HW2 Wenxuan Wang

Exercise 1 OLS estimate

1.Calculate the correlation between Y and X.

```
> #============================================
> # Exercise 1:
> #============================================
> #1 Calculate the correlation between Y and X.
> datind2009=read.csv("datind2009.csv")
> datind2009_1=datind2009[complete.cases(datind2009$wag), ]
> a=rep(1,20232)
> b=cbind(a,datind2009_1$age)
> cor(datind2009_1$wage,datind2009_1$age)
[1] -0.1788512
```

The correlation between Y and X is -0.1788512.

2.Calculate the coefficients on this regression.

```
> #2 Calculate the coefficients on this regression
> X=b
> Y=datind2009_1$wage
> result=solve(t(X)%*%X)%*%t(X)%*%Y
> result
        [,1]
a 22075.1066
   -180.1765
```

The coefficient of the age is -180.1765, and the intercept of this model is 22075.1066.

3.Calculate the standard errors of beta

Using the standard formulas of the OLS.

```
> #Using the standard formulas of the OLS.
>
> df=length(datind2009_1$age)-2
> yhat=result[1,1] + result[2,1]*datind2009_1$age
> error_term=datind2009_1$wage-yhat
> theta2=sum((datind2009_1$wage-yhat)^2)/df
> se1=sqrt(theta2*solve(t(X)%*%X)[1,1])
> se2=sqrt(theta2*solve(t(X)%*%X)[2,2])
> se1
[1] 357.8275
> se2
[1] 6.968652
```

The standard error of the intercept is 357.8275, and the standard error of the coefficient is 6.968652.

Using bootstrap with 49 and 499 replications respectively. Comment on the difference between the two strategies

Using bootstrap I can get that with 49 replications, the standard error for the intercept is 359.0323, the standard error for the coefficient is 6.990535, with 499 replications, the standard error for the intercept is 357.8023, the standard error for the coefficient is 6.96879. I think the standard error using 499 replications is more accurate than using 49 replications, which is more closed to the result of OLS.

For 49

```
> #3.2 Using bootstrap with 49 and 499 replications respectively. Comment on the difference between the two st
rategies.
>
> for (i in 1:49) {
+ sample1=sample(1:20232,size=20232,replace=T)
+ data_choose =datind2009_1[sample1,]
+ df=length(datind2009_1$age)-2
+
+ X=cbind(rep(1,20232),data_choose$age)
+ Y=data_choose$wage
+ beta=solve(t(X)%*%X)%*%t(X)%*%Y
+ yhat=beta[1,1] + beta[2,1]*data_choose$age
+ error_term=data_choose$wage-yhat
+ theta2=sum((data_choose$wage-yhat)^2)/df
+ se1[i]=sqrt(theta2*solve(t(X)%*%X)[1,1])
+ se2[i]=sqrt(theta2*solve(t(X)%*%X)[2,2])
+ }
>
> se49_1=mean(se1)
> se49_1
[1] 359.0323
> se49_2=mean(se2)
> se49_2
[1] 6.990535
```

For 499

```
> for (i in 1:499) {
+   sample1=sample(1:20232,size=20232,replace=T)
+   data_choose =datind2009_1[sample1,]
+   df=length(datind2009_1$age)-2
+
+   X=cbind(rep(1,20232),data_choose$age)
+   Y=data_choose$wage
+   beta=solve(t(X)%*%X)%*%t(X)%*%Y
+   yhat=beta[1,1] + beta[2,1]*data_choose$age
+   error_term=data_choose$wage-yhat
+   theta2=sum((data_choose$wage-yhat)^2)/df
+   se1[i]=sqrt(theta2*solve(t(X)%*%X)[1,1])
+   se2[i]=sqrt(theta2*solve(t(X)%*%X)[2,2])
+ }
>
> se499_1=mean(se1)
> se499_1
[1] 357.8023
> se499_2=mean(se2)
> se499_2
[1] 6.96879
```

Exercise 2 Detrend Data

1.Create a categorical variable *ag*, which bins the age variables into the following groups:

```
> #=============================================
> # Exercise 2:
> #=============================================
> #1.Plot the wage of each age group across years. Is there a trend?
> datind2005=read.csv("datind2005.csv")
> datind2006=read.csv("datind2006.csv")
> datind2007=read.csv("datind2007.csv")
> datind2008=read.csv("datind2008.csv")
> datind2009=read.csv("datind2009.csv")
> datind2010=read.csv("datind2010.csv")
> datind2011=read.csv("datind2011.csv")
> datind2012=read.csv("datind2012.csv")
> datind2013=read.csv("datind2013.csv")
> datind2014=read.csv("datind2014.csv")
> datind2015=read.csv("datind2015.csv")
> datind2016=read.csv("datind2016.csv")
> datind2017=read.csv("datind2017.csv")
> datind2018=read.csv("datind2018.csv")
> Append1=rbind(datind2005,datind2006,datind2007,datind2008,datind2009,datind2010,datind2011,datind2012,da
tind2013,datind2014,datind2015,datind2016,datind2017,datind2018)
```

```
> Append1=Append1 %>%mutate(ag = 0,
+ ag = ifelse(18 <= age & 25 >= age, 1, ag),
+ ag = ifelse(26 <= age & 30 >= age, 2, ag),
+ ag = ifelse(31 <= age & 35 >= age, 3, ag),
+ ag = ifelse(36 <= age & 40 >= age, 4, ag),
+ ag = ifelse(41 <= age & 45 >= age, 5, ag),
+ ag = ifelse(46 <= age & 50 >= age, 6, ag),
+ ag = ifelse(51 <= age & 55 >= age, 7, ag),
+ ag = ifelse(56 <= age & 60 >= age, 8, ag),
+ ag = ifelse(60 < age, 9, ag))
```

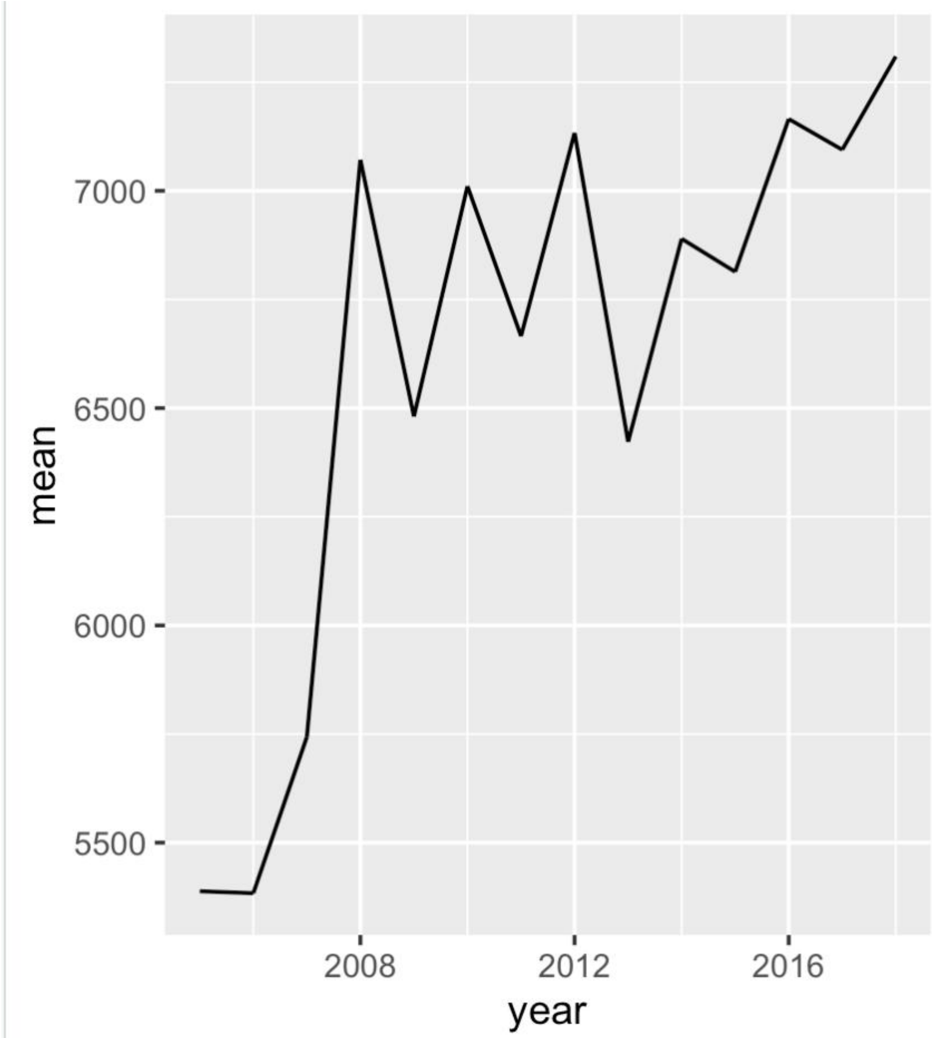
```
> Append1
    X        idind        idmen year    empstat respondent profession gender age  wage ag
1   1 1.120001e+18 1.200010e+15 2005   Inactive          1            Female  31 12334  3
2   2 1.120001e+18 1.200010e+15 2005   Inactive          0            Female  10    NA  0
3   3 1.120001e+18 1.200010e+15 2005   Employed          1         38   Male  32 50659  3
4   4 1.120001e+18 1.200010e+15 2005   Employed          0         45 Female  28 19231  2
5   5 1.120001e+18 1.200010e+15 2005    Retired          1            Female  90     0  9
6   6 1.120001e+18 1.200010e+15 2005   Employed          1         34   Male  37 31511  4
7   7 1.120001e+18 1.200010e+15 2005   Employed          0         42 Female  35 24873  3
8   8 1.120001e+18 1.200010e+15 2005   Employed          1         55 Female  41 30080  5
9   9 1.120001e+18 1.200010e+15 2005   Inactive          0            Female  16     0  0
10 10 1.120001e+18 1.200010e+15 2005   Employed          1         37   Male  55 43296  7
```

