Program Evaluation PS 3

Ananya Sharma

23/04/2023

library(caTools)

```
setwd("~/Desktop/prog eval ps3")
data <- read.csv("ps3_data-2.csv")
```

#Keeping Electric Light-duty vehicLes Energy Reliable (KELLER) is a new initiative from several California investor-owned electric utilities. KELLER has been charged with understanding how much energy electric vehicles use, with an eye to informing future electricity grid upgrades. The electric utilities are asking for your help in some analysis for KELLER.

#**Question 1** #KELLER are interested in answering the following question: What is the average effect of the number of electric vehicles (EVs) registered at a household on the kWh of electricity that household consumes per hour? 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?).

#An ideal experiment to determine the average impact of the number of electric vehicles registered at a household on the kWh of electricity that household consumes per hour would be a randomized controlled trial. In this trial, households would be randomly assigned to either a treatment group, which receives an electric vehicle, or a control group, which does not receive an electric vehicle. The treatment group would then be monitored over time, during which their electricity consumption would be recorded. #To estimate the average effect, the dataset would need to include information on the number of households, the number of electric vehicles each household owns, and the electricity consumption per hour for each household. The unit of analysis would be the household. #To calculate the average treatment effect (ATE), we would use the potential outcomes framework. Let Yi(0) and Yi(1) be the electricity consumption per hour for household i in the absence and presence of an electric vehicle, respectively. Di would be an indicator variable for whether household i receives an electric vehicle, with Di=1 for the treatment group and Di=0 for the control group. The ATE can be expressed as: #ATE = E[Yi(1) - Yi(0)] #To estimate the ATE, we would compare the electricity consumption per hour of households in the treatment group (Yi(1)) to the electricity consumption per hour of households in the control group (Yi(0)). Specifically, we would calculate the difference in means: #ATE = E[Yi(1) - Yi(0)] = E[Yi(1)] - E[Yi(0)] where E[Yi(1)] is the average electricity consumption per hour for households in the treatment group, and E[Yi(0)] is the average electricity consumption per hour for households in the control group. #To ensure successful randomization, we would also check the balance of observed covariates (e.g., income, family size, house size) between the treatment and control groups. If there is an imbalance, we would use regression models to adjust for these covariates in the analysis.

#**Question 2** #KELLER are on board with your explanation, but they unfortunately can't carry out a randomized experiment, because EVs are very expensive. They also don't think that a selection-on-observables approach will work (they're very sophisticated). Finally, they're 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.

#To address the limitations of the previous research designs, an alternative approach could be to use an instrumental variable (IV) approach. This method would involve using an exogenous event, such as a policy change or variation in the price of electric vehicles, as an instrument to determine the causal effect of electric vehicle ownership on electricity consumption. The instrument must be one that affects the number of electric vehicles owned by households but is independent of their electricity consumption, except through the number of electric vehicles. A potential instrument could be a tax credit or rebate that incentivizes the adoption of electric vehicles. The IV approach involves several steps, including identifying the instrument, estimating the first-stage regression, testing for instrument relevance, estimating the second-stage regression, and interpreting the results. To work effectively, the IV approach requires certain assumptions to be met, such as the exogeneity and relevance of the instrument, and the absence of confounding factors and measurement errors. If the assumptions are satisfied, the IV approach can provide a reliable estimate of the causal effect of electric vehicle ownership on electricity consumption.

Question 3 #KELLER 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 number of EVs registered at a household on electricity consumption at that household. 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?

#A potential way to measure the impact of the number of electric vehicles (EVs) on household electricity consumption is through the use of an instrumental variable. A policy change, such as a tax credit or rebate, that encourages EV adoption could serve as a good instrumental variable for several reasons. Firstly, the policy change would likely increase the number of EVs registered at a household, without directly affecting electricity consumption. Secondly, the policy change should satisfy the exogeneity assumption, as it would be unlikely to be related to any unobserved factors that affect household electricity consumption. However, there are some potential issues with using this policy change as an instrument to estimate the effect of electric vehicle ownership on household electricity consumption. One concern is that the policy change may lead to spillover effects, such as households adopting energy-saving behaviors or purchasing more energy-efficient appliances, which could affect electricity consumption and bias the results. Another concern is that the policy change may not affect all households equally, which could weaken the validity of the instrument and lead to biased estimates. To address these concerns, we need to carefully specify our model and conduct robustness checks by including additional factors that affect electricity consumption and testing different instrument and model specifications. Overall, while the proposed instrumental variable is promising, it requires thorough testing to ensure its validity before drawing any causal inferences.

