

Problem Set 3

Emanuel Mejía

27/02/2024

Contents

1 Peruvian Recycling	2
1.1 Recycling bin ATE	2
1.2 SMS ATE	2
1.3 What outcomes does a recycling bin affect?	2
1.4 What outcomes does a SMS affect?	3
1.5 Marginal effects	3
1.6 Covariates or confounders?	3
1.7 Bad control or useful subset?	3
1.8 What happens if you remove “has cell phone”?	4
2 Multifactor Experiments	5
2.1 Experiment design?	5
2.2 Baseline for interpretation	5
2.3 Bin without sticker effect	5
2.4 With or without a sticker?	5
2.5 Statistical significantly different with or without a sticker?	5
2.6 Fully saturated?	6
3 Now! Do it with data	7
3.1 Treatment only model	8
3.2 Treatment and pre-treatment values	9
3.3 Add street fixed effects	10
3.4 Test for block fixed effects	12
3.5 Feature (no) cell phone	12
3.6 Add the sms treatment	14
3.7 Reproduce Table 4B, Column (2)	15

4 A Final Practice Problem	17
4.1 Simple treatment effect of Zmapp	17
4.2 Add baseline covariates	17
4.3 Interpret estimates	18
4.4 Add day fourteen temperature	18
4.5 Interpret estimates	19
4.6 Look at temperature	19
4.7 Compare health outcomes	20
4.8 Collaborating with others, Part (1)	21
4.9 Collaborating with others, Part (2)	21
4.10 Collaborating with others, Part (3)	21

1 Peruvian Recycling

1.1 Recycling bin ATE

In Column 3 of Table 4A, what is the estimated ATE of providing a recycling bin on the average weight of recyclables turned in per household per week, during the six-week treatment period? Provide a 95% confidence interval. Provide a short narrative using inline R code, such as r inline_reference.

```
# Use this code chunk to show your code work (if needed)
pe_bin_wei <- 0.187
se_bin_wei <- 0.032

cil_bin_wei <- pe_bin_wei - 2 * se_bin_wei
cih_bin_wei <- pe_bin_wei + 2 * se_bin_wei
```

Answer: ... For this particular setting we have a point estimate of 0.187 kilograms of recyclables turned in per week for people who received any bin compared to those who didn't receive one, with a standard error of 0.032. Therefore an estimated 95% confidence interval would be 2 standard errors away from the point estimate on any direction: [0.123, 0.251]. Since it doesn't contain zero in it there seems to be a positive and significant effect among the recipients of bins (they turn in higher quantities of recyclables).

1.2 SMS ATE

In Column 3 of Table 4A, what is the estimated ATE of sending a text message reminder on the average weight of recyclables turned in per household per week? Provide a 95% confidence interval and provide a short narrative using inline R code.

```
# Use this code chunk to show your code work (if needed)
pe_sms_wei <- -0.024
se_sms_wei <- 0.039

cil_sms_wei <- pe_sms_wei - 2 * se_sms_wei
cih_sms_wei <- pe_sms_wei + 2 * se_sms_wei
```

Answer: ... We have a point estimate of -0.024 kilograms of recyclables turned in per week for people who received any sms compared to those who didn't receive one, with a standard error of 0.039. Therefore an estimated 95% confidence interval would be: [-0.102, 0.054]. Since it contains zero we can't find evidence of a significant effect among the sms recipients.

1.3 What outcomes does a recycling bin affect?

Which outcome measures in Table 4A show statistically significant effects (at the 5% level) of providing a recycling bin? How are you dealing with the issue that there are several different tests that have been run, and that you are reading? How, if at all, do the authors deal with this?

Answer: ... Every outcome measure but Average percentage of contamination shows a significant effect. In order to deal with this we could use any kind of multi-comparison correction which the authors' don't seem to use, but since most coefficients are significant, even to a 1% level, perhaps doing this won't turn them insignificant while it would give us more certainty of the validity of the results.

1.4 What outcomes does a SMS affect?

Which outcome measures in Table 4A show statistically significant effects (at the 5% level) of sending text messages? Now that you have read across two different treatments, and many outcomes, what, if anything do the p-values mean to you? Does this feel like p-hacking or doing careful investigation?

Answer: ... None of the outcome measures shows statistically significant effects at the 5% level for sending SMS. And even when this seems like a planned experiment and not just throwing random variables (outcomes, treatments or covariates), there are a lot of them and this could give a similar effect to p-hacking and one relationship could turn significant just by mere chance, although results seem consistent, bin treatment seems significant in almost each experiment while sms doesn't (although this could be captured by the significant *Has Cell Phone* coefficients). So p-values should be treated with care, perhaps using a multi-comparison corrector as said above, and if one result seems particularly interesting, even run another experiment to see if the results replicate.

1.5 Marginal effects

Suppose that, during the two weeks before treatment, household A turns in 2kg per week more recyclables than household B does, and suppose that both households are otherwise identical (including being in the same treatment group). From the model, how much more recycling do we predict household A to have than household B, per week, during the six weeks of treatment? Provide only a point estimate, as the confidence interval would be a bit complicated. This question is designed to test your understanding of slope coefficients in regression.

```
# Use this code chunk to show your code work (if needed)
pe_pre_wei <- 0.281
```

Answer: ... Based on the regression table we would expect in average for house A to return 0.562 kg more per week compared to house B during the next six weeks.

