PS3_12265092

12265092

11/05/2022

```
library(knitr)

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(haven)

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v readr   2.0.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stargazer)
```

```
## Warning: package 'stargazer' was built under R version 4.1.2

##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```r
library(broom)

library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 4.1.3

##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```r
library("ivreg")
```

```
## Warning: package 'ivreg' was built under R version 4.1.2
```

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.1.2

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.1.2

##
## Attaching package: 'car'

## The following object is masked from 'package:purrr':
##
##     some

## The following object is masked from 'package:dplyr':
##
##     recode
```

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?).

For the same set of farmers, we have to observe the below

Consumption - Measured in Acre-feet Costs unit - Measured in dollars per acre-foot

Ideal Experiment:

We need to observe

1. Total groundwater consumption at average costs x

2. Total groundwater consumption at average costs x+1
3. Effect of Average groundwater costs between April, September on the total groundwater consumption in the same period - difference between 1 and 2

Let i be the individual farmer where $i \in \{1, 2, ...N\}$. Treatment indicator D_i where $D_i \in \{0, 1\}$ Treated: $D_i = 1$: Average ground water costs increased by 1 unit
Untreated: $D_i = 0$: Average ground water costs are not changed

Outcome treated: $Y_i(D_i = 1)$: Total froundwater consumption for a farmer i between April and September when average groundwater costs increased by 1 unit -Treatment Outcome untreated: $Y_i(D_i = 0)$: Total ground water consumption for a farmer i between April and September in case of average groundwater costs are not changed - Control

We get the impact of treatment(i.e disconnecting household's electricity) τ_i from the difference between the above outcomes.

- $\tau_i = Y_i(D_i = 1) - Y_i(D_i = 0)$

The impact of treatment τ_i is the difference between the two outcomes i.e difference between Total froundwater consumption for a farmer i between April and September when average groundwater costs increased by 1 unit vs Total ground water consumption for a farmer i between April and September in case of average groundwater costs are not changed.

From above: - $\tau_i = Y_i(D_i = 1) - Y_i(D_i = 0)$

While we need both the outcomes at a given time to compute the impact of treatment, the problem is that at a given time, we cannot observe both the outcomes $ or $ we can only observe either $Y_i(D_i = 1) $ or $ Y_i(D_i = 0)$ at a given time.

In case a farmer is treated (i.e average costs increased by 1 unit), then the observed outcome would be $Y_i(D_i = 1) $(Total groundwater consumption for a farmer #i# between April and September when average costs increased by 1 unit), and $ Y_i(D_i = 0)$ (Total ground water consumption for a farmer *i* between April and September in case of average groundwater costs are not changed) would become an unobserved outcome. Due to the un-observable outcome *or* not being able to observe both the outcomes at a given time, measuring τ_i is impossible.

Average Treatment Effect $ tau^{ATE}$

- $ tau^ATE = E[Y_i(D_i = 1)] - E[Y_i(D_i = 0)]$

ATE measures the average effect of treatment across a population of farmers. ATE measures the effect of average groundwater costs between April and September on total groundwater consumption. The problem is same in case of ATE. At the same time, for a farmer i, we cannot observe both outcomes. Hence it is impossible to measure ATE.

How would a realistic experiment look like?

An RCT where the treatment is assigned to farmers randomly. Then we can calculate the effect of treatment i.e average costs increasing by 1 unit on the average groundwater consumption. When the treatment assigned randomly and the distribution of the observables and the unobservables are same across the treated and untreated, we can take that there is no selection problem by design.

Hence we get, $E[Y_i(1)|D_i = 1] = E[Y_i(1)]$ and $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$

As a result, $tau^{ATE} = E[Y_i(D_i = 1)] - E[Y_i(D_i = 0)]$

Then the ATE will be equal to Naive estimator $\tau_N = \bar{Y}(D = 1) - \bar{Y}(D = 0)$

For this to workout, we assume that the outcome is solely affected by the treatment and there is 100% compliance and there are no spillover effects among the treated or control groups.

Question 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.

As we have only one wave of daya and no repeated observations, we run Regression of Interest„ design with an Instrument Variable.

Regression of Interest:

$Y_i = \alpha \ + \tau D_i \ + \beta X_i + \epsilon_i$

Y_i - Outcome - Total ground water consumption for a farmer i between April and September
D_i - Treatment indicator - Average groundwater costs increased by 1 unit or not
X_i - Covariates ϵ_i - Error

Also, Cov(X_i, ϵ_i) = 0

We assume randomm assignment of treatment and hence that means the treatment variable D_i is exogenous. We also need that D_i is not correlated with the error term.

