Project 6

David Contento

Problem 1

1a

Do not need to remove firms since no NA's are contained in the data.

```r
count(USAirlines,'firm')
```

```
##   firm freq
## 1    1   15
## 2    2   15
## 3    3   15
## 4    4   15
## 5    5   15
## 6    6   15
```

1b

```r
summary(USAirlines)
```

```
##  firm        year        output             cost              price
##  1:15   1970   : 6   Min.   :0.03768   Min.   :  68978   Min.   : 103795
##  2:15   1971   : 6   1st Qu.:0.14213   1st Qu.: 292046   1st Qu.: 129848
##  3:15   1972   : 6   Median :0.30503   Median : 637001   Median : 357434
##  4:15   1973   : 6   Mean   :0.54499   Mean   :1122524   Mean   : 471683
##  5:15   1974   : 6   3rd Qu.:0.94528   3rd Qu.:1345968   3rd Qu.: 849840
##  6:15   1975   : 6   Max.   :1.93646   Max.   :4748320   Max.   :1015610
##         (Other):54
##       load
##  Min.   :0.4321
##  1st Qu.:0.5288
##  Median :0.5661
##  Mean   :0.5605
##  3rd Qu.:0.5947
##  Max.   :0.6763
##
```

the outpuit variable ranges from zero to 2 with a median out of .031 . Cost for the firms range from approximately 70,000 to a little less than 5,000,000. Price is variosu greatly similar to cost. The minimum price of a firm is around 100,000 with a max of 1,000,000 . Lastly load ranges from .4 to a little less than .7 with a median value of .56 .

1c

The betas suggest output and price have a positive influence on cost. Load has a negative effect on cost.

```
#estimating equation
logout2 <- (log(USAirlines$output))^2
mod = lm(log(cost) ~ log(output) + logout2 + log(price) + load, data = USAirlines)
summary(mod)
```

```
##
## Call:
## lm(formula = log(cost) ~ log(output) + logout2 + log(price) +
##     load, data = USAirlines)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24060 -0.06740 -0.01145  0.06233  0.32458
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.42058    0.23035  40.896  < 2e-16 ***
## log(output)  0.93543    0.02929  31.941  < 2e-16 ***
## logout2      0.02254    0.01122   2.009   0.0477 *
## log(price)   0.45767    0.02004  22.838  < 2e-16 ***
## load        -1.53744    0.34232  -4.491 2.21e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1225 on 85 degrees of freedom
## Multiple R-squared:  0.9888, Adjusted R-squared:  0.9883
## F-statistic:  1880 on 4 and 85 DF,  p-value: < 2.2e-16
```

1d

Fewer statistically significant columns when more are added, but better R-squared. The betas suggest that output a a postivie effect on cost, some years have a positive effect on cost, and some firms have a negative influence on cost.

```
library(fastDummies)
full = dummy_cols(USAirlines,select_columns = c('year','firm'))
head(full)
```

```
##   firm year   output    cost  price     load year_1970 year_1971 year_1972
## 1    1 1970 0.952757 1140640 106650 0.534487         1         0         0
## 2    1 1971 0.986757 1215690 110307 0.532328         0         1         0
## 3    1 1972 1.091980 1309570 110574 0.547736         0         0         1
## 4    1 1973 1.175780 1511530 121974 0.540846         0         0         0
## 5    1 1974 1.160170 1676730 196606 0.591167         0         0         0
## 6    1 1975 1.173760 1823740 265609 0.575417         0         0         0
##   year_1973 year_1974 year_1975 year_1976 year_1977 year_1978 year_1979
## 1         0         0         0         0         0         0         0
## 2         0         0         0         0         0         0         0
```

```
## 3         0         0         0         0         0         0         0
## 4         1         0         0         0         0         0         0
## 5         0         1         0         0         0         0         0
## 6         0         0         1         0         0         0         0
##   year_1980 year_1981 year_1982 year_1983 year_1984 firm_1 firm_2 firm_3
## 1         0         0         0         0         0      1      0      0
## 2         0         0         0         0         0      1      0      0
## 3         0         0         0         0         0      1      0      0
## 4         0         0         0         0         0      1      0      0
## 5         0         0         0         0         0      1      0      0
## 6         0         0         0         0         0      1      0      0
##   firm_4 firm_5 firm_6
## 1      0      0      0
## 2      0      0      0
## 3      0      0      0
## 4      0      0      0
## 5      0      0      0
## 6      0      0      0
```

```r
#estimating with time effects only
year = dummy_cols(USAirlines,select_columns = c('year'))
mod1 = lm(log(cost) ~ . +
           log(output) + logout2 + log(price) + load -
           firm - year - output - cost - price - load - year_1970, data = year)
summary(mod1)
```

```
##
## Call:
## lm(formula = log(cost) ~ . + log(output) + logout2 + log(price) +
##     load - firm - year - output - cost - price - load - year_1970,
##     data = year)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24921 -0.07594 -0.01167  0.06632  0.31569
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.72165    4.56678   4.537 2.23e-05 ***
## year_1971    0.07397    0.07807   0.948  0.34654
## year_1972    0.09004    0.07882   1.142  0.25708
## year_1973    0.18719    0.08688   2.155  0.03454 *
## year_1974    0.62987    0.26282   2.397  0.01915 *
## year_1975    0.88778    0.36568   2.428  0.01769 *
## year_1976    0.97995    0.40054   2.447  0.01686 *
## year_1977    1.13614    0.45546   2.495  0.01491 *
## year_1978    1.22601    0.49280   2.488  0.01517 *
## year_1979    1.54658    0.63838   2.423  0.01792 *
## year_1980    2.03284    0.80658   2.520  0.01394 *
## year_1981    2.23186    0.86983   2.566  0.01237 *
## year_1982    2.20276    0.84551   2.605  0.01115 *
## year_1983    2.13188    0.81300   2.622  0.01065 *
## year_1984    2.11518    0.79640   2.656  0.00973 **
## log(output)  0.90657    0.03295  27.513  < 2e-16 ***
```

```
## logout2      0.02940    0.01215   2.419  0.01809 *
## log(price)  -0.59006    0.39346  -1.500  0.13807
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1333 on 72 degrees of freedom
## Multiple R-squared:  0.9888, Adjusted R-squared:  0.9861
## F-statistic: 373.3 on 17 and 72 DF,  p-value: < 2.2e-16
```

```r
#estimating with firm effects only
firm = dummy_cols(USAirlines,select_columns = c('firm'))
mod2 = lm(log(cost) ~ . +
           log(output) + logout2 + log(price) + load -
           firm - year - output - cost - price - load - firm_1, data = firm)
summary(mod2)
```

```
##
## Call:
## lm(formula = log(cost) ~ . + log(output) + logout2 + log(price) +
##     load - firm - year - output - cost - price - load - firm_1,
##     data = firm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16093 -0.04774 -0.01154  0.03969  0.19679
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.52062    0.22362  42.574  < 2e-16 ***
## firm_2       0.03463    0.03380   1.025  0.30860
## firm_3      -0.13817    0.07552  -1.830  0.07099 .
## firm_4       0.29196    0.10171   2.871  0.00522 **
## firm_5       0.09075    0.11988   0.757  0.45124
## firm_6       0.18803    0.12310   1.527  0.13056
## log(output)  0.97751    0.06292  15.537  < 2e-16 ***
## logout2      0.02265    0.01136   1.994  0.04950 *
## log(price)   0.38011    0.01877  20.256  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06812 on 81 degrees of freedom
## Multiple R-squared:  0.9967, Adjusted R-squared:  0.9964
## F-statistic:  3061 on 8 and 81 DF,  p-value: < 2.2e-16
```

```r
#estimating with both firm and time effects
mod3 = lm(log(cost) ~ . +
           log(output) + logout2 + log(price) + load -
           firm - year - output - cost - price - load - year_1970 - firm_1, data = full)
summary(mod3)
```