1.6 Covariates or confounders?

Suppose that the variable “percentage of visits turned in bag, baseline” had been left out of the regression reported in Column 1. What would you expect to happen to the results on providing a recycling bin? Would you expect an increase or decrease in the estimated ATE? Would you expect an increase or decrease in the standard error? Explain your reasoning.

```
# Use this code chunk to show your code work (if needed)
```

Answer: ... Since we're running a randomized experiment we shouldn't expect any difference in the estimated ATE for providing a recycling bin when adding/removing any covariate. Although since we're capturing more variation by including it, we could expect that leaving that baseline out would increase our standard error.

1.7 Bad control or useful subset?

In column 1 of Table 4A, would you say the variable “has cell phone” is a bad control? Explain your reasoning, and engage both with the definition of a bad control, and also the implications of including a bad control in a model.

Answer: ... It seems like adding the *Has cell Phone* variable is indeed a bad control. We might think that it is not, because a bad control is said to be adding an outcome as a covariate (so a bad control result is

moved by the experiment as well). And even when it's not properly one of our outcomes, nor the experiment will change people to have more cell phones, because of the way the experiment is defined it doesn't really allow a good randomization of our *Any SMS* treatment, in fact they seem dependent (while a covariate should be the other way around).

1.8 What happens if you remove “has cell phone”?

If we were to remove the “has cell phone” variable from the regression, what would you expect to happen to the coefficient on “Any SMS message”? Would it go up or down? Explain your reasoning.

Use this code chunk to show your code work (if needed)

Answer: ... As we said, since *Has cell phone* doesn't seem independent with *Any SMS*, removing it might move the *Any SMS* coefficient. I would think they're positively correlated, because even when having a cell phone isn't going to ensure you receive a text message we know that not having a cell phone will definitely prevent you from receiving one. So if they're indeed positively correlated, part of this positive coefficient in *Has Cell Phone* could be capturing part of a positive effect on the *Any SMS* variable, and removing the first one would increase the coefficient on the *Any SMS* variable.

2 Multifactor Experiments

2.1 Experiment design?

What is the full experimental design for this experiment? Tell us the dimensions, such as 2x2x3. The full results appear in Panel 4B. We'll note that the dimensions of an experiment are defined in terms of the *treatments that the experiment assigns*, not in terms of other features about the data.

Answer: ... The first variable would be the bin treatment which has 3 possible values namely regular bin, bin with sticker and no bin (control). On the other hand we have another treatment, SMS, which also takes 3 possible assignments namely generic SMS, personal SMS and no SMS (control). These two treatments are applied to 5 measured outcomes (percentage turned in, number of bins turned, avg weight, avg value and avg contamination percentage). Therefore we have a 3x3x5 experiment.

2.2 Baseline for interpretation

In the results of Table 4B, describe the baseline category. That is, in English, how would you describe the attributes of the group of people for whom all dummy variables are equal to zero?

Answer: ... The baseline category would be the ones who didn't receive any bin nor any message text (and have no cell phone by the way).

2.3 Bin without sticker effect

In column (1) of Table 4B, interpret the magnitude of the coefficient on "bin without sticker." What does it mean?

Answer: ... There is a coefficient of 0.035 (significant at the 5% level) for bins without a sticker, which means that the group that received a bin without a sticker turned in recyclables 0.035 more times (or 3.5% since we're talking about a percentage), compared to the baseline group (the ones who didn't receive any bin at all)

2.4 With or without a sticker?

In column (1) of Table 4B, which seems to have a stronger treatment effect, the recycling bin with message sticker, or the recycling bin without sticker? How large is the magnitude of the estimated difference?

Answer: ... It seems like the ones who received a bin with a sticker have a stronger treatment effect ($\widehat{ATE}_s = 0.055$), by about 0.02 more than the ones who received a bin without a sticker ($\widehat{ATE}_g = 0.035$).

2.5 Statistical significantly different with or without a sticker?

Is this difference you just described statistically significant? Explain which piece of information in the table allows you to answer this question.

Answer: ... The difference doesn't seem statistically significant and, in order to answer this question, we'll be using the standard errors as follows:

- The ones who received a bin with a sticker had a point estimate $\widehat{ATE}_s = 0.055$ with a standard error of $\hat{\sigma}_s = 0.015$. Therefore a 95% confidence interval would be [0.025, 0.085].

- On the other hand, the ones who received a bin without a sticker had a point estimate $\widehat{ATE}_g = 0.035$ with a standard error of $\hat{\sigma}_g = 0.015$. Therefore a 95% confidence interval would be $[0.005, 0.065]$. Both intervals contain the other treatment's point estimate, so we cannot conclude there's a statistically significant difference between them.

2.6 Fully saturated?

Notice that Table 4C is described as results from “fully saturated” models. What does this mean? What does David Reiley propose this definition means to him in the async lecture? What do the authors seem to think it means to them? Looking at the list of variables in the table, explain in what sense the model is “saturated.”

