final project

Zhang Zhiyuan 15220162202517

2019/6/9

ABSTRACT:

Generally, we would expect that having a child decreases mother's working time, especially for those who already have some children. The question is, how big the effect? How to capture a good estimate with the existence of so many confounding factors? Is there any difference when the already existing children are boys? Or girls? This project uses the method of instrument variable to answer the above questions.

1. Introduction

This project is an extension of the previous homework7. In homework 7, I explored the causal effect of having extra child or children on the working time of those who already have two children. In this project, I mainly further explored this three things:

1. the comparison of the predictive accuracy of classic and IV regression.
2. the effect of extra child or children when the first two children are boys.
3. the effect of extra child or children when the first children are girls.

Besides, in this project I used the full sample with 254654 observations, instead of the samll sample of 30000 in homework7. SO the estimates is theoratically much more precise and close to the real value.

the data comes from the 1980 Census of America.

detailed description of variables:

extrakid =1 if mom had more than 2 children

ssex =1 if 1st two children same sex

age age of mom at census

color =1 if mom is black

hispan =1 if mom is Hispanic

orace =1 if mom is not black, Hispanic or white

workingweek mom's weeks worked in 1979

twoboy =1 if 1st two children are all boys

twogirl =1 if 1st two children are all girls

2. Data import and processment

To compare the accuracy of predictive power of classic regression and IV regression, I use the first 220000 observations as train set, and the remaining 34654 observations for validation.

```
data <- read.csv("D:/i_love_learning/fertility.csv",header=T)
library(forecast)

workingweek_t<-data$weeksm1[1:220000]
extrakid_t<-data$morekids[1:220000]
color_t<-data$black[1:220000]
age_t<-data$agem1[1:220000]
hispan_t<-data$hispan[1:220000]
```

```
orace_t<-data$othrace[1:220000]
ssex_t<-data$samesex[1:220000]
boy1_t<-data$boy1st[1:220000]
boy2_t<-data$boy2nd[1:220000]

workingweek_v<-data$weeksm1[220001:254654]
extrakid_v<-data$morekids[220001:254654]
color_v<-data$black[220001:254654]
age_v<-data$agem1[220001:254654]
hispan_v<-data$hispan[220001:254654]
orace_v<-data$othrace[220001:254654]
ssex_v<-data$samesex[220001:254654]
boy1_v<-data$boy1st[220001:254654]
boy2_v<-data$boy2nd[220001:254654]
```

3. model construction and processment

3.1
use the classic regression to capture the effect, the estimate is naturally biased, but can be left for comparision in the magnitude of the effect and the accuracy in predictive power.

```
fit1<-lm(workingweek_t~extrakid_t+color_t+age_t+hispan_t+orace_t)
summary(fit1)
```

```
##
## Call:
## lm(formula = workingweek_t ~ extrakid_t + color_t + age_t + hispan_t +
##     orace_t)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36.12 -17.77 -10.71  22.90  45.46
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.71747    0.41657 -11.325   <2e-16 ***
## extrakid_t  -6.22868    0.09522 -65.414   <2e-16 ***
## color_t     11.57936    0.20389  56.793   <2e-16 ***
## age_t        0.83277    0.01363  61.079   <2e-16 ***
## hispan_t     0.10874    0.20324   0.535    0.593
## orace_t      2.43091    0.22041  11.029   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.4 on 219994 degrees of freedom
## Multiple R-squared:  0.04343,    Adjusted R-squared:  0.04341
## F-statistic:  1998 on 5 and 219994 DF,  p-value: < 2.2e-16
```

```
valid_pre <- predict(fit1,newdata=data.frame(extrakid_t=extrakid_v,color_t=color_v,age_t=age_v,hispan_t=
accuracy(valid_pre,workingweek_v)
```

```
##                 ME     RMSE      MAE  MPE MAPE
## Test set 0.3852793 21.26799 19.18659 -Inf  Inf
```

