Problem Set 5

Christine Sako

11/07/2025

```
library(data.table)

library(sandwich)
library(lmtest)

library(AER)

library(ggplot2)
library(patchwork)
```

Vietnam Draft Lottery

Observational esteimate

Suppose that you had not run an experiment. Estimate the "effect" of each year of education on income as an observational researcher might, by just running a regression of years of education on income (in R-ish, `income ~ years_education`). What does this naive regression suggest?

```
# Naive linear model
model_observational <- lm(income ~ years_education, data = d)

# Extracting coefficients for inline reference
obs_coefs <- coef(model_observational)

summary(model_observational)
```

```
##
## Call:
## lm(formula = income ~ years_education, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -91655 -17459   -837  16346 141587
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -23354.64    1252.74  -18.64   <2e-16 ***
## years_education   5750.48      83.34   69.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26590 on 19565 degrees of freedom
## Multiple R-squared:  0.1957, Adjusted R-squared:  0.1957
## F-statistic:  4761 on 1 and 19565 DF,  p-value: < 2.2e-16
```

Answer: This naive observational regression suggests that each additional year of education is associated with an average increase of about $5,750 in annual income. This estimate is likely to overstate the causal effect of schooling because it does not take into account self-selection into education (i.e. those who would have gone to school anyway are correlated with both education and income) or variables that covary with education and income (e.g. motivation, prior financial status, etc.)

Evaluating observational estimate

Continue to suppose that we did not run the experiment, but that we saw the result that you noted in part 1. Tell a concrete story about why you don't believe that observational result tells you anything causal.

Answer: Suppose there are two different people, Taylor and Avery who have very different backgrounds. Taylor grew up in a wealthy family who always prioritized education and placed Taylor in various education programs and the best schools. Avery grew up less financially fortunate and was not planning to go to college due to not being able to afford it, instead Avery planned to work directly after high school. Let's say Taylor goes to a 4-year college and ends up having 16 total years of education, whereas Avery only completes 12 total years of education. Let's further say Taylor ends up earning $73k a year and Avery earns $50K a year. Without accounting for all the background covariates, a naive regression would take just the years of education and income and assume each year of additional schooling leads to an additional $5,750 in income ($5,750 * 4 = $23,002). The problem with this regression is that it does not consider all of the background information that may have lead to Taylor making more money regardless of years of schooling, and similarly, information that may have indicated that Avery would have earned less even with more years of schooling.

Natural experiment effect on education

Now, let's get to using the natural experiment. Define "having a high-ranked draft number" as having a draft number between 1-80. For the remaining 285 days of the year, consider them having a "low-ranked" draft number). Create a variable in your dataset called `high_draft` that indicates whether each person has a high-ranked draft number or not. Using a regression, estimate the effect of having a high-ranked draft number on years of education obtained. Report the estimate and a correctly computed standard error. (*Hint: How is the assignment to having a draft number conducted? Does random assignment happen at the individual level? Or, at some higher level?)

```r
library(sandwich)
library(lmtest)

# Defining high-ranked draft indicator
d$high_draft <- ifelse(d$draft_number <= 80, 1, 0)

# Regressing years of education on the high-ranked draft indicator
model_education <- lm(years_education ~ high_draft, data = d)
summary(model_education)
```

```
##
## Call:
## lm(formula = years_education ~ high_draft, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5601 -1.4343 -0.4343  1.5657  5.5657
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.43431    0.01691  853.40   <2e-16 ***
## high_draft   2.12576    0.03790   56.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.117 on 19565 degrees of freedom
## Multiple R-squared:  0.1385, Adjusted R-squared:  0.1384
## F-statistic:  3145 on 1 and 19565 DF,  p-value: < 2.2e-16
```

```
# Computing cluster-robust standard errors since assignment happens at the birthday/draft-number level
cluster_se_edu <- sqrt(diag(vcovCL(model_education, cluster = ~draft_number)))

# Extracting coefficients for inline reference
coef_estimate_edu <- coef(model_education)["high_draft"]
se_clustered_edu  <- cluster_se_edu["high_draft"]
```

Answer: Having a high-ranked draft number (1-80) is associated with an increase of approximately 2.1258 years of education, with a cluster-robust standard error of 0.0382. We use cluster-robust standard errors since assignment to the high-ranked draft group is done at the birthday/draft number level vs. the at the individual level.