Answer: ... A fully saturated model means that each categorical variable is decomposed into each of its categories through individual dummy variables, and every interaction between these is also decomposed in an interaction term per each category. In this case the authors are using only every possible interaction between these but no individual dummy variables so they're capturing the effect of every interaction but not the common effect of an individual dummy variable, so in this way it would be really hard to determine if a specific value has an interesting effect by itself, and every interaction coefficient would be hard to interpret as well, because you'd be comparing the effect vs having all categories in 0.

3 Now! Do it with data

```
# Check type of data in each column
sapply(d, function(x) typeof(x))

##          street      havecell avg_bins_treat base_avg_bins_treat
## "double"      "double"      "double"           "double"
##          bin         sms      bin_s             bin_g
## "double"      "double"      "double"           "double"
##         sms_p       sms_g      nocell
## "double"      "double"      "double"

# Check for NAs
sapply(d, function(x) sum(is.na(x)))

##          street      havecell avg_bins_treat base_avg_bins_treat
##            3            1            0                  0
##          bin         sms      bin_s             bin_g
##            0            0            0                  0
##         sms_p       sms_g      nocell
##            0            0            1

# Check minimum values
sapply(na.omit(d), function(x) min(x))

##          street      havecell avg_bins_treat base_avg_bins_treat
##        -999            0            0                  0
##          bin         sms      bin_s             bin_g
##            0            0            0                  0
##         sms_p       sms_g      nocell
##            0            0            0

# Check maximum values
sapply(na.omit(d), function(x) max(x))

##          street      havecell avg_bins_treat base_avg_bins_treat
## 263.000000      1.000000      4.166667      6.375000
##          bin         sms      bin_s             bin_g
## 1.000000      1.000000      1.000000      1.000000
##         sms_p       sms_g      nocell
## 1.000000      1.000000      1.000000

# Check how many unique values each column has
sapply(na.omit(d), function(x) length(unique(x)))

##          street      havecell avg_bins_treat base_avg_bins_treat
##          180            2            67                  36
##          bin         sms      bin_s             bin_g
##            2            2            2                  2
##         sms_p       sms_g      nocell
##            2            2            2
```

By taking a quick look at the data we see that the type of each column is double, so perhaps we'll need to turn some of them into a factor.

We also have 3 N/A in the street column and 1 in the have cell column. We won't drop these for the moment since we don't know whether we'll need them to reproduce the exact results we have in the authors' table.

Finally about the values each variable takes, all of the binary ones seem to take only values of 0 or 1 which is correct. The street variable might take values of -999 which once again won't be dropped since we don't know if this value is for denoting a special street case.

Besides this it's also worth noticing that the maximum for average bins is higher in the baseline period (before treatment) than during the actual experiment.

3.1 Treatment only model

A. For simplicity, let's start by measuring the effect of providing a recycling bin, ignoring the SMS message treatment (and ignoring whether there was a sticker on the bin or not). Run a regression of Y on only the bin treatment dummy, so you estimate a simple difference in means. Provide a 95% confidence interval for the treatment effect, using **of course** robust standard errors (use these throughout) and provide a brief narrative using R inline statements.

```
mod_basic <- lm(avg_bins_treat ~ bin, data = d)

# Robust Standard Errors
mod_basic$vcovHC_ <- vcovHC(mod_basic)
rob_SE_bas <- sqrt(diag(mod_basic$vcovHC_))

# Confidence Interval
basic_robust_ci <- coefci(mod_basic,
                           parm = "bin",
                           vcov. = vcovHC(mod_basic),
                           level = 0.95
                           )

# Regression Table
stargazer(
  mod_basic,
  type = 'latex',
  se=list(rob_SE_bas),
  header=F,
  omit = 'street',
  title="Basic Model",
  dep.var.labels = "Avg. Bins per Week",
  column.labels = c("Basic"),
  order="Constant",
  covariate.labels = c("(Intercept)","Any Bin")
)
```

Narrative: ... By using this basic model we find that the coefficient for Any Bin is significant at the 0.01 level and its confidence interval, using robust standard errors, is (0.0945, 0.1762), which doesn't include 0 so we can also reject the null of no effect.

Table 1: Basic Model

<i>Dependent variable:</i>	
	Avg. Bins per Week
	Basic
(Intercept)	0.635*** (0.011)
Any Bin	0.135*** (0.021)
Observations	1,785
R ²	0.024
Adjusted R ²	0.024
Residual Std. Error	0.405 (df = 1783)
F Statistic	44.516*** (df = 1; 1783)

Note: *p<0.1; **p<0.05; ***p<0.01

3.2 Treatment and pre-treatment values

Now add the pre-treatment value of Y as a covariate. Provide a 95% confidence interval for the treatment effect. Explain how and why this confidence interval differs from the previous one.

```

mod_pretreat <- lm(avg_bins_treat ~ bin + base_avg_bins_treat, data = d)

# Robust Standard Errors
mod_pretreat$vcovHC_ <- vcovHC(mod_pretreat)
rob_SE_pre <- sqrt(diag(mod_pretreat$vcovHC_))

# Confidence Interval
pretreat_robust_ci <- coefci(mod_pretreat,
                                parm = "bin",
                                vcov. = vcovHC(mod_pretreat),
                                level = 0.95
                               )

# Regression Table
stargazer(
  mod_basic, mod_pretreat,
  type = 'latex',
  se=list(rob_SE_bas, rob_SE_pre),
  header=F,
  omit = 'street',
  title="Pre-Treatment Model",
  dep.var.labels = "Avg. Bins per Week",
  column.labels = c("Basic", "Pre-Treatment"),
  order="Constant",
  covariate.labels = c("(Intercept)", "Any Bin",
                      "Baseline - Avg Bins per Week")
)

