Assignment 3

May 12, 2022

Contents

Background

The California Agricultural Lobby's Bureau for Evidence-based Aquifer Recharge Sustainability (CALBEARS) is very concerned about the upcoming drought year, which looks to be one of the worst on record. In order to know how to best manage the state's water systems, they need to understand how the cost of extracting groundwater – which depends on (i) how efficient each farmer's pump is, (ii) how expensive their electricity is, and (iii) how deep the aquifer is – will impact farmers' groundwater usage. Their analysts are a bit in over their heads, so they've called you in to help.

Question 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?).

Answer

To estimate the effect of average groundwater costs between April and September on total groundwater consumption, ideally, we would have an experiment where each California farmer is randomly assigned different prices for groundwater. With a sufficient sample size and elimination of selection bias through randomization, we would ensure the only systematic change between farmers are the price they pay for groundwater. We might want to have multiple price points to get a better understanding of how incremental changes in price affect consumption. We would want to have data on price paid by customers for groundwater before and after the treatment, as well as their baseline and endline groundwater consumption. In this dataset, an individual farmer would ideally be the unit of analysis and we would do the randomization on the level of this unit. After random assignment of price changes, we could estimate the following regression to estimate ATE:

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

Where D_i is the price charged to the farmer i for groundwater. Thanks to randomization, this would allow us to estimate an unbiased of τ^{ATE}.

Under the potential outcomes framework, we would want to estimate:

$$\mathbb{E}[Y_i|D_i = p_1] - \mathbb{E}[Y_i|D_i = p_2]$$

But the farmer i can only have one average price for the April-September period, so the next best approach would be randomization to estimate $\hat{\tau}^{ATE}$ using the naive estimator at different price points.

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.

Answer

Given these limitation, we could implement an instrumental variable design. If we find an instrumental variable that is as good as randomly assigned and correlated with groundwater prices in some way, we can use this variable to isolate the effect of groundwater prices on consumption even though the prices do not fluctuate randomly in real world. For this design to be valid, we need to make some assumptions in addition to the assumption that changes electricity prices are as good as random:

Firstly, we need to assume that our instrument Z_i is correlated with the actual treatment (groundwater prices) D_i:

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

Secondly, we need an to assume that our instrument Z_i is only affects the groundwater consumption Y_i through the groundwater prices and nothing else. In other words, if ϵ_i is the error term we get by regressing groundwater prices on groundwater consumption, then we would assume:

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

This is called the exclusion restriction–Z_i cannot affect Y_i through any other channel, therefore, it can be excluded from a regression of Y_i on D_i. We could also show this through the omitted variable bias (OVB). With OVB we have:

$$\hat{\tau} = \tau + \beta \frac{Cov(D_i, X_i)}{Var(D_i)}$$

Using a valid instrument we would have make a prediction of D_i and retrieve $\hat{D}_i$. We could use it to estimate:

$$\tau^{IV} = \tau + \beta \frac{Cov(\hat{D}_i, X_i)}{Var(\hat{D}_i)}$$

The numerator in the second term is our exclusion restriction. As we assume that it is 0, we effectively assume that $\tau^{IV} = \tau$.

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?

Answer

Since the groundwater prices are affected by the cost of electricity that is used to extract it, we can think of electricity use as a plausible instrumental variable. By choosing electricity price as an instrument, we would first assume that electricity prices are correlated with groundwater prices, which is known to be true given the background information and can also be tested using data. Secondly, we would make the exclusion restriction assumption, which would claim that electricity prices do not affect groundwater consumption through anything other than groundwater prices. This claim is fundamentally untestable. One of our concerns could be that electricity prices are likely to be non-random, they are affected by other patterns in economy. Another concern could be that electricity prices could also increase farmers' other costs such as fertilizers or operating costs of agricultural machinery, which in turn would decrease the disposable income a farmer could spend on irrigation, therefore leading to a lower consumption of groundwater resources not necessarily through prices of groundwater. This would invalidate our second assumption related to exclusion restriction.

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.

Answer

Using the information on randomly varied electricity prices, we could estimate τ^{IV} as:

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

where Z_i is electricity prices, D_i is groundwater prices, and X_i is other covariates we might add to the model. This is also a test of our first assumption–we would expect to see some statistically significant relationship between D_i and Z_i. From this model, we would store the predicted values of $\hat{D}_i$. This would give us the part of treatment of insterest D that we could explain using our instrumental variable Z.

Using the results from the prediction above, we would regress outcome Y_i on predicted $\hat{D}_i$ and other Xs we included in the first stage:

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

This would give us an estimate for τ^{IV} and we could interpret it as our effect of interest.

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?

