UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Faculty of Business and Economics

The Travelling Reviewer Problem - Location-Based Anchoring Effects In Online Ratings

---

Research In Progress

by
Jürgen Neumann
Student of Business Information Systems (Master's Program)

31st December 2016

1 Motivation

"Life begins at the end of your comfort zone." Saying this sentence, Neale Donald Walsch encourages people to try out something new in life. This gives rise to the following question: Do people have different expectations when being out of their comfort zone, i.e., in an environment that is unusual for them? Naturally, consumer expectations are highly important for businesses. Meeting a consumer's taste is essential to build a good reputation and increase the sold quantity in the future. Thus business owners need to know whether people assess a business's quality differently just because the business is located in an *unusual area* for the consumer. As the quote of Walsch suggests, people should feel more alive in a new area and therefore assess quality more positively. Consequently, a business with a high share of people from other cities (*unusual raters*) should have a better reputation than one with a high share of local residents (*usual raters*). Online rating platforms like Yelp are a popular way to find out about a business's reputation and use this information to decide on a business. Consumers highly rely on such online ratings to reduce information asymmetry between business owners who possess complete information on their offered quality and consumers who cannot assess quality until buying and experiencing a product or a service. Does a reviewer change her rating behaviour if she travels away from her *usual rating area*[1]?

In this context, I use the Yelp Academic Dataset of the Yelp Dataset Challenge (Round 8) to answer the following research question:

Do people systematically rate a business's quality differently when it is not located in their usual rating area?

Answering this question provides several management implications:

Business owners: If people tend to rate businesses better which are not located in their usual area, business owners could specifically target (e.g., create discounts for) people who come to visit from other cities. If they tend to rate businesses worse, business owners could target local residents instead. Contrary, if business owners know that a person tends to give a worse rating, they could try to improve service quality for them.

Consumers: As a consumer, one could use the answer to the research question to adjust one's perception of online ratings. If a consumer wants to decide on a business in

[1] Note that, in my code *usual rating area* is called *home* or *hometown*.

an unusual area, she should give a higher weight to reviews from unusual raters.

Online rating platforms Platforms like Yelp could improve recommendations or a consumer's search, for instance by introducing a badge which helps distinguishing between usual and unusual raters.

To answer my research question, I define a concept for the term of *usual rating area* and prepare the data accordingly. I use econometric models like the instrumental variables or the fixed effects approach (see Angrist and Pischke (2009) for instance) to correctly estimate the effect of rating in a usual rating area on review stars. On a business level, I estimate models to measure the effect of a high share of reviews from usual raters on a business's overall rating. I enrich the Yelp Academic Dataset of Round 8 by incorporating data from city-data.com and the United States Census Bureau. My results suggest that usual raters systematically post reviews with a higher star rating. Additionally, I also find that usual raters write more useful reviews. Furthermore, I find that, with a growing number of reviews a user has posted for businesses in the same city, future reviews in this city have more stars on average. In other words, if a consumer posts a lot of ratings in a city, later reviews in this city tend to be more positive.

2 Previous Literature

Literature research still needs to be done. Literature streams which deal with quality expectation formation and expectation disconformation need to be analyzed.

3 Research Setup

To answer my research question, the concept of a user's *usual rating area* needs to be defined. Since there is no information on current place of residence or hometown in the dataset, I define this area/city with three aspects:

1. The user has posted at least 50% of her total review count in this city.

2. The date difference between any two reviews or tips the user has posted is at least 360 days.

3. The user has posted at least three tips and/or reviews in this city.

Firstly, the area needs to be a city in which the user has posted at least 50% of his total review count. This ensures that the user has a clear attachement to the city which others do not. It is also advantegous to use such an endogenous definition of this area since the hometown or current place of residence provided by any user could be intentionally wrong. However, this aspect also postulates that users who do not exhibit such a strong concentration of reviews in a single city do not have an attachement to any city. This may be a limitation to this research setup but it is also reasonable to state that those users are less attached to such a city. In further variations to my econometric model, I relax this restriction.

Secondly, between any two reviews or tips the user has posted in this city, there needs to be a date difference of 360 days. Consider a situation in which a user visits a new city for holiday and posts lots of reviews or tips since she is trying out a lot of different businesses. This aspect ensures that short holidays do not determine the usual rating area and that the area is visited more frequently.

