PPHA 34600 Program Evaluation

Problem Set 3

5/12/2022

(1) CALBEARS are interested in answering the following question: What is the effect of average groundwater costs between April and September (measured in dollars per acre-foot) on total groundwater consumption (measured in acre-feet) during the same time period? 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 (ie, what is "i" here?).

In the ideal experiment, to find the effect of *average groundwater costs* on total groundwater consumption, we would conduct a Randomized Control Trial where we would randomize a shock in groundwater costs among farmers, for example by varying electricity prices or distributing efficient water pumps among farmers.

Here, treatment would be the shock/change in groundwater costs and the control group would be composed of farmers whose groundwater extraction costs remained unchanged.

Through a well-conducted randomization that results in balanced baseline characteristics, farmers in the treatment group would be statistically similar to farmers in the control group, allowing us to recover the Average Treatment Effect (ATE) of groundwater costs between April and September on consumption during the same period:

$$\tau^{ATE} = E[Y_i(1)] - E[Y_i(0)]$$

Where:

- Y is the outcome variable of interest, groundwater consumption (measured in acre-feet)
- i is each individual farmer

Under randomization, the **expected** groundwater consumption for farmers who are subject to a shock in groundwater costs is equivalent to the **mean** groundwater consumption for those same farmers, and viceversa for the control group. This property enables us to estimate the ATE by simply taking the difference in means between both groups:

$$\tau^{ATE} = \overline{Y_i}(D = 1) - \overline{Y_i}(D = 0)$$

Where:

- D refers to treatment status: $D = 1$ for farmers subject to a shock in consumption prices and $D = 0$ for farmers who are not
- Y groundwater consumption (measured in acre-feet)
- i is each individual farmer

The ideal dataset would contain **1)** baseline and endline farmer groundwater consumption (measured in acre feet), **2)** groundwater extraction costs and its cost determinants including pump efficiency, electricity prices and aquifer depth, and, **3)** other baseline characteristics needed to check for balance in the randomization process to ensure treated and control farmers are statistically similar. We would need this data at the farmer level, covering the time period between April and September. For the baseline characteristics, we would need the data to be collected before the experiment starts, for example in March.

(2) CALBEARS are on board with your explanation, but, as they've discussed with you, they won't be able to implement your preferred solution. They don't think that a selection-on-observables approach will work (they're very sophisticated). They're also limited by state privacy laws: they will only be able to give you one wave of data (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.

In the ideal RCT experiment, we can recover τ^{ATE} with a simple difference in means or refression because the key assumption $E[\varepsilon \mid D_i] = 0$ holds.

However, in the absence of randomization, $E[\varepsilon \mid D_i] \neq 0$, which means that the error term, conditional on treatment is non-zero. Therefore, we have selection bias and our τ^{ATE} would be biased. Other reasons why $E[\varepsilon \mid D_i] \neq 0$, include omitted variable bias and/or reverse causality.

As a solution to this problem, and given that CALBEARS provides me with *cross-sectional data*, I would propose implementing an Instrumental Variable approach, part of the Selection on Unobservables (SOU) toolkit.

This design requires finding an appropriate instrument, which is essentially a source of "good" variation in the treatment status (ie. change groundwater costs) and is *quasi random.* Conceptually, we need to isolate exogenous variation by separating selection into groundwater costs D in two parts:

$$D_i = B_i\varepsilon_i + C_i$$

Where:

- D is the treatment status, $D = 1$ for farmers subject to a shock in consumption prices and $D = 0$ for farmers with unchanged costs
- $B_i\varepsilon_i$ refers to the self-selection into groundwater costs (treatment) that is correlated with the individual farmer's choice or decision-making
- C_i is an arbitrary component of selection into groundwater costs (treatment)
- i is each individual farmer

Hypothetically, if we could observer C_i, we could regress groundwater consumption Y_i on it to recover τ^{ATE}. In reality, we cannot statistically distinguish C_i from $B_i\varepsilon_i$. Instead, we can carefully select an instrumental variable Z_i that varies the C_i component and is uncorrelated with the error term ε_i.

Assumptions required for an IV approach

1. First Stage: The instrument Z_i needs to be correlated with treatment status D_i, (ie. groundwater costs) which means their covariance is non-zero:

$$Cov(Z_i, D_i) \neq 0$$

$\longrightarrow$ This first assumption is **testable**.

2. Exclusion Restriction: The instrument is *as good as randomly assigned* because if it only affects groundwater consumption Y_i through treatment, ie. groundwater costs D_i.

$$Z_i \rightarrow D_i \rightarrow Y_i \tag{1}$$

And is not directly related to groundwater consumption Y_i:

$$Z_i \not\rightarrow Y_i \tag{2}$$

This occurs when the instrument Z_i is uncorrelated with the error term ε_i.

$$Cov(Z_i, \varepsilon_i) = 0$$

$\longrightarrow$ This second assumption is fundamentally **untestable**.

(3) CALBEARS 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 the cost of groundwater pumping on acre-feet of groundwater consumption. 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?

Groundwater extraction costs is a function of 1) electricity prices, 2) water pump efficiency and 3) aquifer depth. Potentially, farmers can be invited to participate in a lottery of efficient groundwater pumps distribution as part of a giveaway state program.

