Average Treatment Effects (ATEs), Multi-Factor Experiments (MFEs), and Linear Regressions (LRs)

Ethan Moody

October 2023

Contents

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)

# Define ATE of bin on avg. weight of recyclables (given in paper, p. 20)
bin_weight_ATE        <- 0.187

# Define SE of bin on avg. weight of recyclables (given in paper, p. 20)
bin_weight_SE         <- 0.032

# Define confidence level for 95% CI
bin_weight_CL_val     <- 0.95

# Define critical t-value for 95% CI (approx. z-score given Obs [N] > 1000)
bin_weight_CI_critval <- qnorm((1 + bin_weight_CL_val) / 2)

# Calculate 95% CI around ATE (LB = Lower Bound, UB = Upper Bound)
bin_weight_ATE_CI_LB  <- bin_weight_ATE - (bin_weight_CI_critval * bin_weight_SE)
bin_weight_ATE_CI_UB  <- bin_weight_ATE + (bin_weight_CI_critval * bin_weight_SE)

# Display ATE and 95% CI
paste("Estimated ATE:", bin_weight_ATE)
```

```
## [1] "Estimated ATE: 0.187"
```

```
paste("95% CI:", "[",
      round(bin_weight_ATE_CI_LB, 4), ",",
      round(bin_weight_ATE_CI_UB, 4), "]")
```

```
## [1] "95% CI: [ 0.1243 , 0.2497 ]"
```

Answer: According to Table 4A (see p. 20 of the paper), 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 is **0.187** (measured in kg). Using the standard error of 0.032 also provided in the table, we can estimate a 95% confidence interval around the ATE. This confidence interval can be written as [**0.1243, 0.2497**]. Both the ATE and 95% confidence interval indicate that providing a recycling bin has a positive effect on (i.e., increases) the average weight in kg of recyclables for a household. Based on the table, this effect appears to be statistically significant at the 1% confidence level.

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)

# Define ATE of text/SMS on avg. weight of recyclables (given in paper, p. 20)
sms_weight_ATE        <- -0.024

# Define SE of text/SMS on avg. weight of recyclables (given in paper, p. 20)
```

```r
sms_weight_SE         <- 0.039

# Define confidence level for 95% CI
sms_weight_CL_val     <- 0.95

# Define critical t-value for 95% CI (approx. z-score given Obs [N] > 1000)
sms_weight_CI_critval <- qnorm((1 + sms_weight_CL_val) / 2)

# Calculate 95% CI around ATE (LB = Lower Bound, UB = Upper Bound)
sms_weight_ATE_CI_LB  <- sms_weight_ATE - (sms_weight_CI_critval * sms_weight_SE)
sms_weight_ATE_CI_UB  <- sms_weight_ATE + (sms_weight_CI_critval * sms_weight_SE)

# Display ATE and 95% CI
paste("Estimated ATE:", sms_weight_ATE)
```

```
## [1] "Estimated ATE: -0.024"
```

```r
paste("95% CI:", "[",
  round(sms_weight_ATE_CI_LB, 4), ",",
  round(sms_weight_ATE_CI_UB, 4), "]")
```

```
## [1] "95% CI: [ -0.1004 , 0.0524 ]"
```

Answer: According to Table 4A (see p. 20 of the paper), the estimated ATE of sending a text message reminder (SMS) on the average weight of recyclables turned in per household per week during the six-week treatment period is **-0.024** (measured in kg). Using the standard error of 0.039 also provided in the table, we can estimate a 95% confidence interval around the ATE. This confidence interval can be written as [**-0.1004, 0.0524**]. On the surface, the ATE suggests that sending a text message reminder has a slightly negative effect on (i.e., decreases) the average weight of recyclables in kg for a household. *However, since this result is not statistically significant, we could say that it's plausible to observe this effect simply due to chance, and not because of the influence of a text/SMS message; it's also (arguably) practically insignificant, being such a small number.*

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: In Table 4A, the outcome measures that show a statistically significant effect of providing a recycling bin are (1) "percentage of visits turned in bag" (i.e., the proportion of weeks in which the household had an opportunity to turn in a bag or bin of recyclables and in which they actually did so), (2) "avg. no. of bins turned in per week" (i.e., the average volume of recyclables turned in to the collector, measured by the number of full standard-sized bins given to the collector over the six weeks of post-treatment data collection), (3) "avg. weight (in kg) of recyclables turned in per week" (i.e., the average weight of recyclables given over the six weeks of post-treatment data collection), and (4) "avg. market value of recyclables given per week" (i.e., the average value of the recyclables that were given in terms of prices received by the collector for the items collected over the six weeks of post-treatment data collection). The authors of this study run numerous different tests, which incorporate several different independent/predictor variables and several different dependent/outcome variables, yet they ultimately do not explicitly address the multiple-comparisons problem. As a result, we *might* consider their research to be at risk of drawing conclusions from a "fishing expedition," where significant results MAY arise simply due to chance from the sheer magnitude of different tests run. To avoid this problem, the authors could have either (A) run a Bonferroni or some other type of p-value correction to account for the multiple comparisons they're investigating and to help avoid overstating significance in their results (i.e., by "raising the bar" or threshold for what constitutes a significant difference), and/or (B) considered their findings to be interesting hypotheses that are worth testing individually to check for replicability in a follow-up study. Because they did not do either (A) or (B), the significant effects noted above in (1) through (4) *might* actually be chance noise (unlikely since they all corroborate each other, yet still possible), and it'd be well worth it to perform a second study that evaluates one of the tested treatments individually (e.g., providing bins vs. not providing bins) to see if the authors' results can be reproduced.

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: In Table 4A, there are no outcome measures that show a statistically significant effect of sending texts/SMS messages. This is confirmed on p. 18 of the paper, where the authors comment that "our findings do not indicate a clear impact of SMS reminders" *when important control variables — like whether a given household has a cell phone, or street characteristics — are omitted.* Because the authors are running so many different comparisons (i.e., multiple different treatments, many different outcomes) without doing any correction/adjustment to p-values, there is *potentially* some p-hacking going on with this study — but I wouldn't say that's certainly the case. When a large number of different comparisons are performed, it's possible to see one or more significant effects simply due to chance. However, it's also possible to see several significant effects that essentially "reinforce" each other (i.e., one test looking at a behavior from one angle may reveal significance, and if another test looking at that same behavior from another angle reveals significance, too, there could be a compelling result there). As a result, I'm only putting *some* stock in the p-values I'm seeing in Table 4A. Had the authors raised the threshold required to establish significance and reject their null hypotheses (e.g., considered only p-values of < 0.01 to be significant) and *still* observed significant results, and/or had they run a follow-up experiment to corroborate their results, then I would feel better about taking the asterisked p-values in their paper more seriously. For now, I think they point to an interesting hypothesis that providing a recycling bin *does* impact household recycling behavior (across a number of different dimensions), which — I think — should be further validated by a follow-up study.

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.

