Experimental_Methodology

Mark Paluta, Krysten Thompson, Chris Ventura

April 19, 2019

Contents

1 Introduction / Background / Motivation

###–> CHRIS, can you link any of comments below to the articles under "References"? I see there's reference to NYT article but not sure if others referenced below. (You are likely most familiar w article content.)

In a recent study, Zillow found that the share of adults living with a non-spouse, non-partnered roomate increased from 21% in 2005 to 30% at the end of 2017. Whether it is due to changing attitudes about living with roommates, a neccesity due to increased rents in major cities, or some other factor, the trend is clear: more Americans are cohabitating with someone other than a spouse or partner. While college and post-college aged urbanites often expect to have roommates, the trend is increasing for those older than 30, as a recent New York Times profile depicts. As employment opportunities continue to cluster in dense metropolitan areas such as New York and San Francisco, an older cohort may be forced to cohabitate in order to pursue economic opportunity.

In spite of this demographic trend, little research has been done into whether the age of someone looking for a place to live with roommates has an impact on their options. As the New York Times profile describes, until recently the idea of having a roommate was mostly relegated to young professionals in their 20s.

2 Research Question

Through this experiment, we set out to determine if older professionals face a penalty when seeking a roommate compared to someone in their 20's. We investigate whether age affects replies when responding to a roommate posting on Craigslist as well as determine whether age has a statistically significant effect on the odds of what we will call a "favorable reply".

We hypothesized that the "older" age treatment would receive fewer favorable replies than the 20-something treatment.

3 Initial Experiment Idea

Our initial experiment sought to test gender and age discrimination. We created two email accounts for a female (age 27 and 43) and two email accounts for a male (age 27 and 43). We chose these ages to avoid round numbers, have the 20-something presumed to be slightly more established in a job/career than someone who is 22 or 23 (too young of an age may have caused bias), and have the older person at least over 40 but not yet 45 which may have caused bias as well.

We also considered running this experiment in several different cities in the U.S. to see if regional differences exist in terms of gender and age.

A pilot was conducted in early March to determine if our initial approach would work and identify potential obstacles. We discovered issues with Craigslist and Gmail fraud mechanisms that led us to modify our approach. See Pilot Study below for further details.

3.1 Pilot Study and Outcomes

We conducted a pilot study for three days to identify any problems that may surface during the full study. Major learnings from the pilot:

1) VPNs are not particularly compatible with Craigslist. We initially wanted to use a VPN to avoid being flagged as fraud by Craigslist. We tried several but Craigslist still flagged us as fraud when it detected that we changed email accounts.

2) Google sent one of our accounts a notification that it was shutting down one of our accounts because it detected the account was not being used according to Google's guidelines.

We concluded that maintaining too many personas would be difficult due to concerns about tripping fraud algorithms and getting our real IP addresses blacklisted. As a result, we decided to limit our full study to 2 females. We theorized that females would have higher response rates due to more postings seeming open to females, and suspected that males would be more willing to live with females than females would be willing to live with males.

Not all posts seemed appropriate for measuring effects on a typical applicant. The following types of posts were excluded:

a) Posts with sexual implications including "friends with benefits", reduced rent in exchange for sex
b) Posts soliciting other favors in exchange for reduced rent such as babysitting
c) Posts for students only
d) Posts without an email option or posts explicitly stating that email replies will not be answered
e) Posts explicitly stating males only

4 Experiment Design & Methodology

The experiment design consisted of sending emails to listings under the Craigslist heading "Housing" > "Rooms/Shared".

In order to be able to draw conclusions that could be somewhat generalized to the US population, Indianapolis was selected as the target market. Indianapolis is a mid-sized city located in the midwest.

We sent emails every 1-3 days across a two-week time period. We were not particularly concerned with the exact times or days we collected data since we were collecting a number of time-based covariates in order to not confound our results with any temporal effects. We did make a conscious effort to cover a broad range of variation in our temporal covariates by sending emails on every day of the week and including a mix of morning, afternoon, and evening collections.

The following procedure was implemented each time emails were sent:

1) Open all postings since the last data collection (or fewer if subject to time constraints, prioritizing recent postings)
2) Go through each post and throw out any that meet our exclusion criteria
3) Obtain a count of the number of included (remaining) posts
4) Randomize an array assigning these posts to our 27-year-old or 43-year-old persona.
5) Beginning with the first persona in the array, log into their respective email account and begin emailing posters in order from most recent to least, recording covariates as we go

Emails were sent only to listings that were new, or listings that had been posted days or weeks ago but were "refreshed" in the postings list. There were only 1-2 instances where a listing received more than one "Katie" email and those observations were deleted from the final data set.