Natural experiment effect on income

Using linear regression, estimate the effect of having a high-ranked draft number on income. Report the estimate and the correct standard error.

```
# Regressing income on the the `high_draft` indicator
model_income <- lm(income ~ high_draft, data = d)
summary(model_income)
```

```
##
## Call:
## lm(formula = income ~ high_draft, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67399 -21140  -3002  18005 151306
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  60761.9      235.9  257.56   <2e-16 ***
## high_draft    6637.6      528.7   12.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29530 on 19565 degrees of freedom
## Multiple R-squared:  0.007992,   Adjusted R-squared:  0.007941
## F-statistic: 157.6 on 1 and 19565 DF,  p-value: < 2.2e-16
```

```
# Computing cluster-robust standard errors since assignment happens at the birthday/draft-number level
cluster_se_inc <- sqrt(diag(vcovCL(model_income, cluster = ~draft_number)))

# Extracting coefficients for inline reference
coef_estimate_inc <- coef(model_income)["high_draft"]
se_clustered_inc  <- cluster_se_edu["high_draft"]
```

Answer: High-ranked draft numbers (1-80) are associated with an increase of approximately $6,638 in income, with a cluster-robust standard error of 0.04. We use cluster-robust standard errors since assignment to the high-ranked draft group is made at the birthday/draft number level.

Instrumental variables estimate of education on income

Now, estimate the Instrumental Variables regression to estimate the effect of education on income. To do so, use `AER::ivreg`. After you evaluate your code, write a narrative description about what you learn.

```r
library(AER)

# Regressing income on years of education using `high_draft` as an IV
model_iv <- ivreg(income ~ years_education | high_draft, data = d)
summary(model_iv)
```

```
##
## Call:
## ivreg(formula = income ~ years_education | high_draft, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -78140 -18762  -2145  16461 147217
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      15691.6     3416.4   4.593  4.4e-06 ***
## years_education   3122.4      229.6  13.601  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27260 on 19565 degrees of freedom
## Multiple R-Squared: 0.1548,  Adjusted R-squared: 0.1548
## Wald test:   185 on 1 and 19565 DF,  p-value: < 2.2e-16
```

```r
# Computing clustered variance-covariance matrix
cluster_se_iv <- sqrt(diag(vcovCL(model_iv, cluster = ~draft_number)))

# Extracting cluster-robust standard error for years_education
se_years_education <- cluster_se_iv["years_education"]
```

Answer: Using `high_draft` as an instrument in a 2SLS regression, we estimate that each additional year of education increases annual income by approximately $3,122, with a cluster-robust standard error of SE = 225.8844. The smaller coefficient on `years_education`, compared to the naive observational regression, suggests that the simple OLS model overstates the effect of education because it fails to account for both selection bias and omitted variables. In contrast, the IV estimate leverages random variation in education induced by the draft lottery to more credibly isolate the causal effect of education on income, as well as incorporation of variables that are correlated with both schooling and income.

Evaluating the exclusion restriction

Give one reason this requirement might not be satisfied in this context. In what ways might having a high draft rank affect individuals' income **other** than nudging them to attend more school?

Answer: The exclusion could be violated in the case where having a high-draft rank influences income via participation/non-participation in the military service itself. Those who do not have a high-draft number and participate in the military might have certain induced experiences that have lasting effects on their physical and mental health, which can in turn affect their participation and experience in the labor force. This chain of effects could alter their lifetime earnings independent of education, which violates the assumption that the high-draft rank as an IV affects the outcome (earnings) only via its effect on education.

