iv_cz363

Chuyuan Zhong

2023-05-05

Question 1 Load the data into R and confirm that you have 100,000 observations and that the variables are lnearn (log earnings), female (indicator for female), S (years of schooling), xpr (years of experience), xpr2 (years of experience squared), and compulsoryS (years of compulsory schooling in the individual's state).

```
df <- read.csv('iv_problem_set.csv')
head(df)
```

```
##    S compulsoryS female       xpr      xpr2    lnearn
## 1 15          10      0 19.021330 361.81110  9.220398
## 2 17          12      1 10.137720 102.77330  9.762558
## 3 13          10      0 19.962130 398.48660  8.191334
## 4 11          10      0 27.220020 740.92930  4.216195
## 5 18           9      1  5.174110  26.77142 10.677460
## 6  9           9      0  9.748647  95.03611  9.396896
```

```
colnames(df)
```

```
## [1] "S"           "compulsoryS" "female"      "xpr"         "xpr2"
## [6] "lnearn"
```

```
cat("The dataset has", nrow(df), "observations and", ncol(df), "variables.")
```

```
## The dataset has 100000 observations and 6 variables.
```

Question 2 Regress log earnings on female, year of schooling, years of experience and years of experience squared. What is the impact on earnings of a 1 year increase in schooling? Can you reject the hypothesis that the return to schooling is 0?

```
ols <- lm(lnearn ~ S + female +xpr + xpr2, data = df)
olstable1 = data.frame(xtable(ols))
kable(olstable1, caption="OLS Results")
```

Table 1: OLS Results

	Estimate	Std..Error	t.value	Pr...t..
(Intercept)	6.9226136	0.0240561	287.76957	0
S	0.1513946	0.0015006	100.88631	0

	Estimate	Std..Error	t.value	Pr...t..
female	1.3215400	0.0076354	173.08067	0
xpr	0.0490351	0.0016386	29.92434	0
xpr2	-0.0079619	0.0000442	-180.04476	0

One year increase in schooling is associated with a 0.15 increase in log earnings. `S` has a p-value of 0, suggesting that we can reject the null hypothesis that the return to schooling is 0.

Question 3 Explain why this estimate may be biased due to endogeneity.

The endogeneity may from family background or ability, which can influence the level of schooling and subsequently affect earnings. For instance, individuals from higher-income families or those with social connections may be more likely to invest in education or receive a better education, leading to higher earnings.

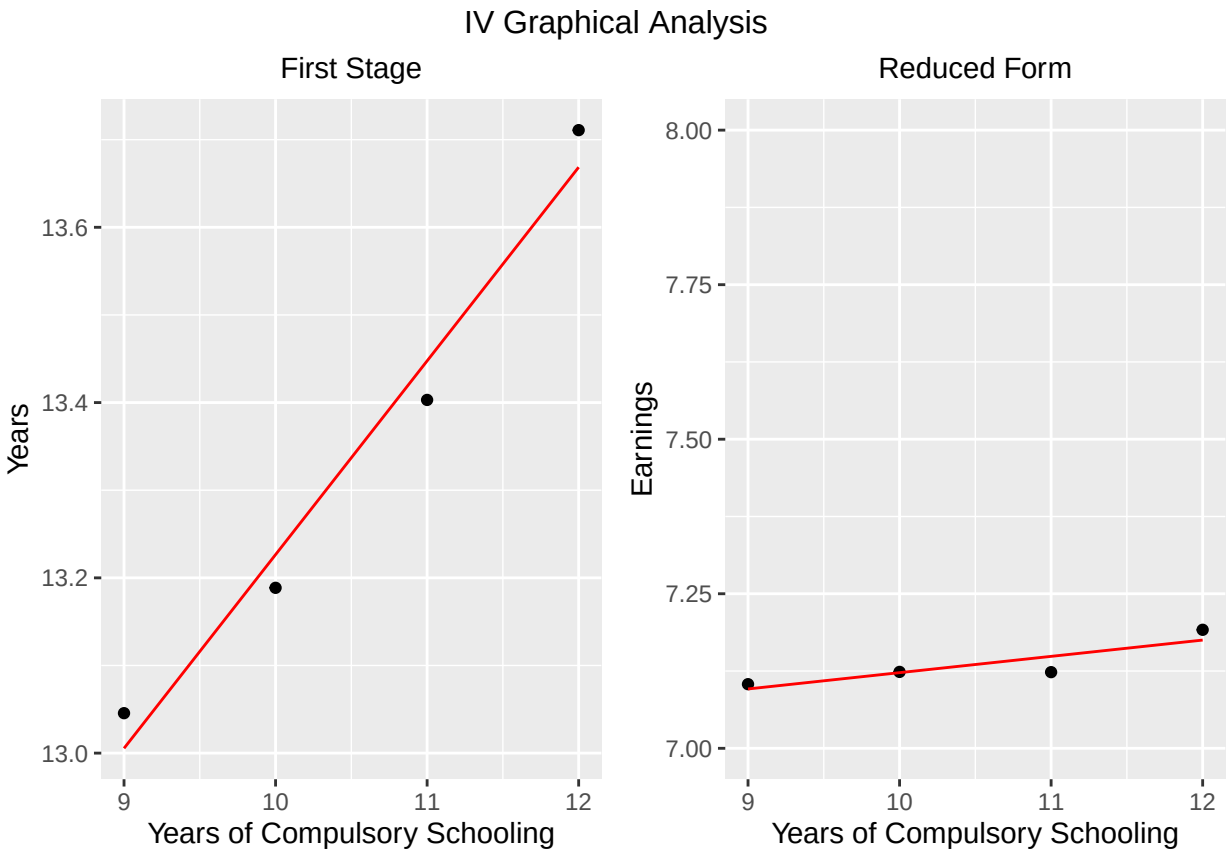
Question 4 Now suppose that we think state compulsory schooling is an instrument for years of schooling. Explain the intuition behind the statistical conditions that have to be satisfied for this variable to be a valid instrument for years of schooling.