A short R radomization function was run prior to sending out the emails to radomize which listing would receive a "Katie 27" or "Katie 43" email.

The following was collected by manually entering data in a Google Sheet:

- Listing post date/time
- Email sent to listing date/time
- Treatment (Katie 27, Katie 43)
- Listing title, description
- Age and gender of person posting listing (if avail)
- Whether reply was received, along with date/time
- Whether reply was favorable

Favorable replies were defined as those received within 72 hours of a "Katie" email being sent and suggesting a time to show the unit or asking "Katie" for more information. An unfavorable reply stated the unit was already rented.

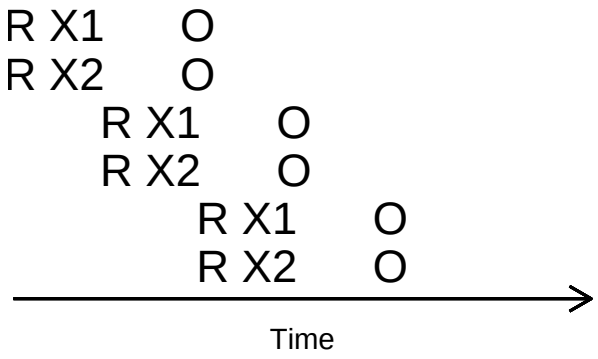
- Add flow diagram <Add diagram of response flows (emails/listing -> replies -> favorable replies)> **Clear statement of the experiment**

  Research Design (using ROXO grammar) Randomization engineering Experimental materials (e.g. treatment materials) Measurement of variables Modeling choices **end**

4.1 ROXO Grammar

Treatment was applied at various times since our data was collected over several weeks. We did not always have completed data on the previous group (group of 10-20 emails sent) prior to treating a new group of listings due to our 72 hour waiting period. This was not a concern as we believed there to be little chance posters would interact with each other. We demonstrate these characteristics using ROXO grammar in the following diagram.

Figure 1: Katie Email



5 Data

5.1 Data Types

The following variables were collected: spanning posting metadata, contents for archiving purposes, time information for temporal effects, covariates, and treatment applied. As we were not sure how clean our data would be and what would prove useful, we erred on the side of collecting more variables than would likely be needed.

Variable	Description	Example	% Missing
Post Title	Title of the Craigslist post	Roommate/shared house ASAP	0%
Post Body	Body of the Craigslist post	ALL INCLUSIVE 10 minutes from downtown...	0%
Post URL	URL of the Craigslist post	https://indianapolis.craigslist.org...	0%
Poster Age	Age of the resident who posted	34	81%
Poster Gender	Gender of the resident who posted	Female	64%

Variable	Description	Example	% Missing
Listing Type	Type of residence	Apartment	8%
Listing Price	Price of the listing	450	0%
Posting Timestamp	Timestamp that the post went up	3/21/2019 22:38	0%
Email Sent Timestamp	Timestamp that we emailed	3/22/2019 6:59	0%
Treatment	Which alias we emailed from	Katie 27	0%
Reply	Did we receive a reply?	Y	0%
Favorable Reply	Did we receive a favorable reply?	Y	0%
Reply Timestamp	Timestamp of a reply, if received	3/23/2019 16:53	0%
Weekend	Was our email sent on a weekend?	Y	0%

In total, we sent 113 emails - 59 for "Katie 27"" and 54 for "Katie 43". Several variables were missing a significant portion of data and were note used in our models. In particular, the variables 'poster age' and 'gender' could have been predictive had more data been collected. However, most postings did not include the poster gender or age. Our hypothesis would be that all else being equal, people prefer to live with those of the same gender and/or age.

Poster.Gender	Katie.27
Female	23.00
Male	55.00

Table 2: Response rate by treatment and gender

Since some replies we received did contain further information, which sometimes included a name, age, or gender, we considered filling in some missing data from replies where possible. We eventually decided against this for a few reasons. First, we still would not get close to 100% completeness on those variables and would need a separate category for "Unknown" and introducing imputations from responses only might even require an additional category. We thought that getting more covariate data from *the types of people* who were replying to our emails would distort the relationship between that covariate and probability of a reply. This could bias the effect of the covariate and in turn bias our estimate of the treatment effect. Finally, since this information is not readily available to someone browsing Craigslist, it is of limited use to them because by the time they receive it, they already have their reply.

