UNIVERSITY OF CALIFORNIA
SANTA CRUZ

CAUSAL INFERENCE IN REPEATED OBSERVATIONAL STUDIES: A STUDY OF EBAY PRODUCT RELEASES

A project submitted in partial satisfaction
of the requirements for the degree of

MASTER OF SCIENCE
in
STATISTICS AND APPLIED MATHEMATICS

by

Vadim von Brzeski

September 2015

The Project of Vadim von Brzeski
is approved:

Professor David Draper

Professor Matt Taddy

Abstract

Causal inference in observational studies is notoriously difficult, due to the fact that the experimenter is not in charge of the treatment assignment mechanism. Many potential confounding factors (PCFs) likely exist in such a scenario, and if one seeks to estimate the causal effect of the treatment on some response, one needs to control for such factors. Identifying PCFs may be very difficult (or impossible) given a *single* observational study. However, the task becomes significantly easier if one can observe a *sequence* of similar treatments over the course of a lengthy time period, because one can identify patterns of behavior of the experimental subjects that need to be controlled for. Specifically, one key pattern (PCF) that may easily emerge is the *early adopter effect*: the scenario where the magnitude of the response is highly correlated with how quickly one adopts a treatment after its release.

We detail a methodology to control for the early adopter effect by jointly modeling multiple observational studies (treatment events) simultaneously. We show that in the presence of the early adopter effect, it is nearly impossible to obtain robust estimates of the treatment effect by analyzing a single observational study in isolation. Furthermore, we show how flexible hierarchical models that account for user heterogeneity are necessary to accurately estimate the factual and counterfactual user responses.

We illustrate the power of this approach through a large-scale case study involving product updates (newer versions of the same product) from eBay, Inc. The product updates we study are not done in a randomized fashion, and users upgrade (or not) to a new version of the product at their own volition and timing. Therefore, we are by definition in an observational study setting. Our response variable is a measure of *user actions*, and we study the behavior of a large set of users in a targeted subset of eBay categories over a period of one year.

Contents

List of Figures

List of Tables

Acknowledgment

I would like to thank my advisor Professor David Draper for inspiring in me a passion for (Bayesian) statistics, for allowing me the opportunity to work on this project at eBay Inc., and for his continuous support and guidance. I would also to like to send a *very special thank you* to Matt Taddy (University of Chicago), my co-advisor on this project, without whom this work would not have been possible. In addition, I'd like to thank Matt Gardner (eBay) for the initial formulation of the problem. Last but certainly not least, I am greatly indebted to Steve Tadelis (eBay) for his help in ensuring this work saw the light of day.

Chapter 1

Introduction

The problem of *causal inference* has a long history in statistics. Nearly one hundred years ago, Jerzy Neyman (Neyman, 1990) introduced the *potential outcomes approach* to causal inference, namely that causal effects can be defined as comparisons of potential outcomes. Donald Rubin (Rubin, 1973) was responsible for giving the potential outcomes approach a solid foundation for *observational studies*. The potential outcomes approach is currently the standard approach to inferring causal effects of a *treatment* or *manipulation*, and before describing our specific problem, we will briefly review this approach and formalize some terminology.

1.1 The Fundamental Problem of Causal Inference

Causal inference attempts to answer the following question: given an observable response Y, a measurable treatment (manipulation, intervention, etc.) Z, a set of n subjects $i = 1, \ldots, n$, partitioned into distinct *treatment* and *control* groups, how much of the observed response was *caused* by the treatment? The *treatment* group (T) contains subjects that were exposed to the treatment, and the *control* group (C) contains those not exposed to the treatment; thus we associate with each subject i an *action* Z_i, where an action is one of treatment or no-treatment. For example, suppose the treatment is "a single Pizza-Hut ad during half-time at the SuperBowl" and the response is the "number of Pizza-Hut pizzas ordered

following half-time at the SuperBowl". The action is then "watching the ad"; the treated group would consist of all folks who did see the ad, and the control group would consist of those who did not see the ad. The causal inference question then becomes: how much effect did the ad have on the number of ordered pizzas, i.e., how much of the measured response was caused by that particular ad?

In the case of binary treatments (actions), the potential outcomes approach defines for each subject i two *potential* outcomes: the response of the subject under treatment $Y_i(Z_i = 1) \equiv Y_i(1)$, and the response of the subject under no-treatment (control) $Y_i(Z_i = 0) \equiv Y_i(0)$. However, for any individual subject i, we cannot observe both outcomes $Y_i(0)$ and $Y_i(1)$, hence the designation "potential" outcomes. This then brings us to the "fundamental problem of causal inference" (Holland, 1986): to estimate the causal effect of a treatment, we need to compare the two potential outcomes for each individual, namely $Y_i(1) - Y_i(0)$, but we get to observe only one of those quantities: either $Y_i(1)$ *or* $Y_i(0)$. Continuing with the pizza ad example above, once a person has watched the ad, he cannot "un-watch" it and erase it from his memory; also, suppose that a person who missed the ad during half-time will not see it again for the remainder of the game, since it was only scheduled to be shown at half-time.

In order to make progress in the face of this problem, Ronald Fisher (Fisher, 1935) introduced the idea of randomization as the "reasoned basis" for inference (Imbens and Rubin, 2015). The idea behind randomization is to remove any *potential confounding factors* (PCFs) that could bias the estimate of the causal effect. PCFs are attributes of the subjects (usually covariates) that are correlated with *both* the treatment assignment and the response. In the presence of PCFs, *who* is treated is not independent of the response, leading to a biased estimate of the causal effect. To see this in the pizza ad example above, suppose that *unbeknownst to the analyst*, the overwhelming majority of folks who saw the ad were men, whereas the overwhelming majority of folks who did not see the ad were women. The treatment is thus correlated with gender. Suppose also that men on average order more pizzas than women; the response is thus also correlated with gender. Gender is therefore a PCF. Having observed a large difference in the number of ordered pizzas between

treatment and control groups, the analyst may conclude that the ad had a huge positive effect. However, had the analyst known and controlled for the PCF of gender, he may have found that the increase in pizza orders was mainly due to gender differences rather than the ad. By having full decision power over which subjects will be treated and which ones will not, researchers running *randomized control trials* ensure a-priori that the distributions of any and all PCFs in the control and treatment groups are as equal as possible.

However, randomized control trials are not the only studies in which one is interested in causal effects. The other types of studies, *where the treatment assignment is not under the control of the researcher*, are called *observational studies*. A treatment event occurs at some point in time, and data is collected on subjects before and after the treatment. Since the treatment assignment is unknown, it is very likely that PCFs exist. Causal inference in observations studies is therefore much more challenging, and it is exactly the topic of this work.

1.2 Estimating Causal Effects in Observational Studies

We review some of the popular methods in practice for causal inference in observational studies, and provide references for further investigation. As mentioned above, to estimate the causal effect we need to compare potential outcomes. However, the problem is complicated by the presence of unknown PCFs. In other words, there is a belief that the treatment and control groups are different along some attributes (covariates), and these attributes are correlated with both the treatment assignment and the response. Thus the central problem in observational studies is to control for these attributes when comparing treatment and control groups. There are several ways to do this.

One of the most widely techniques relies on *matching* (Rubin, 1973) treated subjects and control subjects on the hypothesized covariates, with the goal of achieving a *balance in the covariate distributions* in the treatment and control groups (Rosenbaum and Rubin, 1985). Having defined a set of covariates, matching algorithms will attempt to find the "closest" match to a treatment subject in the control group. "Closest" can mean any number of

things, e.g., Mahanalobis metric matching if the covariate space is a vector space in $\Re^n$. Having identified the best control subject for each treatment subject, the algorithm computes the average difference between the pairs of treatment and control subjects. Subjects for which suitably close matches cannot be found may be dropped from the analysis. For a good review of matching techniques, see Stuart (2010).

A popular method for balancing out covariate distributions is known as *propensity score* (Rosenbaum and Rubin, 1983) matching. The propensity score for a subject i is defined as the probability of receiving the treatment given the observed covariates. There are two important properties of propensity scores. First, at each value of the propensity score, the distribution of the covariates defining the score is the same in the treatment and control groups, i.e., they act as balancing scores. Second, if treatment assignment is ignorable given the covariates, it is also ignorable given the propensity score. Thus to compute the causal effect, one can compare the mean responses of treated and control subjects having the same propensity score. However, we must caution that the above two properties only hold if one has found the *true* propensity score model: a poor estimate of the *true* propensity score will again lead to biased causal effect estimates (Kang and Schafer, 2007). Besides matching on the propensity score, other techniques involve using the propensity score in subclassification (Rosenbaum and Rubin, 1984), weighting (Rosenbaum, 1987), regression (Heckman et al., 1997), and/or combinations of the above (Rubin and Thomas, 2000). Bayesian analyses using the propensity scores also exist (McCandless et al., 2009).

Instrumental variables (Angrist et al., 1996) also have a very long history, and are widely used in econometrics as a way to approach unbiased causal estimates in the presence of PCFs. In the OLS regression setting where the response $y = \beta x + \varepsilon$, if the regressor x is uncorrelated with the error ε, then one obtains unbiased estimates of β. However, in some situations x and ε may indeed be correlated, i.e. suppose $\varepsilon = \varepsilon(x)$. In that case, $dy/dx = \beta + d\varepsilon/dx \neq \beta$ and thus the OLS estimator is biased. If we can find an *instrumental variable* z such that it affects y *only through its effect* on x and is uncorrelated with ε, then we can still estimate β in an unbiased fashion. The issue is that such variables - whose only impact on the response is indirectly through some other covariate - are not easy to find in

most situations.

The above approaches do not rely on any specific model of the data; they compare mean responses between specially constructed samples of subjects from treatment and control groups. Model based approaches (such as ours in this paper), on the other hand, attempt to jointly model the treatment and the response in a flexible way so that the unknown counterfactual potential outcomes can be estimated (predicted) by the model. The models are typically linear regression models of the response, but can also be sophisticated non-parametric models (decision trees) (Hill, 2011), (Karabatsos and Walker, 2012). A recent method utilizes a Bayesian time-series approach and a diffusion-regression state-space model to estimate the causal effect of an advertising campaign (Brodersen et al., 2015); this approach is closest in spirit to our methodology, but it analyzes the effect of only a single intervention.

1.3 Our Contribution

Without loss of generality, let us imagine an observational study in which the response is some measure of user activity (e.g. miles jogged, items bought, ads clicked, etc.), and where the availability of a treatment is announced at some point in time. Users take advantage (or not) of the treatment at their own volition over the subsequent days, weeks, etc., and the response of each user is recorded over time. Furthermore, suppose that (a) the majority of users who adopt the treatment at all do so in a relatively short time period after its release, and (b) those users who are the earliest adopters exhibit a higher average response compared to those who wait longer to try the treatment. In other words, the *waiting time* is (negatively) correlated with the treatment and the response, making it a PCF (the longer a user waits, the less likely he will try the treatment, and if he does, the lower his response). Therefore, by definition, *waiting time* is a PCF. We refer to the above situation as the *early adopter effect*: the overall response is a (confounded) combination of the *characteristics of* the treatment and the simple *existence/appearance of* the treatment.

Situations where the early adopter effect occurs are not hard to imagine. For example,

suppose a new weight-loss diet is introduced and marketed to the general public. It is reasonable to assume that a majority of people who are already health-conscious and driven to lose weight adopt the diet in the first couple of weeks after its introduction; those who are not so concerned with their health and/or weight procrastinate and wait months before finally trying it. Supposing that a large change in average weight loss was observed in users who tried the new diet, the causal inference question becomes: how much did the new diet have to do with it? This causal effect question is *impossible to answer unless one can control for the early adopter effect*: it may well be that the health conscious folks would have lost the same amount of weight without the new diet anyway. The key problem is that the effect of the diet itself (the treatment) is confounded with the habits of the health-conscious (the early adopters).

This brings us to the major contribution of this paper. We demonstrate that in observational studies where the early adopter effect exists, it is very difficult to obtain a reasonable estimate of the treatment effect (on the treated) when one only considers a *single treatment event*. However, we also show that the task is made considerably easier when one studies a *sequence of similar treatments* over an extended period of time. In the single treatment event scenario, one's only option is to discover "static" user attributes that control for the early adopter effect, i.e. what is it about a user that makes him or her an early adopter? This may be a difficult or impossible task altogether if little or no data (e.g. demographics, etc.) is available on the users. (This is akin to modeling the *propensity score* (Rosenbaum and Rubin, 1983), but note that in this case we would need a *time-varying* propensity score that modeled a user's probability to adopt the treatment over time - not a trivial undertaking and one that also requires additional user data).

On the other hand, given a sequence of similar treatments, the problem is greatly simplified if we assume that the (unknown) early adopter behavior is relatively consistent from one treatment event to the next. It is easier because we do not need to know the true characteristics (true PCFs) that make a user an early adopter. Instead, we simply include a set of covariates (PCFs) that encode a user's waiting time into our models, and thus account for the early adopter portion of the total response, leading to a less biased (un-confounded)

estimate of the treatment effect. The two conditions (assumptions) required to make this work are as follows:

1. Early adopters should show a similar response pattern from one treatment to the next; i.e. for any pair of treatments (T_i, T_j), the average response of early adopters of treatment T_i should not differ greatly from the average response of early adopters of treatment T_j.

2. The sets of users classified as early adopters in treatment T_i and T_j should not show a high degree of overlap; i.e. the early adopters cannot be the same (or very nearly the same) set of users from one treatment to the next.

These conditions can easily be verified by exploratory data analysis (see below). The first condition allows us to learn the contribution of the early adopter effect to the overall response. The second condition ensures that we have an identifiable model with which to do so.