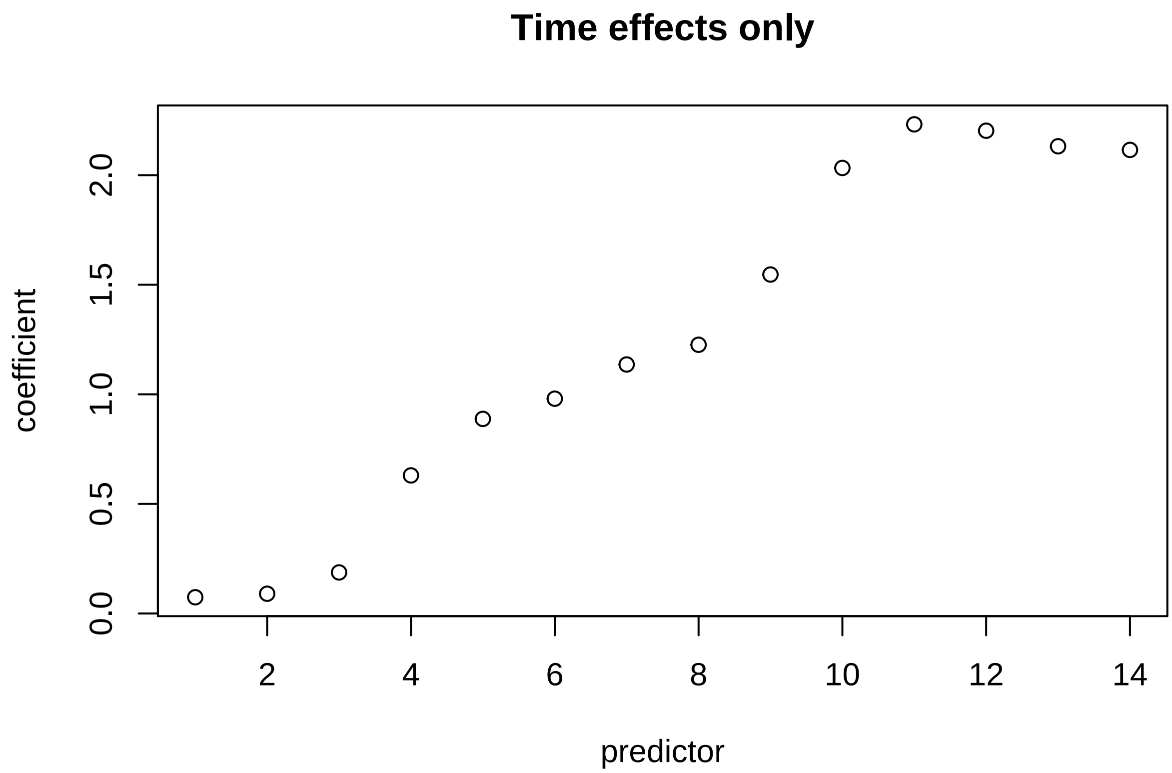
```
##
## Call:
## lm(formula = log(cost) ~ . + log(output) + logout2 + log(price) +
```

```
##     load - firm - year - output - cost - price - load - year_1970 -
##     firm_1, data = full)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.101063 -0.025828 -0.004443  0.024204  0.128254
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.598513   2.064591   6.102 5.89e-08 ***
## year_1971    0.050315   0.032370   1.554  0.12481
## year_1972    0.061955   0.033685   1.839  0.07031 .
## year_1973    0.118959   0.039027   3.048  0.00329 **
## year_1974    0.184177   0.119328   1.543  0.12743
## year_1975    0.252996   0.165886   1.525  0.13194
## year_1976    0.283448   0.181942   1.558  0.12397
## year_1977    0.341743   0.207012   1.651  0.10345
## year_1978    0.368920   0.224355   1.644  0.10479
## year_1979    0.429767   0.290640   1.479  0.14391
## year_1980    0.609910   0.366691   1.663  0.10093
## year_1981    0.694077   0.395641   1.754  0.08395 .
## year_1982    0.710715   0.384827   1.847  0.06919 .
## year_1983    0.703444   0.370569   1.898  0.06197 .
## year_1984    0.719535   0.363440   1.980  0.05184 .
## firm_2      -0.005068   0.031571  -0.161  0.87295
## firm_3      -0.270914   0.079078  -3.426  0.00105 **
## firm_4       0.084853   0.108491   0.782  0.43690
## firm_5      -0.168070   0.129338  -1.299  0.19824
## firm_6      -0.088758   0.133305  -0.666  0.50781
## log(output)  0.850992   0.066656  12.767  < 2e-16 ***
## logout2      0.012344   0.010608   1.164  0.24870
## log(price)   0.112701   0.178276   0.632  0.52942
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05501 on 67 degrees of freedom
## Multiple R-squared:  0.9982, Adjusted R-squared:  0.9976
## F-statistic:  1710 on 22 and 67 DF,  p-value: < 2.2e-16
```

