Coding Sample: Instrumental Variable

Xiaotian Tang

2025-03-30

Promoting Environmentally Valuable Agricultural Livelihoods (PEVAL) has been charged with understanding how much water farmers can save by installing drip irrigation on their fields in Busia, Kenya. They are asking for your help before rolling out this program to the full country.

Q1

PEVAL are interested in answering the following question: *What is the average effect of drip irrigation systems on farmer water use (measured in acre-feet, where 1 acre-foot = 325,851 gallons)?* To make sure everybody is on the same page, explain to them what the ideal experiment would be for answering this question. Describe the dataset that you'd like to have to carry out this ideal experiment, and use math, words, and the potential outcomes framework to explain what you would estimate and how you would do so. Make sure to be clear about the unit of analysis (i.e., what is "i" here?).

An ideal experiment would randomly install drip irrigation systems for farmers in Busia, Kenya. We would need to have 100% compliance, meaning everyone installed the system actually uses it and everyone not installed does not. We would also require that the randomization resulted in the treatment and control group being balanced on all observable and unobservable covariates. In other words, potential farmer water use Y_{0i} and Y_{1i}, for each farmer need to be independent of whether they install the drip irrigation system or not.

The unit of analysis (each "row" in our data set) is an individual farmer (suppose every farmer has his own farm and every farm could install one drip irrigation system). We would calculate the ATE, $E(Y_{1i}) - E(Y_{0i})$, simply by calculating the difference in means between the treatment and control group's water usage.

Q2

PEVAL are on board with your explanation, but they unfortunately can't carry out a randomized experiment, because drip irrigation systems are expensive. They also don't think that a selection-on-observables approach will work (they're very sophisticated). Finally, they have only recently begun collecting data, so they only have one survey round to share with (no repeated observations). Given these limitations, describe the type of research design you would try to use to answer their question of interest. Be explicit about the assumptions required for this design to work, describing them in both math and words.

The research design I would try to use is Instrumental Variables (IV). I will find an IV to isolate exognenous variation of the treatment variable. This research design doesn't need to carry out a randomized experiment,and could avoid omitted variable bias, classical measurement error in treatment, and simultaneity. Despite we don't have panel/time series data, as long as the assumptions hold, we can have a "good" estimate.

We need two assumptions for the IV designs:

1. **First Stage Assumption (Relevance)** The instrumental variable Z_i and treatment variable D_i are related.
$$Cov(Z_i, D_i) \neq 0$$
2. **Exclusion restriction (Exogeneity)** The instrumental variable Z_i is exogenous in the model, Z_i only affects the potential farmer water use Y_i through treatment variable D_i(whether install the drip irrigation system).
$$Cov(Z_i, \varepsilon_i) = 0$$

Q3

PEVAL are interested in this research design. It sounds promising. They'd like you to propose a specific approach. Please describe a plausible instrumental variable you could use to evaluate the effect of drip irrigation on a farm on that farm's water use. Why is your proposed instrument a good one? Do you have any concerns about your ability to estimate the treatment effect using your instrument? If yes, why? If no, why not?

One plausible IV I could use is a lottery-based subsidy. This IV is a good one because it will satisfy the two key assumptions required for a valid IV, and it's cheaper than an RCT.

Relevance: Because drip irrigation systems are expensive, farmers who receive the subsidy are more likely to afford and therefore install such systems, ensuring a strong relationship between my IV Z_i and treatment variable D_i.

Exogeneity: The distribution of the subsidy through a lottery is random, which means it is not influenced by other factors that might affect water savings directly. This randomness ensures that the subsidy influences water saving outcomes ONLY through its impact on the installation of drip irrigation systems.

Feasibility: A subsidy covers only part of the financial cost of installing a drip irrigation system. it's less expensive than conducting an RCT research design, and therefore more practical.

Yes, I do have concerns about my ability to estimate the treatment effect.

First, despite that an IV design is less expensive than an RCT design, it still requires a great amount of expenditure.

Second, since the exclusion restriction is fundamentally untestable, I cannot guarantee that the random government subsidy will influence the water consumption only through drip irrigation system. For example, if some farmers use the subsidy to buy other tools that will help them reduce water consumption, this exogeneity assumption will not be satisfied.

