homework7

Zhang Zhiyuan 15220162202517

2019/6/8

this exercise uses instrument variable to explore the effect of having more children on women's working time. it aims to find out the causal effect of having extra child or children on the working time of those women who already have two children.

the data is a small sample from the 1980 Census of America.

detailed data description:

Variable Description

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

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

workingweek<-data$weeksm1
extrakid<-data$morekids
color<-data$black
age<-data$agem1
hispan<-data$hispan
orace<-data$othrace
ssex<-data$samesex
```

first try to use the classic linear regression model to capture the causal effect.

```
fit1<-lm(workingweek~extrakid+color+age+hispan+orace)
summary(fit1)
```

```
##
## Call:
## lm(formula = workingweek ~ extrakid + color + age + hispan +
##     orace)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36.47 -18.04 -10.41  23.10  44.90
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.52276    1.12257  -4.029 5.62e-05 ***
## extrakid    -6.89600    0.25769 -26.761  < 2e-16 ***
## color       11.53423    0.55252  20.876  < 2e-16 ***
```

```
## age          0.84175    0.03685  22.843  < 2e-16 ***
## hispan      -0.24731    0.52149  -0.474    0.635
## orace        3.33469    0.58920   5.660 1.53e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.41 on 29994 degrees of freedom
## Multiple R-squared:  0.04767,    Adjusted R-squared:  0.04751
## F-statistic: 300.3 on 5 and 29994 DF,  p-value: < 2.2e-16
```

but through rational analysis we can expect this result to be highlt biased, since there are so many other factors that may affect working time and pregnancy, such as disease, education, income, etc. so, in order to get a better result, i tried to use the variable "ssex" as instrument.

```
fit2<-lm(extrakid~ssex+color+age+hispan+orace)
summary(fit2)
```

```
##
## Call:
## lm(formula = extrakid ~ ssex + color + age + hispan + orace)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6984 -0.3859 -0.3037  0.5791  0.8278
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.1730167  0.0252522  -6.852 7.45e-12 ***
## ssex         0.0678624  0.0055269  12.279  < 2e-16 ***
## color        0.0962641  0.0123369   7.803 6.24e-15 ***
## age          0.0164380  0.0008182  20.090  < 2e-16 ***
## hispan       0.1484327  0.0116251  12.768  < 2e-16 ***
## orace        0.0235263  0.0131685   1.787    0.074 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4786 on 29994 degrees of freedom
## Multiple R-squared:  0.02552,    Adjusted R-squared:  0.02536
## F-statistic: 157.1 on 5 and 29994 DF,  p-value: < 2.2e-16
```

```
extrakid_ins<-fit2$fitted.values
```

from model fit2 we can see that ssex does has predictive power on extrakid, with a t-statistic of over 12. and it is common sense that the sex of a child is purely random. so the two conditions for a good IV, relevance and exogeneity, are both astisfied.

```
fit3<-lm(workingweek~extrakid_ins+color+age+hispan+orace)
summary(fit3)
```

```
##
## Call:
## lm(formula = workingweek ~ extrakid_ins + color + age + hispan +
##     orace)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33.00 -18.82 -12.28  23.24  41.01
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.3703     1.2421  -3.519 0.000435 ***
## extrakid_ins  -5.7808     3.6869  -1.568 0.116914
## color         11.4263     0.6628  17.240  < 2e-16 ***
## age            0.8235     0.0708  11.631  < 2e-16 ***
## hispan        -0.4118     0.7567  -0.544 0.586313
## orace          3.3078     0.6027   5.488  4.1e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.67 on 29994 degrees of freedom
## Multiple R-squared:  0.02501,    Adjusted R-squared:  0.02485
## F-statistic: 153.9 on 5 and 29994 DF,  p-value: < 2.2e-16
```

from the IV result we can see, the true effect of extra child or children on women's working time is -5.78, which means the existence of extra child or children on average decreases mother's working time by 5.78 weeks, instead of -6.90, as suggested by classic biased regression.