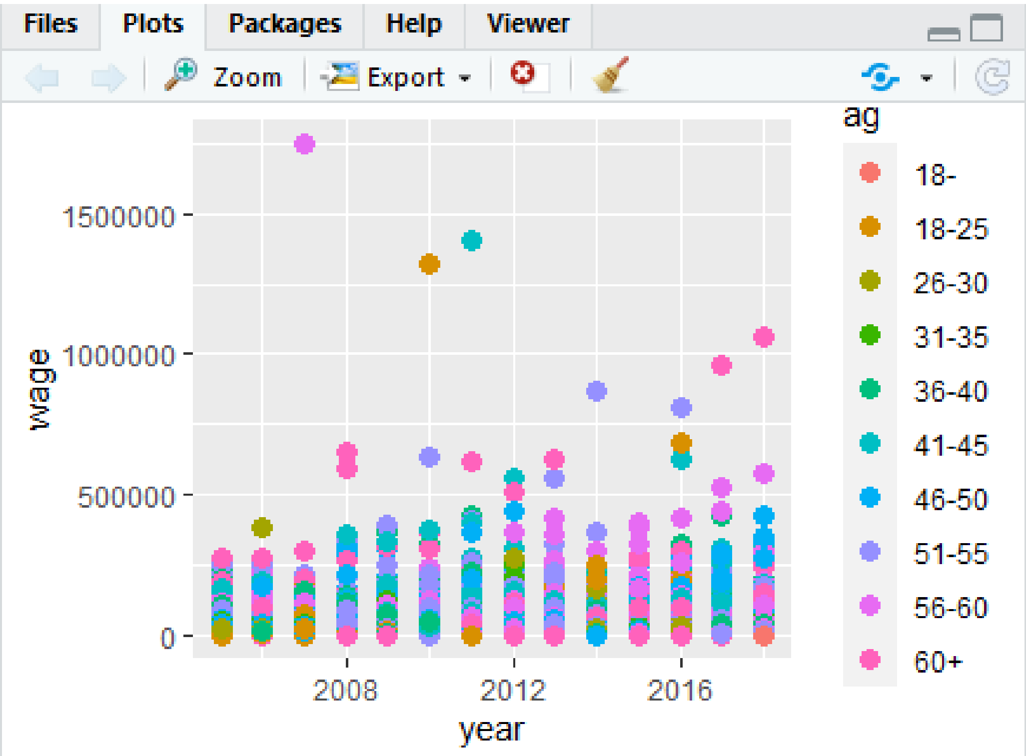
Econ613 HW2

```
setwd("D:/R . Data/hw2/Data")
library(dplyr)
library(data.table)
library(AER)
library(ggplot2)
library(mfx)

# Exercise1
(1)
data1<-fread("datind2009.csv")
data2<-data1[complete.cases(data1[,10])] #Drop NA in wage
Y<-data2$wage
X<-data2$age
cor(Y,X)
##-0.1788512

#or cor = sum((X-mean(X))*(Y-mean(Y))) /(sqrt(sum((X-mean(X))^2))*sqrt(sum((Y-mean(Y))^2)))

(2)
X<- cbind(matrix(1,20232,1),X)    # include intercept
 β <- solve(t(X)%*%X)%*%t(X)%*%Y
 β
##β=-180.1765 , intercept=22075.1

(3)OLS
resid<- Y-X%*%β
sigma2<-as.numeric(t(resid) %*% resid) / (nrow(X) - ncol(X))
sqrt(sigma2)
#18622.31
se_β <-diag(sqrt(sigma2 * solve(t(X) %*% X)))
se_β
#357.8275    ; 6.9687

#bootstrap 49
reg = lm(wage ~ age,data = data2)
R       = 49;                              # number of bootstraps
nind = nrow(X);                    # number of individuals
nvar = length(reg$coefficients)    # number of variables
outs = mat.or.vec(R,nvar)
set.seed(123)
```

```
for (i in 1:R)
{
    samp        = sample(1:nind,nind,rep=TRUE)
    dat_samp = data2[samp,]
    reg1        = lm(wage ~ age,data = dat_samp)
    outs[i,] = reg1$coefficients
}
mean_est = apply(outs,2,mean)
sd_est     = apply(outs,2,sd)
est = cbind(summary(reg)$coefficients[,1],
                    summary(reg)$coefficients[,2],
                    mean_est,
                    sd_est)
colnames(est) = c("CF: est","CF: sd","BT: est","BT: sd")
est
#295.9，  5.68

#bootstrap 499
R      = 499;                                    # number of bootstraps
nind = nrow(X);                        # number of individuals
nvar = length(reg$coefficients)    # number of variables
outs = mat.or.vec(R,nvar)
set.seed(123)
for (i in 1:R)
{
    samp        = sample(1:nind,nind,rep=TRUE)
    dat_samp = data2[samp,]
    reg1        = lm(wage ~ age,data = dat_samp)
    outs[i,] = reg1$coefficients
}
mean_est = apply(outs,2,mean)
sd_est     = apply(outs,2,sd)
est = cbind(summary(reg)$coefficients[,1],
                    summary(reg)$coefficients[,2],
                    mean_est,
                    sd_est)
colnames(est) = c("CF: est","CF: sd","BT: est","BT: sd")
est
#305.614962,   5.364861
#As replications increased, bootstrap results were closer to OLS.
```

```
# exercise2(1)
dati1<-fread("datind2005.csv",colClasses=c(idind="character",idmen="character"))
dati2<-fread("datind2006.csv",colClasses=c(idind="character",idmen="character"))
dati3<-fread("datind2007.csv",colClasses=c(idind="character",idmen="character"))
dati4<-fread("datind2008.csv", colClasses=c(idind="character",idmen="character"))
dati5<-fread("datind2009.csv", colClasses=c(idind="character",idmen="character"))
dati6<-fread("datind2010.csv", colClasses=c(idind="character",idmen="character"))
dati7<-fread("datind2011.csv", colClasses=c(idind="character",idmen="character"))
dati8<-fread("datind2012.csv", colClasses=c(idind="character",idmen="character"))
dati9<-fread("datind2013.csv", colClasses=c(idind="character",idmen="character"))
dati10<-fread("datind2014.csv", colClasses=c(idind="character",idmen="character"))
dati11<-fread("datind2015.csv", colClasses=c(idind="character",idmen="character"))
dati12<-fread("datind2016.csv", colClasses=c(idind="character",idmen="character"))
dati13<-fread("datind2017.csv", colClasses=c(idind="character",idmen="character"))
dati14<-fread("datind2018.csv", colClasses=c(idind="character",idmen="character"))
dat_total<-rbind(dati1,dati2,dati3,dati4,dati5,dati6,dati7,dati8,dati9,dati10,dati11,dati12,
          dati13,dati14)
dat_total<-dat_total[complete.cases(dat_total[,10])]     #drop NA
dat_t=dat_total[,c(4,9,10)]
breaks <- c(0,18,25,30,35,40,45,50,55,60,110)
labels <- c("18-","18-25", "26-30", "31-35", "36-40", "41-45","46-50","51-55", "56-60", "60+");
dat_t[,"ag"] <- cut(dat_t$age, breaks = breaks, labels = labels)
##"ag" in dat_t
#(2)
plot <- ggplot(data=dat_t, aes(x=year, y=wage, color=ag))+geom_point(size=3)
plot
#The income of people aged between 18 and 40 tends to rise year by year.
# Other age groups seems to have no change
```



