Homework 4

Minxuan Wang

November 2017

Question 1

a.

```
data.cost<-read.table("TableF4-4.txt",head=TRUE)
attach(data.cost)

## Compute each terms
lnCPf<-log(cost/pf)
lnQ<-log(q)
lnQ2<-0.5*(lnQ)^2
lnPkPf<-log(pk/pf)
lnPlPf<-log(pl/pf)

## Linear regression
fitlm_cost<-lm(lnCPf~lnQ+lnQ2+lnPkPf+lnPlPf)
summary(fitlm_cost)

##
## Call:
## lm(formula = lnCPf ~ lnQ + lnQ2 + lnPkPf + lnPlPf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42576 -0.08891 -0.00223  0.08404  0.37363
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.818163   0.252439 -27.009  < 2e-16 ***
## lnQ          0.402745   0.031483  12.792  < 2e-16 ***
## lnQ2         0.060895   0.004325  14.079  < 2e-16 ***
## lnPkPf       0.162034   0.040406   4.010 9.46e-05 ***
## lnPlPf       0.152445   0.046597   3.272  0.00132 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1378 on 153 degrees of freedom
## Multiple R-squared:  0.9922, Adjusted R-squared:  0.992
## F-statistic:  4880 on 4 and 153 DF,  p-value: < 2.2e-16
```

```
## Asymptotic covariance matrix
cov_matrix<-vcov(fitlm_cost)
cov_matrix
```

```
##               (Intercept)           lnQ          lnQ2        lnPkPf
## (Intercept)  0.0637255474 -0.0023838181  3.104204e-04  3.994585e-03
## lnQ         -0.0023838181  0.0009911868 -1.335824e-04  1.002255e-04
## lnQ2         0.0003104204 -0.0001335824  1.870819e-05 -1.493338e-05
## lnPkPf       0.0039945854  0.0001002255 -1.493338e-05  1.632609e-03
## lnPlPf      -0.0104712922 -0.0001995679  2.453652e-05 -1.019813e-03
##                    lnPlPf
## (Intercept) -1.047129e-02
## lnQ         -1.995679e-04
## lnQ2         2.453652e-05
## lnPkPf      -1.019813e-03
## lnPlPf       2.171313e-03
```

b.

```
## Compute delta f
delta_f<-1-coefficients(fitlm_cost)[4]-coefficients(fitlm_cost)[5]
delta_f
```

```
##    lnPkPf
## 0.6855215
```

```
estmean<-coef(fitlm_cost)[4:5]
estvar<-vcov(fitlm_cost)[4:5,4:5]


## Estimate the asymptotic standard error
library("msm")
```

```
## Warning: package 'msm' was built under R version 3.4.2
```

```
deltamethod(~1-x1-x2,estmean,estvar)
```

```
## [1] 0.04200352
```

c.

```
beta<-coefficients(fitlm_cost)[2]
gamma<-coefficients(fitlm_cost)[3]
est.Q<-exp((1-beta)/gamma)
est.Q
```

```
##     lnQ
## 18177.1
```

```
est.Q_mean<-coef(fitlm_cost)[2:3]
est.Q_var<-vcov(fitlm_cost)[2:3,2:3]


## Standard error
se.Q<-deltamethod(~exp((1-x1)/x2),est.Q_mean,est.Q_var)
```

```
lowerbound<-est.Q-qnorm(0.975)*se.Q
upperbound<-est.Q+qnorm(0.975)*se.Q

## 95% confidence interval
IC<-c(lowerbound,upperbound)
IC

##      lnQ      lnQ
## 10537.96 25816.25
```

d.

```
## Pick out the firms sets
firms1<-subset(data.cost,data.cost$q>=lowerbound)
firms2<-subset(firms1,firms1$q<=upperbound)

## Compute the number of firms that reached the efficient scale
length(firms2$q)

## [1] 28
```