2.Plot the wage of each age group across years. Is there a trend?

Trend:the salary tends to increase over years, but when people get older, the overall salary is getting smaller.

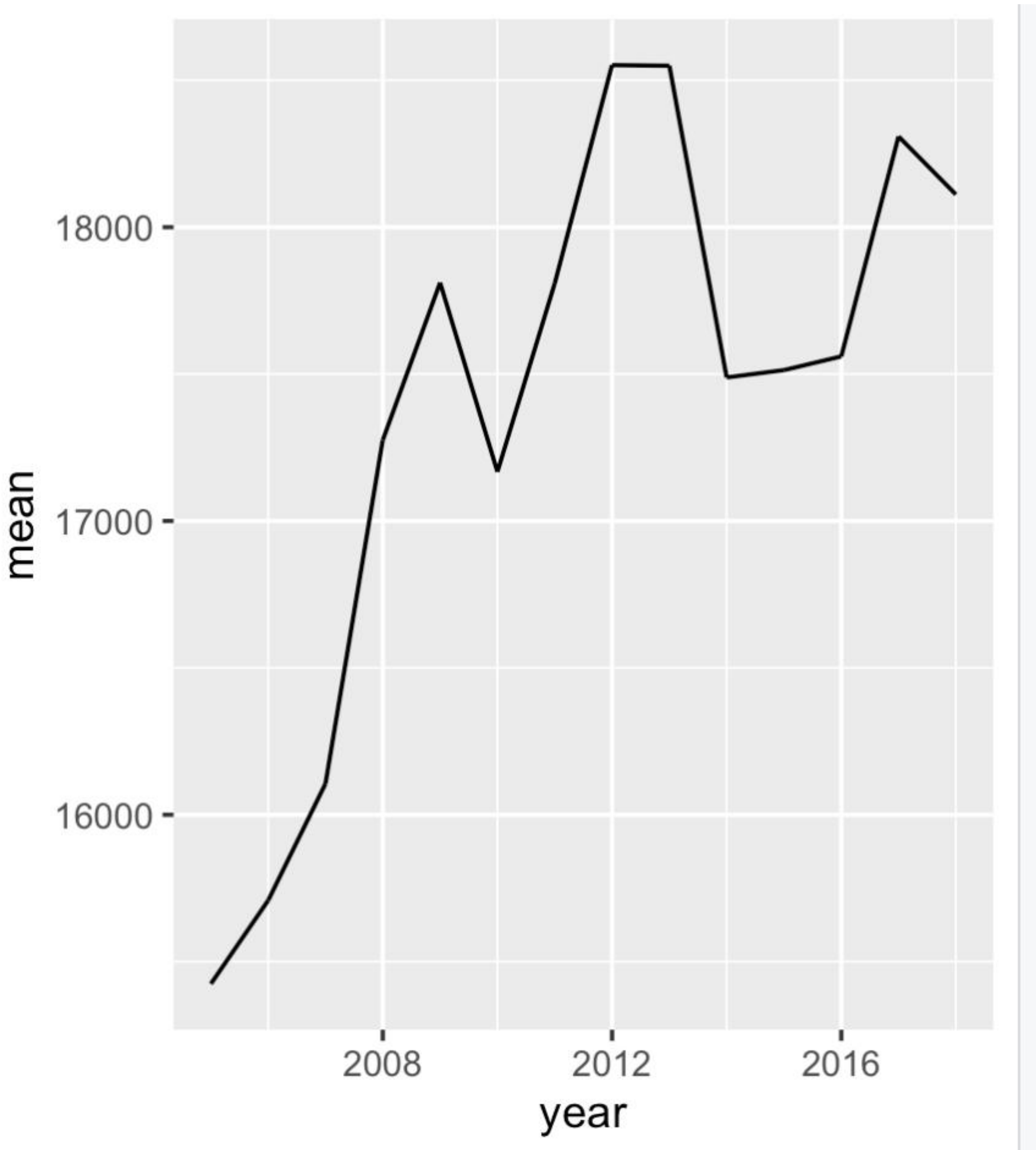
Group 18-25

```
> #2.Plot the wage of each age group across years. Is there a trend?
> Append1=Append1[complete.cases(Append1$wage), ]
> #group 18-25
> group1=Append1%>%filter(Append1$ag=="1")
> group1=group1%>%
+   group_by(year) %>%
+   mutate(mean= mean(wage))
> pic1=ggplot(group1,aes(x=year,y=mean) )+geom_line()
> pic1
```



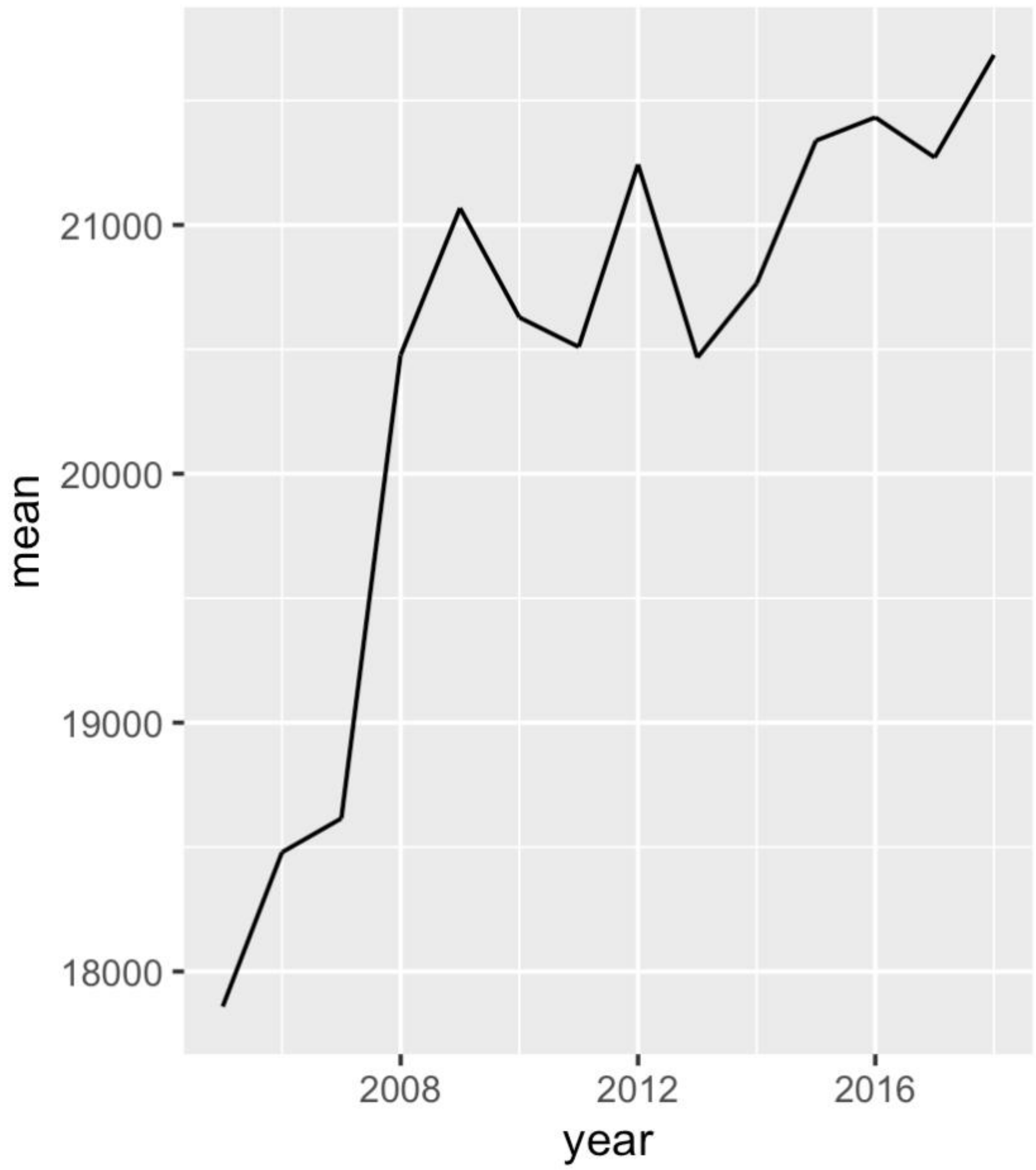
Group 26-30

```
> #group 26-30
> group2=Append1%>%filter(Append1$ag=="2")
> group2=group2%>%
+   group_by(year) %>%
+   mutate(mean= mean(wage))
> pic2=ggplot(group2,aes(x=year,y=mean) )+geom_line()
> pic2
```