(3) dat_total$"2005" = as.numeric(dat_total$year==2005)

dat_total$"2006" = as.numeric(dat_total$year==2006)

dat_total$"2007" = as.numeric(dat_total$year==2007)

dat_total$"2008" = as.numeric(dat_total$year==2008)

dat_total$"2009" = as.numeric(dat_total$year==2009)

dat_total$"2010" = as.numeric(dat_total$year==2010)

dat_total$"2011" = as.numeric(dat_total$year==2011)

dat_total$"2012"= as.numeric(dat_total$year==2012)

dat_total$"2013" = as.numeric(dat_total$year==2013)

dat_total$"2014" = as.numeric(dat_total$year==2014)

dat_total$"2015" = as.numeric(dat_total$year==2015)

dat_total$"2016" = as.numeric(dat_total$year==2016)

dat_total$"2017" = as.numeric(dat_total$year==2017)

#Put these dummy variables into A

A = cbind(rep(1,length(dat_total$age)),dat_total$age) %>%

cbind(dat_total$"2005",dat_total$"2006",dat_total$"2007",dat_total$"2008",dat_total$"2009",dat_total$"2010",

dat_total$"2011",dat_total$"2012",dat_total$"2013",dat_total$"2014",dat_total$"2015",

dat_total$"2016",dat_total$"2017")

B = dat_total$wage

β _a<- solve(t(A)%*%A)%*%t(A)%*%B

β _a

```
          [,1]
 [1,] 24311.2098
 [2,]  -186.8793
 [3,] -3636.1515
 [4,] -3614.2143
 [5,] -3341.3490
 [6,] -2210.9609
 [7,] -1915.7910
 [8,] -1766.6265
 [9,] -1520.1339
[10,] -1034.9240
[11,] -1157.3081
[12,]  -886.4765
[13,]  -515.1823
[14,]  -226.0382
[15,]  -157.1196
```

```
#exercise3(1)
data3<-fread("datind2007.csv")
data3<-data3[-which(data3$empstat=="Inactive")]
data3<-data3[-which(data3$empstat=="Retired")]
data3<-data3[complete.cases(data3[,10])]
data3
```

(2)

```
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))}
```

(3)

```
set.seed(123)
x1 = data3$age
yvar = as.numeric(data3$empstat == "Employed")
ntry = 500       #ntry can be larger
out = mat.or.vec(ntry,3)
for (i in 1:ntry){
   start = runif(2,-5,5)
   res = optim(start,fn = flike,method = "BFGS",
         control = list(trace=6,maxit=1000),
         x1 = x1,
         yvar = yvar)
   out[i,c(1,2)] = res$par
   out[i,3] = res$value}
out = data.frame(out)
colnames(out) = c("theta", "bar_age", "likelihood")
out[which(out$likelihood == min(out$likelihood)),]
```

```
       theta      bar_age likelihood
254 1.052278 0.006743782   3545.692
```

Age coefficient is positive. It only means age has positive effects on employment.

```
#(4) We have 2 variables here.
flike1 = function(par,x1,x2,yvar)
{xbeta = par[1] + par[2]*x1 +par[3]*x2
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))}
set.seed(123)
x1 = data3$age
x2=data3$wage
yvar = as.numeric(data3$empstat == "Employed")
ntry = 100        #ntry can be larger
out1 = mat.or.vec(ntry,4)
for (i in 1:ntry){
   start = c(runif(1,-0.1,0.1),runif(1,-0.01,0.01),runif(1,-0.0001,0.0001))
   res = optim(start,fn = flike1,method = "BFGS",
                        control = list(trace=6,maxit=1000),
                        x1 = x1,x2=x2,
                        yvar = yvar)
   out1[i,c(1,2,3)] = res$par
   out1[i,4] = res$value}
out1 = data.frame(out1)
colnames(out1) = c("theta", "bar_age","bar_wage", "likelihood")
out1[which(out1$likelihood == min(out1$likelihood)),]
```

```
        theta     bar_age     bar_wage likelihood
23 0.06147933 0.006254232 7.741589e-05   2814.323
```

#Yes. Both coefficients are positive.