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

# Define pre-treatment avg. recycling weight difference between HH A and HH B
baseline_avg_weight_diff   <- 2

# Define coefficient for baseline avg. recyclable weight (given in paper, p. 20)
coef_baseline_avg_weight   <- 0.281

# Calculate prediction for avg. recycling weight difference between HH A and HH B
pred_recycling_weight_diff <- baseline_avg_weight_diff * coef_baseline_avg_weight

# Display predicted point estimate
paste("Predicted Difference in Avg. Recycling Weight (in kg) per Week:",
      pred_recycling_weight_diff)
```

```
## [1] "Predicted Difference in Avg. Recycling Weight (in kg) per Week: 0.562"
```

Answer: Given a baseline difference of 2kg in avg. recyclables per week between household A and household

B, and based on the assumption that both households are otherwise identical to each other, we could use the results from Table 4A to predict that household A will turn in 2kg * 0.281 = **0.562kg** more of recycling per week than household B during the six weeks of treatment.

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)
# See written answer below
```

Answer: If the variable "percentage of visits turned in bag, baseline" had been left out of the regression reported in Column 1 of Table 4A, I'd expect the estimated ATE for providing a recycling bin to *remain the same* and the estimated standard error to *increase*. This variable is a key pre-treatment predictor/covariate of recycling behavior because it measures preexisting recycling tendencies among households (i.e., tendencies that exist prior to the experimental treatment period and that, based on the regression results, appear to influence treatment outcomes). However, we know from the paper that households were equally likely to be randomly assigned to the control and treatment conditions/groups, *regardless of their baseline behaviors*. As a result, assuming the random assignment was executed cleanly, it'd be reasonable to expect that the control and treatment groups would have had a roughly equivalent make-up of households with higher or lower baseline recycling behaviors in the pre-treatment period. So, if this variable were omitted from the regression, we really shouldn't expect to see any impact to the ATE; it'd be the same either way, due to the fact that differences in this pre-treatment predictor/covariate were neutralized (or controlled for) up-front by random assignment. Separately, though, we could expect to see a change in the standard error if this variable were omitted. This omission would force the model to be fit on fewer predictors to explain the variability in outcomes between the control and treatment groups, which would reduce the precision of its estimates and increase the amount of "signal" (effect) flowing into the error term of the linear regression formula. As a result, leaving out this variable would increase the estimated 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: I wouldn't necessarily say the variable "has cell phone" is a bad control. However, this is a tricky question because "has cell phone" is essentially connected with (i.e., not really independent from) the text/SMS message treatment. In other words, the text/SMS message treatment presupposes the presence of a cell phone. A bad control is generally defined as a variable that could be affected by the treatment, or could be considered an outcome variable. Including a bad control in a regression — especially when it's a post-treatment variable — could lead to biased estimates, as the model would attribute some of the "signal" (effect) of the ATE to the bad control variable rather than to the treatment. In this study, I don't see any clear-cut evidence that households changed their possession of a cell phone in response to receiving a recycling bin or text/SMS message; instead, the "has cell phone" variable remained fixed throughout the experiment and was not responsive to or affected by any experimental treatment. For this reason, it doesn't seem to be a bad control on the basis of the definition of a bad control given above. I think the comparison of results between households that did have a cell phone and those that did not is actually, on the whole, useful for understanding the treatment effects in the study, since it allows us to tease out some insights around what having a cell phone means with respect to recycling behaviors. Now, interestingly, the authors note the following on p. 6 of the paper: *"Although the use of cell phones has grown recently in Peru, phones remain relatively expensive and are not affordable to all households. The full impact of the campaign was only experienced by households that both owned a cell phone and were willing to share their phone number because only these households received the reinforcement text messages prior to the marketing agent visit. As shown* *in supplemental table S1.1, cell phone owners in the participation study were slightly richer, more educated, and more interested in local affairs (especially recycling matters) than non-cell phone owners."* I think this statement suggests that mere possession of a cell phone isn't necessarily what drives recycling behaviors to be different for some households than others; instead, it's possible that those households who have cell phones are more active recyclers because they're "slightly richer, more educated, and more interested in local affairs (especially recycling matters)" than those households who don't have cell phones. That is to say, other characteristics/attributes of these households may underlie both why they have cell phones and why they engage in more pro-recycling behaviors. So, including "has cell phone" is useful for seeing differences in the recycling behaviors between groups in this study, but — based on the authors' admission — the *mechanism* by which those differences come about could be related to factors/variables other than cell phone ownership (such as higher interest in local affairs, or even an unmeasured factor underlying that higher interest).

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)
# See written answer below
```

Answer: If we were to remove "has cell phone" from the regression reported in Column 1 of Table 4A, I would expect an *increase* in the coefficient on "Any SMS message." This is because some of the "signal" (or difference) in observed recycling outcomes between the control and treatment groups is being picked up through the "has cell phone" variable, and it would be attributed to remaining predictors — i.e., those that are left in scope, like the "Any SMS message" variable — if "has cell phone" was omitted. As noted in my prior answer (above), the authors mentioned that households with a cell phone tended to be richer, more educated, and more interested in local affairs — especially recycling matters — than households without a cell phone, which suggests that the possession of a cell phone could be associated with greater recycling tendencies. If "has cell phone" was removed, these greater tendencies to recycle — and, more specifically, the resulting effect of these tendencies on measured recycling outcomes — would be attributed to the "Any SMS message" variable, since only those households that already possessed a cell phone at the time of the experiment actually received any text/SMS messages during the study.

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 full experimental design for this experiment (specifically, the "Participation Intensity Study" referred to in question 1 of this problem set) is a little nuanced. On the surface, it appears to be a *3x3 multifactor between-subjects experiment* (where comparisons are drawn between control/treatment groups and also across pre-/post-measures of recycling behaviors). However, there's also something *extra* going on with the "has cell phone" variable, as described below. The study outlines three possibilities for the first treatment condition (recycling bins): (1) receiving a bin with an informative sticker, (2) receiving a bin without an informative sticker, or (3) receiving no bin. It also outlines three possibilities for the second treatment condition (text/SMS messages): (1) receiving a generic SMS reminder message, (2) receiving a personalized SMS reminder message, or (3) receiving no SMS reminder message. That initially gives us solid grounds to consider the design a 3x3, with 9 possible experimental groups. But wait, there's more! The authors mention that they evaluated results across households having a cell phone and households not having a cell phone (with roughly 50% of the total household sample providing valid cell phone numbers). Since the possession of a cell phone is not technically an experimental *treatment variable* (though it's certainly logically connected to the text/SMS message treatment variable), it's not counted in that 3x3 design descriptor above. However, as shown in Table 4C on p. 23, the authors' proposed "fully saturated" model lists out a few additional experimental groups/conditions that are formed by a portion of participating households having no cell phone. This leads me to believe they conceived of more distinct experimental groups/conditions in the study than just 9 (3x3) and that we should, too. If we account for the "has cell phone" variable then, we might say the "best" representation/descriptor of the full design is a **3x3 between-subjects experiment for households with a cell phone** and a **3x1 (3-level) between-subjects experiment for households without a cell phone** (which gives us something like 12 distinct experimental groups in total). Tricky!

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 group of people for whom all dummy (indicator) variables are equal to zero could be described as a sort of "pure" control group (across all treatment conditions). This "baseline category" represents *households that do not receive a recycling bin, do not receive a text/SMS message, and also do not have a cell phone.*

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: In Column 1 of Table 4B, the coefficient on "Bin without sticker" is 0.035. This coefficient means that when a household received a recycling bin without a sticker, the percentage of visits in which it turned in a bag was — on average — 3.5% higher than the baseline category/group.

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: Based on Column 1 of Table 4B, the recycling bin with a sticker appears to have a stronger treatment effect than the recycling bin without a sticker. The coefficient on "Bin with sticker" is 0.055, while the coefficient on “Bin without sticker” is 0.035. The magnitude of the estimated difference between these two conditions is 0.055 - 0.035 = 0.020, or 2.0% points.

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: No, this difference is *not* statistically significant. The F-test p-value for (1) = (2) toward the bottom of Table 4B shows the result of an F-test between the coefficients for these two variables (“Bin with sticker” and “Bin without sticker”). Because the resulting p-value is 0.31 (which is not less than the customary significance threshold of 0.05), we can conclude that the coefficients are statistically the same and that the difference between them may be simply due to random chance.

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: According to David Reiley’s proposed definition in the async lecture, a “fully saturated” model is one in which each value of a covariate gets a different dummy variable to represent it (allowing the model to be fully non-linear). When multiple covariates are included, a “fully saturated” model would need to include all possible interactions between the dummy variables representing each value of each covariate. In a footnote at the bottom of Table 4C on p. 23 of the paper, the authors state that the results shown in the table represent a “fully saturated” model, which they suggest is a model “with indicators for each unique combination of *treatments*” (*emphasis added*). This is an interesting — albeit slightly subtle — departure from Reiley’s definition; the authors do *not* make any mention of indicators for covariate values, but only mention indicators for treatment values. Looking at the list of variables in the table, the model they propose appears to be “saturated” in the sense of including every combination of treatment variables: recycling bins (with sticker, without sticker, or none), SMS messages (generic message, personal message, or none), and cell phone possession (yes or no). However, their model does not have a different dummy variable representing each unique value of each unique covariate included in their analysis — and (arguably) nor should it, since the pre-treatment baseline variables (which could be considered covariates) are continuous variables that would wildly “blow up” the terms in the model if assigned to individual dummy indicators.

3 Now! Do it with data

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.