Answer

```
water <- read_csv("ps3_data.csv")

stage_1 <- lm(groundwater_cost ~ electricity_price_pilot, data = water)
stargazer(stage_1, type = "latex", header = FALSE, title = "First stage regression results")
```

As we suspected, electricity prices seem to be positively correlated with groundwater prices. This estimate tells us that with each additional increase by $1 in electricity costs, we can expect groundwater prices to go up by $0.74. This statistically significant relationship and high F-statistic suggests that our first assumption is valid and we can move forward with our instrumental variable estimation if we have reason to believe our second assumption as well.

Table 1: First stage regression results

	Dependent variable:
	groundwater_cost
electricity_price_pilot	0.740***
	(0.043)
Constant	291.173***
	(3.519)
Observations	4,000
R^2	0.068
Adjusted R^2	0.068
Residual Std. Error	176.849 (df = 3998)
F Statistic	292.738*** (df = 1; 3998)
Note:	*p<0.1; **p<0.05; ***p<0.01

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?

Answer

```
reduced_form <- lm(groundwater_use ~ electricity_price_pilot,
    data = water)
stargazer(reduced_form, type = "latex", header = FALSE, title = "Reduced form regression results")
```

This model gives us the reduced form regression. If our assumptions are valid, we can assume a causal relationship between the dependent variable `groundwater_use` and `electricity_price_pilot`. With this model we estimate the θ in the reduced form:

$$Y_i = \alpha + Z_i + \eta_i$$

which is the effect of our electricity prices on groundwater consumption.

We see that each additional dollar in electricity costs result in 147.498 acre-feet less consumption of groundwater. This estimation is useful for policy as it gives us a clear relationship between electricity prices and groundwater consumption. For instance, it suggests we can increase groundwater consumption by subsidizing electricity production.

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

Table 2: Reduced form regression results

	Dependent variable:
	groundwater_use
electricity_price_pilot	−147.498***
	(16.077)
Constant	108,301.500***
	(1,307.493)
Observations	4,000
R^2	0.021
Adjusted R^2	0.020
Residual Std. Error	65,717.690 (df = 3998)
F Statistic	84.172*** (df = 1; 3998)
Note:	*p<0.1; **p<0.05; ***p<0.01

your standard errors, so using a canned routine is a good idea here. Interpret your effect. Do groundwater costs matter for consumption?

Answer

Step-by-step

```
# First stage
stage_1 <- lm(groundwater_cost ~ electricity_price_pilot, data = water)
stargazer(stage_1, type = "latex", header = FALSE, title = "First stage regression results")
```

Table 3: First stage regression results

	Dependent variable:
	groundwater_cost
electricity_price_pilot	0.740***
	(0.043)
Constant	291.173***
	(3.519)
Observations	4,000
R^2	0.068
Adjusted R^2	0.068
Residual Std. Error	176.849 (df = 3998)
F Statistic	292.738*** (df = 1; 3998)
Note:	*p<0.1; **p<0.05; ***p<0.01

```r
# Second stage
stage_2 <- lm(water$groundwater_use ~ stage_1$fitted.values,
    data = water)
stargazer(stage_2, type = "latex", header = FALSE, title = "Second stage regression results")
```

Table 4: Second stage regression results

	Dependent variable:
	groundwater_use
fitted.values	−199.263***
	(21.719)
Constant	166,321.400***
	(7,193.070)
Observations	4,000
R^2	0.021
Adjusted R^2	0.020
Residual Std. Error	65,717.690 (df = 3998)
F Statistic	84.172*** (df = 1; 3998)
Note:	*p<0.1; **p<0.05; ***p<0.01

Canned routine

```r
iv_reg <- ivreg(groundwater_use ~ groundwater_cost | electricity_price_pilot,
    data = water)
stargazer(iv_reg, type = "latex", title = "Canned routine IV results",
    header = F)
```

Table 5: Canned routine IV results

	Dependent variable:
	groundwater_use
groundwater_cost	−199.263***
	(18.032)
Constant	166,321.400***
	(5,972.083)
Observations	4,000
R^2	0.325
Adjusted R^2	0.325
Residual Std. Error	54,562.450 (df = 3998)
Note:	*p<0.1; **p<0.05; ***p<0.01