3.2 use the variable “ssex” as instrument, and recapture the causal effect. also relevant tests are enforced to ensure the relevacne and exogeneity.

```r
fit2<-lm(extrakid_t~ssex_t+color_t+age_t+hispan_t+orace_t)
summary(fit2)
```

```
##
## Call:
## lm(formula = extrakid_t ~ ssex_t + color_t + age_t + hispan_t +
##     orace_t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7006 -0.3822 -0.3020  0.5865  0.8230
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.1510527  0.0093648 -16.130  < 2e-16 ***
## ssex_t       0.0645888  0.0020388  31.679  < 2e-16 ***
## color_t      0.1049727  0.0045494  23.074  < 2e-16 ***
## age_t        0.0156234  0.0003028  51.599  < 2e-16 ***
## hispan_t     0.1353188  0.0045313  29.863  < 2e-16 ***
## orace_t      0.0333278  0.0049233   6.769  1.3e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4781 on 219994 degrees of freedom
## Multiple R-squared:  0.02286,    Adjusted R-squared:  0.02284
## F-statistic:  1029 on 5 and 219994 DF,  p-value: < 2.2e-16
```

```r
extrakid_t_ins<-fit2$fitted.values

fit3<-lm(workingweek_t~extrakid_t_ins+color_t+age_t+hispan_t+orace_t)
summary(fit3)
```

```
##
## Call:
## lm(formula = workingweek_t ~ extrakid_t_ins + color_t + age_t +
##     hispan_t + orace_t)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32.88 -18.63 -12.39  22.80  41.46
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -4.67436    0.45234 -10.334   <2e-16 ***
## extrakid_t_ins -5.86014    1.42664  -4.108    4e-05 ***
## color_t        11.54085    0.25398  45.440   <2e-16 ***
## age_t           0.82703    0.02610  31.691   <2e-16 ***
## hispan_t        0.05881    0.28158   0.209    0.835
## orace_t         2.41867    0.22749  10.632   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.61 on 219994 degrees of freedom
## Multiple R-squared:  0.0249, Adjusted R-squared:  0.02488
```

```
## F-statistic:  1123 on 5 and 219994 DF,  p-value: < 2.2e-16
```

```
resid<-fit3$residuals
test<-lm(resid~ssex_t+color_t+age_t+hispan_t+orace_t)
summary(test)
```

```
##
## Call:
## lm(formula = resid ~ ssex_t + color_t + age_t + hispan_t + orace_t)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32.88 -18.63 -12.39  22.80  41.46
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.589e-13  4.232e-01       0        1
## ssex_t      -1.794e-13  9.214e-02       0        1
## color_t      1.207e-13  2.056e-01       0        1
## age_t        1.009e-14  1.368e-02       0        1
## hispan_t     1.116e-13  2.048e-01       0        1
## orace_t     -1.437e-13  2.225e-01       0        1
##
## Residual standard error: 21.61 on 219994 degrees of freedom
## Multiple R-squared:  4.748e-29,  Adjusted R-squared:  -2.273e-05
## F-statistic: 2.089e-24 on 5 and 219994 DF,  p-value: 1
```

```
fit4<-lm(extrakid_v~ssex_v+color_v+age_v+hispan_v+orace_v)
extrakid_v_ins<-fit4$fitted.values

valid_pre_ins <- predict(fit3,newdata=data.frame(extrakid_t_ins=extrakid_v_ins,color_t=color_v,age_t=ag
accuracy(valid_pre_ins,workingweek_v)
```

```
##                 ME     RMSE      MAE  MPE MAPE
## Test set 0.3668273 21.48916 19.51076 -Inf  Inf
```