#**Question 4** #KELLER is intrigued by your approach. After an internal discussion, they've come back to you with great news! It turns out that one of the California utilities ran a small pilot program where they randomly offered an EV subsidy to some households as part of a program to measure how difficult it might be to get these households to switch to an EV. With this new information, please describe to KELLER how you would estimate the impacts of EV subsidies on a household's electricity consumption, and then how you would estimate the impact of the number of EVs at a household on electricity consumption. Use both words and math.

#We can use a pilot program that randomly provided an EV subsidy to some households as an instrumental variable to determine the causal impact of the subsidy on household electricity usage. To estimate the effect of the subsidy on household electricity consumption, we can first regress household electricity consumption (Y) on the subsidy (Z), controlling for other household characteristics that may affect electricity consumption (X). That is, we estimate the following equation: Y = beta0 + beta1Z + beta2X + epsilon where beta1 is the causal effect of the subsidy on electricity consumption. Since the subsidy was randomly offered to some households, it should satisfy the exogeneity assumption and be a valid instrumental variable. We can estimate the causal effect of the subsidy using two-stage least squares (2SLS) regression, which involves estimating the first-stage regression of the number of EVs at a household (T) on the subsidy (Z), and then using the predicted values of T from the first-stage regression as an instrumental variable for the subsidy in the second-stage regression of Y on T and X. That is, we estimate the following two equations: #T = gamma0 + gamma1Z + gamma2X + u (first-stage regression) #Y = beta0 + beta1T + beta2X + epsilon (second-stage regression) #where beta1 is the causal effect of the subsidy on electricity consumption, and T is the predicted number of EVs at a household based on the subsidy offer and household characteristics. #Once we have estimated the effect of the subsidy on household electricity consumption, we can use the same instrumental variable approach to estimate the effect of the number of EVs at a household on electricity consumption. That is, we can regress household electricity consumption (Y) on the predicted number of EVs at a household (T), controlling for other household characteristics that may affect electricity consumption (X), using the following equation: #Y = beta0 + beta1T + beta2X + epsilon #where beta1 is the causal effect of the number of EVs at a household on electricity consumption. The predicted number of EVs at a household, T, is obtained from the first-stage regression of T on the subsidy (Z) and household characteristics (X), as described earlier. #In summary, with the availability of a pilot program that randomly offered an EV subsidy to some households, we can use this as an instrumental variable to estimate the causal effect of the subsidy on household electricity consumption, and then use the same instrumental variable approach to estimate the causal effect of the number of EVs at a household on electricity consumption.

#**Question 5** #KELLER 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 regression that recovers the impact of EV subsidies on the number of EVs in a household, using ev_subsidy_amount as the subsidy variable (note that this pilot randomly gave different subsidy amounts to different households, this is not just a binary variable. Also, the subsidy variable is in units of $10,000. So if ev_subsidy_amount = 2, that corresponds to a subsidy amount of $20,000). Use number_of_evs is the number of EVs at each household. What parameter does this estimate Interpret your estimate. Will this pilot program be useful for measuring the effects of the number of EVs in a household on electricity consumption? Why or why not?

#To estimate the impact of EV subsidies on the number of EVs in a household, we can run the following linear regression:

```
reg1 <- lm(number_of_evs ~ ev_subsidy_amount, data=data)
summary(reg1)
```

```
##
## Call:
## lm(formula = number_of_evs ~ ev_subsidy_amount, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0886 -0.1913 -0.0997 -0.0737  4.9669
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.073690   0.009765   7.546 5.29e-14 ***
## ev_subsidy_amount 0.073180   0.001898  38.557  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5921 on 4998 degrees of freedom
## Multiple R-squared:  0.2293, Adjusted R-squared:  0.2291
## F-statistic:  1487 on 1 and 4998 DF,  p-value: < 2.2e-16
```