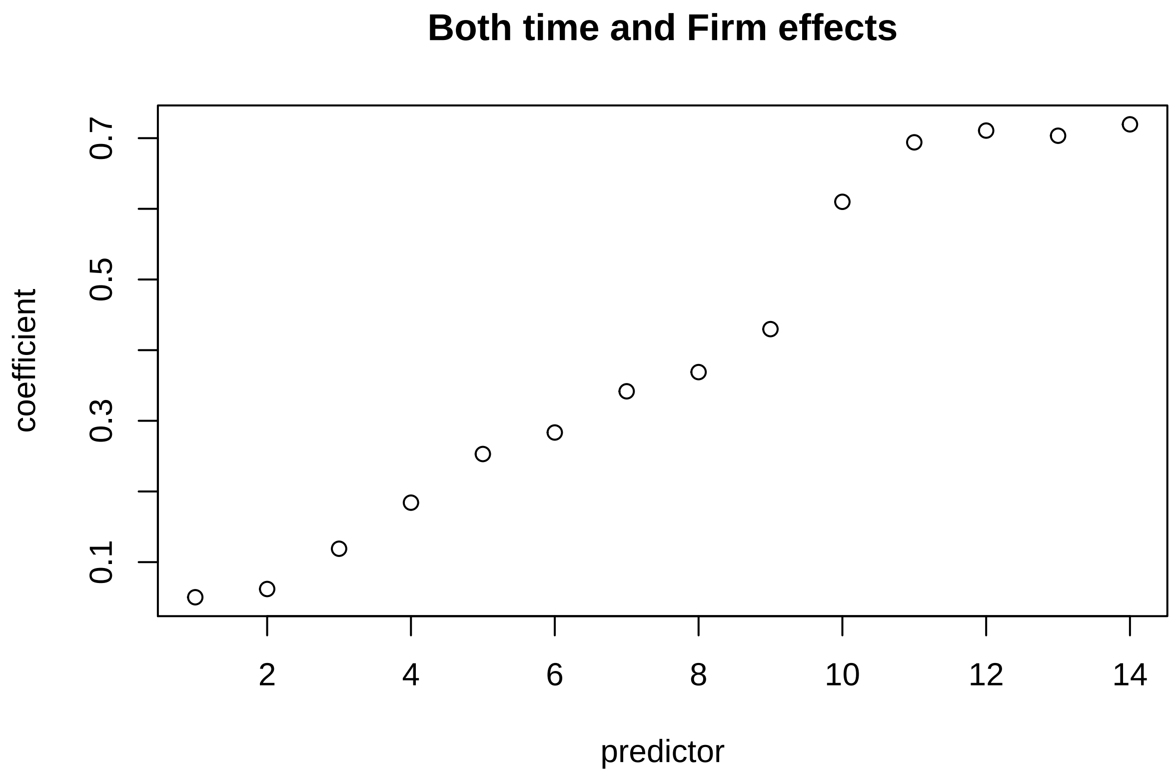
1e

With both time and firm effects, the time effect coefficients are less significant and have a smaller absolute value

```r
#comparing model with plots
plot(mod1$coefficients[grepl('year',names(mod1$coefficients))],xlab='predictor', ylab='coefficient', ma
```




```
plot(mod3$coefficients[grepl('year',names(mod3$coefficients))],xlab='predictor', ylab='coefficient', mai
```



1f

According to the betas output and price positively impact cost while load negatively impacts cost.

```
#random effects model
with <- plm(data = full,log(cost) ~ log(output) + logout2 + log(price) + load)
rand <- plm(data = full,log(cost) ~ log(output) + logout2 + log(price) + load,model='random')
summary(rand)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = log(cost) ~ log(output) + logout2 + log(price) +
##     load, data = full, model = "random")
##
## Balanced Panel: n = 6, T = 15, N = 90
##
## Effects:
##                    var  std.dev share
## idiosyncratic 0.003352 0.057896 0.403
## individual    0.004964 0.070456 0.597
## theta: 0.7924
##
## Residuals:
```

```
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.1317670 -0.0452097 -0.0050006  0.0429753  0.1972573
##
## Coefficients:
##               Estimate Std. Error t-value  Pr(>|t|)
## (Intercept)  9.6141178  0.1898836 50.6316 < 2.2e-16 ***
## log(output)  0.9484650  0.0374441 25.3302 < 2.2e-16 ***
## logout2      0.0150745  0.0086673  1.7392   0.08561 .
## log(price)   0.4243102  0.0135018 31.4261 < 2.2e-16 ***
## load        -1.0595477  0.2055261 -5.1553 1.626e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    42.578
## Residual Sum of Squares: 0.33212
## R-Squared:      0.9922
## Adj. R-Squared: 0.99183
## F-statistic: 2703.03 on 4 and 85 DF, p-value: < 2.22e-16
```

```
phtest(with, rand)
```

```
##
##  Hausman Test
##
## data:  log(cost) ~ log(output) + logout2 + log(price) + load
## chisq = 5.6153, df = 4, p-value = 0.2298
## alternative hypothesis: one model is inconsistent
```

1g

We fail to reject the null hypothesis so the random effects model is appropriate.

