8-2012

Analysis of the Potential Market for Out-of-Print eBooks

Michael D. Smith
Carnegie Mellon University, mds@cmu.edu

Rahul Telang
Carnegie Mellon University, rtelang@andrew.cmu.edu

Yi Zhang
Carnegie Mellon University, yizhang1@andrew.cmu.edu

Analysis of the Potential Market for Out-Of-Print eBooks

Michael D. Smith, Rahul Telang, Yi Zhang

(mds@andrew.cmu.edu, rtelang@andrew.cmu.edu, yzh.cmu@gmail.com)

Heinz College,
School of Information Systems and Management
Carnegie Mellon University

This Version：August 2012

Acknowledgements: While this research was conducted independently, the authors thank Google University Research grant for the financial support.

Analysis of the Potential Market for Out-Of-Print eBooks

ABSTRACT

The growth of the electronic book market has allowed publishers to make many previously out-of-print titles available, cost-effectively, in an electronic format. However, as of January 2012, there were still nearly 2,700,000 out-of-print titles that are unavailable as eBooks. The goal of this paper is to generate estimates of how much producer and consumer surplus could be created by making these out-of-print titles available in eBook markets.

To do this, we first collect a unique dataset, comprising a random sample of all out-of-print titles that are and that are not available in eBook markets. We then use Bayesian Propensity Score Matching techniques to match books in these two samples based on their observable characteristics. Using these matched titles, we estimate that making the remaining 2.7 million out-of-print books available as eBooks could create $740 million in revenue and $860 million in consumer surplus in the first year after their debut. We also estimate that $460 million of this revenue would accrue directly to publishers and authors as profit.

Keywords: *eBooks, digital distribution, propensity score analysis, consumer surplus.*

I. Introduction

Although it has been 40 years since the first eBook was created by Michael S. Hart[1] in 1971, and more than 10 years since the first eBook was sold online in 1998, eBooks did not show significant market share growth until the first Kindle was introduced by Amazon in November 2007. Since then, sales of eBooks have grown rapidly as reflected in a variety of industry statistics. For example, in 2010, sales of Kindle titles at Amazon exceeded the sales of hardcover titles for the first time (Miller, 2010) and in February 2011 the Association of American Publishers reported that sales of eBooks surpassed the sales of all other book formats (Sporkin 2011). In terms of number of eBook readers, a study conducted by International Data Corporation (IDC) in March 2011[2] found that worldwide sales of eBook readers numbered 12.8 million in 2010, of which 48% were Kindles. Finally, in terms of revenue, the AAP reported that revenue from eBooks increased 120% from 2010 to 2011, reaching $970 million.[3] Annual eBook revenue from 2002 to 2011 by year is shown in Figure 1.

With the growth in readers and eBook sales, there has also been a dramatic growth in the number of available eBook titles, greatly expanding the consumers' choice set. When the Kindle was introduced in late 2007, only 88,000 Kindle titles were available. This number grew to 275,000 by late 2008 and exceeded 500,000 in the spring of 2010.[4] As of April 21, 2012, there were approximately 1.4 million Kindle titles available in Amazon.

Among these 1.4 million titles, many represented digitized versions of titles that had been unavailable in print versions for some time. Our data show that, out of 3,720 to-be-released Kindle titles from October 1, 2011 to December 31, 2011, 596 titles (16%) were already available in physical format (the remainder were new books released in both eBook and print formats or new books only available in eBook format).

[1] Source: http://en.wikipedia.org/wiki/eBook
[2] Source: http://www.idc.com/about/viewpressrelease.jsp?containerId=prUS22737611
[3] Year-end AAP sales report represents data provided by 84 U.S. publishing houses.
[4] Source: http://www.ebookreaders.org.uk/amazon-Kindle/

Digitization of catalog titles — books that have been available in print for some time — has become a new revenue source for publishers and writers, especially for independent writers. For example, novelist Barbara Freethy re-released 12 of her out-of-print titles priced at $0.99 each. The re-release was well received by the market, with her book, "Don't Say A Word," which had been out-of-print in 1995, climbing to No.2 on the Barnes & Noble's NOOK bestseller list in 2011 (Owen, 2011).

There are several potential reasons that previously out-of-print titles might perform well as eBooks. First, eBooks have a very different cost structure than print titles. Because of the fixed costs associated with physical printing runs, after an initial print run sells out, it only makes sense to reprint the book if the expected residual demand exceeds 500 to 1,000 copies. Books with expected demand below these numbers will be allowed to go out of print. However, this leaves a great deal of demand unmet — demand that could be filled in an electronic format where the fixed costs associated with digitization are very low, and where the marginal costs of delivery are near zero. A second reason out-of-print titles might do well as eBooks is the increased opportunities for discovery afforded by digital marketplaces. Physical bookstores only stock 20,000-100,000 unique titles, whereas online retailers can stock as many books as are available. Add to this, increased opportunities to use recommendation engines, peer reviews, and personalized advertisements, and you have a recipe to allow consumers to discover a broader selection of titles than they could in a physical storefront (Zentner, Smith, and Kaya 2012; Brynjolfsson, Hu, and Smith 2010; Kumar, Smith, and Telang 2011). Finally, it is possible that some titles will benefit disproportionately from the convenience and immediate gratification offered through electronic delivery of eBooks.

Together these arguments suggest that the eBook marketplace might give new life to previously out-of-print titles. The goal of this paper is to produce estimates of the producer and consumer surplus that could be created by bringing the world's 2.7 million out-of-print titles back into print as eBooks. To do this, we first generate a random sample of out-of-print books that are available as eBooks, and out-of-print titles that are not available as eBooks. We then use Bayesian Propensity Score Matching techniques to match

titles across these two groups based on observable characteristics. Based on these techniques, we estimate that making the world's out-of-print titles available as eBooks could create $740 million in revenue in the first year after publication, $460 million of which would accrue to the publishers and authors. In addition, we estimate that making these books available would create $860 million in consumer surplus in the first year after publication.

II. Literature Review

This paper draws on a variety of literatures, notably the marketing and information systems literatures on how electronic markets influence variety, sales, and welfare. In this context, Brynjolfsson, Hu, and Smith (2003) find that estimate the consumer surplus gain from access to increased product variety in online stores versus physical stores. Based on 2000 data, they find an increase of nearly $1 billion in consumer surplus from increased product variety in books alone. Brynjolfsosn, Hu, and Rahman (2009) extend this result to show that there is very little competition between online and offline retailers in niche product settings, and Brynjolfsson, Hu, and Simester (2011) show how electronic marketplaces decrease consumer search costs for products relative to search costs that would be seen in physical marketplaces. Finally, Brynjolfsson, Hu, and Smith (2010) find that the consumer surplus gain from "long tail" markets is significantly larger in 2008 than it was in 2000.

Our paper also draws heavily on the statistical literature on Propensity Score Matching techniques. These techniques were first proposed by Rosenbaum and Rubin (1983) as a way to remove bias due to observed covariates. An (2010) estimated that from 1983 to 2010, Propensity Score Matching techniques were used by more than 200 papers in the American Sociological Review and the American Journal of Sociology alone. However, to our knowledge, there are very few applications of Propensity Score Matching in the fields of marketing and information systems. Rubin and Waterman (2006) criticized the under use of Propensity Score Matching by marketing researchers, saying that the tradition of researching on marketing intervention is "to use generally inappropriate techniques."

One important recent extension to propensity score matching techniques is the incorporation of Bayesian approaches to determine the uncertainty in propensity score estimates. Specifically, McCandless, Gustafson and Austin (2009) proposed a Bayesian based model that takes into account the uncertainty in propensity score estimates, and showed that the Bayesian credible interval for the treatment effect is 10% wider than that using conventional method. Kaplan and Chen (2011) acknowledged the value of Bayesian model proposed by McCandless, Gustafson and Austin (2009). However, they argue that the model itself is problematic since the propensity score is treated as a latent variable that is affected by the treatment effect. Instead, they proposed a two-step Bayesian Propensity Score Matching model, a model that uses a Bayesian Probit model when calculating the propensity score, and then uses a traditional method to match multiple sets of propensity scores.

III. Methodology

Our research is focused on estimating the consumer surplus gain from introducing previously out-of-print books to the eBook market. By out-of-print, we mean books that are not stocked by new book retailers and distributors (i.e. they potentially only available through used book markets). We operationalize this in our study by considering books that are not available directly from Amazon (even if they may be available from an Amazon marketplace seller) as being out-of-print. Figure 2 provides an example of such an out-of-print title.

Using this definition, our proposed methodology relies on generating random sample of books that are out-of-print, but available in eBook format, books that we refer to as "Kindle Out-Of-Print" or KOOP titles; and books that are out-of-print and not available in eBook formats, books that we refer to as Non-Kindle Out-Of-Print" or NOOP titles.

In our study, we are interested in predicting the potential sales of NOOP titles if they were made available in the Kindle marketplace, which can be given as

$$ATU = E(Y_1 \mid D = 0) - E(Y_0 \mid D = 0) \quad (1)$$

where ATU is the average treatment effect of moving a book from being unavailable (Y_0) to available (Y_1) in the Kindle marketplace when the book was previously unavailable in the Kindle marketplace (D=0), and where $E(Y_1 \mid D = 0)$ is the expected revenue generated by digitizing a NOOP title and $E(Y_0 \mid D = 0)$ is the current sale of NOOP titles, which is by definition 0.

Unfortunately, we do not know sales or pricing information for a potential NOOP title, given that they are not yet available in the Kindle marketplace. Moreover, we cannot directly conduct an experiment to randomly choose NOOP titles and bring them into the Kindle marketplace. However, we can observe the sales and price of titles that have already been re-released. Thus a tentative solution to the problem of estimating (1) is to infer the sales and price of NOOP titles from sales and price of those KOOP titles that have been re-released, as in (2)

$$TE \quad without \quad Adjustment = E(Y_1 \mid D = 1) - E(Y_0 \mid D = 0) \quad (2)$$

The challenge to directly calculating the treatment effect using (2) is that one must assume that there is no difference between the KOOP and NOOP samples. If this is not true, the bias of inferring the true estimates to (1) by using the estimates of (2) is given by

$$bias = E(Y_1 \mid D = 1) - E(Y_1 \mid D = 0) \quad (3)$$

The NOOP and KOOP samples are likely to differ given that publishers may intentionally digitize titles that are more likely to be successful in the eBook market before they will digitize other titles. In our study, we use Propensity Score Matching as a way to match NOOP titles to similar KOOP titles in an effort to remove this bias. Specifically, after calculating the propensity score using observable characteristics of the books in our sample, we assume that books with the same propensity score can be seen as being randomly assigned to their respective KOOP or NOOP group.

We use the following steps to calculate the propensity score for books in our sample:

(1) Calculation of Propensity Score using Probit/Logit models

(2) Calculation of Propensity Score using Near Neighborhood Matching, Stratification Matching, Caliper Matching, Mahalanobis Metric Matching, etc

(3) Multivariate analysis on the matched groups

We note that Propensity Score Matching relies on two important assumptions. The first is that any selection bias is only due to observed variables. The other assumption is overlap, which means that there is sufficient overlap between the propensity scores in both samples (NOOP and KOOP in our case) to support matching.

1. Sample selection of KOOP titles and NOOP titles

Our first goal is to find a random sample of all KOOP and NOOP titles. Figure 2 summarizes, in flowchart form, our methodology for obtaining the KOOP sample. Specifically, we first conducted an exhaustive scrape of Amazon's Kindle marketplace for all Kindle titles where the original print book was published before 2005. We excluded books published after 2005 because it significantly simplifies our search space and because we believe that the vast majority of books released after 2005 will likely have been published in an eBook format and also will still be in print. This resulted in 125,509 Kindle books that were published before 2005.

We then cross-matched these books with the print book page at Amazon to determine the International Standard Book Number (ISBN) for the matching print title of each book, and to determine if the book was out of print. After removing all books that were still in print, 4,210 KOOP books remained in our sample. We then determined the physical characteristics of these KOOP titles by search for the ISBN number in Global Books in Print (GBIP), Bing, and Amazon we outlined below and tracked the daily rank and price for these KOOP titles for 8 days from November 22, 2011 to November 29, 2011, and calculated the

weekly average rank and weekly average price for those titles, which we will subsequently use to determine Kindle sales for these titles.

We obtained a random sample of NOOP titles (100,000) by randomly selecting a sample of titles that are no longer in print (from GBIP), We then dropped all titles published after 2005, with significant missing product information, and titles that have Kindle copies available. This yields a sample of 7,930 NOOP titles.

