#Exercise1

#1.1 Calculate the correlation between Y and X.

Case1: I only discard the data with NA. (corr=-0.1788512)

```
               Y       varX
Y      1.0000000 -0.1788512
varX -0.1788512  1.0000000
```

Case2: In addition to discarding NA, I also discard the data with wage=0.

(corr = 0.143492)

```
             Y_no0 varX_no0
Y_no0    1.000000 0.143492
varX_no0 0.143492 1.000000
```

#1.2 Calculate the coefficients on this regression.

```
> coefficient1_2
                [,1]
V1       14141.1794
varX_no0   230.9923
```

V1 indicates the coefficient of constant term, and varX_no0 represents the coefficient of age.

#1.3 Calculate the standard errors

#Method 1 (Use the standard formulas of the OLS.)

```
> ster(Y_matrix,X_matrix)
      V1 varX_no0
645.2348  14.8774
```

#Method 2 (Using bootstrap with 49 and 499 replications respectively. Comment on the difference between the two strategies.)

R=49

```
> bootsd1<-apply(boot1,2,sd)
> bootsd1
[1] 554.01799  15.77478
```

R=499

```
> bootsd2
[1] 568.27237  15.21815
```

In terms of bootstrap, as the number of replications increases, standard error gets closed to what we obtained in Method 1. The basic difference between OLS and bootstrap is the latter has a chance of having the same individual's data.

#Exercise2

#2.1 Create a categorical variable ag, which bins the age variables into the following groups: "18-25", "26-30", "31-35", "36-40","41-45", "46-50","51-55", "56-60", and

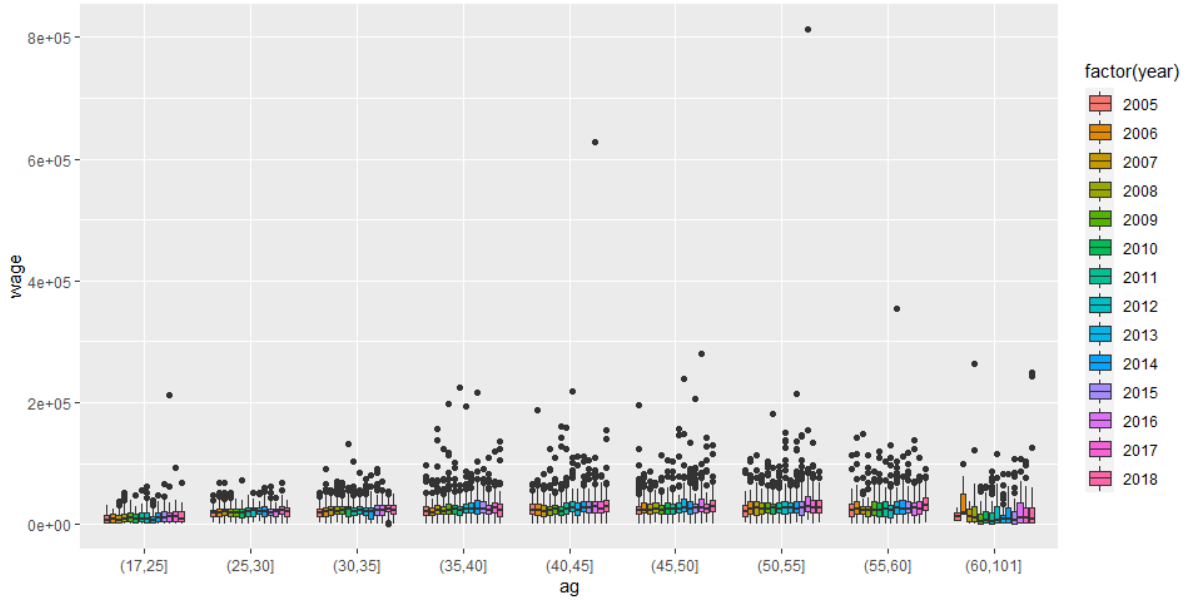
“60+”.