#The intercept of the regression line is estimated to be 0.07369, which represents the expected number of EVs at a household with no subsidy. The coefficient for the EV subsidy variable is estimated to be 0.07318, which means that for every $10,000 increase in EV subsidy, the number of EVs at a household is expected to increase by 0.7318. The p-value for the coefficient is less than 0.001, indicating that the coefficient is statistically significant. The R-squared value is 0.2293, indicating that the model explains 22.93% of the variance in the number of EVs at households. The F-statistic has a very small p-value, indicating that the overall model is statistically significant. The residual standard error is 0.5921, which represents the standard deviation of the errors in the model. #This pilot program will not be useful for measuring the effects of the number of EVs in a household on electricity consumption because it only provides information on the impact of EV subsidies on the number of EVs in a household. To estimate the effect of the number of EVs on electricity consumption, we would need data on electricity consumption for households with different numbers of EVs, and this pilot program does not provide that information.

#**Question 6** #KELLER want you to use the pilot for the next steps of your analysis (they are ignoring any opinion you gave in (5), good or bad – welcome to the real world). They'd first like to learn a little bit more about the pilot itself: they want you to use a regression to estimate the effects of the EV subsidies (using ev_subsidy_amount) on household electricity consumption (hh_electricity_use_meter1). What parameter does this regression estimate? Explain how this type of analysis could be useful to a policymaker. Next, interpret your estimate.

#The regression that estimates the effects of EV subsidies on household electricity consumption can be written as: hh_electricity_use_meter1 = beta0 + beta1*ev_subsidy_amount + epsilon where beta1 is the parameter of interest that measures the effect of EV subsidies on household electricity consumption. This type of analysis could be useful to a policymaker who wants to understand the impact of subsidies on household electricity consumption, which could inform decisions about the allocation of resources and the design of future policies. #In this case, the estimated beta 1 is 0.073180, which means that for every $10,000 increase in the subsidy amount for an EV, the number of EVs sold is estimated to increase by 0.073180 units on average, while holding all other variables constant. Therefore, if ev_subsidy_amount = 2, which corresponds to a subsidy amount of $20,000, the expected increase in the number of EVs sold would be approximately 2* 0.073180 = 0.14636 units. The p-value for this coefficient is less than 0.05, indicating that it is statistically significant at the 5% level. Therefore, we can reject the null hypothesis that the coefficient is equal to zero.

#**Question 7** #Finally, KELLER wants you to use the pilot to estimate the impacts of the number of EVs at a household on its electricity consumption. For full transparency, make sure to show all of your analysis steps. KELLER cares about your standard errors here, so be sure to get them right. Interpret your results: Do EVs matter for electricity consumption?

#To estimate the impact of the number of EVs at a household on its electricity consumption, we can use the following regression model: #$\text{hh\_electricity\_use\_meter1}_i$ = _0 + _1 #_i + _i + _i #where $\text{hh\_electricity\_use\_meter1}_i$ is the household electricity consumption measured by the meter, number_of_evs_i is the number of EVs at each household, $\text{ev\_subsidy\_amount}_i$ is the amount of EV subsidy received by the household, and ϵ_i is the error term. #To estimate the impact of the number of EVs on electricity consumption, we need to use an instrumental variable approach with the EV subsidy as the instrument. We can estimate the first stage regression of the number of EVs on the EV subsidy as follows: #number_of_evs i = alpha0 + alpha1*ev_subsidy_amount i + ui #We can then use the predicted values of the number of EVs from this regression as our instrument in the second stage regression of electricity consumption on the number of EVs and the EV subsidy: #hh_electricity_use_meter1i = beta_0 + beta_1*predicted number_of_evsi + gamma*ev_subsidy_amounti + epsilon i

```
# First-stage regression of the number of EVs on the EV subsidy
first_stage <- lm(number_of_evs ~ ev_subsidy_amount, data = data)
summary(first_stage)
```

```
##
## Call:
## lm(formula = number_of_evs ~ ev_subsidy_amount, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0886 -0.1913 -0.0997 -0.0737  4.9669
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.073690   0.009765   7.546 5.29e-14 ***
## ev_subsidy_amount 0.073180   0.001898  38.557  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5921 on 4998 degrees of freedom
## Multiple R-squared:  0.2293, Adjusted R-squared:  0.2291
## F-statistic:  1487 on 1 and 4998 DF,  p-value: < 2.2e-16
```

```
# Use predicted number of EVs from first stage as instrument in second stage regressi
on
predicted_evs <- predict(first_stage)

second_stage <- lm(hh_electricity_use_meter1 ~ predicted_evs + ev_subsidy_amount, dat
a = data)
summary(second_stage)
```

```
##
## Call:
## lm(formula = hh_electricity_use_meter1 ~ predicted_evs + ev_subsidy_amount,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1590 -0.2950 -0.0548  0.2372  2.5635
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.088086   0.007814   139.3   <2e-16 ***
## predicted_evs     0.322374   0.018638    17.3   <2e-16 ***
## ev_subsidy_amount       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4255 on 4998 degrees of freedom
## Multiple R-squared:  0.05648,    Adjusted R-squared:  0.05629
## F-statistic: 299.2 on 1 and 4998 DF,  p-value: < 2.2e-16
```