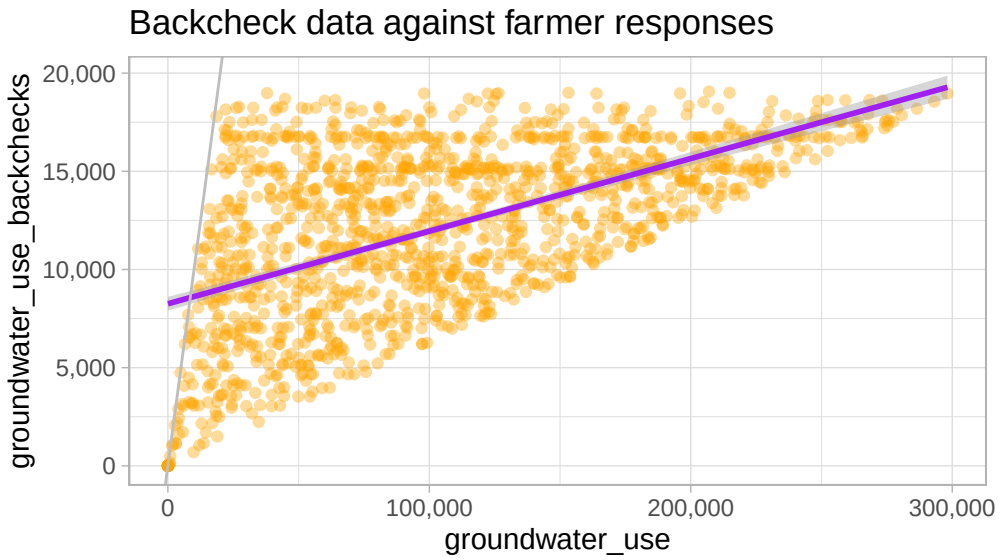
We estimate τ^{IV} as -199.263. This suggests that with each dollar increase in groundwater costs, the groundwater consumption decreases by -199.263 acre-feet. The costs seem to have a strong causal effect on consumption of groundwater.

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.

Answer

```
water %>%
    na.omit() %>%
    ggplot(aes(y = groundwater_use_backchecks, x = groundwater_use)) +
    geom_point(color = "orange", alpha = 0.4) + geom_smooth(method = "lm",
    color = "purple") + scale_y_continuous(labels = comma_format()) +
    scale_x_continuous(labels = comma_format()) + geom_abline(color = "gray") +
    labs(title = "Backcheck data against farmer responses") +
    theme_light()
```



Overall, farmers' estimates seem to be extremely overstated compared to the data from back-checks. Secondly, as consumption increases, farmers start to exaggerate their consumption even more and respond with values farther from their actual use. This pattern should not change the direction and significance of the value we estimated. However, the overall overstatement of use would certainly impact the value we wanted to estimate (possibly shrink it as groundwater use is, in general, not as high as farmers claimed).

```
iv_reg_backchecks <- ivreg(groundwater_use_backchecks ~ groundwater_cost |
    electricity_price_pilot, data = water)
stargazer(iv_reg_backchecks, type = "latex", title = "Canned routine IV results with backcheck data",
    header = F)
```

Table 6: Canned routine IV results with backcheck data

	Dependent variable:
	groundwater_use_backchecks
groundwater_cost	−25.001***
	(0.236)
Constant	20,035.650***
	(75.827)
Observations	1,389
R^2	0.991
Adjusted R^2	0.991
Residual Std. Error	416.957 (df = 1387)
Note:	*p<0.1; **p<0.05; ***p<0.01

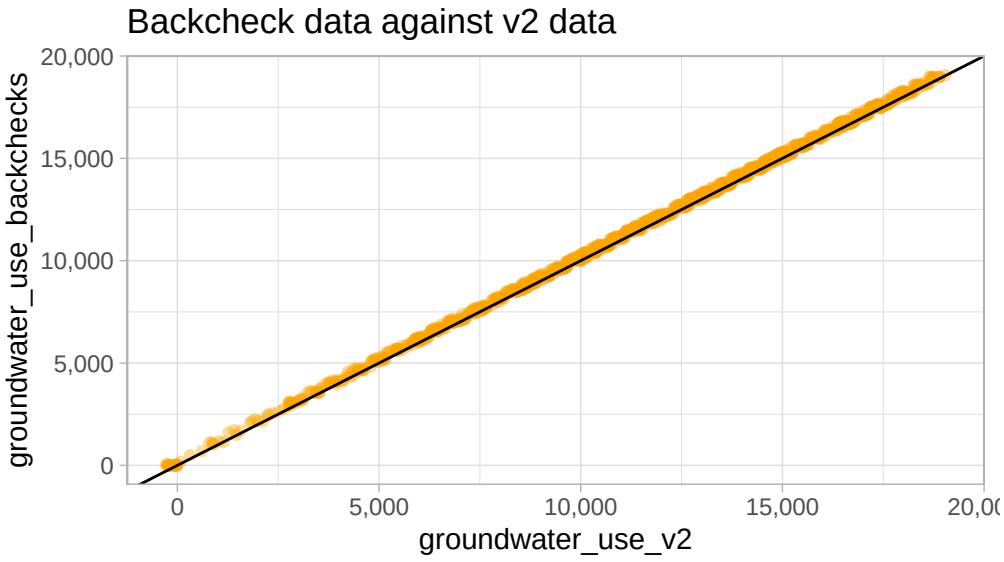
As suspected, the estimates we get from backcheck data is much lower than the prior estimate. This is due to the overall tendency of farmers to overstate their use. The gray line in the figure above show the 1:1 relationship, and all estimates are either on the line or to the right of the, indicating upwards bias. We should also note that, possibly due to high costs of backchecking, the number of observations in the second analysis drop to 1,389 from 4,000.