After examining our logged data, we noticed four postings for which we had sent emails from both "Katie 27" and "Katie 43", a total of eight observations in our dataset. While a few of these emails garnered replies, in some cases for both "Katies", we removed all eight records from our dataset as we could not assess any possible spillover that may have occurred. Given our randomization scheme, we do not believe this impacted any potential results and removing the observations was the best way to prevent any bias.

5.2 Data Transformations

- Describe any transformations or cleanup steps that would be interesting to a reader

```
##         sent_timestamp
## 1  2019-03-21 20:44:00
## 67 2019-03-27 19:07:00
## 75 2019-03-28 15:57:00
## 76 2019-03-28 16:07:00
## 87 2019-03-29 19:31:00
##                                                                                                  post_url
## 1  https://indianapolis.craigslist.org/roo/d/indianapolis-share-house-near-airport/6847331771.html
## 67                          https://indianapolis.craigslist.org/roo/d/carmel-room/6852070945.html
## 75  https://indianapolis.craigslist.org/roo/d/indianapolis-mastersuite-for-rent-in/6852727902.html
## 76       https://indianapolis.craigslist.org/roo/d/indianapolis-little-flower-near/6852753420.html
## 87    https://indianapolis.craigslist.org/roo/d/indianapolis-looking-for-roommates/6853682073.html
##                                                                post_title
## 1  $600 Share house near airport - month to month (West Washington street)
## 67                                                          Room (Carmel)
## 75                    MasterSuite for rent in Owner share 4 bedroom home.
## 76                                          Little Flower near Irvington
## 87                           $400 Looking for roommates (Indianapolis)\n
##
## 1                                                                Share house, private baseme
## 67
## 75
## 76 Furnished bedroom in 2 Bedroom house for rent. Furniture can be moved out if need be. All utiliti
## 87
##         post_timestamp list_price list_type poster_gender poster_age
## 1  2019-03-21 22:38:00        600     house          <NA>         NA
## 67 2019-03-27 19:50:00        600   unknown          <NA>         NA
## 75 2019-03-28 16:52:00        985     house        female         NA
## 76 2019-03-28 17:23:00        400     house          <NA>         NA
## 87 2019-03-29 20:13:00        400     house          <NA>         NA
##    treatment reply fave_reply     reply_timestamp gender_in_email
## 1   katie_43     Y          1 2019-03-22 08:49:00
## 67  katie_27     N          0                <NA>
## 75  katie_27     N          0                <NA>
## 76  katie_43     N          0                <NA>
## 87  katie_43     Y          1 2019-03-29 22:34:00
##    age_in_email email_sent_dow post_dow reply_dow hours_post_to_email
## 1            NA              5        5         6          -1.9000000
## 67           NA              4        4        NA          -0.7166667
## 75           NA              5        5        NA          -0.9166667
## 76           NA              5        5        NA          -1.2666667
## 87           NA              6        6        NA          -0.7000000
##    hours_email_to_reply sent_is_weekend
## 1              12.08333               0
## 67                   NA               0
## 75                   NA               0
## 76                   NA               0
## 87               3.05000               1
```

- Do some initial data descriptions - Include any plots that make sense and comment on them

We noticed a right-skew tendancy in our continuous variables. This is suggestive that multiplicative differences are more meaningful than additive differences. Given this observation we decided to perform log transformations to facilitate this interpretation. It also may aid our modeling because a lopsided clustering of data points at one end of the range of possible values could reduce the quality of our fit at the other end. This will to some extent even out the distribution of values at either end.

```
## Warning in log(hours_post_to_email): NaNs produced

## Warning in log(hours_post_to_email): NaNs produced

## Warning: Removed 5 rows containing non-finite values (stat_bin).
```

6 Results

6.1 Summary Results

- Two bars, side by side, with probability of favorable replies and error margins - needs code and interpretation **- line chart showing probability of favorable reply based on sent time for treatment and control** - needs code and interpretation

6.2 Regression

To estimate our treatment effects, we employed a variety of regression specifications. All of the regression specifications followed the following equations:

$$Y = \beta_0 + \beta_1(Treatment) + \epsilon$$

For regressions with additional covariates, the specification included additional beta coefficients for each covariate.