Since the lottery to distribute efficient pumps would be random, the lottery itself could be a plausible instrument to evaluate the effect of groundwater costs on groundwater consumption. For this instrument to be valid, it needs to satisfy both identifying assumptions.

1) Assessing the First Stage Assumption The giveaway lottery Z_i would be correlated with treatment D_i, where treatment would be *lower groundwater costs as a function of efficient water pump usage*, because only farmers that participate in the lottery would have access to a free efficient pump, which would lower groundwater extraction costs.

2) Assessing the Exclusion Restriction Assumption If the lottery is in fact random, this assumption is likely to hold. We can expect the lottery to affect groundwater assumption **only** through the distribution of efficient water pump. There is no reason to believe that the lottery itself would be correlated with groundwater consumption.

I wouldn't be concerned with my ability to estimate τ^{ATE} because both requires assumptions are likely to hold for this plausible instrument, making it *as good as randomly assigned* to recover an unbiased estimate of $\hat{\tau}^{IV}$. As long as the lottery is random and transparent, there shouldn't be ethical concerns of why some farmers receive a free efficient water pump over others. The general public may believe that only farmers with lower economic means should be considered for a giveaway, but this depends on how the program is pitched and implemented.

(4) CALBEARS is intrigued by your approach. After an internal discussion, they've come back to you with great news! It turns out that two of the California utilities ran a small pilot program where they randomly varied electricity prices to different farms as part of a new policy proposal. With this new information, please describe to CALBEARS how you would estimate the impacts of electricity prices on groundwater consumption, and how you would estimate the impacts of groundwater costs on groundwater consumption. Use both words and math.

Given this information, the random variation in electricity prices could be used as an instrumental variable to measure the effect of groundwater costs on groundwater consumption.

We can do this by employing two approaches **A) Two Stage Least Squares (2SLS)** or **B) Reduced Form**.

A) 2SLS APPROACH: This method consists of the following two steps.

1. **First Stage** We run a regression to estimate the effect of our instrument Z_i (ie. randomized electricity price variation) on treatment D_i (ie. groundwater cost variation).

$$D_i = \alpha + \gamma Z_i + \beta X_i + \eta_i$$

Where:

- D is the treatment status $D = 1$ groundwater cost variation and $D = 0$ if no variation
- Z_i is the instrument, randomized electricity price variation
- X_i refers to baseline covariates
- i is each individual farmer

After running the regression, we have to store the predicted values of D_i and $\hat{D}_i$. Next, we check if the F-stat > 20 to confirm whether the **First Stage Assumption**, ie. $Cov(Z_i, D_i) \neq 0$.

2. **Second Stage** We regress the effect of the "good variation" in treatment $\hat{D}_i$ (ie. the change in groundwater costs that is random, akin to C_i), on our outcome, Y_i (ie. groundwater consumption).

$$Y_i = \alpha + \tau \hat{D}_i + \delta X_i + \varepsilon_i$$

Where:

- τ is the instrumental variable estimate
- Y_i is the outcome variable, groundwater consumption
- $\hat{D}_i$ is treatment, stripped of selection bias
- Z_i is the instrument, ie. randomized electricity price variation
- X_i refers to baseline covariates
- i is each individual farmer

By the exclusion restriction assumption, the instrument is *as good as randomly assigned.*

With these two assumption in place, 2SLS lets us recover an unbiased estimate of $\hat{\tau}^{IV}$.

$\longrightarrow$ When running the 2SLS method in a statistical software, the standard errors need to be adjusted because the software does not know this regression uses predictions from a previous regression.

A) REDUCED FORM APPROACH:

Unlike the 2SLS approach, which gives us τ^{IV}, the Reduced Form indicates how groundwater consumption Y_i varies with randomized electricity variation, our instrument Z_i.

The Reduced Form can be mathematically expressed as:

$$Y_i = \alpha + \theta Z_i + \pi X_i + \eta_i$$

Where:

- θ is the effect of the instrument (randomized electricity price variation) on the outcome (groundwater consumption)
- Y_i is the outcome variable, groundwater consumption
- Z_i is the instrument, ie. randomized electricity price variation
- X_i refers to baseline covariates
- i is each individual farmer

The Reduced Form (RF) should have a causal interpretation and can be very informative.

(5) CALBEARS 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.csv. Please report the results of an analysis of the impact of electricity prices on groundwater costs, using electricity_price_pilot as the price variable and groundwater_cost as the cost variable. What parameter does this regression estimate? Interpret your estimate. Will this utility pilot be a helpful way forward to estimating the impacts of groundwater costs on groundwater usage? Why or why not?

```r
# Regress electricity price variation on groundwater costs
reg <- lm(groundwater_cost ~ electricity_price_pilot, data = data)
summary(reg)
```

```
##
## Call:
## lm(formula = groundwater_cost ~ electricity_price_pilot, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -278.11 -131.85  -42.74  104.32  796.25
##
## Coefficients:
##                          Estimate Std. Error t value            Pr(>|t|)
## (Intercept)             291.17320    3.51851   82.75 <0.0000000000000002 ***
## electricity_price_pilot   0.74022    0.04326   17.11 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 176.8 on 3998 degrees of freedom
## Multiple R-squared:  0.06823,    Adjusted R-squared:  0.06799
## F-statistic: 292.7 on 1 and 3998 DF,  p-value: < 0.00000000000000022
```

This regression provides an estimate of $\hat{\gamma}$, or the effect of randomized electricity price variation (ie. our instrument) on groundwater costs, our treatment. In other words, this is a **First Stage** regression.