The key intuition is that state compulsory schooling is not directly related to earnings, but rather its effect on earnings is through its influence on years of schooling. At the same time, state compulsory schooling should not be correlated with the error term in the earnings equation, as this would violate the exclusion restriction assumption and render the instrument invalid. It must be exogenous and only affect earnings through its impact on years of schooling.

Question 5 Present a graphical analysis to plot the first stage and reduced form results. How does this graphical analysis motivate the instrumental variables strategy?

```
c_means <- aggregate(df$compulsoryS,list(df$compulsoryS), mean)
S_mean <- aggregate(df$S,list(df$compulsoryS), mean)
learn_mean <-aggregate(df$lnearn,list(df$compulsoryS), mean)
meandata = data.frame(c_means[,2], learn_mean[,2], S_mean[,2])

s_dist_reg = lm(meandata[,3] ~ meandata[,1])
predict_s_dist <- data.frame(s_dist_pred = predict(s_dist_reg, meandata), meandata[,1])

p1 <- ggplot(meandata, aes(x=meandata[,1], y=meandata[,3])) + geom_point() +
  labs(x = "Years of Compulsory Schooling", y = "Years",title = "First Stage")+
  theme(plot.title = element_text(size = 11, hjust = 0.5))+
  geom_line(color='red',data = predict_s_dist, aes(x=meandata[,1], y=s_dist_pred))

lnearn_dist_reg = lm(meandata[,2] ~ meandata[,1])
predict_lnearn_dist <- data.frame(lnearn_dist_pred =
                                  predict(lnearn_dist_reg, meandata),
                                  meandata[,1])

p2 <- ggplot(meandata, aes(x=meandata[,1], y=meandata[,2])) + geom_point() +
  labs(x = "Years of Compulsory Schooling", y = "Earnings",title = "Reduced Form") +
  theme(plot.title = element_text(size = 11, hjust = 0.5)) +
  geom_line(color='red',data = predict_lnearn_dist, aes(x=meandata[,1],
                                                        y=lnearn_dist_pred)) +
```

```
  ylim(7, 8)
grid.arrange(p1, p2, nrow=1, top="IV Graphical Analysis")
```



Based on the graph, we can see that the instrumental variable is correlated with years of schooling (the dependent variable), and that years of compulsory schooling are positively associated with years of schooling but not correlated with earnings (the outcome) since they didn't change much.

Question 6 Estimate the first stage regression. Is compulsory schooling a statistically significant predictor of schooling?

```
fsm = lm(S ~ compulsoryS + female + xpr + xpr2, data = df)
fsm = data.frame(xtable(fsm))
kable(fsm, caption = 'First Stage Regression -- Years of Schooling')
```

Table 2: First Stage Regression – Years of Schooling

	Estimate	Std..Error	t.value	Pr...t..
(Intercept)	11.0234430	0.0919450	119.8917061	0.0000000
compulsoryS	0.2203517	0.0083646	26.3432717	0.0000000
female	-0.0049834	0.0160348	-0.3107887	0.7559619
xpr	-0.0015573	0.0034414	-0.4525292	0.6508888
xpr2	0.0000399	0.0000929	0.4294973	0.6675623

The p value of 0 suggests that compulsory schooling is statistically significant predictor of schooling.