Question 2

a.

```
setwd("D:/Econ 403A/Homework 4")
merged.data<-read.csv("Koop-Tobias.csv") # Get from NYU Stern

## Define the variables name in R
educ<-merged.data$EDUC
logwage<-merged.data$LOGWAGE
potexper<-merged.data$POTEXPER
ability<-merged.data$ABILITY
mothered<-merged.data$MOTHERED
fathered<-merged.data$FATHERED
brknhome<-merged.data$BRKNHOME
siblings<-merged.data$SIBLINGS

## Linear regression
lm1<-lm(logwage~educ+potexper+ability)
lm1

##
## Call:
## lm(formula = logwage ~ educ + potexper + ability)
##
## Coefficients:
## (Intercept)         educ     potexper      ability
##     1.02723      0.07376      0.03949      0.08289
```

```
lm2<-lm(logwage~-1+mothered+fathered+brknhome+siblings)
lm2

##
## Call:
## lm(formula = logwage ~ -1 + mothered + fathered + brknhome +
##     siblings)
##
## Coefficients:
## mothered  fathered  brknhome  siblings
##  0.11735   0.04222   0.03219   0.11696
```

b/c.

The F-test is the statistic for the hypothesis test with null hypothesis and alternate hypothesis:
H0: All non-constant coefficients in the regression equation are zero
Ha: At least one of the non-constant coefficients in the regression equation is non-zero.

```
summary(lm1)

##
## Call:
## lm(formula = logwage ~ educ + potexper + ability)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.52891 -0.27558  0.02441  0.30914  2.13659
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.0272291  0.0300415   34.19   <2e-16 ***
## educ        0.0737621  0.0022143   33.31   <2e-16 ***
## potexper    0.0394896  0.0008984   43.96   <2e-16 ***
## ability     0.0828907  0.0046000   18.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4803 on 17915 degrees of freedom
## Multiple R-squared:  0.1734, Adjusted R-squared:  0.1733
## F-statistic:  1253 on 3 and 17915 DF,  p-value: < 2.2e-16

summary(lm2)

##
## Call:
## lm(formula = logwage ~ -1 + mothered + fathered + brknhome +
##     siblings)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -2.7575 -0.3187  0.0890  0.4880  3.0361
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## mothered 0.117348   0.001878  62.474   <2e-16 ***
## fathered 0.042223   0.001760  23.990   <2e-16 ***
## brknhome 0.032193   0.013629   2.362   0.0182 *
## siblings 0.116962   0.002011  58.148   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6609 on 17915 degrees of freedom
## Multiple R-squared:  0.9214, Adjusted R-squared:  0.9214
## F-statistic: 5.248e+04 on 4 and 17915 DF,  p-value: < 2.2e-16
```

p-value: < 2.2e-16, which means that we reject the H0, model has predictive capability.

Numerically, we can the defination of F statistic (using the first model as an example):

```
anova(lm1)

## Analysis of Variance Table
##
## Response: logwage
##              Df Sum Sq Mean Sq F value    Pr(>F)
## educ          1  385.5  385.51 1671.19 < 2.2e-16 ***
## potexper      1  406.7  406.67 1762.92 < 2.2e-16 ***
## ability       1   74.9   74.91  324.72 < 2.2e-16 ***
## Residuals 17915 4132.6    0.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## SS(Regression)=SS(Total)-S(Residual)

## Get the SST value
SST_1<-var(logwage)*(nrow(merged.data)-1)

## Get the SSE value
SSE_1<-sum(lm1$residual^2)

## Get the SSR value
SSR_1<-SST_1-SSE_1

## Get the degree of freedom
dfE_1<-lm1$df.residual
dfReg_1<-nrow(merged.data)-1-dfE_1
MSreg_1<-SSR_1/dfReg_1
```

```r
MSE_1<-SSE_1/dfE_1
Fstat_1<-MSreg_1/MSE_1
pvalue_1<-pf(Fstat_1,dfReg_1,dfE_1,lower.tail=FALSE)
```

d. Wald test

```r
library(survey)

## Warning: package 'survey' was built under R version 3.4.2

## Loading required package: grid

## Loading required package: Matrix

## Loading required package: survival

##
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':
##
##     dotchart

regTermTest(lm2,"mothered")

## Wald test for mothered
##  in lm(formula = logwage ~ -1 + mothered + fathered + brknhome +
##     siblings)
## F =  3902.977  on  1  and  17915  df: p= < 2.22e-16

regTermTest(lm2,"fathered")

## Wald test for fathered
##  in lm(formula = logwage ~ -1 + mothered + fathered + brknhome +
##     siblings)
## F =  575.5141  on  1  and  17915  df: p= < 2.22e-16

regTermTest(lm2,"brknhome")

## Wald test for brknhome
##  in lm(formula = logwage ~ -1 + mothered + fathered + brknhome +
##     siblings)
## F =  5.579391  on  1  and  17915  df: p= 0.018184

regTermTest(lm2,"siblings")

## Wald test for siblings
##  in lm(formula = logwage ~ -1 + mothered + fathered + brknhome +
##     siblings)
## F =  3381.22  on  1  and  17915  df: p= < 2.22e-16
```