Every additional USD dollar in electricity price variation, is associated with a 0.74 dollar increase per acre-foot in groundwater cost. This estimate is highly statistically significant at all conventional levels of significance, since the p-value is very small.

Given that every additional dollar increase in electricity price is estimated to increase groundwater costs by 74 cents, this is a considerable amount and reflects that groundwater costs is in fact a function of electricity prices.

The F-statistic is $292.7 > 20$, so the instrument passes the test first stage assumption that the instrument (ie. randomized electricity priced variation) is correlated with treatment (ie. groundwater costs). As to the exclusion restriction, since we are told the variation in electricity prices was random, we can also assume it holds.

Therefore, utility pilot is a valid instrument and is in fact helpful in estimating the impacts of groundwater costs on groundwater usage.

(6) CALBEARS wants you to use the pilot in your analysis (they are ignoring any opinion you gave in (5), good or bad). Please report the results of an analysis of the impact of electricity prices on groundwater consumption, using electricity_price_pilot as the price variable and groundwater_use as the usage variable. What parameter does this regression estimate? Interpret your estimate. Is this estimate useful for policy? Why or why not?

```
# Regress electricity prices on groundwater costs
reg2 <- lm(groundwater_use ~ electricity_price_pilot, data = data)
summary(reg2)
```

```
##
## Call:
## lm(formula = groundwater_use ~ electricity_price_pilot, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -105657  -53706  -10665   46909  191948
##
## Coefficients:
##                          Estimate Std. Error t value            Pr(>|t|)
## (Intercept)             108301.53    1307.49  82.831 <0.0000000000000002 ***
## electricity_price_pilot   -147.50      16.08  -9.175 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65720 on 3998 degrees of freedom
## Multiple R-squared:  0.02062,    Adjusted R-squared:  0.02037
## F-statistic: 84.17 on 1 and 3998 DF,  p-value: < 0.00000000000000022
```

This regression provides a * $\hat{\theta}$ estimate, which captures the effect of the instrument (randomized electricity price variation) on the outcome (groundwater consumption). As described in question (4), this is the **Reduced Form**.

Every additional USD dollar in electricity price variation, is associated with a - 147.50 acre-feet decrease in groundwater consumption. This $\hat{\theta}$ estimate is highly statistically significant at all conventional levels of significance. The sign of the estimate is logical by economic theory, since we can expect an increase in electricity prices to reduce consumption.

On the contrary, in a context where the utilities that ran the pilots do not hold an electricity monopoly of in the state, we would expect a substitution effect where farmers switch to cheaper electricity providers. However, assuming the utilities are the main providers, then the negative sign of the Reduced Form estimate is coherent.

This estimate is useful for public policy. Even though economic theory suggests a negative correlation between electricity prices and groundwater consumption, this estimate provides statistical evidence in support of this claim. This could be helpful in lobbying for electricity-varying policies to reduce groundwater consumption in anticipation of the expected drought.

(7) CALBEARS would like you to use their pilot to estimate the effect of groundwater costs on groundwater consumption. For full transparency, make sure to show all of your analysis steps. CALBEARS cares about your standard errors, so using a canned routine is a good idea here. Interpret your effect. Do groundwater costs matter for consumption?

First we obtain the **First Stage**:

```
# First stage regression
first_stage <- lm(groundwater_cost ~ electricity_price_pilot , data = data)
```

```
predicted_y <- fitted(first_stage)
gamma <- first_stage$coefficients[[2]]
cat("The effect of instrument on treatment is",gamma)
```

```
## The effect of instrument on treatment is 0.7402173
```

```
summary(first_stage)
```

```
## 
## Call:
## lm(formula = groundwater_cost ~ electricity_price_pilot, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -278.11 -131.85  -42.74  104.32  796.25 
## 
## Coefficients:
##                          Estimate Std. Error t value            Pr(>|t|)    
## (Intercept)             291.17320    3.51851   82.75 <0.0000000000000002 ***
## electricity_price_pilot   0.74022    0.04326   17.11 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 176.8 on 3998 degrees of freedom
## Multiple R-squared:  0.06823,    Adjusted R-squared:  0.06799 
## F-statistic: 292.7 on 1 and 3998 DF,  p-value: < 0.00000000000000022
```

This is the same regression as the one estimated in part (5). The first stage $\hat{\gamma}$ estimate shows that every additional USD dollar in electricity price variation (our instrument) is associated with a 0.74 dollar increase per acre-foot in groundwater cost (our treatment). The F-statistic $292.7 > 20$, shows the instrument satisfies the first stage assumption. The exclusion restriction likely holds because the variation in electricity prices was random.