```

Table 2: Pre-Treatment Model

	<i>Dependent variable:</i>	
	Avg. Bins per Week	
	Basic	Pre-Treatment
	(1)	(2)
(Intercept)	0.635*** (0.011)	0.350*** (0.021)
Any Bin	0.135*** (0.021)	0.125*** (0.017)
Baseline - Avg Bins per Week		0.393*** (0.030)
Observations	1,785	1,785
R ²	0.024	0.342
Adjusted R ²	0.024	0.342
Residual Std. Error	0.405 (df = 1783)	0.333 (df = 1782)
F Statistic	44.516*** (df = 1; 1783)	463.891*** (df = 2; 1782)

Note:

*p<0.1; **p<0.05; ***p<0.01

Answer: ... Once again the coefficient of our treatment variable is significant and while it differed from the previous model, by randomizing this treatment we ensure that this treatment is independent to any covariate, so the difference shouldn't be big, in this case the difference doesn't seem that big like to indicate some bad randomization process. The confidence interval for our bin treatment is (0.091, 0.1584), which is smaller than the one in the basic model (we can also see that the SE is smaller in this model adding pre-treatment baselines).

3.3 Add street fixed effects

Now add the street fixed effects. (You'll need to use the R command `factor()`. You can do this either within the `lm` call, or you can move this factoring up in the data pipeline so that it persists through the rest of your analysis. The only thing we would recommend that you *not* do is to engineer a new, persistent feature at this point.) Provide a 95% confidence interval for the treatment effect and provide a brief narrative using r inline statements.

```
mod_fixed_effects <- lm(
  avg_bins_treat ~ bin +
  base_avg_bins_treat +
  factor(street),
  data = d)

# Robust Standard Errors
mod_fixed_effects$vcovHC_ <- vcovHC(mod_fixed_effects)
rob_SE_fix <- sqrt(diag(mod_fixed_effects$vcovHC_))

# Confidence Interval
fixed_robust_ci <- coefci(mod_fixed_effects,
                           parm = "bin",
```

```

vcov. = vcovHC(mod_fixed_effects),
level = 0.95
)

# Regression Table
stargazer(
  mod_basic, mod_pretreat,
  mod_fixed_effects,
  type = 'latex',
  se=list(rob_SE_bas, rob_SE_pre, rob_SE_fix),
  header=F,
  omit = 'street',
  column.sep.width = "1pt",
  font.size = "small",
  title="Street Fixed Effects Model",
  dep.var.labels = "Avg. Bins per Week",
  column.labels = c("Basic", "Pre-Treatment", "Fixed Effects"),
  order="Constant",
  covariate.labels = c("(Intercept)","Any Bin",
                      "Baseline - Avg Bins per Week")
)

```

Table 3: Street Fixed Effects Model

	Dependent variable:		
	Basic (1)	Avg. Bins per Week Pre-Treatment (2)	Fixed Effects (3)
(Intercept)	0.635*** (0.011)	0.350*** (0.021)	0.368*** (0.035)
Any Bin	0.135*** (0.021)	0.125*** (0.017)	0.114*** (0.019)
Baseline - Avg Bins per Week		0.393*** (0.030)	0.374*** (0.030)
Observations	1,785	1,785	1,782
R ²	0.024	0.342	0.436
Adjusted R ²	0.024	0.342	0.372
Residual Std. Error	0.405 (df = 1783)	0.333 (df = 1782)	0.324 (df = 1600)
F Statistic	44.516*** (df = 1; 1783)	463.891*** (df = 2; 1782)	6.840*** (df = 181; 1600)

Note:

*p<0.1; **p<0.05; ***p<0.01

Narrative: ... Once we add the street level fixed effects the treatment effect reduced again, so perhaps we should be checking the randomization process just to ensure everything is well computed. For this model the confidence interval is (0.0768, 0.1509). Which slides a little bit to the left (closer to 0) and it's range is a bit wider than before, so perhaps within each street there's more variation than expected.