Question 3

(i).

```
load("D:/Econ 403A/Homework 4/401ksubs.RData")
attach(data)
nettfa<-data$nettfa
mean(nettfa)

## [1] 19.07168

sd(nettfa)

## [1] 63.96384

max(nettfa)

## [1] 1536.798

min(nettfa)

## [1] -502.302
```

(ii).

```
## T test
nettfa_0<-subset(nettfa,data$e401k==0)
nettfa_1<-subset(nettfa,data$e401k==1)
t.test(nettfa_0,nettfa_1,
alternative="two.side",
paired=FALSE,
var.equal=FALSE,
conf.level=.95)

##
##  Welch Two Sample t-test
##
## data:  nettfa_0 and nettfa_1
## t = -13.099, df = 6072.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -21.68060 -16.03604
## sample estimates:
## mean of x mean of y
##  11.67677  30.53509
```

From the result of t.test, p-value < 2.2e-16, which means we should reject the H0: the average nettfa does not differ by 401(k) eligibility status.
The dollar amount difference is:

```
mean(nettfa_1)-mean(nettfa_0)

## [1] 18.85832
```

(iii).

```
e401k<-data$e401k
inc2<-incsq
age2<-agesq
fitlm=lm(nettfa~inc+inc2+age+age2+e401k)
summary(fitlm)

##
## Call:
## lm(formula = nettfa ~ inc + inc2 + age + age2 + e401k)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -516.66  -15.63   -3.27    6.05 1464.79
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.0852038  9.9597050   2.318 0.020479 *
## inc         -0.2784651  0.0745386  -3.736 0.000188 ***
## inc2         0.0102601  0.0005869  17.481  < 2e-16 ***
## age         -1.9718860  0.4833774  -4.079 4.55e-05 ***
## age2         0.0347637  0.0055487   6.265 3.89e-10 ***
## e401k        9.7046880  1.2774063   7.597 3.32e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57.16 on 9269 degrees of freedom
## Multiple R-squared:  0.2018, Adjusted R-squared:  0.2014
## F-statistic: 468.7 on 5 and 9269 DF,  p-value: < 2.2e-16
```

From the p-value, this regression model is statistically significant. So the linear equations are:
when e401k=1, $nettfa = -23.24 + 0.008109 \times incsq + 0.01221 \times agesq + 8.166$
when e401k=0, $nettfa = -23.24 + 0.008109 \times incsq + 0.01221 \times agesq$
The estimated dollar effect of 401(k) eligibility is 8.166

(iv).

```
age_41<-age-41

## Regress the model with interaction term
fitlm_2=lm(nettfa~inc+inc2+age+age2+I(e401k*age_41)+e401k)
summary(fitlm_2)

##
## Call:
## lm(formula = nettfa ~ inc + inc2 + age + age2 + I(e401k * age_41) +
##     e401k)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -518.62  -14.96    -2.51     4.26 1460.05
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          32.3630674 10.1011940   3.204  0.00136 **
## inc                  -0.2789400  0.0744320  -3.748  0.00018 ***
## inc2                  0.0102339  0.0005861  17.461  < 2e-16 ***
## age                  -2.2068552  0.4847558  -4.553 5.37e-06 ***
## age2                  0.0349726  0.0055409   6.312 2.89e-10 ***
## I(e401k * age_41)     0.6379022  0.1214841   5.251 1.55e-07 ***
## e401k                 9.5846941  1.2757839   7.513 6.32e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57.08 on 9268 degrees of freedom
## Multiple R-squared:  0.2042, Adjusted R-squared:  0.2037
## F-statistic: 396.3 on 6 and 9268 DF,  p-value: < 2.2e-16
```

The interaction term is significant because p-value=1.55e-07(t=5.251), the coeffecient is 0.638.

(v).

The coefficient on e401k at age 41 in these two regressions are 9.705(in part iii the influences at all ages are same) and 9.585, it doesn't differ a lot.