The remainder of this work is organized as follows. In chapter 2, we review the standard estimators of treatment effects found in the literature and describe the estimator we will be using in our work. We also describe the exact nature of the causal inference problem at eBay and give the reader a preview of our major results. Chapter 3 details our models, design matrices, and counterfactual computations. Our results (causal effect estimates) are shown in chapter 4. We conclude with a chapter on model validation (in which we check our assumptions and the out-of-sample performance of our models), and with a summary of our main results.

Chapter 2

Problem Statement and Definitions

2.1 Case Study: eBay Product Releases

Our case study deals with a sequence of observational studies at eBay Inc., in which analysts attempted to infer the causal effect of new versions (releases) of a specific software product, henceforth referred to as the *Product*, on aggregate *User Actions* with said Product (the true response and the true product are not disclosed due to confidentiality reasons).

The exact nature of the Product is not important; however, it possesses a number of very relevant qualities to our study. First, newer versions (upgrades) of the Product are released on a semi-regular basis, with releases happening on the order of $6-8$ weeks apart on average. Second, once a new version of the Product is released and becomes available to the general public, users adopt (upgrade to) the new version *at their own volition and timing*. The new version of the Product is not an *en masse* replacement of the previous version; rather users choose to upgrade to it or not. Some users upgrade immediately (or shortly after) the version becomes available - we will refer to these users as "early adopters"; some users never upgrade and continue to use the same version of the Product throughout our study. The first quality will be important to us as we analyze user behavior and look for PCFs. The second quality is precisely what makes this an observational study.

For the purposes of this paper, *User Action* is a normalized, non-negative, unit-less quantity reported in "user-action units" (UAs). Higher aggregate values of UA imply higher

(aggregate) levels of satisfaction with the Product by the users in our study, and conversely lower UA values imply lower (aggregate) levels of satisfaction with the Product. The graph of weekly aggregate UA over a 52-week period is shown in figure 2.1. The dashed vertical lines in figure 2.1 indicate weeks of Product releases; there were 7 unique Product releases (treatments) in our 52-week time window. Each release corresponds to a new version (upgrade) of the Product, e.g., version 8 to version 9, etc. To give the reader a sense of an individual user's behavior, UA for four random users is shown in figure 2.2. At the individual level, UA is very unpredictable. However, if we plot aggregate (scaled) UA over all 10.5M users, we see that it is more well behaved (figure 2.1).

We would like to estimate what the UA graph in figure 2.1 would look like in a counterfactual setting we shall now describe. Suppose we take two consecutive version releases, say v_1 and v_2 released on weeks t_1 and t_2, respectively, and we take v_2 as the *counterfactual version* (the one whose causal effect we want to estimate). Now, suppose that instead of releasing v_2 in week t_2, *eBay instead releases* v_1 *again but labels it as* "v_2". One can envision this counterfactual universe as eBay releasing a "placebo" version which has a new label, but is in fact identical in functionality to its predecessor. We use this counterfactual construction because we are interested not in what would have happened had a release never happened at all, but what difference did the features of the new release have on User Action? It is worth emphasizing that we can never know the true counterfactual given this data gathering method (observational study); we cannot roll back time and roll it forward again in an alternate universe.

2.2 Our Approach and Data

Instead of looking at any single Product version release in an isolated fashion we approach the observational study problem from a longitudinal perspective and jointly model the sequence of Product releases.

Our dataset consists of the UA response for $\approx 10.5M$ eBay users over an (undisclosed) 52 week period. The data is aggregated week by week, i.e., $t = 1, \ldots, T$, where $T = 52$. For

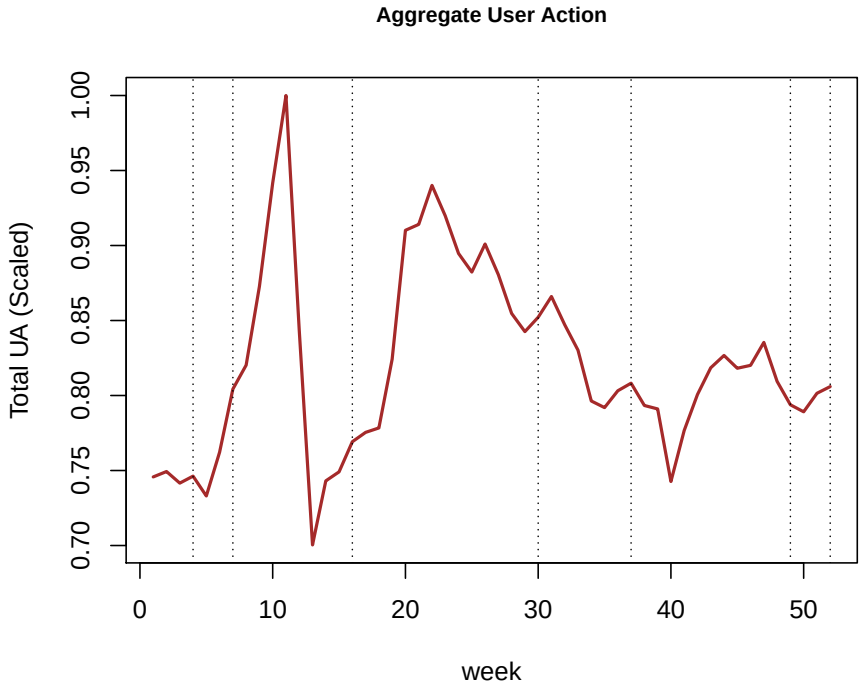

Figure 2.1: Aggregate (scaled) UA for 10.5M users over the 52 week period. The vertical lines indicate the weeks of new Product version releases.

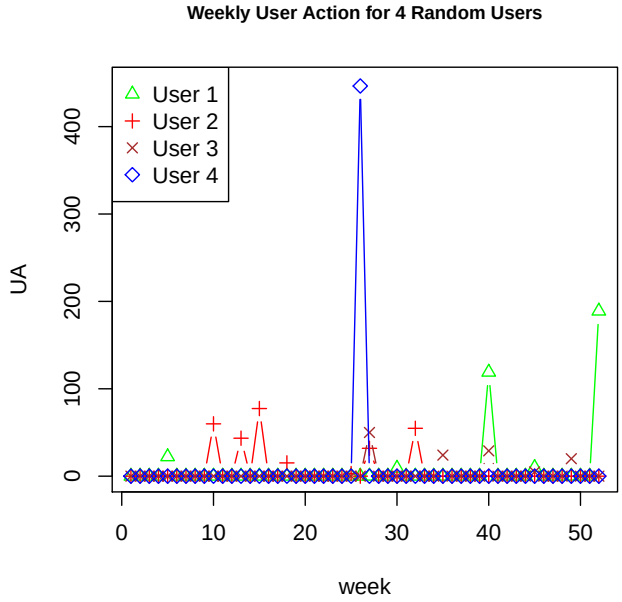

Figure 2.2: UA for four random users over the 52 week period. As one can see, there is great variety in the frequency and intensity of user actions.

each user $i = 1, \ldots, n$, where $n = 10,491,859$, we have Product usage data (session logs) broken out by version; i.e., for each week, we know which version of the Product a user had, and if he (she) upgraded mid-week, we know the relative proportion of each version's usage during that week. A user was included in our study if he (she) was a registered eBay user as of the first day of our study, *and* had at least one Product session logged in our 52-week window. Note: our response UA is correlated with Product usage (number of sessions logged), but it is not the same as Product usage. A frequent user (many logged Product sessions) can still have zero UAs logged.

Our dataset contains 11 distinct versions: 2, 3, 4, 5, **6, 7, 8, 9, 10, 11, 12** (so designated due to confidentiality reasons); the ones in boldface were released during our 52-week window (the others were legacy versions). We also had some users on versions prior to version 2; we lump all these into a *pre_v_1* category. This gives us a total of $R = 12$ version indicators.

To determine if our case study exhibits the early adopter effect, we construct the graph shown in figure 2.3, which shows the average UA per user per week for each individual version. We compute the graph by summing up the UA by Product version in a given week, and then dividing that sum by the number of users of that version in the given week. Two points are made clear by the curves in figure 2.3:

1. Users who upgrade to a new version in its first weeks of availability are the ones who are the most active on average (measured in UA units). These are the early adopters. Average UA per user declines as more and more "late adopters" join the ranks and upgrade to the latest version.

2. The UA pattern from one version release to another is quite consistent and exhibits a similar pattern for each release: the average (scaled) per user UA in the first week of each release is between 0.75 and 0.91, with most of the averages peaking around 0.80.

Therefore we believe the assumptions with respect to the early adopter effect hold in our case study, and we must control for this effect in our models so we can *determine the causal impact of the functionality of the version alone.*

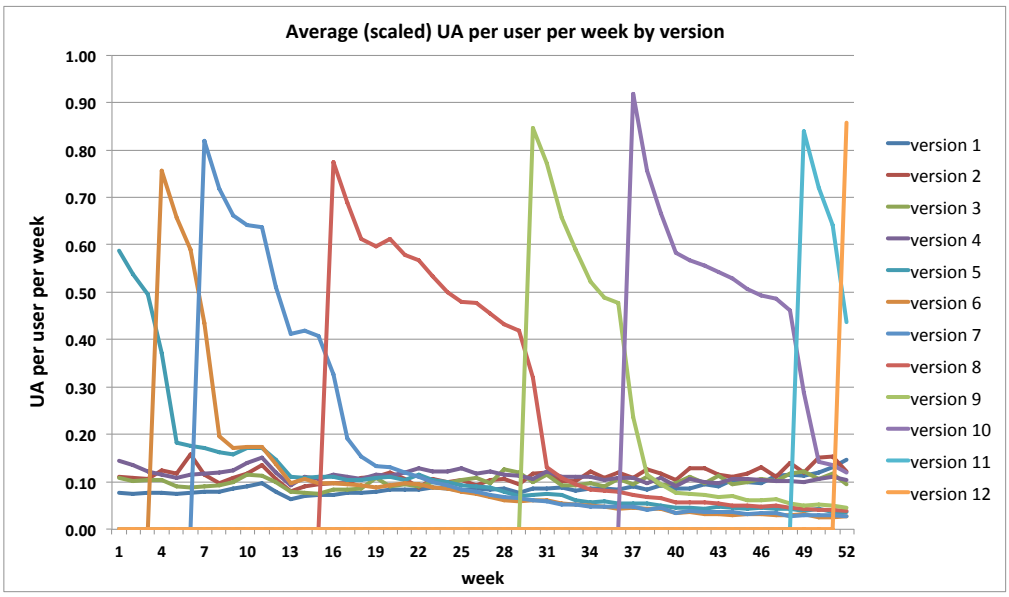

Figure 2.3: This is the average (scaled) true UA per user per week for each individual version. This graph shows that the early adopters of a new release have the highest UA average, and that the early adopter effect is quite regular from version to version.

2.3 Estimates of Treatment Effects and Assumptions

We briefly review the standard estimators for causal effects found in the literature and used in practice (Imbens, 2004), and we discuss the one we chose for our case study. We also discuss some of the assumptions involved when using these estimates of causal effects.

As mentioned above, the basic idea behind estimating causal effects is the comparison of potential outcomes $Y_i(0)$ and $Y_i(1)$. We recall that we have $i = 1, \ldots, n$ units, and a binary treatment assignment variable Z_i, where $Z_i = 1$ means unit i received the treatment. For each unit i we define two potential outcomes (responses): $Y_i(Z_i = 0) = Y_i(0)$, and $Y_i(Z_i = 1) = Y_i(1)$. For each unit i, we also have a set of *pre-treatment* covariates $\mathbf{X}_i$.

- *Average Treatment Effect* (ATE): ATE is defined as $E[Y_i(1) - Y_i(0)]$, and it is a measure of treatment effect over the *population*, where the expectation is with respect to the distribution induced by random sampling (Imbens, 2004). For example, in our eBay case study, the population consists of users of a similar product and similar to those in the study at about the same time as the study.

- *Sample Average Treatment* Effect (SATE): SATE is defined as $\frac{1}{n}\sum_{i=1}^{n}[Y_i(1) - Y_i(0)]$;

it is computed like the ATE, but only for the sample, not the entire population.

- *Conditional Average Treatment Effect* (CATE): CATE is defined as $\frac{1}{n}\sum_{i=1}^{n} E[Y_i(1) - Y_i(0)|\mathbf{X}_i = \mathbf{x}]$, i.e., the treatment effect on the population conditional on some covariates or PCFs.

The first two treatment estimators (ATE and SATE) do not apply in our case since (a) we are not drawing inferences about the population, and (b) we do condition on $\mathbf{X}$ in our models. However, we also do not estimate the CATE here because it requires the estimate of two counterfactuals: $Y_{i:Z_i=0}(1)$, the response of the control if they were treated, and $Y_{i:Z_i=1}(0)$, the response of the treated had they remained in the control. Given our case study, we are able to reliably place the treated (upgraders) into the control group (non-upgraders), but are not yet able to reliably predict *who out of the non-upgraders would upgrade* and *when they would upgrade*.

Therefore, here we employ the Conditional Average Treatment (Effect) on the Treated (CATT) as our measure of causal effect, initially similar to the above estimators but only dealing with the treated group. CATT is defined as:

$$\text{CATT} = \frac{1}{n_T} \sum_{i:Z_i=1} E[Y_i(1) - Y_i(0)|\mathbf{X}_i = \mathbf{x}] \tag{2.1}$$

There are a number of assumptions we need to state before we can use the CATT estimator above:

1. *Ignorability (unconfoundedness) assumption* (Rosenbaum and Rubin, 1983): $Z_i \perp Y_i(0), Y_i(1)|\mathbf{X}_i = \mathbf{x}$.