Hence $E[\ \epsilon_i \mid D_i] = 0$

There can be cases always that shows D_i can be endogenous, Examples: Systematic Errors:Errors in the treatment invariable due to human errors Ommitted Variable Bias: D_i will be endogenous if there are any ommitted variables that are correlated with costs which can show effect on the groundwater consumption. Like equipment costs Simultaneity: Simulatneity or Reverse Causality where the outcome variable affects the treatment variable. i.e Consumption of ground water affecting the costs.

Similar to above can make the D_i endogenous i.e $E[\ \epsilon_i \mid D_i]! = 0$

An alternate approach as we are limited by just one wave of data is by implementing a Instrument Variable

We define Z_i an instrument variable that is not correlated with ϵ_i with the following assumptions: Z_i and D_i are correlated i.e., $Cov(Z_i, D_i)! = 0$ Exclusion restriction: $Cov(Z_i, \epsilon_i) = 0$ i.e as said above the instrument variable Z_i is not correlated with ϵ_i.

While the first condition can be tested using the data we have, the exclusion restriction cannot be tested. The exclusion restriction also means that Y_i can be affected by Z_i, but only through the treatment D_i. Thus we donot include it in the regression.

We implement it as follows:

We follow a two stage approach, we already know that $Cov(Z_i, D_i)! = 0$ and $Cov(Z_i, \epsilon_i) = 0$

1. Isolate exogeneous variation in treatment D_i
   $D_i \ = \alpha \ + \gamma \ .Z_i + \beta \ .X_i + \ \eta_i$

where,
D_i: Treatment variable i.e ground water costs for a farmer i
α: Average cost of groundwater use when there is no Z_i and X_i
γ: Average effect of Z_i on treatment D_i
β: Average effect of known covariates X_i on treatment D_i η = Error term

We get $hatD_i$

2. No we regress the outcome Y_i on $hatD_i$ and the covariates X_i: $Y_i = \alpha + hat\tau.hatD_i + \delta.X_i + \epsilon_i$

Here: - $\tau\hat{}$: is the IV estimate for the ATE τ_{ATE}
- Y_i: Outcome variable i.e Ground water consumption by farmer i
- α: Average cost of groundwater use when there is no Z_i and X_i
- $\tau\hat{}$ = ATE of treatment variable $hatD_i$ on Y_i
- δ = Average effect of known covariates X_i on treatment D_i
- ϵ = Error term

QUestion 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?

Depth of the aquifier can be a plausible IV. Let us test the possibility below:

Assumption $Cov(Z_i, D_i)! = 0$, it is given that the cost of extracting groundwater depends on the depth of the acquifier. Hene this assumption is satisfied.

Assumption $Cov(Z_i, \epsilon_i) = 0$. As said above, we cannot test this assumption in real world. But it can be safe to assume that the groundwater consumption donot vary in aquifier depth.

Thus we can say that the instrument variable Z_i may not affect Y_i any other way except through the treatment variable D_i i.e the groundwater cost. Thus it can be excluded from the regression of Y_i on D_i

As we determined that the Depth of aquifier is an instrument variable, we can now estimate τ_hat using the 2 stage approach.

Pump Efficiency can be a plausible IV. Let us test the possibility:

Assumption $Cov(Z_i, D_i)! = 0$, it is given that the cost of extracting groundwater depends on the efficiency of pump. Hence this assumption is satisfied.

Assumption $Cov(Z_i, \epsilon_i) = 0$, as said above we cannot test this assumption in real world. But if we assume that the groundwater consumption is dependent on crop spread, we can assume that groundwater consumption will not vary with pump efficiency.

Thus we can say that the instrument variable Z_i, the pump efficiency, may not affect Y_i any other way except through the treatment variable D_i i.e the groundwater cost. Thus it can be excluded from the regression of Y_i on D_i

As we determined that the Pump Efficiency is an instrument variable, we can now estimate $hat\tau$ using the 2 stage approach.

Similarly, we can determine Cost of Electricity as a plausible instrument variable and implement the 2 stage approach to determine $hat\tau$

Due to the above said reasons and the choices satisfying to be plausible instrument variables, I dont see concerns unless any of the above said assumptions are proven to be not true.

Question 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.

Electricity Costs as an instrument variable

Electricy Costs - measured in dollars per unit of electricity Ground water costs - measured in dollars per acre-foot Ground water consumption - measured in acre-feet

Checking if the electricity costs satisfies the IV assumptions

1. $Cov(Z_i, D_i)! = 0$ It is already given that the cost of extracting electricity depends on electricity price. Hence this assumption is satisfied
2. $Cov(Z_i, \epsilon_i) = 0$ We cannot test this in real world. But if we assume that the groundwater consumption depends on the crop spread, we can assume that ground water consumption does not vary with the electricity costs Thus we can assume that electricity price cannot affect the outcome in any other way except for the treatment variable. Thus we can exclude electricity costs in the regression of ouctome i.e groundwater consumption on treatment variable groundwater cost.