(vi).

```
## Define the dummy virables
fsize1<-as.numeric(fsize==1)
fsize2<-as.numeric(fsize==2)
fsize3<-as.numeric(fsize==3)
fsize4<-as.numeric(fsize==4)
fsize5<-as.numeric(fsize>=5)

## Add dummy virables to regression model
fitlm_3<-lm(nettfa~inc+inc2+age+age2+e401k+fsize5+fsize2+fsize3+fsize4)
summary(fitlm_3)

##
## Call:
## lm(formula = nettfa ~ inc + inc2 + age + age2 + e401k + fsize5 +
##     fsize2 + fsize3 + fsize4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -517.55  -16.09    -3.16     6.48 1461.84
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 16.3366003 10.1156944   1.615 0.106350
## inc         -0.2398681  0.0754935  -3.177 0.001491 **
## inc2         0.0100454  0.0005894  17.042  < 2e-16 ***
## age         -1.4948962  0.4946402  -3.022 0.002516 **
## age2         0.0289958  0.0056991   5.088 3.69e-07 ***
## e401k        9.4552262  1.2778223   7.399 1.49e-13 ***
## fsize5      -7.3608890  2.1006137  -3.504 0.000460 ***
## fsize2      -0.8589355  1.8180426  -0.472 0.636616
## fsize3      -4.6651683  1.8768488  -2.486 0.012949 *
## fsize4      -6.3147522  1.8679912  -3.381 0.000727 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57.11 on 9265 degrees of freedom
## Multiple R-squared:  0.2037, Adjusted R-squared:  0.2029
## F-statistic: 263.3 on 9 and 9265 DF,  p-value: < 2.2e-16
```

(vii).

```
## Define five conditions
data_fsize1<-subset.data.frame(data,fsize1=="1")
data_fsize2<-subset.data.frame(data,fsize2=="1")
data_fsize3<-subset.data.frame(data,fsize3=="1")
data_fsize4<-subset.data.frame(data,fsize4=="1")
data_fsize5<-subset.data.frame(data,fsize5=="1")

## Run the regression
unreg.1<-lm(nettfa~inc+incsq+age+agesq+e401k,data=data_fsize1)
unreg.2<-lm(nettfa~inc+incsq+age+agesq+e401k,data=data_fsize2)
unreg.3<-lm(nettfa~inc+incsq+age+agesq+e401k,data=data_fsize3)
unreg.4<-lm(nettfa~inc+incsq+age+agesq+e401k,data=data_fsize4)
unreg.5<-lm(nettfa~inc+incsq+age+agesq+e401k,data=data_fsize5)

## review the regression results
anova(unreg.1)
```

```
## Analysis of Variance Table
##
## Response: nettfa
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## inc          1  377482  377482 190.5175 < 2.2e-16 ***
## incsq        1     138     138   0.0698  0.791614
## age          1  167370  167370  84.4727 < 2.2e-16 ***
## agesq        1   16133   16133   8.1426  0.004368 **
## e401k        1   20343   20343  10.2675  0.001375 **
## Residuals 2011 3984498    1981
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(unreg.2)
```

```
## Analysis of Variance Table
##
## Response: nettfa
##             Df   Sum Sq Mean Sq  F value    Pr(>F)
## inc          1  2226203 2226203 419.9206 < 2.2e-16 ***
## incsq        1   417545  417545  78.7599 < 2.2e-16 ***
## age          1   648380  648380 122.3016 < 2.2e-16 ***
## agesq        1    28296   28296   5.3374   0.02097 *
## e401k        1    86407   86407  16.2987 5.596e-05 ***
## Residuals 2193 11626157    5301
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(unreg.3)
```

```
## Analysis of Variance Table
##
## Response: nettfa
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## inc          1 1082008 1082008 293.3108 < 2.2e-16 ***
## incsq        1  172252  172252  46.6939 1.128e-11 ***
## age          1  156847  156847  42.5181 9.042e-11 ***
## agesq        1   10535   10535   2.8558  0.091215 .
## e401k        1   25154   25154   6.8187  0.009095 **
## Residuals 1823 6724953    3689
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(unreg.4)
```

```
## Analysis of Variance Table
##
## Response: nettfa
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## inc          1 1174297 1174297 469.6041 < 2.2e-16 ***
## incsq        1  174827  174827  69.9137 < 2.2e-16 ***
## age          1   84461   84461  33.7760 7.188e-09 ***
## agesq        1     998     998   0.3993 0.5275255
## e401k        1   27402   27402  10.9582 0.0009486 ***
## Residuals 1984 4961213    2501
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(unreg.5)
```

```
## Analysis of Variance Table
##
## Response: nettfa
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## inc          1  543363  543363 249.3919 < 2.2e-16 ***
## incsq        1  133188  133188  61.1306 1.139e-14 ***
```

```
## age           1   12314   12314   5.6520 0.0175870 *
## agesq         1       7       7   0.0034 0.9535194
## e401k         1   31829   31829  14.6090 0.0001389 ***
## Residuals 1234 2688580    2179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fitlm_3)
```

```
## Analysis of Variance Table
##
## Response: nettfa
##             Df   Sum Sq Mean Sq   F value    Pr(>F)
## inc          1  5381009 5381009 1649.9986 < 2.2e-16 ***
## inc2         1   936033  936033  287.0193 < 2.2e-16 ***
## age          1  1043721 1043721  320.0400 < 2.2e-16 ***
## age2         1   107848  107848   33.0698 9.173e-09 ***
## e401k        1   188589  188589   57.8278 3.140e-14 ***
## fsize5       1    19656   19656    6.0272 0.0141055 *
## fsize2       1    11439   11439    3.5075 0.0611213 .
## fsize3       1     2618    2618    0.8028 0.3702758
## fsize4       1    37269   37269   11.4278 0.0007265 ***
## Residuals 9265 30215207    3261
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Calculate sum of squared residuals for each regression
SSR_ur<-sum(anova(unreg.1)[6,2],anova(unreg.2)[6,2],anova(unreg.3)[6,2],
anova(unreg.4)[6,2],anova(unreg.5)[6,2])
SSR_ur
```

```
## [1] 29985400
```

```
SSR_r<-anova(fitlm_3)[10,2]
SSR_r
```

```
## [1] 30215207
```

```
## Computing the Chow test statistic (F-test)
Chow.F.statistic<-((SSR_r-SSR_ur)/SSR_ur)*(9245/20)
Chow.F.statistic
```

```
## [1] 3.542674
```

```
## Calculate P-value
1-pf(Chow.F.statistic,20,9245)
```

```
## [1] 1.424927e-07
```