3.4 Test for block fixed effects

Recall that the authors described their experiment as “stratified at the street level,” which is a synonym for blocking by street. Does including these block fixed effects change the standard errors of the estimates *very much*? Conduct the appropriate test for the inclusion of these block fixed effects, and interpret them in the context of the other variables in the regression.

```
test_fixed_effects <- anova(
  mod_fixed_effects,
  test = "F")
test_fixed_effects

## Analysis of Variance Table
##
## Response: avg_bins_treat
##                               Df  Sum Sq Mean Sq F value    Pr(>F)
## bin                      1   7.079   7.079  67.6214 4.052e-16 ***
## base_avg_bins_treat     1  93.250  93.250 890.7404 < 2.2e-16 ***
## factor(street)          179 29.278   0.164   1.5624 9.514e-06 ***
## Residuals                1600 167.502   0.105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: ... By doing this F-test (anova), we’re trying to determine if each variable inclusion is meaningful, and we can see that including the *Baseline (Avg. Bins)* as compared to the model including only the *Any Bin*, has a significant effect on explaining the variation of the data even when we’re adding 1 extra variable. And we can finally see that adding the *Street* (as factor), has also a significant p-value for the F-statistic, meaning we’re indeed explaining more variation of the data, and we should consider using these fixed effects even when using 179 more variables.

3.5 Feature (no) cell phone

Perhaps having a cell phone helps explain the level of recycling behavior. Instead of “has cell phone,” we find it easier to interpret the coefficient if we define the variable ”no cell phone.” Give the R command to define this new variable, which equals one minus the “has cell phone” variable in the authors’ data set. Use “no cell phone” instead of “has cell phone” in subsequent regressions with this dataset.

Now add “no cell phone” as a covariate to the previous regression. Provide a 95% confidence interval for the treatment effect. Explain why this confidence interval does not differ much from the previous one.

```
mod_cellphone <- lm(
  avg_bins_treat ~ bin +
  base_avg_bins_treat +
  factor(street) +
  nocell,
  data = d)

# Robust Standard Errors
mod_cellphone$vcovHC_ <- vcovHC(mod_cellphone)
rob_SE_cel <- sqrt(diag(mod_cellphone$vcovHC_))

# Confidence Interval
cell_robust_ci <- coefci(mod_cellphone,
```

```

        parm = "bin",
        vcov. = vcovHC(mod_cellphone),
        level = 0.95
    )

# Regression Table
stargazer(
    mod_basic, mod_pretreat,
    mod_fixed_effects, mod_cellphone,
    type = 'latex',
    se=list(rob_SE_bas, rob_SE_pre, rob_SE_fix, rob_SE_cel),
    header=F,
    omit = 'street',
    omit.stat=c("f"),
    column.sep.width = "1pt",
    font.size = "small",
    title="No Cell Phone Model",
    dep.var.labels = "Avg. Bins per Week",
    column.labels = c("Basic",
                      "Pre-Treatment",
                      "Fixed Effects",
                      "No Cell Phone"),
    order="Constant",
    covariate.labels = c("(Intercept)","Any Bin",
                         "Baseline - Avg Bins per Week",
                         "No Cell Phone")
)

```

Table 4: No Cell Phone Model

	Dependent variable:			
	Basic (1)	Avg. Bins per Week		
		Pre-Treatment (2)	Fixed Effects (3)	No Cell Phone (4)
(Intercept)	0.635*** (0.011)	0.350*** (0.021)	0.368*** (0.035)	0.387*** (0.036)
Any Bin	0.135*** (0.021)	0.125*** (0.017)	0.114*** (0.019)	0.115*** (0.019)
Baseline - Avg Bins per Week		0.393*** (0.030)	0.374*** (0.030)	0.373*** (0.030)
No Cell Phone				-0.050*** (0.018)
Observations	1,785	1,785	1,782	1,781
R ²	0.024	0.342	0.436	0.439
Adjusted R ²	0.024	0.342	0.372	0.375
Residual Std. Error	0.405 (df = 1783)	0.333 (df = 1782)	0.324 (df = 1600)	0.323 (df = 1598)

Note:

*p<0.1; **p<0.05; ***p<0.01

Answer: ... By adding the *No Cell Phone* variable the confidence interval is (0.078, 0.1522) which seems

quite similar to the previous one (the standard error for our bin treatment variable is also the same), meaning this variable, even when it is statistically significant, is not capturing much of the data's variation.

3.6 Add the sms treatment

Now let's add in the SMS treatment. Re-run the previous regression with "any SMS" included. You should get the same results as in Table 4A. Provide a 95% confidence interval for the treatment effect of the recycling bin. Explain why this confidence interval does not differ much from the previous one.

```
mod_sms <- lm(
  avg_bins_treat ~ bin +
  sms + nocell +
  base_avg_bins_treat +
  factor(street),
  data = d)

# Robust Standard Errors
mod_sms$vcovHC_ <- vcovHC(mod_sms)
rob_SE_sms <- sqrt(diag(mod_sms$vcovHC_))

# Confidence Interval
sms_robust_ci <- coefci(mod_sms,
                           parm = "bin",
                           vcov. = vcovHC(mod_sms),
                           level = 0.95
                           )