Next, we calculate the **second stage**:

```
# Second Stage regression
second_stage <- lm(groundwater_use ~ predicted_y, data = data)
tau_iv <- second_stage$coefficients[[2]]
cat("The effect of treatment on outcome is", tau_iv)
```

```
## The effect of treatment on outcome is -199.2625
```

```
summary(second_stage)
```

```
## 
## Call:
## lm(formula = groundwater_use ~ predicted_y, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -105657  -53706  -10665   46909  191948 
## 
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)    
## (Intercept) 166321.44    7193.07  23.122 <0.0000000000000002 ***
## predicted_y   -199.26      21.72  -9.175 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65720 on 3998 degrees of freedom
## Multiple R-squared:  0.02062,    Adjusted R-squared:  0.02037
## F-statistic: 84.17 on 1 and 3998 DF,  p-value: < 0.00000000000000022
```

In the second stage, we obtain our $\hat{\tau}^{IV}$ estimate of the effect of groundwater costs (treatment) on groundwater consumption (outcome variable Y_i).

This means, on average, that each additional dollar per acre-foot in groundwater costs is associated with a decrease of -199.26 dollars per acre-foot on groundwater consumption. This coefficient is highly statistically significant at all conventional levels of significance.

To note, the standard errors in this second stage estimate are off because R is not aware that this regression is built from the results of another regression (the first stage).

To correct for this and validate our estimated results, we can use R's AER package:

```
two_sls <- ivreg(groundwater_use ~ groundwater_cost | electricity_price_pilot,
                 data = data)
summary(two_sls)
```

```
##
## Call:
## ivreg(formula = groundwater_use ~ groundwater_cost | electricity_price_pilot,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -130243  -38812   -2124   39336  141213
##
## Coefficients:
##                   Estimate Std. Error t value            Pr(>|t|)
## (Intercept)      166321.44    5972.08   27.85 <0.0000000000000002 ***
## groundwater_cost   -199.26      18.03  -11.05 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54560 on 3998 degrees of freedom
## Multiple R-Squared: 0.3249,  Adjusted R-squared: 0.3247
## Wald test: 122.1 on 1 and 3998 DF,  p-value: < 0.00000000000000022
```

As expected, we obtain the same $\hat{\tau}^{IV}$ estimate as before, ie. -199.26. However, the standard errors decreased from 21.72 to 18.03, showing this estimate is more *precise*. Since the $\hat{\tau}^{IV}$ is highly statistically significant and is of a considerable magnitude, groundwater costs *do* matter for consumption as an increase in groundwater costs is correlated with a corresponding decrease in groundwater consumption.

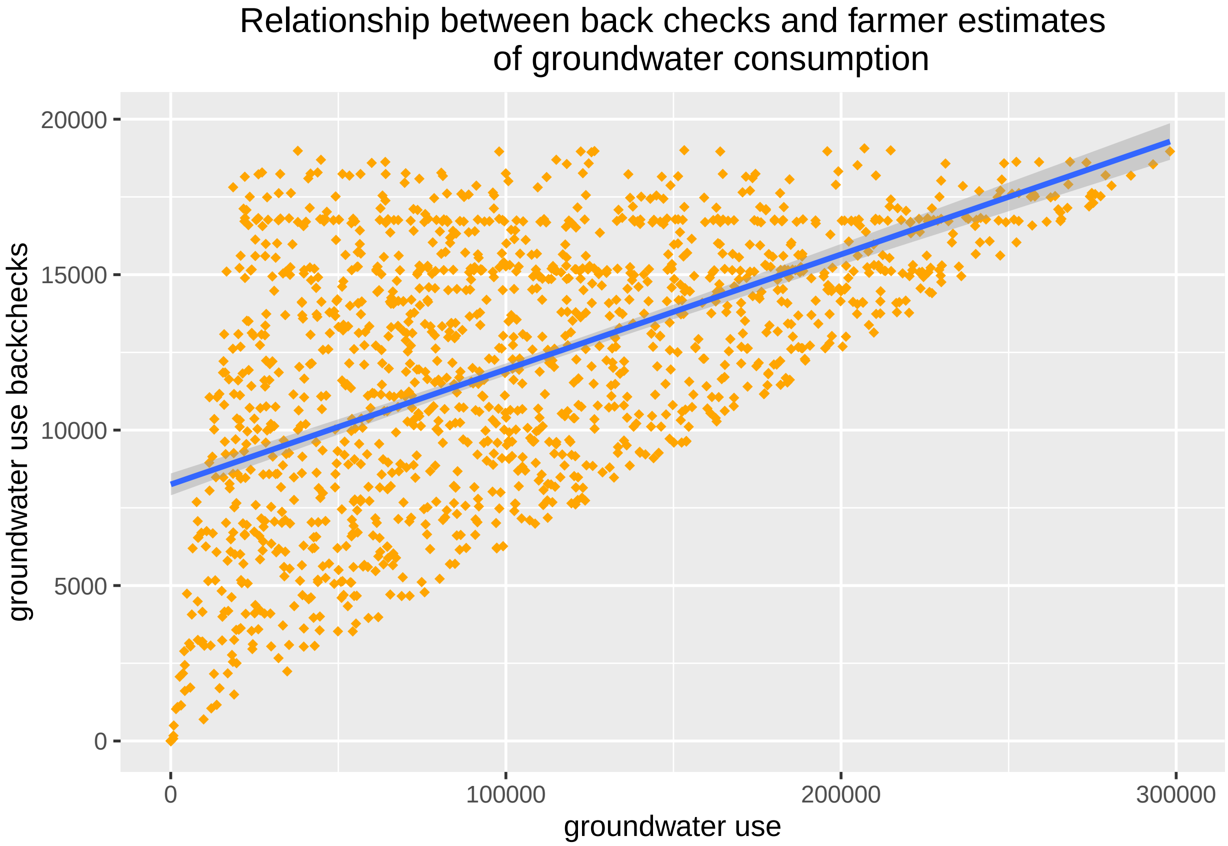
(8) CALBEARS like your analysis, but they're a bit worried about the quality of their data on groundwater consumption. The way they normally collect these data is by surveying the farmers. However, they went and did some back-checks in a subsample of data that they gave you, and noticed that the farmer reports seem to be off. They would like you to make a graph showing the relationship between their back-checks (groundwater_use_backchecks) and the farmers estimates (groundwater_use). Describe to them what you find. Is this likely to be a problem for your analysis? Why or why not? Next, estimate the impacts of groundwater costs on groundwater consumption using the backcheck data instead of the farmer estimates. Report what you find. Do your estimates differ? If no, explain why not. If yes, explain why.

```
ggplot(data, aes(x=groundwater_use, y=groundwater_use_backchecks)) +
geom_point(shape=18, color="orange") + geom_smooth(method=lm) +
ggtitle("Relationship between back checks and farmer estimates
        of groundwater consumption") + theme(plot.title = element_text
        (hjust = 0.5)) + xlab("groundwater use") +
        ylab("groundwater use backchecks")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2611 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2611 rows containing missing values (geom_point).
```