Problem 2

2a

The Breusch-Pagan test indicates there is heteroskedasticity.

```
#creating model and running Breusch-Pagan test
olsmodel=lm(wages$LWAGE~wages$EXPER+wages$WKS+wages$OCC+wages$IND+wages$SOUTH+wages$SMSA+wages$MS+wages$
bptest(olsmodel)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  olsmodel
## BP = 93.641, df = 11, p-value = 3.214e-15
```

2b

The estimates are identical suggesting white standard errors can work with panel data in this particular instance.

```
#computing robust and white errors
library(sandwich)
library(ivpack)
library(plm)
robust.se(olsmodel)
```

```
## [1] "Robust Standard Errors"
```

```
##
## t test of coefficients:
##
##                 Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)  5.44115791  0.07474945  72.7919 < 2.2e-16 ***
## wages$EXPER  0.01037090  0.00058407  17.7563 < 2.2e-16 ***
## wages$WKS    0.00494447  0.00116100   4.2588 2.101e-05 ***
## wages$OCC   -0.14863433  0.01509054  -9.8495 < 2.2e-16 ***
## wages$IND    0.05305779  0.01217719   4.3571 1.350e-05 ***
## wages$SOUTH -0.05321007  0.01316625  -4.0414 5.410e-05 ***
## wages$SMSA   0.14530390  0.01232896  11.7856 < 2.2e-16 ***
## wages$MS     0.06607989  0.02145912   3.0793  0.002088 **
## wages$FEM   -0.35330159  0.02427720 -14.5528 < 2.2e-16 ***
## wages$UNION  0.10207577  0.01245321   8.1967 3.258e-16 ***
## wages$ED     0.05715394  0.00273401  20.9048 < 2.2e-16 ***
## wages$BLK   -0.16712256  0.02179716  -7.6672 2.176e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(olsmodel, vcov=vcovHC(olsmodel, cluster="individual"))
```

```
##
## t test of coefficients:
##
##                 Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)  5.44115791  0.07517312  72.3817 < 2.2e-16 ***
## wages$EXPER  0.01037090  0.00058607  17.6958 < 2.2e-16 ***
## wages$WKS    0.00494447  0.00116862   4.2310 2.376e-05 ***
## wages$OCC   -0.14863433  0.01514100  -9.8167 < 2.2e-16 ***
## wages$IND    0.05305779  0.01221367   4.3441 1.432e-05 ***
## wages$SOUTH -0.05321007  0.01320815  -4.0286 5.712e-05 ***
## wages$SMSA   0.14530390  0.01236627  11.7500 < 2.2e-16 ***
## wages$MS     0.06607989  0.02156975   3.0635  0.002201 **
## wages$FEM   -0.35330159  0.02440536 -14.4764 < 2.2e-16 ***
## wages$UNION  0.10207577  0.01249606   8.1686 4.098e-16 ***
## wages$ED     0.05715394  0.00274356  20.8321 < 2.2e-16 ***
## wages$BLK   -0.16712256  0.02192441  -7.6227 3.060e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2c

The individual effects do not change dramatically. The beta on experience suggests it has a positive impact on wages.

```
#fixed effects without time
plmfe=plm(wages$LWAGE~wages$EXPER+wages$WKS+wages$OCC+wages$IND+wages$SOUTH+wages$SMSA+wages$MS+wages$FE
summary(plmfe)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = wages$LWAGE ~ wages$EXPER + wages$WKS + wages$OCC +
##     wages$IND + wages$SOUTH + wages$SMSA + wages$MS + wages$FEM +
##     wages$UNION + wages$ED + wages$BLK, data = wages, model = "within")
##
## Balanced Panel: n = 595, T = 7, N = 4165
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -1.7984511 -0.0535263  0.0042525  0.0628480  1.9452352
##
## Coefficients:
##                Estimate  Std. Error t-value Pr(>|t|)
## wages$EXPER  0.09657698  0.00119085 81.0992  < 2e-16 ***
## wages$WKS    0.00114223  0.00060316  1.8937  0.05834 .
## wages$OCC   -0.02486403  0.01388776 -1.7904  0.07348 .
## wages$IND    0.02075656  0.01556962  1.3331  0.18257
## wages$SOUTH -0.00319792  0.03457562 -0.0925  0.92631
## wages$SMSA  -0.04372702  0.01958444 -2.2327  0.02563 *
## wages$MS    -0.03025961  0.01913663 -1.5812  0.11391
## wages$UNION  0.03415826  0.01504220  2.2708  0.02322 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    240.65
## Residual Sum of Squares: 83.624
## R-Squared:      0.65251
## Adj. R-Squared: 0.59378
## F-statistic: 836.082 on 8 and 3562 DF, p-value: < 2.22e-16
```

```
#fixed effects with time
plmfet=plm(wages$LWAGE~wages$EXPER+wages$WKS+wages$OCC+wages$IND+wages$SOUTH+wages$SMSA+wages$MS+wages$
summary(plmfet)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = wages$LWAGE ~ wages$EXPER + wages$WKS + wages$OCC +
##     wages$IND + wages$SOUTH + wages$SMSA + wages$MS + wages$FEM +
##     wages$UNION + wages$ED + wages$BLK + wages$time, data = wages,
##     model = "within")
##
```

```
## Balanced Panel: n = 595, T = 7, N = 4165
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -1.8149692 -0.0549163  0.0038679  0.0629619  1.9263486
##
## Coefficients: (1 dropped because of singularities)
##                Estimate  Std. Error t-value  Pr(>|t|)
## wages$EXPER  0.09555975  0.00147693 64.7014 < 2.2e-16 ***
## wages$WKS    0.00094862  0.00060235  1.5749 0.1153799
## wages$OCC   -0.02213145  0.01384344 -1.5987 0.1099771
## wages$IND    0.02235811  0.01551107  1.4414 0.1495515
## wages$SOUTH  0.00228936  0.03443928  0.0665 0.9470031
## wages$SMSA  -0.04318164  0.01951543 -2.2127 0.0269822 *
## wages$MS    -0.02899049  0.01905818 -1.5212 0.1283091
## wages$UNION  0.03067453  0.01498978  2.0464 0.0407940 *
## wages$time2 -0.00592816  0.00822322 -0.7209 0.4710154
## wages$time3  0.02856496  0.00781517  3.6551 0.0002608 ***
## wages$time4  0.03178481  0.00768144  4.1379 3.587e-05 ***
## wages$time5  0.02717390  0.00781871  3.4755 0.0005160 ***
## wages$time6  0.00927219  0.00821690  1.1284 0.2592145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    240.65
## Residual Sum of Squares: 82.751
## R-Squared:      0.65614
## Adj. R-Squared: 0.59746
## F-statistic: 522.098 on 13 and 3557 DF, p-value: < 2.22e-16
```

2d

Random effects model is more appropriate according to the Hausman Test. Experience continues to have an effect on wages although the beta is smaller compated to previous models.