# Regression Table
stargazer(
  mod_sms, mod_sms,
  type = 'latex',
  se=list(NULL, rob_SE_sms),
  header=F,
  omit = 'street',
  no.space=TRUE,
  column.sep.width = "1pt",
  font.size = "small",
  title="SMS Model",
  dep.var.labels = "Avg. Bins per Week",
  column.labels = c("Regular SE", "Robust SE"),
  order="Constant",
  covariate.labels = c("(Intercept)","Any Bin",
                      "Any SMS", "No Cell Phone",
                      "Baseline")
)
```

Answer: ... Now we add the *SMS* treatment variable. For this specific model we're computing the Regular Standard errors (left column) to reproduce the authors' table, but we're also using robust standard errors (right column) which we also use to compute the confidence interval is (0.0779, 0.1522) which is quite similar to the previous one (again), meaning that there's also not much data variation captured by this variable SMS, but in this case the coefficient is not even significant, because the effect of this variable could be already captured by the *No Cell Phone* variable as we expected (bad control).

Table 5: SMS Model

	<i>Dependent variable:</i>	
	Avg. Bins per Week	
	Regular SE	Robust SE
	(1)	(2)
(Intercept)	0.385*** (0.034)	0.385*** (0.038)
Any Bin	0.115*** (0.017)	0.115*** (0.019)
Any SMS	0.005 (0.021)	0.005 (0.024)
No Cell Phone	-0.047** (0.020)	-0.047** (0.023)
Baseline	0.373*** (0.014)	0.373*** (0.030)
Observations	1,781	1,781
R ²	0.439	0.439
Adjusted R ²	0.375	0.375
Residual Std. Error (df = 1597)	0.323	0.323
F Statistic (df = 183; 1597)	6.834***	6.834***

Note:

*p<0.1; **p<0.05; ***p<0.01

3.7 Reproduce Table 4B, Column (2)

Now reproduce the results of column 2 in Table 4B, estimating separate treatment effects for the two types of SMS treatments and the two types of recycling-bin treatments. Provide a 95% confidence interval for the effect of the unadorned recycling bin. Explain how your answer differs from that in question 3.6, and explain why you think it differs.

```

mod_full <- lm(
  avg_bins_treat ~ bin_s +
  bin_g +
  sms_p +
  sms_g +
  nocell +
  base_avg_bins_treat +
  factor(street),
  data = d)

mod_full$vcovHC_ <- vcovHC(mod_full)
rob_SE_full <- sqrt(diag(mod_full$vcovHC_))

# Confidence Interval
full_robust_ci <- coefci(mod_full,
                           parm = "bin_g",
                           vcov. = vcovHC(mod_full),
                           level = 0.95
                           )

# Regression Table
stargazer(
  mod_full, mod_full,
  
```

```

type = 'latex',
se=list(NULL, rob_SE_full),
header=F,
no.space=TRUE,
column.sep.width = "1pt",
omit = 'street',
title="Full Model",
font.size = "small",
dep.var.labels = "Avg. Bins per Week",
column.labels = c("Regular SE", "Robust SE"),
order="Constant",
covariate.labels = c("(Intercept)","Bin with Sticker",
                     "Bin without Sticker", "SMS - Personal",
                     "SMS - Generic", "No Cell Phone",
                     "Baseline - Avg Bins per Week")
)

```

Table 6: Full Model

	<i>Dependent variable:</i>	
	Avg. Bins per Week	
	Regular SE	Robust SE
	(1)	(2)
(Intercept)	0.385*** (0.034)	0.385*** (0.038)
Bin with Sticker	0.128*** (0.022)	0.128*** (0.024)
Bin without Sticker	0.103*** (0.022)	0.103*** (0.025)
SMS - Personal	-0.008 (0.025)	-0.008 (0.028)
SMS - Generic	0.020 (0.025)	0.020 (0.028)
No Cell Phone	-0.046** (0.020)	-0.046** (0.023)
Baseline - Avg Bins per Week	0.374*** (0.014)	0.374*** (0.030)
Observations	1,781	1,781
R ²	0.440	0.440
Adjusted R ²	0.375	0.375
Residual Std. Error (df = 1595)	0.323	0.323
F Statistic (df = 185; 1595)	6.769***	6.769***

Note:

*p<0.1; **p<0.05; ***p<0.01

Answer: ... For this table we use Regular Standard Errors (left column) to reproduce the authors' result, but we also use Robust Standard Errors (right column) to compute more conservative results. We find that both bin treatment versions are significant (although their point estimates suggests a not significant but plausible larger effect). The confidence interval (under robust SE) of the unadorned bin is (0.054, 0.1523) which is different from our previous results because we're not looking at *Any Bin* anymore, and the effect of unadorned bins seems a bit smaller than for the ones with stickers.

4 A Final Practice Problem

4.1 Simple treatment effect of Zmapp

Without using any covariates, answer this question with regression: What is the estimated effect of ZMapp (with standard error in parentheses) on whether someone was dehydrated on day 14? What is the p-value associated with this estimate?

```
zmapp_1 <- lm(dehydrated_day14 ~ treat_zmapp, data = d)

zmapp_1$vcovHC_ <- vcovHC(zmapp_1)
t.zmapp_1 <- coeftest(zmapp_1, vcov. = zmapp_1$vcovHC_)
t.zmapp_1

## 
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.847458   0.047616 17.798   < 2e-16 ***
## treat_zmapp -0.237702   0.091459  -2.599   0.01079 *  
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: ... From looking at this basic model we could interpret that Zmapp treatment reduces chance of being dehydrated by -0.2377(0.0915). This result seems significant at a 5% level (almost at a 1% level) with a p-value $p = 0.0108$