#It shows that both wage and age have a positive effect on employment

```
#Exercise4(1) use data from Exercise2(1)
dat_total1<-rbind(dati1,dati2,dati3,dati4,dati5,
                  dati6,dati7,dati8,dati9,dati10,dati11)

dat_total1<-dat_total1[-which(dat_total1$empstat=="Inactive")]
dat_total1<-dat_total1[-which(dat_total1$empstat=="Retired")]
dat_total1<-dat_total1[complete.cases(dat_total1[,10])]
dat_total1$"2005" = as.numeric(dat_total1$year==2005)
dat_total1$"2006" = as.numeric(dat_total1$year==2006)
dat_total1$"2007" = as.numeric(dat_total1$year==2007)
dat_total1$"2008" = as.numeric(dat_total1$year==2008)
dat_total1$"2009" = as.numeric(dat_total1$year==2009)
dat_total1$"2010" = as.numeric(dat_total1$year==2010)
dat_total1$"2011" = as.numeric(dat_total1$year==2011)
dat_total1$"2012"= as.numeric(dat_total1$year==2012)
dat_total1$"2013" = as.numeric(dat_total1$year==2013)
dat_total1$"2014" = as.numeric(dat_total1$year==2014)
dat_total1
#create year dummy variables

#(2) probit
set.seed(123)
x1 = dat_total1$age
x2 = dat_total1$"2005"
x3 = dat_total1$"2006"
x4 = dat_total1$"2007"
x5 = dat_total1$"2008"
x6 = dat_total1$"2009"
x7 = dat_total1$"2010"
x8 = dat_total1$"2011"
x9 = dat_total1$"2012"
x10 = dat_total1$"2013"
x11 = dat_total1$"2014"
yvar = as.numeric(dat_total1$empstat == "Employed")

flike2 = function(par,x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,yvar)
{xbeta = par[1] + par[2]*x1 +par[3]*x2+par[4]*x3+par[5]*x4+par[6]*x5+
    par[7]*x6+par[8]*x7+par[9]*x8+par[10]*x9+par[11]*x10+par[12]*x11
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))}
```

```
ntry = 10          # I set a small ntry here because my computer runs slowly.
out2 = mat.or.vec(ntry,13)
for (i in 1:ntry){
    start = c(runif(1,-1,1),runif(11,-0.1,0.1))
    res = optim(start,fn = flike2,method = "BFGS",
                         control = list(trace=6,maxit=1000),
                         x1 = x1,x2=x2,x3=x3,x4=x4,x5=x5,x6=x6,x7=x7,x8=x8,
                         x9=x9,x10=x10,x11=x11,
                         yvar = yvar)
out2[i,1:12] = res$par
out2[i,13] = res$value}
out2 = data.frame(out2)
colnames(out2) = c("theta", "bar_age","x2","x3","x4","x5","x6","x7","x8","x9",
                              "x10","x11","likelihood")
out2<-out2[which(out2$likelihood == min(out2$likelihood)),]
out2
```

```
      theta     bar_age         x2         x3        x4        x5         x6         x7        x8         x9
        x10
3 0.6933817 0.01234967 0.05556004 0.07071481 0.1357396 0.1639375 0.08084792 0.07754967 0.1089989 0.06499158
 0.01468691
         x11 likelihood
3 0.02151178   42105.21
```

```
# logit
flike3 = function(par,x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,yvar)
{xbeta = par[1] + par[2]*x1 +par[3]*x2+par[4]*x3+par[5]*x4+par[6]*x5+
    par[7]*x6+par[8]*x7+par[9]*x8+par[10]*x9+par[11]*x10+par[12]*x11
pr    = 1/(1+exp(-x_beta))
pr[pr>0.999999] = 0.999999
pr[pr<0.000001] = 0.000001
like                  = yvar*log(pr) + (1-yvar)*log(1-pr)
return(-sum(like))}
yvar = as.numeric(dat_total1$empstat == "Employed")
ntry = 10        #I set a small ntry here because my computer runs slowly.