   This assumption deals with the PCFs $\mathbf{X}_i$ and states that if indeed you have controlled for all PCFs $\mathbf{X}_i$, then treatment assignment Z_i and response Y_i are conditionally independent given the PCFs. If this is indeed the case, it can be shown that the causal effect estimate will be unbiased (Rosenbaum and Rubin, 1983). The issue is that this assumption can almost never be verified, because one can never know if one has in fact controlled for *all* confounding factors. Thus in practice, we typically

proceed by adding covariates (PCFs) which are reasonable predictors of the response and may be correlated with the treatment. We stop when we have exhausted our supply of "feasible and reasonable" covariates and report our results using this set. It is also possible to perform a sensitivity analysis, examining how "impactful" an omitted PCF would have to be to sharply change your conclusions given your "best" set of PCFs.

2. *Overlap assumption for CATT* (Heckman et al., 1997): $\Pr(Z = 1|\mathbf{X} = \mathbf{x}_i) < 1$

   This assumption states that when conditioning on some $\mathbf{X} = \mathbf{x}_i$, one cannot have all subjects in the treatment group. There must be some subjects in the control group, else one cannot estimate the effect on the treated using the potential outcomes framework. This assumption can be verified in some respects, and we do so in section A.1.

3. *Stable Unit Treatment Value Assumption (SUTVA)* (Imbens and Rubin, 2015):

   This assumption states that the potential outcomes for any unit do not vary with the treatments assigned to other units. In other words, whether a given subject is treated or not has no impact on another subject's response and vice-versa. This can also in principle be verified, but we assume it holds in our scenario because we assume that the particular eBay product under consideration does not have a"viral" (i.e., an exponential adoption rate) nature to it, a fact confirmed by conversations with eBay Product Managers.

Under the ignorability assumption above, and for some flexible function (model) f, $E[Y_i(0)|\mathbf{X}_i = \mathbf{x}] = E[Y_i|Z_i = 0, \mathbf{X}_i = \mathbf{x}] = f(0, \mathbf{x})$, our CATT estimates become:

$$\text{CATT (per treated user)} = \frac{1}{n_T} \sum_{i:Z_i=1} [Y_i - f(0, \mathbf{x}_i)] = \frac{1}{n_T} \sum_{i:Z_i=1} [Y_i - \hat{\mathbf{y}}_i^{CF}] \tag{2.2}$$

and we define the *CATT Causal Ratio (CCR)* as:

$$\text{CATT Causal Ratio (CCR)} = \frac{\sum_{i:Z_i=1} f(0, \mathbf{x}_i)}{\sum_{i:Z_i=1} Y_i} = \frac{\sum_{i:Z_i=1} \hat{\mathbf{y}}_i^{CF}}{\sum_{i:Z_i=1} Y_i} \tag{2.3}$$

Simply put, the CCR is the ratio of the aggregate $\hat{\mathbf{y}}_i^{CF}$ to the aggregate $\mathbf{y}_i$ for the treated group. All results below are reported in terms of CCR. Note that if $\sum_{i:Z_i=1} \hat{\mathbf{y}}_i^{CF} = \sum_{i:Z_i=1} Y_i$, then CCR=1, which means that the treatment had no causal effect on the response of the treated.

2.4 Previous Causal Estimates

We shall now describe a previous approach at eBay to the above causal inference problem. The approach focused on analyzing the causal effect of one release at a time. A pool of users was selected based on activity logs in a -/+ 2 week window around the release in question (the counterfactual release). The pool of users was then divided into treatment and control groups based on their version usage during the pre-release and post-release window. The UA for both groups was computed in a 2 week window before the release, and in a 2 week window that started 3 days (burn-in) after the day of Product release. The results are shown in table 2.1 below. As we can see from this table, a simple ("unadjusted") estimate using the means of the treatment and control groups shows CCR values in excess of 1.30 (very large in relative terms). The "PCF adjusted" CCR (approximately $1.10 - 1.14$) was computed using the SATE version of the CCR estimator described above, using a regression model that included hypothesized PCFs as covariates. No one in the organization believed the "unadjusted" estimates, and although the "PCF adjusted" numbers were more reasonable, no one believed them either because they were still large in relative terms. Therefore, a new approach was necessary. (Notice that the results were similar for *two* releases, suggesting that each release had a significant impact on UA, further eroding the credibility of this approach.)

Version	N Treated	N Control	Unadjusted CCR	PCF Adjusted CCR
5	3638K	561K	1.35	1.14
7	3838K	704K	1.32	1.10

Table 2.1: Previous attempts at answering the causal effect question lead to *unrealistic* causal effect estimates. "Unadjusted" refers to a simple comparison of means (null model); "adjusted" refers to regression models with a variety of PCFs as covariates, and is based on the SATE estimator.

2.5 Preview of Our Final Results

In the sections below, we demonstrate the following:

- If the response to a treatment exhibits the characteristics of an early adopter effect, then it is practically impossible to isolate the causal effect of the treatment by simply analyzing a single treatment event (i.e. a single observational study) in isolation. Attempting to do so (without including the appropriate PCFs) will either yield unrealistic or unstable estimates of the treatment effect due to model identifiability issues.

- To isolate the causal effect of the treatment in such a scenario, the straightforward solution is to model many similar treatment events simultaneously. If the early adopter effect exhibits a relatively consistent pattern from one treatment event to the next, we can infer its contribution to the overall response, thus inferring the contribution of the treatment itself.

- Convincing CATT estimates require doing a good job on the (in-sample) fit of the treated. Clearly, it is necessary to first accurately fit the *factual* response of the treatment group before attempting to predict its counterfactual response. This requires flexible models that account for user heterogeneity.

- Studying two Product versions, version 9 and 10, and starting with initial naive and unbelievable CCR estimates of 0.824 and 0.720, respectively, we show how our approach produces more reasonable and robust mean CCR estimates of 1.004 and 1.028, respectively. (Recall that a CCR value of exactly 1 indicates no causal effect at all.)

Chapter 3

Model and Design Matrices

We fit our data using variations of a Bayesian hierarchical (mixed effects) model with a Gaussian error distribution. We perform sensitivity analysis of this choice of model class in chapter 5.

Many of the particular models include auto-regressive terms of different AR orders p. In such cases, we use the standard conditional-likelihood approach to building the likelihood function with AR terms; this is justified in our case because (a) our outcome variable, with a reasonable number of AR lags, is essentially stationary, and (b) the modeling has the property that the X matrix, when the AR model is estimated via regression, is invertible. This permits us to regard the $\mathbf{f}_i$ matrix for each user i as a matrix of fixed known constants.

The dimensions of all the quantities listed below are as follows, where p denotes the AR order.

- $\mathbf{y}_i$: a $(T-p)$ by 1 vector of user i's response (UA)
- β_i : a d by 1 vector; in random effects models, d is the length of the random effects coefficients vector, and includes the AR coefficients
- $\mathbf{f}_i$: a $(T-p)$ by d matrix of constants and lagged y_i values; see section 3.1.
- $\mathbf{W}_i$: a $(T-p)$ by $(T-p)$ matrix of fixed known constants (typically week indicators)
- γ : a $(T-p)$ by 1 vector of coefficients of the fixed effects

Our primary working model is a mixed-effects hierarchical model with Gaussian error. For user $i = 1, \ldots, n$ =10,491,859, we have:

$$\begin{aligned}
\mathbf{y}_i &= \mathbf{f}_i\beta_i + \mathbf{W}_i\gamma + \varepsilon_i \\
\beta_i &\sim \mathbf{N}(\mu, \mathbf{\Sigma}) \\
\varepsilon_i &\sim \mathbf{N}(\mathbf{0}, \nu\mathbf{I}_{T-p}) \\
p(\mu) &\sim \mathbf{N}(\mathbf{0}, \kappa_\mu\mathbf{I}_d) \\
p(\gamma) &\sim \mathbf{N}(\mathbf{0}, \kappa_\gamma\mathbf{I}_{T-p}) \\
p(\nu) &\sim \text{Inv-Gamma}(\epsilon/2, \epsilon/2) \\
p(\mathbf{\Sigma}) &\sim \text{Inv-Wishart}_{d+1}(\mathbf{I})
\end{aligned}$$

This model assumes that each Product version affects all users *differently*, i.e., the model treats all users in a *heterogeneous* fashion, and allows room for homogeneous fixed effects common to all users in the $\mathbf{W}_i$ matrix. We assume the error distribution to be Gaussian, which we don't believe to be the case in reality for *any single user*, but we believe it to be a very good model in the aggregate - see Appendix A.

We employ very diffuse (yet proper) priors for μ, γ, and ν, namely $\kappa_\mu = \kappa_\gamma = 10^6$, and $\epsilon = 0.001$. For the prior on the unknown covariance matrix $\mathbf{\Sigma}$, we choose a non-informative proper prior distribution (Gelman et al., 2014) which has the nice feature that each single correlation in the $\mathbf{\Sigma}$ matrix has marginally a uniform prior distribution. We fit the above mixed-effects model using MCMC, specifically Gibbs sampling, since all full conditional distributions are available in closed form - see Appendix B.

3.1 $\mathbf{f}_i$ Matrix

For each user i, the design matrix $\mathbf{f}_i$ contains three sets of covariates: (a) the version indicators, (b) the PCFs that encode waiting time to adopt the latest version, and (c) other user covariates. We detail each of these below.

$$\mathbf{f}_i^{version} = \begin{bmatrix} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 & v_8 & v_9 & v_{10} & v_{11} & v_{12} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.7 & 0.3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ & & & & & .... & & & & & & \\ & & & & & .... & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Table 3.1: A portion of a user's sample $\mathbf{f}_i^{version}$ matrix.

Version Indicators

The $R = 12$ version indicator columns of $\mathbf{f}_i$ denote which specific version (treatment) user i had installed during each of the $T = 52$ weeks. In detail, $\mathbf{f}_i^{version} = [\mathbf{x}_{i,1}^{'}, \mathbf{x}_{i,2}^{'}, \ldots, \mathbf{x}_{i,R}^{'}]^{'}$, where $r = 1, \ldots, R = 12$ is the number of unique Product versions in the study, and each column of version indicators $\mathbf{x}_{i,r}$ is defined as follows: $\mathbf{x}_{i,r} = (x_{i,r,1}, x_{i,r,2}, \ldots, x_{i,r,T})^{'}$, where

$$x_{i,r,t} = \begin{cases} a, & 0 < a < 1, \text{ if user } i \text{ has Product version } r \text{ at time } t; \\ 0, & \text{otherwise} \end{cases}$$

We compute the fraction a of version r usage in week t using session data for that week. A sample $\mathbf{f}_i$ for a user is shown in table 3.1. A few things to note about the sample $\mathbf{f}_i$:

- the user first appeared to use the Product on week 4; he started with version v_6
- he used two different versions, v_6 and v_7, during week 6
- he upgraded to version v_7 during week 6 (spending $5/7 \doteq 0.7$ of the week with v_6 and $2/7 \doteq 0.3$ with v_7), and upgraded to version v_9 during week 10, and he remained on v_9 for the remainder of the time period

t	0_wks	1_wks	2_wks	3_wks	4_wks	5_wks	6_wks	7_wks	8_wks	...
0	.	.	.	.	.	.	.	.	.	...
1	.	.	.	.	.	.	.	.	.	...
2	.	.	.	.	.	.	.	.	.	...
3	.	.	.	**1**	.	.	.	.	.	...
4	.	.	.	.	**1**	.	.	.	.	...
5	.	.	.	.	.	**1**	.	.	.	...
6	.	.	.	.	.	.	**1**	.	.	...
7	.	.	.	.	.	.	.	.	.	...
8	.	**1**	.	.	.	.	.	.	.	...
9	.	.	**1**	.	.	.	.	.	.	...
10	.	.	.	**1**	.	.	.	.	.	...
					....					

Table 3.2: Sample of a user's "n-weeks-past-release" indicator columns $\mathbf{f}_i^{waiting\_time}$. The "." entries represent 0. The horizontal lines indicate weeks of new version releases. The user waited 3 weeks to upgrade to the version released in week 0 ($t = 0$), and waited 1 week to upgrade to the version released in week 7 ($t = 7$).

3.1.1 Waiting Time PCFs

To control for the early adopter effect mentioned above, we construct 14 binary indicator variables called "n-weeks-past-release" indicators. For a given user i in a given week t, we calculate how long ago the current latest version was released, relative to the given week t. Suppose the current latest version was shipped in week t_1, and the given week is t; we calculate n-weeks-past-release$(t) = t - t_1$. We then set the $(t - t_1)$-th indicator variable to 1.

For example, if version a was released at $t = 0$ and the next version b was released at $t = 7$, and if a certain user upgraded to a at time $t = 3$ (waited 3 weeks), and to b at $t = 8$ (waited 1 week), his first 9 n-weeks-past-release indicator columns $\mathbf{f}_i^{waiting\_time}$ for $t = 0$ to $t = 11$ would be given by table 3.2. There are 14 such indicators because that is the maximum number of weeks between consecutive releases.

3.1.2 Other Covariates

Looking at all of our 10.5M users, we have approximately 2.83M users whose first recorded Product usage was during our 52 week window. (This is not to say these users had *never*

used the Product before, but we did not find a record of them using the Product in the 12 months prior to the start of our study). Thus we include a binary indicator covariate called "virgin_user" to denote those users who appeared to use the Product for the first time ever in our study window.

We also add a covariate that captures a user's long term behavior, namely the six-month rolling average of UA over *all of eBay's products*, not just using the Product in the study.

Finally, in order to control for a user's behavior during the one week he or she upgrades, we include a binary indicator covariate ("upgrade-week") for the particular week in which an upgrade occurs.

