as a control: if, after the new strategy rollout, sales in the United States are growing faster than those in Canada, you have evidence that the new strategy is working. Even better, instead of just focusing on Canada, you can compare post-treatment sales to a multicountry average that tracks with U.S. sales. Each country could be weighted such that the aggregate is, historically, a good estimate of U.S. sales. That is, you can use the sales in other, untreated, countries to predict what sales would have been in the United States if you had not introduced the new strategy.

This general setting is common in business when you need to evaluate large scale policy decisions. A full randomized controlled trial is not possible because, for example, you can't randomly allocate strategies to individual sales agents (they are paid on commission, so this would be seen as unfair, and they are competing for the same customers so you would get dependence between treated and control agents). In one version of this setup, you have only two observation periods: before and after the treatment has been applied. In another version you have multiple observations periods both before and after treatment. There different techniques that apply for each version—'diff-in-diff' and 'synthetic controls' respectively. These are actually quite similar mathematically, although the estimation algorithms will look different at first glance; see Arkhangelsky et al. (2019) for an overview that ties them together. We cover each technique in what follows, and these will likely be commonly used methods from your business analysis toolbox.

0.4.1 Diff-in-Diff Analysis

A *differences-in-differences* (diff-in-diff, or DiD for short) analysis applies when you have a number of *units* who are observed in two time periods: one before any treatment, and another after a subset of units have received treatment. The diff-in-diff framework consists of nothing more than some basic regression modeling along with a strong assumption of conditional ignorability.

The classic application of this framework has two markets, say Canada and the United States (to extend our example from above), with only one receiving some sort of new sales strategy. For example, suppose you want to see the effect of a free shipping promotion. You can model the trend in sales in both countries before and after treatment—free shipping—is rolled out *only* in the United States. If sales grow in the United States relative to Canada after treatment, then you have a positive treatment effect *if you assume that this difference is not because of external shocks that hit only one of the two countries.* This last assumption is the Achilles heel of diff-in-diff analysis, and there is no way to get around it. For this reason, diff-in-diff results are only as reliable as the two groups are truly comparable.

To introduce some notation, suppose that you have units i for $i = 1, \ldots n$ that are observed at times $t = 0$ and $t = 1$. Treatment versus control group membership is encoded as $d_i = 1$ if unit i is in treatment group, 0 otherwise. The treatment group units receive treatment only in the second time period, such that the actual treatment status is $d_i t$: the interaction between the treatment group indicator and the time period indicator. For example, if unit i has $d_i = 1$ then in period $t = 0$ it

has treatment status $1 \times 0 = 0$ and in period $t = 1$ it has treatment status $1 \times 1 = 1$. The control group units always have treatment status $t \times 0 = 0$.

The diff-in-diff regression model is then

$$y_{it} = \beta_0 + \beta_1 d_i + \beta_2 t + \gamma d_i t + \varepsilon_{it} \tag{27}$$

The treatment effect of interest is γ: the coefficient on the *interaction* between d_i and t. This model can be estimated using your usual regression tools (i.e., `glm`). Since the errors within the same unit (ε_{i0} and ε_{i1}) are possibly correlated, you will want to use the 'clustered standard error' techniques of Chapter **??** to calculate the standard errors on estimated $\hat{\gamma}$ (or use a bootstrap).

Because Equation (27) has a treatment group intercept term, $d_i\beta_1$, it is fine if the groups have different averages in the pre-treatment period. However, our conditional ignorability assumption here is that there is no factor other than the treatment that would cause the treatment group responses to *change* differently from those of the control group. We are requiring that nothing impacting the response has changed, other than treatment status, across observation periods. This is a strong assumption (e.g., what if something else changed in Canada, such as the Canadian economy getting weaker, between treatment periods). But it is an assumption that is often close enough to the truth to be useful.