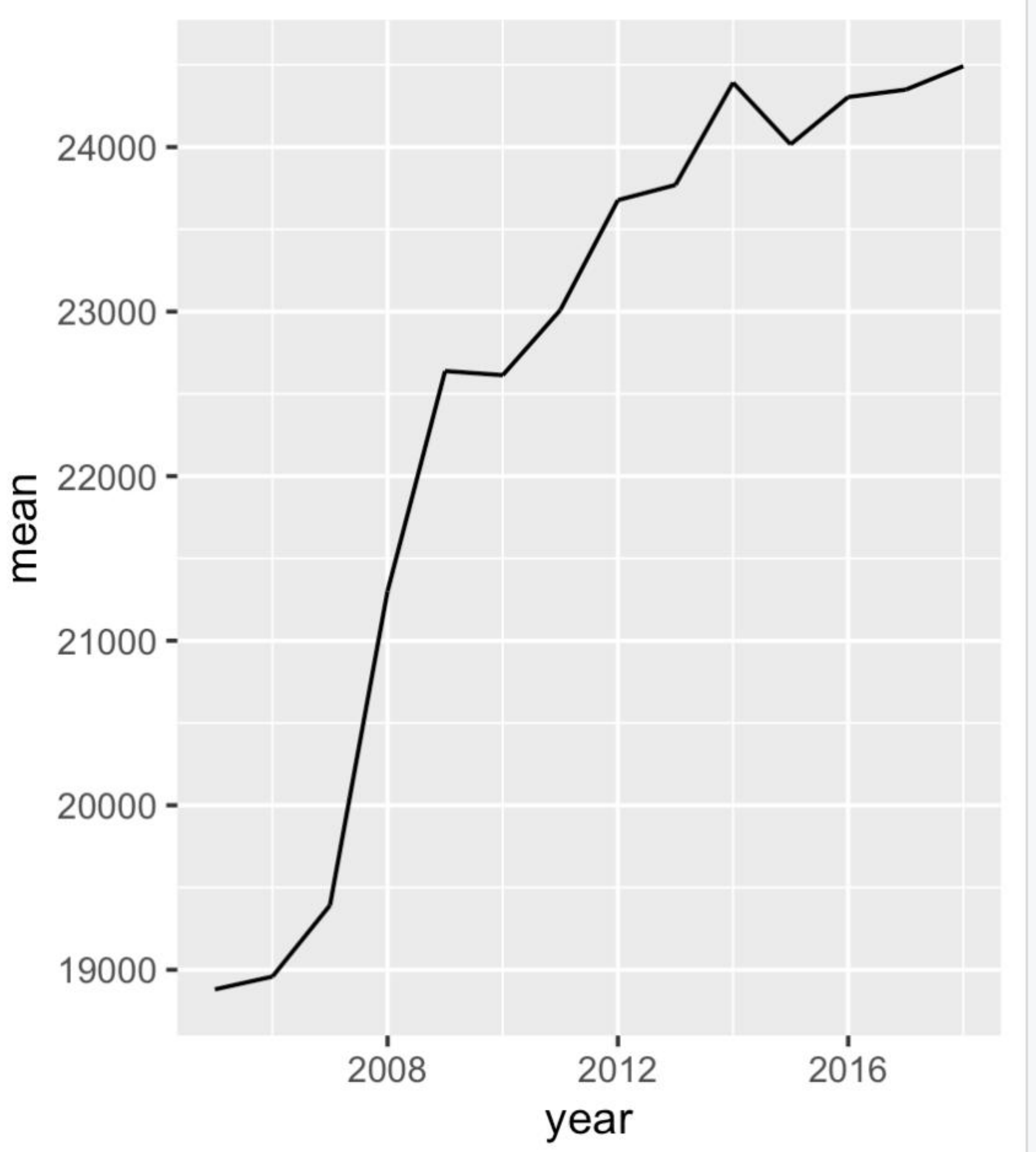
group 31-35

```
> #group 31-35
> group3=Append1%>%filter(Append1$ag=="3")
> group3=group3%>%
+   group_by(year) %>%
+   mutate(mean= mean(wage))
> pic3=ggplot(group3,aes(x=year,y=mean) )+geom_line()
> pic3
```