From the result we can see the p-value is essentially zero. In this case, there is strong evidence that the slopes change across family size.

Question 4

(i).

```
## Estimate simple linear probability model
fitlm_e401k=lm(e401k~inc+inc2+age+age2+male,data=data)
summary(fitlm_e401k)

##
## Call:
## lm(formula = e401k ~ inc + inc2 + age + age2 + male, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6970 -0.3719 -0.2149  0.4870  0.9155
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.063e-01  8.110e-02  -6.243 4.48e-10 ***
## inc          1.245e-02  5.929e-04  20.993  < 2e-16 ***
## inc2        -6.165e-05  4.732e-06 -13.028  < 2e-16 ***
## age          2.651e-02  3.922e-03   6.758 1.49e-11 ***
## age2        -3.053e-04  4.501e-05  -6.782 1.26e-11 ***
## male        -3.533e-03  1.208e-02  -0.292     0.77
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4648 on 9269 degrees of freedom
## Multiple R-squared:  0.09428,    Adjusted R-squared:  0.09379
## F-statistic:   193 on 5 and 9269 DF,  p-value: < 2.2e-16

library(RCurl)

## Loading required package: bitops

## Import the function
url_robust<-"https://raw.githubusercontent.com/IsidoreBeautrelet/econom
ictheoryblog/master/robust_summary.R"
eval(parse(text=getURL(url_robust,ssl.verifypeer=FALSE)),envir=.GlobalE
nv)

## Use new summary function
summary(fitlm_e401k,robust=TRUE)

##
## Call:
## lm(formula = e401k ~ inc + inc2 + age + age2 + male, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6970 -0.3719 -0.2149  0.4870  0.9155
```