Differential attrition

Conduct a test for the presence of differential attrition by treatment condition. That is, conduct a formal test of the hypothesis that the "high-ranked draft number" treatment has no effect on whether we observe a person's income. **(Note, that an earning of $0 *actually* means they didn't earn any money – i.e. earning $0 does not mean that their data wasn't measured. Let's be really, really specific: If you write a model that looks anything like, `lm(income == 0 ~ .)` you've gone the wrong direction.)**

```r
# Defining indicator variable for an observed outcome
d$observed_outcome <- !is.na(d$income)

# Counting missing observed outcomes
cat("Number of missing observed outcomes:", sum(is.na(d$income)),"\n")
```

```
## Number of missing observed outcomes: 0
```

```r
# Regressing observation status on `high_draft`
model_differential_attrition <- lm(observed_outcome ~ high_draft, data = d)
summary(model_differential_attrition)
```

```
##
## Call:
## lm(formula = observed_outcome ~ high_draft, data = d)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.800e-17 -1.800e-17 -1.800e-17 -1.800e-17  2.775e-13
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  1.000e+00  1.585e-17  6.309e+16   <2e-16 ***
## high_draft  -1.771e-17  3.552e-17 -4.990e-01    0.618
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.984e-15 on 19565 degrees of freedom
## Multiple R-squared:    0.5,  Adjusted R-squared:    0.5
## F-statistic: 1.957e+04 on 1 and 19565 DF,  p-value: < 2.2e-16
```

Answer: Regressing an indicator variable for non-null observed outcomes (`observed_outcome`) on `high_draft` to test whether assignment to a high draft number affects the likelihood of having an observed income, we see that the coefficient on `high_draft` negligible (-1.771×10^{-17}) and statistically insignificant ($p = 0.6181$). When inspecting the dataset, we confirm that there are no missing values in `income`, which helps explain why treatment status had no effect on whether an observation's outcome was measured. Therefore, there is no evidence of differential attrition (or any attrition) by draft assignment and attrition does not pose a threat to the validity of our IV analysis.

Evaluate differential attrition

Tell a concrete story about what could be leading to the result in part 7. How might this differential attrition create bias in the estimates of a causal effect?

Answer: We did not find any missing values of `income` in our dataset, which means there was no differential attrition (or any attrition at all). In principle, however, differential attrition could bias our estimates if the likelihood of observing `income` was correlated with draft number and consequently `high_draft`. For example, if individuals with low draft numbers were more likely to be drafted and thus harder to track for income data, the observed sample would over-represent those with high draft numbers, potentially biasing the IV estimate upward (according to our IV analysis). Conversely, if individuals with high draft numbers were more likely to be missing, the observed sample would over-represent those with low draft numbers, potentially biasing the IV estimate downward (according to our IV analysis).

Think about Treatment Effects

Throughout this course we have focused on the average treatment effect. *Why* we are concerned about the average treatment effect. What is the relationship between an ATE, and some individuals' potential outcomes? Make the strongest case you can for why this is a *good* measure.

We focus on the ATE because it provides a single, interpretable figure that represents the causal effect of the experiment. The ATE is the mean difference in potential outcomes between those who received the treatment versus those who did not (control). The ATE is directly related to individual potential outcomes by assigning to each subject a $Y_i(1)$ and $Y_i(0)$. Since we can never simultaneously observe both outcomes for an individual subject, randomization allows us to estimate the average of these individual-level treatment effects by comparing outcome across treated and control groups, providing a measure of the expected causal effect. Averaging provides stability that mitigates the potential influence of extreme individual outcomes, providing a robust representative measure of the treatment effect, even as effects vary across individuals.