2. Calculating Propensity Score for KOOP titles and NOOP titles

After identifying KOOP and NOOP titles, we need to match samples in the NOOP group with samples in the KOOP group. To do this, we first select the variables to be used in the Propensity Score Matching process. Brookhart et al. (2006) suggests that, in selecting Propensity Score Matching variables, researchers should include all variables that might affect outcome, even if they are not related to the exposure. This decreases the variance of estimated exposure without increasing bias. Following this approach, we include the following variables in our Propensity Score calculation:

(1) Price: The list price of the print version of a title. (Source: GBIP)

(2) Year: The year when the title first became available in print format. (Source: GBIP)

(3) Pages: The number of pages of print version. (Source: GBIP)

(4) Category: GBIP divides books into 16 categories based on their topic: Arts, Biography, Business, IT, Education, Fiction, Juvenile, Life, Literature, Medical, Relaxation, Religion, Science, Self-help, Social Science, and Sports. (Source: GBIP)

(5) Audience: GBIP divides books into 4 groups based on Audience: College, General, Professional, and Children. (Source: GBIP)

(6) Bing: The number of Bing search results for each title using the ISBN as the search criteria. (Source: Bing)

(7) Rank: The rank of physical version listed by Amazon. (Source: Amazon)

(8) Format: GBIP divides books into 4 groups based on their binding format: Paperback, Hardcover, Library Binding, and Other. (Source: GBIP)

(9) Large Publisher: This is an indicator variable set to one for "large publishers." The publisher of the book is identified from the second through sixth digits in the ISBN number.

Table 1 and Table 2 present summary statistics for the data. These statistics show clear differences between the two groups, but also show a significant amount of overlap for most variables. The differences between the summary statistics for the two groups suggests a need to use Propensity Score Matching techniques to control for any bias across the two samples, and the overlap in variables suggests an opportunity for these techniques to be successful. To do this, we first calculate the propensity score each title using the following Probit model:

$$\begin{aligned} y_i^* &= \beta_0 + \beta_1 \log(Rank_i) + \beta_2 Price_i + \beta_3 Pages_i + \beta_4 Total_i + \beta_5 (Year_i - 1900) \\ &\quad + \beta_6 \log(Bing_i + 0.1) + \beta_7 \log(Price_i) + \beta_8 \log(Pages_i) + \beta_9 \log(Price_i) * \log(Pages_i) \\ &\quad + \beta_{10} (Year_i - 1900) * \log(Bing_i + 0.1) + \sum_{j=1}^{15} \beta_{11j} Genre.dummies_{ij} \\ &\quad + \sum_{k=1}^{3} \beta_{12k} Binding.dummies_{ik} + \sum_{l=1}^{3} \beta_{13l} Audience.dummies_{il} + \varepsilon_i \\ &= \beta x_i + \varepsilon_i \qquad\qquad where \quad \varepsilon_i \sim N(0,1) \end{aligned} \tag{4}$$

$$y_i = \begin{cases} 0 & if \quad y_i^* \le 0 \\ 1 & if \quad y_i^* > 0 \end{cases} \tag{5}$$

where y_i^* is the latent utility for y_i, the choice made by publisher to publish the book in Kindle format. The predicted value of y_i^* is the propensity score. Here variables of log(Price), log(Pages), and interaction terms of log(Price)*log(Pages) and (Year-1900)*log(Bing+0.1) are added to the model, since

these four terms significantly decrease the AIC. log(Bing+0.1) is used instead of log(Bing) to account for possible zero values.

Table 3 displays the resulting coefficients for this regression. These results suggest that (not surprisingly) publishers decisions regarding which books to bring back into print are not random, but are heavily influenced by the coefficients in our regression: list price, number of pages, and rank of physical format all negatively impact the probability of an OOP title being digitized, while the number of Bing search results and the total number of titles from the publisher in our sample positivity affect the possibility of a title being digitized.

Beyond these individual coefficients, we are also interested in distribution of propensity scores for NOOP and KOOP titles, and whether there is sufficient overlap in these distributions. Figure 4 displays the density plot for the propensity scores from the two samples. From this plot, it is clear that the distribution of propensity scores for KOOP group and NOOP group is quite different, with KOOP titles generally having a higher propensity score, but that there is also significant overlap between the two distributions for propensity score values between 0.2 and 0.8 (see also Figure 9 for a histogram of titles in each group by propensity score values).

3. Calibrating Sales Rank and Sales Quantity

Before matching KOOP and NOOP titles based on these propensity scores, we first must estimate the sales (and revenue) that titles in the KOOP group receive. Unfortunately, Amazon does not publicize its Kindle sales on a per title basis. It does, however, list the "sale rank" of each Kindle title and we use the techniques established in the literature to map these sales ranks to sales levels.

Specifically, prior research has shown that the relationship between Amazon sales and sales ranks approximates a Pareto (Brynjolfsson, Hu and Smith 2003; Chevalier and Goolsbee 2003; Ghose, Smith and Telang 2006), which after a log transformation is given as follows:

$$\log(Sales_i) = \beta_1 + \beta_2 \log(Rank_i) + \varepsilon_i \quad (6)$$

We then calibrate this relationship using data provided by a major publisher matching Kindle weekly sales to observed Kindle sales ranks. This dataset covers weekly sales and sales ranks for 713 eBook titles for 10 weeks.

In our setting it is particularly important that this relationship produces strong fits in the tail of the distribution (titles with lower sales). Our initial exploratory data analysis using (6) found that, consistent with the prior literature (Brynjolfsson, Hu, and Smith 2006), the Pareto distribution doesn't fit well in the tails of the distribution. Because of this we estimated a form of (6) using various different polynomial rank terms, finding that a third degree polynomial best fits our data based on BIC and R^2 measures:

$$\log(Sales_i) = \beta_1 + \beta_2 \log(Rank_i) + \beta_3 \log(Rank_i)^2 + \beta_4 \log(Rank_i)^3 + \varepsilon_i \quad (7)$$

The resulting calibration estimates, and observed sales-rank pairs, are shown in Figure 5. This Figure suggests that, while we obtain reasonably good fit for observations with ranks below 200,000, the fit is not quite as good in the extreme tail (ranks above 200,000).

Because of this, we complement the method outlined above by using a simple experiment (first proposed by Chevalier and Goolsbee 2003) where we order several copies of books with ranks greater than 200,000 and observe their sales rank both before and after purchase. Specifically, we randomly selected 30 Kindle titles with ranks between 200,000 and 1,000,000. We then purchased between 1 and 3 copies of these books and tracked their sales rank before and after this experiment. The resulting ranks are shown in Table 5, where we made our initial purchases at 2:00PM on "Day 1." This table shows that the effect of the sale did not show up in the sales rank until 6-7 hours after the initial purchase, we then use the approximate changes from 1, 2, or 3 purchases to estimate the decrease in rank one would see when a copy of a low selling title is purchased.

In summary, we estimate sales based on observed sales ranks as follows:

(1) For the titles with low ranks (<200,000), we predict the sales using the result from regression shown in Table 4.

(2) If a title has a rank higher than 200,000, we assign sales according to the expected sales that belong to the interval the title's rank falls into based on the experiment described above. For example, if a title has a weekly rank of 231,221, we assign it the sales of 0.838.

(3) If a title does not have a rank, we assume it has no sales.

4. Adjusting Sales Decay Effect for Weekly Rank

Unfortunately, the results above only tell us sales at a particular point in time after the release of a Kindle title. They do not tell us what that title was selling initially after release. Because we are interested in estimating sales after the initial release of a title, we need to attempt to estimate the decay curve of Kindle sales over time.

To do this, we obtain the Kindle release date for each book in our sample and then we attempt to estimate the rank in an arbitrary week based on the distribution of sales by week as follows:

$$Rank_i^T = \Lambda_i + \alpha f(Week_T) + \varepsilon_{iT} \tag{8}$$

$$Rank_i = \Lambda_i + \alpha f(Week_i) + \varepsilon_i \tag{9}$$

where Λ_i is the fixed effect of title i on Rank, and $f(Week_i)$ is a function of $Week_i$. We then combine (8) and (9) to get (10), which can be used to calculate $Rank_i^T$:

$$\begin{aligned} Rank_i^T &= Rank_i + \alpha[f(Week_T) - f(Week_i)] + \varepsilon_{iT} - \varepsilon_i \\ &\approx Rank_i + \alpha[f(Week_T) - f(Week_i)] \end{aligned} \tag{10}$$

Thus, $Rank_i^T$ can be estimated by $Rank_i + \alpha[f(Week_T) - f(Week_i)]$ as long as $\varepsilon_{iT} - \varepsilon_i$ is not large. An assumption of (8) and (9) is that only the fixed variables will affect Rank besides time, which is a

fairly strong assumption, but it is balanced by the fact that these out-of-print titles typically do not receive a lot of promotion and typically have very infrequent price changes.

As noted above, we do not have many books that we observe at the same number of weeks after release. Instead, we can use another method to estimate α , and then use (10) to recover the rank. Specifically, we have both older and newer titles that fall into different release weeks in the sample. If these titles experience similar decay, the average rank of titles that fall into each week is exactly what the sales would look like after a certain time since the debut of a title. As such, we use the following model to estimate α

$$Rank_i = \beta X_i + \alpha f(Week_i) + \varepsilon_i \tag{11}$$

where X_i are book characteristics, and $f(Week_i)$ is a function of $Week_i$

In order to choose the actual model in the form of (11) and $f(Week_i)$, we first calculate the average rank for titles that fall into each week and then plot the average rank for each week. We also conducted a kernel regression of average ranks on time. The data and fitted lines are plotted in Figure 6. From Figure 6, we can see that sales seem to climb after the title's debut, and then drop gradually after around 125 weeks. From Figure 6, we can see that we might need to fit different models to newer titles and older titles separately.

Using this, we used forward step AIC for model selection, where (12) is the initial model, and (13) is the full model:

$$\log(E.Rank_i) = \beta_0 + \varepsilon_i \tag{12}$$

$$\begin{aligned}\log(E.Rank_i) = \beta_0 &+ \beta_1 \log(Rank_i) + \beta_2 log(Price_i) + \beta_3 Pages_i + \beta_4 Total_i + \beta_5 Year_i \\ &+ \beta_6 \log(Bing_i + 0.1) + \sum_{j=1}^{15} \beta_{7j} Genre.dummies_{ij} + \sum_{k=1}^{3} \beta_{8k} Binding.dummies_{ik} \\ &+ \sum_{l=1}^{3} \beta_{9l} Audience.dummies_{il} + \alpha_1 \log(Week_i) + \alpha_2 \log(Week_i)^2 + \alpha_3 Week_i \\ &+ \alpha_4 Week_i^{\,2} + \alpha_5 Week_i^{\,3} + \varepsilon_i\end{aligned} \tag{13}$$

We tried different cut point to divide the dataset into two subsets, and then apply the forward step model selection process based on AIC to each of the subsets. After some trial and error we found week 125 works well. Model (14) and (15) fit each subset as suggested by this approach, with the estimation of model (14) shown in Table 9 and the estimation of model (15) in Table 10.

If $Week_i < 125$

$$\begin{aligned}Rank_i^* = \beta_0 &+ \beta_1 \log(Rank_i) + \beta_2 log(Price_i) + \beta_3 Juvenile_i + \beta_4 \log(Pages_i) + \beta_5 IT_i \\ &+ \beta_6 Paperback_i + \beta_7 Social_i + \beta_8 Business_i + \beta_9 Science_i + \beta_{10} Education_i \\ &+ \beta_{11} Life_i + \beta_{12} Medical_i + \beta_{13} Arts_i + \beta_{14} Religion_i + \beta_{15} Literature_i + \beta_{16} Sports_i \\ &+ \beta_{17} Self\_help_i + \beta_{19} Biography_i + \beta_{20} Relaxation_i + \alpha Week_i + \varepsilon_i\end{aligned} \tag{14}$$

If $Week_i >= 125$

$$\begin{aligned}Rank_i^* = \beta_0 &+ \beta_1 Week_i + \beta_2 \log(Rank_i) + \beta_3 log(Price_i) + \beta_4 \log(Pages_i) + \beta_5 Paperback_i \\ &+ \beta_6 Juvenile_i + \beta_7 Library_i + \beta_8 Self\_help_i + \beta_9 Business_i + \beta_{10} \log(Bing_i + 0.1) \\ &+ \beta_{11} Total_i + \beta_{12} Year_i + \beta_{13} Sports_i + \beta_{14} Education_i + \varepsilon_i\end{aligned} \tag{15}$$

From (14) and (15), we find that $Week_i$ only appears in (14), suggesting that KOOP titles do not show an obvious sign of decay until 75 weeks after their debut. We also estimated two other models on data that has $Week_i >= 125$. One model includes only the variable $Week_i$ and the other includes all variables in model (15) plus $Week_i$ as independent variables. The results for these models are shown in Table 10. These estimates show that $Week_i$ is not significant in either of the models. One explanation for this "stable" period might be that these titles are obscure and rarely receive promotion. Thus, the length of

time it takes for consumers to become informed of a KOOP title's debut might be relatively uniformly distributed for a long period.