Here, I only showed the first ten rows with ag.

	age	wage	year	ag
1	55	49240	2005	(50,55]
2	50	22166	2005	(45,50]
3	59	10973	2005	(55,60]
4	48	25750	2005	(45,50]
5	45	22742	2005	(40,45]
6	24	12001	2005	(17,25]
7	33	25222	2005	(30,35]
8	54	17110	2005	(50,55]
9	43	18941	2005	(40,45]
10	34	16233	2005	(30,35]

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

Yes. [30,35] and [50,55], [40,45] and [55,60] have the similar trend. Overall, except for two age groups: (17,25] and (60,101], the trend is upward.



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

In order to include a time fixed effect, I choose 2005 as a base year, and add 13 year dummies. In terms of significance, the coefficient of age in two models are both significant based on t-value. Given others constant, on average, the effect of age on wage is higher in this case compared with that in Exercise1. (308.2602>230.9923)

```
> tvalue11_OLS
varX_no0
15.52639
> tvalue22time_OLS
  age
19.29
```

```
> coefficient2_2
                wage
age         308.2602
year_2006  1796.2803
year_2007  1067.6359
year_2008  1301.9965
year_2009  1289.7508
year_2010   886.0537
year_2011  1445.2288
year_2012  3157.9475
year_2013  2528.4423
year_2014  3428.5492
year_2015  3094.5637
year_2016  5539.2546
year_2017  3049.5952
year_2018  3545.8030
constant   9307.4379
```

#Exercise3

#Exclude all individuals who are inactive.

I use filter to get rid of the individual with empstat=Inactive, or empstat=Retired.

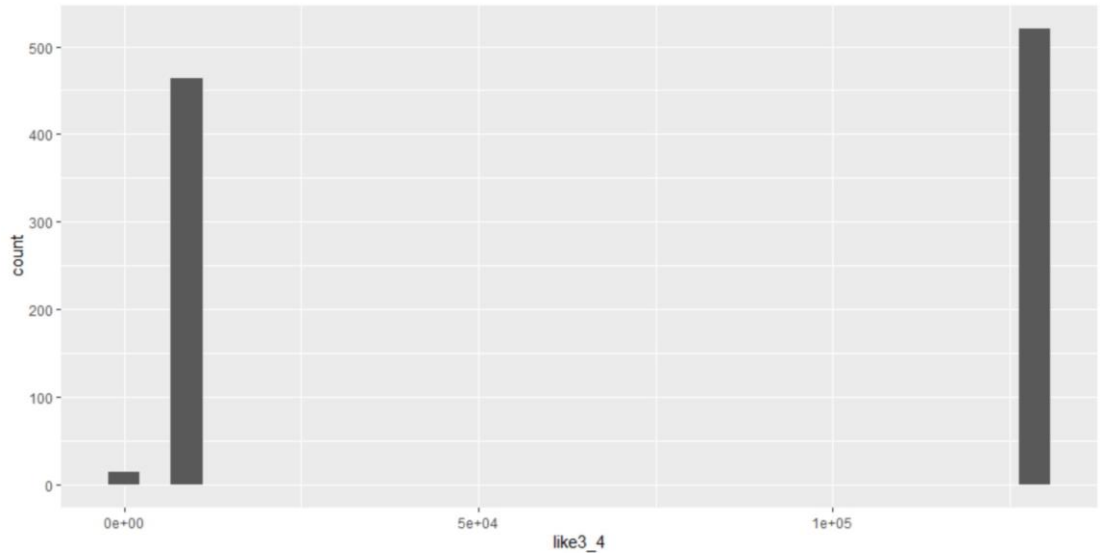
#Write a function that returns the likelihood of the probit of being employed.

```
flike = function(par,x1,x2,yvar)
{xbeta <- par[1]*x1 + par[2]*x2
  prob_XB <- pnorm(xbeta)
  prob_XB[prob_XB>0.999999] = 0.999999
  prob_XB[prob_XB<0.000001] = 0.000001
  return(-sum((1 - yvar) * log(1 - prob_XB) + yvar * log(prob_XB)))
  }
```

Optimize the model and interpret the coefficients. You can use pre-programmed optimization packages.

```
> result3_4[which.min(like3_4),]
[1] 1.09238155 0.01308507
```

Unlike linear regression, the coefficient of probit model cannot be interpreted directly. In other words, only the sign of the coefficient provides us with some useful information. For example, the sign of coefficient age is positive means the individual is likely to be employed if his age increases. The reason why I choose the interval for runif is [-11,11] is because when I enlarge the interval to [-12,12] and try 1000 times. See the following figure.