3.2 $\mathbf{W}_i$ Matrix

From our initial exploratory flat models, we found it necessary to account for time somehow in our models. Models that did not account for time at all or ones that involved a simple linear time variable (t) did very poorly in fitting the aggregate response. We discovered that our best models were those that included an effect for each individual specific week of our 52-week period. Therefore, we included $T - p$ (where p is the AR order) indicator columns as fixed effects as the matrix $\mathbf{W}_i$ (each $\mathbf{W}_i$ is effectively the identity matrix of dimension $T - p$.)

3.3 Counterfactual $\mathbf{f}_i$ Matrix

Throughout our work we will be estimating the counterfactual response $\hat{\mathbf{Y}}_{CF}$, given a certain Product version which we call the *counterfactual version*. Concretely, we will try to estimate the response (UA) if that particular version had not been released, but a "placebo version" had been released in its place. For example, if we take version 9 as our counterfactual version, our counterfactual scenario (alternate universe) is where version 8 (the version immediately preceding 9) is released again instead, on the same date as version 9 actually shipped. Thus we are trying to isolate the effect on UA caused solely by the features of version 9.

$$\mathbf{f}_i^{CF\_version} = \begin{bmatrix} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 & v_8 & v_9 & v_{10} & v_{11} & v_{12} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ & & & & & .... & & & & & & \\ & & & & & .... & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Table 3.3: A portion of a user's sample $\mathbf{f}_i^{CF\_version}$ matrix.

When estimating the counterfactual response, we will need a counterfactual counterpart to the version indicator columns described above, namely $\mathbf{f}_i^{CF\_version}$. A sample one - taking version v_7 as the counterfactual version - for the same sample user above is shown in table 3.3. A few things to note on how $\mathbf{f}_i^{CF\_version}$ was constructed:

- the column corresponding to version v_7 was added to the column corresponding to the user's prior version to produce the latter column; in this case it is version v_6, but it doesn't have to be the one immediately preceding - this is user dependent
- the column for version v_7 was zeroed out
- columns to the right of v_7 and to the left of v_5 were not affected

Note that the counterfactual for a "virgin user" is constructed by assigning such a user to the release immediately preceding the counterfactual release in question.

3.4 Counterfactual Computation

The estimate of the counterfactual $\hat{\mathbf{Y}}_{CF}$ response in our hierarchical mixed-effects model is computed as follows. Given that we have fit the model and run M samples after burn-in, we have the following sets of samples from the posterior distributions above:

- M samples each of μ, $\mathbf{\Sigma}$, γ, and ν.
- $\bar{\beta}_i$: Since we have $n = 10.5$M users in our dataset, and finite memory and disk space, we do not store M samples of each user's d dimensional vector β_i. Instead, we simply store the mean $\bar{\beta}_i$ for each user i, where the mean is taken over the M posterior samples.

Given the above, we calculate the counterfactual estimates as follows. If we are just interested in point estimates, we simply use the point estimates $\bar{\beta}_i$ from the posterior for each user i:

$$\hat{\mathbf{y}}_i^{CF} = \mathbf{f}_i^{CF}\bar{\beta}_i + \mathbf{W}_i\bar{\gamma}$$

If we would like to estimate uncertainty bands around our estimate, we simulate the following. For each user i, we draw β_i^* from the its full conditional given the true $\mathbf{f}_i$ matrix and the posterior means of the other parameters, and then draw $\hat{\mathbf{y}}_i^{CF}$ using the counterfactual $\mathbf{f}_i^{CF}$ matrix:

$$\beta_i^* \sim p(\beta_i|\mathbf{y}_i, \mathbf{f}_i, \bar{\mu}, \bar{\mathbf{\Sigma}}, \bar{\nu}, \bar{\gamma})$$
$$\hat{\mathbf{y}}_i^{CF} \sim \mathbf{N}(\mathbf{f}_i^{CF}\beta_i^* + \mathbf{W}_i\bar{\gamma}, \bar{\nu})$$

We then sum up each user's CF estimate to obtain the aggregate estimate $\hat{\mathbf{Y}}_{CF} = \sum_{i=1}^{n} \hat{\mathbf{y}}_i^{CF}$. Note that our CATT estimates only consider the counterfactual response during the weeks of a release's lifetime, i.e., when it was the latest release on the market. In the case of version 9, this period was from week 30 up to and including week 36, and we make no claims about the counterfactual story thereafter, at which point version 10 comes on the market. The reason for this is as follows. In our counterfactual constructed universe, during the 7 weeks in which version 9 was the latest version, users were shifted onto the release they had immediately prior to version 9 (this varied among users, but the majority were on version 8). When version 10 was released, users who upgraded to version 10 in the true universe were upgraded in the CF universe as well, but users who remained on version 9 in

the true universe were retained on version 8. In the window where version 9 was the latest release on the market (the only game in town, so to speak), this is the only choice available to us. However, when version 10 replaces version 9 as the latest release, we cannot be sure those same users who stuck with version 9 until the end would have also stuck with version 8 until the end.

3.4.1 Counterfactuals in Models with AR(p) Terms

There is a slight twist to computing counterfactual estimates in models that include autoregressive AR(p) terms, as many of our models do. Returning to our sample $\mathbf{f}_i^{CF}$ matrix, suppose we include an AR(1) term in as a random effect. In this case, we have to make an adjustment in the counterfactual computation during the time period in which the counterfactual version was active. Namely, true lagged y values are *replaced* by their (sequentially) *estimated* lagged $\hat{y}$ values, but only during the period of time during which the counterfactual version was used by this user. In the example in table 3.4, $\hat{y}_7^{CF}$ will be computed using the true y_6 because during $t = 6$ the user still had v_6; however, $\hat{y}_8^{CF}$ through $\hat{y}_{11}^{CF}$ will be computed using the predicted (lagged) values of $\hat{y}_7^{CF}$ through $\hat{y}_{10}^{CF}$. This may or may not have a substantial effect on the results, depending on the magnitude of the AR coefficients.

$$\mathbf{f}_i^{CF\_version} = \begin{bmatrix} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 & v_8 & v_9 & v_{10} & v_{11} & v_{12} & y_{-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 4.25 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{\hat{y}_7} \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{\hat{y}_8} \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{\hat{y}_9} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & \mathbf{\hat{y}_{10}} \\ & & & & & .... & & & & & & & \\ & & & & & .... & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Table 3.4: A portion of a user's sample $\mathbf{f}_i^{CF\_version}$ matrix showing the way the counterfactuals are computed in models with AR terms.

Chapter 4

Estimates of Treatment Effects

In order to motivate our methodology, we first demonstrate what happens when one models an individual version release in isolation and also ignores the early adopter effect. Next we show that more reasonable estimates are achieved when one takes the early adopter effect into account, and in order to do so, one must model the entire sequence of version releases. In the following causal effect estimates, we *take version 9 as our counterfactual version* released in week 30 (and replaced in week 37) in all models initially; once we have settled on the "best" model, we apply the same CF estimation technique to version 10 (released in week 37).

4.1 Modeling a Single Version in Isolation

In this scenario, we decide to ignore the early adopter effect completely and decide to estimate the causal effect of version 9 in isolation. Thus, each user's $\mathbf{f}_i$ matrix will not contain the "n-weeks-past-release" indicator columns (the waiting time PCFs). Since the subsequent version, namely version 10, is released in week 37, the time window for our analysis can maximally go through week 36 (+6 weeks). The start of the time window is arbitrary as long as it does not impinge on the previous version, and for symmetry reasons we choose week 24 (-6 weeks). (Note that *both* start and end week choices are arbitrary, leading to different CATT estimates as we show below). In summary, our first *reduced model*

(model M_R^a) includes weeks 24 through 36 only, only versions 1 through 9, and does not include the 14 "n-weeks-past-release" indicators. The results for this are shown in figure 4.1. The estimate of the mean CCR for version 9 from model M_R^a is 0.906, *a relatively large and not very realistic estimate*. It effectively states that without version 9, the UA of the treated would have been around 10% lower in aggregate over weeks 30-36.

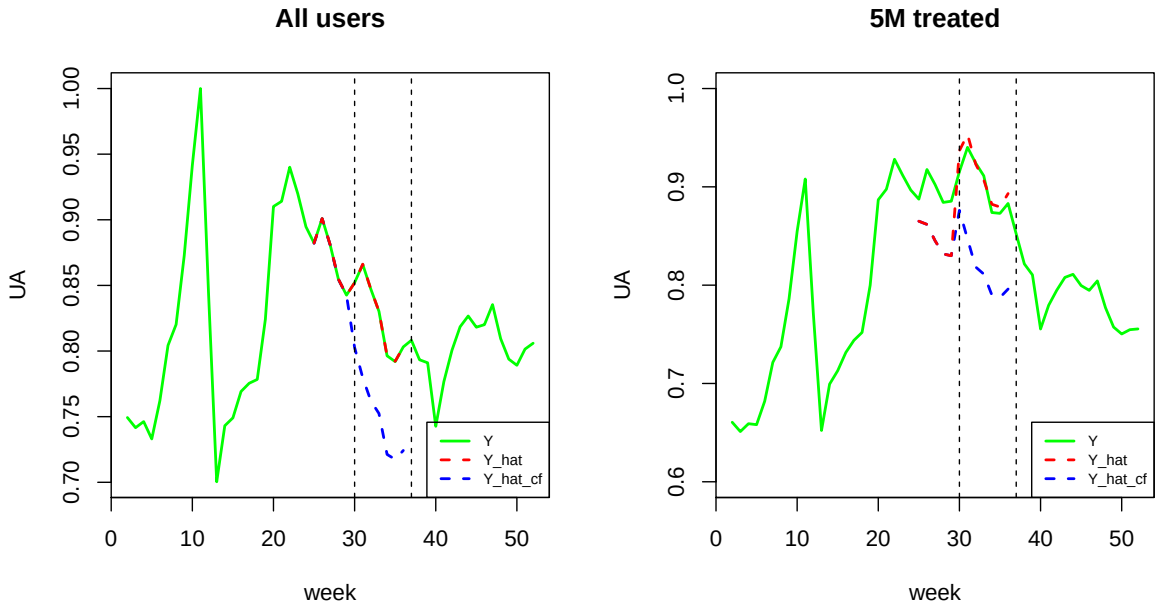

Figure 4.1: Model M_R^a: modeling version 9 in isolation and ignoring the early adopter effect (i.e. n-weeks-past-release indicators not included). The estimate of the mean CCR for **version 9** is **0.906**.

Given the unsatisfactory CATT estimate above, we subsequently decide to model the early adopter effect and include the 14 "n-weeks-past-release" indicators in the model. However, due to computational expediency, we stick with the reduced model, and naively include these waiting time covariates into model M_R^a above, with the other covariates and the -/+6 week time window unchanged. The results for our second reduced model (model M_R^b) are shown in figure 4.2. This has not helped, as the estimate of the mean CCR for version 9 has become more unrealistic, namely 0.878.

Still unhappy with the CATT estimates from model M_R^b, we conclude that in order to nail down the treatment effect, it is necessary to simply extend the time window of our

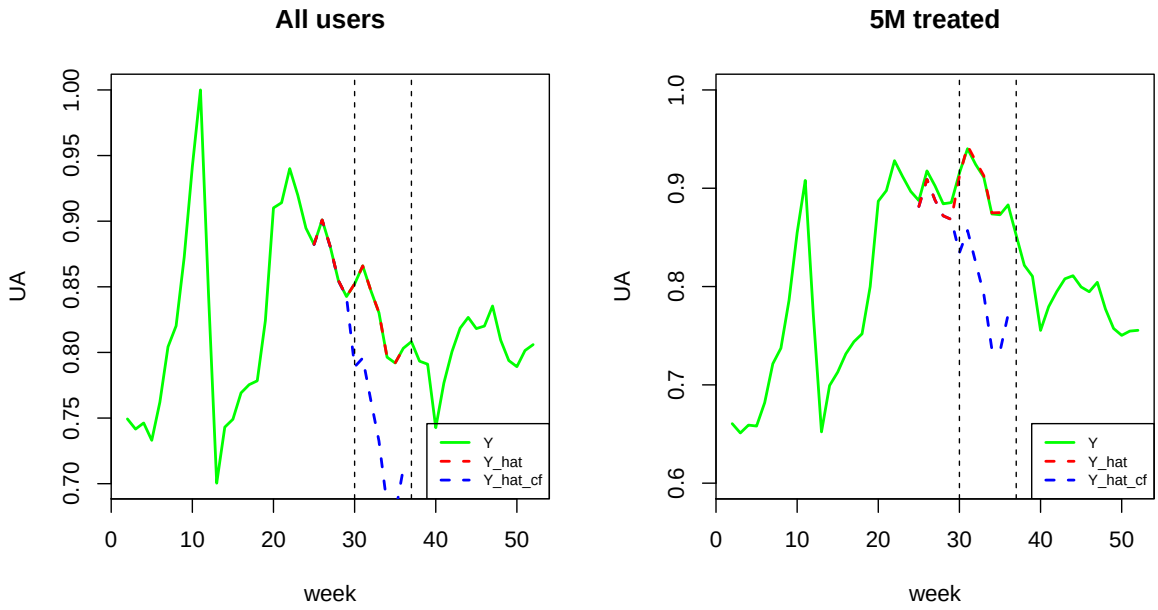

Figure 4.2: Model M_R^b: modeling version 9 in isolation but naively including n-weeks-past-release indicators. The estimate of the mean CCR for **version 9** is **0.878**.

analysis. The earliest week our time window can start is limited by the release week of version 8, namely week 16, since by definition we are treating version 9 in isolation. Thus, our time window for our third reduced model (model M_R^c) will be weeks 17 through 36, with all covariates unchanged from model M_R^b. The results for model M_R^c are shown in figure 4.3. Surprisingly, model M_R^c shows a much different mean CCR estimate of 1.188, *larger in magnitude and in a different causal direction* (CCR greater than 1 as opposed to less than 1 previously). Further perturbation of the time window (not shown), namely weeks 21 through 36, yields a similarly large (and strange) mean CCR estimate of 1.35.

