Homework 2: OLS and Probit

Yuzhe Wang

2022/2/3

Exercise 1 OLS estimate

```
## Import data.
datind2009 = fread('./data/datind2009.csv')
dat = na.omit(datind2009[,c('wage','age')])
dat = dat[dat$wage!=0,]
X = cbind(rep(1,length(dat$age)),dat$age)
Y = dat$wage
```

Calculate the correlation between Y and X(age).

```
age = dat$age
corr = sum((age-mean(age))*(Y-mean(Y))) /
  (sqrt(sum((age-mean(age))^2))*sqrt(sum((Y-mean(Y))^2)))
corr
```

```
## [1] 0.143492
```

Calculate the coefficients on this regression

```
beta_hat = c(solve(t(X)%*%X)%*%t(X)%*%Y)
names(beta_hat) = c('intercept','beta1(age)')
beta_hat
```

```
##  intercept beta1(age)
## 14141.1794   230.9923
```

Calculate the standard errors of beta.

```
# 1.OLS formula
error = Y - X %*% beta_hat
s_sqr = as.numeric((t(error) %*% error) / (length(Y) - 2))
se_beta_hat = diag(sqrt(s_sqr * solve(t(X) %*% X))) # (645.2348  14.8774)
names(se_beta_hat) = c('se_intercept(ols)','se_beta1(ols)')
# 2. Bootstrap
set.seed(123)
boot = function(data, boot_n){
  inter_result = c()
  beta1_result = c()
  for (i in 1:boot_n){
  indices = sample(1:nrow(data) ,nrow(data) ,replace = T)
  d = dat[indices,]
  Y = as.matrix(d[,1])
  X = as.matrix(cbind(rep(1,length(d[,2])),d[,2]))
  beta_hat_f = c(solve(t(X)%*%X)%*%t(X)%*%Y)
  inter_result = c(inter_result, beta_hat_f[1])
  beta1_result = c(beta1_result, beta_hat_f[2])
  }
  return(data.frame(se_intercept = inter_result, se_beta1 = beta1_result))
}
results1 = apply(boot(dat,49),2,function (x) sd(x))
names(results1) = c('se_intercept(boot49)','se_beta1(boot49)')
results2 = apply(boot(dat,499),2,function (x) sd(x))
names(results2) = c('se_intercept(boot499)','se_beta1(boot499)')

se_beta_hat # OLS
```

```
## se_intercept(ols)     se_beta1(ols)
##          645.2348           14.8774
```

```
results1 # 49 boot
```

```
## se_intercept(boot49)     se_beta1(boot49)
##             596.08216             15.72919
```

```
results2 # 499 boot
```

```
## se_intercept(boot499)     se_beta1(boot499)
##               625.2644               16.4461
```

Comment: Two strategies gives similar results. When you bootstrap more, you give more similar results.