```r
#random effects model
rem=plm(wages$LWAGE~wages$EXPER+wages$WKS+wages$OCC+wages$IND+wages$SOUTH+wages$SMSA+wages$MS+wages$FEM
summary(rem)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = wages$LWAGE ~ wages$EXPER + wages$WKS + wages$OCC +
##     wages$IND + wages$SOUTH + wages$SMSA + wages$MS + wages$FEM +
##     wages$UNION + wages$ED + wages$BLK + wages$time, data = wages,
##     model = "random")
##
## Balanced Panel: n = 595, T = 7, N = 4165
##
## Effects:
##                   var std.dev share
## idiosyncratic 0.02326 0.15253 0.243
```

```
## individual    0.07245 0.26917 0.757
## theta: 0.7906
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -1.8674762 -0.0724025  0.0017051  0.0785332  1.9257262
##
## Coefficients:
##                Estimate  Std. Error  t-value  Pr(>|t|)
## (Intercept)  5.38510575  0.07742745  69.5503 < 2.2e-16 ***
## wages$EXPER  0.00827172  0.00109918   7.5253 6.412e-14 ***
## wages$WKS    0.00119300  0.00059911   1.9913 0.0465158 *
## wages$OCC   -0.04705155  0.01291722  -3.6425 0.0002733 ***
## wages$IND    0.03086763  0.01343157   2.2981 0.0216033 *
## wages$SOUTH -0.05843089  0.02079514  -2.8098 0.0049799 **
## wages$SMSA   0.04206960  0.01562189   2.6930 0.0071099 **
## wages$MS    -0.01387155  0.01787754  -0.7759 0.4378402
## wages$FEM   -0.42417887  0.04039835 -10.4999 < 2.2e-16 ***
## wages$UNION  0.04571450  0.01327180   3.4445 0.0005778 ***
## wages$ED     0.06625067  0.00457128  14.4928 < 2.2e-16 ***
## wages$BLK   -0.15241848  0.04575265  -3.3314 0.0008718 ***
## wages$time2  0.08072921  0.00901248   8.9575 < 2.2e-16 ***
## wages$time3  0.20321474  0.00921803  22.0454 < 2.2e-16 ***
## wages$time4  0.29385415  0.00955517  30.7534 < 2.2e-16 ***
## wages$time5  0.37523543  0.00997187  37.6294 < 2.2e-16 ***
## wages$time6  0.44616485  0.01050618  42.4669 < 2.2e-16 ***
## wages$time7  0.52505848  0.01112929  47.1781 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    269
## Residual Sum of Squares: 98.412
## R-Squared:      0.63415
## Adj. R-Squared: 0.63265
## F-statistic: 422.842 on 17 and 4147 DF, p-value: < 2.22e-16
```

```r
phtest(rem, plmfet)
```

```
##
##  Hausman Test
##
## data:  wages$LWAGE ~ wages$EXPER + wages$WKS + wages$OCC + wages$IND +  ...
## chisq = 283.43, df = 13, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

Problem 3

3a

Below we calculated investment (invest) as the difference between income and expenditure.

```r
#calculating investment
investment=datadf$income-datadf$expenditure
datadf[,3]=investment
colnames(datadf)[3]<-"invest"
datadf
```

```
##    income expenditure invest
## 1   751.6       672.1   79.5
## 2   779.2       696.8   82.4
## 3   810.3       737.1   73.2
## 4   864.7       767.9   96.8
## 5   857.5       762.8   94.7
## 6   874.9       779.4   95.5
## 7   906.8       823.1   83.7
## 8   942.9       864.3   78.6
## 9   988.8       903.2   85.6
## 10 1015.7       927.6   88.1
```

3b

Below we calculated the summary statistics and calculated the summary statistics fo each variable.

```
##             vars  n   mean    sd median trimmed   mad   min    max range
## income         1 10 879.24 86.41 869.80  878.14 98.30 751.6 1015.7 264.1
## expenditure    2 10 793.43 84.97 773.65  791.83 93.63 672.1  927.6 255.5
## invest         3 10  85.81  7.93  84.65   86.01  8.30  73.2   96.8  23.6
##             skew kurtosis    se
## income      0.11    -1.39 27.32
## expenditure 0.20    -1.44 26.87
## invest      0.05    -1.50  2.51
```

```
##      income        expenditure        invest
##  Min.   : 751.6   Min.   :672.1   Min.   :73.20
##  1st Qu.: 822.1   1st Qu.:743.5   1st Qu.:80.22
##  Median : 869.8   Median :773.6   Median :84.65
##  Mean   : 879.2   Mean   :793.4   Mean   :85.81
##  3rd Qu.: 933.9   3rd Qu.:854.0   3rd Qu.:93.05
##  Max.   :1015.7   Max.   :927.6   Max.   :96.80
```