It seems that modeling version 9 in isolation *and* naively incorporating the early adopter effect (via the "n-weeks-past-release" indicators), leads to very volatile estimates of the treatment effect. CCR estimates in these models are highly sensitive to the choice of the time window of the model. Why is that the case?

The reason for this has to do with the *identifiability* of the model. A reduced model, which considers a release in isolation *and* which includes the "n-weeks-past-release" covariates to account for the early adopter effect, runs the risk of being non-identifiable because

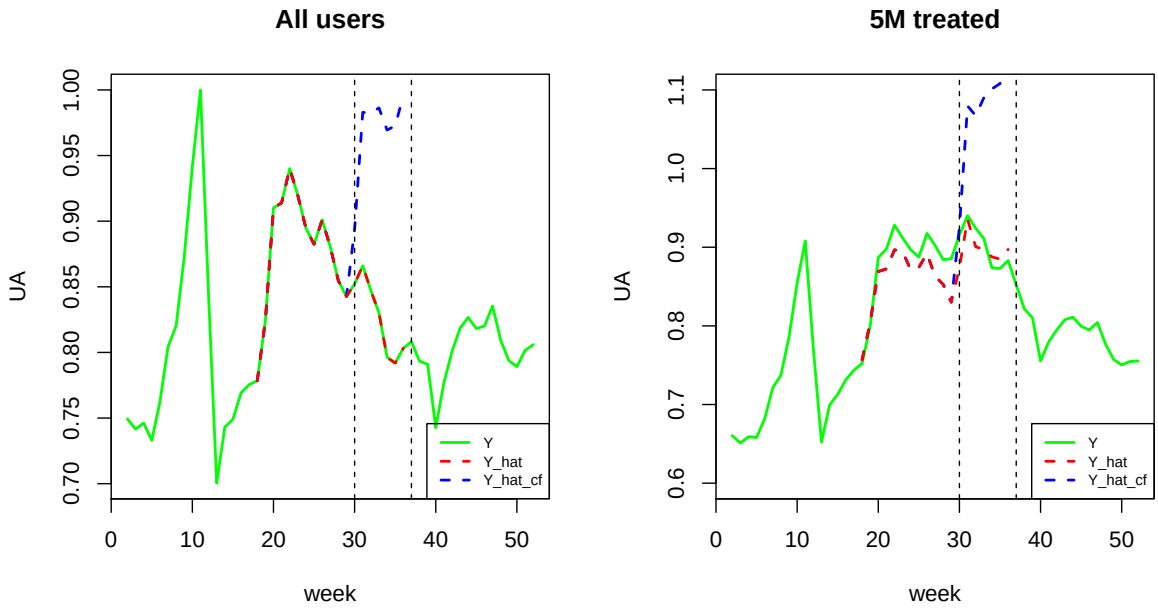

Figure 4.3: Model M_R^c: modeling version 9 in isolation, naively including n-weeks-past-release indicators, and extending the modeling time window to weeks 17-36 (from weeks 24-36 as in figure 4.2). The estimate of the mean CCR for **version 9** is now **1.188**, very different from the estimates obtained from model M_R^b in figure 4.2.

by construction the treatment indicator column becomes a linear combination of the "n-weeks-past-release" indicator columns. In fact, in model M_R^b (weeks 24-36), the (Pearson) correlation coefficient between the version 9 indicator column and the *sum* of the first 7 "n-weeks-past-release" columns is **0.984**. The correlation is computed over all 10.5M users for the 12-week time window (13 weeks minus 1 for AR(1)); the constructed 126M x 1 dimensional vectors differ in fewer than 0.7% locations. Clearly models with such collinear covariates are "on the edge of identifiability" so to speak, leading to wildly varying estimates of parameters and consequently treatment effects. Furthermore, this makes it is almost impossible to isolate the early adopter effect from the version treatment effect. Nevertheless, the early adopter effect is real (recall figure 2.3) and must be accounted for.

4.2 Modeling All Versions Jointly

The straightforward approach to deal with the identifiability problem that arises when modeling a single treatment individually is to model all treatments jointly in a *full model*, and include covariates (PCFs) that explicitly encode each user's waiting time to adopt the treatment, i.e. the "n-weeks-past-release" indicators. Pooled data from other treatments reduces the collinearity in the design matrices because *different users play the role of early adopters in each release*. In order to get an identifiable model, we are counting on the fact that the set of early adopters varies from release to release. This is indeed the case: computing the correlation coefficient between the same covariate vectors in the full model (as we did in reduced model) shows a value of only 0.216, meaning that there is little overlap between sets of early adopters from one version to the next. This allows for robust estimates of the early adopter effect, i.e. of the values of the "n-weeks-past-release" coefficients, which in turn produces robust estimates of the treatment indicator coefficients, which then leads to a realistic CATT estimate. These results are shown in figure 4.4. The estimated **mean CCR value for version 9 is 0.998**, a negligible causal effect and much more in line with expectations. This estimate is also relatively robust to model perturbation - see chapter 5.

The early adopter effect can also clearly be seen in figures 4.5 and 4.6. Figure 4.5 shows the posterior mean of the components of μ that correspond to the 12 versions (i.e. the version coefficients) resulting from two models: first, a model that ignores the early adopter effect and excludes the "n-weeks-past-release" indicators, and second, a model that does include the "n-weeks-past-release" indicators. Note that in the first model, the version coefficients show an increasing trend because the early adopter effect is confounded with the treatment effect. (Version 12's coefficient does not follow the trend because version 12 appears in only the last week in our study (week 52); in this case, the "upgrade-week" indicator accounts for the first week adopter effect). The confounding occurs because given a fixed 52-week time window, the fewer weeks a Product version has been on the market in the time window, the larger its proportion of early adopters is as a fraction of its total users. This has the effect of artificially driving up the version coefficient in the model, making it

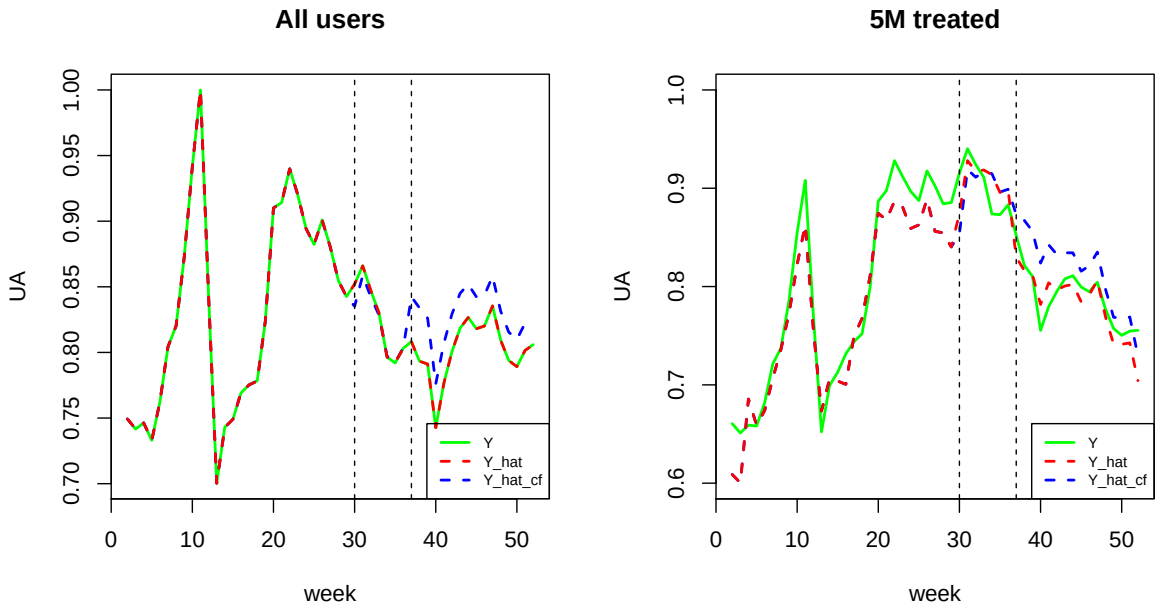

Figure 4.4: Modeling all releases (treatments) jointly; AR(1) model. The estimate of the mean CCR for **version 9** is **0.998**.

appear as if the version had a large positive effect on UA. An obvious way to control for this effect is via the 'n-weeks-past-release" indicators, whose posterior mean values are shown in figure 4.6.

4.3 Simulating the Counterfactual

We estimate the uncertainty bands around our estimate by simulating the following. For each user i, we draw β_i^* from the its full conditional given the true $\mathbf{f}_i$ matrix and the posterior means of the other parameters, and then draw $\hat{\mathbf{y}}_i^{CF}$ using the counterfactual $\mathbf{f}_i^{CF}$ matrix:

$$\beta_i^* \sim p(\beta_i | \mathbf{y}_i, \mathbf{f}_i, \bar{\mu}, \bar{\mathbf{\Sigma}}, \bar{\nu}, \bar{\gamma}) \tag{4.1}$$

$$\hat{\mathbf{y}}_i^{CF} \sim \mathbf{N}(\mathbf{f}_i^{CF}\beta_i^* + \mathbf{W}_i\bar{\gamma}, \bar{\nu}) \tag{4.2}$$

A graph of the uncertainty bands for the counterfactual of version 9 are shown in figure 4.7.

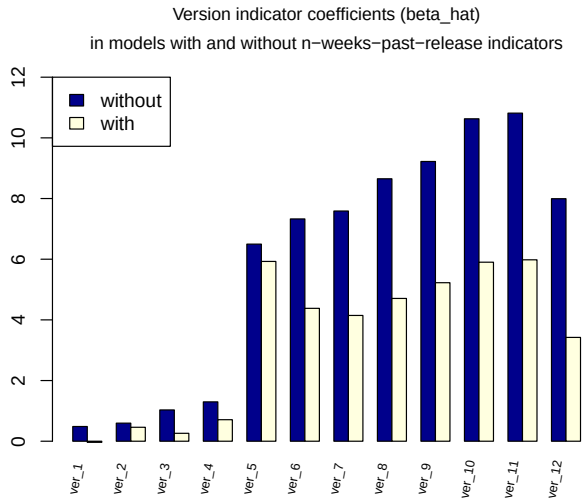

Figure 4.5: Posterior means of the version coefficients in models with and without the "n-weeks-past-release" indicators. Including the "n-weeks-past-release" indicators in the model eliminates the increasing trend in version coefficients and isolates the treatment effect.

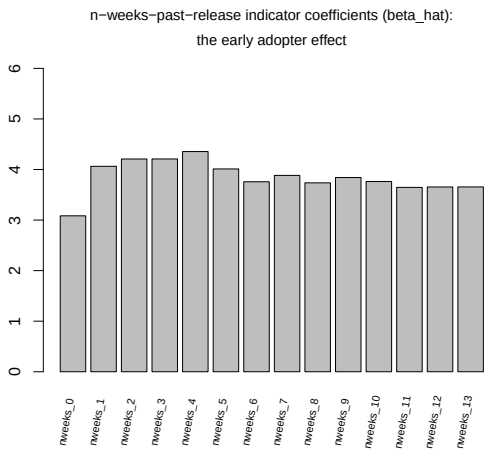

Figure 4.6: Posterior means of the "n-weeks-past-release" indicators. The magnitude of these coefficients shows that a significant portion of the response can attributed to how long a user waited to take the adopt the treatment.

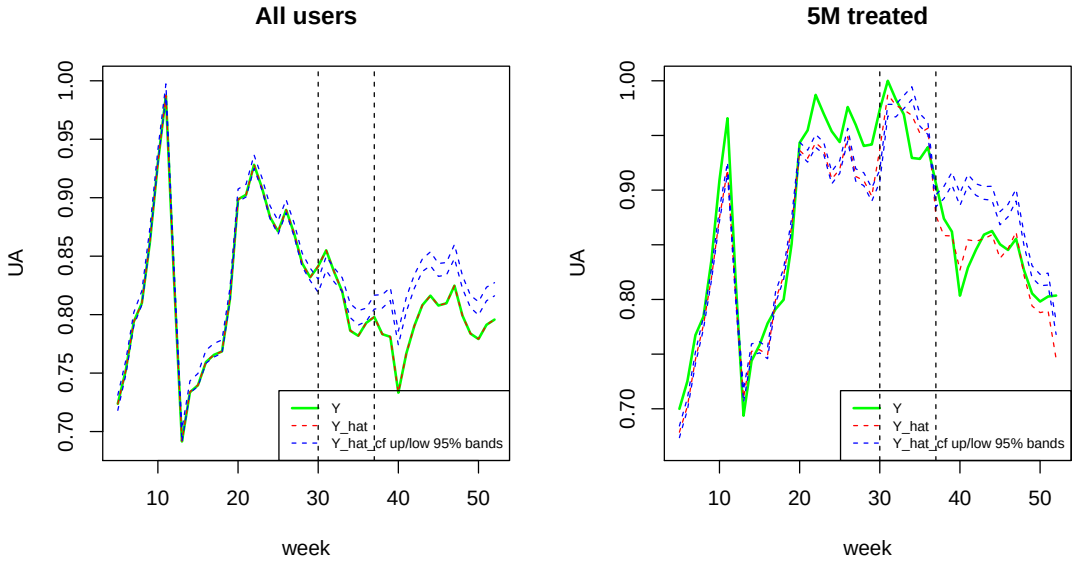

Figure 4.7: Hierarchical model AR(4) model: the 95% uncertainty bands for CCR for version 9 are [0.997,1.010].