From the graph, we can observe that the groundwater back-check variation is very visibly concentrated on lower self-reported groundwater use. In other words, the variation between backchecks and self-reported groundwater use is *not random*, if it were, we wouldn't see the points concentrated in one area as we observe here. It appears that there is systematic discrepancy among farmers that reported lower groundwater use.

This is a case of measurement error in the outcome variable, Y (groundwater consumption). This is not a problem if and only if two conditions are satisfied:

1) The covariance between the measurement error γ_i and, **2)** the error term ε_i is 0 AND the covariance between the measurement error and treatment is also 0.

$$Cov(\gamma_i, \varepsilon_i) = 0 \text{ and } Cov(\gamma_i, D_i) = 0$$

When these two conditions hold, even though we don't observe Y_i, we do observe :

$$\tilde{Y}_i = Y_i + \gamma_i$$

Then, the slope doesn't change and the relationship between X_i (groundwater costs) and Y_i(groundwater consumption) remains the same because the movement in groundwater consumption is *random*.

Based on the scatterplot, condition 1) does not hold because the measurement error is not random and it likely does covary with the error term. For example, it may be the case that reporting groundwater use means filing more taxes so there may be an incentive to lie or under-report groundwater consumption.

As such, this will be a problem for my analysis because the proposed IV design cannot correct for this issue. Alternatively, another instrument variable for the measurement error could be used to account for this problem.

Now, we estimate the effect of groundwater costs on groundwater consumption using the back-check data:

```
backcheck <- ivreg(groundwater_use_backchecks ~ groundwater_cost |
                     electricity_price_pilot, data = data)
summary(backcheck)
```

```
##
## Call:
## ivreg(formula = groundwater_use_backchecks ~ groundwater_cost |
##     electricity_price_pilot, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -50.86  -38.47  -35.08  -31.87 8268.97
##
## Coefficients:
##                    Estimate Std. Error t value            Pr(>|t|)
## (Intercept)      20035.6543    75.8271   264.2 <0.0000000000000002 ***
## groundwater_cost   -25.0009     0.2357  -106.1 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 417 on 1387 degrees of freedom
## Multiple R-Squared: 0.991,   Adjusted R-squared: 0.991
## Wald test: 1.125e+04 on 1 and 1387 DF,  p-value: < 0.00000000000000022
```

We obtain a new $\hat{\tau}^{IV}$ of -25.00. On average, each additional dollar per acre-foot in groundwater costs is correlated with a decrease of -25.00 dollars per acre-foot on groundwater consumption. Like before, this estimate is also highly statistically significant.

In comparison to the IV model in part (7), we find a much smaller estimate (from -199.26 to -25 dollars per acre-foot). This is indicative that the measurement error in reported groundwater consumption is in fact problematic because it was leading us to considerably overestimate the effect of groundwater costs on groundwater consumption.

```
# Number of groundwater consumption back-checks conducted
sum(!is.na(data$groundwater_use_backchecks))
```

```
## [1] 1389
```

Nonetheless, it should be noted that the groundwater checks were done only on a subset of the sample, 1389 out of 4000 farmers, so the magnitude of the estimate may not be entirely accurate.