```
# Model the effect of a recycling bin on recycling outcome (avg. bins per week)
mod_basic             <- d[, lm(
  avg_bins_treat ~ bin)]

# Create model summary output for reference
mod_basic_summary     <- summary(mod_basic)

# Calculate robust standard errors (RSEs)
mod_basic_rses        <- sqrt(diag(vcovHC(mod_basic)))

# Calculate 95% confidence interval with RSEs for estimated effect
mod_basic_95ci_rse_LB <- coefci(
  mod_basic, vcov = vcovHC(mod_basic))[2]
mod_basic_95ci_rse_UB <- coefci(
  mod_basic, vcov = vcovHC(mod_basic))[4]

# Display stargazer table with model results and RSEs
stargazer(
  mod_basic,
  se = list(mod_basic_rses),
  add.lines = list(
    c("Using Robust Standard Errors", "Yes"),
    c("NA Values Removed (No.)", "Yes (4)")),
  type = "latex",
  header = FALSE)
```

```
# Display estimated effect, RSEs, and 95% confidence interval
paste("Estimated bin effect:", round(mod_basic_summary$coefficients[2], 3))
```

```
[1] “Estimated bin effect: 0.133”
```

```
paste("Estimated RSE:", "(", round(mod_basic_rses[2], 3), ")")
```

```
[1] “Estimated RSE: ( 0.021 )”
```

```
paste("Estimated 95% CI:", "[", round(mod_basic_95ci_rse_LB, 3), ",",
      round(mod_basic_95ci_rse_UB, 3), "]")
```

```
[1] “Estimated 95% CI: [ 0.092 , 0.174 ]”
```

Narrative: Our basic model suggests that providing a recycling bin has a positive effect on the average number of bins turned in per week per household. The treatment effect of providing a recycling bin is given by the coefficient for `bin` in the table above (0.133), with the 95% confidence interval around this effect spanning [0.092, 0.174]. The robust standard error (RSE) for this effect is given by the number in parentheses beneath this coefficient (0.021). These results indicate that when a recycling bin is provided to a given household, the average number of recycling bins turned in by that household per week generally increases by somewhere

Table 1:

	Dependent variable:
	avg_bins_treat
bin	0.133***
	(0.021)
Constant	0.636***
	(0.011)
Using Robust Standard Errors	Yes
NA Values Removed (No.)	Yes (4)
Observations	1,781
R^2	0.024
Adjusted R^2	0.023
Residual Std. Error	0.404 (df = 1779)
F Statistic	43.218*** (df = 1; 1779)
Note:	*p<0.1; **p<0.05; ***p<0.01

between 0.092 and 0.174 bins.

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.

```r
# Model the effect of a recycling bin + baseline bins on recycling outcome
mod_pretreat             <- d[, lm(
  avg_bins_treat ~ bin + base_avg_bins_treat)]

# Create model summary output for reference
mod_pretreat_summary     <- summary(mod_pretreat)

# Calculate robust standard errors (RSEs)
mod_pretreat_rses        <- sqrt(diag(vcovHC(mod_pretreat)))

# Calculate 95% confidence interval with RSEs for estimated effect
mod_pretreat_95ci_rse_LB <- coefci(
  mod_pretreat, vcov = vcovHC(mod_pretreat))[2]
mod_pretreat_95ci_rse_UB <- coefci(
  mod_pretreat, vcov = vcovHC(mod_pretreat))[5]

# Display stargazer table with model results and RSEs
stargazer(
  mod_pretreat,
  se = list(mod_pretreat_rses),
  add.lines = list(
    c("Using Robust Standard Errors", "Yes"),
    c("NA Values Removed (No.)", "Yes (4)")),
  type = "latex",
  header = FALSE)
```

Table 2:

	Dependent variable:
	avg_bins_treat
bin	0.124*** (0.017)
base_avg_bins_treat	0.390*** (0.031)
Constant	0.352*** (0.021)
Using Robust Standard Errors	Yes
NA Values Removed (No.)	Yes (4)
Observations	1,781
R^2	0.338
Adjusted R^2	0.337
Residual Std. Error	0.333 (df = 1778)
F Statistic	453.012*** (df = 2; 1778)
Note:	*p<0.1; **p<0.05; ***p<0.01

```r
# Display estimated effect, RSEs, and 95% confidence interval
paste("Estimated bin effect:", round(mod_pretreat_summary$coefficients[2], 3))
```

```
[1] “Estimated bin effect: 0.124”
```

```r
paste("Estimated RSE:", "(", round(mod_pretreat_rses[2], 3), ")")
```

```
[1] “Estimated RSE: ( 0.017 )”
```

```r
paste("Estimated 95% CI:", "[", round(mod_pretreat_95ci_rse_LB, 3), ",",
      round(mod_pretreat_95ci_rse_UB, 3), "]")
```

```
[1] “Estimated 95% CI: [ 0.091 , 0.158 ]”
```

Answer: Adding the pre-treatment value for `base_avg_bins_treat` to our second model reduces our standard error for the treatment effect estimate of providing a recycling bin and tightens the 95% confidence interval around this treatment effect. The new 95% confidence interval around this effect (0.124) spans [0.091, 0.158], which is not as wide of a range as with our first basic model. This suggests that the average number of recycling bins provided per week by a given household pre-treatment is a good predictor of the average number of recycling bins provided per week by that same household post-treatment, and including it in our model increases the precision of our model's predictions/estimates.

3.3 Add state 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.