3.3 use the variable “twoboy” as instrument, and recapture the causal effect. also the tests are enforced.

```
twoboy_t<-boy1_t*boy2_t
twoboy_v<-boy1_v*boy2_v

fit5<-lm(extrakid_t~twoboy_t+color_t+age_t+hispan_t+orace_t)
summary(fit5)
```

```
##
## Call:
## lm(formula = extrakid_t ~ twoboy_t + color_t + age_t + hispan_t +
##     orace_t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6911 -0.3883 -0.3110  0.5956  0.7980
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.1250600  0.0093403 -13.389  < 2e-16 ***
## twoboy_t     0.0306209  0.0023122  13.243  < 2e-16 ***
```

```
## color_t      0.1048288  0.0045580  22.999  < 2e-16 ***
## age_t        0.0155735  0.0003033  51.338  < 2e-16 ***
## hispan_t     0.1355863  0.0045398  29.866  < 2e-16 ***
## orace_t      0.0331662  0.0049326   6.724 1.77e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.479 on 219994 degrees of freedom
## Multiple R-squared:  0.01919,    Adjusted R-squared:  0.01916
## F-statistic: 860.7 on 5 and 219994 DF,  p-value: < 2.2e-16
```

```
extrakid_t_b<-fit5$fitted.values

fit6<-lm(workingweek_t~extrakid_t_b+color_t+age_t+hispan_t+orace_t)
summary(fit6)
```

```
##
## Call:
## lm(formula = workingweek_t ~ extrakid_t_b + color_t + age_t +
##     hispan_t + orace_t)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32.77 -18.90 -12.28  22.88  41.50
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5.17198    0.57925  -8.929  < 2e-16 ***
## extrakid_t_b -10.11430    3.40627  -2.969  0.00298 **
## color_t       11.98545    0.41110  29.155  < 2e-16 ***
## age_t          0.89329    0.05479  16.303  < 2e-16 ***
## hispan_t       0.63511    0.50484   1.258  0.20838
## orace_t        2.55991    0.24960  10.256  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.61 on 219994 degrees of freedom
## Multiple R-squared:  0.02486,    Adjusted R-squared:  0.02484
## F-statistic:  1122 on 5 and 219994 DF,  p-value: < 2.2e-16
```

```
resid_b<-fit5$residuals
test_b<-lm(resid_b~twoboy_t+color_t+age_t+hispan_t+orace_t)
summary(test_b)
```

```
##
## Call:
## lm(formula = resid_b ~ twoboy_t + color_t + age_t + hispan_t +
##     orace_t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6911 -0.3883 -0.3110  0.5956  0.7980
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -2.741e-15  9.340e-03       0        1
## twoboy_t    -8.682e-15  2.312e-03       0        1
## color_t     -5.241e-15  4.558e-03       0        1
## age_t        7.608e-17  3.033e-04       0        1
## hispan_t    -7.566e-15  4.540e-03       0        1
## orace_t     -1.422e-15  4.933e-03       0        1
##
## Residual standard error: 0.479 on 219994 degrees of freedom
## Multiple R-squared:  1.187e-28,  Adjusted R-squared:  -2.273e-05
## F-statistic: 5.221e-24 on 5 and 219994 DF,  p-value: 1
```

```
fit7<-lm(extrakid_v~twoboy_v+color_v+age_v+hispan_v+orace_v)
extrakid_v_b<-fit7$fitted.values

valid_pre_b <- predict(fit6,newdata=data.frame(extrakid_t_b=extrakid_v_b,color_t=color_v,age_t=age_v,his
accuracy(valid_pre_b,workingweek_v)
```

```
##                   ME     RMSE      MAE  MPE MAPE
## Test set 0.5798222 21.49608 19.46993 -Inf  Inf
```