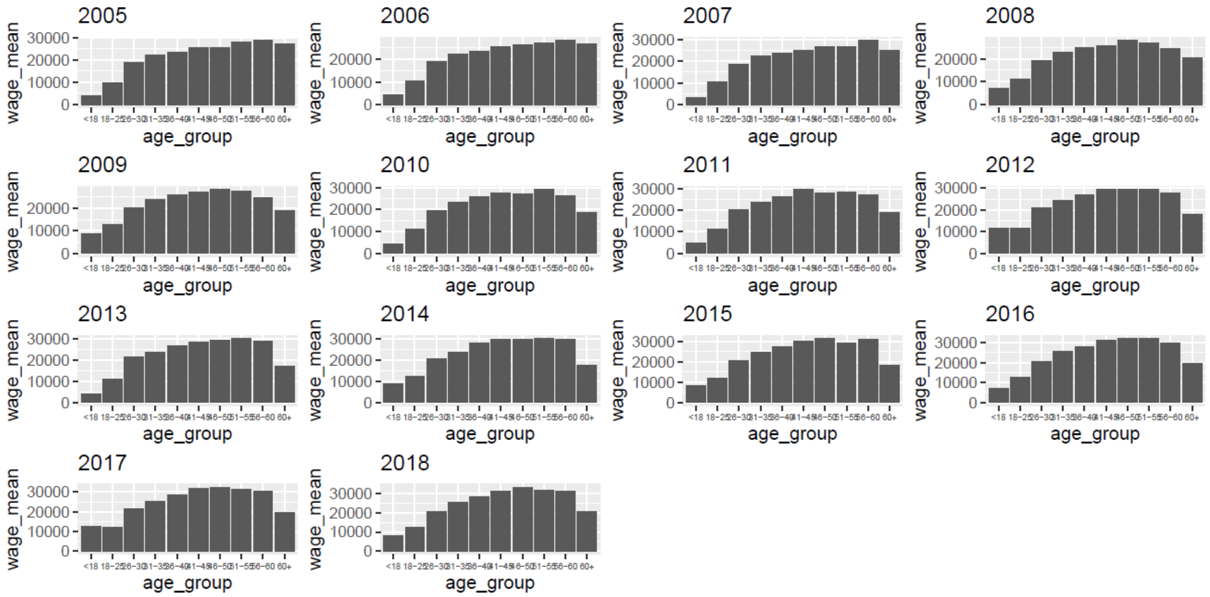
Exercise 2 Detrend Data

Create a categorical variable

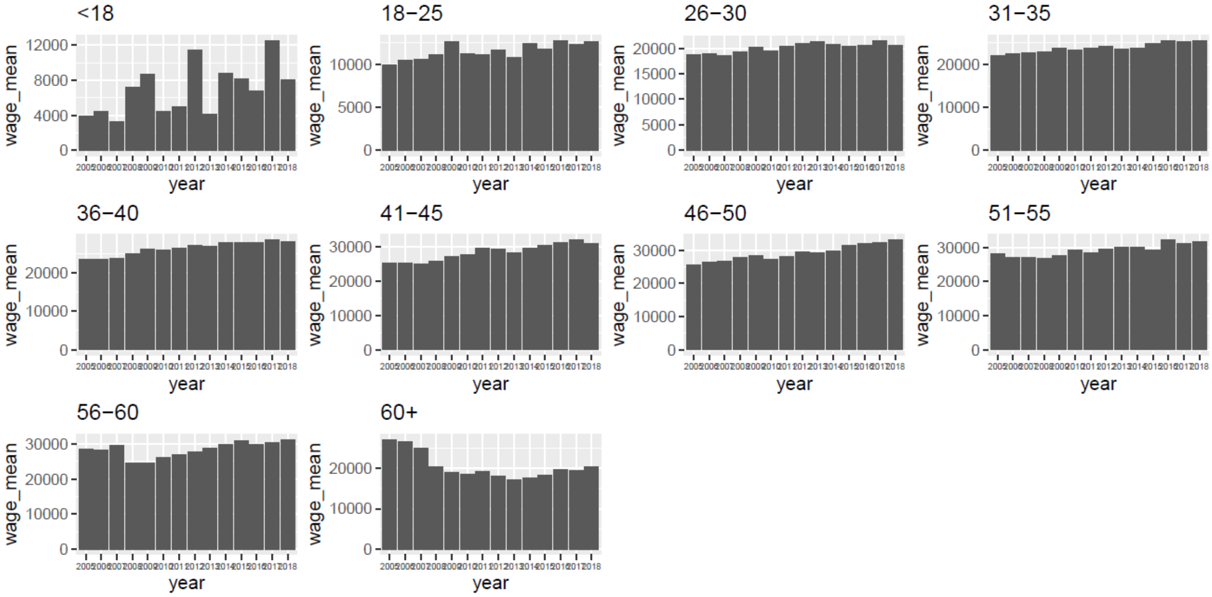
```
datind_dirs = paste("./data/datind",2005:2018,sep = "")%>%paste(".csv",sep = "")
datind_total = fread(datind_dirs[1])
for (i in c(2:length(datind_dirs))){
  dat_temp = fread(datind_dirs[i])
  dat_temp$idind = as.integer64(dat_temp$idind)
  datind_total = rbind(datind_total, dat_temp)
}
datind_total = datind_total[,-1]
datind_total = datind_total[!is.na(datind_total$age)&datind_total$wage!=0,]

ag_list = c()
for (ag in datind_total$age){
  if (ag < 18){ag_list = c(ag_list, "<18")}
  if (ag >= 18 & ag <= 25){ag_list = c(ag_list, "18-25")}
  if (ag >= 26 & ag <= 30){ag_list = c(ag_list, "26-30")}
  if (ag >= 31 & ag <= 35){ag_list = c(ag_list, "31-35")}
  if (ag >= 36 & ag <= 40){ag_list = c(ag_list, "36-40")}
  if (ag >= 41 & ag <= 45){ag_list = c(ag_list, "41-45")}
  if (ag >= 46 & ag <= 50){ag_list = c(ag_list, "46-50")}
  if (ag >= 51 & ag <= 55){ag_list = c(ag_list, "51-55")}
  if (ag >= 56 & ag <= 60){ag_list = c(ag_list, "56-60")}
  if (ag > 60){ag_list = c(ag_list, "60+")}
}
```

Plot the wage of each age group across years

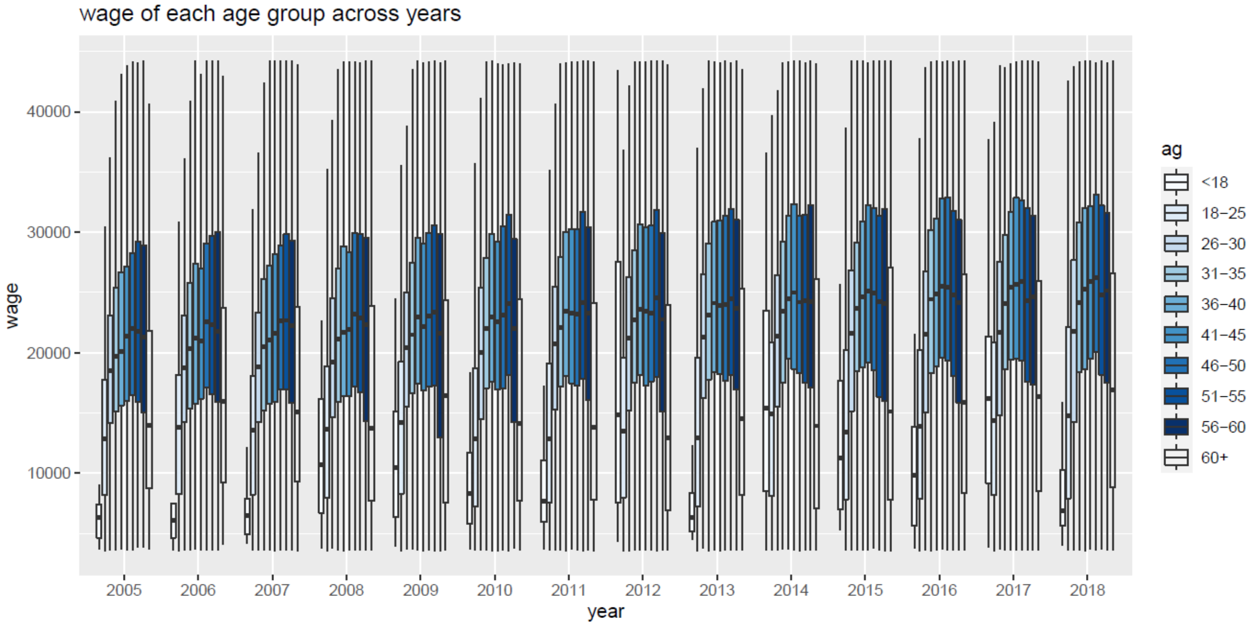
```
datind_total$ag = ag_list
datind_total$ag = factor(datind_total$ag)
datind_total$year = factor(datind_total$year)

age_year_matrix = by(datind_total$wage,datind_total[,c('ag','year')],mean)

for (y in 1:14){
  year = y + 2004
  plot_dat = data.frame(age_group = names(age_year_matrix[,y]),
                        wage_mean = age_year_matrix[,y])
  plot_temp = ggplot(plot_dat,aes(age_group,wage_mean)) +
    geom_bar(stat = 'identity') +
    ggtitle(label=year) +
    theme(axis.text.x = element_text(size=5),
          plot.margin = unit(rep(0.1,4),"cm"))
  if (y == 1){p1 = plot_temp}else{p1 = p1 + plot_temp}
}

for (a in 1:10){
  age_group = names(age_year_matrix[,1])[a]
  plot_dat = data.frame(wage_mean = age_year_matrix[a,],
                        year = names(age_year_matrix[1,]))
  plot_temp = ggplot(plot_dat,aes(year,wage_mean)) +
    geom_bar(stat = 'identity') +
    ggtitle(label=age_group) +
    theme(axis.text.x = element_text(size=5),
          plot.margin = unit(rep(0.1,4),"cm"))
  if (a == 1){p2 = plot_temp}else{p2 = p2 + plot_temp}
}

library(RColorBrewer)
p3 = ggplot(datind_total,aes(x=year,y=wage,fill=ag))+
  geom_boxplot(outlier.shape = NA)+
  scale_y_continuous(limits = quantile(datind_total$wage, c(0.1, 0.9)))+
  scale_fill_brewer()+
  ggtitle(label='wage of each age group across years')

p1 # mean of wage in each year
```



```
p2 # mean of wage in each ag
```



```
p3 # distribution in each year
```