As we have determined the viability of Electricity costs to be a valid Instrument variable, we can implement the two stage aproach to derive $hat\tau$

Isolating exogeneous variation in D_i using Z_i

$groundwatercost_i \ = \alpha \ + \gamma \ .electricityprice_i + \beta \ .X_i + \ \eta_i$

where,
Z_i: Electricity price X_i: Known COvariates $groundwatercost_i$: Ground water costs for a farmer i $groundwater$ α: Average cost of groundwater use when there is no Z_i and X_i
γ: Average effect of Z_i on $groundwatercost_i$
$\beta =$ Average effect of X_i on $groundwatercost_i$
$\eta =$ Error term for a farmer i

We get $hatgroundwatercost_i$, the predicted

Then we go to the second stage

Regress the outcome i.e $groundwaterconsumption_i$ on $hatgroundwatercost_i$ and the same covariates X_i

is $groundwaterconsumption_i \ = \ \alpha + hat\tau.hatgroundwatercost_i + \delta.X_i + \epsilon_i$ where, - $hat\tau$ will be the IV estimate for the ATE τ_{ATE}
- $hat\tau$ ATE of $hatgroundwatercost_i$ on $groundwaterconsumption_i$
- $\delta =$ Average effect X_i on $groundwaterconsumption_i$
- $\epsilon =$ Error term for farmer i

Reduced Form methodology:

Here we will have 3 stages

Stage 1 i.e $groundwatercost_i \ = \alpha \ + \gamma \ .electricityprice_i + \beta \ .X_i + \ \eta_i$

where,
Z_i: Electricity price X_i: Known COvariates $groundwatercost_i$: Ground water costs for a farmer i $groundwater$ α: Average cost of groundwater use when there is no Z_i and X_i

γ: Average effect of Z_i on $groundwatercost_i$
β = Average effect of X_i on $groundwatercost_i$
η_i = Error term for a farmer i

Undertake the above regression and get the values of $hat\gamma$, the estimated average effect of the instrument ($electricityprice_i$) on treatment ($groundwatercost_i$)

Then we perform the following regression to get $hat\gamma$

$groundwatercost_i \ = \alpha \ + hat\gamma \ .electricityprice_i + hat\beta \ .X_i$

where $hat\gamma$ is the estimated average effect of $electricityprice_i$ on $groundwatercost_i$

Stage 2 Reduced form: Estimating effect $electricityprice_i$ on ($groundwaterconsumption_i$)

i.e $groundwaterconsumption_i \ = \ \alpha + \theta.electricityprice_i + \delta.X_i + \epsilon_i$

$groundwaterconsumption_i$: Ground water consumption for a farmer i
α: Average consumption of groundwater use when there is no $hatZ_i$ and X_i
θ: Average treatment effect of Instrument Variable $ electricityprice_i$ on $groundwaterconsumption_i$
δ: Average effect of X_i on $groundwaterconsumption_i$
ϵ_i = Error term for a farmer i

Then we run the below regression and get the values of $hat\theta$ then $groundwaterconsumption_i \ = \alpha + hat\theta.electricityprice_i + hat\delta.X_i$

where $hat\theta$ is the estimated average effect of $electricityprice_i$ on $groundwaterconsumption_i$

Final Stage: Now we calculate the estimated effect of $hat\tau^{IV}$ on the outcome variable $groundwaterconsumption_i$

$\tau IV hat = \frac{hat\theta}{hat\gamma}$

where $hat\tau^{IV}$: IV estimate for ATE τ_{ATE} of the groundwater costs on total groundwater consumption between April and September

Question 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
data <- read_csv('ps3_data.csv')
```

```
## Rows: 4000 Columns: 7
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (7): iou, groundwater_cost, electricity_price_pilot, groundwater_use_bac...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
summary(data)
```

```
##       iou      groundwater_cost  electricity_price_pilot
##  Min.   :1.0   Min.   :  36.61   Min.   :-26.08
##  1st Qu.:1.0   1st Qu.: 190.29   1st Qu.: 12.91
##  Median :2.0   Median : 294.97   Median : 26.72
##  Mean   :1.5   Mean   : 327.71   Mean   : 49.36
##  3rd Qu.:2.0   3rd Qu.: 444.41   3rd Qu.: 58.13
##  Max.   :2.0   Max.   :1132.14   Max.   :600.45
##
##  groundwater_use_backchecks groundwater_use  groundwater_use_v2
##  Min.   :    0              Min.   :    25   Min.   : -325.7
##  1st Qu.: 9215              1st Qu.: 45789   1st Qu.: 8774.2
##  Median :13006              Median : 90191   Median :12458.8
##  Mean   :12082              Mean   :101020   Mean   :11669.5
##  3rd Qu.:15307              3rd Qu.:147532   3rd Qu.:15123.5
##  Max.   :19061              Max.   :299512   Max.   :19040.8
##  NA's   :2611
##   survey_price
##  Min.   :-36.211
##  1st Qu.:  5.259
##  Median : 15.576
##  Mean   : 16.354
##  3rd Qu.: 26.961
##  Max.   : 77.968
##
```

```r
#Removing the -ve values in groundwater_use_v2 as per explanation on piazza
data <- data %>%
            filter( groundwater_use_v2 > 0)
```

```r
#Testing the validity of IV assumptions
#Assumption 1: the IV variable electricitypricepilot_i and groundwatercost_i are correlated

reg_1 <- lm(groundwater_cost ~ electricity_price_pilot, data = data)
summary(reg_1)
```

```
##
## Call:
## lm(formula = groundwater_cost ~ electricity_price_pilot, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -270.6 -125.6  -38.4  104.5  538.8
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             282.79030    3.28772   86.01   <2e-16 ***
## electricity_price_pilot   0.74368    0.04034   18.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 164.3 on 3940 degrees of freedom
## Multiple R-squared:  0.07942,    Adjusted R-squared:  0.07919
## F-statistic: 339.9 on 1 and 3940 DF,  p-value: < 2.2e-16
```