Thirdly, restricting this definition to those users who have posted at least three tips and/or reviews in the city makes sure that a city is only considered a usual rating area if the user reviews often enough.

Generally, I always take tips into consideration. They provide additional information with regards to a user's location and therefore indicate the frequency of visits. Thus, when I calculate the maximum day difference between reviews and tips, I first sort all dates of both, reviews and tips, into one list before I calculate the differences in days.

There are a lot of users for which I am not able to determine a usual rating area. Some of them are those people who rate in a lot of different cities but may be located in one of the cities I can observe. These are the ones for whom it is reasonable to state that they are less attached to any city (see above). But there are also other users whose usual rating area is a city I do not observe. For this subgroup I can reliably tell that their usual rating area is not a city contained in the dataset because, according to a statement from the Yelp-Dataset-Challenge-Team, the dataset contains all recommended reviews for the selected cities from businesses with at least three recommended reviews. Thus, if a user is not observed suffiently often in the dataset to determine her usual rating area, it is highly probable that most of her reviews are posted for businesses in other cities.

4 Data Preparation

I use a C# program to transform the data from JSON-format to CSV-format. Also, I use this transformation step to determine every user's usual rating area according to my description in Section 3. For each business I also count for how many reviews this business is inside, respectively outside, of the corresponding user's usual rating area.

Since average ratings also might be influenced by properties of the city or the ZIP code area, I determine the ZIP code of a business from its full adress. The city attribute is often filled with wrong-spelled city names. Therefore, using data I downloaded from http://federalgovernmentzipcodes.us, I map the ZIP Code to the corresponding city name. This also means that I restrict my dataset to US cities.

To get demografic data for the cities in the dataset, I used a customized C# web crawler to crawl information from http://www.city-data.com/ on 4th of December 2016. This information includes variables for population (POP), percentage change of population since 2000 (POP_CHANGE), median age (MED_AGE), median rent (MED_RENT), median household or condo value (MED_HVAL), median income (MED_INCOME), unemployment rate in percent (UNEMPRATE), share of Hispanic people in percent (HISP_RATE), share of Asian people in percent (ASIAN_RATE), share of black people in percent (BLACK_RATE) and share of American Indian people in percent (AMER_IND_RATE). The number of businesses who offer travel accommodation (e.g., hotels or bed and breakfasts) may influence the share of those reviewers who do not consider this area to be their usual one. I include such data from the United States Census Bureau (https://factfinder.census.gov) on a ZIP code level. I also exclude all businesses with category "Hotels", "Hotel & Travel", "Hostels" and "Bed & Breakfast" to ensure no interference with this additional data.

It could also be useful to examine businesses at multiple points in time. To this end, time-variant data for cities is necessary. Since city-data.com provides data which is as recent as possible, I created another C# program which uses the API of https://archive.org/web/ to create a list of URLs. These URLs lead to earlier snapshots of the city-data.com page for the different cities under consideration. Each snapshot is the one closest to the date I send to the API. The multiple dates I include are the first of January 2012, 2013, 2014 and 2015. For these snapshots I used the web crawler on 18th of December 2016 to collect the same variables as mentioned above. Correspondingly, I

create snapshots of all businesses at these points in time by calculating cumulative averages and review counts. I collected the data from the US Census Bureau for the years 2012, 2013 and 2014. Finally, I use all this data in CSV-format to import it into STATA SE 13 and estimate econometric models.

5 Empirical Methodology

I estimate econometric models on two different levels, namely review level and business level. The first allows to describe a relationship of posting a review in a usual area and review stars. The latter addresses the relationship between the share of usual raters and the business average rating.