We will adjust the rank of all titles using the last week (week=125) of the "stable" period as standard week. Thus, if a title has value of $Week$ bigger than 125, we do not need to adjust the rank. However, if a title has value of $Week$ smaller than 125, we will need to adjust the rank using (14).

In our estimates, we need to pay special attention to the 24 titles in our sample that have sales ranks lowered than 20,000. To be conservative, we do not adjust their ranks. The distribution of ranks after adjustment is shown in Figure 8.

5. Matching Propensity Scores

Next, we attempt to exploit this overlap to match the NOOP to the KOOP samples based on their propensity score. We first note that the smaller overlap between KOOP and NOOP samples is not a problem for propensity score values larger than 0.8 because our goal is the match NOOP titles to KOOP titles, and in this range there are more than enough KOOP titles compared to NOOP titles. The lack of overlap is a problem, however, for propensity score values less than 0.2 since we only have 256 KOOP titles in this range (6.1% of all KOOP titles). Because of this, to be conservative in our analysis we only consider books with propensity scores larger than 0.2 in our analysis, effectively treating NOOP titles with propensity scores from 0 to 0.2 as having no impact on consumer of producer surplus if they were to be digitized. For titles with propensity scores greater than 0.2, we attempt to match titles across groups using two different methods, outlined below.

(1) Nearest Neighbor Matching (NNM)

Using the nearest neighbor matching (NNM) method, we select the KOOP title with the propensity score closest to the score of the NOOP to be matched. Since the number of KOOP titles is much smaller than

the number of NOOP title to be matched, we use NNM without replacement. In order to avoid bad matches, we also only consider pairs that have difference in PS smaller than 0.005.

To check the matching quality using this technique, we note that good propensity score matches should be able to balance the distribution of the relevant variables in both the control and the treatment groups (Caliendo and Kopeinig 2008). From Table 7 and Figure 10, we can see that the distributions of variables after matching are quite similar. In addition to checking the distribution of variables in both groups, Sianesi (2004) suggests that researchers could calculate the propensity score again for both groups after matching, and compare Pseudo R^2 after matching. If the matching is strong, the Pseudo R^2 should be low. We used this method and found that McFadden Pseudo R^2 drops from 0.3566 to 0.0064 after matching.[5] This suggests that, after matching, the variables used for matching can no longer tell the difference between two groups. Thus, the Average Treatment Effect on the untreated group can be calculated as the average of outcomes in the matched KOOP group.

Using these matched samples, and the Kindle sales values estimated above, we find that the ATU of sales would be 1.945 copies/book per week, which is higher than the average copies KOOP samples with propensity scores larger than 0.2 sold during that week, and the ATU of revenue would be $13.74, which is lower than the average revenue of KOOP samples with propensity score larger than 0.2.

(2) Stratification Method

Cochran (1968) shows that five subclasses are often sufficient to remove over 90% of the bias due to the subclassifying variable or covariate. However, as the number of subclassifying variables increases, the number of subclasses would need to increase exponentially (Cochran and Chambers 1965). However, since propensity score is a scalar variable of multiple covariates, using propensity score alone on 5 subclasses would often be enough to remove over 90% of the bias due to each of the covariates (Rosenbaum and Rubin 1984). Thus, to implement this approach we take all titles with propensity scores

[5] We find similar drops using the Maximum Likelihood Pseudo R^2 (0.3689 to 0.0089) and the Cragg and Uhler's Pseudo R^2 (0.5089 to 0.0119).

larger than 0.2 and stratify them into 8 subgroups based on the propensity score: if a title has a propensity score between 0.2 and 0.3, it is assigned to stratum 1, and so on through stratum 8 (propensity score of 0.9 to 1). The average Treatment effect can be calculated using (16) and (17).

$$\overline{Sales}^{ATU} = \frac{\sum_{j=1}^{8}\left(N_j^{NOOP}\frac{\sum_{i=1}^{N_j^{KOOP}} Sales_i}{N_j^{KOOP}}\right)}{\sum_{j=1}^{8} N_j^{NOOP}} \tag{16}$$

$$\overline{Revenue}^{ATU} = \frac{\sum_{j=1}^{8}\left(N_j^{NOOP}\frac{\sum_{i=1}^{N_j^{KOOP}} Sales_i * Price_i}{N_j^{KOOP}}\right)}{\sum_{j=1}^{8} N_j^{NOOP}} \tag{17}$$

The results using these two equations are shown in Table 12. We find that the ATU of sales using the stratification matching is 1.71 copies/book per week, which is higher than the average sales in the KOOP sample with propensity scores larger than 0.2, and the ATU of revenue is $12.94/book, which is smaller than the average revenue of KOOP samples with propensity scores larger than 0.2.

6. Bayesian Propensity Score Matching (BPSM)

Conventional method of Propensity Score Matching discussed earlier does not do a good job of providing a confidence interval for the results Propensity Scores. The variance in Propensity Score estimates is especially important in our study, because of the skewness in sales across titles. Although the number of high-selling titles is relatively small, they could excessively influence our results, particularly if some of the bestselling KOOP are matched multiple times to NOOP samples.

We use Bayesian Propensity Score Matching, which allows us to draw multiple sets of propensity scores from the distribution and repeating the matching process with each set, to help estimate the confidence intervals for propensity scores.

We implement the Bayesian Propensity Score Matching approach using the two-stage method proposed by Gelman et al. (2003) and Kaplan and Chen (2011). Our specific model is the same as the Probit model used above. To estimate this model, we choose a diffuse prior, and set the posterior distribution of the model as

$$y_i^* \mid x_i, \beta, y_i \sim \begin{cases} N(x_i\beta,1)I(y_i > 0) & if \quad y_i = 1 \\ N(x_i\beta,1)I(y_i \le 0) & if \quad y_i = 0 \end{cases} \tag{18}$$

$$\beta \mid x_i, y^* \sim N((x'x)^{-1}(xy^*),(x'x)^{-1}) \tag{19}$$

We run 15,000 iterations of this model with thinning parameters set to 3, and we choose a burn-in period of 2,000, leaving 3,000 propensity scores for use in our estimates. The trace plots for the model variables are shown in Figure 11 and Table 11 shows the estimates of covariates. Table 11 shows that the resulting coefficients are quite similar to those obtained using the conventional propensity score method above.

We then use the resulting 3,000 propensity scores to generate matches based on both the Nearest Neighbor and Stratification methods applied above. This will give us an interval that accounts for the variation due to the uncertainty in the propensity score. The resulting estimates are shown in Table 12, and summarized below:

(1) The expected average sale of NOOP titles is 1.53-1.78 copies/week (25%-75% CI) using the Nearest Neighbor method, and 1.61-1.74 copies/week (25%-75% CI) using the stratification method. This is much higher than the average weekly KOOP sales of 1.43.

(2) The expected average revenue for a NOOP title is $12.01-13.61/week (25%-75% CI) using the Nearest Neighbor method, and $12.6-13.15/week (25%-75% CI) using the stratification method. This is much lower than the average weekly revenue for KOOP titles of $14.69.

Figure 12 shows that the probability of being digitized is strongly correlated with expected revenue. This is especially obvious for titles with very high propensity scores (0.9-1). We can see this more clearly by running (20) on all Kindle titles:

$$Week_i = \beta_0 + \beta_1 PS_i + \beta_2 log(Total_i) + \beta_3 PS_i * log(Total_i) + \varepsilon_i \quad (20)$$

Table 13 displays the results of this regression and shows that publishers tend to release titles with higher propensity scores earlier than other titles (negative β_1). Likewise, larger publishers enter the digital market earlier (negative β_2) than other publishers do.

Before we move on to the next part, we need to examine the robustness of our result. In order to do this, we randomly draw (1) 6,000 (2) 8,000 (3) 10,000 (4) 12000 samples and re-run the previous steps using these samples to compare how result differs. Table 14 shows the result of estimation using different subset of samples. We can see that although the number of ATU estimated varies cross different random subset of samples, the variation is small. We consider the estimation is pretty robust.

7. Welfare Analysis

In this section, we use these propensity score and Kindle sales estimates to calculate estimates of the producer and consumer surplus that could be realized by making current NOOP titles available in Kindle format. In the following analysis we do this by estimating these figures for the first year after Kindle release for a random sample of 100,000 NOOP titles with a propensity score larger than 0.2.

To evaluate the potential impact of this digitization on consumer surplus, we follow the technique developed by Hausman (1981) and applied by Brynjolfsson, Hu, and Smith (2003) and Hausman and Leonard (2002). Specifically, we measure compensating variation as follows:

$$CV = \sum_{i=1}^{N} CV_i = \sum_{i=1}^{N} [e(p_{p0i}, p_{e0i}, u_{1i}) - e(p_{p1i}, p_{e1i}, u_{1i})] \quad (21)$$

where CV represents total net consumer welfare by introducing all NOOPs in a Kindle version, CV_i represents consumer welfare by introducing NOOP title i into a Kindle version, p_{p0i} and p_{p1i} are the price of physical books in the used marketplace before and after the introduction of KOOP respectively, p_{e0i} is the virtual price of KOOP title i,[6] p_{p1i} is post-introductory price of the KOOP title, u_{1i} is the post-introduction utility level, $e(p_{p0i}, p_{e1i}, u_{1i})$ is the consumer's expenditure function before the introduction of the product i, and $e(p_{p1i}, p_{e1i}, u_{1i})$ is the consumer's expenditure function after the introduction of product i.

Our estimates assume that the introduction of a specific Kindle title has a very small impact on physical book sales for that title. This assumption is consistent with Hu and Smith's (2012) finding that delaying the introduction of Kindle titles results in a statistically insignificant increase in print sales. Given this assumption, or CV equation simplifies to:

$$CV_i = \sum_{i=1}^{N} [e'(p_{e0i}, u_{1i}) - e'(p_{e1i}, u_{1i})] \tag{22}$$

Following Hausman (1981) and Brynjolfsson, Hu and Smith (2003) and Ghose, Smith and Telang (2006), we assume the consumer's demand follows the Cobb-Douglas demand function, which is

$$x_i = A p_i^{\alpha} y^{\delta} \tag{23}$$

Using Roy's identity

$$x_i(p_i, y) = -\frac{\partial u_i(p_i, y)/\partial p_i}{\partial u_i(p_i, y)/\partial y} \tag{24}$$

and solving this function, we get

[6] The virtual price is defined by Hausman (1981) as the lowest price that would set demand equal to zero.

$$u_i(p_i, y) = -A\frac{p_i^{1+\alpha}}{1+\alpha} + \frac{y^{1-\delta}}{1-\delta} \tag{25}$$

and

$$e_i(p_i, u_i) = [(1-\delta)(u_i + \frac{Ap_i^{1+\alpha}}{1+\alpha})]^{\frac{1}{1-\delta}} \tag{26}$$

Using (25) and (26), it can be shown that (Hausman 1981):

$$CV_i = [\frac{1-\delta}{1+\alpha} y^{-\delta}(p_{e0i}x_{0i} - p_{e1i}x_{1i}) - y^{1-\delta}]^{\frac{1}{1-\delta}} - y \tag{27}$$

Further, following Brynjolfsson, Hu, and Smith (2003), if we assume zero income elasticity for books ($\delta = 0$) — based on the fact that books make up a relatively small proportion of overall consumer expenditures — and given that $p_{e0i}x_{0i} = 0$, equation (27) simplifies to

$$CV_i = -\frac{p_{e1i}x_{1i}}{1+\alpha} \tag{28}$$

and, assuming a constant elasticity across Kindle titles, the total CV for all titles is given by

$$CV = -\frac{\sum_{i=1}^{N} p_{e1i}x_{1i}}{1+\alpha} \tag{29}$$

where N is still the number of NOOP titles to be digitized.

If we take the average price and initial quantity over a random sample of titles, we can further simplify (29) as

$$CV = -N\frac{\frac{1}{N_1}\sum_{i=1}^{N_1} p_{e1i}x_{1i}}{1+\alpha} = -\frac{N\overline{px}}{1+\alpha} \tag{30}$$

This leaves our main task as estimating price (p_{e1i}), sales (x_{1i}) for the NOOP samples in order to calculate average revenue, and then multiply this by $-\frac{N}{1+\alpha}$ to get consumer surplus brought by digitizing a large number of NOOP titles. We discuss this approach in more detail below.

(1) Total Revenue.

We calculated the expected average revenue from digitizing Kindle titles as part of the Propensity Score matching discussion above. Since sales over the first 75 weeks of a title are relatively stable, we can simply multiply the average expected weekly revenue of one NOOP book by the number of titles to be digitized and the number of weeks. The first column of Table 15 displays the results for this calculation and shows expected revenue of $627.17 to $707.51 (25%-75% CI) per title for the first year after their debut using nearest neighbor matching, and $655.20 to $683.80 (25%-75% CI) using stratification. This is lower than $763.88, the total revenue generated from the same number of randomly selected KOOP titles with a propensity score larger than 0.2, and is lower than the $714.22 and $672.93 estimates that would result from using the traditional estimation approach with nearest neighbor and stratification matching respectively.