The p-value for the variable electricity__price__pilot is 2e-16 at 99% confidence level and the Fstatistic of 339.9, which says that it is statistically significant Thus with a 1 unit increase in electricity prices results in the groundwater costs change by 0.74368 per acre-foot.Thus we can say that the variables $electricitypricepilot_i$ and $groundwatercost_i$ are correlated Thus $Cov(electricitypricepilot_i, groundwatercost_i)! = 0$

Assumption 2: $Cov(Z_i, \epsilon_i) = 0$ We cannot test this in the real world

Assumption 3: Random assignment and Selection bias. This cannot be tested with the pilot data. If so there is a selection bias in the after random assignment, we would get biased estimate of the coefficient of the independed variable.

Due to the above problem, we cannot say for certain that the utility pilot is a helpful way forward to estimating the impacts of groundwater costs on groundwater usage.

Question 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?

We use the reduced regression form

$groundwateruse_i \;=\; \alpha + hat\theta.electricityprice_i$

$hat\theta$: Average effect electricityprice__i on groundwateruse__i, i.e the effect of change of one dollar in electricity prices on groundwater consumption in acre foot

```
reg_2 <- lm(groundwater_use ~ electricity_price_pilot, data = data)
summary(reg_2)
```

```
##
## Call:
## lm(formula = groundwater_use ~ electricity_price_pilot, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -106750  -53071  -10508   46362  190414
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             109840.98    1301.71  84.382   <2e-16 ***
## electricity_price_pilot   -148.60      15.97  -9.305   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65030 on 3940 degrees of freedom
## Multiple R-squared:  0.0215, Adjusted R-squared:  0.02125
## F-statistic: 86.58 on 1 and 3940 DF,  p-value: < 2.2e-16
```

The p-value for variable electricity__price__pilot is 1.6e-11 at 99% confidence level. This says that it is statistically signficant. The coefficient is -148.6. This means that with one dollar increase in the electricty price results in a decrease of ground water consumption by 148.6 units. Thus we can say that there is observable correlation between the two variables $groundwateruse_i$ and $electricitypricepilot_i$. $Cov(electricitypricepilot_i, groundwateruse_i)! = 0$.

Assumption 2: $Cov(Z_i, \epsilon_i) = 0$ We cannot test this in the real world

To belive in this estimate, we need to know that the estimate we got from regressing $groundwaterconsumption_i$ on $groundwatercost_i$ as the treatment variable and the $electricitypricepilot_i$ as the Instrument variable is unbiased.

As said previously, we cannot test the random assignment of the treatment across farmers. As we cannot test it from the pilot data, a selection bias may exist. If there is Selection bias, then the assignment of treatment is not random, that means we have biased estimates in the regression. Thus we cannot trust the previous estimate of -148.6 as to be purely unbiased.

Due to the above problem, we cannot say for certain that the utility pilot is a helpful way forward to estimating the impacts of groundwater costs on groundwater usage. This can only clarified once we know that the $electricitypricepilot_i$ has a good random assignment across farmers and there is no selection bias.

Question 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?

Using the 2 stage approach here, the 2SLS regression methodology

```
reg_3 <- ivreg(groundwater_use ~ groundwater_cost | electricity_price_pilot, data = data)
summary(reg_3)
```

```
##
## Call:
## ivreg(formula = groundwater_use ~ groundwater_cost | electricity_price_pilot,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -130220  -39243   -2498   39854  141209
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      166347.51    5858.78   28.39   <2e-16 ***
## groundwater_cost   -199.82      18.13  -11.02   <2e-16 ***
##
## Diagnostic tests:
##                   df1  df2 statistic p-value
## Weak instruments    1 3940   339.910  <2e-16 ***
## Wu-Hausman          1 3939     0.454   0.501
## Sargan              0   NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 54910 on 3940 degrees of freedom
## Multiple R-Squared: 0.3025,  Adjusted R-squared: 0.3023
## Wald test: 121.4 on 1 and 3940 DF,  p-value: < 2.2e-16
```

In the above, we regressed $groundwaterconsumption$ on $groundwatercost$ as the treatment variable and the $electricitypricepilot_i$ as the Instrument variable.

The p-value on groundwater_cost variable 2e-16 at the 99% confidence level says that the variable is statistically significant in determining groundwater consumption. For a dollar/acre-foot increase in the ground water cost, it results in a decrease of 199.82 acre-foot in groundwater consumption. THus we can say that the two variables $groundwateruse$ and $groundwatercost$ are correlated. $Cov(groundwatercost, groundwateruse)! = 0$, $Cov(groundwatercost, groundwateruse) = -199.82 \ \ acrefoot \ per \ dollar.$

Thus we can say that the groundwater costs matter for groundwater consumption. The relation between the both is as stated above.

from the results in Q5 and Q6, we see

The relation determined between $electricitypricepilot$ and $groundwatercost$ is $Cov(electricitypricepilot, groundwatercost) = 0.74368 dollar $

The relaton determined for $electricitypricepilot$ and $groundwateruse$ is $Cov(electricitypricepilot, groundwateruse) = -148.6 \ \ acre - foot$

To determine $Cov(groundwatercost, groundwateruse)$, we need to take ratio of the above two covariances

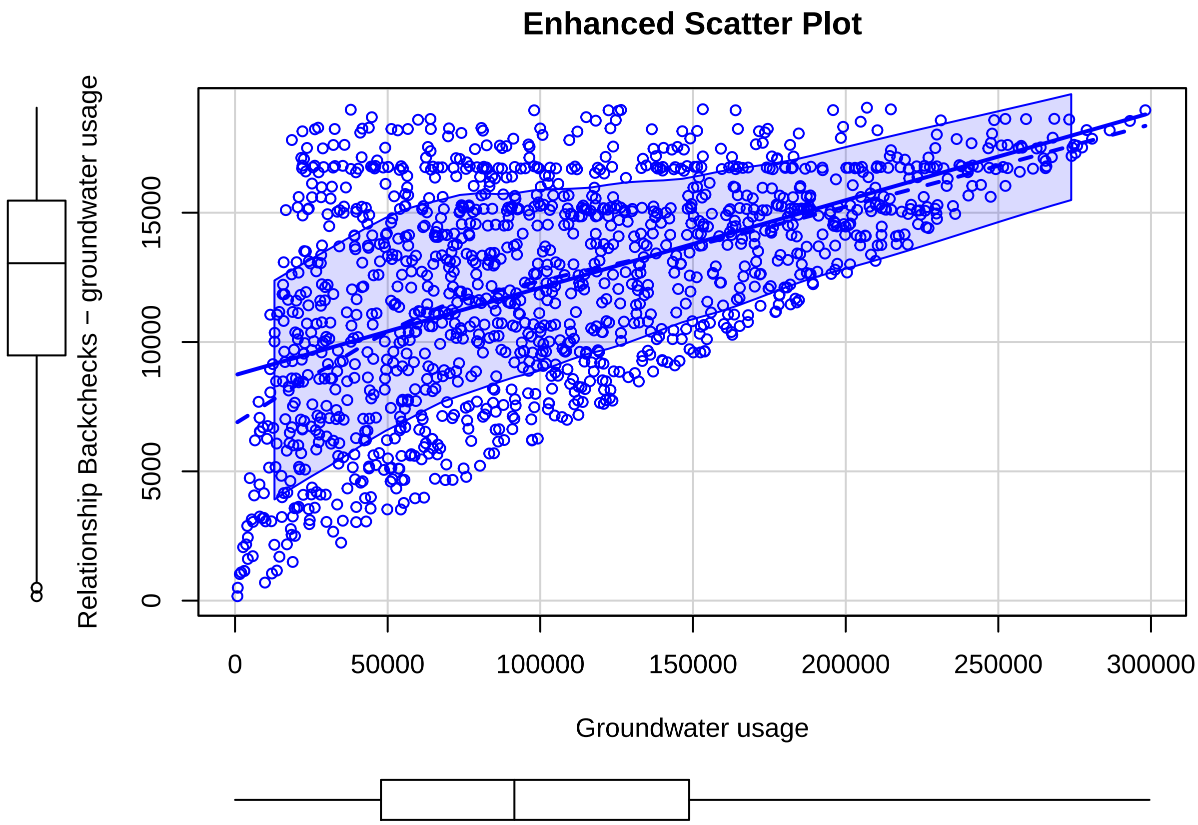
$Cov(groundwatercost, groundwateruse) = \frac{-148.6}{0.74368} = -199.82 \ \ acrefoot \ per \ dollar$