3.4 use the variable “twogirl” as instrument, and recapture the causal effect. also the tests are enforced.

```
twogirl_t<-ssex_t-twoboy_t
twogirl_v<-ssex_v-twoboy_v

fit8<-lm(extrakid_t~twogirl_t+color_t+age_t+hispan_t+orace_t)
summary(fit8)
```

```
##
## Call:
## lm(formula = extrakid_t ~ twogirl_t + color_t + age_t + hispan_t +
##     orace_t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6948 -0.3838 -0.3057  0.5850  0.8037
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.1317017  0.0093338 -14.110  < 2e-16 ***
## twogirl_t    0.0558898  0.0023920  23.365  < 2e-16 ***
## color_t      0.1043274  0.0045541  22.908  < 2e-16 ***
## age_t        0.0156213  0.0003031  51.538  < 2e-16 ***
## hispan_t     0.1351219  0.0045360  29.789  < 2e-16 ***
## orace_t      0.0333716  0.0049284   6.771 1.28e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4786 on 219994 degrees of freedom
## Multiple R-squared:  0.02083,    Adjusted R-squared:  0.02081
## F-statistic: 936.2 on 5 and 219994 DF,  p-value: < 2.2e-16
```

```
extrakid_t_g<-fit8$fitted.values

fit9<-lm(workingweek_t~extrakid_t_g+color_t+age_t+hispan_t+orace_t)
summary(fit9)
```

```
##
## Call:
## lm(formula = workingweek_t ~ extrakid_t_g + color_t + age_t +
##     hispan_t + orace_t)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32.74 -18.86 -12.24  22.80  41.41
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.38208    0.47735  -9.180   <2e-16 ***
## extrakid_t_g -3.36146    1.93231  -1.740   0.0819 .
## color_t      11.27971    0.28820  39.139   <2e-16 ***
## age_t         0.78811    0.03306  23.837   <2e-16 ***
## hispan_t     -0.27968    0.33236  -0.841   0.4001
## orace_t       2.33572    0.23158  10.086   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.61 on 219994 degrees of freedom
## Multiple R-squared:  0.02484,    Adjusted R-squared:  0.02481
## F-statistic:  1121 on 5 and 219994 DF,  p-value: < 2.2e-16
```

```
resid_g<-fit9$residuals
test_g<-lm(resid_g~twogirl_t+color_t+age_t+hispan_t+orace_t)
summary(test_g)
```

```
##
## Call:
## lm(formula = resid_g ~ twogirl_t + color_t + age_t + hispan_t +
##     orace_t)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32.74 -18.86 -12.24  22.80  41.41
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.258e-12  4.214e-01       0        1
## twogirl_t   -2.647e-12  1.080e-01       0        1
## color_t      1.224e-13  2.056e-01       0        1
## age_t       -2.021e-14  1.368e-02       0        1
## hispan_t     2.030e-13  2.048e-01       0        1
## orace_t     -2.459e-13  2.225e-01       0        1
##
## Residual standard error: 21.61 on 219994 degrees of freedom
## Multiple R-squared:  2.929e-27,  Adjusted R-squared:  -2.273e-05
## F-statistic: 1.289e-22 on 5 and 219994 DF,  p-value: 1
```

```
fit10<-lm(extrakid_v~twogirl_v+color_v+age_v+hispan_v+orace_v)
extrakid_v_g<-fit7$fitted.values

valid_pre_g <- predict(fit9,newdata=data.frame(extrakid_t_g=extrakid_v_g,color_t=color_v,age_t=age_v,his
accuracy(valid_pre_g,workingweek_v)
```

```
##                 ME     RMSE      MAE  MPE MAPE
## Test set 0.2417249 21.48814 19.53808 -Inf  Inf
```