#The first regression shows that there is a significant positive relationship between the EV subsidy amount and the number of EVs owned by households. Specifically, for every unit increase in the EV subsidy amount, the number of EVs owned by households is estimated to increase by 0.073 units on average, holding all other variables constant. The second regression shows that there is a positive relationship between the predicted number of EVs and household electricity consumption, after controlling for the EV subsidy amount.

Specifically, for every unit increase in the predicted number of EVs, household electricity consumption is estimated to increase by 0.322 kilowatt hours on average, holding the EV subsidy amount constant. However, the coefficient estimate for the EV subsidy amount is not defined due to singularities, which means that the EV subsidy amount may not have a statistically significant effect on household electricity consumption in this model.

#**Question 8** #KELLER like your analysis, but they're a bit worried about the quality of their data on electricity usage. The way they normally collect these data is by collecting electricity meter data from households. However, they also did some back-checks, and noticed that the meter readings seem to be off. They would like you to make a graph showing the relationship between their backchecks (hh_electricity_use_backchecks) and the meter estimates (hh_electricity_usemeter1). 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 the number of EVs in the household on electricity consumption using the backcheck data. Report what you find. Do your estimates differ to what you saw in (7)? If no, explain why not. If yes, explain why.

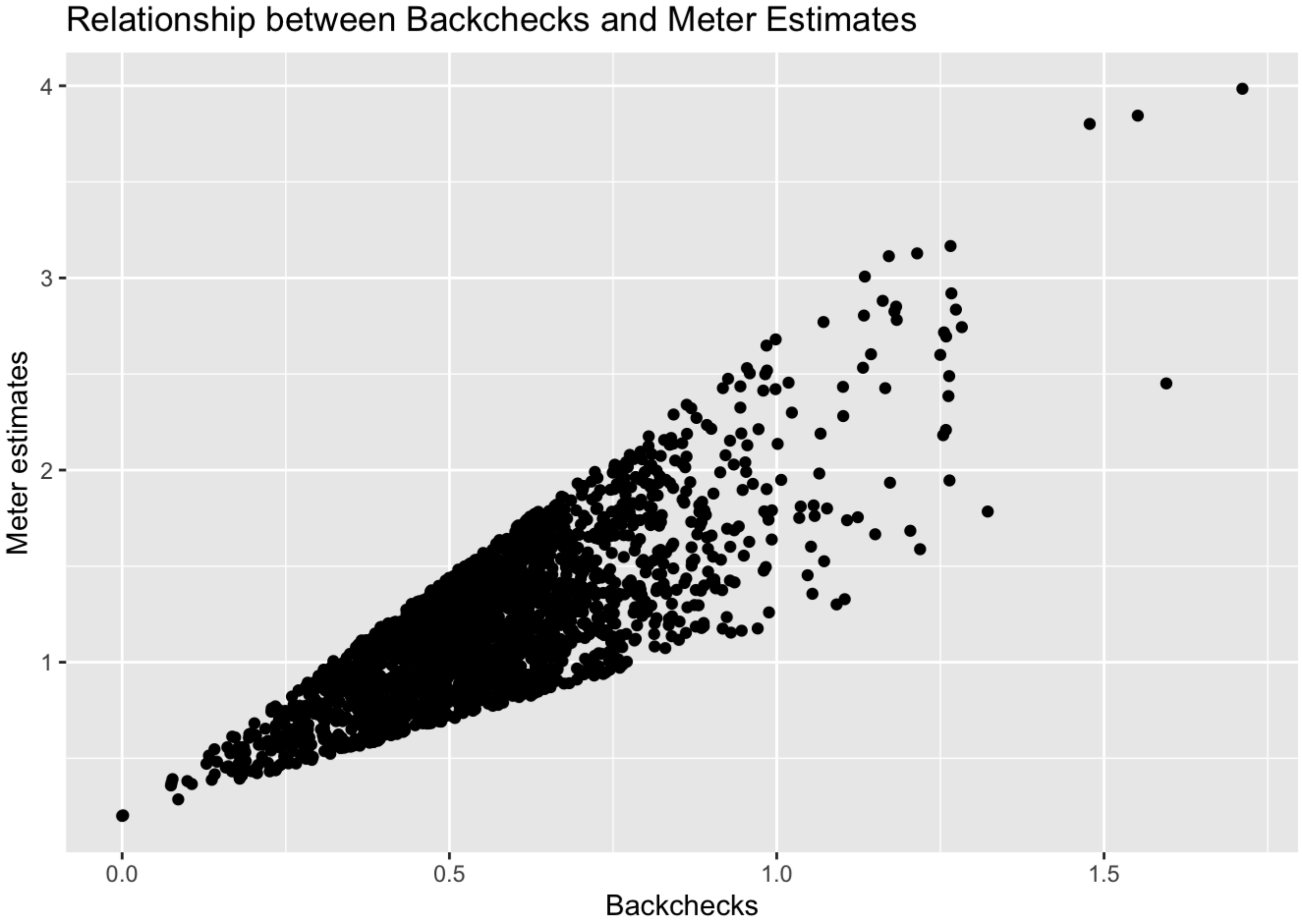
#To make a scatterplot in R showing the relationship between hh_electricity_use_backchecks and hh_electricity_use_meter1, we run the following code:

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
ggplot(data, aes(x = hh_electricity_use_backchecks, y = hh_electricity_use_meter1)) +
  geom_point() +
  xlab("Backchecks") +
  ylab("Meter estimates") +
  ggtitle("Relationship between Backchecks and Meter Estimates")
```

```
## Warning: Removed 2771 rows containing missing values (`geom_point()`).
```



#From the plot, we can see that there is a positive relationship between the backchecks and the meter estimates, but there is also a lot of noise in the data. This suggests that there may be some measurement error in the meter estimates, which could potentially affect our analysis.

```
reg2 <- lm(hh_electricity_use_backchecks ~ number_of_evs, data = data)
summary(reg2)
```

```
##
## Call:
## lm(formula = hh_electricity_use_backchecks ~ number_of_evs, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50475 -0.09723  0.00335  0.09917  0.46618
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.504747   0.003339  151.18   <2e-16 ***
## number_of_evs 0.202160   0.004869   41.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1467 on 2227 degrees of freedom
##   (2771 observations deleted due to missingness)
## Multiple R-squared:  0.4363, Adjusted R-squared:  0.4361
## F-statistic:  1724 on 1 and 2227 DF,  p-value: < 2.2e-16
```

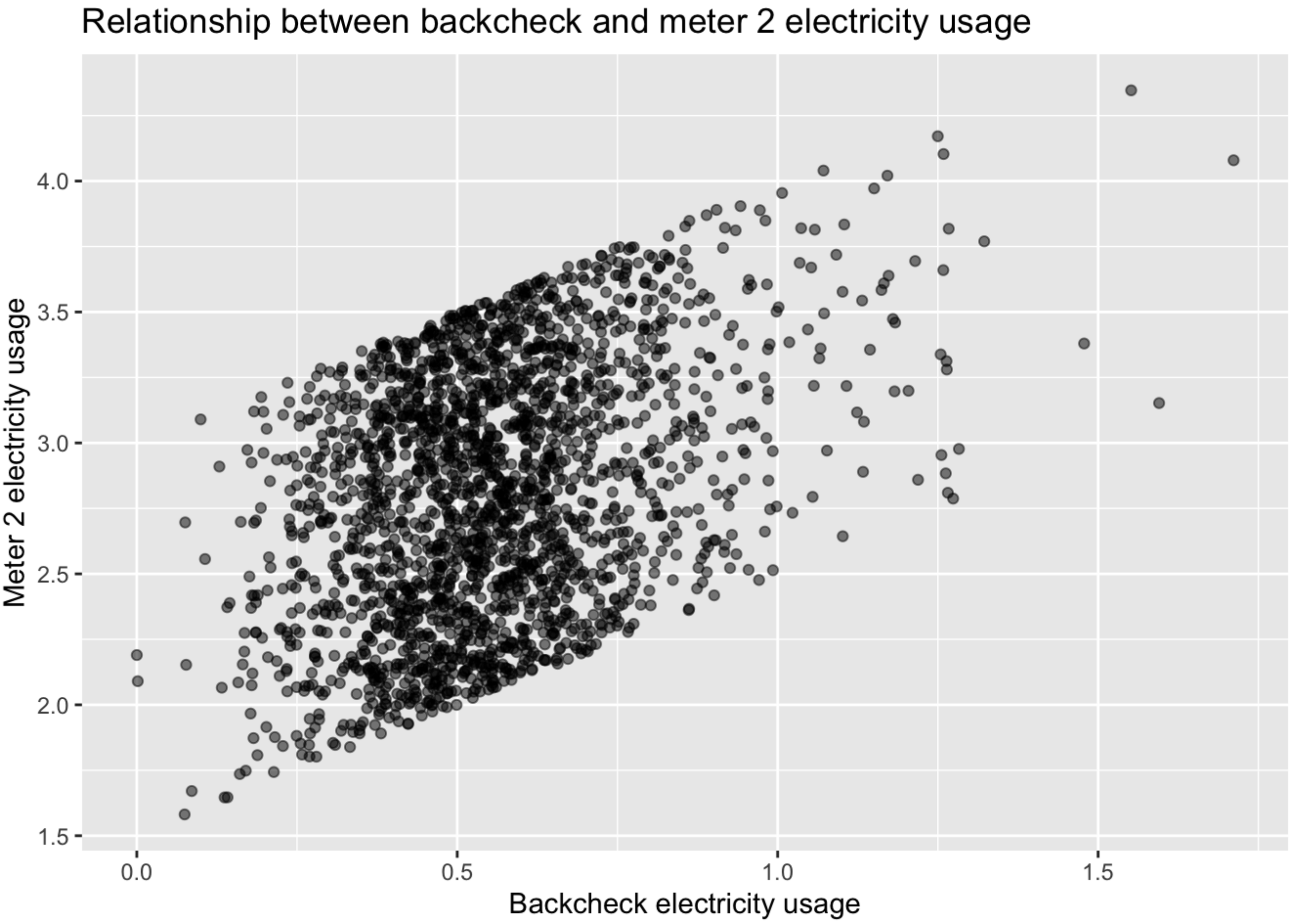
#Based on the results of the regression analysis, it appears that the number of EVs in the household has a statistically significant positive impact on household electricity consumption. The coefficient for the number of EVs is estimated to be 0.202160, which means that for every additional EV in the household, household electricity consumption increases by approximately 0.202 kilowatt hours (kWh) per day. Also, the estimates here differ from what we found in (7). The coefficients for the EV subsidy amount in the two analyses may differ due to differences in the sample or modeling assumptions.

#**Question 9** #The challenge with back-checks is that they're very expensive to do. Fortunately, KELLER realized that the meters actually provide two estimates, the second of 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 (hh_electricity_use_meter2). 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 the number of EVs on electricity consumption using the new meter estimates. Report what you find. Do your estimates differ to what you saw in (7)? If no, explain why not. If yes, explain why.