This converges with the relation between $groundwatercost$ and $groundwateruse$ in this question.

###Question 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.

Graph showing the relationship between their back-checks (groundwater_use_backchecks) and the farmers estimates (groundwater_use).

```
scatterplot(groundwater_use_backchecks ~groundwater_use, data=data,
   xlab="Groundwater usage", ylab="Relationship Backchecks - groundwater usage",
   main="Enhanced Scatter Plot")
```



This is measurement error in outcome. $groundwateruse_i = groundwaterusebackchecks_i + \gamma_i$ where $gamma_i$ is the measurement error.

Provided that the measurement error in outcome is random across the outcome $groundwateruse_i$, the ATE is not affected by this. That means no correlation between $groundwater_i$ and the measurement error γ_i i.e $Cov(\gamma, \epsilon) = 0 \quad and \quad Cov(\gamma, groundwatercost) = 0$.

However, from the scatter plot we see that the measurement error is not random across the outcome $groundwateruse_i$. We observe that the farmers reported higher usage of groundwater as compared to the usage found in the back checks. Reported usage > actual usage. Thus the reported groundwater usage overestimates the effect of groundwater costs on the groundwater use. Hence the above assumptions regarding the relation between $groundwatercost$ and $groundwateruse$ will not hold.

Thus $Cov(\gamma_i, \epsilon_i)\ ! = 0 \quad and \quad Cov(\gamma_i, groundwatercost_i)\ ! = 0$. This means that there can be bias or other errors which can likely cause a problem to our analysis.

This problem can be due to many reaons as the measurement can be affected by errors or as we discussed may have some bias in the data. We can estimate the effect of groundwater cost on actual groundwater consumption, we can run a regression of $groundwaterusebackchecks$ on the treatment variable $groundwatercost$ and the instrument variable $electricitypricepilot$

```
backcheck_reg_4 <- ivreg(groundwater_use_backchecks ~ groundwater_cost | electricity_price_pilot, data =
summary(backcheck_reg_4)
```

```
##
## Call:
## ivreg(formula = groundwater_use_backchecks ~ groundwater_cost |
##     electricity_price_pilot, data = data)
```

```
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -15.09597  -3.25497   0.01751   3.19253  18.85347
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.000e+04  8.629e-01   23176   <2e-16 ***
## groundwater_cost -2.500e+01  2.759e-03   -9062   <2e-16 ***
##
## Diagnostic tests:
##                   df1  df2 statistic p-value
## Weak instruments    1 1365   129.768  <2e-16 ***
## Wu-Hausman          1 1364     0.472   0.492
## Sargan              0   NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.993 on 1365 degrees of freedom
## Multiple R-Squared:     1,   Adjusted R-squared:     1
## Wald test: 8.211e+07 on 1 and 1365 DF,  p-value: < 2.2e-16
```

The p-value for the variable $groundwatercost$ 2e-16 at 99% confidence level says that the variable is statistically significant in determining the groundwater consumption. The coefficient of the variable as we observe means that for every 1 dollar per acre-foot increase in groundwater costs, it results in decrease of 25 acre-foot in groundwater consumption.

The coefficient we found in Q7 with reported usage data is -199.82. The coefficient we found using actual usage data i.e -25 proves the over estimation of the coefficients while regressing on reported data.