We value the ATE because it is a reliable, unbiased estimator of the treatment effect when the treatment assignment does not systematically affect individuals' potential outcomes. When this requirement is not met (e.g. confounding, non-random assignment, selection bias, differential attrition, group imbalance, noncompliance), there are tools to mitigate the potential bias introduced (e.g. instrumental variables, difference-in-differences, better randomization strategies, inverse probability weighting, data imputing, matching, CACE). The flexibility of the ATE allows for generalization across contexts using strategies such as blocking, clustering, and weighting, which help account for group-level differences or sample imbalances, making the estimate relevant for broader populations and policy applications. This is a particularly important feature of the ATE because it allows researchers and policymakers to extend insights and findings from controlled studies to broader populations by helping predict the likely impact if the treatment were implemented more widely.

Optional Online advertising natural experiment.

Cross table of total_ads and treatment_ads

A. Run a crosstab – which in R is `table` – of `total_ads` and `treatment_ads` to sanity check that the distribution of impressions looks as it should. After you write your code, write a few narrative sentences about whether this distribution looks reasonable. Why does it look like this? (No computation required here, just a brief verbal response.)

```
# Creating the cross tab
cross_tab <- table(d$total_ads, d$treatment_ads)
cross_tab
```

```
## 
##         0     1     2     3     4     5     6
##   0 61182     0     0     0     0     0     0
##   1 36754 37215     0     0     0     0     0
##   2 21143 42036 20965     0     0     0     0
##   3 10683 32073 32314 10726     0     0     0
##   4  5044 20003 30432 20223  5115     0     0
##   5  2045 10563 20970 20793 10293  2131     0
##   6   729  4437 10977 14771 11147  4486   750
```

Answer: The table has an upper-triangular structure that reflects the fact that the number of treatment ads never exceeds the total number of homepage visits. Users tend to have treatment exposures distributed roughly symmetrically around half of their total visits, following a binomial pattern that reflects the quasi-random (Bernoulli) assignment of ads to even-second homepage visits.

Placebo test

A colleague of yours proposes to estimate the following model: `d[ , lm(week1 ~ tretment_ads)]` You are suspicious. Run a placebo test with `week0` purchases as the outcome and report the results. Since treatment is applied in week 1, and `week0` is purchases in week 0, *should* there be an relationship? Did the placebo test "succeed" or "fail"? Why do you say so?

```
# Creating placebo regression
model_colleague <- lm(week0 ~ treatment_ads, data = d)
summary(model_colleague)
```

```
## 
## Call:
## lm(formula = week0 ~ treatment_ads, data = d)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -3.248 -2.196 -1.670  2.430  8.330 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   1.669685   0.006027   277.0   <2e-16 ***
## treatment_ads 0.263099   0.003155    83.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.796 on 499998 degrees of freedom
## Multiple R-squared:  0.01372,    Adjusted R-squared:  0.01372 
## F-statistic:  6955 on 1 and 499998 DF,  p-value: < 2.2e-16
```

Answer: The placebo regression of `week0` on `treatment_ads` yields a coefficient of 0.2631 with a p-value of 0. Since treatment is applied starting in week 1, there should be no causal effect on week 0 purchases. The fact that the coefficient is positive and highly significant indicates that users with more treatment exposures already had higher baseline spending. Therefore, the placebo test fails and suggests that the treatment assignment is correlated with pre-existing differences in purchasing behavior and that simple regressions of `week1` on `treatment_ads` could lead to biased estimates of the treatment effect.

What has gone wrong?

Here's the tip off: the placebo test suggests that there is something wrong with our experiment (i.e. the randomization isn't working) or our data analysis. We suggest looking for a problem with the data analysis. Do you see something that might be spoiling the "randomness" of the treatment variable? (Hint: it should be present in the cross-tab that you wrote in the first part of this qusetion.) How can you improve your analysis to address this problem? Why does the placebo test turn out the way it does? What one thing needs to be done to analyze the data correctly? Please provide a brief explanation of why, not just what needs to be done.