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.

Answer

```
water %>%
    na.omit() %>%
    ggplot(aes(y = groundwater_use_backchecks, x = groundwater_use_v2)) +
    geom_point(color = "orange", alpha = 0.4) + scale_y_continuous(labels = comma_format()) +
    scale_x_continuous(labels = comma_format()) + geom_abline() +
    labs(title = "Backcheck data against v2 data") + theme_light()
```



The `groundwater_use_v2` variable is much better aligned with the backcheck data, although it is still slightly below the backcheck values (as most of the points fell above the diagonal line). In general, this should not be a big problem for our analysis as the bias seems to be negligible.

```
iv_reg_v2 <- ivreg(groundwater_use_v2 ~ groundwater_cost | electricity_price_pilot,
    data = water)
stargazer(iv_reg_v2, type = "latex", title = "Canned routine IV results with v2 data",
    header = F)
```

Table 7: Canned routine IV results with v2 data

	Dependent variable:
	groundwater_use_v2
groundwater_cost	−25.014***
	(0.122)
Constant	19,866.880***
	(40.553)
Observations	4,000
R^2	0.993
Adjusted R^2	0.993
Residual Std. Error	370.500 (df = 3998)
Note:	*p<0.1; **p<0.05; ***p<0.01

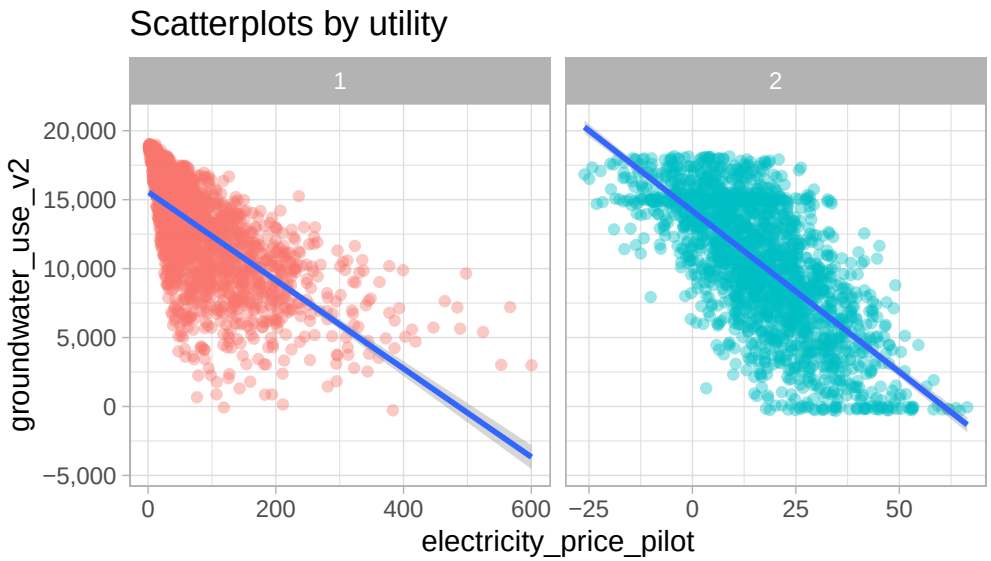
As the v2 data was much more closely aligned with the actual backcheck values, the τ^{IV} estimate we get from it is much closer to the one we get from the model that used the backchecked data.

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.