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.063e-01  7.855e-02  -6.445 1.21e-10 ***
## inc          1.245e-02  6.003e-04  20.734  < 2e-16 ***
## inc2        -6.165e-05  5.004e-06 -12.320  < 2e-16 ***
## age          2.651e-02  3.823e-03   6.932 4.41e-12 ***
## age2        -3.053e-04  4.375e-05  -6.977 3.22e-12 ***
## male        -3.533e-03  1.205e-02  -0.293    0.769
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4648 on 9269 degrees of freedom
## Multiple R-squared:  0.09428,    Adjusted R-squared:  0.09379
## F-statistic: 209.5 on 5 and 9269 DF,  p-value: < 2.2e-16
```

All the parameters are statistical significant:

$e4\hat{0}1k = -0.506 + 0.0124 \times inc - 0.000062 \times inc^2 + 0.0265 \times age - 0.00031 \times age^2 - 0.0035 \times male$

From the two summaries, we can see the Std.Error are almostly the same. So there are no important differences.

(ii).

Notice that the approximate estimator of the random error term μ_i is expressed by the residual e_i, such that we get:

$$Var(\mu_i) = E(\mu_i^2) \approx e_i^2$$

$$e_i = Y_i - (\hat{Y_i})_{ols}$$

We can write this as a regression model in a simple way.

$$e_i^2 = \alpha_0 + \alpha_1 X_i + \alpha_2 X^2 + v$$

The restrictions are $\alpha_0 = 0$,$\alpha_1 = 1$, and $\alpha_2 = -1$. In the oringinal linear probability model:

$$\hat{Y_i} = \beta_0 + \beta_i X_i$$

So, when we run the regression e_i^2 on $\hat{y}_i$ and $\hat{y_i^2}$, the intercept estimates should be close to zero, the coefficient on $\hat{y}_i$ should be close to 1, and the coefficient on $\hat{y_i^2}$ should be close to −1.

(iii).

```
## Get the residual squared sequence
u2<-fitlm_e401k$residuals^2

## Do the linear regression ui2 on yi and yi2
y<-fitted(fitlm_e401k)
summary(lm(u2~y+I(y^2)))
```

```
##
## Call:
## lm(formula = u2 ~ y + I(y^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13158 -0.11178 -0.07017  0.06353  0.76870
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.009033   0.010915  -0.828    0.408
## y            1.009682   0.057717  17.494   <2e-16 ***
## I(y^2)      -0.970286   0.069728 -13.915   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.154 on 9272 degrees of freedom
## Multiple R-squared:  0.06274,    Adjusted R-squared:  0.06254
## F-statistic: 310.3 on 2 and 9272 DF,  p-value: < 2.2e-16
```

The White F statistic is about 310.3, which is very significant. The coefficient on $e4\hat{0}1k$ and $e4\hat{0}1k^2$ is 1.0097 and -0.9703, the intercept is -0.009. The coefficient estimates roughly correspond to the theoretical values described in part (ii).

(iv).

```
## Compute the upper bound and lower bound of fitted values
max(y)
```

```
## [1] 0.6971899
```

```
min(y)
```

```
## [1] 0.02991716
```

```
## Fit a WLS model using weights=1/(fitted values)
fitlm_e401k.wls=lm(e401k~inc+inc2+age+age2+male,data=data,weights=1/y)
summary(fitlm_e401k.wls)
```

```
##
## Call:
## lm(formula = e401k ~ inc + inc2 + age + age2 + male, data = data,
##     weights = 1/y)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8569 -0.6056 -0.4495  0.6743  3.1585
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.841e-01  7.089e-02  -6.829 9.08e-12 ***
## inc          1.277e-02  5.192e-04  24.588  < 2e-16 ***
## inc2        -6.166e-05  3.963e-06 -15.557  < 2e-16 ***
## age          2.500e-02  3.489e-03   7.165 8.38e-13 ***
## age2        -2.901e-04  3.992e-05  -7.269 3.93e-13 ***
## male        -3.239e-03  1.127e-02  -0.287    0.774
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7757 on 9269 degrees of freedom
## Multiple R-squared:  0.103,  Adjusted R-squared:  0.1025
## F-statistic: 212.8 on 5 and 9269 DF,  p-value: < 2.2e-16
```

$\widehat{e401k} = -0.4841 + 0.01277 \times inc - 0.000062 \times inc^2 + 0.025 \times age - 0.00029 \times age^2 - 0.00324 \times male$

They doesn't differ in important ways from the OLS estimates.

Question 5

(i).