Example 6. eBay Sponsored Search Marketing: diff-in-diff Our diff-in-diff example is taken from the paper by Blake et al. (2014), where researchers from eBay studied the effect of sponsored search marketing (SSM). *Sponsored* or *paid* search refers to the advertised links that you see around search results on, for example, Google or Amazon. Figure 5 shows an example web page returned after a Google search, dominated by paid search results. The research question is simple: 'What is the effect of *paid search advertising?*' Or, to turn it around, what would happen to sales revenue if eBay stopped paying for SSM? Since a big website like eBay will show up anyway in the 'organic' results (those which are not sponsored; e.g., see Zappos in both organic and paid results in Figure 5), do they get any benefit from also appearing in sponsored slots? And how big is the benefit? Is it worth the cost?

Questions about marketing return on investment (ROI) are generally tough to answer. The sponsored results get clicked and lead to conversions, but you have no idea if these users would have followed the organic result if there was no sponsored option. And you can't compare the pages where eBay ads don't appear to those where they do: the ads appear with the searches that eBay and Google think are most likely to lead to clicks. That is, the pages where ads don't appear will expect to see fewer clicks on eBay links for search-relevance reasons independent of the presence or absence of sponsored results.

Blake et al. managed to convince the leadership at eBay to run a large-scale experiment where SSM was turned off for a portion of users. This created a unique opportunity to measure the treatment effect of paid search (for a single company), something that had never before been reliably measured. In particular, eBay stopped bidding on any AdWords (the marketplace through which Google

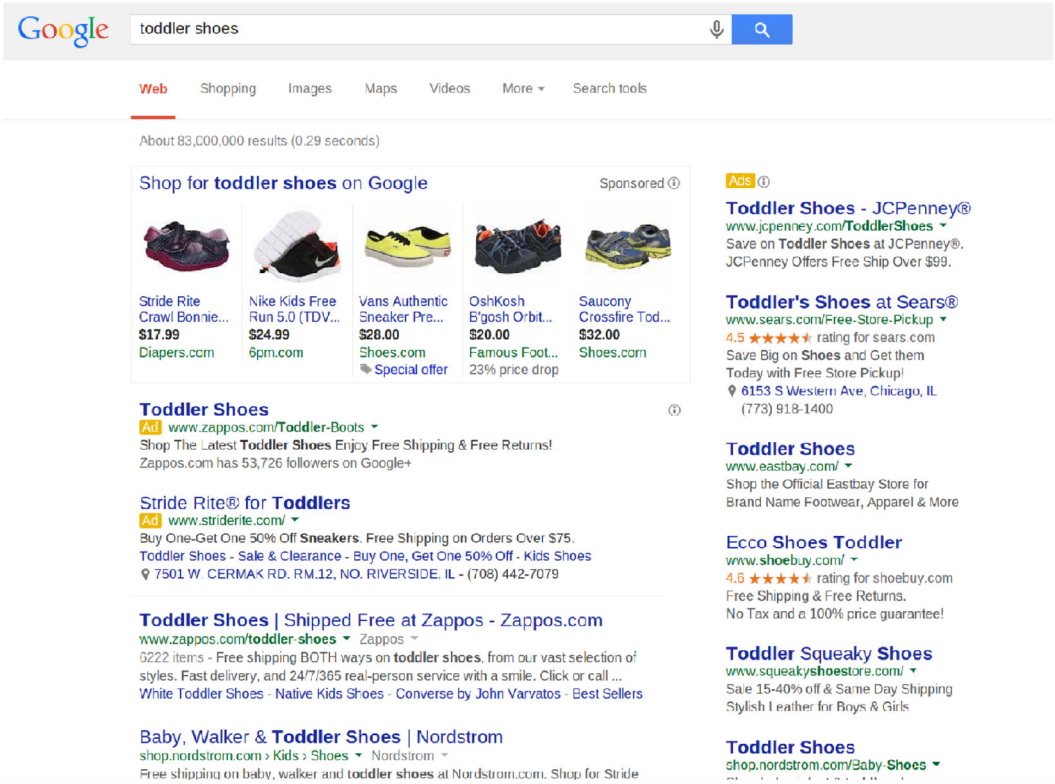

Figure 5: SSM around search results on Google. Almost everything in the screenshot is 'sponsored'—it has been paid for and has not risen *organically* through Google's relevance metrics. The only organic results are the bottom two listings in the main column, first for Zappos and second for Nordstrom.