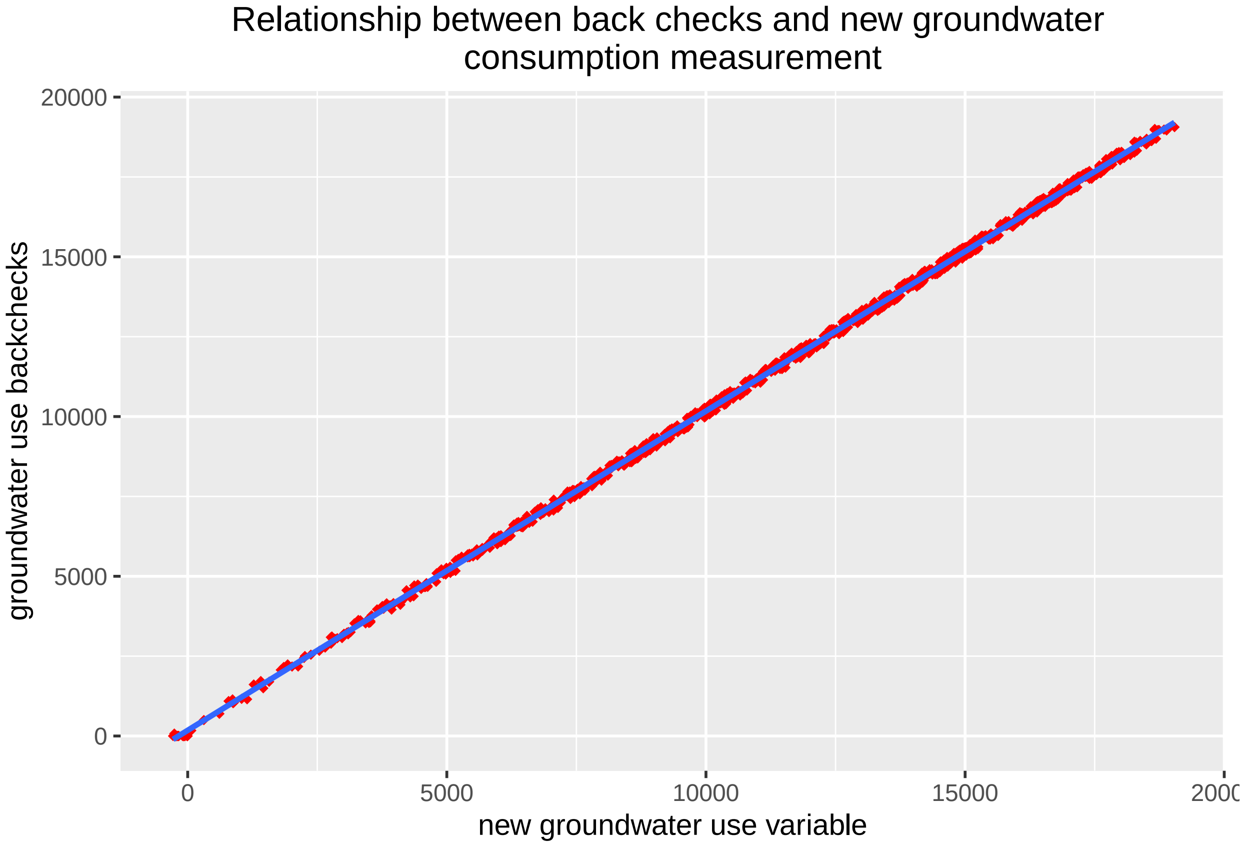
(9) The challenge with back-checks is that they're very expensive to do. Fortunately, CALBEARS realized that they have another dataset on groundwater consumption which seems to match the back-checks much better. They'd like you to make a graph showing the relationship between their back-checks and this new measurement (groundwater_use_v2). Describe to them what you find. Is this likely to be a problem for your analysis? Why or why not? Next, estimate the impacts of groundwater costs on groundwater consumption using the backcheck data and using the new estimates. Report what you find. Do your estimates differ? If no, explain why not. If yes, explain why.

```
ggplot(data, aes(x=groundwater_use_v2, y=groundwater_use_backchecks)) +
geom_point(shape=18, color="red") + geom_smooth(method=lm) +
ggtitle("Relationship between back checks and new groundwater
consumption measurement") + theme(plot.title = element_text
        (hjust = 0.5)) + xlab("new groundwater use variable") +
        ylab("groundwater use backchecks")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2611 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2611 rows containing missing values (geom_point).
```



From the graph, we can see that the new measurement for groundwater use is extremely close to the back-checks data. This provides evidence in favor that the back_checks data, even if performed on a subset of the sample, is very close to the new measurement variable which is provided for the 4000 observations.

This is unlikely a problem for my analysis. On the contrary, this contributes to the analysis by providing a groundwater consumption variable that is more or less clean of non-random measurement error and it can help to validate the $\hat{\tau}^{IV}$ estimated with the back-checks.

Next, we run the IV model with the new variable:

```r
groundwater_use2 <- ivreg(groundwater_use_v2 ~ groundwater_cost |
                         electricity_price_pilot, data = data)
summary(groundwater_use2)
```

```
##
## Call:
## ivreg(formula = groundwater_use_v2 ~ groundwater_cost | electricity_price_pilot,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -218.33 -117.25  -29.61   58.51 8251.50
##
## Coefficients:
##                    Estimate Std. Error t value            Pr(>|t|)
## (Intercept)      19866.8826    40.5528   489.9 <0.0000000000000002 ***
## groundwater_cost   -25.0137     0.1224  -204.3 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 370.5 on 3998 degrees of freedom
## Multiple R-Squared: 0.9932,  Adjusted R-squared: 0.9932
## Wald test: 4.173e+04 on 1 and 3998 DF,  p-value: < 0.000000000000000022
```

As a result, the estimated newly $\hat{\tau}^{IV}$ is -25.01, marginally different from the -25.00 obtained in part (8). Since, from the scatterplot, the back_checks data is extremely close to the new groundwater consumption measurement, it makes sense that the estimates do not change. This validates that the back-checks subset data is representative of the sample as a whole and that these last two $\hat{\tau}^{IV}$ estimates are likely more reliable.