```r
# Model the effect of a recycling bin + baseline bins + street on recycling outcome
mod_fixed_effects              <- d[, lm(
  avg_bins_treat ~ bin + base_avg_bins_treat + street)]
## NOTE: Factor applied to `street` right after loading in original data

# Create model summary output for reference
mod_fixed_effects_summary      <- summary(mod_fixed_effects)

# Calculate robust standard errors (RSEs)
mod_fixed_effects_rses         <- sqrt(diag(vcovHC(mod_fixed_effects)))

# Calculate 95% confidence interval with RSEs for estimated effect
mod_fixed_effects_95ci_rse_LB <- coefci(
  mod_fixed_effects, vcov = vcovHC(mod_fixed_effects))[2]
mod_fixed_effects_95ci_rse_UB <- coefci(
  mod_fixed_effects, vcov = vcovHC(mod_fixed_effects))[184]

# Display stargazer table with model results and RSEs
stargazer(
  mod_fixed_effects,
  se = list(mod_fixed_effects_rses),
  omit = "street",
  add.lines = list(c("Using Robust Standard Errors", "Yes"),
                   c("Including Street Fixed Effects", "Yes"),
                   c("NA Values Removed (No.)", "Yes (4)")),
  type = "latex",
  header = FALSE)
```

Table 3:

	Dependent variable:
	avg_bins_treat
bin	0.114***
	(0.019)
base_avg_bins_treat	0.374***
	(0.030)
Constant	0.368***
	(0.035)
Using Robust Standard Errors	Yes
Including Street Fixed Effects	Yes
NA Values Removed (No.)	Yes (4)
Observations	1,781
R^2	0.436
Adjusted R^2	0.372
Residual Std. Error	0.324 (df = 1599)
F Statistic	6.833*** (df = 181; 1599)
Note:	*p<0.1; **p<0.05; ***p<0.01

```r
# Display estimated effect, RSEs, and 95% confidence interval
paste("Estimated bin effect:", round(mod_fixed_effects_summary$coefficients[2], 3))
```

[1] "Estimated bin effect: 0.114"

```r
paste("Estimated RSE:", "(", round(mod_fixed_effects_rses[2], 3), ")")
```

[1] "Estimated RSE: (0.019)"

```r
paste("Estimated 95% CI:", "[", round(mod_fixed_effects_95ci_rse_LB, 3), ",",
      round(mod_fixed_effects_95ci_rse_UB, 3), "]")
```

[1] "Estimated 95% CI: [0.077 , 0.151]"

Narrative: This time, adding another variable to our model (in this case, one that represents street fixed effects) increases our standard error for the treatment effect estimate of providing a recycling bin and slightly opens up the 95% confidence interval around this treatment effect. The new 95% confidence interval around this effect (0.114) spans [0.077, 0.151], which is a slightly wider range than what we observed with our first or second models. At the same time, including street fixed effects reduces the estimated treatment effect of providing a recycling bin. These two observations together suggest that (1) including street fixed effects reduces the risk of overestimating the treatment effect, where overestimation could result from the model incorrectly attributing more of the variation in post-treatment recycling outcomes between control/treatment groups to the treatment itself (providing a recycling bin) than to meaningful preexisting differences in recycling behaviors by street/location, and (2) the random assignment and street-level stratification procedures mentioned by the authors were appropriately executed and led to well-balanced, sufficiently randomized control/treatment groups. The higher standard error in this case may be a byproduct of the model trying to amend its "fit" to account for street-level variability in recycling behaviors, which — given the granularity of street data — almost certainly has much more variability than the "aggregate group" of households does.

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.

```r
# Run F-test (ANOVA) between prior two models, w/ fixed effects and w/o fixed effects
test_fixed_effects        <- anova(mod_pretreat, mod_fixed_effects, test = "F")

# Display result of F-test (ANOVA)
test_fixed_effects
```

```
## Analysis of Variance Table
##
## Model 1: avg_bins_treat ~ bin + base_avg_bins_treat
## Model 2: avg_bins_treat ~ bin + base_avg_bins_treat + street
##   Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
## 1   1778 196.78
## 2   1599 167.50 179    29.284 1.5618 9.689e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Extract and print estimated F-statistic from F-test
test_fixed_effects_fstat  <- round(test_fixed_effects$F[2], 4)
```

```r
paste("Estimated F-statistic:", test_fixed_effects_fstat)
```

```
## [1] "Estimated F-statistic: 1.5618"
```

```r
# Extract and print estimated p-value from F-test
test_fixed_effects_pvalue <- round(test_fixed_effects$`Pr(>F)`[2], 9)
paste("Estimated p-value:", test_fixed_effects_pvalue)
```

```
## [1] "Estimated p-value: 9.689e-06"
```

Answer: The inclusion of block fixed effects *does* appear to change our model estimates meaningfully, even though the difference in standard error between our `mod_pretreat` model and `mod_fixed_effects` model doesn't appear to be that large (just using the eyeball test). This could be because we're adding the street fixed effects to the model *after* already including the significantly predictive baseline bin return rate variable (which is picking up on household-level differences in recycling behavior, much like streets would/do). An F-test (ANOVA) between our `mod_pretreat` model and `mod_fixed_effects` model shows an F-statistic of 1.5618 with a highly significant p-value (at the <1% level) of 9.689×10^{-6}. This result suggests that our `mod_fixed_effects` model — with street fixed effects included — explains more of the variability in the average number of bins turned in per week per household post-treatment than our `mod_pretreat` model. The higher standard error for our `mod_fixed_effects` model could be related to the assumption of heteroscedastic errors (which is at-play with RSEs, and potentially more pronounced at the street-/block-level), or collinearity/correlation between street-level fixed effects and other predictors/variables in the model, like the baseline bin return measure.

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.

```r
## rather than letting this feature engineering persist here -- which is
## bad practice because it requires that you keep in mind what code you
## have, and what code you have not run
##
## instead, move this recode up to the top of your data loading in this file
## so that it runs everytime that you load your data

############################################################################
## NOTE: The first code chunk at the top of this .rmd file does this step  ##
############################################################################
```

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.