Answer: Yes. Wage increases from 18 to 55 years old and starts to decrease after 56 years old.

After including a time fixed effect, how do the estimated coefficients change?

```
datind_total$year_2005 = as.numeric(datind_total$year==2005)
datind_total$year_2006 = as.numeric(datind_total$year==2006)
datind_total$year_2007 = as.numeric(datind_total$year==2007)
datind_total$year_2008 = as.numeric(datind_total$year==2008)
datind_total$year_2009 = as.numeric(datind_total$year==2009)
datind_total$year_2010 = as.numeric(datind_total$year==2010)
datind_total$year_2011 = as.numeric(datind_total$year==2011)
datind_total$year_2012 = as.numeric(datind_total$year==2012)
datind_total$year_2013 = as.numeric(datind_total$year==2013)
datind_total$year_2014 = as.numeric(datind_total$year==2014)
datind_total$year_2015 = as.numeric(datind_total$year==2015)
datind_total$year_2016 = as.numeric(datind_total$year==2016)
datind_total$year_2017 = as.numeric(datind_total$year==2017)
datind_total$year_2018 = as.numeric(datind_total$year==2018)

X = as.matrix(cbind(rep(1,length(datind_total$age)),datind_total$age)%>%
  cbind(datind_total[,c('year_2006','year_2007','year_2008','year_2009','year_2010',
            'year_2011','year_2012','year_2013','year_2014','year_2015',
            'year_2016','year_2017','year_2018')]))
Y = datind_total$wage
beta_hat = c(solve(t(X)%*%X)%*%t(X)%*%Y)
names(beta_hat) = c('intercept','beta1(age)','year_2006','year_2007','year_2008',
                    'year_2009','year_2010','year_2011','year_2012','year_2013',
                    'year_2014','year_2015','year_2016','year_2017','year_2018')
beta_hat
```

```
##   intercept  beta1(age)   year_2006   year_2007   year_2008   year_2009
## 10235.73310   305.87938   107.85740   118.46871   -79.97919   789.36741
##   year_2010   year_2011   year_2012   year_2013   year_2014   year_2015
##   554.06701  1089.32855  1571.50892  1409.63652  2041.75759  2306.79676
##   year_2016   year_2017   year_2018
##  2971.83787  2954.70265  3042.17066
```

Answer: After including a time fixed effect, estimated coefficient of age is larger.

Exercise 3 Numerical Optimization

Exclude all individuals who are inactive

```
datind2007 = fread('./data/datind2007.csv')[,-1]
datind2007 = datind2007[datind2007$empstat != 'Inactive' & datind2007$empstat != 'Retired',]
```

Write a function that returns the likelihood of the probit of being employed

```
flike = function(beta,x,y)
{
  x_beta = beta[1] + beta[2]*x
  pr = pnorm(x_beta)
  pr[pr>0.999999] = 0.999999
  pr[pr<0.000001] = 0.000001
  likelihood = y*log(pr) + (1-y)*log(1-pr)
  return(-sum(likelihood))
}
```

Optimize the model and interpret the coefficients

```
set.seed(123)
x = datind2007$age
y = as.numeric(datind2007$empstat == 'Employed')

ntry = 100
out = mat.or.vec(ntry,3)
for (i in 1:ntry){
  start = runif(2,-10,10)
  capture.output(res <- optim(start,
              fn = flike,
              method = "BFGS",
              control = list(trace=6,maxit=1000),
              x = x,
              y = y))
  out[i,c(1,2)] = res$par
  out[i,3] = res$value
}
out = data.frame(out)
colnames(out) = c('intercept', 'beta1_hat(age)', '-likelihood')
out[which(out$`-likelihood` == min(out$`-likelihood`)),]
```

```
##    intercept beta1_hat(age) -likelihood
## 76  1.043601    0.006939333    3555.891
```