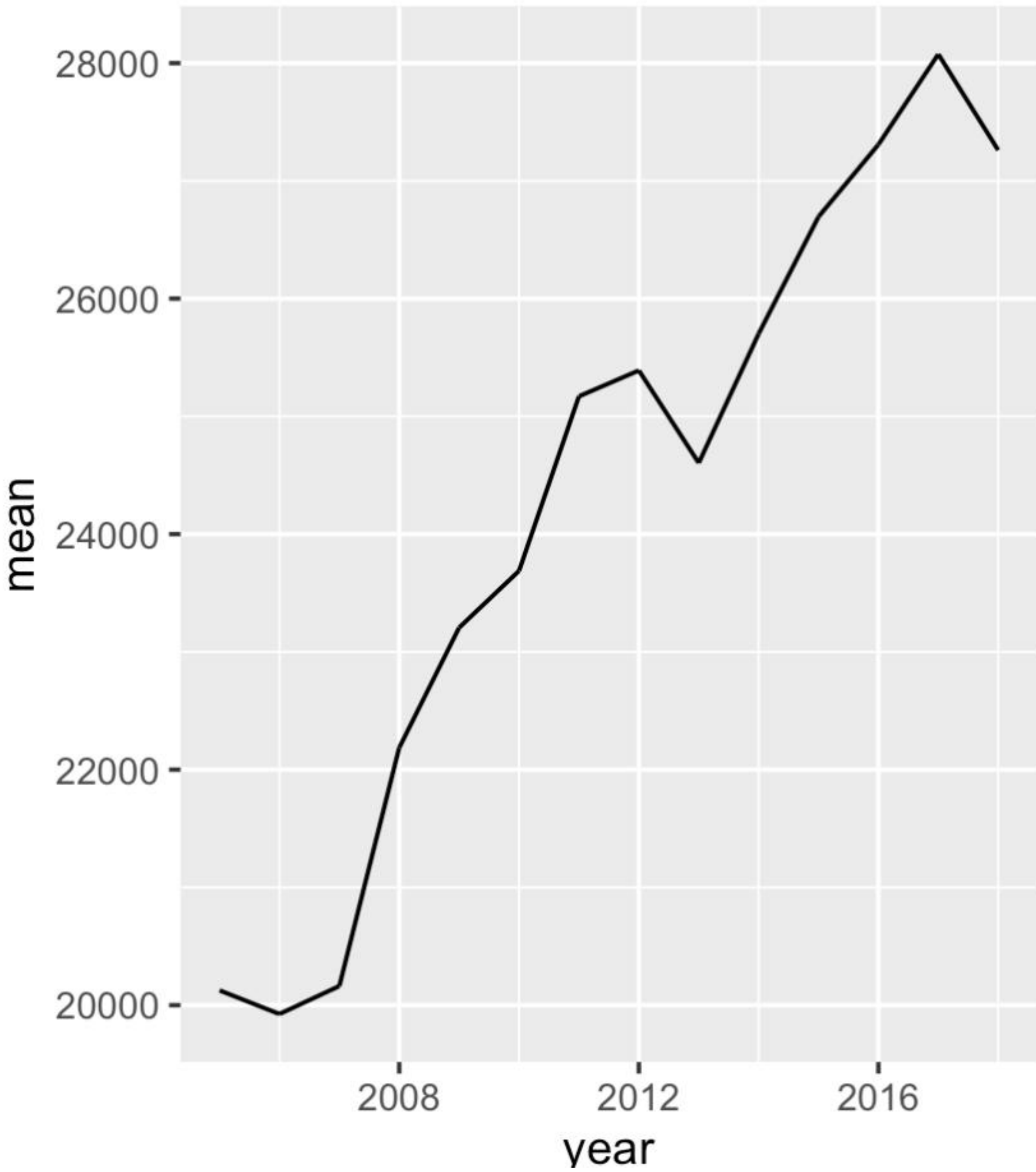
group 36-40

```
> #group 36-40
> group4=Append1%>%filter(Append1$ag=="4")
> group4=group4%>%
+   group_by(year) %>%
+   mutate(mean= mean(wage))
> pic4=ggplot(group4,aes(x=year,y=mean) )+geom_line()
> pic4
>
```



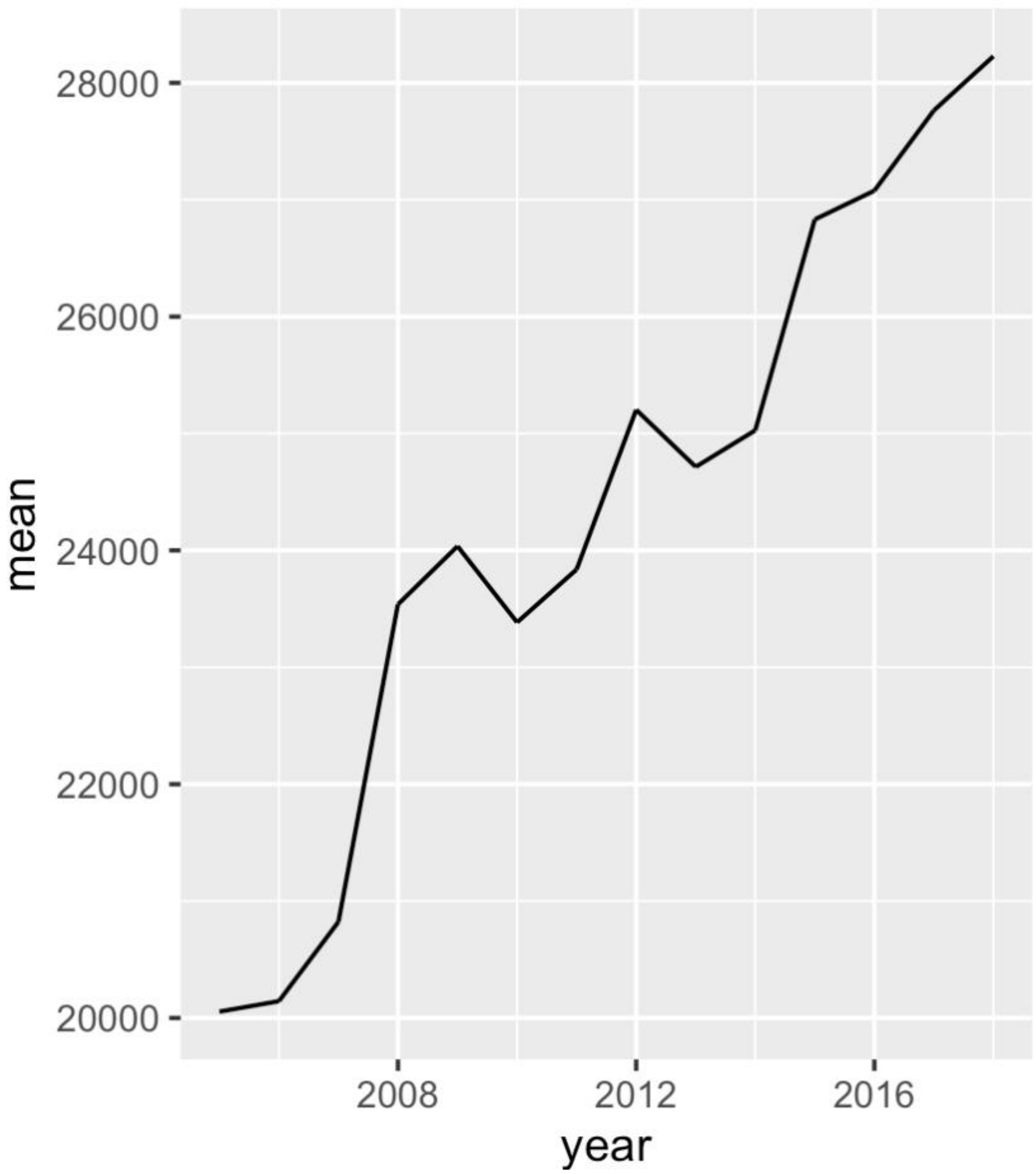
group 41-45

```
> #group 41-45
> group5=Append1%>%filter(Append1$ag=="5")
> group5=group5%>%
+   group_by(year) %>%
+   mutate(mean= mean(wage))
> pic5=ggplot(group5,aes(x=year,y=mean) )+geom_line()
> pic5
>
```



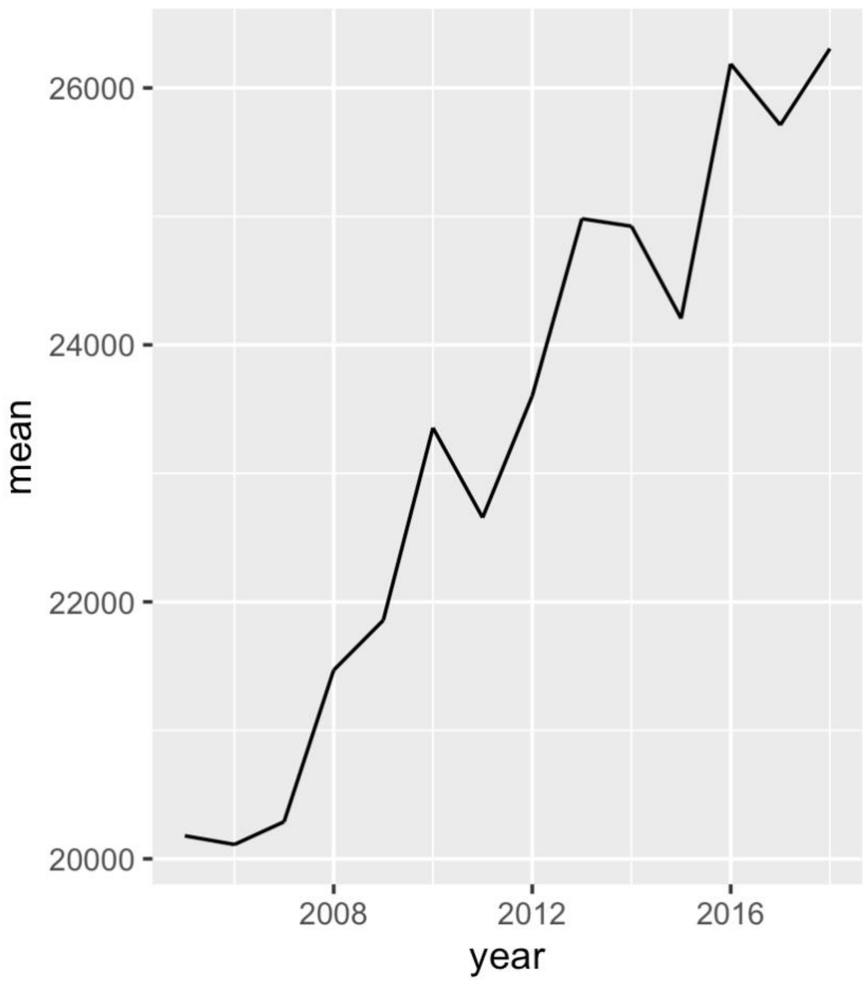
group 46-50

```
> #group 46-50
> group6=Append1%>%filter(Append1$ag=="6")
> group6=group6%>%
+   group_by(year) %>%
+   mutate(mean= mean(wage))
> pic6=ggplot(group6,aes(x=year,y=mean) )+geom_line()
> pic6
```