Answer: From the cross-tab, we can see that users with more total homepage visits inherently tend to have more treatment exposures. This means that `treatment_ads` is mechanically correlated with `total_ads`, which is itself correlated with week 0 purchasing behavior. Users who have more homepage visits during the campaign week (and thusly higher `total_ads`) are more likely to have higher treatment exposures (`treatment_ads`) and also had a higher baseline spending in week 0. Therefore, regressing `week0` on `treatment_ads` picks up the pre-existing correlation despite there being no treatment applied yet. If we adjust for total visits by including `total_ads` as a covariate in the regression, the mechanical correlation between the treatment and baseline purchase behavior will be removed.

Conduct proposed solution and re-evaluate placebo test

Implement the procedure you propose from part 3, run the placebo test for the Week 0 data again, and report the results. (This placebo test should pass; if it does not, re-evaluate your strategy before wasting time proceeding.) How can you tell this this has fixed the problem? Is it possible, even though this test now passes, that there is still some other problem?

```r
# Creating regression that integrates `total_ads` as the covariate
model_passes_placebo <- lm(week0 ~ treatment_ads + total_ads, data = d)
summary(model_passes_placebo)
```

```
##
## Call:
## lm(formula = week0 ~ treatment_ads + total_ads, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.817 -2.079 -1.589  2.455  7.823
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.345375   0.007295 184.436   <2e-16 ***
## treatment_ads -0.002245   0.004629  -0.485    0.628
## total_ads      0.245348   0.003149  77.922   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.779 on 499997 degrees of freedom
## Multiple R-squared:  0.02555,    Adjusted R-squared:  0.02555
## F-statistic:  6556 on 2 and 499997 DF,  p-value: < 2.2e-16
```

Answer: The coefficient on treatment_ads is -0.0022 with an insignificant p-value of 0.6276. This coefficient is effectively zero (in addition to being statistically insignificant), which indicates that once we account for total homepage visits, there is no longer a confounding relationship between treatment exposures and baseline purchase behavior. While this specific issue has been remedied, it is still possible that there are other confounding variables or the possibility of nonlinear treatment effects.

Estimate treatment effect with proposed solution

Now estimate the causal effect of each ad exposure on purchases during Week 1. You should use the same technique that passed the placebo test in part 4. Describe how, if at all, the treatment estimate that your model produces changes from the estimate that your colleague produced.

```r
model_causal <- lm(week1 ~ treatment_ads + total_ads, data = d)
summary(model_causal)
```

```
##
## Call:
## lm(formula = week1 ~ treatment_ads + total_ads, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.003 -2.104 -1.542  2.447  8.110
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.317960   0.007263  181.47   <2e-16 ***
## treatment_ads 0.056340   0.004609   12.22   <2e-16 ***
## total_ads     0.224478   0.003135   71.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.767 on 499997 degrees of freedom
## Multiple R-squared:  0.02782,    Adjusted R-squared:  0.02781
## F-statistic:  7153 on 2 and 499997 DF,  p-value: < 2.2e-16
```

Answer: In this model, the coefficient on `treatment_ads` is 0.0563 with a p-value of 2.334×10^{-34}. This indicates that each additional treatment ad exposure increases purchases in week 1 by approximately 0.0563 dollars on average, after controlling for total homepage visits. Compared to the colleague's model, the adjusted estimate of `treatment_ads` is smaller in magnitude since their model confounded treatment with total visits.

Defend you method

Upon seeing these results, the colleague who proposed the specification that did not pass the placeo test challenges your results – they make the campaign look less successful! Write a short paragraph (i.e. 4-6 sentences) that argues for why your estimation strategy is better positioned to estimate a causal effect.

Answer: While the colleague's model suggested a larger effect of `treatment_ads`, it fails to account for the fact that users with more homepage visits naturally receive more treatment exposures and also tend to spend more even before the campaign begins. In contrast, our estimation strategy adjusts for total homepage visits (total_ads) which isolates the variation in treatment ads that is independent of visit frequency. Since our specification passed the placebo test on week 0 purchases, we have shown that pre-campaign spending is no longer correlated with the treatment variable. By controlling for total visits, our approach yields a less biased, more accurate estimate of the causal effect of each ad exposure on week 1 purchases. Therefore, while the adjusted effect appears smaller, it more faithfully represents the true causal impact of the advertising campaign.