(2) Publisher Welfare

To calculate publisher welfare, we first note that, based on current Kindle sales contracts, publishers receive 70% of the marginal profit generated from Kindle sales, which is price minus delivery cost. The current Amazon delivery cost is $0.15/MB of content[7]. Based on this, our estimate of publisher welfare can be given as follows:

$$\begin{aligned} PW &= TN\frac{1}{N_s}\sum_{i=1}^{N_s}[Q_i RoyaltyRate*(Price_i - DeliveryRate*FilesSize_i) - ScanningCost_i] \\ &\approx TN\frac{1}{N_s}\sum_{i=1}^{N_s}Q_i*RoyaltyRate(Price_i - DeliveryRate*\overline{FilesSize}) - N\overline{ScanningCost} \end{aligned} \tag{31}$$

[7] Source: https://kdp.amazon.com/self-publishing/help?topicId=A29FL26OKE7R7B

where $RoyaltyRate = 70\%$, $DeliveryRate$ =\$0.15/MB, N_s is the sample size of the matched group, Q_i is the weekly sales of title i , $Price_i$ is the price of title i , T is the number of weeks (52 in our case) , N is the number of NOOP titles to be digitized (1 for the estimates below), $\overline{FilesSize}$ is the average size of NOOP titles (which we assume to be 5MB), and $\overline{ScanningCost}$ is the average scanning cost (which was estimated to be \$5-\$10/book, and where we use \$10/book to be conservative).[8]

First note that (31) roughly equals to (32):

$$PW \approx RoyaltyRate(TN\,\overline{Revenue}^{ATU} - TN\,\overline{Sales}^{ATU}\,DeliveryRate * \overline{FilesSize}) - N\,\overline{ScanningCost} \quad (32)$$

where $\overline{Revenue}^{ATU}$ is the average treatment effect using revenue as the outcome, and $\overline{Sales}^{ATU}$ is the average treatment effect using sales as the outcome.

Using estimates for $\overline{Revenue}^{ATU}$ and $\overline{Sales}^{ATU}$ obtained above, we estimate (see Table 14) that average publisher welfare from digitizing one previous unavailable (NOOP) title with PS higher than 0.2 is between \$405 and \$421 (25%-75% CI) using the Nearest Neighbor Method and between \$387 and \$437 (25%-75% CI) using the stratification method.

(3) Retailer Welfare

To estimate retailer welfare, we use the fact that Amazon receives the remaining 30% of marginal profit. Following a similar approach as in (32) above, we then express retailer welfare as

$$\begin{aligned} RW &= TN \frac{1}{N_s} \sum_{i=1}^{N_s} Q_i (1 - RoyaltyRate) * (Price_i - 0.15 * FilesSize_i) \\ &\approx TN \frac{1}{N_s} \sum_{i=1}^{N_s} Q_i (1 - RoyaltyRate * Price_i) * (Price_i - 0.15 * \overline{FilesSize}) \end{aligned} \quad (33)$$

Again, following the approximation above, (33) can be approximated as follows

[8] Source: http://www.opencontentalliance.org/2009/03/22/economics-of-book-digitization/

$$RW \approx (1 - RoyaltyRate)(TN\overline{Revenue}^{ATU} - TN\overline{Sales}^{ATU} DeliveryRate * \overline{FilesSize}) \quad (34)$$

Our result using this equation is shown in the third column of Table 14. We find that retailer welfare per title is between $170 and $191 (25%-75% CI) using the Nearest Neighbor Method and between $178 and $185 (25%-75% CI) using the stratification method.

(4) Consumer Surplus

Following equation (30), consumer surplus can be calculated as follows:

$$CV \approx -\frac{TN\overline{Revenue}^{ATU}}{1+\alpha} \quad (35)$$

where all parameters are known except for price elasticity (α).

To calculate price elasticity, we start with the following relationship between price and sales:

$$\log(Sales_i) = \beta_0 + \Lambda_i + \alpha \log(Price_i) + \varepsilon_i \quad (36)$$

$$\log(Sales_i^t) = \beta_0 + \Lambda_i + \alpha \log(Price_i^t) + \varepsilon_i^t \quad (37)$$

where Λ_i captures the book fixed effect. Combining these two equations gives

$$\Delta \log(Sales_i) = \alpha \Delta \log(Price_i^t) + \Delta \varepsilon_i \quad (38)$$

We calibrate (38) using price and rank data collected in late March 2012 and again in April 2012 on the same sample of titles. This collection found 685 titles that experienced a price change during this period.

We estimate (38) on both the whole sample and on samples with $\Delta \log(Price_i^t)$ >0.1, 0.2, 0.3, and 0.4 separately. Our results, shown in Table 16, suggest that Kindle price elasticity is between -1.53 and -1.86. We note that this is similar to the price elasticity of physical books found in previous studies (for example, Brynjolfsson, Hu and Smith (2003) estimated print book elasticity between −1.56 and −1.79, and Ghose and Gu (2006) print price elasticity between -1.49 and -1.89.

To be conservative, we use a price elasticity of -1.86 in (35), which results in a consumer surplus estimate of between $729.27 to $822.69 (25%-75% CI) per title using the Nearest Neighbor Method and between $761.86 and $795.12 (25%-75% CI) using the stratification method.

IV. Discussion

As noted above, the growth of the eBook market has created a significant potential opportunity for publishers and authors to bring previously out-of-print titles back into the marketplace through electronic distribution. The goal of this paper is to attempt to generate economic estimates of the producer and consumer surplus that could be created by digitizing and selling the 2.7 million books that are currently unavailable in eBook format.

In this paper we attempted to generate these estimates by converting the known sales rank into estimates of sales of a random sample of out-of-print titles that are available on the Kindle marketplace. We then used propensity score matching techniques to match these Kindle-available (KOOP) titles to a similar random sample of out-of-print titles that were not available on the Kindle marketplace (NOOP). We then estimated that the sales of NOOP titles would approximate the estimated sales for their matched KOOP title if the NOOP titles were made available in an electronic marketplace. We then use these estimates, along with established methods for calculating surplus generated by new goods, to estimate the consumer and producer surplus that would be generated by digitizing randomly selected NOOP titles with PS larger than 0.2. These estimates are presented above.

With these estimates, we can then generate a total estimate of the consumer and producer surplus that could be created by digitizing all the world's 2.7 million out-of-print titles and making them available as eBooks by multiplying the 2.7 million and then scaling these estimates to account for the fact that 41.3% of our titles (40.8% to 41.8% with a 25% confidence interval) have propensity scores above 0.2, the cutoff point for obtaining reliable estimates in our data.

After doing this, we find that bringing the world's 2.7 million out-of-print titles back into print as eBooks could create $740 million in revenue in the first year after publication, $460 million of which would accrue to the publishers and authors. In addition, we estimate that making these books available would create $860 million in consumer surplus in the first year after publication.

However, we wish to note carefully that our methodology for obtaining these estimates has several important limitations. First, our estimates rely on accuracy of the propensity score matching across NOOP and KOOP titles, which is based on observable book characteristics. If these observable characteristics do not adequately capture publisher's decisions about which out-of-print titles to bring into the Kindle market, it could bias our results. In order to check how this selection might affect our prediction, we eliminate all top 10% bestselling KOOP samples, which one might argue to titles that were deliberately and successfully selected by publishers. We then use the rest of the samples to match with our NOOP samples. The average weekly sale per title drops to 0.23, and the average weekly revenue per title drops to $2.65. Based on this calculation, making the remaining 2.7 million out-of-print books available as eBooks could create $150 million in revenue and $177 million in consumer surplus in the first year after their debut. Out of the revenue, $55 million would accrue directly to publishers and authors as profit. These numbers are much lower than what we get using all KOOP samples. However, the numbers suggest that the surplus created by digitization is still pretty large even if top selling titles were those that were successfully selected by publishers. Second, lacking publicly available Kindle sales data, our estimates rely on our ability to properly map observed sales ranks for Kindle titles to actual sales levels. While we tried to be both careful and conservative in this estimation, as noted above, the fit between sales rank and sales is relatively poor for low selling titles — the focus of our research, and this might also bias our results. Third, we only considered sales of titles with PS bigger than 0.2 when calculating this surplus generated from releasing 2.7 million titles, causing our estimate underestimated. A final category of limitations arise from the fact that our estimates are (of necessity) based on the current size and scope of the eBook market. Our estimates could change (and indeed would likely increase) as the penetration of

eBook readers increases. Previous research shows that the cannibalization of physical book from eBook for the same title is negligible (Hu and Smith, 2011). However, we may be overestimating the true surplus generated by digitizing these new titles if the sales of these new titles cannibalize sales of existing titles (titles that are currently available in Kindle format). However, in spite of these limitations, we believe that our estimates provide a useful first effort to estimate changes in consumer surplus resulting from the introduction of new goods in this strategic market.

We also note that the method proposed in this paper could also be applied by publishers to decide which of their titles they should focus on first when digitizing out-of-print catalog titles. We note that publishers could also adapt our proposed methods to take into account other, unobservable, book characteristics that might influence the decision to introduce books into the Kindle market.

References:

Allen, T., Feb 22, 2011. Kindle, We Have a Problem: Amazon's Pricing Policies Affect Publishers Publishers Weekly http://www.publishersweekly.com/pw/by-topic/digital/content-and-eBooks/article/46244-kindle-we-have-a-problem-amazon-s-pricing-policies-affect-publishers-.html.

An, W., 2010. Bayesian Propensity Score Estimators: Incorporating Uncertainties in Propensity Scores into Causal Inference, *Sociological Methodology* 40, 151-189.

Bittlingmayer, G., 1992. The Elasticity of Demand for Books, Resale Price Maintenance and the Lerner Index, *Journal of Institutional and Theoretical Economics* 148, 588-606.

BLog, L. L., March 12, 2012. Google Book Scan Project Slows Down, Law Librarian Blog http://lawprofessors.typepad.com/law_librarian_blog/2012/03/googleBook-scan-project-slows-down.html.

Brookhart, M. A., S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, St, and T. rmer, 2006. Variable Selection for Propensity Score Models, *American Journal of Epidemiology* 163, 1149-1156.

Brynjolfsson, E., Y. Hu, and M. D. Smith, 2003. Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers, *Management Science* 49, 1580-1596.

Brynjolfsson, E., Y. J. Hu, and M. S. Rahman, 2009. Battle of the Retail Channels: How Product Selection and Geography Drive Cross-Channel Competition, *Management Science* 55, 1755-1765.

Brynjolfsson, E., Y. J. Hu, and D. Simester, 2011. Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales, *Management Science*, Forthcoming.

Brynjolfsson, E., Y. J. Hu, and M. D. Smith, 2010. The Longer Tail: The Changing Shape of Amazon s Sales Distribution Curve, SSRN eLibrary.

Caliendo, M., and S. Kopeinig, 2008. Some Practical Guidance for the Implementation of Propensity Score Matching, *Journal of Economic Surveys* 22, 31-72.

Chevalier, J., and A. Goolsbee, 2003. Measuring Prices and Price Competition Online: Amazon.com and BarnesandNoble.com, *Quantitative Marketing and Economics* 1, 203-222.

Cochran, W. G., 1968. The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies, *Biometrics* 24, 295-313.

Cochran, W. G., and S. P. Chambers, 1965. The Planning of Observational Studies of Human Populations, *Journal of the Royal Statistical Society*. Series A (General) 128, 234-266.

Deahl, R., Jul 22, 2010. Random House Prepared to Challenge Wylie Agency's New Publishing Biz publishers Weekly http://www.publishersweekly.com/pw/by-topic/digital/content-and-eBooks/article/43925-random-house-prepared-to-challenge-wylie-agency-s-new-publishing-biz.html.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2003. Bayesian Data Analysis, Second edition. (Chapman and Hall, London).

Ghose, A., and B. Gu, 2006. Search Costs, Demand Structure and Long Tail in Electronic Markets: Theory and Evidence, SSRN eLibrary.

Ghose, A., M. D. Smith, and R. Telang, 2006. Internet Exchange for Used Books: An Empirical Analysis of Product Cannibalization and Welfare Impact, *Information Systems Research* 17, 3-19.

Ghose, A., M. D. Smith, and R. Telang, 2006. Internet Exchanges for Used Books: An Empirical Analysis of Product Cannibalization and Welfare Impact, *Information Systems Research* 17, 3-19.

Hausman, J. A., 1981. Exact Consumer's Surplus and Deadweight Loss, *The American Economic Review* 71, 662-676.