```r
# Model the effect of a recycling bin + baseline bins + street + no cell on recycling outcome
mod_cellphone               <- d[, lm(
  avg_bins_treat ~ bin + base_avg_bins_treat + street + nocell)]

# Create model summary output for reference
mod_cellphone_summary       <- summary(mod_cellphone)

# Calculate robust standard errors (RSEs)
mod_cellphone_rses          <- sqrt(diag(vcovHC(mod_cellphone)))

# Calculate 95% confidence interval with RSEs for estimated effect
mod_cellphone_95ci_rse_LB <- coefci(
  mod_cellphone, vcov = vcovHC(mod_cellphone))[2]
mod_cellphone_95ci_rse_UB <- coefci(
  mod_cellphone, vcov = vcovHC(mod_cellphone))[185]

# Display stargazer table with model results and RSEs
stargazer(
  mod_cellphone,
  se = list(mod_cellphone_rses),
  omit = "street",
  add.lines = list(c("Using Robust Standard Errors", "Yes"),
                   c("Including Street Fixed Effects", "Yes"),
                   c("NA Values Removed (No.)", "Yes (4)")),
  type = "latex",
  header = FALSE)
```

Table 4:

	Dependent variable:
	avg_bins_treat
bin	0.115***
	(0.019)
base_avg_bins_treat	0.373***
	(0.030)
nocell	−0.050***
	(0.018)
Constant	0.387***
	(0.036)
Using Robust Standard Errors	Yes
Including Street Fixed Effects	Yes
NA Values Removed (No.)	Yes (4)
Observations	1,781
R^2	0.439
Adjusted R^2	0.375
Residual Std. Error	0.323 (df = 1598)
F Statistic	6.875*** (df = 182; 1598)
Note:	*p<0.1; **p<0.05; ***p<0.01

```r
# Display estimated effect, RSEs, and 95% confidence interval
paste("Estimated bin effect:", round(mod_cellphone_summary$coefficients[2], 3))
```

```
[1] “Estimated bin effect: 0.115”
```

```r
paste("Estimated RSE:", "(", round(mod_cellphone_rses[2], 3), ")")
```

```
[1] “Estimated RSE: ( 0.019 )”
```

```r
paste("Estimated 95% CI:", "[", round(mod_cellphone_95ci_rse_LB, 3), ",",
      round(mod_cellphone_95ci_rse_UB, 3), "]")
```

[1] "Estimated 95% CI: [0.078 , 0.152]"

Answer: Adding an indicator for `nocell` to our model minimally impacts both our standard error for the treatment effect estimate of providing a recycling bin and the 95% confidence interval around this treatment effect. The new 95% confidence interval around this effect (0.115) spans [0.078, 0.152], which is about as wide of a range as with our previous model including street fixed effects. This suggests that including an indicator for possession or lack of a cell phone doesn't really change the precision of our model's predictions/estimates for the treatment effect of providing a recycling bin, especially when considered alongside the other variables already included in our model. While it does pull some "signal" (effect) out of the `Constant` (intercept) model term, it doesn't meaningfully impact our standard error estimate or confidence interval because households were randomly assigned to recycling bin experimental groups, making all subsequent cell phone-/text-related variables in this study balanced across bin experimental groups and independent from bin group assignment.

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.

```r
# Model the effect of a recycling bin + baseline bins + street + no cell + sms on recycling outcome
mod_sms              <- d[, lm(
  avg_bins_treat ~ bin + base_avg_bins_treat + street + nocell + sms)]

# Create model summary output for reference
mod_sms_summary      <- summary(mod_sms)

# Calculate robust standard errors (RSEs)
mod_sms_rses         <- sqrt(diag(vcovHC(mod_sms)))

# Calculate 95% confidence interval with RSEs for estimated effect
mod_sms_95ci_rse_LB <- coefci(
  mod_sms, vcov = vcovHC(mod_sms))[2]
mod_sms_95ci_rse_UB <- coefci(
  mod_sms, vcov = vcovHC(mod_sms))[186]

# Display stargazer table with model results and RSEs
stargazer(
  mod_sms,
  se = list(mod_sms_rses),
  omit = "street",
  add.lines = list(c("Using Robust Standard Errors", "Yes"),
                   c("Including Street Fixed Effects", "Yes"),
                   c("NA Values Removed (No.)", "Yes (4)")),
  type = "latex",
  header = FALSE)
```

```r
# Display estimated effect, RSEs, and 95% confidence interval
paste("Estimated bin effect:", round(mod_sms_summary$coefficients[2], 3))
```

[1] "Estimated bin effect: 0.115"

```r
paste("Estimated RSE:", "(", round(mod_sms_rses[2], 3), ")")
```

[1] "Estimated RSE: (0.019)"

```r
paste("Estimated 95% CI:", "[", round(mod_sms_95ci_rse_LB, 3), ",",
      round(mod_sms_95ci_rse_UB, 3), "]")
```

Table 5:

	Dependent variable:
	avg_bins_treat
bin	0.115^{***}
	(0.019)
base_avg_bins_treat	0.373^{***}
	(0.030)
nocell	-0.047^{**}
	(0.023)
sms	0.005
	(0.024)
Constant	0.385^{***}
	(0.038)
Using Robust Standard Errors	Yes
Including Street Fixed Effects	Yes
NA Values Removed (No.)	Yes (4)
Observations	1,781
R^2	0.439
Adjusted R^2	0.375
Residual Std. Error	0.323 (df = 1597)
F Statistic	6.834^{***} (df = 183; 1597)
Note:	*p<0.1; **p<0.05; ***p<0.01

[1] "Estimated 95% CI: [0.078 , 0.152]"

Answer: Similarly, adding an indicator for `sms` (any SMS) to our model minimally impacts both our standard error for the treatment effect estimate of providing a recycling bin and the 95% confidence interval around this treatment effect. The new 95% confidence interval around this effect (0.115) spans [0.078, 0.152], which is virtually identical to what we observed with our previous model including street fixed effects. Just like in the previous scenario, this suggests that including an indicator for text/SMS messages doesn't meaningfully change the precision of our model's predictions/estimates for the treatment effect of providing a recycling bin, especially when considered alongside the other variables already included in our model. Again, like before, it does pull a little bit of "signal" (effect) out of the `Constant` (intercept) model term and `nocell` model term, but it doesn't meaningfully affect our standard error estimate or confidence interval because text/SMS treatments were entirely independent from bin treatments.

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.

```r
# Model the effect of a recycling bin and all other predictors on recycling outcome
mod_full             <- d[, lm(
  avg_bins_treat ~ bin_s + bin_g + sms_p + sms_g + nocell + base_avg_bins_treat + street)]

# Create model summary output for reference
mod_full_summary     <- summary(mod_full)

# Calculate robust standard errors (RSEs)
mod_full_rses        <- sqrt(diag(vcovHC(mod_full)))

# Calculate 95% confidence interval with RSEs for estimated effect
mod_full_95ci_rse_LB <- coefci(
  mod_full, vcov = vcovHC(mod_full))[3]
mod_full_95ci_rse_UB <- coefci(
  mod_full, vcov = vcovHC(mod_full))[189]

# Display stargazer table with model results and RSEs
stargazer(
  mod_full,
  se = list(mod_full_rses),
  omit = "street",
  add.lines = list(c("Using Robust Standard Errors", "Yes"),
                   c("Including Street Fixed Effects", "Yes"),
                   c("NA Values Removed (No.)", "Yes (4)")),
  type = "latex",
  header = FALSE)
```

```r
# Display estimated effect, RSEs, and 95% confidence interval
paste("Estimated bin w/o sticker effect:", round(mod_full_summary$coefficients[3], 3))
```

[1] "Estimated bin w/o sticker effect: 0.103"

```r
paste("Estimated RSE:", "(", round(mod_full_rses[3], 3), ")")
```

[1] "Estimated RSE: (0.025)"

Table 6:

	Dependent variable:
	avg_bins_treat
bin_s	0.128***
	(0.024)
bin_g	0.103***
	(0.025)
sms_p	−0.008
	(0.028)
sms_g	0.020
	(0.028)
nocell	−0.046**
	(0.023)
base_avg_bins_treat	0.374***
	(0.030)
Constant	0.385***
	(0.038)
Using Robust Standard Errors	Yes
Including Street Fixed Effects	Yes
NA Values Removed (No.)	Yes (4)
Observations	1,781
R^2	0.440
Adjusted R^2	0.375
Residual Std. Error	0.323 (df = 1595)
F Statistic	6.769*** (df = 185; 1595)
Note:	*p<0.1; **p<0.05; ***p<0.01