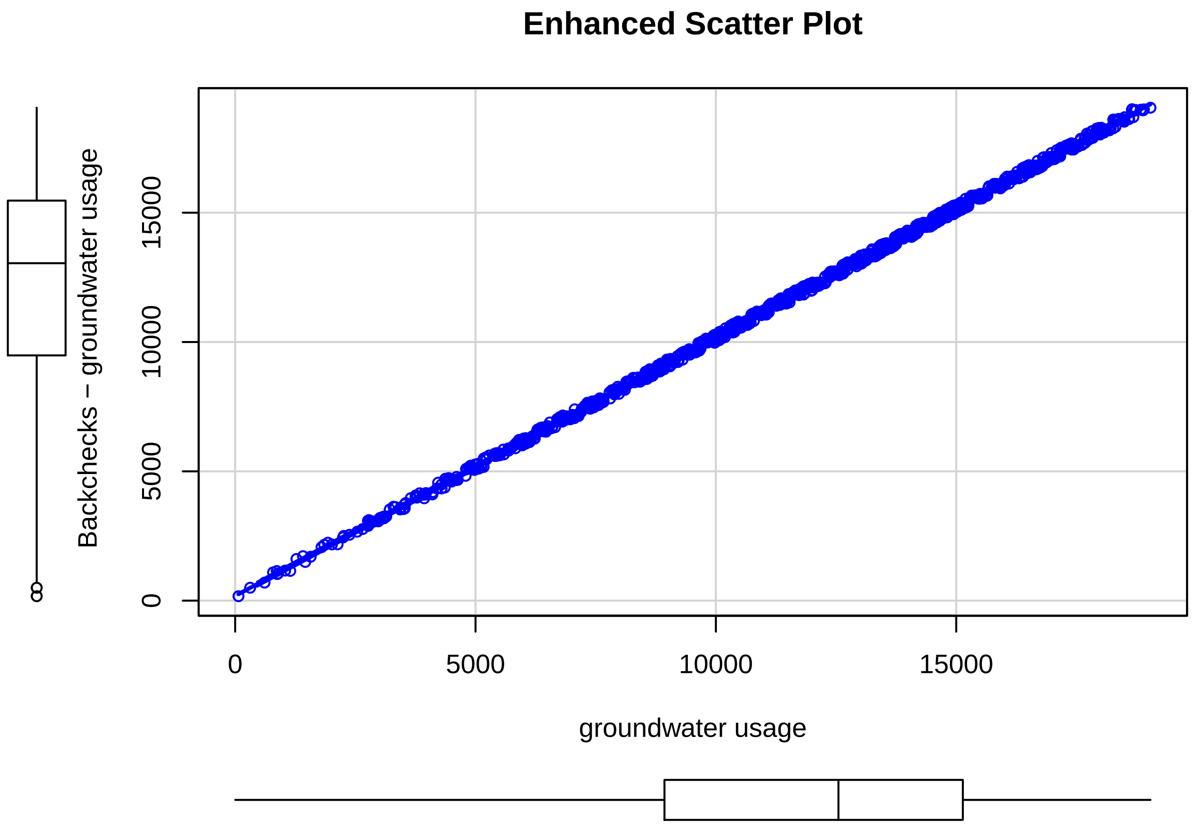
The relation between $groundwaterusebackchecks$ and $groundwatercost$ is $Cov(groundwatercost, groundwaterusebackchecks$ -25. i.e $Cov(groundwatercost_i, groundwaterusebackchecks_i)! = 0$. The assumption holds.

Question 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.

Graph showing the relationship between their back-checks and this new measurement (groundwater_use_v2)

```
scatterplot(groundwater_use_backchecks ~ groundwater_use_v2 , data=data,
   xlab="groundwater usage", ylab=" Backchecks - groundwater usage",
   main="Enhanced Scatter Plot")
```



This is again the case of measurement error. $groundwateruse2_i = groundwaterusebackchecks_i + \gamma_i$ where $gamma_i$ is the measurement error.

Provided that the measurement error in outcome is random across the outcome $groundwateruse_i$, the ATE is not affected by this. That means no correlation between $groundwater_i$ and the measurement error γ_i i.e $Cov(\gamma, \epsilon) = 0 \;\; and \;\; Cov(\gamma, groundwatercost) = 0$.

This time, we see that the measurement error is random across the $groundwaterusev2$ which means that the there is no observable correlation between γ_i and $groundwatercost_i$. i.e $Cov(\gamma, groundwatercost) = 0$. Thus we can say that the new reported usage data values are similar to the backcheck data i.e the actual usage data. And there is no bias observed. Hence there can be no over reporting and under reporting. Assumption is satisfied.

$Cov(\gamma_i, \epsilon_i) \; = 0 \;\; and \;\; Cov(\gamma_i, groundwatercost_i) \; = 0$

This is not likely a problem for our analysis as it supports the argument of random assignment.

Now we perform a regression using the new reported usage data using the treatment variable $groundwatercost_i$ and Instrument variable $electricitypricepilot_i$

```
reg_5 <- ivreg(groundwater_use_v2 ~ groundwater_cost | electricity_price_pilot, data = data)
summary(reg_5)
```

```
##
## Call:
## ivreg(formula = groundwater_use_v2 ~ groundwater_cost | electricity_price_pilot,
##     data = data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -205.725  -87.226   -1.372   87.123 1560.000
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.986e+04  1.116e+01  1779.2   <2e-16 ***
## groundwater_cost -2.508e+01  3.454e-02  -726.1   <2e-16 ***
##
## Diagnostic tests:
##                  df1  df2 statistic p-value
## Weak instruments    1 3940   339.910  <2e-16 ***
## Wu-Hausman          1 3939     5.546  0.0186 *
## Sargan              0   NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 104.6 on 3940 degrees of freedom
## Multiple R-Squared: 0.9994,  Adjusted R-squared: 0.9994
## Wald test: 5.273e+05 on 1 and 3940 DF,  p-value: < 2.2e-16
```

The p-value 2e-16 at 99% confidence level says that the coefficient for groundwater_cost -25.08 is statistically significant. This is same as what we found in the previous question8 where the coefficient is -25. This shows that the new data is infact matches the actual usage as we saw in the backchecks data. The coefficient of -25.08 means that for every dollar per acre-foot increase in the groundwater cost , it results in a decrease of 25.0972 acre-foot decrease in the groundwater consumption.

Question 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.