Question 7 Use ivreg to implement the IV estimator in which we instrument for schooling using compulsory schooling. What are your results? How does the IV estimate for the return to schooling compare to the OLS estimate?

```
ivmodel<-ivreg(lnearn ~ female + xpr + xpr2 | S | compulsoryS, data = df)
summary(ivmodel)
```

```
##
## Call:
## ivreg(formula = lnearn ~ female + xpr + xpr2 | S | compulsoryS,
##     data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -5.155304 -0.816166 -0.001561  0.819575  5.728230
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  6.777e+00  2.414e-01   28.077   <2e-16 ***
## S            1.623e-01  1.808e-02    8.976   <2e-16 ***
## female       1.322e+00  7.638e-03  173.025   <2e-16 ***
## xpr          4.904e-02  1.639e-03   29.920   <2e-16 ***
## xpr2        -7.962e-03  4.423e-05 -179.996   <2e-16 ***
##
## Diagnostic tests:
##                     df1   df2 statistic p-value
## Weak instruments      1 99995   693.968  <2e-16 ***
## Wu-Hausman            1 99994     0.365   0.546
## Sargan                0    NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.208 on 99995 degrees of freedom
## Multiple R-Squared: 0.8169,  Adjusted R-squared: 0.8169
## Wald test: 1.09e+05 on 4 and 99995 DF,  p-value: < 2.2e-16
```

The results indicate a positive association between earnings and variables such as female, years of schooling, and experience since they are statistical significance. Compared to OLS, the p value 0.546 here does not reject the null hypothesis of exogeneity, suggesting a significant improvement in the consistency of the estimates.

Question 8 Suppose that we think the return to schooling varies by gender and we want to instrument for the female*S interaction term using compulsory schooling interacted with gender. Estimate the first stage regressions (note that we have 2 variables that need to be instrumented). Do we have a valid instrument for each variable that needs to be instrumented?

```
fsmodel2 <- lm(female*S ~ female*compulsoryS + xpr + xpr2, data = df)
fsmodel3 <- data.frame(xtable(fsmodel2))
kable(fsmodel3, caption = 'First Stage Regression -- Interaction')
```

Table 3: First Stage Regression – Interaction

	Estimate	Std..Error	t.value	Pr...t..
(Intercept)	0.0026693	0.0901086	0.0296229	0.9763679
female	10.9882009	0.1247729	88.0656302	0.0000000
compulsoryS	0.0000125	0.0083839	0.0014946	0.9988075
xpr	-0.0002405	0.0024353	-0.0987545	0.9213334
xpr2	0.0000036	0.0000657	0.0550929	0.9560646
female:compulsoryS	0.2221624	0.0118381	18.7667813	0.0000000

```
fsmodel4 <- lm(S ~ female*compulsoryS + xpr + xpr2, data = df)
fsmodel5 <- data.frame(xtable(fsmodel4))
kable(fsmodel5, caption = 'First Stage Regression -- Schooling')
```

Table 4: First Stage Regression – Schooling

	Estimate	Std..Error	t.value	Pr...t..
(Intercept)	11.0430609	0.1273347	86.7246799	0.0000000
female	-0.0440867	0.1763195	-0.2500389	0.8025578
compulsoryS	0.2184832	0.0118475	18.4413168	0.0000000
xpr	-0.0015584	0.0034414	-0.4528438	0.6506622
xpr2	0.0000399	0.0000929	0.4299836	0.6672085
female:compulsoryS	0.0037254	0.0167287	0.2226980	0.8237710

In the first regression, we observe that female and the interaction effect between female and years of compulsory schooling are statistically significant. However, the high p-values of `xpr`, `xpr2`, and `compulsoryS` suggest that they are not statistically significant, indicating that there is no significant relationship between these variables and the instrument.But `female` and the interaction between female and compulsory `female:compulsoryS` are statistically significant.

In the second regression, we can see that only years of compulsory schooling `compulsoryS` is statistically significant and it is positively correlated with years of schooling.

Question 9 Estimate the IV results related to the first stage regressions in (8). Can we reject the hypothesis that the IV estimate of the coefficient on female*S is 0? What are the conclusions about whether the return to schooling varies based on gender or not?

```
ivmodel1<-ivreg(lnearn ~ xpr + xpr2 | female*S | compulsoryS*female, data = df)
summary(ivmodel1)
```

```
##
## Call:
## ivreg(formula = lnearn ~ xpr + xpr2 | female * S | compulsoryS *
##     female, data = df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -5.1943454 -0.8125815 -0.0001014  0.8170772  5.6894355
##
## Coefficients:
```

```
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  7.093e+00  3.433e-01   20.664  < 2e-16 ***
## female       6.979e-01  4.801e-01    1.454    0.146
## S            1.386e-01  2.573e-02    5.387 7.18e-08 ***
## xpr          4.901e-02  1.633e-03   30.015  < 2e-16 ***
## xpr2        -7.961e-03  4.407e-05 -180.647  < 2e-16 ***
## female:S     4.681e-02  3.603e-02    1.299    0.194
##
## Diagnostic tests:
##                                df1   df2 statistic p-value
## Weak instruments (S)             2 99994   347.005  <2e-16 ***
## Weak instruments (female:S)      2 99994   353.308  <2e-16 ***
## Wu-Hausman                       2 99992     1.146   0.318
## Sargan                           0    NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.203 on 99994 degrees of freedom
## Multiple R-Squared: 0.8183,  Adjusted R-squared: 0.8183
## Wald test: 8.789e+04 on 5 and 99994 DF,  p-value: < 2.2e-16
```

The p-value of 0.194 indicates that the results are not statistically significant, so we fail to reject the null hypothesis. We can therefore conclude that the return to schooling does not vary based on gender.