5.1 Review Level

On a review level, I describe the model with the regression formula in Equation 1.

$$Y_i = \alpha_0 + \alpha_1 \cdot in\_usual_i + \alpha_2 \cdot city\_controls_i + \alpha_3 \cdot user\_controls_i + \alpha_4 \cdot business\_controls_i + \epsilon_i \quad (1)$$

For the outcome variable Y_i, I test two different outcomes: Review i's star rating and its usefulness rating. α_1 is the value of interest since it measures the effect of a rating in a usual area on the outcome variable. Accordingly, *in_usual* is an indicator variable. The vector *city_controls* contains all variables mentioned in Section 4 for the city of the review's corresponding business. User information and preferences is stored in the vector *user_controls*. This includes a user's cumulative average rating up to the time of review i, her overall average rating corrected by the review's current rating, the cumulative review count, the number of fans, the number of years a user has been an elite yelper and the number of reviews a user has made in his usual rating area. ϵ captures the effect of unobservable variables which affect the outcome. *business_controls* contains the business's cumulative average rating, its corrected overall average rating, its cumulative review count and an indicator variable whether the business is still operating.
If unobservable variables correlate with *in_usual*, the estimation of α_1 is biased (omitted variable bias). For instance, the age of a user could affect how often she leaves her usual rating area and it may also affect how useful her reviews are. To overcome this

shortcoming, I use the instrumental variables approach and estimate a 2-Stage-Least-Squares-Estimator. To this end, I use the number of establishments offering travel accommodation in the corresponding ZIP code area as an instrument. Thus, I estimate the model in Equation 2 to predict a value for in_usual.

$$\begin{aligned} in\_usual_i = \beta_0 + \beta_1 \cdot number\_of\_establishments_i \\ + \beta_2 \cdot city\_controls_i + \beta_3 \cdot user\_controls_i + \beta_4 \cdot business\_controls_i + \zeta_i \end{aligned} \quad (2)$$

$number\_of\_establishments$ contains the number of establishments offering travel accommodation. After running this regression, I use the estimates to obtain the prediction $\widehat{in\_usual}_i$. Then, I find the predictor $\hat{\alpha_1}$ with Equation 3.

$$\begin{aligned} Y_i = \tilde{\alpha_0} + \tilde{\alpha_1} \cdot \widehat{in\_usual}_i \\ + \tilde{\alpha_2} \cdot city\_controls_i + \tilde{\alpha_3} \cdot user\_controls_i + \tilde{\alpha_4} \cdot business\_controls_i + \nu_i \end{aligned} \quad (3)$$

The instrumental variables approach solves omitted variable bias and reverse causality. The latter would mean that review stars would have an effect on whether the review is posted in a usual area. This seems unreasonable. However, even though the instrumental variables approach solves omitted variable bias, it bears several assumptions which I discuss later.

To relax the restrictions imposed by my definition of usual rating area, I also measure the effect of the number of reviews a user has made so far in the city the business is located in ($reviews\_in\_city$). This model tests whether a user systematically alters her reviews if she has already posted a lot of reviews in the corresponding city, i.e. the user has „invested “ time by writing reviews for businesses in an area and may feel more or less attached to it. I also use the instrumental variables approach and describe this model with the structural Equation 4. Note that I exclude the number of reviews a user has made in his usual rating area from the vector $user\_controls$ for this model, since it would interfere with the effect I want to measure.

$$\begin{aligned} Y_i = \gamma_0 + \gamma_1 \cdot reviews\_in\_city_i \\ + \gamma_2 \cdot city\_controls_i + \gamma_3 \cdot user\_controls_i + \gamma_4 \cdot business\_controls_i + \epsilon_i \end{aligned} \quad (4)$$

5.2 Business Level

On a business level, I also use the instrumental variables approach but also incorporate a fixed effects approach. The variable of interest is the share of reviews posted by usual raters (i.e. the ratio of reviews from usual raters to the number of observed reviews in total). The model is depicted in Equation 5.

$$Y_i = \delta_0 + \delta_1 \cdot usual\_share_i + \delta_2 \cdot city\_controls_i + \delta_3 \cdot business\_controls_i + \epsilon_i \quad (5)$$

The outcome variable Y is now the observed average rating of a business. I use the same city-specific control variables as in Section 5.1. As business-specific control variables, I use the number of reviews which can be observed in the dataset, the number of not recommended reviews (i.e. the number of reviews which cannot be observed) and an indicator whether the business is still operating. Again, I use the number of establishments offering travel accommodation to estimate the 2-Stage-Least-Squares-Estimator (see Equation 6).

$$\begin{aligned} usual\_share_i = \theta_0 + \theta_1 \cdot number\_of\_establishments_i \\ + \theta_2 \cdot city\_controls_i + \theta_3 \cdot business\_controls_i + \zeta_i \end{aligned} \quad (6)$$