```r
paste("Estimated 95% CI:", "[", round(mod_full_95ci_rse_LB, 3), ",",
      round(mod_full_95ci_rse_UB, 3), "]")
```

```
[1] "Estimated 95% CI: [ 0.054 , 0.152 ]"
```

Answer: Our full model suggests that providing a recycling bin still has a positive effect on the average number of bins turned in per week per household, even when the bin is generic and has no sticker. The treatment effect of providing an "unadorned" recycling bin is given by the coefficient for `bin_g` in the table above (0.103), with the 95% confidence interval around this effect spanning [0.054, 0.152]. The main reason why this confidence interval is different (and wider) than the confidence interval in the previous question is because we've now split the generic `bin` (any bin) variable into separate `bin_s` (bin-with-sticker) and `bin_g` (bin-without-sticker) variables in the full model. This splitting of the treatment variable into two "flavors" of treatment forces the model to fit its predictions on both of these variables separately and impacts prediction errors. Random assignment should limit the extent of differences in variability across street fixed effects and pre-treatment baseline recycling behaviors between treatment groups receiving the `bin_g` vs. `bin_s` bin treatment, though there still could be some differences impacting the treatment effect standard error and the resulting confidence interval, causing a small difference in the interval for this model relative to the previous version.

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?

```r
# Create linear model of ZMapp treatment against day 14 dehydration indicator
zmapp_1 <- d[, lm(
  dehydrated_day14 ~ treat_zmapp)]

# Store model summary results
zmapp_1_summary <- summary(zmapp_1)

# Display result/summary of model
# zmapp_1_summary

# Extract and print estimated effect value from model
zmapp_1_summary_effect <- round(zmapp_1_summary$coefficients[2], 4)
paste("Estimated Effect:", zmapp_1_summary_effect)
```

```
## [1] "Estimated Effect: -0.2377"
```

```r
# Extract and print estimated standard error value from model
zmapp_1_summary_serror <- round(zmapp_1_summary$coefficients[4], 4)
paste("Estimated Standard Error:", zmapp_1_summary_serror)
```

```
## [1] "Estimated Standard Error: 0.0856"
```

```r
# Extract and print estimated p-value from model
zmapp_1_summary_pvalue <- round(zmapp_1_summary$coefficients[8], 4)
paste("Estimated p-value:", zmapp_1_summary_pvalue)
```

```
## [1] "Estimated p-value: 0.0066"
```

Answer: The estimated effect of ZMapp on whether someone was dehydrated on day 14 is -0.2377 (0.0856), with the standard error noted in parentheses. The p-value associated with this estimate is 0.0066, which is significant at the 1% level.

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.

```r
# Expand linear model to include covariates for day 0 dehydration and temperature
zmapp_2 <- d[, lm(
  dehydrated_day14 ~ treat_zmapp + dehydrated_day0 + temperature_day0)]

# Store model summary results
zmapp_2_summary <- summary(zmapp_2)

# Display result/summary of model
# zmapp_2_summary

# Extract and print estimated effect value from model
zmapp_2_summary_effect <- round(zmapp_2_summary$coefficients[2], 4)
paste("Estimated Effect:", zmapp_2_summary_effect)
```

```
## [1] "Estimated Effect: -0.1655"
```

```
# Extract and print estimated standard error value from model
zmapp_2_summary_serror <- round(zmapp_2_summary$coefficients[6], 4)
paste("Estimated Standard Error:", zmapp_2_summary_serror)
```

```
## [1] "Estimated Standard Error: 0.0757"
```

```
# Extract and print estimated p-value from model
zmapp_2_summary_pvalue <- round(zmapp_2_summary$coefficients[14], 4)
paste("Estimated p-value:", zmapp_2_summary_pvalue)
```

```
## [1] "Estimated p-value: 0.0311"
```

Answer: With covariates for dehydration on day 0 and patient temperature on day 0 included in the model, the estimated effect of ZMapp on whether someone was dehydrated on day 14 is -0.1655 (0.0757), with the standard error noted in parentheses. The p-value associated with this estimate is 0.0311, which is significant at the 5% level.

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.

```
# Run ANOVA to perform an F-test between previous two models
zmapp_test_object <- anova(zmapp_1, zmapp_2, test = "F")

# Display result of F-test (ANOVA)
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
```

```
# Extract and print estimated F-statistic from F-test
zmapp_test_object_fstat <- round(zmapp_test_object$F[2], 4)
paste("Estimated F-statistic:", zmapp_test_object_fstat)
```

```
## [1] "Estimated F-statistic: 16.5919"
```

```
# Extract and print estimated p-value from F-test
zmapp_test_object_pvalue <- round(zmapp_test_object$`Pr(>F)`[2], 10)
paste("Estimated p-value:", zmapp_test_object_pvalue)
```

```
## [1] "Estimated p-value: 6.472e-07"
```

Answer: An F-test (ANOVA) between the two linear models produces an F-statistic value of 16.5919 and a highly significant p-value (at more than the 0.1% level) of 6.472×10^{-7}. These results indicate that we should reject the null hypothesis and conclude that there *is* evidence to suggest a significant difference between our alternative model's ability to explain the variance in the outcome variable and our base/original model's ability to do so. Practically, this means that the inclusion of additional predictors/covariates in our alternative model — the second one we produced — ultimately helped improve that model's "fit" (the closeness of its predictions to actual observed values) relative to our original model. In other words, certain pre-treatment (day 0) measures — specifically patient temperature (given the significant p-value observed for this variable's coefficient in our second model) — appear to be highly predictive of post-treatment (day 14) patient dehydration. Based on these results, I prefer the estimate of the ATE reported by our alternative/second model more than our original/first model.

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.

```r
# Expand linear model to include an additional covariate for day 14 temperature
zmapp_3 <- d[, lm(
  dehydrated_day14 ~ treat_zmapp + dehydrated_day0 + temperature_day0 + temperature_day14)]

# Store model summary results
zmapp_3_summary <- summary(zmapp_3)

# Display result/summary of model
# zmapp_3_summary

# Extract and print estimated effect value from model
zmapp_3_summary_effect <- round(zmapp_3_summary$coefficients[2], 4)
paste("Estimated Effect:", zmapp_3_summary_effect)
```

```
## [1] "Estimated Effect: -0.1201"
```

```r
# Extract and print estimated standard error value from model
zmapp_3_summary_serror <- round(zmapp_3_summary$coefficients[7], 4)
paste("Estimated Standard Error:", zmapp_3_summary_serror)
```

```
## [1] "Estimated Standard Error: 0.0777"
```

```r
# Extract and print estimated p-value from model
zmapp_3_summary_pvalue <- round(zmapp_3_summary$coefficients[17], 4)
paste("Estimated p-value:", zmapp_3_summary_pvalue)
```