4.2 Add baseline covariates

Add covariates for dehydration on day 0 and patient temperature on day 0 to the regression from part (a) and report the ATE (with standard error). Also report the p-value.

```
zmapp_2 <- lm(dehydrated_day14 ~ treat_zmapp +
                 dehydrated_day0+
                 temperature_day0,
                 data = d)

zmapp_2$vcovHC_ <- vcovHC(zmapp_2)
t.zmapp_2 <- coeftest(zmapp_2, vcov. = zmapp_2$vcovHC_)
t.zmapp_2

## 
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -19.469655   7.607812 -2.5592 0.012054 *  
## treat_zmapp    -0.165537   0.081976 -2.0193 0.046242 *  
## dehydrated_day0   0.064557   0.178032  0.3626 0.717689    
## temperature_day0  0.205548   0.078060  2.6332 0.009859 ** 
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: ... By adding Day-0 covariates, the treatment effect $-0.1655(0.082)$ seems less efficient. And, even when it is significant at a 5% level with a p-value $p = 0.0462$, it is not as significant anymore.

4.3 Interpret estimates

Do you prefer the estimate of the ATE reported in the chunk called `dehydration model` or `add pre-treatment measures`? Why? Report the results of the F-test that you used to form this opinion.

```
zmapp_test_object <- anova(zmapp_1, zmapp_2, test = "F")
zmapp_test_object
```

```
## Analysis of Variance Table
##
## Model 1: dehydrated_day14 ~ treat_zmapp
## Model 2: dehydrated_day14 ~ treat_zmapp + dehydrated_day0 + temperature_day0
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1     98 17.383
## 2     96 12.918  2    4.4653 16.592 6.472e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: ... By adding day-0 covariates our treatment effect seems lower and less significant, although I would prefer the result after adding them. On one side, they're both pre-treatment covariates so they don't seem like bad controls, on the other hand when running an F-Test it seems that adding them is indeed explaining more variance since the p-value of the F-Test is highly significant. It remains to be seen whether our randomization is good enough because the coefficient of our treatment perhaps shouldn't move that much when we add these covariates.

4.4 Add day fourteen temperature

The regression from part `add pre-treatment measures` suggests that temperature is highly predictive of dehydration. Add, temperature on day 14 as a covariate and report the ATE, the standard error, and the p-value.

```
zmapp_3 <- lm(dehydrated_day14 ~ treat_zmapp +
                 dehydrated_day0 +
                 temperature_day0 +
                 temperature_day14,
                 data = d)

zmapp_3$vcovHC_ <- vcovHC(zmapp_3)
t.zmapp_3 <- coeftest(zmapp_3, vcov. = zmapp_3$vcovHC_)
t.zmapp_3

##
## t test of coefficients:
##
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -22.591585   7.746036 -2.9165 0.004416 **
## treat_zmapp          -0.120101   0.085798 -1.3998 0.164829
## dehydrated_day0       0.046038   0.173177  0.2658 0.790934
```

```

## temperature_day0    0.176642   0.077024  2.2933 0.024034 *
## temperature_day14   0.060148   0.025831  2.3286 0.022002 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Answer: ... By adding Day-14 Temperature, the treatment effect seems even less efficient -0.1201(0.0858). And with a p-value $p = 0.1648$, it is not statistically significant anymore.

4.5 Interpret estimates

Do you prefer the estimate of the ATE reported in part `add pre-treatment measures` or `add day 14 temperature`? What is this preference based on?

Answer: ... Besides the fact of not being statistically significant, I wouldn't use this last estimated ATE, because adding Day-14 temperature seems like a bad control by definition, since this "covariate" resembles an outcome (its result could be driven by the experiment), and therefore we shouldn't include it.

4.6 Look at temperature

Now let's switch from the outcome of dehydration to the outcome of temperature, and use the same regression covariates as in the chunk titled `add pre-treatment measures`. Test the hypothesis that ZMapp is especially likely to reduce mens' temperatures, as compared to womens', and describe how you did so. What do the results suggest?

```

zmapp_4 <- lm(temperature_day14 ~ treat_zmapp +
               male + male * treat_zmapp +
               dehydrated_day0 +
               temperature_day0,
               data = d)

zmapp_4$vcovHC_ <- vcovHC(zmapp_4)
t.zmapp_4 <- coeftest(zmapp_4, vcov. = zmapp_4$vcovHC_)
t.zmapp_4

```

```

##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 48.712690  10.194000  4.7786 6.499e-06 ***
## treat_zmapp -0.230866   0.118272 -1.9520  0.05391 .
## male         3.085486   0.121773 25.3379 < 2.2e-16 ***
## dehydrated_day0 0.041131   0.194539  0.2114  0.83301
## temperature_day0 0.504797   0.104511  4.8301 5.287e-06 ***
## treat_zmapp:male -2.076686   0.198386 -10.4679 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Answer: ... To test whether men are more responsive to treatment than women we'll be computing the interaction term between Zmapp treatment and being male. By looking at the results it seems like treatment decreases women's temperature by -0.2309(0.1183) and it is significant at a 10% level ($p = 0.0539$). But when looking at the interaction, treatment effect decreases men -2.0767(0.1984) more, as compared to women and statistically significant ($p = 1.8693698 \times 10^{-17}$), giving a total estimated effect of -2.3076 on men.