The frequency on the left column is 15 out of 1000 and the corresponding like3_4 is higher than that of chosen value of [-11,11].

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

No. Because most of the individuals who are employed have the wage>0 and most of the individuals who are unemployed have the wage=0, it represents the high correlation between labor market participation and wage. Although I did find some individuals whose employment status is retired has wage>0, those of them account for the small proportion of the data.

#Exercise 4

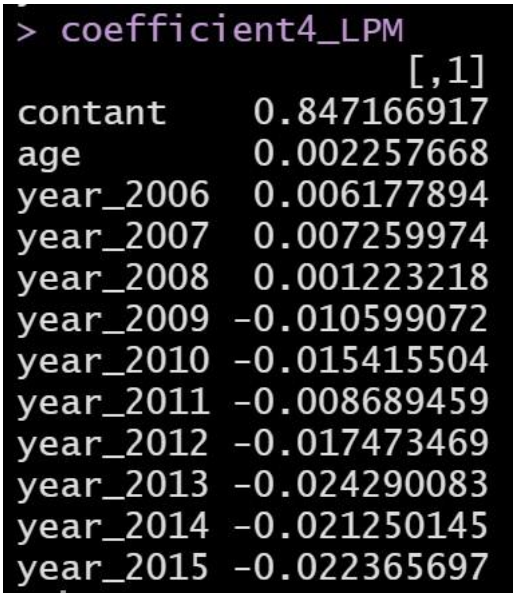
#Exclude all individuals who are inactive.

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

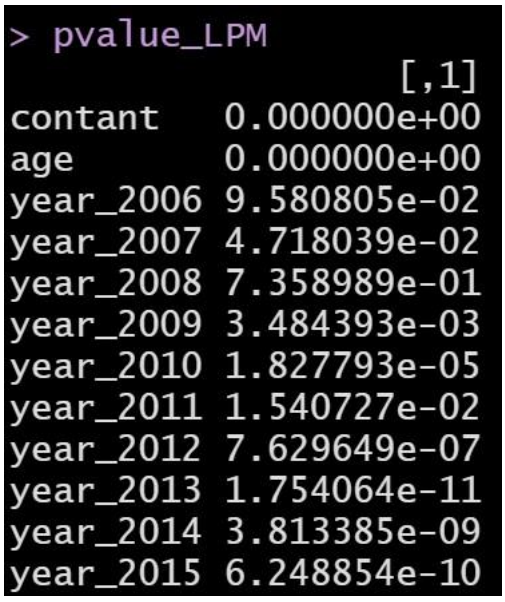
#Part (i) Write LPM

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

<Part1:LPM>



For the interpretation, I choose the estimated coefficient of age. An increase in a year of age, the probability of being employed increase by 0.23% giving others constant.
Here, I choose alpha=0.05, so if the value < alpha, the coefficient is significant.