Hausman, J. A., and G. K. Leonard, 2002. The Competitive Effects of a New Product Introduction: A Case Study, *The Journal of Industrial Economics* 50, 237-263.

Hawkins, R., Aug 25, 2010. Wylie Agency & Random House Come to Agreement on eBooks, American Booksellers Association http://news.bookweb.org/news/wylie-agency-random-house-come-agreement-eBooks.

Helft, M., April 3, 2009. Google's Plan for Out-of-Print Books is Challenged, *The New York Times* http://www.nytimes.com/2009/04/04/technology/internet/04books.html?pagewanted=all.

Hu, Y. J., and M. D. Smith, 2011. The Impact of Ebook Distribution on Print Sales: Analysis of a Natural Experiment, SSRN eLibrary.

III, J. H., and D. Wiley, 2010. The Short-Term Influence of Free Digital Versions of Books on Print Sales, *The Journal of Electronic Publishing* 13.

Kahn, B. E., and D. R. Lehmann, 1991. Modeling choice among assortments, *Journal of Retailing* 67, 274-299.

Kaplan, D., and C. J. S. Chen, 2011. Bayesian Propensity Score Analysis: Simulation and Case Study, *Society for Research on Educational Effectiveness*.

Lerner, A. P., 1934. The Concept of Monopoly and the Measurement of Monopoly Power, *The Review of Economic Studies* 1, 157-175.

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

Miller, C. C., July 19, 2010. eBooks Top Hardcovers at Amazon, The New York Times http://www.nytimes.com/2010/07/20/technology/20kindle.html.

Owen, L. H., May 20, 2011. The Bestsellers: Out-of-Print Romance Title Lands New Profits As eBook, Paid Content http://paidcontent.org/article/419-the-bestsellers-out-of-print-romance-title-lands-new-profits-as-eBook/.

Rosenbaum, P. R., and D. B. Rubin, 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika* 70, 41-55.

Rosenbaum, P. R., and D. B. Rubin, 1984. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score, *Journal of the American Statistical Association* 79, 516-524.

Rubin, D. B., and R. P. Waterman, 2006. Estimating the Causal Effects of Marketing Interventions Using Propensity Score Methodology, *Statistical Science* 21, 206-222.

Schinittman, E., Sep 27, 2010. Ebooks Don't Cannibalize Print, People Do, Black Plastic Glasses (Blog) http://www.blackplasticglasses.com/2010/09/27/ebooks-don%E2%80%99t-cannibalize-print-people-do/.

Sianesi, B., 2004. An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s, *Review of Economics and Statistics* 86, 133-155.

Sporkin, A., April 14, 2011 eBooks Rank as #1 Format among All Trade Categories for the Month, Association of American Publishers http://www.publishers.org/press/30/.

Stein, L., and P. Lehu, 2008. Literary Research and the American Realism and Naturalism Period: Strategies and Sources (Scarecrow Press, Lanham, MD).

Weekly, P., Sep 15, 2011. Judge Adopts Trial Schedule At Google Status Conference, but Settlement Talks Continue Publishers Weekly http://www.publishersweekly.com/pw/by-topic/digital/copyright/article/48709-judge-adopts-pre-trial-schedule-at-google-status-conference-but-settlement-talks-continue.html.

Zhao, Z., 2008. Sensitivity of propensity score methods to the specifications, *Economics Letters* 98, 309-319.

Figures and Tables:

Figure 1 Growth in eBook Revenue[9]

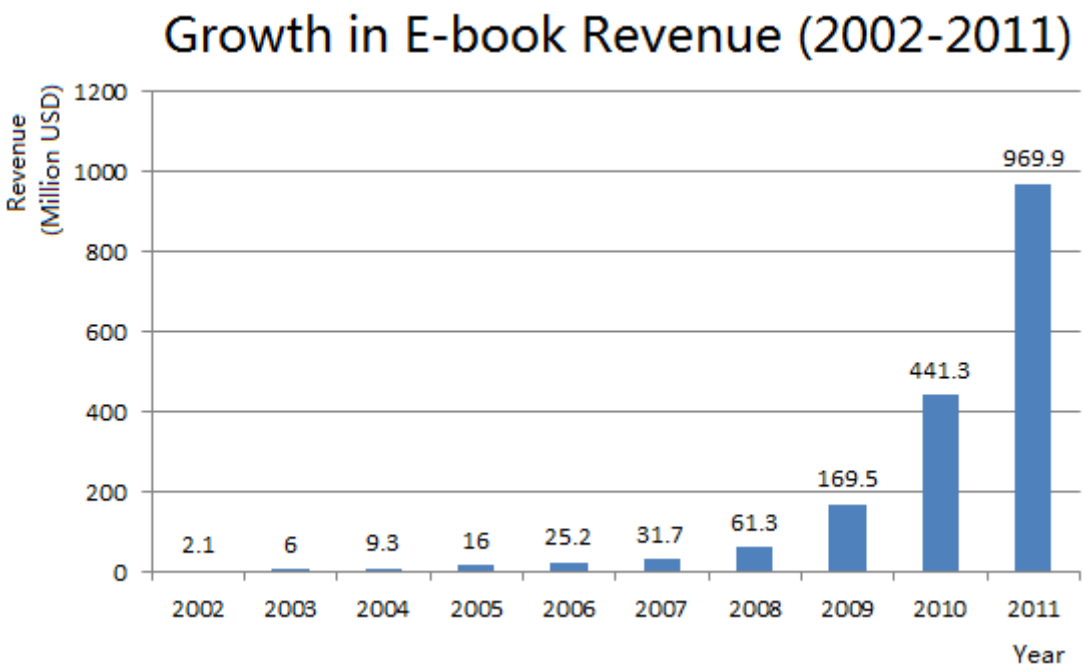
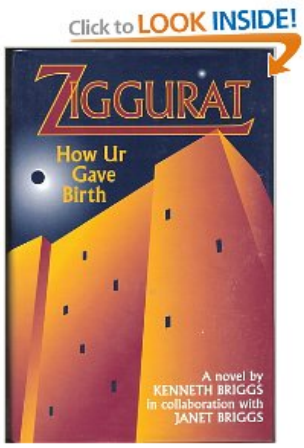


Figure 2 Example of OOP Book



See 1 customer image

Share your own customer images

I own the rights to this title and would like to make it available again through Amazon.

Ziggurat: How Ur Gave Birth [Hardcover]

Kenneth Briggs (Author), Janet Briggs (Author)

Be the first to review this item | Like (0)

Available from these sellers.

9 new from $9.90 7 used from $0.01 3 collectible from $11.85

Formats	Amazon Price	New from	Used from
Kindle Edition	$7.99	--	--
Hardcover	--	$9.90	$0.01
Unknown Binding	--	--	--

[9] Graph plotted based on data from http://www.book-fair.com/pdf/buchmesse/buchmarkt_usa.pdf and http://www.publishersweekly.com/pw/by-topic/industry-news/financial-reporting/article/50805-aap-estimates-eBook-sales-rose-117-in-2011-as-print-fell.html