As stated in Section 4, it could be useful to create a panel dataset out of the Yelp dataset. Since businesses can be observed at multiple points in time, I aggregate data for the first of January for each year from 2010 to 2015. For every business, the average rating at this point, its review count and the share of usual raters can be computed. Since I also collected time-variant data for each city, I can estimate a fixed effects model as depicted in Equation 7. Note that the number of not recommended is not included in the business-specific control variables as it cannot be observed over time.

$$\begin{aligned} Y_{it} = \pi_0 + \pi_1 \cdot usual\_share_{it} \\ + \pi_2 \cdot city\_controls_{it} + \pi_3 \cdot business\_controls_{it} + \pi_4 \cdot business\_ind_i + \epsilon_{it} \end{aligned} \quad (7)$$

The control variables now vary other time. Business fixed effects are included with an indicator variable for each business. Thus, each unobserved variable which is time-invariant for each business is accounted for and does not lead to omitted variable bias. I also combine the instrumental variables approach with the fixed effects approach for the years

2012 to 2014 with the same instrument as before.

6 Empirical Results

In this section, I present the results to all the models mentioned in Section 5. Moreover, I discuss the identifying assumptions of the instrumental variables approach. The results also include the regular regression (ordinary least squares) as a baseline model.

6.1 Review Level

Results on a review level are presented in Table 1 and Table 2. The results from the first (column 2 and 4) suggest that a review, which is posted for a business located in a usual area, has on average 0.6754 stars and 0.5310 usefulness votes more than those posted for businesses from an unusual area. The results in the second table (column 2 and 4) explain these results in detail. They indicate that one more review in the business's corresponding city leads to an increase in review stars by 0.0636 stars (i.e., 0.636 stars for 10 additional reviews) and an increase in usefulness votes by 0.0460 votes (i.e., 1 vote for 22 additional reviews). Thus, posting more reviews in a city, as it happens in the case of a usual rating area, leads to assessing the quality rather positive and to posting more useful reviews.

6.2 Business Level

The results in Table 3 are in line with those presented in Section 6.1. An increase in the share of reviews posted by usual raters leads to an increase in the observed average rating.

Table 1: Coefficients for rating in usual rating area on review level from ordinary least squares and instrumental variables regressions

	(1) Ordinary Least Squares	(2) Instrumental Variables	(3) Ordinary Least Squares	(4) Instrumental Variables
Outcome Variable	*review_stars*	*review_stars*	*votes_useful*	*votes_useful*
in_usual	-0.0686*** (0.0025)	0.6754*** (0.0154)	0.2053*** (0.0037)	0.5310*** (0.0268)
user_controls	✓	✓	✓	✓
business_controls	✓	✓	✓	✓
city_controls	✓	✓	✓	✓
Observations	2,269,139	2,269,139	2,269,139	2,269,139

Robust standard errors in parantheses. Instrument: Number of establishments offering travel accommodation.
***: 1%-significance; **: 5%-significance; *: 10%-significance

Table 2: Coefficients for reviews posted in the city on review level from ordinary least squares and instrumental variables regressions

	(1) Ordinary Least Squares	(2) Instrumental Variables	(3) Ordinary Least Squares	(4) Instrumental Variables
Outcome Variable	*review_stars*	*review_stars*	*votes_useful*	*votes_useful*
reviews_in_city	-0.0003*** (0.0000)	0.0573*** (0.0022)	0.0038*** (0.0002)	0.0460*** (0.0026)
user_controls	✓	✓	✓	✓
business_controls	✓	✓	✓	✓
city_controls	✓	✓	✓	✓
Observations	2,269,139	2,269,139	2,269,139	2,269,139

Robust standard errors in parantheses. Instrument: Number of establishments offering travel accommodation.
***: 1%-significance; **: 5%-significance; *: 10%-significance

Table 3: Coefficients for share of reviews from usual raters on business level from ordinary least squares and instrumental variables regressions

