(df=', dfdiff,'), p =',
    pchisq(devdiff,dfdiff,lower.tail=FALSE))
```

```
## Chi-square = 11.2655 (df= 1 ), p = 0.0007896086
```

And we also test the random effects in the model by comparing the model to a model fitted with just the fixed effects and excluding the random effects. (they are the same in depth)

```
model.fixed = gls(CVH~gender+insurance+age+education,
                  data=data,
                  method="ML")

anova(model.fixed,mixed)
```

```
##             Model df      AIC      BIC    logLik   Test L.Ratio p-value
## model.fixed     1  6 4302.088 4332.898 -2145.044
## mixed           2  7 4292.823 4328.767 -2139.411 1 vs 2 11.2655   8e-04
```

We can see that the random effects are significant, and the mixed model has smaller AIC and BIC and larger loglik

```
summary(mixed)
```

```
## Linear mixed-effects model fit by maximum likelihood
##  Data: data 
##        AIC      BIC    logLik
##   4292.823 4328.767 -2139.411
## 
## Random effects:
##  Formula: ~1 | age_group
##         (Intercept) Residual
## StdDev:   0.2336385  1.32387
## 
## Fixed effects: CVH ~ gender + insurance + age + education 
##                 Value  Std.Error   DF   t-value p-value
## (Intercept)  4.289836 0.27577966 1243 15.555303  0.0000
## gender      -0.373854 0.07588635 1243 -4.926494  0.0000
## insurance    0.217944 0.10244091 1243  2.127509  0.0336
## age         -0.021504 0.00445181 1243 -4.830343  0.0000
## education    0.177984 0.03302760 1243  5.388953  0.0000
##  Correlation: 
##           (Intr) gender insrnc age   
## gender    -0.163                     
## insurance -0.084  0.100              
## age       -0.805 -0.024 -0.152       
## education -0.422  0.024 -0.263  0.093
## 
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max 
## -3.55389551 -0.60902890  0.01652693  0.69647353  2.56213854 
## 
## Number of Observations: 1255
## Number of Groups: 8
```

To conclude, factors related to a favorable CVH score included insurance covered, younger age, female sex, and a higher level of education.

So the answer to the question we brought up is yes, women tend to have a better cardiovascular health condition than men in the US.