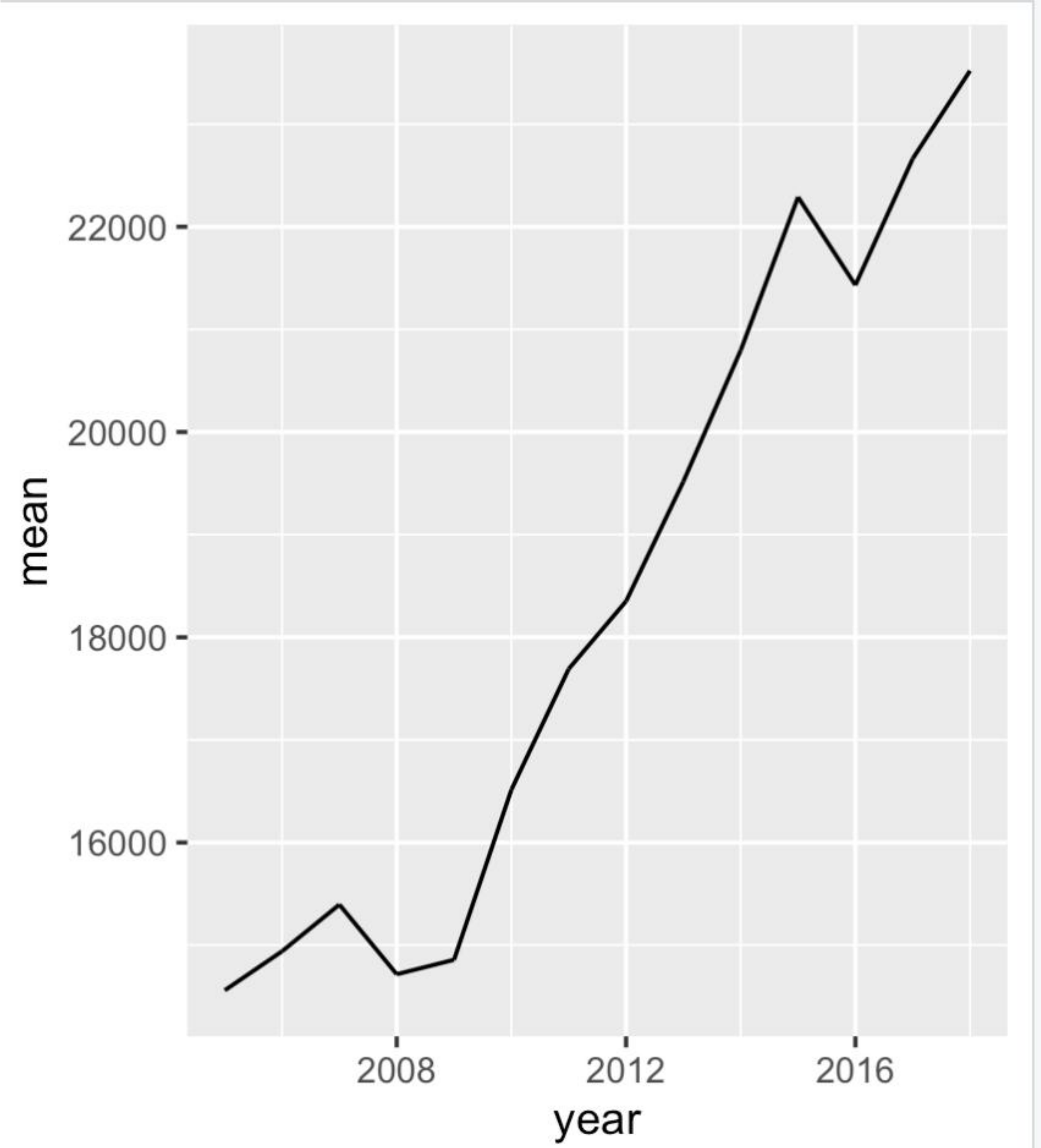
group 51-55

```
> #group 51-55
> group7=Append1%>%filter(Append1$ag=="7")
> group7=group7%>%
+   group_by(year) %>%
+   mutate(mean= mean(wage))
> pic7=ggplot(group7,aes(x=year,y=mean) )+geom_line()
> pic7
```



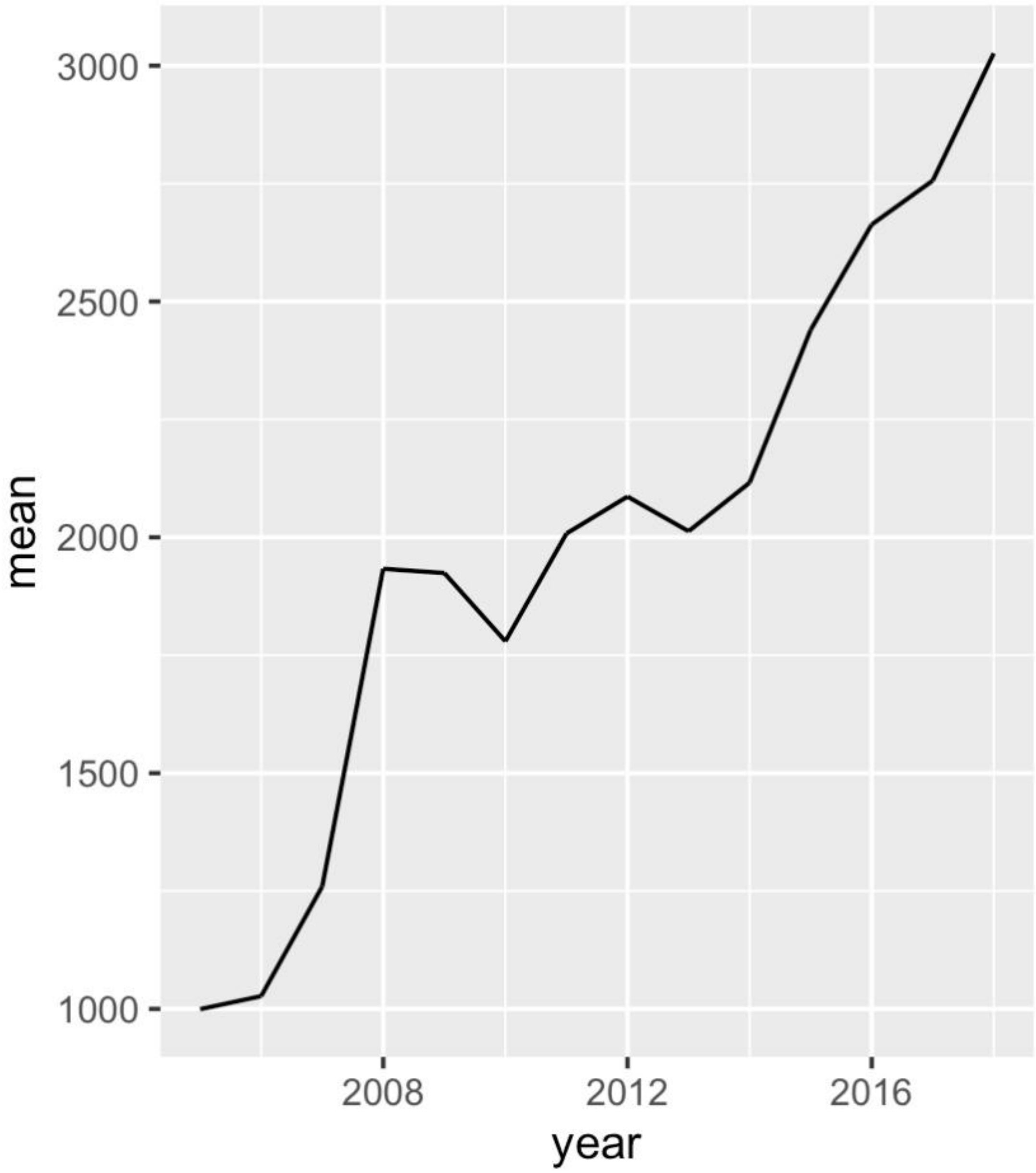
group 56-60

```
> #group 56-60
> group8=Append1%>%filter(Append1$ag=="8")
> group8=group8%>%
+   group_by(year) %>%
+   mutate(mean= mean(wage))
> pic8=ggplot(group8,aes(x=year,y=mean) )+geom_line()
> pic8
```