4.7 Compare health outcomes

Which group – those that are coded as `male == 0` or `male == 1` have better health outcomes (temperature) in control? What about in treatment? How does this help to contextualize whatever heterogeneous treatment effect you might have estimated?

```
mod_control <- d[treat_zmapp == 0,
                  lm(temperature_day14 ~ male + dehydrated_day0 + temperature_day0)]

mod_control$vcovHC_ <- vcovHC(mod_control)
t.mod_control <- coeftest(mod_control, vcov. = mod_control$vcovHC_)
rob_SE_ctr <- sqrt(diag(mod_control$vcovHC_))

mod_treat <- d[treat_zmapp == 1,
                 lm(temperature_day14 ~ male + dehydrated_day0 + temperature_day0)]

mod_treat$vcovHC_ <- vcovHC(mod_treat)
t.mod_treat <- coeftest(mod_treat, vcov. = mod_treat$vcovHC_)
rob_SE_tre <- sqrt(diag(mod_treat$vcovHC_))

stargazer(
  mod_control, mod_treat,
  type = 'latex', header=F,
  font.size = "small", no.space=TRUE,
  column.sep.width = "1pt",
  se = list(rob_SE_ctr, rob_SE_tre),
  title="Control VS Treatment by Gender",
  column.labels = c("Control", "Treat")
)
```

Table 7: Control VS Treatment by Gender

	<i>Dependent variable:</i>	
	temperature_day14	
	Control	Treat
	(1)	(2)
male	3.087*** (0.124)	1.010*** (0.170)
dehydrated_day0	0.035 (0.246)	0.042 (0.350)
temperature_day0	0.500*** (0.131)	0.514*** (0.190)
Constant	49.193*** (12.779)	47.534** (18.552)
Observations	59	41
R ²	0.927	0.704
Adjusted R ²	0.923	0.680
Residual Std. Error	0.455 (df = 55)	0.459 (df = 37)
F Statistic	232.688*** (df = 3; 55)	29.370*** (df = 3; 37)

Note:

*p<0.1; **p<0.05; *** p<0.01

Answer: ... To give an answer to this question we'll first test whether being male has an effect on temperature of the control group (Table 7 left column), and we find that men are expected to have a highly

significant difference ($p = 1.436082 \times 10^{-31}$) of 3.0873(0.1243). And receiving Zmapp treatment (Table 7 right column) still holds a significant difference ($p = 7.4314288 \times 10^{-7}$) of 1.0102(0.1699) in temperature of men as compared to women. These results, in addition to the previous answer, let us know that there's a heterogeneous treatment effect by gender, and even when men's health receives more benefits from the Zmapp treatment I would still say that women seem healthier (because their overall temperature is lower in either group, even when after treatment the difference is smaller).

4.8 Collaborating with others, Part (1)

Suppose you speak with a colleague to learn about heterogeneous treatment effects.

This colleague has access to a non-anonymized version of the same dataset and reports that they looked at heterogeneous effects of the ZMapp treatment by each of 80 different covariates to examine whether each predicted the effectiveness of ZMapp on each of 20 different indicators of health.

Across these regressions your colleague ran, the treatment's interaction with sex on the outcome of temperature is the only heterogeneous treatment effect that he found to be statistically significant. They reason that this shows the importance of sex for understanding the effectiveness of the drug, because nothing else seemed to indicate why it worked. Bolstering your colleague's confidence, after looking at the data, they also returned to their medical textbooks see the whispers of a theory about why ZMapp interacts with processes only present in men to cure.

Another doctor, unfamiliar with the data, hears your colleague's theory and finds it plausible. How likely do you think it is ZMapp works especially well for curing Ebola in men, and why? (This question is conceptual can be answered without performing any computation.)

Answer: ... This seems a lot like a fish expedition to me, and I would be dubious about their results because this is the only HTE effect they've found, and it could be just a matter of chance. In order to have more convincing results we might apply a multi-comparison correction (as Bonferroni correction) and even run another experiment specifically on this effect to test whether the results replicate or not

4.9 Collaborating with others, Part (2)

Suppose that your colleague conducted their research looking at the interaction of 80 covariates with ZMapp, but that you on your own tested this and only this HTE, and discovered a positive result. How, if at all, does your colleague's behavior change the interpretation of your test? Does this seem fair or reasonable?

Answer: ... Since both experiments are using the same data, I would even doubt of my results. Which doesn't seem fair because the experiment we did tested only for that specific HTE. But after looking at their experiment I think we'd still need to follow the previous suggestion (use a multi-comparison correction and even run another experiment to determine if the results are consistent).

4.10 Collaborating with others, Part (3)

Now, imagine that your colleague had not conducted the 80 different regressions. Instead, they tested this heterogeneous treatment effect, and only this heterogeneous treatment effect, of their own accord. Would you be more or less inclined to believe that the heterogeneous treatment effect really exists? Why? Is there a general principle that is guiding your reasoning?

Answer: ... If that was the case I think I would be more inclined to believe that heterogeneous treatment effect really exists. Besides of the hard/numeric reasons: in case we're just running one HTE we don't even need to compensate for extra variables, I would even be more willing to believe it because dedicating resources to this kind of experiment for just one interaction doesn't seem like something random (unlike a fishing expedition), I would even believe there's some knowledge backing the theory that this might work to begin with.