Are verified online reviews on Amazon more trustworthy?

Li Liu*

3/18/2020

Abstract

Online reviews, usually written by the customers who previously purchased the products, are important indicators for product qualities. However, customers are losing trust in these reviews as they think many reviews are unauthentic. As a result, in November 2016, Amazon introduced the "Amazon Verified Purchase" label so that customers could voluntarily add this label to their review only if they purchased the products at Amazon and didn't receive the product at a deep discount. This project estimates the causal effect of having the verified purchase label on gaining other customers' trust by propensity score matching. Using reviews of a keyboard product as sample, my preliminary result suggests verified online reviews are not necessarily more trustworthy as they seem to be.

JEL classification: M31, C21, L81

Keywords: Online Reviews, Verified Purchase Label, Trust Propensity

*M.A. Student in Computational Social Science, The University of Chicago
Email: liu431@uchicago.edu
I would like to thank Prof. Guanglei Hong, Prof. Kazuo Yamaguchi, Simon Shachter, Arvind Ilamaran, Mengyuan Liang, and classmates in CHDV 30102 for helpful comments and suggestions. All remaining errors are mine.
Code and replication files are available at Github.

1 Research Question

The research question of this project is : what is the treatment effect of adding verified purchase labels for online reviews on gaining trust from other customers?

1.1 Background

Online reviews, usually written by the customers who previously purchased the products, are important indicators for product qualities. However, customers are losing trust in these reviews as they think many reviews are unauthentic. As a result, in November 2016, Amazon introduced the "Amazon Verified Purchase" label so that customers could voluntarily add this label to their review only if they purchased the products at Amazon and didn't receive the product at a deep discount.

Even though this policy aims at making the authentic reviews stand out, many people still complain that verified reviews are not more trustworthy. Thus, I plan to evaluate this policy by causal inference methods using the observational data.

1.2 Implication

Understanding how verified reviews gains trust have several important implications. For Amazon, their data scientists and economists could use the finding and research design to further help customers make better decisions from reading reviews. For customers, the finding could guide them to read reviews more smartly as there are too much information to digest when they are shopping on Amazon.

1.3 Road map

In the following sections, I will explain the data (sample and population), study design, measures, methods, identification assumptions, and external validity in detail.

2 Sample and Population

The data of Amazon product reviews has two parts: the reviews and product metadata. The data is scraped by several computer science researchers at the University of California, San Diego and made available for the public ((Ni, Li, & McAuley, n.d.)). There are two major advantages of using this data. Firstly, it has great breadth, as it contains 233 million reviews for various products in almost 30 product categories. Secondly, it has excellent depth, as it has reviews ranging from May

1996 to October 2018.

There are a lot of variables in the data. For each review, it has the reviewer's name, ratings, text, summary, attached image URL, and helpfulness votes (ratio of votes they received from all viewers), the label for verified purchase, etc. For each product, the data provides detailed information on its characteristics, such as color, price, package type, descriptions, technical details, similar products, image features, categories information, sales rank, etc.

For this project, the relevant population is reviews of all products. However, the reviews data is too large in scale and have many potential con-founders, such as product types. Thus, the sample would be all reviews for one particular popular product. I chose the product: Microsoft Natural Ergonomic Keyboard 4000 as my sample data, which is priced around $46 with a high average rating as shown in 1. There are 2567 reviews for this product in the data shown in 2.

Figure 1: Product page of Microsoft Natural Ergonomic Keyboard 4000 on Amazon

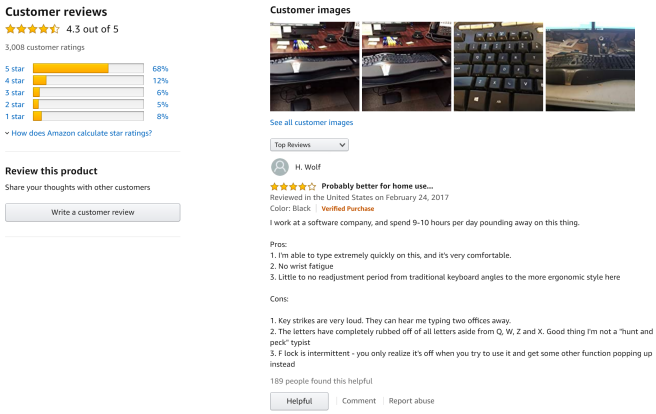

Figure 2: Sample online reviews of Microsoft Natural Ergonomic Keyboard 4000 on Amazon

3 Study Design

Since this data is scraped from Amazon website, it is observational and comes with many limitations. It contains almost all information that customers would see on the Amazon product pages. However, since the data was collected at one time by the researchers around November 2018, I do not know whether different customers might see different information under Amazon's many online experiments (such as A/B tests).