out3 = mat.or.vec(ntry,13)
for (i in 1:ntry){
    start = c(runif(1,-1,1),runif(11,-0.1,0.1))
    res = optim(start,fn = flike2,method = "BFGS",
                         control = list(trace=6,maxit=1000),
                         x1 = x1,x2=x2,x3=x3,x4=x4,x5=x5,x6=x6,x7=x7,x8=x8,
                         x9=x9,x10=x10,x11=x11,
                         yvar = yvar)
    out3[i,1:12] = res$par
    out3[i,13] = res$value}
out3 = data.frame(out2)
```

```
colnames(out3) = c("theta", "bar_age","x2","x3","x4","x5","x6","x7","x8","x9",
                   "x10","x11","likelihood")
out3<-out3[which(out3$likelihood == min(out3$likelihood)),]
out3
```

```
      theta    bar_age         x2         x3        x4        x5         x6         x7        x8
3 0.6933817 0.01234967 0.05556004 0.07071481 0.1357396 0.1639375 0.08084792 0.07754967 0.1089989
          x9        x10        x11 likelihood
3 0.06499158 0.01468691 0.02151178   42105.21
>
```

```
#linear

A = cbind(rep(1,length(dat_total1$age)),x1) %>%
  cbind(x2,x3,x4,x5,x6,x7,x8,x9,x10,x11)
B = yvar
β_a1<- solve(t(A)%*%A)%*%t(A)%*%B #
c(β_a1)
```

```
c(β_a1)
[1] 0.786470643 0.002338625 0.011407479 0.013938534 0.025220991 0.029545180 0.015210997 0.014717030
[9] 0.019929195 0.012126947 0.002822538 0.004169451
```

```
(3)#probit
#yvar is dummy variable showing empstat
dat_total2<-as.data.frame(cbind(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,yvar))
glm1<-
glm(yvar~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11,family=binomial(link=probit),data=dat_total2)
summary(glm1)    #
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.6964434  0.0230329  30.237  < 2e-16 ***
x1          0.0122809  0.0004197  29.262  < 2e-16 ***
x2          0.0554654  0.0223276   2.484 0.012986 *
x3          0.0705998  0.0222380   3.175 0.001500 **
x4          0.1356287  0.0224082   6.053 1.43e-09 ***
x5          0.1638175  0.0226315   7.238 4.54e-13 ***
x6          0.0806879  0.0221415   3.644 0.000268 ***
x7          0.0774933  0.0219446   3.531 0.000413 ***
x8          0.1089177  0.0219883   4.953 7.29e-07 ***
x9          0.0649324  0.0214489   3.027 0.002467 **
x10         0.0146201  0.0216909   0.674 0.500299
x11         0.0214798  0.0216801   0.991 0.321802
```

```
#logit
glm2<-
glm(yvar~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11,family=binomial(link=logit),data=dat_total2)
summary(glm2)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) 1.0060444  0.0437042  23.019  < 2e-16 ***
x1          0.0253196  0.0008146  31.082  < 2e-16 ***
x2          0.1162300  0.0428295   2.714 0.006652 ** 
x3          0.1437959  0.0427417   3.364 0.000767 ***
x4          0.2724696  0.0435031   6.263 3.77e-10 ***
x5          0.3259580  0.0441057   7.390 1.46e-13 ***
x6          0.1590113  0.0425783   3.735 0.000188 ***
x7          0.1538549  0.0422063   3.645 0.000267 ***
x8          0.2133826  0.0424773   5.023 5.07e-07 ***
x9          0.1266626  0.0411719   3.076 0.002095 ** 
x10         0.0288622  0.0413770   0.698 0.485463    
x11         0.0425398  0.0414328   1.027 0.304554    
```