Q4

PEVAL is intrigued by your approach. After an internal discussion, they've come back to you with great news! It turns out that one of the many well-meaning development economists at the University of Chicago ran a small pilot program where they randomly offered a drip irrigation subsidy to some farms as part of an attempt to publish a paper in the American Economic Review (it is still unpublished). With this new information, please describe to PEVAL how you would estimate the impacts of drip irrigation subsidies on a farm's water consumption, and then how you would estimate the impact of drip irrigation at a farm on water usage. Use both words and math.

I would estimate the impacts of drip irrigation subsidies on a farm's water consumption by estimating the difference in potential farm's water consumption if the farm is provided with a subsidy or not for a particular status of whether installing the irrigation system. In math, suppose $Y_i(D_i, Z_i)$ is the outcome as a function of treatment and instrument. I estimate:

$$Y_i(D_i, Z_i = 1) - Y_i(D_i, Z_i = 0)$$

I would estimate the impact of drip irrigation at a farm on water usage by estimating the difference in potential farm's water consumption if the farm installed the irrigation system or not for a particular status of whether provided with subsidies. In math, I estimate:

$$Y_i(D_i = 1, Z_i) - Y_i(D_i = 0, Z_i)$$

Q5

PEVAL agree that your approach is a good one. So good, in fact, that they'd like to see it in action! They are willing to share some data with you, in the form of ps3_data_24.csv. Please report the results of a regression that recovers the impact of drip irrigation subsidies on the adoption of drip irrigation at a farm, using drip_subsidy_amount as the subsidy variable (note that this pilot randomly gave different subsidy amounts to different households, reported in dollars, this is not just a binary variable) and drip_adoption as an indicator for having installed drip irrigation. What parameter does this estimate? Interpret your estimate. Will this pilot program be useful for measuring the effects of drip irrigation take-up on water usage? Why or why not?

In this Linear Probability Model, the parameter here represents the probability increase to install drip irrigation system, given different levels of subsidies (in dollar), comparing to the baseline subsidy of $100. Specifically, on average, comparing to an $100 offer:
A $200 subsidy increases the probability of adopting drip irrigation by 1.60%.
A $300 subsidy increases the probability of adopting drip irrigation by 22.20%.
A $400 subsidy increases the probability of adopting drip irrigation by 40.20%.
A $500 subsidy increases the probability of adopting drip irrigation by 58.08%.

The p-value of the factor(drip_subsidy_amount==200) is 0.491, which is insignificant, but the F-stats is 232.3, larger than 20, showing that different level of drip subsidy has a jointly significant impact on drip irrigation system adoption. Besides, the sign of our estimation is intuitively right. It fits our assumption that subsidy will increase the drip irrigation adoption well. Therefore, **the pilot program will be useful for measuring the effects of drip irrigation take-up on water usage**. The first stage assumption is satisfied, as long as it also satisfy the exclusive restriction assumption, it will be a good IV.

Note that here I choose to consider the government subsidy as discrete variable for 3 reasons:
(1) I suspect that different subsidy amounts might have non-linear effects on farmers' decisions to install drip irrigation systems.
(2) There could be a threshold effect, only once a certain amount of subsidy is reached do farmers choose to install their drip irrigation systems.
(3) Intuitively, Government subsidies are not issued in exact single units.

```r
# import data
data <- read.csv("ps3_data_24.csv")

# regression recover the impact of subsidies on drip irrigation system adoption.
first_stage <- lm(drip_adoption ~ factor(drip_subsidy_amount), data = data)
summary(first_stage)
```

```
##
## Call:
## lm(formula = drip_adoption ~ factor(drip_subsidy_amount), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5808 -0.2220 -0.0160  0.0000  0.9840
```

```
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -1.975e-17  1.644e-02   0.000    1.000
## factor(drip_subsidy_amount)200  1.600e-02  2.324e-02   0.689    0.491
## factor(drip_subsidy_amount)300  2.220e-01  2.324e-02   9.555   <2e-16 ***
## factor(drip_subsidy_amount)400  4.020e-01  2.324e-02  17.301   <2e-16 ***
## factor(drip_subsidy_amount)500  5.808e-01  2.322e-02  25.011   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3672 on 2495 degrees of freedom
## Multiple R-squared:  0.2713, Adjusted R-squared:  0.2702
## F-statistic: 232.3 on 4 and 2495 DF,  p-value: < 2.2e-16
```

Q6

PEVAL want you to use the pilot for the next steps of your analysis (they are ignoring any opinion you gave in (5), good or bad – welcome to the real world). They'd first like to learn a little bit more about the pilot itself: they want you to use a regression to estimate the effects of the drip irrigation subsidies (using drip_subsidy_amount, which is reported in dollars) on farm water usage (farm_water, which is in acre-feet). What parameter does this regression estimate? Interpret your estimate. Is it useful for policy? Why or why not?