By the figure above, only year_2006 and year_2008 are not significant.
The coefficient in LPM after optimization.

```
> est_logit
                    prop_sigma  pvalue_logit
 [1,]  1.37054632 0.058399189 8.541033e-122
 [2,]  0.03498805 0.001062756 1.065174e-237
 [3,]  0.11595487 0.062715700  6.447268e-02
 [4,]  0.13763492 0.062032089  2.650240e-02
 [5,]  0.02603787 0.059989797  6.642606e-01
 [6,] -0.16771321 0.057904060  3.774733e-03
 [7,] -0.23823657 0.056799110  2.736291e-05
 [8,] -0.13520219 0.057781923  1.929045e-02
 [9,] -0.26714414 0.055760220  1.659977e-06
[10,] -0.36312949 0.056040912  9.189219e-11
[11,] -0.32301515 0.056504155  1.086370e-08
[12,] -0.33854674 0.056563328  2.160324e-09
```

<Part II: Probit>

```
> est_probit
                    prop_sigma     zvalue pvalue_probit
 [1,]  0.90380562 0.0282506375 31.9923975 1.391036e-224
 [2,]  0.01634174 0.0005046033 32.3853165 4.418327e-230
 [3,]  0.05537559 0.0297337308  1.8623828  6.254915e-02
 [4,]  0.06399474 0.0293656335  2.1792393  2.931389e-02
 [5,]  0.01105627 0.0286319853  0.3861509  6.993849e-01
 [6,] -0.08107122 0.0279604803 -2.8994932  3.737665e-03
 [7,] -0.11624110 0.0275190076 -4.2240294  2.399727e-05
 [8,] -0.06457612 0.0278321249 -2.3202007  2.033002e-02
 [9,] -0.12910514 0.0270289534 -4.7765497  1.783285e-06
[10,] -0.17535581 0.0273111682 -6.4206630  1.356821e-10
[11,] -0.15703454 0.0274437745 -5.7220461  1.052488e-08
[12,] -0.16230746 0.0275061326 -5.9007736  3.618012e-09
```

Year_2006, Year_2008 are insignificant.

Again, we can only utilize the information of the coefficient. Take the coefficient of age for example, a unit increase in age increases the probability of being employed.

<Part III: Logit>

```
> est_logit
                    prop_sigma  pvalue_logit
 [1,]  1.37054632 0.058399189 8.541033e-122
 [2,]  0.03498805 0.001062756 1.065174e-237
 [3,]  0.11595487 0.062715700  6.447268e-02
 [4,]  0.13763492 0.062032089  2.650240e-02
 [5,]  0.02603787 0.059989797  6.642606e-01
 [6,] -0.16771321 0.057904060  3.774733e-03
 [7,] -0.23823657 0.056799110  2.736291e-05
 [8,] -0.13520219 0.057781923  1.929045e-02
 [9,] -0.26714414 0.055760220  1.659977e-06
[10,] -0.36312949 0.056040912  9.189219e-11
[11,] -0.32301515 0.056504155  1.086370e-08
[12,] -0.33854674 0.056563328  2.160324e-09
```

Year_2006, Year_2008 are insignificant.

Again, we can only utilize the information of the coefficient. Take the coefficient of age for example, a unit increase in age increases the probability of being employed. Three methods give the same sign of each estimated coefficient.

#Exercise5

#5.1 Compute the marginal effect of the previous probit and logit models.

Here, I choose to compute MEM (marginal effect at the mean)

Probit model

```
> mem_probit[-1]
 [1]  0.002101242  0.007120265  0.008228527  0.001421629 -0.010424242
 [6] -0.014946430 -0.008303280 -0.016600505 -0.022547476 -0.020191704
[11] -0.020869702
```

Logit model

```
> mem_logit[-1]
 [1]  0.002131621  0.007048055  0.008369474  0.001570753 -0.010225935
 [6] -0.014520037 -0.008246562 -0.016287318 -0.022137346 -0.019690464
[11] -0.020639660
```

#5.2 Construct the standard errors of the marginal effects.

Probit

```
> bootsdmepro5_2_1
 [1] 0.01663978 0.01532462 0.02578576 0.02260384 0.01857693 0.02441091
 [7] 0.01762611 0.01665592 0.02667372 0.02335689 0.02462302 0.01939023
```

Logit

```
> bootsdlog5_2_1
 [1] 0.0278393519 0.0007105635 0.0023504916 0.0027923128 0.0005250439
 [6] 0.0034096562 0.0048407622 0.0027501673 0.0054284367 0.0073810375
[11] 0.0065637612 0.0068816749
```