(10) CALBEARS comes back to you again with yet another data problem. This time, they're worried that the utilities aren't reporting electricity prices very well. They'd like you to focus on the effect of electricity price on groundwater consumption (you can ignore groundwater costs for the remainder of the problem set). CALBEARS explain to you that, in one utility (labeled iou == 1 in the data, because #privacy), something was going wrong with the price information they were using. Farms facing low prices had these prices recorded correctly in the data, but the higher the price, the more inflated utility 1's record is. In the other utility (labeled iou == 2), there are still imperfect measurements, but CALBEARS is convinced that the measurement problems are random. Explain the implications of these data issues in each utility to CALBEARS. Are these measurement issues going to be a problem for your analysis? Use words and math to explain why or why not. Despite any misgivings you might have, run your analysis anyway, separately for each utility this time (using your preferred groundwater consumption variable from the three described above), and report your findings.

Given the new focus, our treatment D_i is now electricity price and the outcome Y_i is still groundwater consumption.

Measurement error in the treatment, as is described in this question, is problematic. There are two types of this measurement error:

1. Classical: This occurs when the measurement error is *random*, in this case, when iou == 2. Here, the estimated relationship between groundwater consumption and electricity prices gets closer to 0 and the slope becomes flatter.

We are interested in estimating:

$$Y_i = \alpha + \tau D_i + \varepsilon_i$$

Where:

- Y_i is the outcome variable, groundwater consumption
- i is the individual farmer

Because of the classical measurement error, we don't observe D_i, but rather D_i plus the measurement error:

$$\tilde{D}_i = D_i + \gamma_i$$

Since the measurement error is random, we assume that groundwater consumption and electricity are uncorrelated. Then the covariance term disappears, as follows:

$$Var(D_i + \gamma_i) = Var(D_i) + Var(\gamma_i) + 2Cov(D_i, \gamma_i)$$

$$Var(D_i + \gamma_i) = Var(D_i) + Var(\gamma_i)$$

As a result, our estimate becomes:

$$\hat{\tau}^{IV} = \tau \left(\frac{Var(D_i)}{Var(D_i) + Var(\gamma_i)} \right)$$

The component within the parenthesis is less than 1, so that it biases the $\hat{\tau}^{IV}$ estimate towards 0. This is known as **attenuation bias**, where even though we have a biased estimate, the sign on the estimate is accurate.

If we can find a new valid instrument, Z_i for the noisy $\tilde{D}_i$, then we can resolve the problematic aspect of this classical measurement error. Otherwise, we will get an attenuated and biased towards 0 $\hat{\tau}^{IV}$ estimate.

2. Non-Classical: This is a very problematic scenario in which the measurement error is *NOT random* and occurs in a systematic way. This is the case of iou == 1 because the higher the price, the more inflated the record of utility 1.

In other words, there is correlation between higher electricity prices and measurement error. Then, the covariance term remains and the $\hat{\tau}^{IV}$ is scaled by this term, where we now get:

$$\hat{\tau} = \tau \left(\frac{Var(D_i) + Cov(D_i, \gamma_i)}{Var(D_i) + Var(\gamma_i) + 2Cov(D_i, \gamma_i)} \right)$$

Once again, we obtain a *biased* estimate except that the sign can flip and is no longer accurate. Here, there is nothing that can be done to recover an unbiased $\hat{\tau}^{IV}$.

Running IV models to estimate the effect of electricity price on groundwater consumption

For the following regressions, I am going to use the groundwater_use_v2 variable because it is as accurate as the back_checks measurement with the added bonus that it's provided for every observation in the sample.

A) Classical measurement error, utility 2

```
utility2 <- filter(data, iou == 2)
classical <- lm(groundwater_use_v2 ~ electricity_price_pilot, data = utility2)
summary(classical)
```

```
##
## Call:
## lm(formula = groundwater_use_v2 ~ electricity_price_pilot, data = utility2)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12078.0  -2208.4    158.1   2395.2   9273.9
##
## Coefficients:
##                         Estimate Std. Error t value            Pr(>|t|)
## (Intercept)            14178.246    114.702  123.61 <0.0000000000000002 ***
## electricity_price_pilot  -233.421      5.343  -43.69 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3405 on 1999 degrees of freedom
## Multiple R-squared:  0.4885, Adjusted R-squared:  0.4882
## F-statistic:  1909 on 1 and 1999 DF,  p-value: < 0.00000000000000022
```

For the classical measurement error, we obtain that every additional dollar in the electricity price is associated with a -233.42 decrease in acre-feet groundwater consumption. Even though this coefficient is highly statistically significant, this is misleading as the estimated $\hat{\tau}^{IV}$ is *biased* towards 0, which suggests that the actual τ has a higher magnitude.

However, the negative sign of the estimate is reliable, and tells us there is negative correlation between electricity prices and groundwater consumption, as expected by economic theory.