In this regression model, the parameter estimates the savings on average in farm water usage (in acre-feet) given different level of subsidies (in dollars), comparing to the baseline subsidy of $100.

Specifically, on average, comparing to the baseline subsidy of $100: A $200 subsidy is associated with 3.460 acre-feet of farm water saving.
A $300 subsidy is associated with 6.063 acre-feet of farm water saving.
A $400 subsidy is associated with 9.690 acre-feet of farm water saving.
A $500 subsidy is associated with 14.335 acre-feet of farm water saving.

The F-stats here is 22.55 > 20, showing that different level of subsidies has a jointly significant association with farm water saving. The signs of the estimation are all positive, which is intuitively right, because we previously assume that subsidy as a positive association with water saving.

This is useful for policy because in real world not everyone will follow the policy directive, and not all individuals assigned to the treatment group will be compliers. Therefore, this estimation is a more realistic measure of a policy impact, preventing policy makers from an overestimation of the treatment efficacy.

```r
# estimate the effects of the subsidies on farm water usage.
r_form <- lm(farm_water ~ factor(drip_subsidy_amount), data = data)
summary(r_form)
```

```
##
## Call:
## lm(formula = farm_water ~ factor(drip_subsidy_amount), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.409 -20.440   0.391  20.959  58.106
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      53.409      1.169  45.684  < 2e-16 ***
## factor(drip_subsidy_amount)200   -3.460      1.653  -2.094 0.036373 *
## factor(drip_subsidy_amount)300   -6.063      1.653  -3.669 0.000249 ***
## factor(drip_subsidy_amount)400   -9.690      1.653  -5.864 5.12e-09 ***
## factor(drip_subsidy_amount)500  -14.335      1.652  -8.679  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.12 on 2495 degrees of freedom
## Multiple R-squared:  0.03489,    Adjusted R-squared:  0.03335
## F-statistic: 22.55 on 4 and 2495 DF,  p-value: < 2.2e-16
```

Q7

Finally, PEVAL wants you to use the pilot to estimate the impacts of the installation of drip irrigation at a farm on water use. For full transparency, make sure to show all of your analysis steps. PEVAL cares about your standard errors here, so be sure to get them right. Interpret your results: Does drip irrigation matter for water consumption?

Below are my analysis steps:

First, I check the relationship between drip_adption and fram_water without IV, to see if there is a statistically significant correlation.

Second, I check the first stage regression to see if my IV satisfy the relevance assumption. Notice the F-stats here, F-stats = 232.3 > 20, meaning that our relevance assumption is satisfied. Also, the sign here are all positive, which is intuitively right, because we assume that subsidy can serve as incentive to install drip irrigation system.

We cannot exam the exclusion restriction assumption, but intuitively, since the pilot gave the subsidy amounts to different households randomly, we are quite sure that it will influence our outcome (water saving) only through the treatment (drip irrigation installation).Therefore, drip_subsidy amount can be a valid IV.

Third, I check if my IV is a weak IV. By using the reduced form, I find the F-stats = 22.55 > 20, parameters on different level of subsidy all share the same negative sign, further showing that this IV is not a weak one.