group 60+

```
> #group 60+
> group9=Append1%>%filter(Append1$ag=="9")
> group9=group9%>%
+   group_by(year) %>%
+   mutate(mean= mean(wage))
> pic9=ggplot(group9,aes(x=year,y=mean) )+geom_line()
> pic9
```



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

```
> #After including a time fixed effect, how do the estimated coefficients change?
>
> reg2=lm(Append1$wage ~ Append1$age)
> summary(reg2)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22559.3      104.3  216.25   <2e-16 ***
Append1$age   -182.5        2.0  -91.25   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> reg3= plm(Append1$wage ~ Append1$age, data=Append1,index=c("year"), model="within")
> summary(reg3)
```

```
Coefficients:
             Estimate Std. Error t-value  Pr(>|t|)
Append1$age -186.8793     2.0016 -93.366 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

from -182.4896 to -186.8793

Exercise 3 Numerical Optimization

1.Exclude all individuals who are inactive.

```
> #1Exclude all individuals who are inactive.
> datind2007=read.csv("datind2007.csv")
> datind2007_1=datind2007[complete.cases(datind2007$empstat), ]
> datind2007_2=filter(datind2007_1,datind2007_1$empstat!="Inactive")
> datind2007_2
      X         idind         idmen year    empstat respondent profession gender age  wage
1     1 1.140001e+18 1.400010e+15 2007 Unemployed          1         NA   Male  49     0
2     2 1.140001e+18 1.400010e+15 2007   Employed          0         52 Female  49 22744
3     4 1.140001e+18 1.400010e+15 2007   Employed          1         21   Male  40  1243
4     8 1.140001e+18 1.400010e+15 2007   Employed          1         22   Male  57     0
5     9 1.140001e+18 1.400010e+15 2007 Unemployed          0         NA Female  54     0
6    12 1.140001e+18 1.400010e+15 2007    Retired          1         NA   Male  71     0
7    13 1.140001e+18 1.400010e+15 2007   Employed          0         45 Female  63 19739
8    14 1.140001e+18 1.400010e+15 2007   Employed          0         38   Male  28 29561
```

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

The probit likelihood is -6582.155

```
> #2Write a function that returns the likelihood of the probit of being employed.
> datind2007_2$status = ifelse(datind2007_2$empstat == "Employed", 1, 0)
> datind2007_3=datind2007_2[complete.cases(datind2007_2$empstat), ]
>
> flike = function(par,x1,yvar)
+ {
+   xbeta           = par[1] + par[2]*x1
+   pr              = pnorm(xbeta)
+   pr[pr>0.999999] = 0.999999
+   pr[pr<0.000001] = 0.000001
+   like           = yvar*log(pr) + (1-yvar)*log(1-pr)
+   return(-sum(like))
+ }
>
> reg4 = glm(datind2007_3$status~datind2007_3$age,family = binomial(link = "probit"))
> summary(reg4)

Call:
glm(formula = datind2007_3$status ~ datind2007_3$age, family = binomial(link = "probit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4654  -0.5284   0.2938   0.7033   2.5822

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.8291873  0.0505722   75.72   <2e-16 ***
datind2007_3$age -0.0678642  0.0009246  -73.40   <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: 21944  on 16638  degrees of freedom
Residual deviance: 13164  on 16637  degrees of freedom
AIC: 13168

Number of Fisher Scoring iterations: 6

> test_coefficients = reg4$coefficients
> x=datind2007_3$age
> y = datind2007_3$status
> like(reg4$coefficients,x,y)
[1] -6582.155
> logLik(reg4)
'log Lik.' -6582.155 (df=2)
```

3.Optimize the model and interpret the coefficients.(解释)

The coefficient -0.0678642 means that, all else be equal, when age increases, the probability of labor market participation will decrease.

```
> #3
> #-0.0678642 means that when age increases, the labor market participation will decrease.
> opt1  = optim(reg4$coefficients,fn=flike,method="BFGS",control=list(trace=5,REPORT=1,maxit=10000),x=datind20
07_3$age,y=datind2007_3$status,hessian=TRUE)
initial  value 6582.154620
iter   1 value 6582.154620
final  value 6582.154620
converged
> opt1$par
     (Intercept) datind2007_3$age
       3.8291873       -0.0678642
```

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

```
> reg5 = glm(datind2007_3$status~datind2007_3$age+datind2007_3$wage,family = binomial(link = "probit"))
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

No. We cannot, because the algorithm did not converge and the fitted probabilities numerically 0 or 1 occurred.

Exercise 4 Discrete choice

1.Exclude all individuals who are inactive.

```
> #==========================================
> # Exercise 4:
> #==========================================
> #1
> datind2005=read.csv("datind2005.csv")
> datind2006=read.csv("datind2006.csv")
> datind2007=read.csv("datind2007.csv")
> datind2008=read.csv("datind2008.csv")
> datind2009=read.csv("datind2009.csv")
> datind2010=read.csv("datind2010.csv")
> datind2011=read.csv("datind2011.csv")
> datind2012=read.csv("datind2012.csv")
> datind2013=read.csv("datind2013.csv")
> datind2014=read.csv("datind2014.csv")
> datind2015=read.csv("datind2015.csv")
>
> Append2=rbind(datind2005,datind2006,datind2007,datind2008,datind2009,datind2010,datind2011,datind2012,datind
2013,datind2014,datind2015)
> Append2_1=Append2[complete.cases(Append2$empstat), ]
> Append2_2=filter(Append2_1,Append2_1$empstat!="Inactive")
> Append2_2$status = ifelse(Append2_2$empstat == "Employed", 1, 0)
> #creat the fixed effect variables
> Append2_2$y2006=ifelse(Append2_2$year == "2006", 1, 0)
> Append2_2$y2007=ifelse(Append2_2$year == "2007", 1, 0)
> Append2_2$y2008=ifelse(Append2_2$year == "2008", 1, 0)
> Append2_2$y2009=ifelse(Append2_2$year == "2009", 1, 0)
> Append2_2$y2010=ifelse(Append2_2$year == "2010", 1, 0)
> Append2_2$y2011=ifelse(Append2_2$year == "2011", 1, 0)
> Append2_2$y2012=ifelse(Append2_2$year == "2012", 1, 0)
> Append2_2$y2013=ifelse(Append2_2$year == "2013", 1, 0)
> Append2_2$y2014=ifelse(Append2_2$year == "2014", 1, 0)
> Append2_2$y2015=ifelse(Append2_2$year == "2015", 1, 0)
> Append2_2
```

```
> Append2_2
    X        idind        idmen year    empstat respondent profession gender age  wage status y2006
1   3 1.120001e+18 1.200010e+15 2005   Employed          1         38   Male  32 50659      1     0
2   4 1.120001e+18 1.200010e+15 2005   Employed          0         45 Female  28 19231      1     0
3   5 1.120001e+18 1.200010e+15 2005    Retired          1            Female  90     0      0     0
4   6 1.120001e+18 1.200010e+15 2005   Employed          1         34   Male  37 31511      1     0
5   7 1.120001e+18 1.200010e+15 2005   Employed          0         42 Female  35 24873      1     0
6   8 1.120001e+18 1.200010e+15 2005   Employed          1         55 Female  41 30080      1     0
7  10 1.120001e+18 1.200010e+15 2005   Employed          1         37   Male  55 43296      1     0
8  11 1.120001e+18 1.200010e+15 2005   Employed          0         54 Female  55 20426      1     0
9  12 1.120002e+18 1.200020e+15 2005   Employed          1         11   Male  57     0      1     0
10 13 1.120002e+18 1.200020e+15 2005   Employed          0         11 Female  52     0      1     0

   y2007 y2008 y2009 y2010 y2011 y2012 y2013 y2014 y2015
1      0     0     0     0     0     0     0     0     0
2      0     0     0     0     0     0     0     0     0
3      0     0     0     0     0     0     0     0     0
4      0     0     0     0     0     0     0     0     0
5      0     0     0     0     0     0     0     0     0
6      0     0     0     0     0     0     0     0     0
7      0     0     0     0     0     0     0     0     0
8      0     0     0     0     0     0     0     0     0
9      0     0     0     0     0     0     0     0     0
10     0     0     0     0     0     0     0     0     0
```