Figure 3 Identification of KOOP title

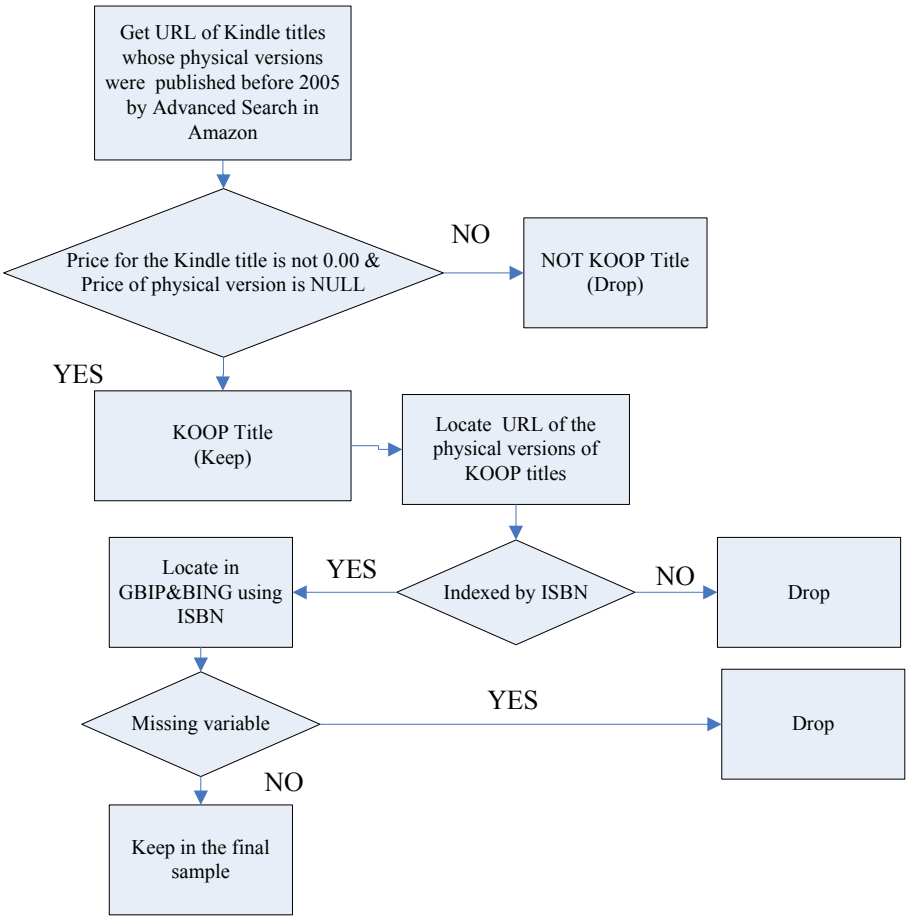


Figure 4 Density Plot of Propensity Score

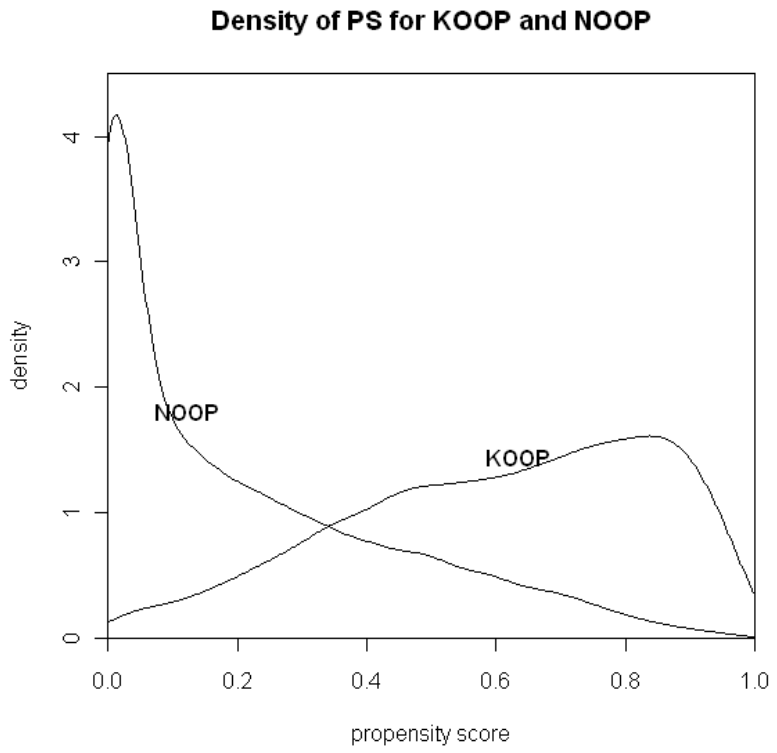


Figure 5 Calibrations between Sales and Rank

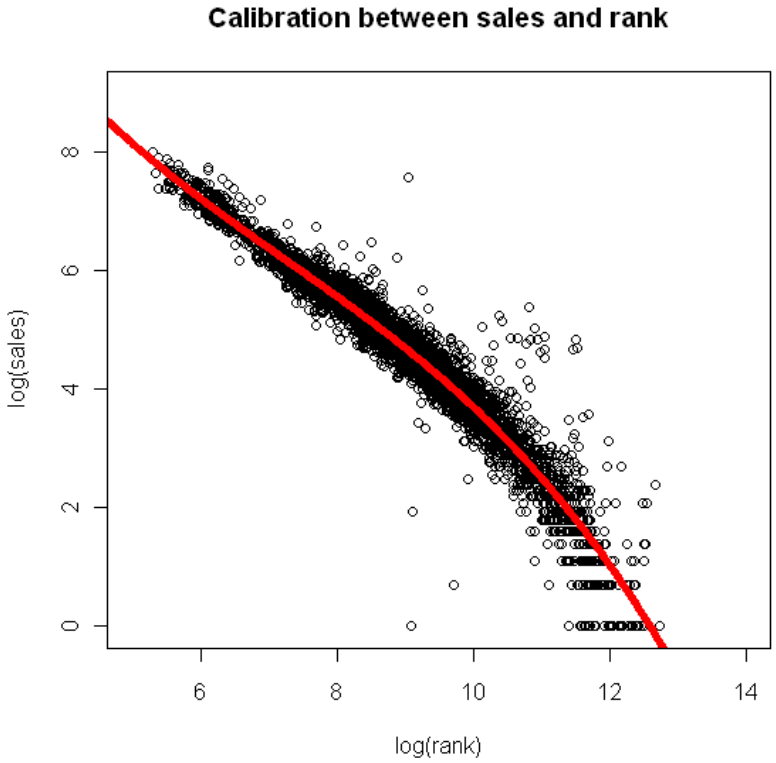


Figure 6 Number of KOOP released

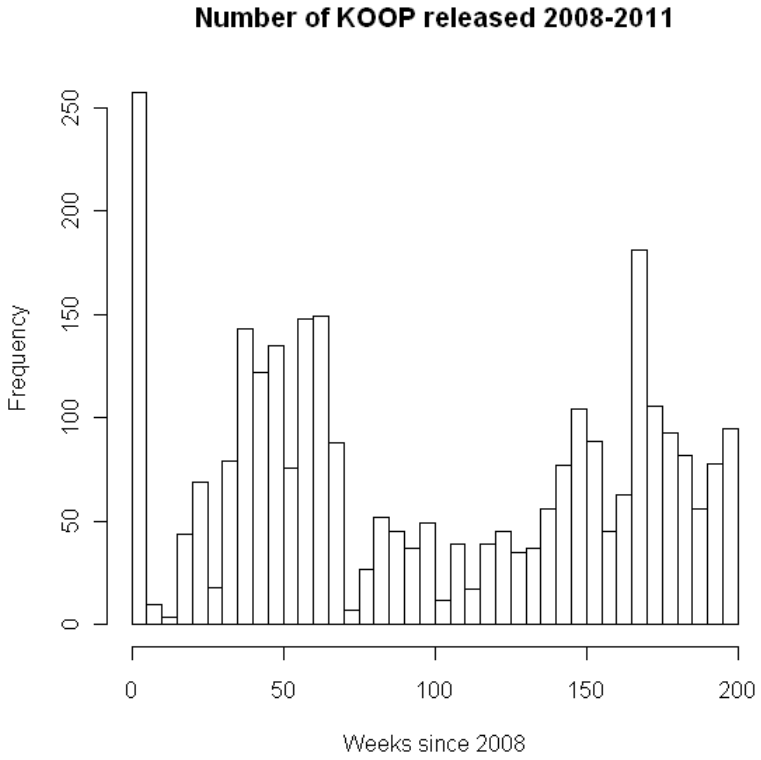


Figure 7 First Look at Difference in Rank for Older/New titles

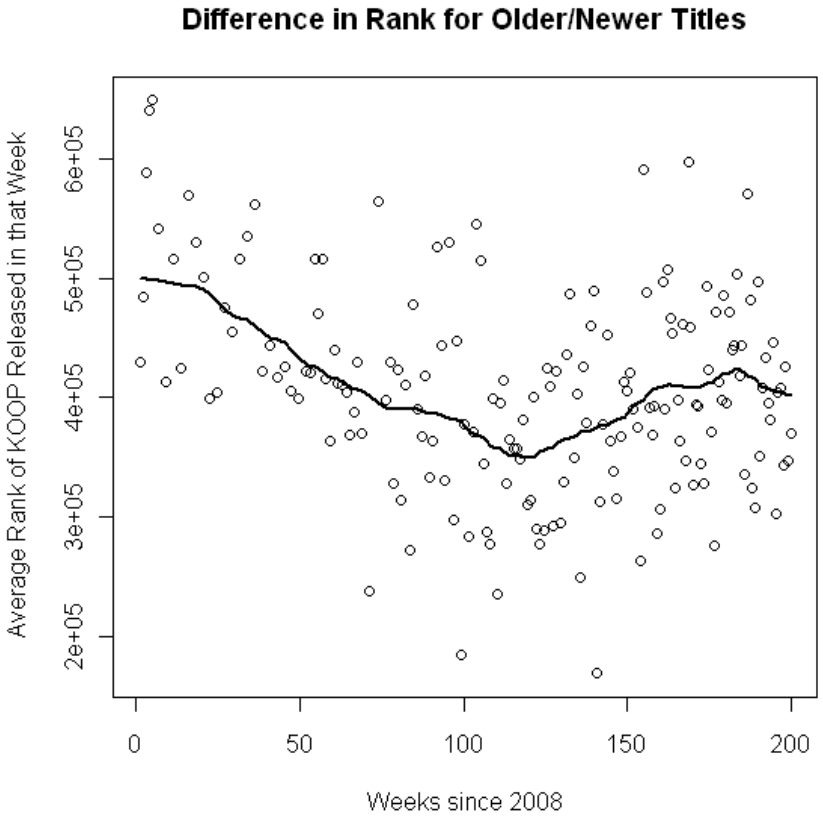


Figure 8 Distribution of Adjusted Rank

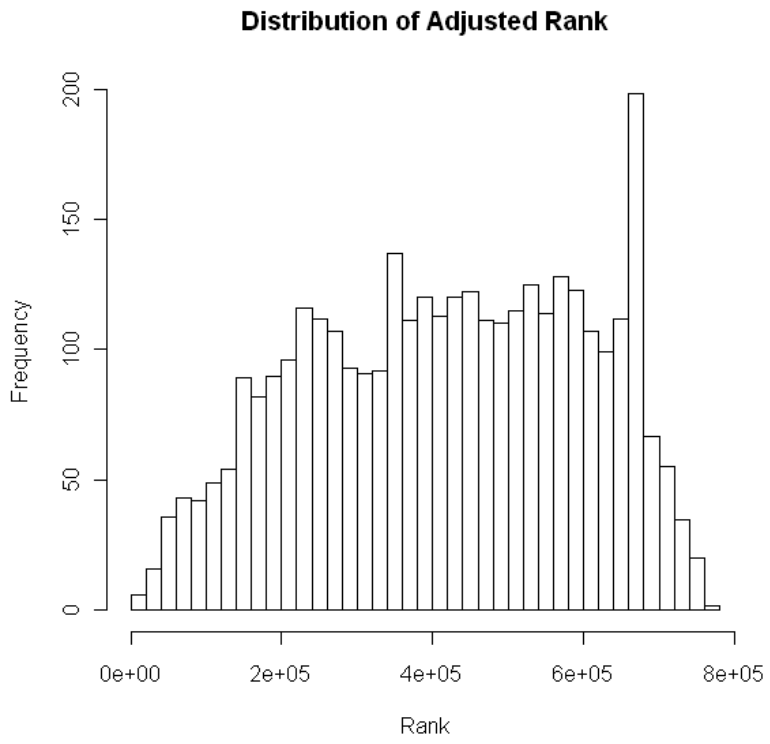


Figure 9 Number of KOOP and NOOP titles in each Stratum

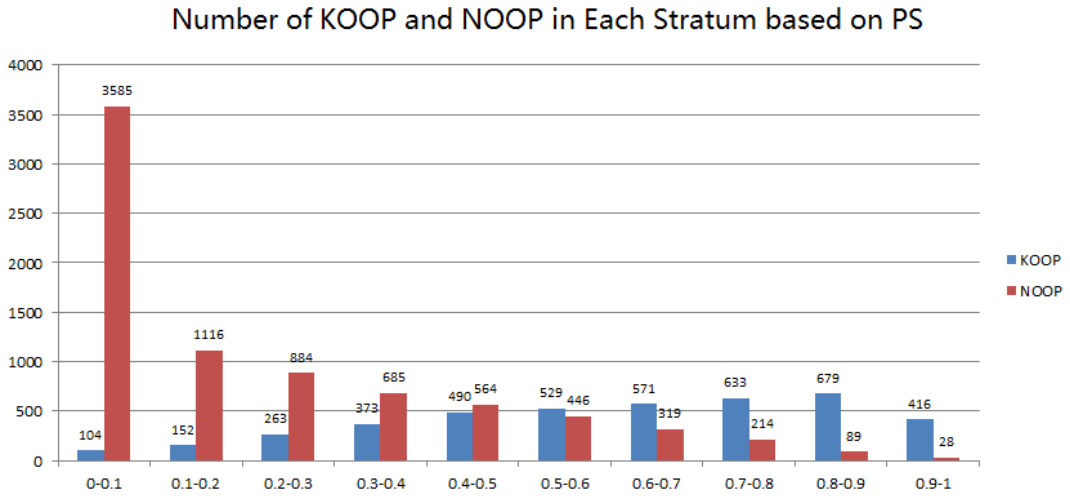


Figure 10 Boxplot for Variables before and after Matching

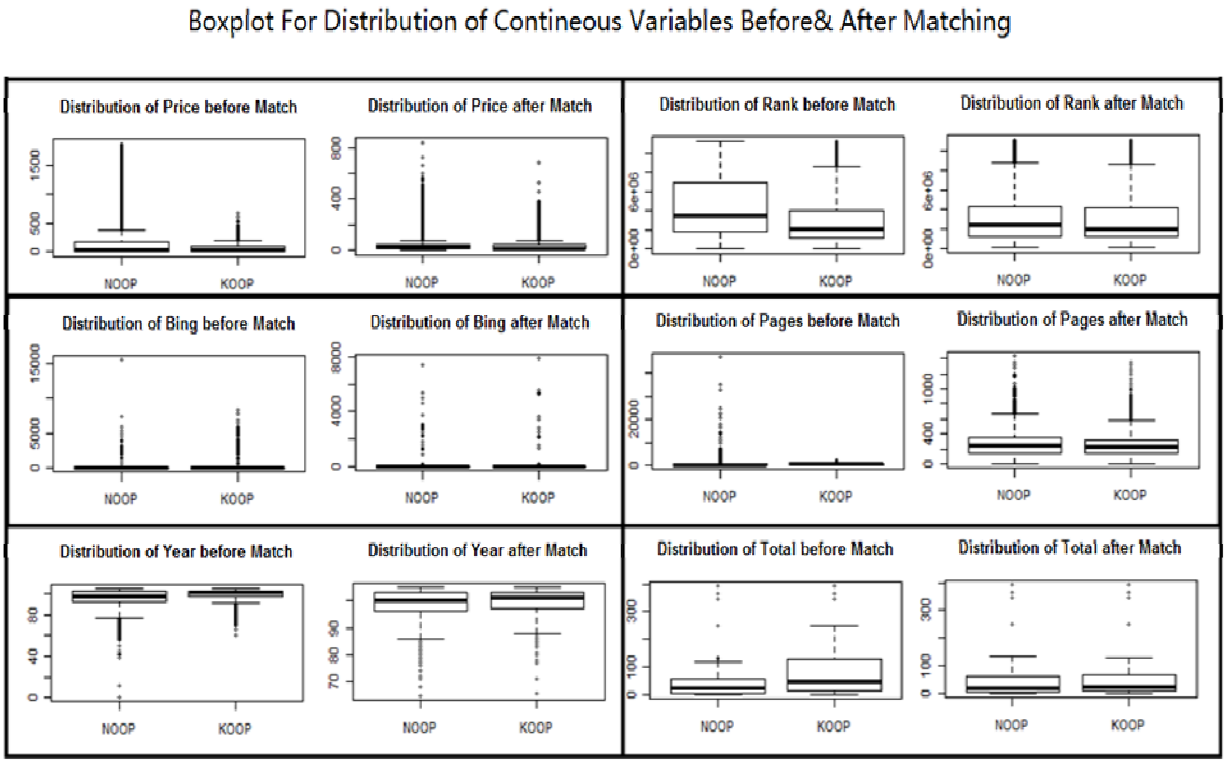


Figure 11 Trace Plots for Variables in Bayesian Probit Model

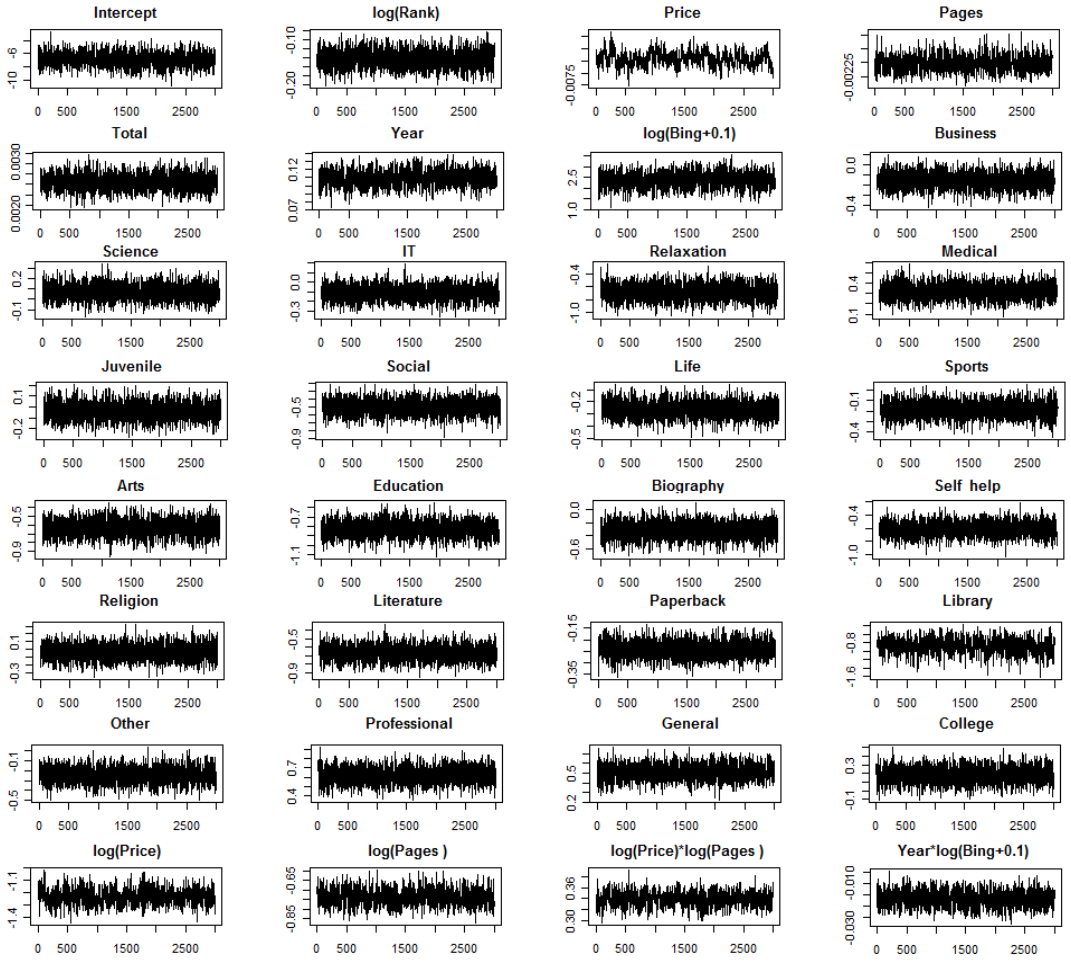


Figure 12 Relationship between PS and Revenue

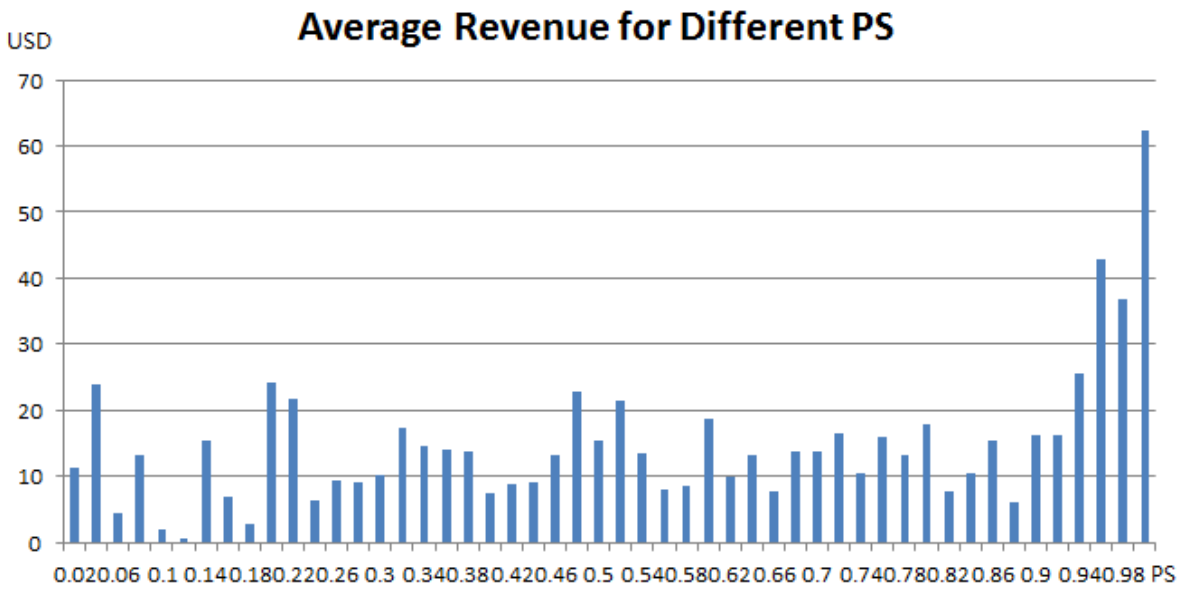


Table 1 Statistics for continuous variables (Aggregate)

Variables	Min	1st Qu	Median	Mean	3 rd Qu	Max
			Aggregate			
Price	0.5	13.95	24.95	129.3	109	1883
Rank	4342	1469161	2976172	3973408	5783096	11211166
Bing	0	9	14	49.38	22	15500
Pages	1	111	224	379.2	360.2	46864
Year	1901	1994	1999	1997	2002	2005
Total	1	8	29	70.03	75	393
			KOOP			
Price	0.95	15	29.25	63.11	89.95	679
Rank	4342	1124821	2161949	3023458	4095268	11191057
Bing	1	12	18	87.02	29	8390
Pages	1	175	256	286.6	352	2192
Year	1960	1998	2001	2000	2003	2005
Total	1	14	46	109	129	393
			NOOP			
Price	0.5	12.95	22.99	164.44	162	1883
Rank	46109	1789708	3535421	4477731	6918296	11211166
Bing	0	8	12	29.29	18	15500
Pages	1	64	194.5	428.4	368	46864
Year	1901	1992	1997	1995	2002	2005
Total	1	6	22	50.95	55	393

Table 2 Statistics for Categorical Variables

	Variables	Aggregate	KOOP	NOOP	Percentage (KOOP)	Percentage (NOOP)
Genre	Arts	680	126	554	2.99%	6.99%
	Biography	277	72	205	1.71%	2.59%
	Business	801	392	409	9.31%	5.16%
	IT	1108	576	532	13.68%	6.71%
	Education	201	65	136	1.54%	1.72%
	Fiction	1114	333	781	7.91%	9.85%
	Juvenile	1272	156	1116	3.71%	14.07%
	Life	620	236	384	5.61%	4.84%
	Literature	692	125	567	2.97%	7.15%
	Medical	948	440	508	10.45%	6.41%
	Relaxation	192	40	152	0.95%	1.92%
	Religion	710	369	341	8.76%	4.30%
	Science	1129	557	572	13.23%	7.21%
	Self-help	330	113	217	2.68%	2.74%
	Social Science	1643	519	1124	12.33%	14.17%
	Sports	423	91	332	2.16%	4.19%
Format	Paperback	5690	1799	3891	42.73%	49.07%
	Hardcover	5667	2251	3416	53.47%	43.08%
	Library Binding	308	7	301	0.17%	3.80%
	Other	475	153	322	3.63%	4.06%
Audience	College	1254	382	872	9.07%	11.00%
	General	6728	2409	4319	57.22%	54.46%
	Children	1609	208	1401	4.94%	17.67%
	Professional	2549	1211	1338	28.76%	16.87%

Table 3 Regression Result using Probit Model for PS Calculation

Coefficients:	Estimate	Std.	P-value	Significance
(Intercept)	-6.4844	1.1169	0	***
log(Rank)	-0.1506	0.0187	0	***
Price	-0.0065	0.0003	0	***
Pages	-0.0031	0.0002	0	***
Total	0.0025	0.0001	0	***
Year	0.1087	0.0101	0	***
log(Bing+0.1)	2.3871	0.3678	0	***
Business	-0.2041	0.0758	0.0071	**
Science	0.0084	0.0733	0.9087	
IT	-0.127	0.0735	0.0841	.