For utility 1, i.e iou = 1

$groundwateruse_i = \alpha + \tau.electricityprice_i + \epsilon_i$

Given there is a issue of measurement error in the $electricityprice_i$ which is our treatment variable. This means we are observing something else. Lets call is observed electricity price. Hence the ATE is

ATE $hat\tau = \frac{Cov(groundwateruse_i, electricityprice_i + \gamma_i)}{Var(electricityprice_i + \gamma_i)}$

gives $tauhat = \frac{Cov(\alpha + \tau.electricityprice_i + \epsilon_i, electricityprice_i + \gamma_i)}{Var(electricityprice_i + \gamma_i)}$

gives tau _hat = $[\frac{\tau.Var(electricityprice_i) + \tau.Cov(electricityprice_i, \gamma_i) + Cov(electricityprice_i, \epsilon_i) + Cov(\gamma_i, \epsilon_i)}{Var(electricityprice_i) + Var(\gamma_i) + 2.Cov(electricityprice_i, \gamma_i)}]$

We see that farms are observing low prices. But as the prices recorded are not accurate, we see higher prices. This is Nonclassical measurement error in $electricityprice_i$, the treatment variable.

Assumptions:
Treatment is random $Cov(electricityprice_i, \epsilon_i) = 0$ Measurement error is not related to the actual error term $Cov(\gamma_i, \epsilon_i) = 0$
Measurement error is correlated with treatment $Cov(electricityprice_i, \gamma_i)! = 0$

We get

get $hat\tau = \tau.[\frac{Var(electricityprice_i)+Cov(electricityprice_i,\gamma_i)}{Var(electricityprice_i)+Var(\gamma_i)+2.Cov(electricityprice_i,\gamma_i)}]$

Now we run regression to observe this. we would get a bias in $hat\tau$ and the sign of bias depends on the sign of $Cov(electricityprice_i, \gamma_i)$.

```
cleaned_data_iou1 <-
  data %>%
  filter(iou == 1)



reg_6 <- lm(groundwater_use_v2 ~ electricity_price_pilot, data = cleaned_data_iou1)
summary(reg_6)
```

```
##
## Call:
## lm(formula = groundwater_use_v2 ~ electricity_price_pilot, data = cleaned_data_iou1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12443.9  -1673.3    597.1   2119.1   9970.1
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            15557.2034    94.9781  163.80   <2e-16 ***
## electricity_price_pilot   -31.9033     0.8426  -37.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2891 on 1995 degrees of freedom
## Multiple R-squared:  0.4181, Adjusted R-squared:  0.4178
## F-statistic:  1434 on 1 and 1995 DF,  p-value: < 2.2e-16
```

The p-value for $electricity_price_pilot$ 2e-16 at 99% confidence level says that the coeficient estimated is statistically significant. Coefficient = -31.9033. This coefficient means that for every dollar increase in the electricity prices, it results in a decrease of 31.9033 acre-foot in the groundwater consumption. Thus we can say that there is observable correlation between these two variables. i.e $Cov(electricitypricepilot_i, groundwaterusev2_i)! = 0$.

Now for utility 2

It is told that the farms report impertect measurements, but CALBEARS think the measurement problems are random. This is Classical measurement error in $electricityprice_i$, the treatment variable.

Assumptions: Treatment is random $Cov(electricityprice_i, \epsilon_i) = 0$ Measurement error is not related to the actual error term $Cov(\gamma_i, \epsilon_i) = 0$
Measurement error is not correlated with treatment $Cov(electricityprice_i, \gamma_i) = 0$

Thus we get $\hat{\tau} = \left[\frac{\tau.Var(electricityprice_i)}{Var(electricityprice_i)+Var(\gamma_i)}\right]$

Here we get an Attenuation bias in $\hat{\tau}$ and that if $Cov(electricityprice_i, \gamma_i)! = 0$, then we get Ommitted variable bias.

Now we run regression for this

```
cleaned_data_iou2 <-
  data %>%
  filter(iou == 2)
```

```
reg_7 <- lm(groundwater_use_v2 ~ electricity_price_pilot, data = cleaned_data_iou2)
summary(reg_7)
```

```
##
## Call:
## lm(formula = groundwater_use_v2 ~ electricity_price_pilot, data = cleaned_data_iou2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12043.7  -2185.5    155.2   2358.2   8996.4
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            14098.501    113.714  123.98   <2e-16 ***
## electricity_price_pilot  -219.806      5.557  -39.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3328 on 1943 degrees of freedom
## Multiple R-squared:  0.4461, Adjusted R-squared:  0.4458
## F-statistic:  1565 on 1 and 1943 DF,  p-value: < 2.2e-16
```

The p-value 2e-16 at 99% confidence level says that the coefficient of *electricity price pilot* is statistically significant. The coefficient value -219.806 means that for every dollar increase in electricity prices, it results in a decrease of 219.806 acre-foot of groundwater consumption. Thus the two variables are correlated i.e $Cov(electricitypricepilot_i, groundwaterusev2_i)! = 0$, $Cov(electricitypricepilot_i, groundwaterusev2_i) = -219.806$ for utility 2.

Question 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 for 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?

We can use $surveyprice_i$ as an Instrument Variable for the current variable *electricity price pilot* which has errors. $surveyprice_i = electricitypricepilot_i + \zeta_i$

To determine $\hat{\tau}^{IV}$, we do $\hat{\tau}^I V = \frac{Cov(groundwateruse_i, surveyprice_i)}{Cov(observedelectricityprice_i, surveyprice_i)}$

We get

$\hat{\tau}^{IV} = \frac{\tau.Cov(electricitypricepilot_i, surveyprice_i) + Cov(\epsilon_i, surveyprice_i)}{Cov(electricitypricepilot_i, surveyprice_i) + Cov(\gamma_i, surveyprice_i)}$

For utility 1

It is given in the question that 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.