Intertemporal substitution?

One concern raised by David Reiley is that advertisements might just shift *when* people purchase something – rather than increasing the total amount they purchase. Given the data that you have available to you, can you propose a method of evaluating this concern? Estimate the model that you propose, and describe your findings.

```
# Defining a cumulative revenue variable
d$cumulative_week_1_10 <- rowSums(d[, paste0("week", 1:10)])
```

```
# Create cumulative model
model_overall <- lm(cumulative_week_1_10 ~ treatment_ads + total_ads, data = d)
summary(model_overall)
```

```
##
## Call:
## lm(formula = cumulative_week_1_10 ~ treatment_ads + total_ads,
##     data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.597  -7.372  -0.731   6.654  59.782
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.15081    0.02771 618.949   <2e-16 ***
## treatment_ads  0.01274    0.01758   0.724    0.469
## total_ads      2.22834    0.01196 186.307   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.56 on 499997 degrees of freedom
## Multiple R-squared:  0.1321, Adjusted R-squared:  0.1321
## F-statistic: 3.804e+04 on 2 and 499997 DF,  p-value: < 2.2e-16
```

Answer: The cumulative model yields a coefficient on treatment_ads of 0.0127 with a p-value of 0.4688. This coefficient is small and statistically insignificant, indicating that additional treatment ads do not significantly increase total spending over the ten weeks following the campaign. Comparing these results to the week 1 model (`model_causal`), where `treatment_ads` had a statistically significant positive effect, suggests that the campaign mainly accelerates purchases into week 1 rather than generating additional spending overall. This can be interpreted as the ads shifting the timing of purchases rather than increasing total revenue.

Weekly effects

If you look at purchases in each week – one regression estimated for each outcome from week 1 through week 10 (that's 10 regression in a row) – what is the relationship between treatment ads and purchases in each of those weeks. This is now ranging into exploring data with models – how many have we run in this question alone!? – so consider whether a plot might help make whatever relationship exists more clear.