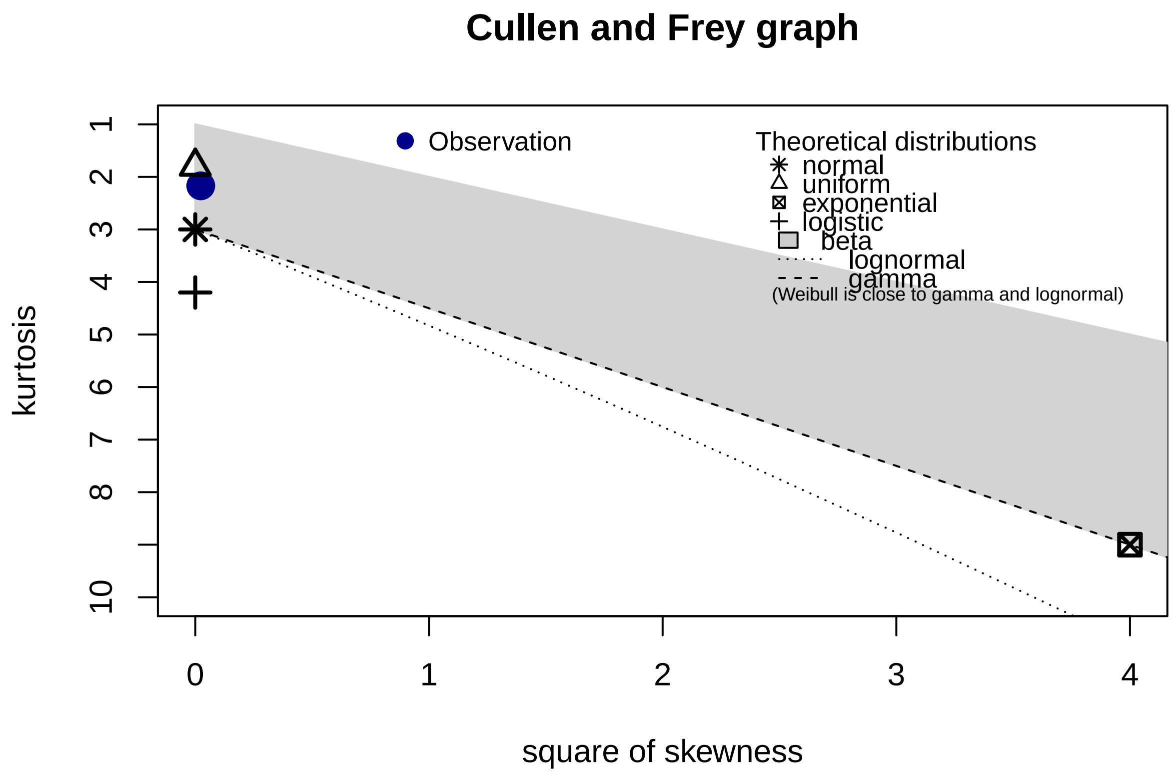
```r
library(fitdistrplus)
```

```
## Loading required package: MASS
```

```
## Loading required package: npsurv
```

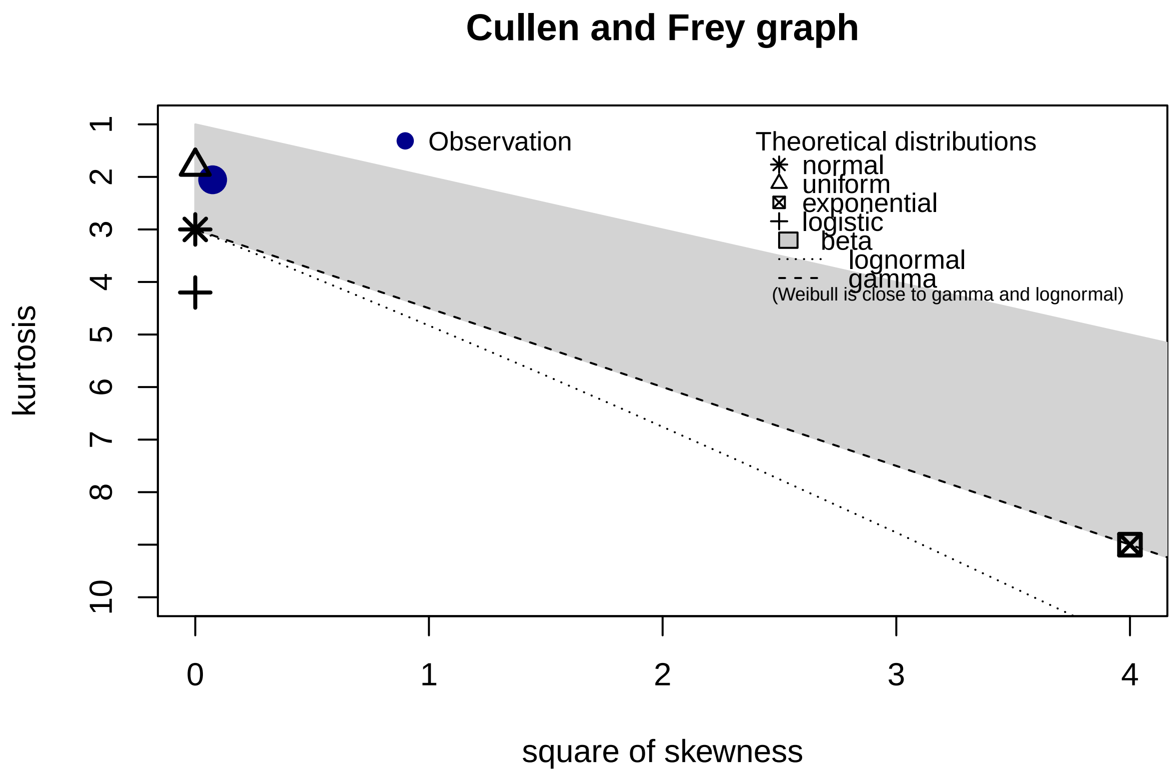
```
## Loading required package: lsei
```

```r
#distribution for income
descdist(datadf$income)
```