We used a linear regression for each of our four regression models, regressing the binary variable of a favorable reply on our treatment and, for the first three models, our chosen covariates. While our outcome variable is binary, using a linear regression allows us to directly measure the expected change in probability of getting a favorable reply, with the coefficients of the independent variables corresponding to the probability increase or decrease.

The first regression contains only the treatment variable.

The second regression contains treatment and the log of hours between post time and the email being sent. We will detail this in the next section, but we wanted to ensure the only statistically significant covariate had no impact on the treatment effect. We chose to take the log of hours between post time and email sent for ease of interpretability.

The third regression contains the treatment variable, the log of hours between post time and the email sent, and the binary indicator of the email being sent on the weekend. This regression specification is meant to help ensure that we randomized effectively and that varibales related to the listing type were not correllated with the treatment.

Finally, the fourth regression regresses favorable reply on the treatment, the log of list price, binary indicators for house and unknown list types (leaving apartments as the default), the log of hours between the post time and email sent, and a binary value indicating if the treament or control email was sent on a weekend. Given that we have only 113 observations, we were wary of including all of our variables for fear of overfitting our regressions and to preserve degrees of freedom. A number of our variables were sparsely populated, so we selected these variables based on their near-complete observations as well as likelihood of explaining variations in a poster's response. We chose to take the log of list price due to a skew in the values.

Examining the results of the four regression specifications, we see a fairly consistent treatment effect ranging from -0.07 to -0.02. This suggests that our randomization worked appropriately, and that the effect of being a 43 year old female as opposed to a 27 year old females was to reduce the probability of receiving a response to roommate listing inquiries by 0.02 to 0.07. In spite of this practical significance, however, this result lacked statistical signifiance for all four model specifications with p-values ranging from 0.46 to 0.81.

In fact, the only variable to show any statistical significance in our model specifications is the log of hours between the post time and the sending of our treatment or control email. The coefficients of -0.07 to -0.07 suggests that a 100% increase in response time corresponds with a -7% to -7% change in likelihood of a response.

6.3 Power Calculations

To better understand if and how our experiment might be lacking in power, we ran a 2-Sample Z-test and Power Tests. The Z-test indicates any differences in the proportions between Katie_27 and Katie_43 receiving a favorable reply. Z-test results:

Estimated probability of success:

Katie_27	0.559
Katie_43	0.537

P-value: 0.961

Confidence Interval at 95%: -0.179 - 0.224

The high p-value of 0.96, combined with a confidence interval wide enough to have both estimates of the probability of success overlap, leads to a conclusion that the proportion of favorable replies between the two Katies' is not significantly different.

Table 3:

	Dependent variable:			
	fave_reply			
	(1)	(2)	(3)	(4)
treatmentkatie_43	−0.022	−0.059	−0.062	−0.070
log(list_price)				0.066
list_typehouse				0.112
list_typeroommate				0.426
list_typeunknown				0.067
log(hours_post_to_email)		−0.069	−0.072	−0.074
sent_is_weekend			−0.045	−0.042
Constant	0.559***	0.788***	0.811***	0.333***
	(0.065)	(0.094)	(0.103)	(0.093)
Observations	113	108	108	108
R^2	0.001	0.069	0.071	0.087
Adjusted R^2	−0.009	0.051	0.044	0.023
Residual Std. Error	0.502 (df = 111)	0.486 (df = 105)	0.488 (df = 104)	0.494 (df = 100)
F Statistic	0.056 (df = 1; 111)	3.884** (df = 2; 105)	2.633* (df = 3; 104)	1.356 (df = 7; 100)
Note:				*p<0.1; **p<0.05; ***p<0.01

We then ran two power tests, one to calculate the number of observations we would need to obtain for *each* treatment, the other to determine the required difference between the outcome variable (favorable replies) in order to observe a statistically significant effect from only 113 observations.

We would need 32434 subjects per treatment condition for our results to be statistically significant. In order to be able to observe a statistically significant effect in age on the outcome variable "favorable replies" using only 113 observations, the treatment effect would need to be 0.37, nearly twenty times our observed effect.