Chapter 5

Model Selection and Validation

Our goals in model selection and validation are twofold: (a) choose the single "best" model so that we can efficiently estimate the CCR for other versions, and (b) make sure that our results are relatively robust (i.e. no over-fitting). Therefore, we study the sensitivity of our results to different AR orders, as well as investigate two additional model classes: a simpler flat model with a Gaussian error distribution, and a more complex model with a non-parametric error distribution. Finally, we perform 5-fold cross-validation to verify that our results are not sensitive to some random subset of our users.

Our primary selection criterion for a "best" model is the *the fit-of-the-treated*, namely the *in-sample* root-mean-squared-error (RMSE) on the *treatment group*. The reasons for this are two-fold. First, we are interested in computing CATT: the effect on the treated group. There are a number of models that at first glance do very well fitting the aggregate response over a *random out-of-sample* (OOS) subset of users. However, on closer inspection, one finds that those same models do a very poor job at fitting the aggregate response of a very non-random group of individuals - the treatment group - even on an in-sample subset! Such models are useless to us since they do not accurately model the real behavior of the treated over the course of our study. Second, we are (currently) not interested in predicting aggregate UA into the *future*, nor are we interested in predicting how another totally different set of users would behave. We are focused on accurately modeling *our in-sample* 10.5M users (and the approximately 5M treated ones in particular) so that we

can ascertain what *their* behavior would have been like in the counterfactual world. A hypothetical model that had great predictive power on a different set of users would also be of dubious value if it could not accurately model *our* treatment group.

The secondary selection criteria is scalability (computational ease-of-fit). Given that we are doing Bayesian inference over 10.5M users in an approximately 30-dimensional space, the ability to fit the model in a reasonable amount of time is important. If a model takes an inordinate amount of time to fit, its performance advantage over competing models should be commensurate with this increased time and effort.

5.1 AR Order Sensitivity

Table 5.1 shows the results for three identical hierarchical models where we perturbed the AR order: $p = \{0, 1, 4\}$. We see that the overall estimates of the causal effect - as measured by the posterior mean of CCR for version 9 - are not very sensitive to the AR order. However, since our goal is to minimize the RMSE on the treated users, we see that the AR(4) model performs better than the AR(0) and AR(1) models (with minimal computational overhead), and so we elect to proceed with the AR(4) model in all subsequent analyses.

Model Class	Error Dist	AR order	RMSE Before	RMSE During	Mean CCR
hierarchical	Normal	0	57.88	68.61	0.984
hierarchical	Normal	1	56.55	66.96	0.998
hierarchical	Normal	4	53.64	63.35	1.004

Table 5.1: The table shows the RMSE $[\frac{1}{n_T}(\hat{\mathbf{Y}}_T - \mathbf{Y}_T)^2]^{1/2}$ for the *treated* users over 7 weeks before and 7 weeks after the release of version 9. Note that the (posterior) mean CCR results are not very sensitive to the AR order of the model, but more heterogeneity (i.e. a larger AR order) leads to a better fit of the treated.

5.2 Sensitivity to Model Class

Having selected an AR(4) hierarchical model as our "best" candidate so far, we now study results from a simpler flat model with a Gaussian error distribution and a more complex model with a non-parametric error distribution.

First, the flat model with Gaussian error assumes that each Product version affects all

users equally, i.e., the model treats all users in a *homogeneous* fashion by having a single β parameter for all users. It is basically an ordinary least-squares (OLS) regression. In the flat OLS model, for user $i = 1, \ldots, n$ =10,491,859, we have:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{f}_i\beta + \mathbf{W}_i\gamma + \varepsilon_i \\ \varepsilon_i &\sim \mathbf{N}(\mathbf{0}, \nu\mathbf{I}_{T-p}) \\ p(\beta, \gamma, \nu) &\propto 1 \end{aligned}$$

We do not consider the OLS model a true reflection of reality, because we don't believe all 10.5M users can be treated in a homogeneous fashion. However, it is a limiting case of our hierarchical model in which $\Sigma \to \mathbf{0}$, and thus a good candidate for analysis. We assume an (improper) diffuse prior for all the parameters. All of these marginal prior choices are standard for a diffuse prior except the marginal for ν, which would usually be Jeffreys prior $p(\nu) \propto \nu^{-1}$. However, with n on the order of 10 million, it doesn't matter if you take $p(\nu) \propto 1$, or $p(\nu) \propto \nu^{-1}$, or $p(\nu)$ equal to a point mass on the MLE, so we have user $p(\nu) \propto 1$ for computational convenience.

Second, the non-parametric error model below is an extension of our current hierarchical model, in that it allows the error distribution to have a (more realistic) non-parametric form, namely a Dirichlet process (DP) mixture of Gaussians. Specifically, the ε_i errors for each user come from a Dirichlet process mixture model, and each ε_i is the same for all weeks t, i.e. the $\varepsilon_{i,t}$ are $(T - p)$ IID draws from $\mathbf{N}(\theta_i, \nu)$. The idea is that the non-parametric DP (location) mixture form of the error distribution will do a better job fitting the small percentage of high activity users (as measured by UA units). We fit the above model using the *marginal Gibbs sampler* (Gelman et al., 2014). In the non-parametric error model, for

user $i = 1, \ldots, n$ =10,491,859, we have:

$$
\begin{aligned}
\mathbf{y}_i &= \mathbf{f}_i\beta_i + \mathbf{W}_i\gamma + \varepsilon_i \\
\varepsilon_{i,t} &\sim \mathbf{N}(\theta_i, \nu) && t = (p+1)\ldots, T \\
\theta_i &\sim \mathcal{P} \\
\mathcal{P} &\sim \mathrm{DP}(\alpha, \mathcal{P}_0) && \alpha = 3 \text{ (fixed)} \\
\mathcal{P}_0 &= \mathbf{N}(0, \tau_0) && \tau_0 = 100 \text{ (fixed)} \\
\beta_i &\sim \mathbf{N}(\mu, \mathbf{\Sigma}) \\
p(\nu, \mu, \mathbf{\Sigma}, \gamma) &: \text{as above in chapter 3}
\end{aligned}
$$

The results for different model classes are shows in table 5.2; we have included the Normal hierarchical model as well for comparison. We see that the given the same covariates, all three models produce reasonable CCR estimates. However, note that the flat model does very poor job at fitting the treated users. Figure 5.1 graphically illustrates this lack of in-sample fit to the treated. Without the benefit of similar CCR estimates from the hierarchical model, its CCR estimate of 1.024 would be *very hard to believe given such a poor fit to the treated*. This illustrates the importance of taking user heterogeneity into account, especially when one is trying to estimate the causal effect on the treated.

At the other end of the model complexity spectrum, table 5.2 shows that the non-parametric error distribution model did not have a significant impact on the results. The RMSE on the treated set of users improved very little, and the mean CCR moved from 1.004 to 1.011. However, this came at a great computational cost - it took 2-3 times the amount of time to fit the DP mixture model as it did the Gaussian hierarchical model with the same exact covariates, while not producing very different CCR estimates.

Model Class	Error Dist	AR order	RMSE Before	RMSE During	Mean CCR
flat	Normal	4	89.46	96.62	1.024
hierarchical	Normal	4	53.64	63.35	1.004
hierarchical	DP-mixture	4	53.34	62.75	1.011

Table 5.2: The table shows the RMSE $[\frac{1}{n_T}(\hat{\mathbf{Y}}_T - \mathbf{Y}_T)^2]^{1/2}$ for the *treated* users over 7 weeks before and 7 weeks after the release of version 9. Note that although the flat (homogeneous) model yields similar (posterior) mean CCR estimates, it exhibits a poor fit to the true response of the treated (see figure 5.1). The non-parametric error model also produces similar results, but at a very high computational cost.

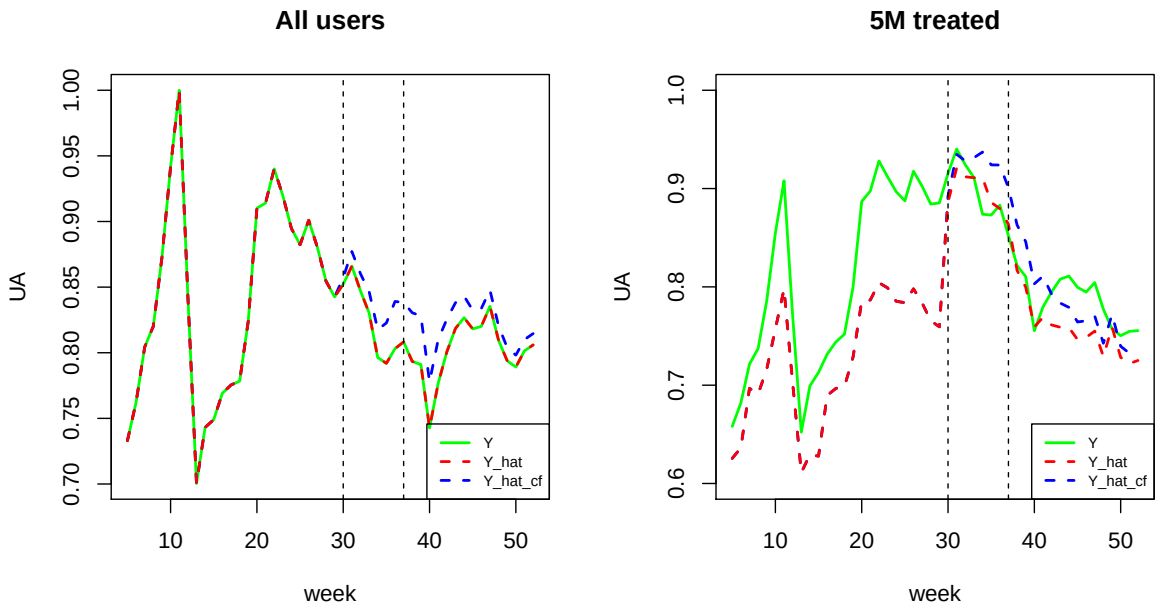

Figure 5.1: Flat AR(4) model. Note the poor fit to the true response of the treated on the right hand side. The estimate of the mean CCR for **version 9** is **1.024**.

5.3 5-fold Cross Validation

Given that the hierarchical AR(4) model is our chosen model, we performed 5-fold cross-validation to check its "out-of-sample" (OOS) predictions. We fit 5 different variants of it using only 80% of users each time, and we examined the fit and computed the CCR estimates on the remaining 20% of users. We did this to ensure that our results were not sensitive to any particular subset of the users.

Table 5.3 shows the estimated CCR for the treated users in each of the 5 folds. Each fold had approximately 2.1M out-of-sample users (and approximately 1M OOS treated users). As we can see from table 5.3, each of the 95% uncertainty bands includes 1, and the mean posterior CCR in each fold is similar to the one obtained using the whole set of users (see for example table 5.2.).

Fold	N Treated Users	Upper 95%	Mean CCR	Lower 95%
1	1015302	0.994	1.006	1.019
2	1014554	0.997	1.009	1.023
3	1015416	0.987	1.001	1.014
4	1013886	0.994	1.007	1.020
5	1014807	0.987	1.001	1.014

Table 5.3: Simulated means and 95% uncertainty bands for the CCR for version 9 for five different held-out sets of treated users.

To further demonstrate the advantages of the hierarchical model, we compare the ability of the flat and the hierarchical models to predict the OOS aggregate response for a relatively small set of users, namely *just 1000 users* drawn randomly from the larger 2.1M OOS set. The two models under comparison have exactly the same covariates. The results of this comparison are shown in figures 5.2 and 5.3. Figure 5.2 shows the fit of the flat model on the set of 1000 users, whereas figure 5.3 shows the fit of a hierarchical model on that same set of 1000 users. Note how much better the hierarchical model is able to capture the aggregate user response compared to the flat model in figure 5.2, and this improved fit is important in obtaining more accurate counterfactual estimates. As we have mentioned before, one must *first accurately fit the factual*, true response *before one can being to predict the counterfactual* response, and doing so requires flexible (hierarchical) models that account

for user heterogeneity.

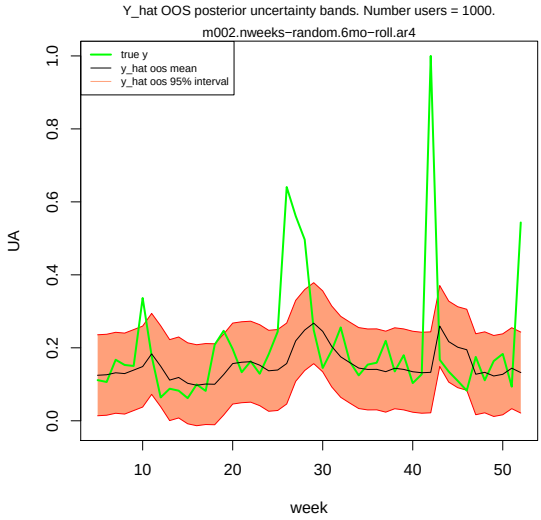

Figure 5.2: Model fit results on 1000 OOS users using *flat* model. The flat model, using the same covariates as its hierarchical counterpart in figure 5.3, cannot capture the nuances (non-Normality) of the aggregate response for a small sample of users.

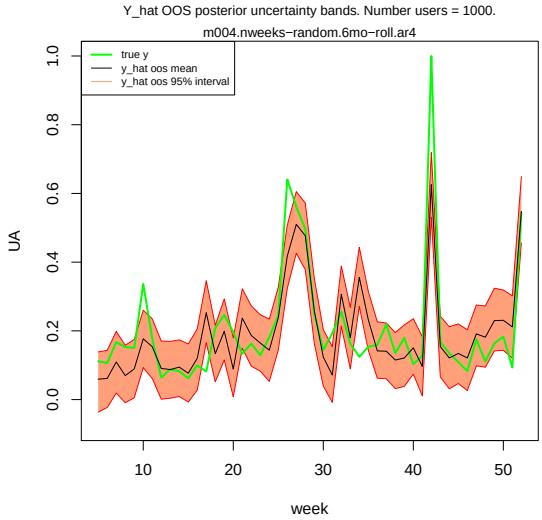

Figure 5.3: Model fit results on 1000 OOS users using *hierarchical* model. Note how much better the hierarchical model is able to capture the aggregate user response compared to the flat model in figure 5.2.