4. results analysis and Conclusion

```
accuracy(valid_pre,workingweek_v)
```

```
##                 ME     RMSE      MAE  MPE MAPE
## Test set 0.3852793 21.26799 19.18659 -Inf  Inf
```

```
accuracy(valid_pre_ins,workingweek_v)
```

```
##                 ME     RMSE      MAE  MPE MAPE
## Test set 0.3668273 21.48916 19.51076 -Inf  Inf
```

```
accuracy(valid_pre_b,workingweek_v)
```

```
##                 ME     RMSE      MAE  MPE MAPE
## Test set 0.5798222 21.49608 19.46993 -Inf  Inf
```

```
accuracy(valid_pre_g,workingweek_v)
```

```
##                 ME     RMSE      MAE  MPE MAPE
## Test set 0.2417249 21.48814 19.53808 -Inf  Inf
```

```
summary(fit3)
```

```
##
## Call:
## lm(formula = workingweek_t ~ extrakid_t_ins + color_t + age_t +
##     hispan_t + orace_t)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32.88 -18.63 -12.39  22.80  41.46
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -4.67436    0.45234 -10.334   <2e-16 ***
## extrakid_t_ins -5.86014    1.42664  -4.108    4e-05 ***
## color_t        11.54085    0.25398  45.440   <2e-16 ***
## age_t           0.82703    0.02610  31.691   <2e-16 ***
## hispan_t        0.05881    0.28158   0.209    0.835
## orace_t         2.41867    0.22749  10.632   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.61 on 219994 degrees of freedom
## Multiple R-squared:  0.0249, Adjusted R-squared:  0.02488
## F-statistic:  1123 on 5 and 219994 DF,  p-value: < 2.2e-16
```

```
summary(fit6)
```

```
##
## Call:
## lm(formula = workingweek_t ~ extrakid_t_b + color_t + age_t +
##     hispan_t + orace_t)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -32.77 -18.90 -12.28  22.88  41.50
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5.17198    0.57925  -8.929  < 2e-16 ***
## extrakid_t_b -10.11430    3.40627  -2.969  0.00298 **
## color_t       11.98545    0.41110  29.155  < 2e-16 ***
## age_t          0.89329    0.05479  16.303  < 2e-16 ***
## hispan_t       0.63511    0.50484   1.258  0.20838
## orace_t        2.55991    0.24960  10.256  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.61 on 219994 degrees of freedom
## Multiple R-squared:  0.02486,    Adjusted R-squared:  0.02484
## F-statistic:  1122 on 5 and 219994 DF,  p-value: < 2.2e-16
```

```r
summary(fit9)
```

```
##
## Call:
## lm(formula = workingweek_t ~ extrakid_t_g + color_t + age_t +
##     hispan_t + orace_t)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32.74 -18.86 -12.24  22.80  41.41
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.38208    0.47735  -9.180   <2e-16 ***
## extrakid_t_g -3.36146    1.93231  -1.740   0.0819 .
## color_t      11.27971    0.28820  39.139   <2e-16 ***
## age_t         0.78811    0.03306  23.837   <2e-16 ***
## hispan_t     -0.27968    0.33236  -0.841   0.4001
## orace_t       2.33572    0.23158  10.086   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.61 on 219994 degrees of freedom
## Multiple R-squared:  0.02484,    Adjusted R-squared:  0.02481
## F-statistic:  1121 on 5 and 219994 DF,  p-value: < 2.2e-16
```

From the above results we can see that all the three instruments are quiet good. And the captured effects are much different among different methods. We believe that the classic estimate is biased, the three IV estimates are seemingly unbiased and consistent, since they all satisfy the conditions for relevance and exogeneity.

The results above mainly lead to three conclusions:

1. compared to classic regression, the IV regression does not show great improvment in predictive accuracy.
2. on average, having extra kid reduces mother's working time by 5.86 weeks.
3. for mothers with two boys, having extra kid reduces working time by 10.11 weeks.
4. for mothers with two girls, having extra kid reduces working time by 3.36 weeks.

5. Interpretation

We can see that the predictive accuracy of all the four regressions are similar. But this is not absolute, since it can be due to this specific dataset. Also in capturing causal effect, this accuracy is not what we care most. Since we are trying to find causal effect here, instead of predicting, the the true effect matters most. Rigorously speaking, the IV estimates are still likely to be biased. Because there are sex-revealing technology available, so the estimates will be biased according to the parents' different sex preference. And the IV of "twoboy" and "twogirl" does show that the sex preference are not neutral regeading boys and girls. For mothers with two boys, the average decreased number of weeks is 10.11, while for mothers with two girls, the number is just 3.36. This huge difference indicates the high possibility of discrimination between genders inside the household. But sexual preference alone is also rather unlikely to generate so big a difference. Even with the sex-revealing technology, the cost of sex selection is generally high enough to block most thoughts of abortion. We can testify this from the mean of the variable "twoboys" and "twogirls"

```
mean(twoboy_t)
```

```
## [1] 0.2656955
```

```
mean(twogirl_t)
```

```
## [1] 0.2392227
```

We can see that the mean of the two variable is very close, suggesting that human control effect is not so big. So why the effects under this two conditions differ so much? I had two explanations:

1. average time input from mothers is much less for girls compared to boys.
2. elder sisters would share a good amount of burden for mothers in taking care of new babies.

This two reasons would explain the results that the effect of having extra kid conditional on two boys is larger than the average effect, while conditional on two girls the effect is much smaller. But this two explanations also point out the potential problem of a discriminative fostering within households. It may draw some attention regarding the issue of gender equality. Although the data is of year 1980, the problem it reflects may still exist today. So might leave it to further discussion. By someone else, of course.