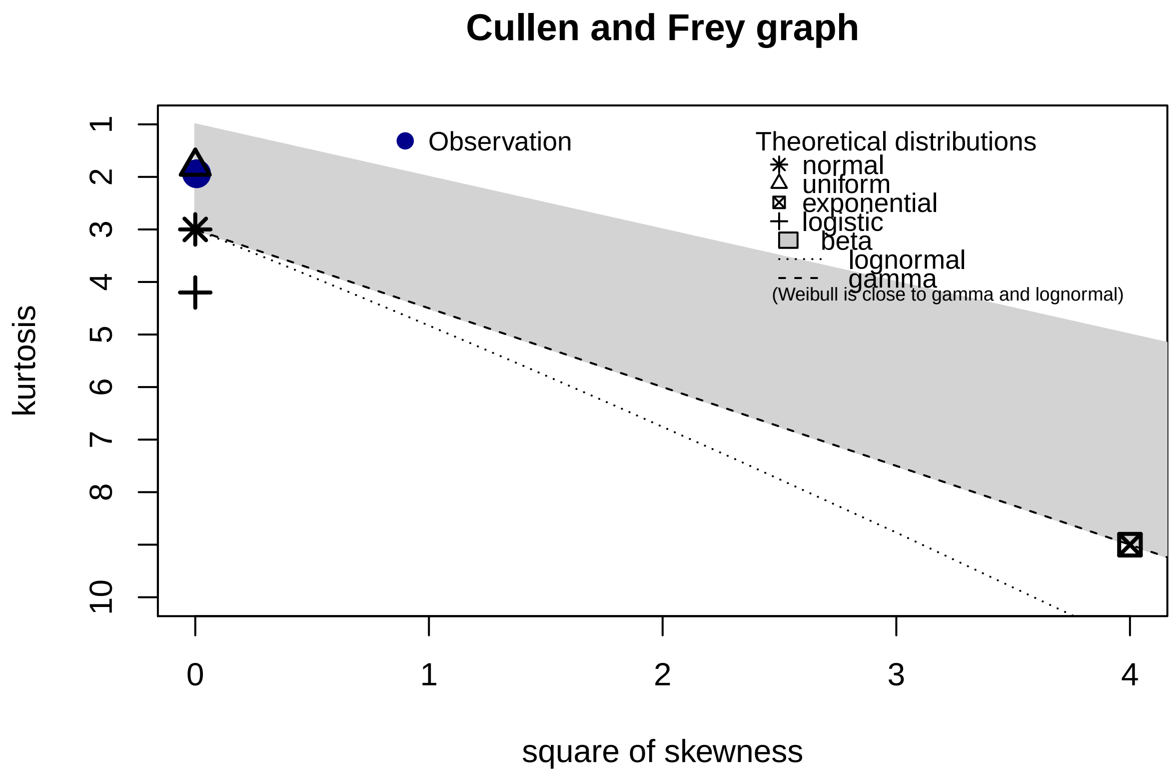

```
## summary statistics
## ------
## min:  751.6   max:  1015.7
## median:  869.8
## mean:  879.24
## estimated sd:  86.40502
## estimated skewness:  0.1535735
## estimated kurtosis:  2.170257
```

```r
#distribution for expenditure
descdist(datadf$expenditure)
```




```
## summary statistics
## ------
## min:  672.1   max:  927.6
## median:  773.65
## mean:  793.43
## estimated sd:  84.96543
## estimated skewness:  0.2727997
## estimated kurtosis:  2.055203
```

```r
#distribution for investment
descdist(datadf$invest)
```




```
## summary statistics
## ------
## min:  73.2   max:  96.8
## median:  84.65
## mean:  85.81
## estimated sd:  7.929474
## estimated skewness:  0.07565637
## estimated kurtosis:  1.940239
```

From the graph above we can see that a uniform distribution fits all the variables in the data extremely close.

3c

```r
#regressing income on expenditure
reg=lm(datadf$income~datadf$expenditure)
summary(reg)
```

```
##
## Call:
## lm(formula = datadf$income ~ datadf$expenditure)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -11.893  -4.196  -1.894   7.105  11.315
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          75.70815   26.06849   2.904   0.0198 *
## datadf$expenditure  1.01273    0.03269  30.983 1.28e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.332 on 8 degrees of freedom
## Multiple R-squared:  0.9917, Adjusted R-squared:  0.9907
## F-statistic: 959.9 on 1 and 8 DF,  p-value: 1.28e-09
```

3d

```r
#stage 1 (creating instrument)
reg1=lm(expenditure~investment)

#stage 2 (using instrument)
reghat=fitted.values(reg1)
reg2=lm(income~reghat)

cor(datadf$expenditure,datadf$invest)
```

```
## [1] 0.1364239
```

```r
cor(reg1$residuals,datadf$invest)
```

```
## [1] -2.399166e-16
```

```r
summary(reg2)
```

```
##
## Call:
## lm(formula = income ~ reghat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -112.11  -43.12  -33.05   69.25  130.82
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -456.966   2037.209  -0.224    0.828
## reghat         1.684      2.567   0.656    0.530
##
## Residual standard error: 89.28 on 8 degrees of freedom
## Multiple R-squared:  0.05104,    Adjusted R-squared:  -0.06758
## F-statistic: 0.4303 on 1 and 8 DF,  p-value: 0.5303
```

From the summary and correlations above we can see that the instrument (invest) is correlated with expenditure, but not correlated with the errors of the orginial regression. This suggests that the instrument is good and would help us understand the effect of expenditure on income. However, using the instrument takes away any statistical significance we had in the regression so perhaps we have a bad instrument.