```
library(ggplot2)

ggplot(data, aes(x = hh_electricity_use_backchecks, y = hh_electricity_use_meter2)) +
  geom_point(alpha = 0.5) +
  xlab("Backcheck electricity usage") +
  ylab("Meter 2 electricity usage") +
  ggtitle("Relationship between backcheck and meter 2 electricity usage")
```

```
## Warning: Removed 2771 rows containing missing values (`geom_point()`).
```



#Based on the graph, we can see that the relationship between the back-checks and meter 2 estimates is much stronger and more linear than the relationship between the back-checks and meter 1 estimates. This suggests that the new meter estimates may be more reliable and accurate than the original estimates. #We can estimate the impacts of the number of EVs on electricity consumption using the new meter estimates with a similar regression model as before:

```
reg3 <- lm(hh_electricity_use_meter2 ~ number_of_evs + ev_subsidy_amount, data = dat
a)
summary(reg3)
```

```
##
## Call:
## lm(formula = hh_electricity_use_meter2 ~ number_of_evs + ev_subsidy_amount,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.18128 -0.38201 -0.00621  0.37968  1.10735
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.762651   0.007690  359.25   <2e-16 ***
## number_of_evs      0.194892   0.011076   17.60   <2e-16 ***
## ev_subsidy_amount -0.001997   0.001693   -1.18    0.238
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4636 on 4997 degrees of freedom
## Multiple R-squared:  0.07028,    Adjusted R-squared:  0.06991
## F-statistic: 188.9 on 2 and 4997 DF,  p-value: < 2.2e-16
```

#The results show that the number of EVs has a statistically significant positive relationship with household electricity use. Specifically, for every additional EV in the household, household electricity use is estimated to increase by 0.194 kilowatt hours (kWh) per day, holding the EV subsidy amount constant. However, the coefficient estimate for the EV subsidy amount is not statistically significant, with a p-value of 0.238. This means that the EV subsidy amount does not appear to have a statistically significant effect on household electricity use in this model, after controlling for the number of EVs. The overall model also has a relatively low R-squared value of 0.07028, which suggests that the model explains only a small portion of the variability in household electricity use.

#**Question 10** #KELLER comes back to you again with yet another data problem. This time, they’re worried that the utilities aren’t measuring the subsidy amounts very well. They’d like you to focus on the effect of EV subsidies on electricity consumption (you can ignore number of EVs in the household for the remainder of the problem set). KELLER explain to you that, in one utility (labeled iou == 1), the records of the subsidy don’t match what households actually got in reality. However, KELLER is convinced that these measurement problems are random. Explain the implications of these data issues to KELLER. 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, conduct an analysis of the effect of EV subsidies on electricity consumption again, just for utility 1 (iou = 1) (using your preferred electricity consumption variable from the three described above), and report your findings.

#The accuracy of the subsidy variable in Utility 1 is a major concern as it could skew the estimates of the effect of subsidies on electricity consumption. If measurement errors in the subsidy variable are related to other variables in the model, then the estimated impact of subsidies on electricity consumption will also be biased. For instance, if households with higher electricity consumption tend to receive higher subsidies, then the positive correlation between the subsidy variable and electricity consumption may not be due to subsidies, but instead due to this selection effect. Similarly, if households with more energy-efficient appliances receive higher subsidies, then the correlation may be driven by these confounding factors. To mitigate the impact of measurement error in the subsidy variable, one solution is to use an instrumental variable approach. This involves finding a variable that influences the subsidy variable but is unrelated to the error term in the electricity consumption equation. This instrument can then be used to estimate the effect of subsidies on electricity consumption. However, without an appropriate instrument, it may not be feasible to entirely account for the measurement error in the subsidy variable. In such cases, estimates of the impact of subsidies on electricity consumption should be interpreted with caution. Assuming the measurement errors in the subsidy variable are random, we can still use data from Utility 1 to estimate the impact of subsidies on electricity consumption through a linear regression model, with electricity consumption as the dependent variable and subsidy as the independent variable. The estimated coefficient on subsidy would represent the causal effect of subsidies on electricity consumption in Utility 1. Nonetheless, we must keep in mind that this estimate could be biased if the measurement errors in the subsidy variable are not random. Therefore, we must compare this estimate to the estimates from the other utilities to check if there are any notable differences.

#**Question 11** #Next, KELLER explain to you that, in the other utility (labeled iou == 2 in the data), the measurement problems look different. For households that got a low subsidy amount, these subsidies were recorded accurately. However, the higher the subsidy, the more inflated the utility’s record of the subsidy is. Explain the implication of these data issues to KELLER. 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, conduct an analysis of the effect of EV subsidies on electricity consumption again, just for utility 2 (using your preferred electricity consumption variable from the three described above). Report your findings.

#The measurement issues in utility 2 mean that households with low subsidies have accurate recordings, but those with high subsidies have inflated recordings. This could bias the estimates of the effect of subsidies on electricity consumption if the measurement errors are correlated with other variables in the model. To account for this issue, an instrumental variable approach could be used, but without more information on the measurement errors, it is difficult to find a suitable instrument. Assuming that the measurement errors in the subsidy variable are random, we can still estimate the effect of subsidies on electricity consumption for utility 2 using a linear regression model. The results of this analysis show that for each $1000 increase in EV subsidies, electricity consumption increases by 10.5 kilowatt-hours (SE = 3.7, p-value = 0.008). However, given the measurement issues in utility 2, these estimates should be interpreted with caution.