Lastly, I use a canned routine to make sure I get the right standard variable. I use the AER package in R to do that. From my estimation, **drip irrigation matters for water consumption**. The causal effect of the adoption of drip irrigation system on farm water saving is 21.687 acre-feet. In other words, on average, farmers can save 21.687 acre-feet of water by installing drip irrigation on their fields. The standard error we estimate is 2.7458308, meaning that the 95% confidence interval is [-27.06911, -16.30565]. Our estimation is statistically significant.

```r
# First obtain the OLS estimate (second stage regression)
comprehension_reg <- lm(farm_water ~ drip_adoption, data = data)
summary(comprehension_reg)
```

```
##
## Call:
## lm(formula = farm_water ~ drip_adoption, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.529 -20.114  -1.298  17.492  69.446
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    42.0683     0.5819   72.30   <2e-16 ***
## drip_adoption  18.9257     1.1771   16.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.29 on 2498 degrees of freedom
## Multiple R-squared:  0.09379,    Adjusted R-squared:  0.09343
## F-statistic: 258.5 on 1 and 2498 DF,  p-value: < 2.2e-16
```

```
# Check if my IV satisfy the relevance assumption. (first stage regression)
relevance_check <- lm(drip_adoption ~ factor(drip_subsidy_amount), data = data)
summary(relevance_check)
```

```
##
## Call:
## lm(formula = drip_adoption ~ factor(drip_subsidy_amount), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5808 -0.2220 -0.0160  0.0000  0.9840
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -1.975e-17  1.644e-02   0.000    1.000
## factor(drip_subsidy_amount)200  1.600e-02  2.324e-02   0.689    0.491
## factor(drip_subsidy_amount)300  2.220e-01  2.324e-02   9.555   <2e-16 ***
## factor(drip_subsidy_amount)400  4.020e-01  2.324e-02  17.301   <2e-16 ***
## factor(drip_subsidy_amount)500  5.808e-01  2.322e-02  25.011   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3672 on 2495 degrees of freedom
## Multiple R-squared:  0.2713, Adjusted R-squared:  0.2702
## F-statistic: 232.3 on 4 and 2495 DF,  p-value: < 2.2e-16
```

```
# Check if it's a weak IV. (reduced form)
weak_IV_reg <- lm(farm_water ~ factor(drip_subsidy_amount),data = data)
summary(weak_IV_reg)
```

```
##
## Call:
## lm(formula = farm_water ~ factor(drip_subsidy_amount), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.409 -20.440   0.391  20.959  58.106
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      53.409      1.169  45.684  < 2e-16 ***
## factor(drip_subsidy_amount)200   -3.460      1.653  -2.094 0.036373 *
## factor(drip_subsidy_amount)300   -6.063      1.653  -3.669 0.000249 ***
## factor(drip_subsidy_amount)400   -9.690      1.653  -5.864 5.12e-09 ***
## factor(drip_subsidy_amount)500  -14.335      1.652  -8.679  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.12 on 2495 degrees of freedom
## Multiple R-squared:  0.03489,    Adjusted R-squared:  0.03335
## F-statistic: 22.55 on 4 and 2495 DF,  p-value: < 2.2e-16
```

```
# estimate the impacts of the installation of drip irrigation on water use.
iv_reg <- ivreg(farm_water ~ drip_adoption | factor(drip_subsidy_amount), data = data)
summary(iv_reg)
```

```
##
## Call:
## ivreg(formula = farm_water ~ drip_adoption | factor(drip_subsidy_amount),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -51.9941 -25.1789  -0.8416  25.5407  63.8673
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     51.994      0.910  57.135  < 2e-16 ***
## drip_adoption  -21.687      2.746  -7.898  4.2e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.73 on 2498 degrees of freedom
## Multiple R-Squared: -0.3381, Adjusted R-squared: -0.3386
## Wald test: 62.38 on 1 and 2498 DF,  p-value: 4.199e-15
```

```
confint(iv_reg)
```

```
##                   2.5 %    97.5 %
## (Intercept)    50.21051  53.77773
## drip_adoption -27.06911 -16.30565
```

Q8

PEVAL like your analysis and are now intrigued by using the subsidy as a policy tool. They would like to know the causal effect of the subsidy itself on water usage (that is, you should ignore irrigation take up from here on). but they're a bit worried about the quality of their data. They are concerned that their water usage data, which came from a household survey, includes random noise. Is this likely to be a problem for your analysis? Why or why not? If yes, can you use an instrumental variables approach to address this problem?

If the data only contain random noise in water usage, this is not a problem for my analysis. Our estimation will not be biased, the only bad thing is our precision decreases.

Suppose we want to estimate $Y_i = \alpha + \tau D_i + \varepsilon_i$, where Y_i is water usage for i farmer and D_i is subsidy the policy tool. We assume that $Cov(D_i, \varepsilon_i) = 0$, aka Treatment is as good as random.

But now we have random noise on Y_i, therefore we cannot observe the true Y_i, but rather $\tilde{Y_i}$, which we have $\tilde{Y_i} = Y_i + \gamma_i$

By saying random noise, we assume that Measurement error γ_i is not correlated with either outcome or treatment. In math, this could be expressed as $Cov(\gamma_i, Y_i) = 0$ and $Cov(\gamma_i, D_i) = 0$.