```
fitlm2<-lm(nettfa~inc+inc2+age+age2+male+e401k)
summary(fitlm2)
```

```
##
## Call:
## lm(formula = nettfa ~ inc + inc2 + age + age2 + male + e401k)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -516.00  -15.84   -3.19    6.09 1465.14
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 21.1977926  9.9922112   2.121 0.033912 *
## inc         -0.2702243  0.0746105  -3.622 0.000294 ***
## inc2         0.0102160  0.0005871  17.400  < 2e-16 ***
## age         -1.9397708  0.4834769  -4.012 6.06e-05 ***
## age2         0.0345662  0.0055482   6.230 4.86e-10 ***
## male         3.3690485  1.4858129   2.267 0.023384 *
## e401k        9.7134817  1.2771269   7.606 3.11e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57.15 on 9268 degrees of freedom
## Multiple R-squared:  0.2022, Adjusted R-squared:  0.2017
## F-statistic: 391.6 on 6 and 9268 DF,  p-value: < 2.2e-16
```

$\hat{nettfa} = 21.198 - 0.27 \times inc + 0.0102 \times inc^2 - 1.940 \times age + 0.0346 \times age^2 + 3.369 \times male + 9.713 \times e401k$

The coefficient on e401k is 9.713, which means when other terms are fixed, the mean of net financial assets of a family with e401k=1 is about 9713 greater than the family with e401k=0.

(ii).

Same as the previous question, we pick out the residuals first and then do the regression of $\hat{\mu}_i^2$ on inc, inc^2, age, age^2, $male$ and $e401k$.

```
## Define the square of residuals
u2_2<-fitlm2$residuals^2

## Regress the linear model
summary(lm(u2_2~inc+inc2+age+age2+male+e401k))
```

```
##
## Call:
## lm(formula = u2_2 ~ inc + inc2 + age + age2 + male + e401k)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -119297   -2835    -615     816 2133263
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14762.8872  7708.2376   1.915   0.0555 .
## inc          -433.6569    57.5564  -7.534 5.36e-14 ***
## inc2            5.7980     0.4529  12.801  < 2e-16 ***
## age          -525.2654   372.9660  -1.408   0.1591
## age2            8.0599     4.2800   1.883   0.0597 .
## male          928.3227  1146.1926   0.810   0.4180
## e401k         399.3052   985.2071   0.405   0.6853
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44090 on 9268 degrees of freedom
## Multiple R-squared:  0.03737,	Adjusted R-squared:  0.03675
## F-statistic: 59.97 on 6 and 9268 DF,  p-value: < 2.2e-16
```

R^2=0.0374, F-Statistic is 59.97, p-value: < 2.2e-16. So this model could have heteroskedasticity, which means given the explanatory variablees, the variance of series error is not equal to 0.

(iii).

```
library(L1pack)

## Warning: package 'L1pack' was built under R version 3.4.2

fitlad<-lad(nettfa~inc+inc2+age+age2+male+e401k)
summary(fitlad)

## Call:
## lad(formula = nettfa ~ inc + inc2 + age + age2 + male + e401k)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -506.360   -5.576    0.000    8.925 1501.617
##
## Coefficients:
##             Estimate Std.Error Z value p-value
## (Intercept) 12.4912   7.0686    1.7671  0.0772
## inc         -0.2616   0.0528   -4.9559  0.0000
## inc2         0.0071   0.0004   17.0604  0.0000
## age         -0.7227   0.3420   -2.1130  0.0346
## age2         0.0111   0.0039    2.8214  0.0048
## male         1.0188   1.0511    0.9693  0.3324
## e401k        3.7373   0.9035    4.1367  0.0000
##
## Degrees of freedom: 9275 total; 9268 residual
## Scale estimate: 28.58673
## Log-likelihood: -43588.01 on 8 degrees of freedom
```

The lad estimate model is:

$$net\hat{t}fa = 12.4912 - 0.2616 \times inc + 0.0071 \times inc^2 - 0.7227 \times age + 0.0111 \times age^2 + 1.0188 \times male + 3.7373 \times e401k$$

β_6 is the cofficient of e401k, β=3.7373, which means when other terms are fixed, the median net financial assets of a family whoes e401k=1 is about 3737 greater than the family with e401k=0.

(iv).

401(k) eligibility has a larger effect on mean wealth than on median wealth, which means the 401(k) eligibility has a larger effect when net financial assets are in a high level.