Problem 4

4a

there is a negative relationship between fertility and education holding all other factors constant/fixed

```
## The following object is masked _by_ .GlobalEnv:
##
##     year
```

```
#Original OLS model
reg=lm(kids~age+agesq+educ+black+east+west+northcen+farm+othrural+town+smcity+y74+y76+y78+y80+y82+y84)
summary(reg)
```

```
##
## Call:
## lm(formula = kids ~ age + agesq + educ + black + east + west +
##     northcen + farm + othrural + town + smcity + y74 + y76 +
##     y78 + y80 + y82 + y84)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9878 -1.0086 -0.0767  0.9331  4.6548
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.742457   3.051767  -2.537 0.011315 *
## age          0.532135   0.138386   3.845 0.000127 ***
## agesq       -0.005804   0.001564  -3.710 0.000217 ***
## educ        -0.128427   0.018349  -6.999 4.44e-12 ***
## black        1.075658   0.173536   6.198 8.02e-10 ***
## east         0.217324   0.132788   1.637 0.101992
## west         0.197603   0.166913   1.184 0.236719
## northcen     0.363114   0.120897   3.004 0.002729 **
## farm        -0.052557   0.147190  -0.357 0.721105
## othrural    -0.162854   0.175442  -0.928 0.353481
## town         0.084353   0.124531   0.677 0.498314
## smcity       0.211879   0.160296   1.322 0.186507
## y74          0.268183   0.172716   1.553 0.120771
## y76         -0.097379   0.179046  -0.544 0.586633
## y78         -0.068666   0.181684  -0.378 0.705544
## y80         -0.071305   0.182771  -0.390 0.696511
## y82         -0.522484   0.172436  -3.030 0.002502 **
## y84         -0.545166   0.174516  -3.124 0.001831 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.555 on 1111 degrees of freedom
## Multiple R-squared:  0.1295, Adjusted R-squared:  0.1162
## F-statistic: 9.723 on 17 and 1111 DF,  p-value: < 2.2e-16
```

4b

```r
#stage 1 (creating instrument)
iveduc=lm(educ~meduc+feduc+age+agesq+black+east+west+northcen+farm+othrural+town+smcity+y74+y76+y78+y80+
educfitted=iveduc$fitted.values

#stage 2 (Implementing Instrument )
ivreg1=lm(kids~educfitted+age+agesq+black+east+west+northcen+farm+othrural+town+smcity+y74+y76+y78+y80+y
summary(ivreg1)
```

```
##
## Call:
## lm(formula = kids ~ educfitted + age + agesq + black + east +
##     west + northcen + farm + othrural + town + smcity + y74 +
##     y76 + y78 + y80 + y82 + y84)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9611 -1.0591 -0.0576  0.9432  4.8706
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.241244   3.181487  -2.276 0.023032 *
## educfitted  -0.152739   0.039784  -3.839 0.000130 ***
## age          0.523554   0.141023   3.713 0.000215 ***
## agesq       -0.005716   0.001593  -3.588 0.000347 ***
## black        1.072952   0.176199   6.089 1.56e-09 ***
## east         0.228555   0.135767   1.683 0.092572 .
## west         0.207640   0.170054   1.221 0.222336
## northcen     0.374419   0.123806   3.024 0.002550 **
## farm        -0.077002   0.153536  -0.502 0.616104
## othrural    -0.195245   0.184147  -1.060 0.289252
## town         0.081810   0.126465   0.647 0.517831
## smcity       0.212500   0.162719   1.306 0.191846
## y74          0.272129   0.175417   1.551 0.121107
## y76         -0.094548   0.181795  -0.520 0.603110
## y78         -0.057254   0.185164  -0.309 0.757220
## y80         -0.053248   0.187358  -0.284 0.776307
## y82         -0.496215   0.179114  -2.770 0.005692 **
## y84         -0.521360   0.180464  -2.889 0.003940 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.578 on 1111 degrees of freedom
## Multiple R-squared:  0.103,  Adjusted R-squared:  0.0893
## F-statistic: 7.507 on 17 and 1111 DF,  p-value: < 2.2e-16
```

```r
#Checking for correlation
cor(educ,meduc+feduc)
```

```
## [1] 0.5180762
```

```r
cor(reg$residuals,meduc+feduc)
```

```
## [1] -0.01692679
```

From the output above we can see that the instruments (meduc and feduc) are correlated with education while remaining uncorrelated with the errors of the first regression. This suggests that the instruments works and is good for the purpose of understanding the effects our endogenous variables have on number of kids.

4c

Below we added interaction terms for education over time. From the summary we can see that over time education has increasingly negative effect on the number of kids a woman has.

```r
#including interaction term for education overtime
ivreg1=lm(kids~educfitted+age+agesq+black+east+west+northcen+farm+othrural+town+smcity+y74+y76+y78+y80+y
y74educ+y76educ+y78educ+y80educ+y82educ+y84educ)
summary(ivreg1)
```

```
##
## Call:
## lm(formula = kids ~ educfitted + age + agesq + black + east +
##     west + northcen + farm + othrural + town + smcity + y74 +
##     y76 + y78 + y80 + y82 + y84 + y74educ + y76educ + y78educ +
##     y80educ + y82educ + y84educ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4159 -1.0403 -0.0748  0.9765  4.6073
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.944574   3.155170  -2.518 0.011944 *
## educfitted  -0.043058   0.043126  -0.998 0.318289
## age          0.495686   0.139443   3.555 0.000394 ***
## agesq       -0.005404   0.001575  -3.431 0.000623 ***
## black        1.071781   0.173600   6.174 9.35e-10 ***
## east         0.219515   0.133920   1.639 0.101465
## west         0.188457   0.167770   1.123 0.261551
## northcen     0.362599   0.122060   2.971 0.003036 **
## farm        -0.107202   0.151306  -0.709 0.478777
## othrural    -0.237686   0.181355  -1.311 0.190262
## town         0.085308   0.124531   0.685 0.493466
## smcity       0.205046   0.160098   1.281 0.200550
## y74          1.109122   0.635629   1.745 0.081276 .
## y76          1.206165   0.599846   2.011 0.044590 *
## y78          1.991496   0.700167   2.844 0.004533 **
```

```
## y80          1.282804   0.625625   2.050 0.040558 *
## y82          1.392194   0.593191   2.347 0.019103 *
## y84          1.887372   0.622813   3.030 0.002499 **
## y74educ     -0.069319   0.049777  -1.393 0.164026
## y76educ     -0.107083   0.046907  -2.283 0.022626 *
## y78educ     -0.166304   0.053678  -3.098 0.001996 **
## y80educ     -0.109787   0.046935  -2.339 0.019506 *
## y82educ     -0.151824   0.043588  -3.483 0.000515 ***
## y84educ     -0.190992   0.045674  -4.182 3.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.552 on 1105 degrees of freedom
## Multiple R-squared:  0.1371, Adjusted R-squared:  0.1191
## F-statistic: 7.634 on 23 and 1105 DF,  p-value: < 2.2e-16
```