**Effects of Makeup of Females on Males’ Willingness to Help**



BA830 Experimental Report

Team 6: Fernanda Lin, Xiaoqi Hu, Haolan Ma, Yi Yu, Chenhang Niu

Professor Fradkin

Dec 13th, 2019

1. **Experiment Description**

The purpose of this experiment is to investigate whether makeup on females influences the decision of random stranger male’s willingness to help. Our hypothesis is female with makeup is more likely to receive help from male’s compared to a female without makeup, when everything else stays the same. The background of this experiment is based on Boston. We picked ­­­­­2 big areas on the map of Boston, then divided into 4 small areas and assigned to each individual who will perform this experiment. The four small areas are Kenmore & Fenway, Harvard Square & Kendall Square. Also in order to prevent spill over we used switchback strategy, meanwhile we also set time gaps between each observation in order to prevent contaminations.

1. **Experimental Design**

We exploit an experiment on two regions in Boston, Kenmore/Fenway, and Harvard/Kendal on Nov 23rd, 2019. The experimental condition was assigned at the individual-level involving 64 randomly picked male participants. Since our experiment location is close to each other, we decided to use the switchback testing to prevent sample spill over. In switchback testing, the core concept is that we switch back and forth between control and treatment group in a certain region at alternating time periods. We switch back and forth between the woman with makeup and the woman without makeup from the morning to the afternoon. We then compare the participant’s response between the control time bucket and treatment time bucket during the two time periods in the same location. We randomized the variant used for each time period, rather than simply randomizing the individual variant. We further split across different geographical regions, Kenmore/Fenway, and Harvard/Kendal. These locations are known to have similar characteristics, with mostly college students and working professionals in the area. We wanted to maintain demographics similar across the board. After the selection of locations, we randomize them independently, Kenmore and Harvard could be tested by the woman without makeup in the morning and the woman with makeup the next in the afternoon, while Fenway/Kendal could be tested by the woman with makeup in the morning and the woman without makeup the next in the afternoon. This in a sense actually creates a series of ‘time-region units.’ We not only randomized based on individual level (individual male), but we randomized based on time-region units. Lastly, in order to prevent contamination between each observed male, we conduct each observation with a five to ten mins time gap or walking distance from the previous one in the same small area, just to make sure that participants don’t know if this is an experiment when they make their decision.

*Implementation*

The implementation of the experiment, including the following steps: the female experimenter would randomly approach a male participant on the street during the experiment period and ask for help. The help is defined as borrowing the participant’s smartphone. The words used are consistent across both the control group and the treatment group.

*“Hi, excuse me! My phone is out of charge, could I borrow your phone to make a quick phone call? ”*

We tried to avoid words like “need” and “emergency” that could suggest an urgency, which we thought would have affected a person’s decision. As mentioned, the unit of randomization is stranger male participants walking on the street. The logistic of selecting males who were walking alone is to ensure a non-inference since fellows could potentially affect one’s decision. Additionally, allowing experimenters to pick one man from a group of people contaminate the sample selecting. When it comes to sampling choosing, to achieve the randomization as much as possible, the female experimenters would either walk around 400 meters radius distance decided with a reference or wait for approximately five minutes from the previous location to eliminate the inference that next participants accidentally exposed to the experiment. To control the effects of time and weather, we applied blocking by experimenting with both the morning and afternoon sections to account for the fact that people in the morning are different from the people in the afternoon. Moreover, the outfits are uniform as much as possible including jeans, black coats, and sneakers (Appendix 1). We also avoided wearing luxury items with apparent logos such as Canada Goose to account for the influence of social class.