Can you estimate the same model including wages as a determinant of labor market participation?

```r
datind2007 = datind2007[!is.na(datind2007$wage),]
# Write a function that returns the likelihood of the probit of being employed
flike = function(beta,x1,x2,y)
{
  x_beta = beta[1] + beta[2]*x1 + beta[3]*x2
  pr = pnorm(x_beta)
  pr[pr>0.999999] = 0.999999
  pr[pr<0.000001] = 0.000001
  likelihood = y*log(pr) + (1-y)*log(1-pr)
  return(-sum(likelihood))
}

# Optimize the model and interpret the coefficients
set.seed(123)
x1 = datind2007$age
x2 = datind2007$wage
y = as.numeric(datind2007$empstat == 'Employed')

ntry = 100
out = mat.or.vec(ntry,4)
for (i in 1:ntry){
  start = c(runif(1,-0.5,0.5),runif(1,-0.01,0.01),runif(1,-0.0001,0.0001))
  capture.output(res <- optim(start,
              fn=flike,
              method="BFGS",
              control=list(trace=6,maxit=1000),
              x1=x1,
              x2=x2,
              y=y))
  out[i,c(1,2,3)] = res$par
  out[i,4] = res$value
}
out = data.frame(out)
colnames(out) = c('intercept', 'beta1_hat(age)','beta2_hat(wage)', '-likelihood')
out[which(out$`-likelihood` == min(out$`-likelihood`)),]
```

```
##    intercept beta1_hat(age) beta2_hat(wage) -likelihood
## 46 0.1598384    0.006436114    7.296155e-05    2797.277
```

Answer: No. From the results, we can interpret that both age and wage have the positive effect on participation but age is not significant any more. The reason is that there are some unemployed individuals who have zero wage. This significantly influence the contribution of age. So, it is not reasonable to include the wage as a determinant.

Exercise 4 Discrete choice

Exclude all individuals who are inactive

```r
datind_dirs = paste("./data/datind",2005:2015,sep = "")%>%paste(".csv",sep = "")
datind_total = fread(datind_dirs[1])
for (i in c(2:length(datind_dirs))){
  dat_temp = fread(datind_dirs[i])
  dat_temp$idind = as.integer64(dat_temp$idind)
  datind_total = rbind(datind_total, dat_temp)
}
datind_total = datind_total[,-1]
datind_total = datind_total[!is.na(datind_total$age)
                            &!is.na(datind_total$wage)
                            &datind_total$empstat!='Inactive'
                            &datind_total$empstat!='Retired',]
datind_total$year_2005 = as.numeric(datind_total$year==2005)
datind_total$year_2006 = as.numeric(datind_total$year==2006)
datind_total$year_2007 = as.numeric(datind_total$year==2007)
datind_total$year_2008 = as.numeric(datind_total$year==2008)
datind_total$year_2009 = as.numeric(datind_total$year==2009)
datind_total$year_2010 = as.numeric(datind_total$year==2010)
datind_total$year_2011 = as.numeric(datind_total$year==2011)
datind_total$year_2012 = as.numeric(datind_total$year==2012)
datind_total$year_2013 = as.numeric(datind_total$year==2013)
datind_total$year_2014 = as.numeric(datind_total$year==2014)
datind_total$year_2015 = as.numeric(datind_total$year==2015)
```