```
## [1] "Estimated p-value: 0.1254"
```

Answer: Now, with an additional covariate for patient temperature on day 14 included in the model, the estimated effect of ZMapp on whether someone was dehydrated on day 14 is -0.1201 (0.0777), with the standard error noted in parentheses. The p-value associated with this estimate is 0.1254, which is no longer significant at the 5% or even 10% level.

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: I prefer the estimate of the ATE reported in our second model (with only pre-treatment covariates added) over the estimate of the ATE reported in our third model (with a post-treatment covariate added). This is because patient temperature on day 14 could be considered a "bad control" in the third regression model. A bad control is generally defined as a variable that could be affected by the treatment, or could be considered an outcome variable. Including a bad control in a regression — *especially when it's a post-treatment variable* — could lead to biased estimates, as the model might attribute some of the "signal" (effect) of the treatment to the bad control variable. In this hypothetical scenario, it'd be reasonable to believe that patients' temperatures on day 14 are affected by taking/not taking the ZMapp drug *and/or by their dehydration at both day 0 and day 14* (or even vice-versa — it'd be reasonable to believe that patients' dehydration is affected by temperature). Since there are numerous potential causal relationships at play between post-treatment temperature, post-treatment dehydration, and the ZMapp drug, it doesn't actually make sense to treat the post-treatment temperature reading as a predictor of post-treatment dehydration in the regression model.

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?

```
# Create linear model with day 14 temperature outcome and pre-treatment covariates
zmapp_4 <- d[, lm(
  temperature_day14 ~ treat_zmapp + dehydrated_day0 + temperature_day0 + male + treat_zmapp*male)]

# Store model summary results
zmapp_4_summary <- summary(zmapp_4)

# Display result/summary of model
zmapp_4_summary
```

Call: lm(formula = temperature_day14 ~ treat_zmapp + dehydrated_day0 + temperature_day0 + male + treat_zmapp * male)

Residuals: Min 1Q Median 3Q Max -0.70157 -0.37725 -0.02702 0.34687 0.73968

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 48.71269 9.26618 5.257 9.14e-07 ***treat_zmapp -0.23087 0.11871 -1.945 0.0548 .
dehydrated_day0 0.04113 0.18208 0.226 0.8218
temperature_day0 0.50480 0.09508 5.309 7.34e-07*** male 3.08549 0.12644 24.403 < 2e-16
treat_zmapp:male -2.07669 0.19164 -10.836 < 2e-16 — Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ” 0.05 ‘.’
0.1 ’ ’ 1

Residual standard error: 0.4518 on 94 degrees of freedom Multiple R-squared: 0.9059, Adjusted R-squared: 0.9009 F-statistic: 181 on 5 and 94 DF, p-value: < 2.2e-16

```
# Extract and print estimated effect value from model
zmapp_4_summary_effect <- round(zmapp_4_summary$coefficients[6], 4)
paste("Estimated Effect:", zmapp_4_summary_effect)
```

[1] "Estimated Effect: -2.0767"

```
# Extract and print estimated standard error value from model
zmapp_4_summary_serror <- round(zmapp_4_summary$coefficients[12], 4)
paste("Estimated Standard Error:", zmapp_4_summary_serror)
```

[1] "Estimated Standard Error: 0.1916"

```
# Extract and print estimated p-value from model
zmapp_4_summary_pvalue <- round(zmapp_4_summary$coefficients[24], 22)
paste("Estimated p-value:", zmapp_4_summary_pvalue)
```

[1] "Estimated p-value: 3.1092e-18"

```
############################################################################

# Alternate method: Partition data into males/females and create separate linear models
zmapp_4m <- d[male == 1, lm(
  temperature_day14 ~ treat_zmapp + dehydrated_day0 + temperature_day0)]
```

```r
zmapp_4f <- d[male == 0, lm(
  temperature_day14 ~ treat_zmapp + dehydrated_day0 + temperature_day0)]

# Display output of linear models in stargazer table
stargazer(
  zmapp_4m, zmapp_4f,
  add.lines = list(c("Model for Males Subset", "Yes", "No")),
  type = "latex",
  header = FALSE)
```

Table 7:

	Dependent variable:	
	temperature_day14	
	(1)	(2)
treat_zmapp	−2.239***	−0.229*
	(0.154)	(0.119)
dehydrated_day0	−0.379	0.265
	(0.297)	(0.227)
temperature_day0	0.767***	0.374***
	(0.161)	(0.116)
Constant	26.232	61.463***
	(15.698)	(11.341)
Model for Males Subset	Yes	No
Observations	37	63
R^2	0.918	0.554
Adjusted R^2	0.910	0.532
Residual Std. Error	0.440 (df = 33)	0.452 (df = 59)
F Statistic	122.877*** (df = 3; 33)	24.460*** (df = 3; 59)
Note:		*p<0.1; **p<0.05; ***p<0.01

Answer: To test the hypothesis that ZMapp is especially likely to reduce mens' temperatures as compared to womens' temperatures, I compared two scenarios/tests and evaluated the results: (1) I produced a new linear model including the `male` indicator variable along with the same pre-treatment covariates in our second linear model (above) *and* an interaction term (interacting `treat_zmapp` with `male`), and (2) I ran two separate linear models including the same pre-treatment covariates in our second linear model (above) that are identical to each other apart from their partitioning of the data, with one partitioning the data into males only (`male == 1`) and the other partitioning the data into females only (`male == 0`). The first scenario showed a significant p-value of 3.1092×10^{-18} for the interaction term, indicating that being male and receiving the ZMapp treatment significantly impacts patients' post-treatment day 14 temperatures. The second scenario corroborates this finding by showing a *more extreme coefficient* — that's also more statistically significant — for `treat_zmapp` for males than for females, as shown in the stargazer output table. The `treat_zmapp` coefficient is statistically significant only at the 10% level for females (which is generally regarded as *non-significant*) versus significant at the <1% level for males. Given these findings, we could say that the hypothesis that ZMapp is especially likely to reduce mens' temperatures as compared to womens' temperatures *appears to be supported* by both the larger `treat_zmapp` coefficient (in absolute terms) for males than females and the greater statistical significance with this value across the two scenarios/tests above. However, as discussed in the answer to the next question, the difference we see for treated mens' temperatures vs. treated womens' temperatures could be affected by other factors — like their baseline/starting temperatures (which we can infer from the male/female control groups) — rather than just being driven by some sort of special effectiveness of the drug for 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?