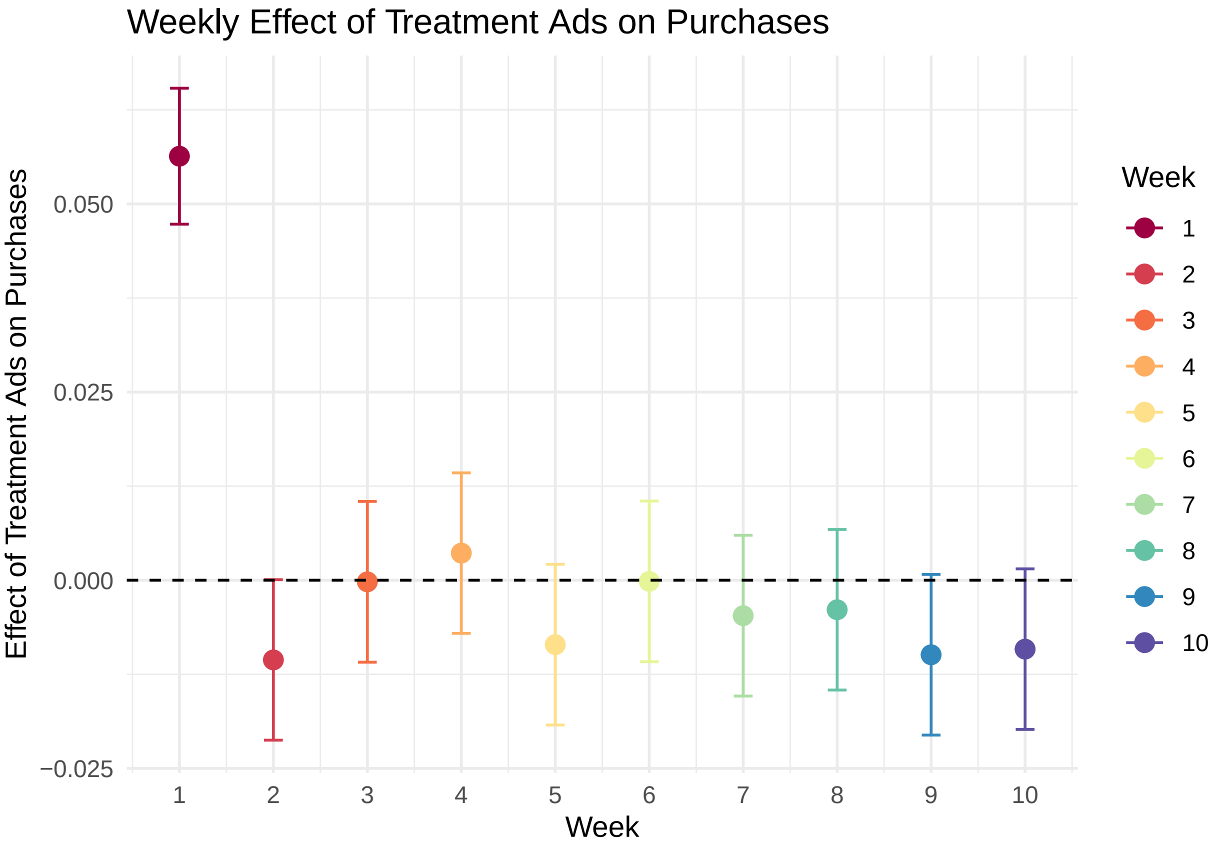
```
library(ggplot2)

# Running one regression per week and storing results
weeks <- paste0("week", 1:10)
coef_treat <- numeric(length(weeks))
se_treat <- numeric(length(weeks))

for(i in seq_along(weeks)) {
  model <- lm(as.formula(paste(weeks[i], "~ treatment_ads + total_ads")), data = d)
  coef_treat[i] <- coef(model)["treatment_ads"]
  se_treat[i] <- summary(model)$coefficients["treatment_ads", "Std. Error"]
}
```

```
# Creating data frame for plotting
plot_data <- data.frame(
  week = 1:10,
  coef = coef_treat,
  lower = coef_treat - 1.96 * se_treat,
  upper = coef_treat + 1.96 * se_treat
)

# Plotting treatment effect over time
ggplot(plot_data, aes(x = week, y = coef, color = factor(week))) +
  geom_point(size = 3) +
  geom_errorbar(aes(ymin = lower, ymax = upper), width = 0.2) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  scale_x_continuous(breaks = 1:10) +
  scale_color_brewer(palette = "Spectral") +
  labs(x = "Week",
       y = "Effect of Treatment Ads on Purchases",
       color = "Week",
       title = "Weekly Effect of Treatment Ads on Purchases") +
  theme_minimal()
```



Answer: We create a plot of `treatment_ads` coefficient estimates and 95% confidence intervals per week, where confidence intervals straddling 0 indicated statistically insignificant estimates. From the plot, we can see that week 1 sees a positive and statistically significant coefficient, showing that treatment ads increase purchases during this week. However, all of weeks 2-10 have statistically insignificant coefficients near 0. Together, these results suggest that treatment ads mainly accelerate purchases into week 1, rather than increasing overall spending throughout weeks 1-10.

Evaluating what is happening in the data

I. What might explain this pattern in your data. Stay curious when you're writing models! But, also be clear that we're fitting a **lot** of models and making up a theory/explanation after the fact.

Answer: The observed pattern of a positive effect in week 1, but near-zero effects in weeks 2–10, could indicate a timing-shift where ads accelerate purchases that would have otherwise occurred later. This suggests that users respond immediately to ads, concentrating revenue in the campaign week rather than increasing overall spending. Other factors such as seasonality or user heterogeneity could also contribute to these results, so further analysis would be needed to confirm the specific mechanism at play.