Most of the existing study with this data is for training better recommendation system by natural language processing. The provider of the data Jianmo Ni and Julian McAuley have three published paper in top CS conferences with this data. (Ni & McAuley, 2018), (Ni, Lipton, Vikram, & McAuley, 2017), (Ni et al., n.d.)).

My project is among the very few studies to find causal effect using this data without randomized experiment. It also shows this data could be analyzed for evaluating important business policy decision.

4 Measures

The outcome variable is the trustworthiness of the reviews. I conceptualize trustworthiness as the number of helpful votes that each review received from other customers. For example, in the sample review 2, there is a sentence on the bottom of the review text: "189 people found this useful", which suggests this review is very trustworthy as 189 other customers clicked the button "Helpful" after reading it in the past.

The treatment variable would be having the verified purchase labels. From the official definition on Amazon (*Amazon.com Help: About Amazon Verified Purchase Reviews*, n.d.), customers could voluntarily add this label to their review only if they purchased the products at Amazon and didn't receive the product at a deep discount. Since this policy was introduced around November 2016, I would expect there is a clear cutoff point around this time that reviews afterwards get more trust if the verified purchase label has a positive effect.

5 Methodological challenges

One challenge is that the reviews are conditional on product features and prices which are continuously changing. From this data, I do not know the associated features and prices when one review was written in the past. As a result, product features and prices are unobservable confounders in

my study design.

Another challenge is that customers who chose to write reviews might not be a representative sample of the whole population who have purchased the products. They are usually regarded as people who have extreme sentiments of the experience. Also, customers might be less likely to write reviews if there are already thousands of reviews for the product.

6 Evaluation of alternative methods

There are various candidate causal methods to estimate the true causal effect δ. However, different methods require different assumptions and variables. In this session, I will briefly evaluate the applicability of these methods to my research question.

6.1 Experimental Designs

In an experimental design, researchers have the power to randomly assign customers who have purchased to add the label or not. However, this is not the story behind this data. Amazon lets all eligible customers to have the labels.

The prima facie effect of the verified purchase label is the differences in means of vote numbers between the two groups, which could be expressed as $\delta_{PF} = E[Y(1)|Z = 1] - E[Y(0)|Z = 0] = 19.35 - 9.29 = 10$. Since this data is not randomized, the selection bias of this estimate is $\delta_{PF} - \delta$

6.2 Propensity Score Matching

Propensity score matching estimates the causal effect by adjusting for the observed observed covariates. This method is particularly helpful for my research question as I want to make sure the compared reviews are similar in terms of probability of having the verified purchased label.

6.3 Propensity Score Stratification

Propensity score stratification provides a coarse way of matching the observations by their propensity score. I will use method as an alternative way to validate the result from closest neighbor matching.

6.4 Inverse-Probability-of-Treatment Weighting

IPTW is an improvement to propensity score by applying weights to samples observations. However, I have not applied this method on the data and would expect the result to be similar.

6.5 Marginal Mean Weighting through Stratification

MMWS is an improvement to IPTW by stratification. Again, I have not applied this method on the data and would expect the result to be similar.

6.6 The Instrumental Variable Method

IT method is only good when there is a good exogenous IV candidate that only affects teh treatment. However, I have not yet found such IV candidate in the data.

6.7 Regression Discontinuity Designs

I tried RDD at the beginning. However, it turns out many reviews before November 2016 also have the verified purchase labels. Thus, the assumption for this method is not satisfied.

6.8 Difference-in-Differences Analysis

Similar to RDD, I have not figured out the reason why reviews written before November 2016 also could have the labels. So DID approach would also not work well when the time doesn't matter for the treatment assignment.

7 Optimal method

The optimal method is propensity score matching with Stratification

7.1 Pretreatment variables

7.1.1 Topics

Different reviews might talk about different topics. Firstly, I preprocessed the review text and used TF-IDF (frequency-inverse document frequency) (Li, 2018) to re-calculate each word's frequencies. Secondly, using the Latent Dirichlet Allocation model (Blei, Ng, & Jordan, 2003), I extracted the proportions of 5 latent topics for each review, where each topic is a combination of keywords and each keyword contributes a certain weight to one topic. (*Topic Modeling in Python with Gensim*, 2018). In this visualization 3, each bubble on the left-hand side plot represents a latent topic and the right-hand shows the weights of the associated words in a bar chat.