#linear

lm1<-lm(yvar~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11,data=dat_total2)

summary(lm1)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 7.865e-01  4.226e-03 186.107  < 2e-16 ***
x1          2.339e-03  7.445e-05  31.414  < 2e-16 ***
x2          1.141e-02  4.047e-03   2.819 0.004821 ** 
x3          1.394e-02  4.014e-03   3.473 0.000516 ***
x4          2.522e-02  3.976e-03   6.343 2.27e-10 ***
x5          2.955e-02  3.985e-03   7.414 1.23e-13 ***
x6          1.521e-02  3.985e-03   3.817 0.000135 ***
x7          1.472e-02  3.951e-03   3.725 0.000195 ***
x8          1.993e-02  3.925e-03   5.077 3.84e-07 ***
x9          1.213e-02  3.871e-03   3.133 0.001733 ** 
x10         2.823e-03  3.959e-03   0.713 0.475850    
x11         4.169e-03  3.945e-03   1.057 0.290544    
```

#The parameters calculated by the three models are different.

##For probit and logit model, positive coefficient only means variables have positive effects on employment. On the contrary, Ols coefficient shows the change of dependent variable when the independent variables change 1 unit.

#βwage is most significant. In different models, some years are significant and some are not.

#Exercise5 (1)

x1_ave = mean(x1)

x2_ave = mean(x2)

x3_ave = mean(x3)

x4_ave = mean(x4)

x5_ave = mean(x5)

x6_ave = mean(x6)

x7_ave = mean(x7)

x8_ave = mean(x8)

x9_ave = mean(x9)

x10_ave = mean(x10)

x11_ave = mean(x11)

x_ave = c(1,x1_ave,x2_ave,x3_ave,x4_ave,x5_ave,x6_ave,
x7_ave,x8_ave,x9_ave,x10_ave,x11_ave)

#marginal effect of probit

β x_ave = sum(out2[,-13] *x_ave)   #out2 is from exercise4,probit

margi_pro = dnorm(βx_ave) * out2[,-13]

margi_pro