	(1) Ordinary Least Squares	(2) Instrumental Variables	(3) Ordinary Least Squares	(4) Instrumental Variables
Outcome Variable	*observed_avg*	*observed_avg*	*observed_avg*	*observed_avg*
usual_share	-0.0010*** (0.0002)	0.0073*** (0.0009)	0.0005*** (0.0002)	0.0384 (0.0370)
city_controls	✓	✓	✓	✓
business_controls	✓	✓	✓	✓
Business Fixed Effects	-	-	✓	✓
Observations	66,152	66,152	55,450	46,981

Robust standard errors in parantheses. Instrument: Number of establishments offering travel accommodation.
***: 1%-significance; **: 5%-significance; *: 10%-significance

Results in column 2 suggest that a 10%-increase of this share increases the average rating on average by 0.072. For businesses with an average rating of 4.43, such an increase would lead from a 4- to a 4.5-star rating. The fixed effects model (column 3) also suggests that reviews posted by usual raters are more positive, but the economic effect is small. The effect grows when I use the instrumental variables approach but no statistical significance (column 4) is found. Reasons for this could be the decrease in observations since observations on the instrument are only available for 2012 to 2014. Furthermore, there could be only little variation in the instrument over time which would make it difficult to estimate the within-variation of a business.

6.3 Discussion Of Identifying Assumptions

For the instrumental variables approach to work properly and consequently measure a causal effect, three assumptions with regards to the instrument need to hold. I discuss all three assumptions and their validity.

Exclusion restriction: For the exclusion restriction to hold, there should not be a direct

influence of the number of establishments offering travel accommodation on the review stars/number of usefulness votes/observed average rating. A high number of competitors surrounding a business could influence the rating. Since a user has additional choices, the decision could be more challenging or it could make her happier to choose from a broader set of businesses. But since the establishments offering travel accommodation are not competing with the businesses in my adjusted dataset, I argue that the exclusion restriction holds.

Instrument strength: The number of establishments offering travel accommodation should be highly correlated with the outcome variables. As commonly done, I conducted a F-test to check for this correlation. All values highly exceeded the usual bound of 10.

Independence: The instrument should be as good as randomly assigned. Whether this statement is true, is debatable. It could be the case that businesses, which are located close to a lot of establishments offering travel accommodation, generally offer a better quality. It could also be the case that users who visit such businesses tend to give more positive ratings. If this would be true, the effect estimated by the instrumental variables approach would be an overestimation. To this end, I also estimated a fixed effects model. Even though the coefficient is rather small, it is not biased by any location-based, time-invariant aspect. The estimated effect of the fixed effects approach points into the same direction as the one estimated by the instrumental variables approach. Thus, concluding that reviews posted in a rater's usual rating area have a higher star rating is not debatable even though the results based on instrumental variables may be an overestimation.

7 Conclusion

Using econometric models, I answered the question posed in the introduction to the Yelp Dataset Challenge: Do reviewers change their behaviour when they travel? My results indicate that they do. Leaving the area they usually rate in leads to a decrease in stars given to a business. Moreover, the more reviews a user posts in a city, the more positive her future reviews in this city tend to be. Thus, users seem to are more attached to the cities they post more reviews in. Consequently, businesses might need to adapt to this

phenomenon in order to build a more postive reputation on Yelp.

Naturally, a lot of work has to be done to further improve this research. For instance, more instruments need to be tested in the instrumental variables approach, the main drivers in the dataset for this phenomenon need to be found and an extensive literature research needs to be done.

Appendix

Explanation of source data files

- aggregated_businesses.csv: Panel data generated from the Yelp Academic Dataset
- businesses.csv: Business data generated from the Yelp Academic Dataset
- Controls.csv: City-specific variables crawled from http://www.city-data.com/ on 4th of December 2016
- reviews.csv: Review data generated from the Yelp Academic Dataset
- TravelAccommodationData (Folder): Business Patterns 2012, 2013, 2014 obtained from http://factfinder.census.gov/ using NAICS code 7211.
- users.csv: User data generated from the Yelp Academic Dataset
- WaybackCrawl.csv: City-specific data from http://www.city-data.com/ obtained through snapshots created by https://archive.org/web/ with a web crawler on 18th of December 2016
- WaybackList.txt: List of urls obtained from API of https://archive.org/web/ with the attached C# program

References

Angrist, Joshua David and Jörn-Steffen Pischke, *Mostly harmless econometrics: an empiricist's companion*, Vol. 1, Princeton university press Princeton, 2009.