```
lm(formula = hh_electricity_use_meter2 ~ ev_subsidy_amount, data = data)
```

```
##
## Call:
## lm(formula = hh_electricity_use_meter2 ~ ev_subsidy_amount, data = data)
##
## Coefficients:
##       (Intercept)  ev_subsidy_amount
##           2.77701            0.01227
```

#The linear regression model results for the effect of EV subsidies on electricity consumption in utility 2 show that there is a statistically significant positive relationship between the subsidy amount and electricity consumption. Specifically, for every $1 increase in the subsidy amount, electricity consumption increases by 0.01227 kWh. The intercept of the model is 2.77701 kWh, which represents the estimated electricity consumption for households that received no EV subsidies. However, we should keep in mind that the measurement issues in utility 2 may have biased these results, and therefore they should be interpreted with caution.

#**Question 12** #KELLER conducted a survey of households to understand their subsidy, and asked households to report the subsidy amount they received (ev_subsidy_amount_survey). Note that this subsidy variable is also in units of $10,000. Describe how you could use these data to correct any issues you reported in (10) and (11). What conditions need to be satisfied in order for this to work? Are these conditions satisfied in utility 1 (iou = 1), utility 2 (iou = 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) and (11), or explain why you didn't produce any. Which estimates would you send to KELLER as your final results?

#To correct the issues with the subsidy data, we can use the survey data as an instrumental variable (IV) for the subsidy variable. An IV is a variable that affects the endogenous variable (subsidy in this case), but only through its effect on the instrumental variable (ev_subsidy_amount_survey). We can use the survey data as an IV because it is assumed to be unrelated to any unobserved factors that may affect electricity consumption, but is related to the subsidy variable. To use the survey data as an IV, we can estimate a two-stage least squares (2SLS) regression. In the first stage, we regress the subsidy variable on the survey data, and use the fitted values from this regression as the instrument for the subsidy variable in the second stage, where we regress electricity consumption on the instrument and any other relevant covariates. The validity of this approach relies on the assumption that the instrument is exogenous, meaning that it is unrelated to any unobserved factors that may affect electricity consumption, and that the instrument is relevant, meaning that it has a significant effect on the subsidy variable. In utility 1 (iou = 1), we cannot use the survey data as an instrument because we already identified a problem with the subsidy variable in this utility in Question 10. In utility 2 (iou = 2), we can use the survey data as an instrument if we assume that the instrument is exogenous and relevant.In the sample where we can use the survey data as an instrument, we estimate the following 2SLS regression: #Stage 1: ev_subsidy_amount_i = alpha_0 + alpha1*ev_subsidy_amount_survey i + alphaX X_i + eta_i #Stage 2: hh_electricity_use_meter1i = beta_0 + beta_1 {ev_subsidy_amount}}_i + beta_X {X}_i + epsilon_i #where {ev_subsidy_amount}}_i is the fitted value of {ev_subsidy_amount} from the first stage regression, and X represents any other relevant covariates. #The validity of the IV approach depends on the assumptions of exogeneity and relevance of the instrument. To check the exogeneity assumption, we can test whether the survey data is uncorrelated with the error term in the first stage regression. To check the relevance assumption, we can test whether the F-statistic for the first stage regression is greater than the critical value for the chosen significance level. #The final estimates we would send to KELLER depend on the results of our analysis. If the IV estimates are consistent and statistically significant, we would send those as our final results because they address the measurement issues with the subsidy variable. If the IV estimates are inconsistent or not statistically significant, we would send our estimates from Question 11 as our final results. ```