Between 10 am and 12 pm, one female experimenter would implement the experiment with full makeup. At the same time, one female experimenter would experiment without any makeup in the same area. Between 2 pm and 4 pm, the female experimenter who wore makeup in the morning section would remove the makeup holding the clothes, shoes, hairstyle, accessories, and the text constant and implement the experiment in the same location, and vice versa for the other experimenter. During the experimental period, the participants who were asked by experimenters with full makeup are the treatment group, while the participants who were asked by the experimenters without makeup are considered the control group (Exhibit 1). Each of the female experimenters would start at 10 am and not stop until they record eight responses, and once in the afternoon, beginning at 2 pm, continue the experiments until they got eight responses.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Kenmore** | **Fenway** | **Harvard** | **Kendal** |
| **Morning** | Control  (Experimenter A) | TreatmenT  (Experimenter B) | Control  (Experimenter C) | Treatment  (Experimenter D) |
| **Afternoon** | Treatment  (Experimenter A) | Control  (Experimenter B) | Treatment  (Experimenter C) | Control  (Experimenter D) |

Exhibit 1

*Data Collection*

The intervention denoted as 1 for the treatment group, 0 for the control group. The outcome measured is binary and defined as whether the participants are willing to land the phone, Yes, as 1, No as 0. Record time for each participant and locations are recorded as well. Additionally, we also collect information about the race of participants, including White, Balck, and Asian, according to their skin tone. Another observation is whether participants are in a rush defining by either the participant's expression or by the experimenters' judgment. For example, the participants who explicitly stated that they are on their way to a meeting or doctor appointment would be categorized as rush denoted as 1. If the participants did not expressly declare, the experiment would observe from their walking speed, facial expressions, and speaking rate. Besides the race and rush, age is another thing we record by experimenters' judgment. Additionally, We also tried to record the marriage status by observing the ring but failed to have all samples because of implementing difficulty.

1. **Results**

*Randomization Check*

We performed the randomization check with the variable age (Appendix 3). The result shows that the treatment group is not statistically different from the control group.

*Result Interpretation*

Since it’s hard to observe certain characteristics about locations and times which we want to control, so we decided to treat them as fixed effects. Be more specific in our experiment during the morning on Nov 23rd 2019, there are relatively fewer passengers on a certain street than others, and it’s hard to observe the reason behind it, so it’s better to put time and location as fixed effects.

*The Effect of the Treatment*

The baseline treatment indicate whether the women with makeup. As shown in Appendix 4, Table 1, women with makeup increase the possibility of a participant's willingness to borrow the phone, but the effect is not statistical significant. We learned that the effect of makeup of females on males’ Willingness to help is not significant in general.

*The Effect of the Treatment with Fixed Effect of Record Time and Location*

The Effect of treatment with time as fixed effect. The coefficient estimate shows us when we consider there is no difference on time, women with makeup on their face will more likely receive help from stranger males on the street, however the P-value shows us the result is not statistically significant, since p equals to 0.5938. Also the adjusted R-squared is 0.04545 which indicates that this model can not describe the majority of our data. Therefore, this model fails.（Appendix 4 : Table 2）

When we run a linear regression on the treatment and control for the location, we see that wearing makeup increases the probability of successfully being lent a phone by a male stranger. However, when we look at the significance level of these results, we see that it is low as it does not even have a p-value of 0.05, p-value: 0.3325 and an adjusted R-squared value of 0.01079 (Appendix 4: Table 3). The results can help us understand if there is a difference by the different locations that we have picked, Kenmore & Fenway, Harvard Square & Kendall Square, from this, we can see that there wasn't an effect from the locations.

We can see similar results when we run a regression on the treatment controlling on the recorded time and the location, in that these do not have an effect on the significance level of the results. By controlling on these two variables we do see that there is also a positive effect of wearing makeup on the reaction of the men in this experiment. However, the p-value for these results is still not high enough to reach 0.1%, p-value: 0.2979 and an adjusted R-squared value of 0.248 (Appendix 4: Table 4).