Chapter 6

Conclusion and Summary of Results

Table 6.1 summarizes all of the CCR estimates for each model we have examined and provides 95% posterior intervals on the CCR for two version releases, version 9 and 10, using our chosen "best" model. We see that careful consideration of long term patterns of user behavior combined with flexible models yield causal estimates that are more reasonable and much different than initial naive estimates. The same results are graphically displayed in figure 6.1 for version 9 and in figure 6.2 for version 10. Figure 6.3 shows how well our best hierarchical model fits the average-UA-per-user-per-week curves shown originally in figure 2.3. This in-sample fit quality is important, because without it, it would be difficult to believe any of our counterfactual estimates.

Version	Model	Lower 95%	Mean CCR	Upper 95%
ver. 9	Null	NA	0.824	NA
ver. 9	Hierarchical AR(4), Normal errors	0.998	1.004	1.010
ver. 10	Null	NA	0.720	NA
ver. 10	Hierarchical AR(4), Normal errors	1.022	1.028	1.035

Table 6.1: Summary of our casual effect estimates for version 9 and version 10. The "Null" model is a just a naive comparison of means for the treated group in a (arbitrary) 6-week time window before and after the respective version release.

In conclusion, we have shown that jointly modeling all treatment events (version releases) using a flexible hierarchical Bayesian model produces realistic estimates of the causal effect

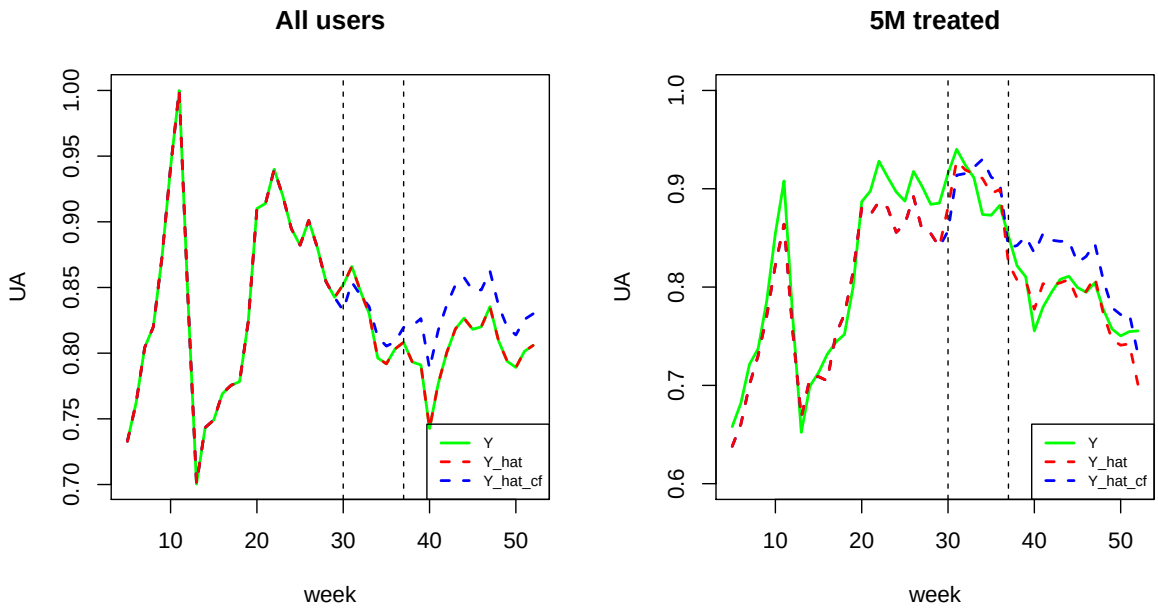

Figure 6.1: Hierarchical AR(4) model: the estimate of the mean CCR for **version 9** is **1.004**.

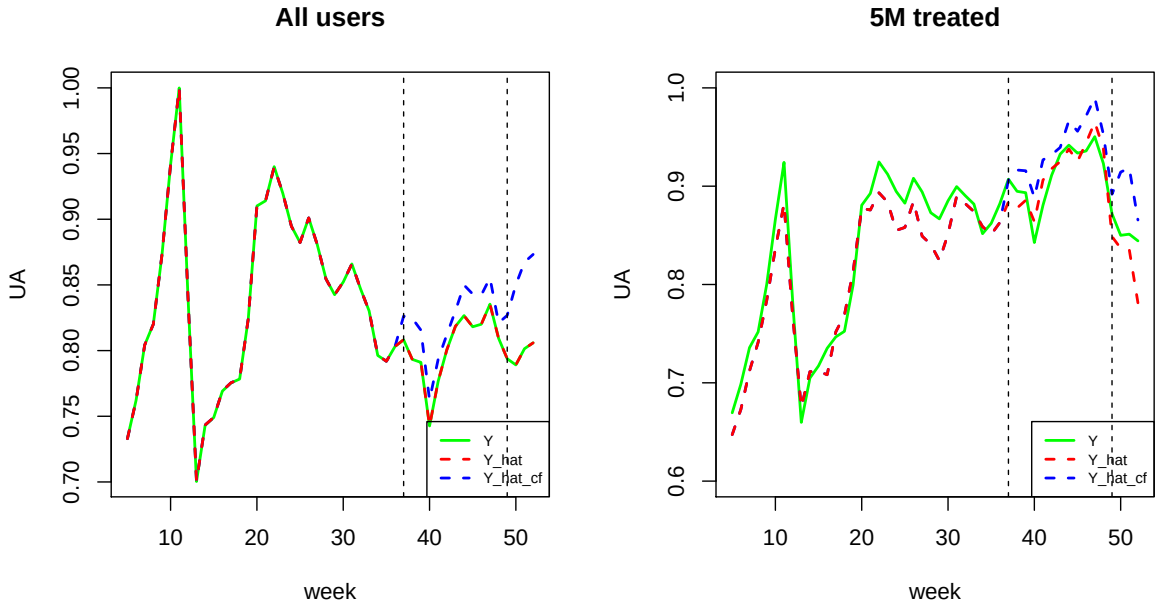

Figure 6.2: Hierarchical AR(4) model: the estimate of the mean CCR for **version 10** is **1.028**.

on the treated. The random-effect component of our model takes user heterogeneity into account, which is crucial to producing good estimates of the true factual response of the treated users. On the other hand, the hierarchical nature of our model, which shrinks all users' parameters to a common mean, allows us to borrow strength across multiple treatment events. It is this sharing of information that allows us to isolate the early adopter effect from the treatment effect, resulting in robust CATT estimates. We close by cautioning that causal inference in observational studies is very difficult by nature, and it is very easy to reach false causal conclusions using overly simplistic models or by ignoring the presence of the early adopter effect when it clearly exists.

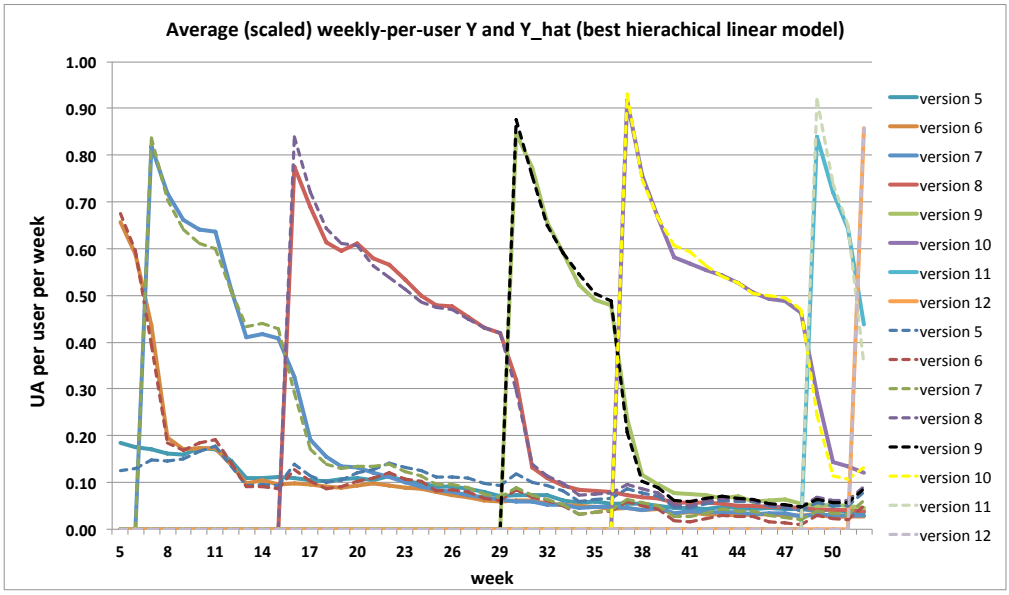

Figure 6.3: Comparing the true average UA per user per week (solid Y) for each version to the estimated $\hat{Y}$ (dashed) from the best hierarchical model.

Appendix A

Assumption Validation

A.1 Assessing the (Weak) Overlap Assumption

For CATT estimates, it is sufficient that we verify $\Pr(W = 1|X) < 1$ (Heckman et al., 1997). To do so, we build a simple (linear) logistic regression model of the treated and control groups using our best covariates. We collapse all of the n-weeks binary indicators into a single "summed" integer value in the range of $[0, 13]$. The estimated density plot of the resulting fitted probabilities from the logistic regression model is shown in figure A.1. We can see that for the most part, $\Pr(W = 1|X) < 1$ is indeed true, except for a very few cases. How few? If we look at the quantiles of the fitted probability scores, we see the distribution in table A.1. We see that for the *vast* majority of treated users the assumption holds; it fails in only 0.0069% of the treated users.

20%	30%	50%	70%	90%	95%	96%	99.90%	99.99%
0.00004	0.00005	0.00012	0.15333	0.41694	0.60879	0.68938	0.95315	0.99991

Table A.1: Quantiles of $\Pr(W = 1|X)$. Fewer than 0.01% (0.0069% to be exact) of the treated violate the $\Pr(W = 1|X) < 1$ condition.

A.2 Gaussian Error Assumption

Except for the model with the non-parametric error distribution, we have tacitly assumed that the error distribution of the response $\mathbf{y}_i$ for any *single user* is Gaussian, *even though*

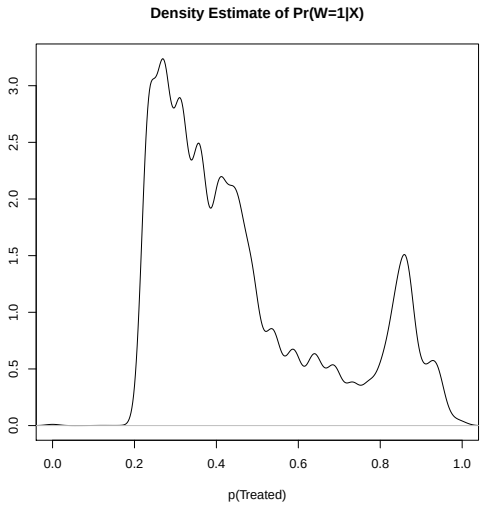

Figure A.1: Density estimate of linear model of propensity score, i.e. the probability of being in treatment group conditional on the covariates.

we know full well it is not. We do not rely on the Normality of the errors for any of our estimates above, *except* when we simulate the counterfactual 95% uncertainty bands as shown in figure 4.7 and in table 6.1 above. The simulation of the bands as shown in equation 4.2 relies on the assumption of Normality. We believe that the aggregate (and thus the mean) of the response does indeed follow an approximately Normal distribution, which allows us to estimate 95% bands for the aggregate using equation 4.2.

We believe this because of the Central Limit Theorem coming into play in our dataset. We run the non-parametric bootstrap algorithm (Efron, 1979) over a number of subsets of the observations to see if indeed the mean of the bootstrapped samples follows an approximately Normal distribution. (We discuss how the subsets were selected below.) We perform the following bootstrap procedure for each of 5 subsets of observations. For each subset S_r of size r observations:

- For $m = 1, \ldots, M = 1000$, do:

    1. Draw a random sample of size r, with replacement, from the set of r responses in subset S_r.
    2. Calculate and store the mean μ_m of the sample.

- Compare statistics over the M bootstrapped means μ_m to Normal distributions.

If we can show that the distribution of mean of the response is approximately Normal given our sample sizes, then we will have support for using our models to estimate uncertainty bands for our aggregate estimates. The primary moments of interest are the skewness and the kurtosis, since the original response is highly skewed and has very large kurtosis values. The results are shown in table A.2 below and pictorially in figure A.2. We can see that indeed the CLT has come into play - in all cells, all of the huge skewness and kurtosis values have been driven down to values approaching a standard Normal distribution.

	Skewness		**Kurtosis - 3**	
r observations	Raw Response $\mathbf{y}_i$	Bootstrapped Mean	Raw Response $\mathbf{y}_i$	Bootstrapped Mean
30.3M	351.1	0.0676	338728	0.0430
9.8M	130.6	-0.0203	49898	-0.1796
5.6M	199.0	-0.0319	72529	-0.0558
3.0M	737.4	0.4890	838477	0.5528
1.2M	50.4	0.0726	5173	0.2240

Table A.2: Bootstrapped estimates of the skewness and kurtosis of the mean of the response $\mathbf{y}_i$ for various numbers of observations. The huge skewness and kurtosis values have been driven down to values approaching those from a standard Normal distribution. The Central Limit Theorem is in effect.