Power increases as sample size or effect size go up, and/or standard deviation is reduced. Because of the requirement of over $32,000$ observations needed for each treatment, in addition to being unsure the treatment effect would be as high as 37% , we conclude that for this particular research project, we do not find any meaningful difference in the rate of favorable replies based on age.

6.4 Randomization Inference

In addition, we also used randomization inference to examine the probability of calculating our treatment effect by chance. Examining the difference in means, we calculate an average treatment effect of -0.0223. Running randomization inference with 1000 iteractions, we find 418 random iterations with at least as significant a treatment effect as our actual. This implies a p-value of 0.418, a non-statistically significant result. This is consistent with our linear regression results showing a non-statistically significant treatment effect.

7 Improvement Opportunities

In retrospect, had we found a concrete effect in this experiment, it would have been easy (and rewarding) to claim that some level of age bias exists. We also speculated on what would have driven an effect had we seen one.

If we saw an effect of 15%, we speculated that some level of age bias exists. While we would have wanted to conclude that the bias was a result of people generally wanting to reside with others of the same gender and age, the lack of data for these two variables would not have allowed us to draw that conclusion. Even having enough data for type of dwelling, rental price, and sent and reply timestamps, we're not convinced these variables play a determinant role in roommate selection.

If an effect of 30% was observed, we would conclude that something went wrong with our experiment. Seeing a 30% effect from only 113 observations is extreme. Note the Power Test indicated a necessary 37% effect to be statistically significant using only 113 observations. In either case, the results would be difficult to accept as valid.

Moving forward, an experiment executed using a similar methodology would require rigorous preparation. Considerations such the population demographics of the area of study, unemployment rates, renting/roommate trends, supply/demand of rental units would all have an impact on the outcome.

In addition, the manual collection of data is tedious at best, and often leaves room for error. Because of this, a more controlled experiment is recommended.

A controlled experiment could leverage an online platform that starts by providing a description of an apartment or house to participants. Participants will be informed that they have placed this ad on a free website and that they are looking for a roommmate. (The roommate would pay their equal share, etc.) Participants would then be asked to review email inquiries from potential renters and indicate how likely they (the participant) would be to respond to the email. Email inquiries would be virtually identical with the exception of age.

This manner of execution would also allow for blocking by participant gender and clustering by participant age. With a large enough, randomly selected participant group, other variables such as renter gender could also be tested.

Although specifics would need to be flushed out in greater detail, this type of experiment would provide a "contained" environment, potentially reduce the amount of confounding variables, and allow researchers to avoid manual data collection.

8 Conclusions

As evidenced by the outputs of our various analyses, we do not find significant evidence that being a 43 year old female garners fewer responses to roommate inquires than being a 27 year old female in Indianapolis. While we found an impact of between-0.07 to -0.02, a practically signficant result, the result lacked statistical signficance in all four of our model specifications, as well as our randomization inference scenario. Thus, we fail to reject our null hypothesis that both ages of females garner the same response rate.

While our experiment found no effect, we believe this result is significant in its own right. We originally set out to study age bias in roommate searches and our experiment provides evidence against such bias. This result could mean that there is one less barrier posed by age in pursuing economic opportunities, and perhaps public policy programs should focus on more pronounced room for bias.

We recommend a variety of future areas of study to expand upon this research. First and foremost, we recommend this study be replicated over a longer time period in a variety of markets, not just Indianapolis. This could help prove the generalizability of our findings. Second, we'd recommend testing a variety of ages. We kept our experiment simple due to time constraints, but examining a continuous span of ages rather than simply two could elucidate more interesting results. Finally, while we limited our experiment to examining age, this experiment could easily be replicated to examine race, gender, sexual orientation, or a variety of other demographic factors. Perhaps some segments of these groups does face significant descrimination when searching for a roommate. Given the feasibility of conducting these experiments on platforms such as Craigslist we would highly recommend this as an area of future research.

9 References

- Fair housing act does not explicitly ban age discrimination in housing (https://www.justice.gov/crt/fair-housing-act-1)
- NYT article on rise of roommates for adults over 30 (https://www.nytimes.com/2016/05/06/fashion/mens-style/adult-men-roommates-new-york.html)
- Atlantic article on rise of roomates (https://www.theatlantic.com/family/archive/2018/08/the-strange-unique-intimacy-of-the-roommate-relationship/567296/)
- Zillow research article on doubling up (https://www.zillow.com/research/rising-rents-more-roommates-17618/)