This is Non - Classical measurement error in $electricityprice_i$, the treatment variable.

Assumptions: $Cov(surveyprice_i, observedelectricityprice_i)$!= 0 $Cov(surveyprice_i, \epsilon_i) = 0$ Measurement error is correlated with treatment i.e $Cov(\zeta_i, electricitypricepilot_i)! = 0$
Measurement error is uncorrelated with actual error i.e $Cov(\zeta_i, \gamma_i) = 0$ Measurement errorζ_i is uncorrelated with actual error ϵ_i i.e $Cov(\zeta_i, \epsilon_i) = 0$

We get, $\hat{\tau}^I V = \frac{\tau.Cov(electricitypricepilot_i, surveyprice_i) + Cov(\epsilon_i, surveyprice_i)}{Cov(electricitypricepilot_i, surveyprice_i) + Cov(\gamma_i, surveyprice_i)}$

to $\hat{\tau}^{IV} = \frac{\tau.Cov(electricitypricepilot_i, surveyprice_i)}{Cov(electricitypricepilot_i, surveyprice_i) + Cov(\gamma_i, electricitypricepilot_i)}$

And $\hat{\tau}^{IV}$ $! =$ τ. This is for the Non Classical Measurement problem.

The estimated $\hat{\tau}^{IV}$ cannot help in determining the accurate estimate of ATE of electricity price on groundwater consumption.

Now we run regression for this:

```
reg_8 <- ivreg(groundwater_use_v2 ~ electricity_price_pilot | survey_price, data = cleaned_data_iou1)
summary(reg_8)
```

```
##
## Call:
## ivreg(formula = groundwater_use_v2 ~ electricity_price_pilot |
##     survey_price, data = cleaned_data_iou1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12748.6  -2382.6   -688.4   1532.6  34266.4
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             19736.957    329.056   59.98   <2e-16 ***
## electricity_price_pilot   -82.547      3.764  -21.93   <2e-16 ***
##
## Diagnostic tests:
##                   df1  df2 statistic p-value
## Weak instruments    1 1995     327.0  <2e-16 ***
## Wu-Hausman          1 1994     841.7  <2e-16 ***
## Sargan              0   NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4847 on 1995 degrees of freedom
## Multiple R-Squared: -0.6355, Adjusted R-squared: -0.6363
## Wald test: 480.9 on 1 and 1995 DF,  p-value: < 2.2e-16
```

The p-value for variable eletricity_price_pilot 2e-16 at 99% says that the coefficient of the variable electricity_price_pilot is statistically significant. The coefficient value -82.547 means that for every one dollar increase in electricity prices, it results in a decrese of 82.547 acre-foot in groundwater consumption.THis means there is visible correlation between the two variables. i.e $Cov(electricitypricepilot_i, groundwaterusev2_i)! = 0$

For Utility 2

It is Classical measurement error in $electricityprice_i$, the treatment variable

Assumptions $Cov(surveyprice_i, observedelectricityprice_i)$!= 0
$Cov(surveyprice_i, \epsilon_i) = 0$
Measurement error is uncorrelated with treatment $Cov(\zeta_i, electricitypricepilot_i) = 0$
Measurement error is uncorrelated with the erro in the variable $Cov(\zeta_i, \gamma_i) = 0$ Measurement error is uncorrelated with original error $Cov(\zeta_i, \epsilon_i) = 0$

So we get $\hat{\tau}^{IV}$ as below, $\hat{\tau}^{IV} = \frac{\tau.Cov(electricitypricepilot_i, surveyprice_i) + Cov(\epsilon_i, surveyprice_i)}{Cov(electricitypricepilot_i, surveyprice_i) + Cov(\gamma_i, surveyprice_i)}$

to $\hat{\tau}^{IV} = \frac{\tau.Cov(electricitypricepilot_i, surveyprice_i)}{Cov(electricitypricepilot_i, surveyprice_i)}$

And $\hat{\tau}^{IV} = \tau$

From above, we can say that $\hat{\tau}^{IV}$ is helpful in determining an accurate estimate of the τ i.e ATE of $electricityprice$ on $groundwaterconsumption$

Regression for the above

```
reg_9 <- ivreg(groundwater_use_v2 ~ electricity_price_pilot | survey_price, data = cleaned_data_iou2)
summary(reg_9)
```

```
##
## Call:
## ivreg(formula = groundwater_use_v2 ~ electricity_price_pilot |
##     survey_price, data = cleaned_data_iou2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -15165.92  -3271.97     52.96   3376.87  15289.13
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             18092.03     356.57   50.74   <2e-16 ***
## electricity_price_pilot  -480.70      22.15  -21.70   <2e-16 ***
##
## Diagnostic tests:
##                    df1  df2 statistic p-value
## Weak instruments     1 1943     301.5  <2e-16 ***
## Wu-Hausman           1 1942     414.8  <2e-16 ***
## Sargan               0   NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4863 on 1943 degrees of freedom
## Multiple R-Squared: -0.1823, Adjusted R-squared: -0.183
## Wald test: 470.9 on 1 and 1943 DF,  p-value: < 2.2e-16
```

The p-value for electricity_price_pilot 2e-16 at the 99% confidence level says that the coefficient is statistically significant. THe coefficient value -480.7 means that for every dollar incrase in electricity prices, it results in a decrease of 480.7 acre-foot in groundwater consumption. Thus we can say ther is visible correlation between the two variables. i.e $Cov(electricitypricepilot_i, groundwaterusev2_i)! = 0$ $Cov(electricitypricepilot_i, groundwaterusev2_i) = -480.7$

I would send the Utility 2 estimates to CALBEARS as final results.