Answer

```
water %>%
    ggplot(aes(x = electricity_price_pilot, y = groundwater_use_v2)) +
    geom_point(aes(color = as.factor(iou)), alpha = 0.4) + geom_smooth(method = "lm") +
    facet_wrap(~iou, scales = "free_x") + scale_y_continuous(labels = comma_format()) +
    scale_x_continuous(labels = comma_format()) + labs(title = "Scatterplots by utility") +
    theme_light() + theme(legend.position = "none")
```




```
reduced_form_1 <- lm(groundwater_use_v2 ~ electricity_price_pilot,
    data = filter(water, iou == 1))
reduced_form_2 <- lm(groundwater_use_v2 ~ electricity_price_pilot,
    data = filter(water, iou == 2))

stargazer(reduced_form_1, reduced_form_2, type = "latex", header = F,
    title = "OLS results by utility")
```

Table 8: OLS results by utility

	Dependent variable:	
	groundwater_use_v2	
	(1)	(2)
electricity_price_pilot	-32.031^{***}	-233.421^{***}
	(0.843)	(5.343)
Constant	$15{,}560.000^{***}$	$14{,}178.250^{***}$
	(95.249)	(114.702)
Observations	1,999	2,001
R^2	0.420	0.488
Adjusted R^2	0.419	0.488
Residual Std. Error	2,902.612 (df = 1997)	3,404.903 (df = 1999)
F Statistic	$1{,}444.535^{***}$ (df = 1; 1997)	$1{,}908.880^{***}$ (df = 1; 1999)
Note:		*p<0.1; **p<0.05; ***p<0.01

For `iou == 1`, we are underestimating the relationship between electricty prices and groundwater use due to the measurement error in electricity prices. As electricity prices increase, the recorded values become more inflated. As measurements get higher than it should be, we underestimate the effect of electricity prices on groundwater use since one unit increase in electricity use is actually having a bigger impact than what we see in the data with measurement error. In this case, the orthogonality assumption is violated and the measurement error is correlated with treatment. Therefore, we estimate a noisy version of Z_i,

$$\tilde{Z}_i = Z_i + \gamma_i$$

Therefore, we have

$$\hat{\tau} = \frac{Cov(Y_i, Z_i + \gamma_i)}{Var(Z_i + \gamma_i)}$$

Which can be rewritten as

$$\tau \frac{Var(Z_i) + Cov(Z_i, \gamma_i)}{Var(Z_i) + Var(\gamma_i) + 2Cov(Z_i, \gamma_i)}$$

Since we know the direction of the correlation between Z and γ to be positive, we can say that the estimate of τ is underestimated.

For `iou == 2`, we are again underestimating the relationship between electricty prices and groundwater use due to the measurement error in electricity prices. Though this measurement error is different from the error we see in the first utility. Here, measurement error leads to attenuation bias and causing $\hat{\tau}$ to be biased towards zero. We are suing a noisy version of electricity prices, $\tilde{Z}_i$, to run

$$Y_i = \alpha + \tau \tilde{Z}_i + \epsilon_i$$

which estimates

$$\begin{aligned}
\hat{\tau} &= \frac{Cov(Y_i, \tilde{Z}_i)}{Var(\tilde{Z}_i)} \\
&= \frac{Cov(\alpha + \tau D_i + \epsilon_i, D_i + \gamma_i)}{Var(D_i) + Var(\gamma_i)} \\
&= \frac{\tau Var(D_i)}{Var(D_i) + Var(\gamma_i)}
\end{aligned}$$

through our assumptions and linear algebra. The denominator in the final expression causes our estimates of τ to shrink.

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?

Answer

In utility 1, since the orthogonality assumption is violated, we cannot use an instrumental variable approach to overcome this non-classical measurement error.

In utility 2, however, we have a classical measurement error which can be overcome using the `survey_price` variable as an instrument. For this to work, we need to assume that: (1) measurement error is uncorrelated with treatment, (2) measurement error in `electricity_price_pilot` (Z_i) is uncorrelated with error in `survey_price`, and (3) measurement error is uncorrelated with original error. These assumptions allow us to solve this problem with two independently noisy measures of our variable.

```
utility_2 <- ivreg(groundwater_use_v2 ~ electricity_price_pilot |
    survey_price, data = filter(water, iou == 2))
stargazer(utility_2, type = "latex", header = FALSE, title = "IV results with survey prices as instrume
```

Table 9: IV results with survey prices as instrument

	Dependent variable:
	groundwater_use_v2
electricity_price_pilot	−460.070***
	(17.498)
Constant	17,818.450***
	(299.990)
Observations	2,001
R^2	0.028
Adjusted R^2	0.027
Residual Std. Error	4,693.730 (df = 1999)
Note:	*p<0.1; **p<0.05; ***p<0.01

As we have suspected, once we use another noisy measure of electricity prices (survey_price) as instrument, we are able to mitigate the underestimation seen in Question 10. Based on these results, each additional dollar in electricity prices seem to decrease the groundwater consumption by 460.07 acre-feet. Because of the measurement errors in 10, this model is the most reliable model to understand the relationship between electricity prices and groundwater consumption by California farmers.