*The Effect of the Treatment with Covariate Variables with Fixed Effect*

Table 5 displays the treatment effect and the standard error together with the effect of the covariate variables, including rush, race, and age. The rush variables indicate whether the participants are in a rush. As shown in Table 5, being in a rush decrease the possibility of a participant's willingness to help, but the effect is not significant. This result is expected since people who are in a rush are less available in general. The interaction effect of treatment and rush increase the outcome and result in a positive conditional treatment effect, but the results are not statistically significant as well. That being said, neither the treatment and rush plays an essential role in affecting one's willingness to help. We also investigate the race as a covariate variable displayed in Table 5. Similar to the rush, race does not show any influence on the outcome, as well. However, although all the results are not statistically significant, we found that the black participant compared to white and Asian participants are more likely to help. Additionally, the white participants are the least willing to help and are negatively related to the willingness to help.

Additionally, Table 5 also displays the treatment effect and the standard error together with the effect of the covariate variables age. The age variables indicate whether the participants are above 40 years old. As shown in Table 5, being in age above 40 years old increase the possibility of a participant's willingness to help, but the effect is not significant. This result is expected since older people deal with more loneliness and love to have company. Older people are more generous to help because they hope their children could get help when needed. The interaction effect of treatment and age increase the outcome, which is not statistically significant. That being said, neither the treatment and age plays an essential role in affecting men's willingness to help. With treatment and fixed effects being considered, treatment alone still has no statistically significant effect on response possibility, but it decreases the possibility of a participant's willingness to help. It’s interesting to find that older men like to help women without makeup.

Last but not least, we run the regression model of response on all variables by adding all variables together and treated race, age, and rush as convirates since we thought that these three variables could all be correlated with other variables as shown in Table 6. For example, with no defense, a young white boy is more likely to help than an old Asian man. As for the result, there are many NAs in the result, and that means this coefficient is not estimable with given information. Moreover, since our dataset contains categorical data and the regression has interaction terms, the NAs also indicate that there are no observations with that combination of levels of the factors. With all variables and fixed effects being considered, treatment alone still has no statistically significant effect on response possibility, which is consistent with our previous estimation results. However, it is interesting to find that age and white race have interaction effects: white people will be less willing to borrow their phones when they are treated with make-up on, and the elder people are more likely to help when they are treated with make-up on.

1. **Discussion**

*Conclusion*

In summary, we found out the treatment of our experiment has no effect on the outcome, since the results are statistically significant. Therefore, we conclude that for female, wearing makeup does not affect the willingness to help of a random male on the street. However, we summarize the following limitations of our experiment for the future improvement.

*Limitation*

* Lack of Information on Interactions

Obversely, we lack much information on interactions. For instance, there is a lack of information on age since age in a private information for people. We only collect the information on age based on one threshold: below 40 or above 40. If we want to have more robust result, we definitely need to collect the exact age on the sample. Moreover, we think that we should come up with more interactions that could potentially affect the treatment outcome.

* Sexual Preference

One of the original ideas for our experiment problem is that we assume and think males would be in favor of females in common case. That is, our design is based on the sexual orientation of heterosexuality, which is man loves women so that man is more willing to help women. However, it seems that we did not take other sexual orientation into account. What if one of the guys in our sample prefer dating boys than girls? Would it cause bias to our experiment? The answer to the second question is yes, it will bring bias to our experiment as well as the result.

* Unethical

In our sampel, some people might be in a rush, but he was so kind and willing to help. However, it turned out that was an experiment. We did meet several situations that people got a little mad after being told it was an experiment.

* Deception and Ignoring

We experienced some situations that people made up some excuses for not helping, such as not having their phone with them, but we saw that they took out their phone right after they walked away from us. Also, several people did not even stop when we reached out to them with saying sorry. These two facts could be in the case of non-compliers in treatment. However, we did not record any information for these non-compliers, so cannot compute the CACE and ITT in our result, which prevent our experiment from accurate.

* Small Sample Size

Overall, due to limited time and short of experimenter, we only 64 samples in total, which is a very small sample size, and it could make our estimations have much lower statistical power. Additionally, the disadvantage of small sample size is that it could affect the reliability of the results of an experiment because it can lead to a higher variability, which may lead to bias.