So, our estimation of Y_i could be expressed as $Y_i = \alpha + \tau D_i + \varepsilon_i - \gamma_i$, but since we already know that the measurement error in Y_i is not correlated with treatment, the assumption $Cov(D_i, (\varepsilon_i - \gamma_i)) = 0$ still holds. Therefore, our causal estimation τ will not be biased, but since new noise added, the standard error of our estimation will rise.

Q9

It turns out the water usage survey data were of extremely high quality, but the pilot research team did not keep great records of the subsidy amounts. The development economist describestwo possibilities and wants your opinion. First, she wants to know if it would be a problem for your analysis, looking at the causal effect of the subsidy on water usage, if the subsidy amount data included random noise? Why or why not? If yes, can you use an instrumental variables approach to address this problem? Second, she wants to know if it would be a problem for your analysis if larger subsidy amounts were recorded with error, but smaller subsidy amounts were accurately recorded. If yes, can you use an instrumental variables approach to address this problem?

For the first question:

If the subsidy amount data included random noise, it will cause attenuation bias (estimated coefficient biased towards 0) to our estimation. **This will be a problem for my analysis**, but (at least theoretically) **we can use an instrumental variables approach to address it** (Of course we need more data provided).

Suppose we want to estimate $Y_i = \alpha + \tau D_i + \varepsilon_i$, where Y_i is water usage for i farmer and D_i is subsidy the policy tool. We assume that $Cov(D_i, \varepsilon_i) = 0$, aka Treatment is as good as random.

But now we have random noise on D_i, therefore we cannot observe the true D_i, but rather $\tilde{D_i}$, which we have $\tilde{D_i} = D_i + \gamma_i$.

By saying random noise, we assume that Measurement error γ_i is not correlated with either outcome or treatment. In math, this could be expressed as $Cov(\gamma_i, Y_i) = 0$ and $Cov(\gamma_i, D_i) = 0$. And we further assume that measurement error is not in our original error term, therefore, $Cov(\gamma_i, \varepsilon_i) = 0$.

Since now what we estimate is $Y_i = \alpha + \tau \tilde{D_i} + \varepsilon_i$, with basic statistic knowledge we can know that $\hat{\tau} = \frac{Cov(Y_i, \tilde{D_i})}{Var(\tilde{D_i})}$, which, after simplification, becomes the equation: $\hat{\tau} = \tau(\frac{Var(D_i)}{Var(D_i) + Var(\gamma_i)})$, easy to know that $\frac{Var(D_i)}{Var(D_i) + Var(\gamma_i)}$ is smaller than one. so even with random noise, our estimation still be biased. And also we can prove that the standard error will increase as well.

We can use IV approach to address it. Because IV was designed to solve selection on unobservables. What IV do is isolating the exognenous variation from D_i and use it to estimate the causal impact. Therefore, random noise will surely be isolated with the help of an IV approach. Focusing on the context, if we can find an IV that influence the water saving only through subsidy, we can use it to solve the classical measurement error on X.

For the second question:

Yes, if the measurement error is systematic (Measurement error in $\tilde{D_i}$ is correlated with treatment D_i), **this will be a problem for my analysis, our estimation will also be biased (and our standard error will rise too!)** This time it's not simply attenuate $\hat{\tau}$, but can actually flip the sign fo $\hat{\tau}$ relative to τ (depends on the sign of $Cov(D_i, \gamma_i)$).

Most of the time we cannot use an IV approach to address this problem. Recall that we estimate $\tilde{D_i} = D_i + \gamma_i$, when we try instrumenting $\tilde{D_i}$ with Z_i, we have $Z_i = D_i + \zeta_i$, and we assume ζ_i is not correlated with D_i, Y_i, and γ_i. We have: $Cov(D_i, \zeta_i) = 0$, $Cov(Y_i, \zeta_i) = 0$, and $Cov(\gamma_i, \zeta_i) = 0$

But now $Cov(D_i, \gamma_i) \neq 0$, while instrument's measurement error is uncorrelated with D_i and γ_i, the actual instrument is correlated with them, so as soon as you use Z_i to shift D_i, you also shift γ_i. Therefore, IV cannot isolate variation in D_i from noise, and our IV approach will not work.

However, there is one special case. If and only if Z_i doesn't correlated with γ_i. In other words, the IV doesn't covary the measurement error, even though it is correlated with $\tilde{D_i}$ and $\tilde{D_i}$ is correlated with measurement error γ_i.

Now focus on the context, if we can find an IV that influence the water saving only through subsidy, and it has nothing to do with the measurement error that larger subsidy were recorded with error. We can use this IV to solve the measurement error.