We select the subsets S_r as follows. Except for the AR(p) terms and the 6-month-rolling-average term, all of the remaining covariates are binary indicators. Focusing on the binary covariates for the moment, we can easily cross-tabulate the binary covariates against each other by computing cross-products of our design matrices. Each cell $[i, j]$ of this cross-product will be equal to the number of data points (observations) that have a 1 for both indicators i and j. Since we are primarily interested in the main effects (we are not considering interactions between versions and/or weeks), we focus on the diagonal cells of the above cross-product matrix. We take 5 representative cells of the following sizes: 30.3M units, 9.8M units, 5.6M units, 3.0M units, and 1.2M units (we have no diagonal cell with fewer than 1.2M observations).

As a final note: our model with a non-parametric error distribution showed only minor differences in the CCR estimates, further lending credence to the fact that although the DP

mixture model may be more realistic at the individual user level, its results are very similar to a simpler Normal model when considering the aggregate response.

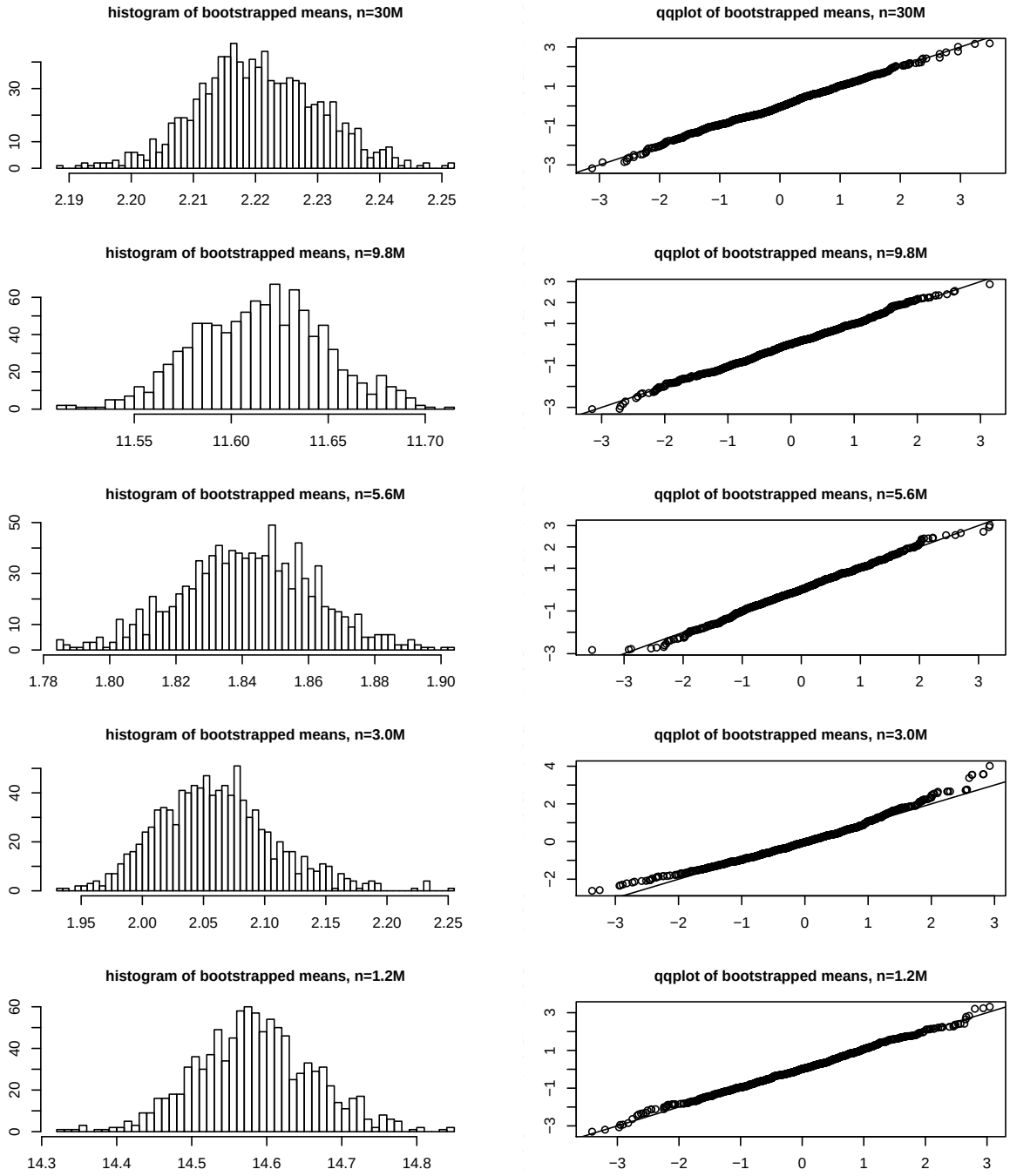

Figure A.2: CLT in action: the distributions of the bootstrapped means look very similar to a Normal distribution, save for the case with 3.0M cells, where the third and fourth raw sample moments were so large that 3.0M observations was not quite enough to drive them both close to 0. Nevertheless, it is quite impressive that the CLT drove kurtosis down from 838K to a value less than 1 - see table A.2.

Appendix B

MCMC Sampling Equations

B.1 Hierarchical model: Gaussian Errors

$$
\begin{aligned}
p(\beta_i|\mathbf{y}_i, \mathbf{f}_i, \mu, \mathbf{\Sigma}, \nu, \gamma) &= \mathbf{N}(\beta_i; \mathbf{m}_i, \mathbf{C}_i) \\
\mathbf{C}_i &= (\mathbf{\Sigma}^{-1} + \nu^{-1}\mathbf{f}_i^{'}\mathbf{f}_i)^{-1} \\
\mathbf{m}_i &= \mathbf{C}_i[\nu^{-1}\mathbf{f}_i^{'}(\mathbf{y}_i - \mathbf{W}_i\gamma) + \mathbf{\Sigma}^{-1}\mu] \\
p(\mu|\beta_1, \beta_2, \ldots, \beta_n, \mathbf{\Sigma}, \kappa_\mu) &= \mathbf{N}(\mu; \mathbf{a}, \mathbf{B}) \\
\mathbf{B} &= [(\kappa_\mu \mathbf{I})^{-1} + n\mathbf{\Sigma}^{-1}]^{-1} \\
\mathbf{a} &= \mathbf{B}(n\mathbf{\Sigma}^{-1}\bar{\beta}) \\
p(\gamma|\beta_1, \beta_2, \ldots, \beta_n, \nu, \mathbf{F}, \mathbf{Y}, \mathbf{W}, \kappa_\gamma) &= \mathbf{N}(\gamma; \mathbf{c}, \mathbf{D}) \\
\mathbf{D} &= [\nu^{-1}\mathbf{W}'\mathbf{W} + (\kappa_\gamma \mathbf{I})^{-1}]^{-1} \\
\mathbf{c} &= \nu^{-1}\mathbf{D}\mathbf{W}^{'}(\mathbf{Y} - \mathbf{F}\mathbf{B}) \\
\mathbf{W}'(\mathbf{Y} - \mathbf{F}\mathbf{B}) &= \sum_{i=1}^{n} \mathbf{W}_i^{'}(\mathbf{y}_i - \mathbf{f}_i\beta_i) \quad \text{and} \quad \mathbf{W}'\mathbf{W} = \sum_{i=1}^{n} \mathbf{W}_i^{'}\mathbf{W}_i \\
p(\mathbf{\Sigma}|\mu, \beta_1, \beta_2, \ldots, \beta_n) &= \text{Inv-Wishart}_{n+d+1}(\mathbf{S} + \mathbf{I}) \\
\text{where } \mathbf{S} &= \sum_{i=1}^{n} (\beta_i - \mu)(\beta_i - \mu)' \\
p(\nu|\mathbf{Y}, \mathbf{F}, \beta, \gamma) &= \text{IG}\Big[\nu; \frac{\epsilon + n(T-p)}{2}, \frac{\epsilon}{2} + \frac{1}{2}\sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{f}_i\beta_i - \mathbf{W}_i\gamma)^{'}(\mathbf{y}_i - \mathbf{f}_i\beta_i - \mathbf{W}_i\gamma)\Big]
\end{aligned}
$$

B.2 Hierarchical model: DP-Mixture Model for Errors

The sampling algorithm is based on the *marginal Gibbs sampler* (Gelman et al., 2014), which separately updates the allocation of users to clusters, and the cluster-specific parameters, as follows:

1. **Update cluster allocation**:

   - For $i = 1, \ldots, n$ users, compute the probability user i belongs to one of the existing k clusters, or to a totally new cluster:

     – For $c = 1, \ldots, k$ clusters and the potentially new cluster $(k+1)$, compute the probability of each cluster as follows, where $n_c^{(-i)}$ is the number of users in cluster c excluding user i, and $k^{(-i)}$ is the number of clusters that exist if user i is not in any cluster:

$$\pi_i \equiv p(u_i = c) \propto \begin{cases} n_c^{(-i)} \mathbf{N}(\varepsilon_i; \theta_c, \nu) & c = 1, \ldots, k^{(-i)} \quad \text{(B.1)} \\ \alpha \int \mathbf{N}(\varepsilon_i; \theta, \nu) \mathbf{N}(\theta; 0, \tau_0) d\theta & c = k^{(-i)} + 1 \quad \text{(B.2)} \end{cases}$$

     – Update the user's cluster membership by sampling from the $(k+1)$ dimensional multinomial: $c_i \leftarrow \text{Multinom}(\pi)$

2. **Update cluster parameters**:

   - For $c = 1, \ldots, k$ clusters, sample updated values of θ_c:

$$\theta_c \sim p(\theta_c) \propto \mathbf{N}(\theta_c; 0, \tau) \prod_{i:u_i \in c} \prod_{t=p+1}^{T} \mathbf{N}(\varepsilon_{i,t}; \theta_c, \nu) \tag{B.3}$$

In detail, equation (B.1), the likelihood for an existing cluster c, is:

$$\begin{aligned} n_c^{(-i)} \mathbf{N}(\varepsilon_i; \theta_c, \nu) &= n_c^{(-i)} \prod_{t=p+1}^{T} \mathbf{N}(y_{i,t} - \mathbf{f}_{i,t}' \beta_i - \mathbf{W}_i \gamma; \theta_c, \nu) \\ &= n_c^{(-i)} \prod_{t=p+1}^{T} \frac{1}{\sqrt{2\pi\nu}} \exp\Big\{ -\frac{1}{2\nu} [\theta_c - (y_{i,t} - \mathbf{f}_{i,t}' \beta_i - \mathbf{W}_i \gamma)]^2 \Big\} \end{aligned}$$

In detail, equation (B.2), the marginal likelihood for a new cluster, is:

$$\alpha \int \mathbf{N}(\varepsilon_i; \theta, \nu)\mathbf{N}(\theta; 0, \tau_0)d\theta = \alpha \int \prod_{t=p+1}^{T} \mathbf{N}(\varepsilon_{i,t}; \theta, \nu)\mathbf{N}(\theta; 0, \tau_0)d\theta$$
$$= \frac{\alpha\sqrt{\nu}}{(\sqrt{2\pi\nu})^{(T-p)}\sqrt{(T-p)\tau_0 + \nu}} \exp\Big[-\frac{\sum_t \varepsilon_{i,t}^2}{2\nu}\Big] \exp\Big\{\frac{\frac{\tau_0}{\nu}(T-p)^2(\bar{\varepsilon_i})^2}{2((T-p)\tau_0+\nu)}\Big\}$$

where $\bar{\varepsilon_i}$ is the mean residual value for user i over $t = (p+1), \ldots, T$, and p is the AR order.

In detail, equation (B.3), is:

$$\theta_c \sim p(\theta_c) = \mathbf{N}(\theta_c; a_c, b_c)$$
$$a_c = b_c\Big(\frac{J\bar{\varepsilon_i}}{\nu}\Big) \quad \text{where } J = (T-p)n_c$$
$$b_c = \frac{1}{1/\tau_0 + J/\nu}$$

Bibliography

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumentral variables. *Journal of the American Statistical Association*, 91(1): 444–455, 06 1996.

Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. Inferring causal impact using Bayesian structural time-series models. *Ann. Appl. Stat.*, 9 (1):247–274, 03 2015. doi: 10.1214/14-AOAS788.

B. Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 01 1979. doi: 10.1214/aos/1176344552.

R.A. Fisher. *Design of Experiments*. Oxford and Boyd, 1935.

A. Gelman, B. Carlin, H. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 2014.

James J. Heckman, Hidehiko Ichimura, and Petra E. Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654, October 1997.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1), 2011.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960, 1986.

G.W. Imbens. Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

G.W. Imbens and D.R. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015.

J.D.Y. Kang and J.L. Schafer. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22(4):523, 2007.

George Karabatsos and Stephen G. Walker. A Bayesian nonparametric causal model. *Journal of Statistical Planning and Inference*, 142(4):925 – 934, 2012. ISSN 0378-3758.

L. C. McCandless, P. Gustafson, and P. C. Austin. Bayesian propensity score analysis for observational data. *Statistics in Medicine*, 28(1):94–112, 2009.

Jerzy Neyman. On the application of probability theory to agricultural experiments. Essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990. ISSN 0883-4237. Translated and edited by Dabrowska, D.M. and Speed, T.P.

P. R. Rosenbaum and D. B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39: 33–38, 1985.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

Paul R. Rosenbaum and Donald B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79:516–524, 1984.

P.R. Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82:387–394, 1987.

D. B. Rubin. Matching to remove bias in observational studies (corr: V30 p728). *Biometrics*, 29:159–183, 1973.

Donald B. Rubin and Neal Thomas. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95(450):573–585, 2000.

Elizabeth A. Stuart. Matching methods for causal inference: A review and a look forward. *Statist. Sci.*, 25(1):1–21, 02 2010. doi: 10.1214/09-STS313.