SSM ads are sold) for 65 of the 210 'designated market areas' (DMA) in the United States for eight weeks following May 22, 2012. These DMAs are viewed as roughly independent markets around metropolitan centers ranging from Boston to Los Angeles. Google guesses the DMA on a browser and eBay can track users by their shipping address, allowing for DMA-specific treatment assignment and response tracking.

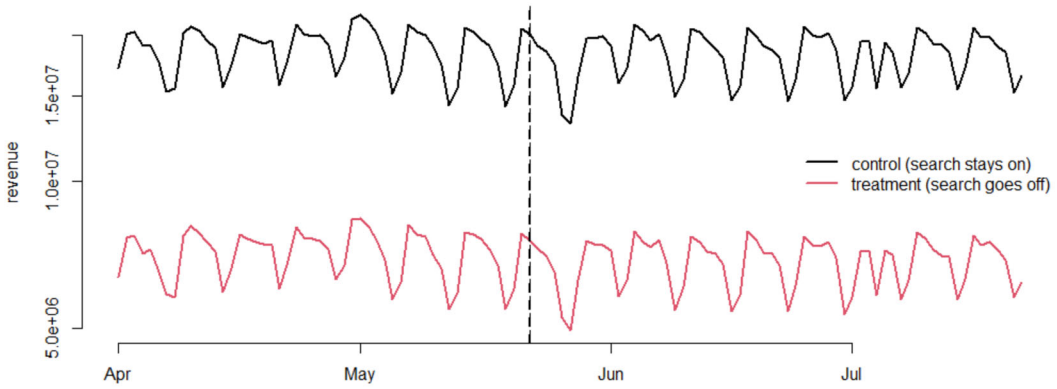

Figure 6: Daily revenue for treatment and control DMAs. The dashed line is May 22, when SSM (bidding on AdWords) was turned off for the treatment group.

The data are in `paidsearch.csv`. Note that this is not the real data; it's a simulated version that obscures real revenue numbers. The data include *daily* revenue totals for each DMA, for 51 days before May 22 and 61 days after and including May 22. The series for treatment and control groups are plotted in Figure 6. The black line corresponds to those DMAs that are never treated (SSM is always on), and the red line is for those where SSM was turned off on starting May 22 (marked with the dashed vertical). It is immediately clear that the daily revenues differ between treatment and control DMAs *before* May 22. The treatment DMAs have about 38% of the revenue of the control DMAs when SSM is on for both. This is not a problem for a DiD analysis, since we allow the groups to have a different baseline expectation (this is due to the $d_i\beta_1$ term in the DiD regression model). However, it illustrates that you wouldn't want to estimate the treatment effect by taking a simple difference in means across treatment groups.

Figure 7 shows the log difference between daily revenues in each group; this is the log of the ratio of the black line over red line from Figure 6. There does appear to be an increase in the difference in the logs following May 22. Is it real (i.e., statistically significant), and what are the implications for the return on investment for SSM?

Assuming that nothing other than the SSM-turnoff changes between the treatment groups after May 22, you can answer these questions with basic regression modeling. After some initial data wrangling, we have the `ebay` data frame consisting of a row for each DMA in each of $t = 0$ and $t = 1$.