Relaxation	-0.741	0.1212	0	***
Religion	0.2898	0.0706	0	***
Medical	-0.0907	0.0754	0.229	
Juvenile	-0.5187	0.0952	0	***
Social	-0.3074	0.0635	0	***
Life	-0.2377	0.0738	0.0013	***
Sports	-0.6742	0.0907	0	***
Arts	-0.89	0.0803	0	***
Education	-0.3587	0.118	0.0024	***
Biography	-0.6416	0.1075	0	***
Self_help	-0.0499	0.1	0.618	
Literature	-0.7024	0.0828	0	***
Paperback	-0.2161	0.0349	0	***
Library	-0.8231	0.1828	0	***
Other	-0.2108	0.0775	0.0065	**
Professional	0.597	0.0838	0	***
General	0.5267	0.0747	0	***
College	0.1699	0.0899	0.0587	.
log(Price)	-1.4794	0.0786	0	***
log(Pages)	-0.7728	0.0437	0	***
Year*log(Bing+0.1))	-0.0218	0.0036	0	***
log(Price)*log(Pages)	0.3954	0.016	0	***
Signif.	*** 0.001	** 0.005	* 0.05	. 0.1
Observations: 12140	AIC:10148	Pseudo R2: 0.357		

Table 4 Calibration between Rank and Sales

	Model1	Model2	Model3	Model4
(Intercept)	13.419	8.656	16.832	21.780
	(0.032)	(0.1474)	(0.682)	(3.254)
log(Rank)	-0.982	0.101	-2.774	-5.111
	(0.004)	(0.033)	(0.237)	(1.521)
$\log(\text{Rank})^2$		-0.060	0.270	0.675
		(0.002)	(0.027)	(0.262)
$\log(\text{Rank})^3$			-0.012	-0.043
			(0.001)	(0.020)
$\log(\text{Rank})^4$				0.0009
				(0.0006)
R^2	0.9303	0.9416	0.9432	0.9432
AIC	4170.8	3178.8	3032.1	3031.7
BIC	4190.7	3205.3	3065.3	3071.5
Observations	5598			

Table 5 Sales impacts on Ranks from Experiment

ASIN	Copies	Rank Before	Day1 (9pm)	Day2 (3am)	Day2 (10am)	Day2 (9pm)	Day3 (9pm)	Day4 (9pm)	Day5 (9pm)	Day7 (9pm)
B002LLOTTO	1	476135	79473	108902	135066	170866	219293	262467	304469	372939
B001JQLTRM	1	479496	79511	108944	135122	170976	219638	263271	306378	382063
B004KSQDL8	1	513311	79537	109000	135237	171198	220159	264151	308057	386963
B005GA9AIW	1	540883	79547	109032	135303	171358	220500	264832	309464	392300
B001DS5EF4	1	634433	79682	109218	135494	171700	221051	265697	310932	395453
B004KKY7FK	1	918462	79633	109175	135521	171855	221472	266516	312642	402867
B000FC26XW	1	946167	79318	108917	135332	171697	221383	266456	312639	403156
B00585MZ8M	2	313210	44664	65708	90224	124602	174282	218961	261615	326795
B001IDYFOU	2	392860	44751	65970	90714	125178	174798	218761	259859	317456
B004P8JQG2	2	566692	44972	66438	91437	126185	176789	222831	267455	339910
B001O9C1N0	2	663721	44669	65980	91056	126003	176852	223153	268210	342875
B004OR1VOY	2	686930	44873	66305	91359	126224	176992	223343	268423	343434
B005JJT88C	2	888061	44712	66065	91143	126109	176989	223479	268774	345097
B002D48Q3E	2	956010	44981	66476	91547	126366	177184	223683	268989	345499
B004TAY1KW	3	291294	31842	47575	65281	103190	148284	191592	235292	292088
B004W0JQU4	3	350516	31923	47788	65669	103826	149242	192632	236729	295082
B0049P1O02	3	487527	32147	48121	66287	104763	151251	195714	242459	309759
B001AV7SRQ	3	603514	32107	48108	66314	104898	151643	196306	243606	312695
B001OW60RU	3	919689	32056	48040	66254	104903	151793	196602	244281	315309

Table 6 Sales estimated for different Rank intervals

Weekly Rank Range	200k-250k	250k-300k	300k-350k	350k-400k	400k-500k	500k-600k	>600k
Average Weekly Sales (Copies)	0.838	0.296	0.174	0.091	0.049	0.009	0.000