Write and optimize the probit, logit, and the linear probability models

```
set.seed(123)
x1 = datind_total$age
year_2006 = datind_total$year_2006
year_2007 = datind_total$year_2007
year_2008 = datind_total$year_2008
year_2009 = datind_total$year_2009
year_2010 = datind_total$year_2010
year_2011 = datind_total$year_2011
year_2012 = datind_total$year_2012
year_2013 = datind_total$year_2013
year_2014 = datind_total$year_2014
year_2015 = datind_total$year_2015
y = as.numeric(datind_total$empstat == 'Employed')

# probit
flike = function(beta,x1,
                 year_2006,year_2007,
                 year_2008,year_2009,
                 year_2010,year_2011,
                 year_2012,year_2013,
                 year_2014,year_2015,
                 y)
{
  x_beta = beta[1] + beta[2]*x1 +
           beta[3]*year_2006+beta[4]*year_2007+
           beta[5]*year_2008+beta[6]*year_2009+
           beta[7]*year_2010+beta[8]*year_2011+
           beta[9]*year_2012+beta[10]*year_2013+
           beta[11]*year_2014+beta[12]*year_2015
  pr = pnorm(x_beta)
  pr[pr>0.999999] = 0.999999
  pr[pr<0.000001] = 0.000001
  likelihood = y*log(pr) + (1-y)*log(1-pr)
  return(-sum(likelihood))
}

ntry = 50
out = mat.or.vec(ntry,13)
hessian_list = list()
for (i in 1:ntry){
  start = c(runif(1,-1,1),runif(11,-0.15,0.15))
  capture.output(res <- optim(start,
              fn=flike,
              method="BFGS",
              control=list(trace=6,maxit=1000),
              x1=x1,
              year_2006=year_2006,year_2007=year_2007,
              year_2008=year_2008,year_2009=year_2009,
              year_2010=year_2010,year_2011=year_2011,
              year_2012=year_2012,year_2013=year_2013,
              year_2014=year_2014,year_2015=year_2015,
              y=y,
              hessian=TRUE))
  out[i,1:12] = res$par
  out[i,13] = res$value
  hessian_list[[i]] = res$hessian
}
out = data.frame(out)
colnames(out) = c('intercept', 'beta1_hat(age)',
                  'year_2006','year_2007','year_2008','year_2009','year_2010',
                  'year_2011','year_2012','year_2013','year_2014','year_2015',
                  '-likelihood')
probit_out = out[which(out$`-likelihood` == min(out$`-likelihood`)),]
fisher = solve(hessian_list[[as.numeric(row.names(probit_out))]])
sigma  = sqrt(diag(fisher))
z = probit_out/sigma
significance = apply(z,2,function(x) if(x>=1.96|x<=-1.96){return('yes')}else{return('no')})
probit_out = rbind(probit_out,sigma,z,significance)[,-13]
rownames(probit_out) = c('coefficient','std.error','z_value','significant_or_not(p=0.05)')

# logit
flike = function(beta,x1,
                 year_2006,year_2007,
                 year_2008,year_2009,
                 year_2010,year_2011,
                 year_2012,year_2013,
                 year_2014,year_2015,
                 y)
{
  x_beta = beta[1] + beta[2]*x1 +
           beta[3]*year_2006+beta[4]*year_2007+
           beta[5]*year_2008+beta[6]*year_2009+
           beta[7]*year_2010+beta[8]*year_2011+
           beta[9]*year_2012+beta[10]*year_2013+
           beta[11]*year_2014+beta[12]*year_2015
  pr = 1/(1+exp(-x_beta))
  pr[pr>0.999999] = 0.999999
  pr[pr<0.000001] = 0.000001
  likelihood = y*log(pr) + (1-y)*log(1-pr)
  return(-sum(likelihood))
}

ntry = 50
out = mat.or.vec(ntry,13)
hessian_list = list()
for (i in 1:ntry){
  start = c(runif(1,-1.5,1.5),runif(11,-0.3,0.3))
  capture.output(res <- optim(start,
              fn=flike,
              method="BFGS",
              control=list(trace=6,maxit=1000),
              x1=x1,
              year_2006=year_2006,year_2007=year_2007,
              year_2008=year_2008,year_2009=year_2009,
              year_2010=year_2010,year_2011=year_2011,
              year_2012=year_2012,year_2013=year_2013,
              year_2014=year_2014,year_2015=year_2015,
              y=y,
              hessian=TRUE))
  out[i,1:12] = res$par
  out[i,13] = res$value
  hessian_list[[i]] = res$hessian
}
out = data.frame(out)
colnames(out) = c('intercept', 'beta1_hat(age)',
                  'year_2006','year_2007','year_2008','year_2009','year_2010',
                  'year_2011','year_2012','year_2013','year_2014','year_2015',
                  '-likelihood')
logit_out = out[which(out$`-likelihood` == min(out$`-likelihood`)),]
fisher = solve(hessian_list[[as.numeric(row.names(logit_out))]])
sigma  = sqrt(diag(fisher))
z = logit_out/sigma
significance = apply(z,2,function(x) if(x>=1.96|x<=-1.96){return('yes')}else{return('no')})
logit_out = rbind(logit_out,sigma,z,significance)[,-13]
rownames(logit_out) = c('coefficient','std.error','z_value','significant_or_not(p=0.05)')

# linear
X = cbind(rep(1,length(x1)),x1,
                year_2006,year_2007,
                year_2008,year_2009,
                year_2010,year_2011,
                year_2012,year_2013,
                year_2014,year_2015)
Y = y
beta_hat = c(solve(t(X)%*%X)%*%t(X)%*%Y)
linear_out = data.frame(t(beta_hat))
```

```r
colnames(linear_out) = c('intercept', 'beta1_hat(age)',
                  'year_2006','year_2007','year_2008','year_2009','year_2010',
                  'year_2011','year_2012','year_2013','year_2014','year_2015')
error = Y - X %*% beta_hat
s_sqr = as.numeric((t(error) %*% error) / (length(Y) - 12))
se_beta_hat = diag(sqrt(s_sqr * solve(t(X) %*% X)))