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

In the probit model,

For Probit model

```
> #Probit
> flike = function(par,x,x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,yvar)
+ {
+   xbeta           = par[1] + par[2]*x + par[3]*x1 + par[4]*x2+par[5]*x3+par[6]*x4+par[7]*x5+par[8]*x6+par
[9]*x7+par[10]*x8+par[11]*x9+par[12]*x10
+   pr              = pnorm(xbeta)
+   pr[pr>0.999999] = 0.999999
+   pr[pr<0.000001] = 0.000001
+   like           = yvar*log(pr) + (1-yvar)*log(1-pr)
+   return(-sum(like))
+ }
```

```
> #Optimize
> res1      = optim(runif(12,min=-0.1,max=0),fn=flike,method="BFGS",control=list(trace=5,REPORT=1,maxit=1000
00),x=Append2_2$age,x1=Append2_2$y2006,x2=Append2_2$y2007,x3=Append2_2$y2008, x4=Append2_2$y2009,x5=Append2_
2$y2010,x6=Append2_2$y2011,x7=Append2_2$y2012,x8=Append2_2$y2013,x9=Append2_2$y2014,x10=Append2_2$y2015,yvar
=Append2_2$status,hessian=TRUE)
initial  value 405023.064014
iter   2 value 136697.448037
iter   3 value 84875.934391
iter   4 value 84834.543441
iter   5 value 84728.978289
iter   6 value 84720.376165
iter   7 value 84714.968174
iter   8 value 83671.247309
iter   9 value 80527.871352
iter  10 value 80521.953203
iter  11 value 80520.963226
iter  12 value 80207.874240
iter  13 value 79915.791061
iter  14 value 79915.681402
iter  15 value 79915.049603
iter  16 value 79900.918683
iter  17 value 79892.072289
iter  18 value 79892.068640
iter  18 value 79892.068640
iter  18 value 79892.068640
final  value 79892.068640
converged
> res1$par[2]
[1] -0.0636211
```

The coefficient is -0.0636211 for probit model

Compare with the glm probit function, we can find that the coefficient is correct

```
> #compare with the glm probit function, we can find that the coefficient is correct
> reg5= glm(Append2_2$status ~ Append2_2$age+Append2_2$y2006+Append2_2$y2007+Append2_2$y2008+Append2_2$y2009
+Append2_2$y2010+Append2_2$y2011+Append2_2$y2012+Append2_2$y2013+Append2_2$y2014+Append2_2$y2015, data=Appen
d2_2,family = binomial(link = "probit"))
> reg5$coefficients[2]
Append2_2$age
  -0.06358962
```

For Logit model

```
> flike2 = function(par,x,x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,yvar)
+ {
+   xbeta          = par[1] + par[2]*x + par[3]*x1 + par[4]*x2+par[5]*x3+par[6]*x4+par[7]*x5+par[8]*x6+pa
r[9]*x7+par[10]*x8+par[11]*x9+par[12]*x10
+   pr             = exp(xbeta)/(1+exp(xbeta))
+   pr[pr>0.999999] = 0.999999
+   pr[pr<0.000001] = 0.000001
+   like          = yvar*log(pr) + (1-yvar)*log(1-pr)
+   return(-sum(like))
+ }
>
> start2    = runif(12,min=-0.1,max=0)
> res2      = optim(start2,fn=flike2,method="BFGS",control=list(trace=5,REPORT=1,maxit=100000),x=Append2_2
$age,x1=Append2_2$y2006,x2=Append2_2$y2007,x3=Append2_2$y2008, x4=Append2_2$y2009,x5=Append2_2$y2010,x6=Ap
pend2_2$y2011,x7=Append2_2$y2012,x8=Append2_2$y2013,x9=Append2_2$y2014,x10=Append2_2$y2015,yvar=Append2_2
$status,hessian=TRUE)
initial  value 369184.076626
iter   2 value 158069.970786
iter   3 value 81529.469390
iter   4 value 81420.501207
iter   5 value 81319.055241
iter   6 value 81268.890576
iter   7 value 81256.311961
iter   8 value 79521.126012
iter   9 value 77837.235989
iter  10 value 77827.505761
iter  11 value 77816.913648
iter  12 value 77044.130069
iter  13 value 76958.976920
iter  14 value 76958.888263
iter  15 value 76950.366247
iter  16 value 76913.358955
iter  17 value 76902.907825
```

```
iter  18 value 76900.756266
iter  19 value 76900.357639
iter  20 value 76900.351293
iter  20 value 76900.351293
iter  20 value 76900.351293
final  value 76900.351293
converged
> res2$par[2]
[1] -0.1241879
```

The coefficient is -0.1241879 for logit model.

Compare with the glm logit function, we can find that the coefficient is correct

```
> #compare with the glm logit function, we can find that the coefficient is correct
> reg6= glm(Append2_2$status ~ Append2_2$age+Append2_2$y2006+Append2_2$y2007+Append2_2$y2008+Append2_2$y2009
+Append2_2$y2010+Append2_2$y2011+Append2_2$y2012+Append2_2$y2013+Append2_2$y2014+Append2_2$y2015, data=Appen
d2_2,family =binomial)
> reg6$coefficients[2]
Append2_2$age
   -0.1241425
```

For linear model

```
> flike3 = function(par,x,x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,yvar)
+ {
+   y_hat          = par[1] + par[2]*x + par[3]*x1 + par[4]*x2+par[5]*x3+par[6]*x4+par[7]*x5+par[8]*x6+par
[9]*x7+par[10]*x8+par[11]*x9+par[12]*x10
+   yvar = as.numeric(y_hat)
+   error                  = Append2_2$status - yvar
+   return(sum(error^2))
+ }
>
> start3=runif(12,min=-0.1,max=0)
> res3 = optim(start3,fn=flike3,method="BFGS",control=list(trace=5,maxit=100000),x=Append2_2$age,x1=Append2_
2$y2006,x2=Append2_2$y2007,x3=Append2_2$y2008, x4=Append2_2$y2009,x5=Append2_2$y2010,x6=Append2_2$y2011,x7=A
ppend2_2$y2012,x8=Append2_2$y2013,x9=Append2_2$y2014,x10=Append2_2$y2015,yvar=Append2_2$status)
initial  value 5732252.897509
iter  10 value 40391.747196
final  value 25694.046354
converged
> res3$par[2]
[1] -0.01846614
```

The coefficient is -0.01846614 for linear model.

Compare with the linear model function, we can find that the coefficient is correct

```
> #compare with the linear model function, we can find that the coefficient is correct
> reg7= lm(Append2_2$status ~ Append2_2$age+Append2_2$y2006+Append2_2$y2007+Append2_2$y2008+Append2_2$y2009+
Append2_2$y2010+Append2_2$y2011+Append2_2$y2012+Append2_2$y2013+Append2_2$y2014+Append2_2$y2015, data=Append
2_2)
> reg7$coefficients[2]
Append2_2$age
  -0.01846614
```

3.Interpret and compare the estimated coefficients. How significant are they?