```
      theta     bar_age          x2         x3         x4         x5         x6         x7         x8
3 0.1224396 0.002180746 0.009810975 0.01248705 0.02396934 0.02894863 0.01427639 0.01369398 0.01924738
         x9         x10         x11
3 0.01147643 0.002593462 0.003798622
```

#marginal effect of logit

β x_ave1 = sum(out3[,-13] *x_ave)

margi_log = exp(βx_ave1)/(1+exp(βx_ave1))^2 *out3[,-13]

margi_log

```
      theta     bar_age          x2         x3         x4         x5        x6         x7         x8
3 0.1182467 0.002106067 0.009474999 0.01205944 0.02314851 0.02795729 0.0137875 0.01322503 0.01858825
         x9        x10         x11
3 0.01108342 0.00250465 0.003668538
```

#(2)standard error

#use packages "mfx"

#probit standard errors

probitmfx(formula = yvar ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11, data=dat_total2,atmean = TRUE)

```
probitmfx(formula = yvar ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 +
    x8 + x9 + x10 + x11, data = dat_total2, atmean = TRUE)

Marginal Effects:
          dF/dx  Std. Err.       z     P>|z|
x1   2.1682e-03 7.3469e-05 29.5123 < 2.2e-16 ***
x2   9.5065e-03 3.7126e-03  2.5606 0.0104482 *
x3   1.2005e-02 3.6382e-03  3.2998 0.0009677 ***
x4   2.2282e-02 3.4123e-03  6.5301 6.572e-11 ***
x5   2.6504e-02 3.3360e-03  7.9449 1.943e-15 ***
x6   1.3650e-02 3.5842e-03  3.8085 0.0001398 ***
x7   1.3136e-02 3.5671e-03  3.6826 0.0002309 ***
x8   1.8164e-02 3.4552e-03  5.2570 1.464e-07 ***
x9   1.1088e-02 3.5393e-03  3.1327 0.0017319 **
x10  2.5616e-03 3.7714e-03  0.6792 0.4969984
x11  3.7501e-03 3.7425e-03  1.0020 0.3163346
```

Std.Err

#logit standard errors

logitmfx(formula = yvar ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11, data=dat_total2,atmean = TRUE)

```
logitmfx(formula = yvar ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 +
    x8 + x9 + x10 + x11, data = dat_total2, atmean = TRUE)

Marginal Effects:
         dF/dx  Std. Err.        z     P>|z|
x1  2.2778e-03 7.1467e-05 31.8717 < 2.2e-16 ***
x2  1.0060e-02 3.5635e-03  2.8230 0.0047570 **
x3  1.2336e-02 3.4925e-03  3.5323 0.0004120 ***
x4  2.2425e-02 3.2612e-03  6.8764 6.140e-12 ***
x5  2.6362e-02 3.1865e-03  8.2731 < 2.2e-16 ***
x6  1.3578e-02 3.4457e-03  3.9406 8.129e-05 ***
x7  1.3165e-02 3.4301e-03  3.8380 0.0001241 ***
x8  1.7917e-02 3.3201e-03  5.3966 6.793e-08 ***
x9  1.0943e-02 3.4129e-03  3.2064 0.0013441 **
x10 2.5721e-03 3.6525e-03  0.7042 0.4813132
x11 3.7742e-03 3.6250e-03  1.0412 0.2978011
```

Std.Err