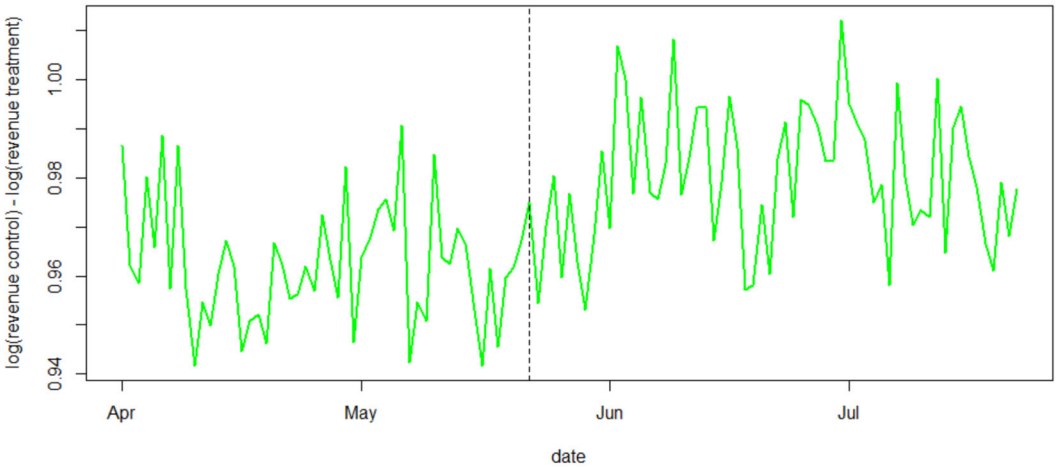

Figure 7: The difference in daily log revenue between treatment and control groups.

```
> head(ebay)
  dma post.treat ssm.turns.off    revenue
1 500          0              1   75866.62
2 501          0              0 2162945.53
3 502          0              0   32718.68
4 503          0              0   36063.90
5 504          0              0  661015.85
6 505          0              1  371153.89
```

Our treatment group indicator, d_i, is `ssm.turns.off`. The observations period, t, is `post.treat`; it is zero before May 22 and 1 on May 22 and afterwards. The revenue column here is the *average daily revenue* for each DMA in each observation period. This format makes it easy to run the regression of Equation 27.

```
> did <- glm(log(revenue) ~ ssm.turns.off*post.treat, data=ebay)
> coef(did)
             (Intercept)  ssm.turns.off     post.treat
            10.963784366    0.011932272   -0.039359359
ssm.turns.off:post.treat
            -0.005775498
> 1-exp(-0.0057755)
[1] 0.005758854
```

Our estimated treatment effect is $\hat{\gamma} = -0.005775$, the fitted coefficient on the *interaction* between treatment group d_i and observation period t. Exponentiating shows that this corresponds to an approximately 0.58% drop in average daily revenue due to SSM having been turned off. You can use the `sandwich` library to get a standard error for $\hat{\gamma}$ that accounts for within-DMA dependence in errors.

```
> library(sandwich)
> library(lmtest)
> coeftest(did, vcov=vcovCL(did, cluster=ebay$dma))
                            Estimate Std. Error  z value Pr(>|z|)
ssm.turns.off:post.treat -0.0057755  0.0057018  -1.0129   0.3111
```

This says that the treatment effect—turning off paid search ads—is not statistically significant (p-value of > 0.3). Even if the result was statistically significant, the estimated effect size is so small that it is doubtful that paid search would have a positive ROI once the *cost* of the marketing is accounted for. A caution, however: this result is for a specific company and for situations where eBay links often occur in the top organic search results. There will likely be a positive ROI for digital marketing in other specific cases, especially when the advertiser is not well known or would not occur in the top organic results.

Before moving on, we note that this DiD analysis is often presented as the analysis of differences between the pre- and post-treatment observations for each DMA. In this presentation, you first calculate the sample of pre-post *differences* for each DMA,

$$r_i = y_{i1} - y_{i0}. \tag{28}$$

You then collect the average difference for the treatment and control groups, say $\bar{r}_1$ and $\bar{r}_0$, and use the difference between these averages as the ATE estimate:

$$\hat{\gamma} = \bar{r}_1 - \bar{r}_0. \tag{29}$$

This routine is the source of the 'difference in differences' name. It gives the exact same answer as we found in our regression analysis.

```
> r <- tapply(log(ebay$revenue), ebay$dma, function(y) y[2]-y[1])
> d <- ebay[match(names(r),ebay$dma),"ssm.turns.off"]
> rBar <- tapply(r,d,mean)
> rBar[2]-rBar[1]
           1
-0.005775498
```

You can also apply the usual formula to get the sampling variance for a difference in means: $\text{var}(\hat{\gamma}) = \text{var}(\bar{r}_1) + \text{var}(\bar{r}_0)$.

```
> rBarVar <- tapply(r, d, function(r) var(r)/length(r))
> sqrt(sum(rBarVar))
[1] 0.00572258
```

The standard error is practically unchanged from our earlier regression analysis.