From Question 2, I get the coefficients of these models. The coefficient is -0.0636211 for probit model. The coefficient is -0.1241879 for logit model. The coefficient is -0.01946614 for linear model. We can find that the signs of the three model are the same. It shows that age has a negative effect on the labor market participation. The probit and logit model's coefficients mean that all else be equal, age has a negative effect on the probability of labor market participation. However, in the linear model, -0.01846614 means that when age increases 1 year, the labor market participation will decrease 0.01846614.

In this question, I calculate the t value. In linear model, it is -381.5322, which is significant at 1% level. In logit model, it is -224.1728, which is significant at 1% level. In probit model, it is -260.0777, which is significant at 1% level.

For linear model

```
> #3 Interpret and compare the estimated coefficients. How significant are they?
> #Calculate the standard error to get the T value
> #For linear model
> # Calculate the correlation between Y and X.
> a=rep(1,190296)
> b=cbind(a,Append2_2$age,Append2_2$y2006,Append2_2$y2007,Append2_2$y2008,Append2_2$y2009,Append2_2$y2010,Appe
nd2_2$y2011,Append2_2$y2012,Append2_2$y2013,Append2_2$y2014,Append2_2$y2015)
> X=b
> Y=Append2_2$status
> result=solve(t(X)%*%X)%*%t(X)%*%Y
> result
           [,1]
a  1.5446069056
  -0.0184661429
   0.0021767159
   0.0065161169
   0.0075729197
  -0.0017816633
  -0.0008924194
   0.0053246825
   0.0035872911
  -0.0022909960
   0.0045556570
   0.0023945445
```

```
> df=length(Append2_2$age)-12
> yhat=result[1,1] + result[2,1]*Append2_2$age+result[3,1]*Append2_2$y2006+result[4,1]*Append2_2$y2007+result
[5,1]*Append2_2$y2008+result[6,1]*Append2_2$y2009+result[7,1]*Append2_2$y2010+result[8,1]*Append2_2$y2011+resu
lt[9,1]*Append2_2$y2012+result[10,1]*Append2_2$y2013+result[11,1]*Append2_2$y2014+result[12,1]*Append2_2$y2015
> error_term=Append2_2$status-yhat
> theta2=sum((Append2_2$status-yhat)^2)/df
> se1=sqrt(theta2*solve(t(X)%*%X)[1,1])
> se2=sqrt(theta2*solve(t(X)%*%X)[2,2])
> se1
[1] 0.003808214
> se2
[1] 4.839996e-05
> T_linear=result[2,1]/se2
> T_linear

-381.5322
> #which is significant at 1% level
```

For logit model

```
> #For Logit model
> Se_l= sqrt(solve(res2$hessian))


> Se_l[2,2]
[1] 0.0005537805
> T_logit=res2$par[2]/Se_l[2,2]
> T_logit
Append2_2$age
    -224.1728
> #which is significant at 1% level
```

For probit model

```
> #For Probit model
> Se_P= sqrt(solve(res1$hessian))


> Se_P[2,2]
[1] 0.0002445024
> T_probit=res1$par[2]/Se_P[2,2]
> T_probit
Append2_2$age
    -260.0777
> #which is significant at 1% level
```

Exercise 5 Marginal Effffects

1.Compute the marginal effect of the previous probit and logit models.

```
> #1.Compute the marginal effect of the previous probit and logit models.
> pdf1=mean(dnorm(predict(reg5, type = "link")))
> marginal.effects1=pdf1*reg5$coefficients[2]
> marginal.effects1
Append2_2$age
  -0.01568521
>

> pdf2=mean(dlogis(predict(reg6, type = "link")))
> marginal.effects2=pdf2*reg6$coefficients[2]
> marginal.effects2
Append2_2$age
  -0.01600454
```

The marginal effect of probit model is -0.01568521, and the marginal effect of logit model is -0.01600454.

2.Construct the standard errors of the marginal effects. Hint: Boostrap may be the easiest way.

The standard errors of the marginal effects in probit model is 3.168908*e^(-5), the standard errors of the marginal effects in logit model is 3.479323*e^(-5)

For probit model (in the next page)

```
> boot=40
> bootvals <- matrix(rep(NA,boot*12), nrow=boot)
```

```
> set.seed(111)
> for(i in 1:boot){
+   samp1 <- Append2_2[sample(1:dim(Append2_2)[1],replace=T,dim(Append2_2)[1]),]
+   res4=optim(reg5$coefficients,fn=flike,method="BFGS",control=list(trace=5,maxit=100000),x=samp1$age,x1=sa
mp1$y2006,x2=samp1$y2007,x3=samp1$y2008, x4=samp1$y2009,x5=samp1$y2010,x6=samp1$y2011,x7=samp1$y2012,x8=samp
1$y2013,x9=samp1$y2014,x10=samp1$y2015,yvar=samp1$status)
+   yhat=res4$par[1] + res4$par[2]*samp1$age+res4$par[3]*samp1$y2006+res4$par[4]*samp1$y2007+res4$par[5]*sam
p1$y2008+res4$par[6]*samp1$y2009+res4$par[7]*samp1$y2010+res4$par[8]*samp1$y2011+res4$par[9]*samp1$y2012+res
4$par[10]*samp1$y2013+res4$par[11]*samp1$y2014+res4$par[12]*samp1$y2015
+   pdf1=mean(dnorm(yhat))
+
+   bootvals[i,] <- pdf1*res4$par
+
+   }
```

```
+   }
initial  value 80250.121225
final  value 80243.674804
converged
initial  value 79420.752611
iter  10 value 79409.518530
iter  10 value 79409.518257
final  value 79409.351308
converged
initial  value 79485.354861
iter  10 value 79477.761096
iter  10 value 79477.760837
final  value 79477.457223
converged
initial  value 79567.980997
iter  10 value 79562.464609
iter  10 value 79562.464609
iter  10 value 79562.464569
final  value 79562.464569
converged
initial  value 80087.704563
iter  10 value 80084.403229
iter  10 value 80084.403229
final  value 80084.242258
converged
initial  value 79848.734149
iter  10 value 79841.609921
iter  10 value 79841.609846
iter  10 value 79841.608863
final  value 79841.608863
converged
```

```
> sd(bootvals[,2] )
[1] 3.168908e-05
```

For logit model

```
> #Logit
> boot=40
> bootvals1 <- matrix(rep(NA,boot*12), nrow=boot)
> set.seed(111)
> for(i in 1:boot){
+   samp2 <- Append2_2[sample(1:dim(Append2_2)[1],replace=T,dim(Append2_2)[1]),]
+   res5=optim(runif(12,min=-0.1,max=0),fn=flike2,method="BFGS",control=list(trace=5,maxit=100000),
=samp2$age,x1=samp2$y2006,x2=samp2$y2007,x3=samp2$y2008, x4=samp2$y2009,x5=samp2$y2010,x6=samp2$y20
1,x7=samp2$y2012,x8=samp2$y2013,x9=samp2$y2014,x10=samp2$y2015,yvar=samp2$status)
+   yhat=res5$par[1] + res5$par[2]*samp2$age+res5$par[3]*samp2$y2006+res5$par[4]*samp2$y2007+res5$p
r[5]*samp2$y2008+res5$par[6]*samp2$y2009+res5$par[7]*samp2$y2010+res5$par[8]*samp2$y2011+res5$par[9
*samp2$y2012+res5$par[10]*samp2$y2013+res5$par[11]*samp2$y2014+res5$par[12]*samp2$y2015
+   pdf2=mean(dlogis(yhat))
+   bootvals1[i,] = pdf2*res5$par
+ }
```

```
initial  value 473472.691374
iter  10 value 77762.239990
iter  20 value 77306.615007
iter  20 value 77306.614986
iter  20 value 77306.614986
final  value 77306.614986
converged
initial  value 157690.682428
iter  10 value 77597.323689
iter  20 value 76408.908183
final  value 76408.879545
converged
initial  value 191491.421054
iter  10 value 83543.852152
iter  20 value 76439.235504
final  value 76439.191808
converged
initial  value 234774.808691
iter  10 value 140377.510052
iter  20 value 78178.553796
iter  30 value 76555.296244
final  value 76551.285184
converged
initial  value 358519.496321
iter  10 value 78612.757700
iter  20 value 77082.533506
iter  20 value 77082.533498
iter  20 value 77082.533311
final  value 77082.533311
converged
```

```
> sd(bootvals1[,2] )
[1] 3.479323e-05
```