B) Non-Classical measurement error, utility 1

```r
utility1 <- filter(data, iou == 1)
non_classical <- lm(groundwater_use_v2 ~ electricity_price_pilot,
                    data = utility1)
summary(non_classical)
```

```
##
## Call:
## lm(formula = groundwater_use_v2 ~ electricity_price_pilot, data = utility1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12436.9  -1676.9    605.4   2125.3  10030.7
##
## Coefficients:
##                          Estimate Std. Error t value            Pr(>|t|)
## (Intercept)            15560.0029    95.2487  163.36 <0.0000000000000002 ***
## electricity_price_pilot   -32.0306     0.8428  -38.01 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2903 on 1997 degrees of freedom
## Multiple R-squared:  0.4197, Adjusted R-squared:  0.4194
## F-statistic:  1445 on 1 and 1997 DF,  p-value: < 0.00000000000000022
```

In the non-classical measurement error, we find that every additional dollar in the electricity price is associated with a -32.03 decrease in acre-feet groundwater consumption. The coefficient is highly statistically significant but that is irrelevant because the estimated $\hat{\tau}^{IV}$ is biased. In fact, the size of the coefficient is much, much smaller than with the classical measurement error (-32.03 vs. -233.42). Since we know that the estimate with classical measurement is biased towards 0, we can infer that the real τ has a bigger size than 233.42, which further suggests how biased this non-classical estimate is.

(11) CALBEARS conducted a survey of farmers to understand their experience with the pricing pilot, and asked the farms to report their electricity prices (survey_price). Describe how you could use these data to correct any issues you reported in (10). What conditions need to be satisfied in order or this to work? Are these conditions satisfied in utility 1, utility 2, both, or neither? Carry out your proposed analysis in the sample where it will work (utility 1, utility 2, both, or neither). Report your results, and describe how they compare to your estimates in (10), or explain why you didn't produce any. Which estimates would you send to CALBEARS as your final results?

In light of the new variable, we can use an Instrumental Variable approach to solve for the classical attenuation bias we obtained for utility 2. Essentially, we would need a new instrument Z_i for the noisy measure of electricity prices we had before.

Mathematically, we would now estimate:

$$\mathring{D}_i = D_i + \zeta_i$$

Where:

- $\mathring{D}_i$ is the measurement error in the price survey
- D_i is electricity prices
- ζ_i is the error
- i is the individual farmer

This approach works as long as the new noisy instrument of electricity prices (ie. price survey) is uncorrelated with the previous noisy measure of treatment (ie. random classical measurement in utility #2's records):

$$\tilde{D}_i \neq \mathring{D}_i$$

Specifically, three assumptions are required:

1. The measurement error ζ_i needs to be uncorrelated with electricity prices

$$Cov(\zeta_i, D_i) = 0$$

2. The measurement error ζ_i in the price survey instrument Z_i must be uncorrelated with the error in $\tilde{D}_i$ (ie. random classical measurement in utility #2's records of electricity prices)

$$Cov(\zeta_i, \gamma_i) = 0$$

3. The measurement error ζ_i has to be uncorrelated with the original error

$$Cov(\zeta_i, \varepsilon_i) = 0$$

If these assumptions are met, then $\hat{\tau}^{IV} = \tau$ and our estimate will be *unbiased*.

Since the measurement error is random for **utility #2** only, these conditions are satisfied because the price survey instrument Z_i and the noisy measure of electricity prices $\tilde{D}_i$ only have the true electricity prices D_i in common. Then, the first stage is:

$$\tilde{D}_i = \alpha + \pi \mathring{D}_i + \epsilon_i$$

This also works in this case because the the treatment, (ie. electricity prices) is a continuous variable. If it were binary, then the measurement error could only be -1 or 0 for $D_i = 1$ and 0 or 1 for $D_i = 0$.

Now, we run the IV analysis to correct for classical measurement error in utility #2:

```
classical_unbiased <- ivreg(groundwater_use_v2 ~ electricity_price_pilot |
                        survey_price, data = utility2)
summary(classical_unbiased)
```

```
##
## Call:
## ivreg(formula = groundwater_use_v2 ~ electricity_price_pilot |
##     survey_price, data = utility2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14961.2  -3127.6     77.6   3217.4  14630.1
##
## Coefficients:
##                         Estimate Std. Error t value            Pr(>|t|)
## (Intercept)              17818.5      300.0   59.40 <0.0000000000000002 ***
## electricity_price_pilot   -460.1       17.5  -26.29 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4694 on 1999 degrees of freedom
## Multiple R-Squared: 0.02793, Adjusted R-squared: 0.02744
## Wald test: 691.3 on 1 and 1999 DF,  p-value: < 0.00000000000000022
```

Here, we find that every additional dollar in the electricity price is associated with a -460.1 decrease in acre-feet groundwater consumption. The coefficient is highly statistically significant at all conventional levels and the negative sign is consistent with economic theory.

As expected, the magnitude of the estimate is bigger than the obtained for the classical measurement in part (10), which was biased towards zero. This unbiased estimate, -460.1, is practically double the size of the previous biased estimate of -233.42. As my final result, I would send this estimate to CALBEARS as it is unbiased, statistically significant and solves the classical measurement problem.

For utility #1, unfortunately it is not possible to implement this approach because with non-classical error, our estimate will be biased no matter what.