t = linear_out/se_beta_hat
p_t = data.frame(t(apply(t,2,function(x) if(x>0){return(2*(1 - pt(x,df = length(y)-12)))}
            else{return(2*pt(x,df = length(y)-12))})))
significance = apply(p_t,2,function(x) if(x<=0.05){return('yes')}else{return('no')})

names(se_beta_hat) = colnames(linear_out)
names(significance) = colnames(linear_out)
colnames(p_t) = colnames(linear_out)
linear_out = rbind(linear_out,se_beta_hat,t,p_t,significance)
rownames(linear_out) = c('coefficient','std.error','t_value','Pr(>|t|)','significant_or_not(p=0.05)')

# outcomes
t(probit_out)
```

```
##                coefficient           std.error              z_value
## intercept      "0.750144549841469"   "0.0228584516268662"   "32.8169450007629"
## beta1_hat(age) "0.0123255213119796"  "0.000407151251988906" "30.2725860519162"
## year_2006      "0.0151237570179179"  "0.0228684381514262"   "0.661337556932138"
## year_2007      "0.0800165082125608"  "0.0230304567440568"   "3.47437782506027"
## year_2008      "0.108331604797681"   "0.0232496760457537"   "4.65948878532726"
## year_2009      "0.0253551870524284"  "0.0227812341923669"   "1.11298566347753"
## year_2010      "0.0219469599064602"  "0.022589985744894"    "0.971534916148447"
## year_2011      "0.0534232451305728"  "0.0226359025844835"   "2.36011110805864"
## year_2012      "0.00946563897531445" "0.0221160773163884"   "0.427998095679483"
## year_2013      "-0.0408188420003304" "0.0223553648282192"   "-1.82590811261576"
## year_2014      "-0.0343917184814785" "0.022346358974736"    "-1.53903007287946"
## year_2015      "-0.0555719812679453" "0.0223259313241514"   "-2.48912264671483"
##                significant_or_not(p=0.05)
## intercept      "yes"
## beta1_hat(age) "yes"
## year_2006      "no"
## year_2007      "yes"
## year_2008      "yes"
## year_2009      "no"
## year_2010      "no"
## year_2011      "yes"
## year_2012      "no"
## year_2013      "no"
## year_2014      "no"
## year_2015      "yes"
```

```r
t(logit_out)
```

```
##                coefficient           std.error              z_value
## intercept      "1.12027849791361"    "0.0442208580375443"   "25.3337123617654"
## beta1_hat(age) "0.0253745658194235"  "0.000814192937881467" "31.1652983449454"
## year_2006      "0.0271040046832894"  "0.0442079896601032"   "0.613101950386815"
## year_2007      "0.156145847002951"   "0.0449456624760302"   "3.47410269202783"
## year_2008      "0.209567476762742"   "0.0455243246426308"   "4.60341758846204"
## year_2009      "0.0424465570171505"  "0.0440510781121095"   "0.963575894990003"
## year_2010      "0.0372877311520324"  "0.0436979889977461"   "0.853305426800205"
## year_2011      "0.0968111385892818"  "0.043962152061804"    "2.2021473938123"
## year_2012      "0.0102247759621268"  "0.0427091126441469"   "0.239405019891653"
## year_2013      "-0.0879396909416147" "0.0429112064763231"   "-2.04934091028498"
## year_2014      "-0.0738575429829539" "0.0429782776548457"   "-1.71848540735151"
## year_2015      "-0.116550239677048"  "0.0428303032360987"   "-2.72120977137552"
##                significant_or_not(p=0.05)
## intercept      "yes"
## beta1_hat(age) "yes"
## year_2006      "no"
## year_2007      "yes"
## year_2008      "yes"
## year_2009      "no"
## year_2010      "no"
## year_2011      "yes"
## year_2012      "no"
## year_2013      "yes"
## year_2014      "no"
## year_2015      "yes"
```

```
t(linear_out)
```

```
##                coefficient            std.error
## intercept      "0.797878122174957"    "0.00421002955662914"
## beta1_hat(age) "0.00233862536052963"  "7.44454771981149e-05"
## year_2006      "0.00253105510487592"  "0.00409834063968949"
## year_2007      "0.013813512147016"    "0.00406159240181189"
## year_2008      "0.0181377017330758"   "0.00406963871455815"
## year_2009      "0.00380351827890235"  "0.0040706281748849"
## year_2010      "0.0033095512558051"   "0.00403745078265211"
## year_2011      "0.00852171659105394"  "0.00401294000199974"
## year_2012      "0.000719467816687747" "0.00396056308854669"
## year_2013      "-0.00858494110775709" "0.00404661934552406"
## year_2014      "-0.00723802774652748" "0.00403425868432918"
## year_2015      "-0.0114074787529691"  "0.00404696205236733"
##                t_value             Pr(>|t|)
## intercept      "189.518413455937"  "0"
## beta1_hat(age) "31.4139347150137"  "0"
## year_2006      "0.617580461800677" "0.536853023222924"
## year_2007      "3.4010089591594"   "0.000671580835559293"
## year_2008      "4.45683339609301"  "8.32494935965045e-06"
## year_2009      "0.934381160718492" "0.350109070696215"
## year_2010      "0.819713089760844" "0.412381208871184"
## year_2011      "2.12355943193952"  "0.0337089090190936"
## year_2012      "0.181657961406632" "0.855851445874079"
## year_2013      "-2.12150943163281" "0.0338808666850796"
## year_2014      "-1.79414071156199" "0.0727930660315456"
## year_2015      "-2.81877581389629" "0.004821455362715"
##                significant_or_not(p=0.05)
## intercept      "yes"
## beta1_hat(age) "yes"
## year_2006      "no"
## year_2007      "yes"
## year_2008      "yes"
## year_2009      "no"
## year_2010      "no"
## year_2011      "yes"
## year_2012      "no"
## year_2013      "yes"
## year_2014      "no"
## year_2015      "yes"
```