Evaluate whether there are non-linear relationships

We started by making the assumption that there was a linear relationship between the treatment ads and purchases. What other types of relationships might exist? After you propose at least two additional non-linear relationships, write a model that estimates these, and write a test for whether these non-linear effects you've proposed produce models that fit the data better than the linear model.

```r
# Creating a quadratic model
model_quadratic <- lm(week1 ~ treatment_ads + I(treatment_ads^2) + total_ads, data = d)
quad_summary <- summary(model_quadratic)

# Creating logarithmic model
model_log <- lm(week1 ~ log(1 + treatment_ads) + total_ads, data = d)
log_summary <- summary(model_log)

quad_summary
```

```
##
## Call:
## lm(formula = week1 ~ treatment_ads + I(treatment_ads^2) + total_ads,
##     data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.977 -2.107 -1.540  2.445  8.115
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.315717   0.007784 169.033  < 2e-16 ***
## treatment_ads       0.063334   0.009874   6.414 1.42e-10 ***
## I(treatment_ads^2) -0.001720   0.002148  -0.801    0.423
## total_ads           0.223935   0.003207  69.820  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.767 on 499996 degrees of freedom
## Multiple R-squared:  0.02782,    Adjusted R-squared:  0.02781
## F-statistic:  4769 on 3 and 499996 DF,  p-value: < 2.2e-16
```

```r
log_summary
```

```
##
## Call:
## lm(formula = week1 ~ log(1 + treatment_ads) + total_ads, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.898 -2.116 -1.528  2.447  8.144
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.302679   0.007380  176.51   <2e-16 ***
## log(1 + treatment_ads) 0.126815   0.010866   11.67   <2e-16 ***
## total_ads              0.224813   0.003195   70.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.767 on 499997 degrees of freedom
## Multiple R-squared:  0.02779,    Adjusted R-squared:  0.02779
## F-statistic:  7146 on 2 and 499997 DF,  p-value: < 2.2e-16
```

```r
# Comparing linear vs. quadratic model
cat("ANOVA comparison for linear vs. quadratic model: \n")
```

```
## ANOVA comparison for linear vs. quadratic model:
```

```r
anova(model_causal, model_quadratic)
```

```
## Analysis of Variance Table
##
## Model 1: week1 ~ treatment_ads + total_ads
## Model 2: week1 ~ treatment_ads + I(treatment_ads^2) + total_ads
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1  5e+05 3826899
## 2  5e+05 3826894  1      4.91 0.6415 0.4232
```

```r
cat("\n")
```

```r
# Comparing linear vs. logarithmic model
cat("ANOVA comparison for linear vs. logarithmic model: \n")
```

```
## ANOVA comparison for linear vs. logarithmic model:
```

```r
anova(model_causal, model_log)
```

```
## Analysis of Variance Table
##
## Model 1: week1 ~ treatment_ads + total_ads
## Model 2: week1 ~ log(1 + treatment_ads) + total_ads
##   Res.Df     RSS Df Sum of Sq F Pr(>F)
## 1  5e+05 3826899
## 2  5e+05 3827000  0   -101.18
```

Answer: We attempt to see if a quadratic (which could model potential diminishing returns) or logarithmic model (which could show potential saturation) is a better for for the data. Using ANOVA to compare the linear model with the quadratic model, we see that the quadratic term is near 0 (-0.0017) and statistically insignificant (p = 0.4232). Comparing the linear model to the logarithmic model, we see a statistically significant coefficient of 0.1268 with a p-value of 1.813×10^{-31}. While this coefficient is statistically significant, comparing the RSS of the linear model to the log-transformed model shows no meaningful improvement (a difference of only 0.0026% in RSS). This confirms that the log transformation does not provide a better fit than the original linear specification. Taken together, these results show both non-linear specifications fail to improve model fit over the original linear model. Therefore the effect of `treatment_ads` on Week 1 purchases is well approximated as linear and positive, conditional on total homepage visits. It appears that non-linearities such as diminishing returns or saturation are not evident in the data.