Table 7 Distribution of Continuous Variables after Matching

	Min	1st Q	Median	Mean	3 rd Q	Max
			KOOP			
Price	0.95	12.95	18	43.54632	39.95	679
Rank	54592	1130294	2024093	2990483	4138825	11121521
Bing	1	11	16	41.35923	24	7860
Pages	1	144	222	258.3271	320	1360
Year	1966	1997	2001	1999.315	2003	2005
Total	1	8	23	52.95281	66	393
			NOOP			
Price	1	13.99	20.95	43.63713	39.95	840
Rank	49842	1193953	2331409	3251640	4254757	11140568
Bing	0	11	15	39.12006	22	7380
Pages	1	145	240	269.2337	352	1434
Year	1965	1996	2000	1998.934	2003	2005
Total	1	8	23	57.00257	64	393

Table 8 Distribution of Categorical Variables Before and After Matching

		Before Matching		After Matching	
	Variables	NOOP	KOOP	NOOP	KOOP
Genre (Percentage)	Arts	2.99%	6.99%	5.36%	5.36%
	Biography	1.71%	2.59%	2.47%	2.89%
	Business	9.31%	5.16%	5.87%	6.10%
	IT	13.68%	6.71%	9.31%	8.15%
	Education	1.54%	1.72%	2.22%	1.64%
	Fiction	7.91%	9.85%	11.81%	12.04%
	Juvenile	3.71%	14.07%	5.46%	6.29%
	Life	5.61%	4.84%	7.16%	7.99%
	Literature	2.97%	7.15%	4.85%	5.36%
	Medical	10.45%	6.41%	7.38%	6.48%
	Relaxation	0.95%	1.92%	2.12%	1.89%
	Religion	8.76%	4.30%	7.35%	6.52%
	Science	13.23%	7.21%	8.12%	7.48%
	Self-help	2.68%	2.74%	2.89%	2.54%
	Social Science	12.33%	14.17%	13.96%	16.18%
	Sports	2.16%	4.19%	3.66%	3.08%
Format (Percentage)	Paperback	42.73%	49.07%	54.22%	55.51%
	Hardcover	53.47%	43.08%	41.44%	40.51%
	Library Binding	0.17%	3.80%	0.10%	0.13%
	Other	3.63%	4.06%	4.24%	3.85%
Audience (Percentage)	College	9.07%	11.00%	8.99%	8.83%
	General	57.22%	54.46%	66.00%	67.26%
	Children	4.94%	17.67%	7.51%	8.35%
	Professional	28.76%	16.87%	17.50%	15.57%

Table 9 Regression Result for Titles Released More than 75 Weeks ago

Coefficients	Estimate	Std.	P-value	Significance
(Intercept)	-854058	65089.89	0	***
log(Rank)	77845.88	3768.93	0	***
log(Price)	62681.67	4585.42	0	***
Week	-623.84	83.75	0	***
Juvenile	-14534.9	4236.31	0.001	***
log(Pages)	196193.1	24657.22	0.002	***
IT	156674.6	16645.03	0	***
log(Bing+0.1)	-12854.6	3296	0	***
Paperback	16502.98	6814.57	0.016	*
Social	130166.1	15642.92	0	***
Business	133148.6	16349.64	0	***
Science	131302.1	17310.76	0	***
Education	153671.1	25778.72	0	***
Life	122917.6	18466.1	0	***
Medical	116677.9	16991.07	0	***
Arts	132098.1	22167.58	0	***
Religion	101158.5	21363.48	0	***
Literature	85915.35	16480.03	0	***
Sports	103885.2	23747.28	0	***
Self_help	95553.75	21938.12	0	***
Biography	80463.95	24438.1	0.001	**
Relaxation	70156.98	29691.02	0.018	*
Signif.	*** 0.001	** 0.005	* 0.05	. 0.1
Observations: 2419		AIC: 64816.49		R2:0.455

Table 10 Regression Result for Titles Released fewer than 75 Weeks

Coefficients	Model1	Model2	Model 3
(Intercept)	332465.4 ***	-5568353.02 **	-5571879.05 **
	(49959.5)	(1967673.42)	(1968690.41)
Week	412.3		56.71
	(303.1)		(278.4)
log(Rank)		66286.69 ***	66211.2 ***
		(6588.35)	(6601.92)
log(Price)		57611.05 ***	57502.49 ***
		(7982.79)	(8004.37)
log(Pages)		-18857.1 *	-18792.86 *
		(7679)	(7689.14)
Paperback		44785.54 ***	44593.98 ***
		(11553.56)	(11597.28)
Juvenile		72919.56 ***	73292.12 ***
		(19625.98)	(19720.37)
Library		-234903.51 **	-235646.38 **
		(74235.25)	(74360.24)
Self_help		66643.53 *	66535.29 *
		(26728.77)	(26746.83)
Business		52205.44	51935.37
		(32650.03)	(32692.54)
log(Bing+0.1)		-13357.06 .	-13340.03 .
		(7647.22)	(7651.33)
Total		152.89 **	152.99 **
		(59.1)	(59.13)
Year		2475.17 *	2472.83 *
		(980.52)	(981.06)
Sports		-42775.73	-42607.18
		(27337.03)	27362.62
Education		-69979.09	-69901.5
		(45845.75)	45869.25
Signif. *** 0.001	** 0.005	* 0.05	. 0.1
Observations 1003			
R2	0.000836	0.186	0.185
AIC	27513.78	27317.26	27319.21

Table 11 Estimates of Coefficients using Bayesian Probit Model

Coefficients	Mean	S.D.	2.5% Quantile	Median	97.5% Quantile
(Intercept)	-6.9883	1.0491	-9.0559	-6.9747	-4.8952
log(Rank)	-0.1431	0.0185	-0.1801	-0.1429	-0.1071
Price	-0.0065	0.0003	-0.0071	-0.0064	-0.0058
Pages	-0.0022	0	-0.0023	-0.0022	-0.0021
Total	0.0024	0.0001	0.0022	0.0025	0.0027
Year	0.1101	0.0095	0.0913	0.1102	0.1286
log(Bing+0.1)	2.3925	0.345	1.714	2.3897	3.0684
Business	-0.155	0.0737	-0.3009	-0.155	-0.0095
Science	0.0734	0.0711	-0.0628	0.0724	0.2169
IT	-0.1127	0.07	-0.2502	-0.1123	0.0235
Relaxation	-0.6812	0.1183	-0.9163	-0.6805	-0.453
Religion	0.3171	0.068	0.1855	0.3176	0.451
Medical	-0.0315	0.0744	-0.1739	-0.0328	0.1146
Juvenile	-0.4961	0.094	-0.677	-0.4963	-0.3108
Social	-0.2646	0.0613	-0.3836	-0.2653	-0.1472
Life	-0.1878	0.0733	-0.3354	-0.1869	-0.0435
Sports	-0.6328	0.0909	-0.8065	-0.6317	-0.4553
Arts	-0.8236	0.0772	-0.9773	-0.8224	-0.6697
Education	-0.3184	0.1166	-0.5475	-0.3173	-0.0896
Biography	-0.6144	0.1051	-0.8179	-0.6137	-0.415
Self_help	-0.0401	0.0983	-0.2352	-0.0409	0.1505
Literature	-0.6651	0.0824	-0.8277	-0.6673	-0.5062
Paperback	-0.2429	0.0349	-0.3109	-0.2435	-0.1752
Library	-0.8805	0.1747	-1.2394	-0.8712	-0.5642
Other	-0.246	0.0786	-0.4035	-0.2476	-0.0875
Professional	0.6163	0.0835	0.4524	0.6152	0.7802
General	0.5102	0.0741	0.3679	0.512	0.656
College	0.1805	0.0883	0.0093	0.182	0.3549
log(Price)	-1.2262	0.0655	-1.3552	-1.2269	-1.1016
log(Pages)	-0.7372	0.0421	-0.8201	-0.737	-0.6549
Year*log(Bing+0.1)	-0.022	0.0034	-0.0286	-0.022	-0.0153
log(Price)*log(Pages)	0.3393	0.013	0.3141	0.3394	0.3648
Observations:	12140				

Table 12 Result from Bayesian PSM

Different Matching		Average Sale/Week (COPIES)	Average Revenue/Week (USD)	Number of NOOP matched
Without Matching		1.43	14.69	--
Conventional PSM	NNM	1.945	13.735	3115
	Stratification	1.71313	12.9409	3229
Nearest Neighbor Matching (Bayesian)	Min	1.168	9.478	2945
	25%	1.528	12.061	3148
	Median	1.649	12.798	3188
	Mean	1.668	12.876	3188
	75%	1.782	13.606	3226
	Max	2.6	17.994	3380
Stratification (Bayesian)	Min	1.443	11.62	3076
	25%	1.606	12.6	3235
	Median	1.659	12.87	3276
	Mean	1.671	12.88	3276
	75%	1.742	13.15	3317
	Max	1.975	14.7	3469

Table 13 Regression Result for PS and Release Decision

Coefficients	Estimate	Std.	P-value	Significance
(Intercept)	147.218	3.52	0	***
PS	-64.392	5.99	0	***
log(Total)	-13.558	10.6	0	***
PS*log(Total)	3.835	2.496	0.135	
Signif.	*** 0.001	** 0.005	* 0.05	. 0.1
Observations: 4207				R2:0.148

Table 14 Estimation of ATU using Subset of Samples

Different Matching		Average Sale/Week (COPIES)						Average Revenue/Week (USD)					
		Whole Sample	12000	10000	8000	6000	Avg of (2-5)	Whole Sample	12000	10000	8000	6000	Avg of (2-5)
		(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
Without Matching		1.43	1.43	1.39	1.55	1.56	1.48	14.69	14.72	14.60	15.56	15.19	15.02
Conventional PSM	NNM	1.95	1.80	1.74	1.48	1.72	1.69	13.74	15.53	14.05	11.21	14.24	13.76
	Stratification	1.71	1.70	1.57	1.73	1.78	1.69	12.94	13.85	13.08	13.45	13.89	13.57
Nearest Neighbor Matching (Bayesian)	Min	1.17	1.14	0.99	1.12	1.09	1.08	9.48	9.16	9.02	8.49	9.19	8.96
	25%	1.53	1.47	1.33	1.49	1.58	1.47	12.06	11.87	11.49	11.96	12.55	11.97
	Median	1.65	1.57	1.41	1.61	1.71	1.58	12.80	12.58	12.26	12.83	13.56	12.81
	Mean	1.67	1.57	1.42	1.62	1.72	1.58	12.88	12.63	12.30	12.87	13.63	12.86
	75%	1.78	1.67	1.51	1.74	1.86	1.69	13.61	13.35	13.04	13.71	14.62	13.68
	Max	2.60	2.18	2.03	2.26	2.59	2.27	17.99	16.96	16.88	18.49	19.78	18.03
Stratification (Bayesian)	Min	1.44	1.41	1.28	1.39	1.45	1.38	11.62	11.38	11.05	11.44	11.80	11.42
	25%	1.61	1.55	1.39	1.56	1.66	1.54	12.60	12.40	12.05	12.49	13.11	12.51
	Median	1.66	1.59	1.43	1.62	1.72	1.59	12.87	12.63	12.33	12.78	13.49	12.81
	Mean	1.67	1.59	1.43	1.62	1.72	1.59	12.88	12.63	12.32	12.80	13.50	12.81
	75%	1.74	1.63	1.46	1.67	1.79	1.64	13.15	12.86	12.58	13.10	13.89	13.11
	Max	1.98	1.76	1.64	1.87	2.02	1.82	14.70	13.69	13.86	14.70	15.45	14.43

Table 15 Result from Welfare Analysis*

Different Matching		Total Revenue	Retailer Profit	Publisher Profit	Consumer Surplus
Without Matching		763.88	212.43	485.68	888.23
Conventional PSM	NNM	714.22	191.51	436.86	830.49
	Stratification	672.93	181.83	414.28	782.47
Nearest Neighbor Matching (Bayesian)	Min	492.86	134.19	303.11	573.09
	0.25	627.17	170.27	387.31	729.27
	Median	665.50	180.36	410.83	773.83
	Mean	669.55	181.35	413.15	778.55
	0.75	707.51	191.40	436.61	822.69
	Max	935.69	250.29	574.00	1088.01
Stratification (Bayesian)	Min	604.24	164.39	373.57	702.60
	0.25	655.20	177.77	404.80	761.86
	Median	669.24	181.36	413.18	778.19
	Mean	669.76	181.38	413.21	778.79
	0.75	683.80	184.76	421.10	795.12
	Max	764.40	206.21	471.16	888.84

* Calculated based on randomly re-releasing a NOOP title with PS higher than 0.2. The unit for all the values in the table is USD.

Table 16 Regression Result for Price Elasticity

$\Delta \log(Price_i^t)$	>0	>0.1	>0.2	>0.3	>0.4
Elasticity	-1.86	-1.69	-1.53	-1.56	-1.66
	(0.52)	(0.52)	(0.53)	(0.62)	(0.74)
Observation	685	402	117	52	33
R2	0.017	0.023	0.058	0.095	0.109

Appendix: Summary of Method