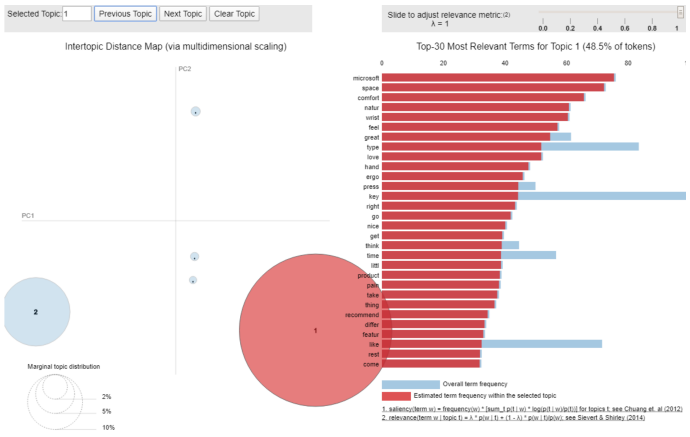

Figure 3: Visualization of 5 topics from LDA model

7.1.2 Number of associated images

Customers have the option to link images as part of the reviews. As a result, the more images that a review has, the more likely it has the verified purchase label with it.

7.1.3 Indicator of using a real name

Customers can either keep their name or remain anonymous when their reviews are shown to others. As a result, I expect that customers who choose to use their true name are more likely to add the verified purchase label. In order to get this indicator of using a real name, I require a "real name" to meet two conditions: 1. The gender of the first name could be guessed by gender-guesser package; 2. The name has at least two parts (ex. first name, middle name, and last name) (PyPI, 2016)

7.1.4 Sentiment

The customers' sentiment might change their propensity of adding the verified purchase label as the more extreme ones are more likely to add the label to earn trustworthiness.

7.1.5 Gender

Gender is an objective information that restaurants would usually ask the customers in the questionnaire. There is no variable on gender in the original data. We applied the gender-guesser package (PyPI, 2016) to guess gender from customers' first names (assuming they are the same as the user names). We labelled 43% of the customers as females.

I applied the VADER algorithm (Valence Aware Dictionary and sEntiment Reasone) to infer the sentiment level in the review, which is a lexicon and rule-based sentiment analysis tool trained with social media data (Hutto & Gilbert, 2014). The final score is a metric for magnitude of the sentiment intensity normalized between -1 and 1.

7.1.6 Length and Sentence Count

I suspect that longer reviews are more likely to have the verified purchase label. I used two ways of calculating the length of the reviews. The first is by counting the number of lexicons. The second is by counting the number of sentences.

7.1.7 Flesch reading ease score

Flesch score is a statistic of reading easiness of the text. This variable could potentially change the probability of having the verified purchase label. The intuition is that people who are more confident of their writing ability chose to looks trustworthy more.

7.1.8 Days since the first review

The review written time ranges from October 2005 to May 2015. As a result, I calculated the days of each review written date since the oldest review. This variable is important to control for in the propensity score as newer reviews are more likely to be verified.

7.1.9 Before/After verified purchase policy

This variable indicates whether the review was written after the introduction of verified purchase label policy (Nov 1st, 2016). If it is true, it is more likely to have verified purchase label with it.

7.1.10 Styles

There are two styles of the selected keyboard: retail or business. I created a dummy variable of indicating whether a review for the retail-version keyboard or not.

7.1.11 Summary of pretreatment variables

For the table 4, I noticed that verified reviews tend to higher customer ratings, fewer images, stronger sentiment, shorter in length, more readable, older, and more be written after Nov 2016.

verified	overall	Topic1	Topic2	Topic3	Topic4	Topic5	imagenum	nameverified	Sentiment	Length	Sentences	Flesch	Days	verified_option	Retail
0	3.59	0.49	0.04	0.4	0.05	0.03	1.26	0.08	0.42	140.30	4.76	35.79	2485.26	0.09	0.83
1	4.30	0.48	0.04	0.4	0.05	0.03	0.43	0.07	0.46	72.01	2.85	57.27	3234.58	0.13	0.74

Figure 4: Differences between pretreatment variables in verified and non-verified groups