Answer: In all three method, the age is positively significant. However, the coefficients are different. But this is not a problem, because the coefficient doesn't mean anything here. We care about the marginal effect.

Exercise 5 Marginal Effects

```r
x1_bar = mean(x1)
year_2006_bar = mean(datind_total$year_2006)
year_2007_bar = mean(datind_total$year_2007)
year_2008_bar = mean(datind_total$year_2008)
year_2009_bar = mean(datind_total$year_2009)
year_2010_bar = mean(datind_total$year_2010)
year_2011_bar = mean(datind_total$year_2011)
year_2012_bar = mean(datind_total$year_2012)
year_2013_bar = mean(datind_total$year_2013)
year_2014_bar = mean(datind_total$year_2014)
year_2015_bar = mean(datind_total$year_2015)
x_bar = c(1,x1_bar,
            year_2006_bar,year_2007_bar,
            year_2008_bar,year_2009_bar,
            year_2010_bar,year_2011_bar,
            year_2012_bar,year_2013_bar,
            year_2014_bar,year_2015_bar)
```

marginal effect

```r
# probit marginal effect
x_bar_beta = sum(as.numeric(probit_out[1,]) * x_bar)
probit_me = dnorm(x_bar_beta) * as.numeric(probit_out[1,])
names(probit_me) = c('intercept_me', 'beta1_hat(age)_me',
                  'year_2006_me','year_2007_me','year_2008_me',
                  'year_2009_me','year_2010_me','year_2011_me',
                  'year_2012_me','year_2013_me','year_2014_me',
                  'year_2015_me')


# logit marginal effect
x_bar_beta = sum(as.numeric(logit_out[1,]) * x_bar)
logit_me = exp(-x_bar_beta)/(1+exp(-x_bar_beta))^2 * as.numeric(logit_out[1,])
names(logit_me) = c('intercept_me', 'beta1_hat(age)_me',
                  'year_2006_me','year_2007_me','year_2008_me',
                  'year_2009_me','year_2010_me','year_2011_me',
                  'year_2012_me','year_2013_me','year_2014_me',
                  'year_2015_me')
```

```
# results
probit_me #probit me
```

```
##      intercept_me beta1_hat(age)_me      year_2006_me      year_2007_me
##       0.132436372       0.002176044       0.002670066       0.014126739
##      year_2008_me      year_2009_me      year_2010_me      year_2011_me
##       0.019125707       0.004476403       0.003874688       0.009431757
##      year_2012_me      year_2013_me      year_2014_me      year_2015_me
##       0.001671138      -0.007206477      -0.006071782      -0.009811111
```

```
logit_me #logit me
```

```
##      intercept_me beta1_hat(age)_me      year_2006_me      year_2007_me
##      0.1007808492      0.0022827094      0.0024382907      0.0140469634
##      year_2008_me      year_2009_me      year_2010_me      year_2011_me
##      0.0188528016      0.0038185148      0.0033544241      0.0087091815
##      year_2012_me      year_2013_me      year_2014_me      year_2015_me
##      0.0009198263     -0.0079111013     -0.0066442638     -0.0104849215
```