```
# Compute mean `temperature_day14` grouped by `male` and `treat_zmapp`
group_means_temp_post <- d[, .(mean_temp14 = mean(temperature_day14)),
                           keyby = .(male, treat_zmapp)]

# Display group means output
# group_means_temp_post

# Compute mean `temperature_day14` for female patients in control group
f_control_mean_temp_post <- d[(male == 0 & treat_zmapp == 0), round(mean(temperature_day14), 4)]

# Compute mean `temperature_day14` for male patients in control group
m_control_mean_temp_post <- d[(male == 1 & treat_zmapp == 0), round(mean(temperature_day14), 4)]

# Compute mean `temperature_day14` for female patients in treatment group
f_treat_mean_temp_post   <- d[(male == 0 & treat_zmapp == 1), round(mean(temperature_day14), 4)]

# Compute mean `temperature_day14` for male patients in treatment group
m_treat_mean_temp_post   <- d[(male == 1 & treat_zmapp == 1), round(mean(temperature_day14), 4)]

# Print `temperature_day14` means for male and female patients in control/treatment groups
paste("Mean day 14 temperature for female patients in control group:", f_control_mean_temp_post)
```

```
## [1] "Mean day 14 temperature for female patients in control group: 98.4865"
```

```
paste("Mean day 14 temperature for male patients in control group:", m_control_mean_temp_post)
```

```
## [1] "Mean day 14 temperature for male patients in control group: 101.6917"
```

```
paste("Mean day 14 temperature for female patients in treatment group:", f_treat_mean_temp_post)
```

```
## [1] "Mean day 14 temperature for female patients in treatment group: 98.1631"
```

```
paste("Mean day 14 temperature for male patients in treatment group:", m_treat_mean_temp_post)
```

```
## [1] "Mean day 14 temperature for male patients in treatment group: 99.1007"
```

Answer: Mean patient temperatures at day 14 for males and females across control and treatment groups are shown in the table below:

male	treat_zmapp	mean_temp14
0	0	98.48654
0	1	98.16308
1	0	101.69167
1	1	99.10071

These values indicate that patients coded as `male == 0` (females) tend to have better day 14 temperatures in the control group. The mean temperature at day 14 for female patients in the control group is 98.4865; meanwhile, the mean temperature at day 14 for male patients in the control group is 101.6917 (pretty different). These values also indicate that patients coded as `male == 0` (females) tend to show much less change in day 14 temperature in the treatment group (on average). The mean temperature at day 14 for female patients in the treatment group is 98.1631 — not substantially different from the mean temperature for female patients in the control group; meanwhile, the mean temperature at day 14 for male patients in the treatment group is 99.1007 — certainly different from the mean temperature for male patients in the control group. Based on these mean differences in temperature across male/female and control/treatment groups, we could conclude one of three things: (1) random assignment of patients to control/treatment groups wasn't executed cleanly, and the pronounced difference in mean temperatures across control and treatment groups for males was due to a procedure that was not entirely random (or due to an unlucky accident); (2) the ZMapp treatment actually *does* produce a larger effect on males' temperatures than females' temperatures; or (3) perhaps most likely, the ZMapp treatment is most effective when administered to patients with fevers (temperatures >101 degrees Fahrenheit) and not really effective when administered to patients who don't have fevers (temperatures around 98-99 degrees Fahrenheit). If (1) were true, we'd probably want to run the RCT/experiment a second time — with a different pool of patients randomly assigned to control/treatment groups — to see if our initial results were reproducible. If (2) were true, the mean differences noted above would provide evidence of a heterogeneous effect and support the conclusion that ZMapp is especially likely to reduce mens' temperatures as compared to womens' temperatures (given the larger difference observed between control/treatment mean temperatures). If (3) were true — which, again, seems most likely given the different mean day 14 temperatures for males and females — the mean differences noted above wouldn't necessarily prove that the heterogeneous treatment effect is because the ZMapp drug is more effective for males than females; instead, it could point to the fact that ZMapp is most effective when administered to *patients in general* who have fevers and less effective when administered to *patients in general* who don't have fevers.

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: On the one hand, the combination of my colleague's post-hoc test results, *and* the results of the data from the actual experiment/RCT performed above, *and* the acknowledgement of plausibility from a doctor with (presumably) strong domain expertise lends credibility to the hypothesis that ZMapp works especially well for curing Ebola in men. Typically, when data and domain knowledge agree, that's a good indicator of a plausible result. However, that's not always the case. It's possible that we could have seen a significant result in the RCT simply due to chance. In fact, with so many tests/checks run by my colleague (i.e., 80 covariates on 20 indicators of health!!!), there's pretty much a statistical guarantee that we'll happen upon *some* test/*some* relationship that's significant, given that we typically consider one-in-every-twenty outcomes (5%) to be significant. This is a situation where we'd need to caution our colleague against p-hacking or engaging in a "fishing expedition" where they're either (A) only reporting the one finding out of many that happened to show significance, or (B) not making a correction to their reported p-values to account for the multiple comparisons they're investigating (i.e., by tightening the threshold for what constitutes a significant difference). I'd invite my colleague to tell me more about how we're avoiding the pitfall of (A) and would also encourage them to consider implementing a p-value correction (like Bonferroni) to their results *or* reproducing the experimental results with a follow-up RCT to avoid the pitfall of (B), before confidently concluding that we do, in fact, see good empirical evidence that ZMapp works especially well for curing Ebola in men.

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: In this scenario, my cautionary remarks on the conclusion that ZMapp works especially well for men would be the same as above. It doesn't ultimately matter if it was my colleague or me who happened upon the "winning" (significant) HTE; if we're analyzing this data and doing the research together, we'd both need to reckon with — and safeguard against — potential pitfalls with the multiple comparisons problem (specifically p-hacking and "fishing expeditions"). Simply put, my colleague's behavior *should* change the interpretation of my own test and make me more hesitant/skeptical to accept this HTE as an indisputable significant result. I'd say this is both fair and reasonable. We're in the pursuit of truth with scientific experimentation, and we shouldn't allow our own preconceived notions or desires for "exciting" results to color our final conclusions, or we may lead interpreters of our work astray. With medical trials, being fast-and-loose with design, methodology, or statistical analysis considerations can be especially dangerous, so I would want to err on the side of caution here and work with my colleague to interpret the results of our study carefully and/or attempt to replicate them — if we really believe them — in a second study.

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: In this scenario, I would be *more inclined* to believe that this HTE really exists. As a general rule, it should be relatively challenging to achieve statistical significance simply due to chance; typically, that'll happen only one out of twenty times. While it's possible that we could have simply stumbled upon a significant HTE here (for instance, perhaps this outcome represents the one-in-twenty time that we encounter significance by chance alone), it's objectively more likely that this significant outcome is real, or at least worth investigating further, *especially* if we were engaging in honest scientific research and specified the HTE as a core hypothesis of our study up-front. We could do this closer inspection and investigation by reproducing the study and seeing if the same result turns up. We could also share our findings with statistics professionals or other researchers as part of a peer-review — and/or with others like the doctor described earlier who have substantial expertise in the domain of interest (e.g., Ebola treatments) — to check for any evidence/explanations to the contrary of what the discovered HTE suggests. If the effect we observed were to hold up under scrutiny and rigorous review, I think we could feel reasonably confident in our outcome. After all, what even is the *full set* of plausible covariates and regression specifications that could be examined/tested for any experimental problem? Is that set even bounded? Can it really be quantified? And should we actually endeavor to test every member of that set, *every possible variable*, within every experiment to try to determine if the one significant HTE we report may have been significant simply due to chance? I think that would be largely unfeasible. Instead, we can follow some of the general practices above to lend credibility to our claims of significant findings. "Mistakes" or "unlucky accidents" can still happen when reporting significant HTEs — just as David Reiley admitted with his own research described in "The Multiple-Comparisons Problem" 7.21 async lecture — but that shouldn't stop well-executed, well-structured, well-thought out research from being published. We wouldn't want to make the threshold for statistical significance or for publishing so high that it's practically unattainable, or the pursuit of truth through science such that we're paralyzed by endless doubts and what-ifs.