7.2 Propensity score

The propensity score summarizes all the selected important pretreatment information that are predictive of the treatment. By using logistic regression, I calculated $\theta_i = \theta_i(X) = pr(Z_i = 1|X_i = x)$. The results from the logistic regression are shown in table 5. All 5 topics, sentiment, days, and verified option are significant at predicting the treatment (verified purchase label). I also saved the propensity score for each review.

```
glm(formula = verified ~ Topic1 + Topic2 + Topic3 + Topic4 +
    Topic5 + imagenum + nameverified + Sentiment + Length + Sentences +
    Flesch + Days + verified_option + Retail, family = binomial(link = "logit"),
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5882   0.3328   0.4124   0.5148   2.1378

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.044e+04  9.490e+03  -2.154 0.031272 *
Topic1           2.043e+04  9.491e+03   2.153 0.031338 *
Topic2           2.040e+04  9.490e+03   2.150 0.031574 *
Topic3           2.044e+04  9.490e+03   2.154 0.031245 *
Topic4           2.049e+04  9.490e+03   2.159 0.030842 *
Topic5           2.050e+04  9.490e+03   2.160 0.030764 *
imagenum        -4.655e-03  4.933e-03  -0.944 0.345350
nameverified    -2.583e-03  2.277e-01  -0.011 0.990948
Sentiment        2.872e-01  1.054e-01   2.725 0.006429 **
Length          -1.835e-03  1.241e-03  -1.479 0.139155
Sentences       -4.732e-02  2.982e-02  -1.587 0.112519
Flesch           8.577e-04  1.529e-03   0.561 0.574913
Days             6.419e-04  6.459e-05   9.937  < 2e-16 ***
verified_option -7.902e-01  2.336e-01  -3.383 0.000717 ***
Retail          -1.731e-01  1.703e-01  -1.016 0.309658
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5: Result from logistic regression

7.3 Identify common support

In the histogram 6, I found the distributions of the propensity scores for verified and non-verified groups are similar.

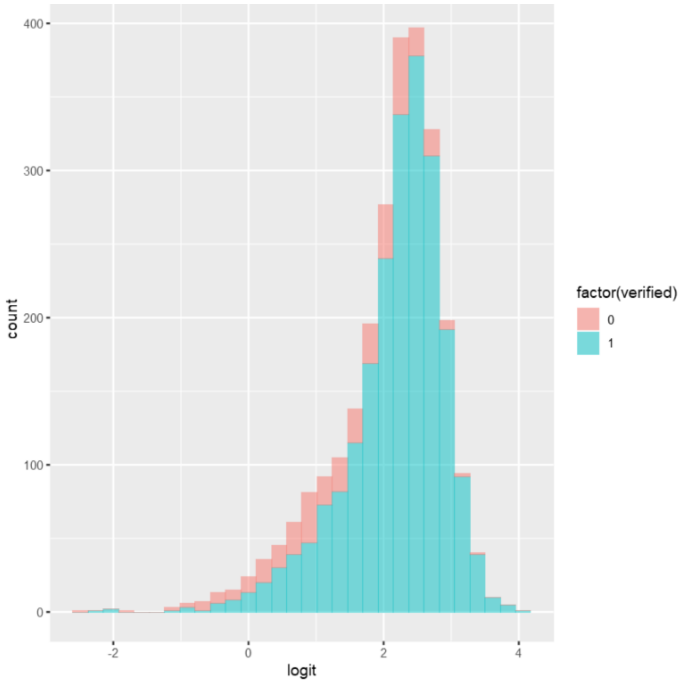

Figure 6: Histogram of propensity scores of the two groups

7.4 ATT estimate by closest neighbor matching

For each review with the treatment group, I matched it 1 by 1 with the review in the control group which has the closest propensity score. The caliper is set to 20% of the scores' standard deviation.

Then I calculated the average difference of the matched pairs in their number of votes, which is the average treatment effect on the treated. The result is 0.86, which is surprisingly small and suggests reviews with verified purchase label are not more trustworthy.

7.5 ATT estimate by stratification

Instead of 1 by 1 matching, I stratified the sample into 5 segments on the logit score and then estimated the within-stratum mean difference in the number of votes between treatment groups. The result is summarized in table 8. Even though the ATT is slightly different at different strata, all ATT estimates are all quite small, which also suggests reviews with verified purchase label are not more trustworthy.

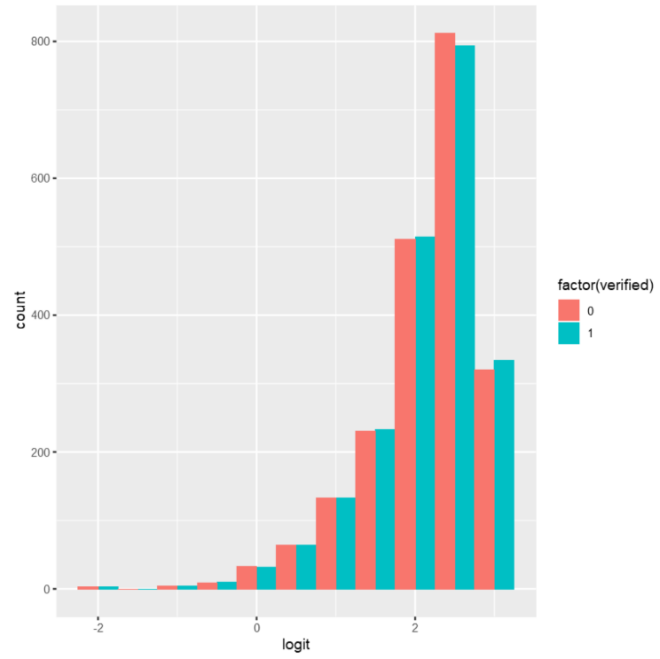

Figure 7: Histogram of matched propensity scores of the two groups