standard error

```
# probit
flike = function(beta,x1,
                 year_2006,year_2007,
                 year_2008,year_2009,
                 year_2010,year_2011,
                 year_2012,year_2013,
                 year_2014,year_2015,
                 y)
{
  x_beta = beta[1] + beta[2]*x1 +
           beta[3]*year_2006+beta[4]*year_2007+
           beta[5]*year_2008+beta[6]*year_2009+
           beta[7]*year_2010+beta[8]*year_2011+
           beta[9]*year_2012+beta[10]*year_2013+
           beta[11]*year_2014+beta[12]*year_2015
  pr = pnorm(x_beta)
  pr[pr>0.999999] = 0.999999
  pr[pr<0.000001] = 0.000001
  likelihood = y*log(pr) + (1-y)*log(1-pr)
  return(-sum(likelihood))
}

for (boot_n in 1:10){
  datind_boot = datind_total[sample(1:nrow(datind_total),nrow(datind_total),replace = T),]
  datind_boot$year_2005 = as.numeric(datind_boot$year==2005)
  datind_boot$year_2006 = as.numeric(datind_boot$year==2006)
  datind_boot$year_2007 = as.numeric(datind_boot$year==2007)
  datind_boot$year_2008 = as.numeric(datind_boot$year==2008)
  datind_boot$year_2009 = as.numeric(datind_boot$year==2009)
  datind_boot$year_2010 = as.numeric(datind_boot$year==2010)
  datind_boot$year_2011 = as.numeric(datind_boot$year==2011)
  datind_boot$year_2012 = as.numeric(datind_boot$year==2012)
  datind_boot$year_2013 = as.numeric(datind_boot$year==2013)
```

```r
datind_boot$year_2014 = as.numeric(datind_boot$year==2014)
datind_boot$year_2015 = as.numeric(datind_boot$year==2015)

x1 = datind_boot$age
year_2006 = datind_boot$year_2006
year_2007 = datind_boot$year_2007
year_2008 = datind_boot$year_2008
year_2009 = datind_boot$year_2009
year_2010 = datind_boot$year_2010
year_2011 = datind_boot$year_2011
year_2012 = datind_boot$year_2012
year_2013 = datind_boot$year_2013
year_2014 = datind_boot$year_2014
year_2015 = datind_boot$year_2015
y = as.numeric(datind_boot$empstat == 'Employed')

ntry = 50
out = mat.or.vec(ntry,13)
for (i in 1:ntry){
  start = c(runif(1,-5,5),runif(11,-1,1))
  capture.output(res <- optim(start,
              fn=flike,
              method="BFGS",
              control=list(trace=6,maxit=1000),
              x1=x1,
              year_2006=year_2006,year_2007=year_2007,
              year_2008=year_2008,year_2009=year_2009,
              year_2010=year_2010,year_2011=year_2011,
              year_2012=year_2012,year_2013=year_2013,
              year_2014=year_2014,year_2015=year_2015,
              y=y))
  out[i,1:12] = res$par
  out[i,13] = res$value
}
out = data.frame(out)
colnames(out) = c('intercept', 'beta1_hat(age)',
                'year_2006','year_2007','year_2008','year_2009','year_2010',
                'year_2011','year_2012','year_2013','year_2014','year_2015',
                '-likelihood')
out = out[which(out$`-likelihood` == min(out$`-likelihood`)),]
out = out[1,-13]

x1_bar = mean(x1)
year_2006_bar = mean(datind_total$year_2006)
year_2007_bar = mean(datind_total$year_2007)
year_2008_bar = mean(datind_total$year_2008)
year_2009_bar = mean(datind_total$year_2009)
year_2010_bar = mean(datind_total$year_2010)
year_2011_bar = mean(datind_total$year_2011)
year_2012_bar = mean(datind_total$year_2012)
year_2013_bar = mean(datind_total$year_2013)
year_2014_bar = mean(datind_total$year_2014)
year_2015_bar = mean(datind_total$year_2015)
```

```
  x_bar = c(1,x1_bar,
              year_2006_bar,year_2007_bar,
              year_2008_bar,year_2009_bar,
              year_2010_bar,year_2011_bar,
              year_2012_bar,year_2013_bar,
              year_2014_bar,year_2015_bar)

  x_bar_beta = sum(as.numeric(out[1,]) * x_bar)
  me = dnorm(x_bar_beta) * as.numeric(out[1,])
  names(me) = c('intercept_me', 'beta1_hat(age)_me',
                    'year_2006_me','year_2007_me','year_2008_me',
                    'year_2009_me','year_2010_me','year_2011_me',
                    'year_2012_me','year_2013_me','year_2014_me',
                    'year_2015_me')
  if (boot_n==1){
    probit_me = me
  }else{
    probit_me = rbind(probit_me,me)
  }
}

# logit
flike = function(beta,x1,
                 year_2006,year_2007,
                 year_2008,year_2009,
                 year_2010,year_2011,
                 year_2012,year_2013,
                 year_2014,year_2015,
                 y)
{
  x_beta = beta[1] + beta[2]*x1 +
           beta[3]*year_2006+beta[4]*year_2007+
           beta[5]*year_2008+beta[6]*year_2009+
           beta[7]*year_2010+beta[8]*year_2011+
           beta[9]*year_2012+beta[10]*year_2013+
           beta[11]*year_2014+beta[12]*year_2015
  pr = 1/(1+exp(-x_beta))
  pr[pr>0.999999] = 0.999999
  pr[pr<0.000001] = 0.000001
  likelihood = y*log(pr) + (1-y)*log(1-pr)
  return(-sum(likelihood))
}

for (boot_n in 1:10){
  datind_boot = datind_total[sample(1:nrow(datind_total),nrow(datind_total),replace = T),]
  datind_boot$year_2005 = as.numeric(datind_boot$year==2005)
  datind_boot$year_2006 = as.numeric(datind_boot$year==2006)
  datind_boot$year_2007 = as.numeric(datind_boot$year==2007)
  datind_boot$year_2008 = as.numeric(datind_boot$year==2008)
  datind_boot$year_2009 = as.numeric(datind_boot$year==2009)
  datind_boot$year_2010 = as.numeric(datind_boot$year==2010)
  datind_boot$year_2011 = as.numeric(datind_boot$year==2011)
  datind_boot$year_2012 = as.numeric(datind_boot$year==2012)
```

```
datind_boot$year_2013 = as.numeric(datind_boot$year==2013)
datind_boot$year_2014 = as.numeric(datind_boot$year==2014)
datind_boot$year_2015 = as.numeric(datind_boot$year==2015)

x1 = datind_boot$age
year_2006 = datind_boot$year_2006
year_2007 = datind_boot$year_2007
year_2008 = datind_boot$year_2008
year_2009 = datind_boot$year_2009
year_2010 = datind_boot$year_2010
year_2011 = datind_boot$year_2011
year_2012 = datind_boot$year_2012
year_2013 = datind_boot$year_2013
year_2014 = datind_boot$year_2014
year_2015 = datind_boot$year_2015
y = as.numeric(datind_boot$empstat == 'Employed')

ntry = 50
out = mat.or.vec(ntry,13)
for (i in 1:ntry){
  start = c(runif(1,-5,5),runif(11,-1,1))
  capture.output(res <- optim(start,
              fn=flike,
              method="BFGS",
              control=list(trace=6,maxit=1000),
              x1=x1,
              year_2006=year_2006,year_2007=year_2007,
              year_2008=year_2008,year_2009=year_2009,
              year_2010=year_2010,year_2011=year_2011,
              year_2012=year_2012,year_2013=year_2013,
              year_2014=year_2014,year_2015=year_2015,
              y=y))
  out[i,1:12] = res$par
  out[i,13] = res$value
}
out = data.frame(out)
colnames(out) = c('intercept', 'beta1_hat(age)',
                'year_2006','year_2007','year_2008','year_2009','year_2010',
                'year_2011','year_2012','year_2013','year_2014','year_2015',
                '-likelihood')
out = out[which(out$`-likelihood` == min(out$`-likelihood`)),]
out = out[1,-13]

x1_bar = mean(x1)
year_2006_bar = mean(datind_total$year_2006)
year_2007_bar = mean(datind_total$year_2007)
year_2008_bar = mean(datind_total$year_2008)
year_2009_bar = mean(datind_total$year_2009)
year_2010_bar = mean(datind_total$year_2010)
year_2011_bar = mean(datind_total$year_2011)
year_2012_bar = mean(datind_total$year_2012)
year_2013_bar = mean(datind_total$year_2013)
year_2014_bar = mean(datind_total$year_2014)
```

```
  year_2015_bar = mean(datind_total$year_2015)
  x_bar = c(1,x1_bar,
              year_2006_bar,year_2007_bar,
              year_2008_bar,year_2009_bar,
              year_2010_bar,year_2011_bar,
              year_2012_bar,year_2013_bar,
              year_2014_bar,year_2015_bar)

  x_bar_beta = sum(as.numeric(out[1,]) * x_bar)
  me = exp(-x_bar_beta)/(1+exp(-x_bar_beta))^2 * as.numeric(out[1,])
  names(me) = c('intercept_me', 'beta1_hat(age)_me',
                    'year_2006_me','year_2007_me','year_2008_me',
                    'year_2009_me','year_2010_me','year_2011_me',
                    'year_2012_me','year_2013_me','year_2014_me',
                    'year_2015_me')
  if (boot_n==1){
    logit_me = me
  }else{
    logit_me = rbind(logit_me,me)
  }
}

results1 = apply(probit_me,2,sd)
results2 = apply(logit_me,2,sd)

# results
results1 #probit me sd
```

```
##      intercept_me beta1_hat(age)_me      year_2006_me      year_2007_me
##      3.213032e-03      7.089422e-05      3.865826e-03      5.470761e-03
##      year_2008_me      year_2009_me      year_2010_me      year_2011_me
##      3.079075e-03      3.719427e-03      3.465068e-03      6.375708e-03
##      year_2012_me      year_2013_me      year_2014_me      year_2015_me
##      4.218715e-03      3.234041e-03      4.950309e-03      5.340185e-03
```

```
results2 #logit me sd
```

```
##      intercept_me beta1_hat(age)_me      year_2006_me      year_2007_me
##      5.478954e-03      5.125078e-05      5.748080e-03      6.796910e-03
##      year_2008_me      year_2009_me      year_2010_me      year_2011_me
##      5.908529e-03      6.470457e-03      7.037623e-03      5.614941e-03
##      year_2012_me      year_2013_me      year_2014_me      year_2015_me
##      6.415935e-03      6.666737e-03      7.251274e-03      6.624104e-03
```