```
Call:
lm(formula = vote_fillna ~ verified + strata2 + strata3 + strata4 +
    strata5, data = strata_data)

Residuals:
    Min      1Q  Median      3Q     Max
  -6.01   -0.42   -0.31   -0.07 1106.99

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.358      1.470   2.964 0.003068 **
verified       1.656      1.476   1.122 0.261994
strata2       -5.596      1.578  -3.548 0.000396 ***
strata3       -5.702      1.578  -3.614 0.000307 ***
strata4       -5.887      1.605  -3.667 0.000250 ***
strata5       -5.940      1.612  -3.686 0.000233 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.37 on 2464 degrees of freedom
Multiple R-squared:  0.008262,  Adjusted R-squared:  0.006249
F-statistic: 4.105 on 5 and 2464 DF,  p-value: 0.001027
```

Figure 8: Result from logistic regression on effect of different strata

8 Identification assumptions

Propensity score matching method requires two assumption. The first on is conditional independence assumption. Under this assumption, the potential outcome of being treated is independent of being treated or not given the same propensity score. This assumption is likely to be held.

The second assumption is strong ignorability assumption. The assumption requires the treatment is as if randomized at different levels of observed pretreatent covariates. This assumption is likely to be plausible because the treatment might be much more likely at some levels of the covariates.

9 External validity

I would expect the results to be applicable to other similar products. The results might be different for products in other categories. However, the methods and analysis procedures could be easily applied for all products on Amazon. As a result, my future research agenda is to calculate the causal effects of having verified purchase labels for a wide range of products. Then I can run regression analysis to understand how product metadata might change the causal effects.

References

Amazon.com Help: About Amazon Verified Purchase Reviews. (n.d.). Retrieved 2020-03-18, from https://www.amazon.com/gp/help/customer/display.html?nodeId=202076110

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Hutto, C., & Gilbert, E. (2014, June). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Ann Arbor, MI.

Li, S. (2018, June). *Topic Modeling and Latent Dirichlet Allocation (LDA) in Python.* Retrieved 2020-03-18, from https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24 (Library Catalog: towardsdatascience.com)

Ni, J., Li, J., & McAuley, J. (n.d.). Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. , 10.

Ni, J., Lipton, Z. C., Vikram, S., & McAuley, J. (2017, November). Estimating Reactions and Recommending Products with Generative Models of Reviews. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 783–791). Taipei, Taiwan: Asian Federation of Natural Language Processing. Retrieved 2019-10-30, from https://www.aclweb.org/anthology/I17-1079

Ni, J., & McAuley, J. (2018, July). Personalized Review Generation By Expanding Phrases and Attending on Aspect-Aware Representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 706–711). Melbourne, Australia: Association for Computational Linguistics. Retrieved 2019-10-30, from https://www.aclweb.org/anthology/P18-2112 doi: 10.18653/v1/P18-2112

PyPI. (2016). *gender-guesser 0.4.0.* Retrieved 2019-05-21, from https://pypi.org/project/gender-guesser/

Topic Modeling in Python with Gensim. (2018, March). Retrieved 2020-03-18, from https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/ (Library Catalog: www.machinelearningplus.com Section: NLP)