* Information Efficiency

Besides what we included in our dataset, we also collected some information about people’s characteristics. We took some side notes to describe the person we went through to make our data collection more enriched. For instance, for one Asian guy, we also noticed he is Indian and we recorded this information. However, not everyone did this record, so we cannot included this piece of information in our dataset. Also, we tried to observe some information to distinguish whether or not the person is married by observing if he wears a ring or not, but it turned out that we could not get the observation successfully all the time, and it caused that we only have partial information on marriage. Finally, we had to discard all information about marriage. These two facts show that we have low efficiency on information collection since we cannot make every information we collected in real use, and it is a waste of both time and effort.

**Appendix**

Appendix 1: Outfits

Appendix 2: Experimental Procedure

Step 1: Walk on the street with or without makeup.

Step 2: Randomly pick a stranger male participant on the street.

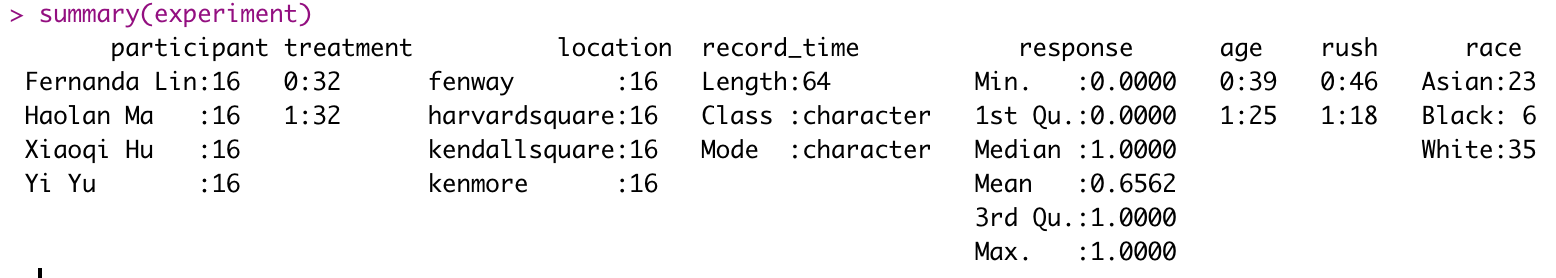
Step 3: Approach him, and asking for borrowing his personal phone, see sentence above

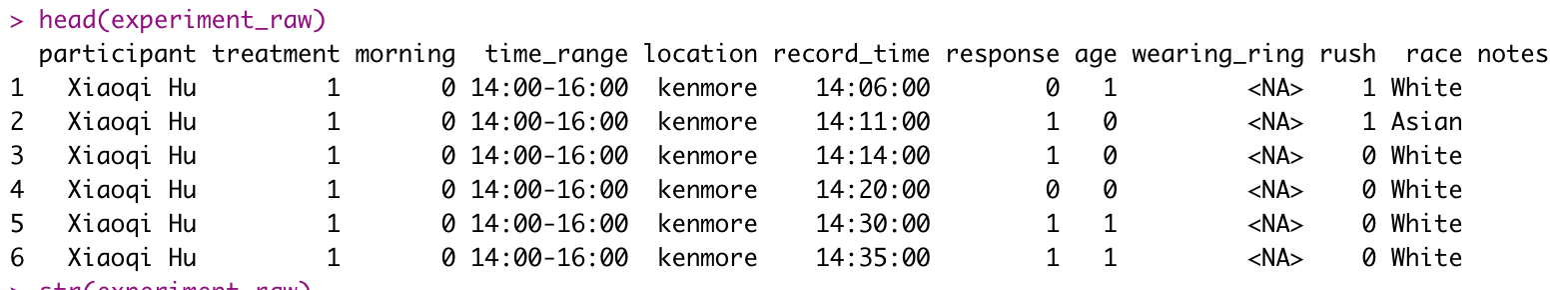
Step 4: Observe his reaction, whether he lends out his phone.

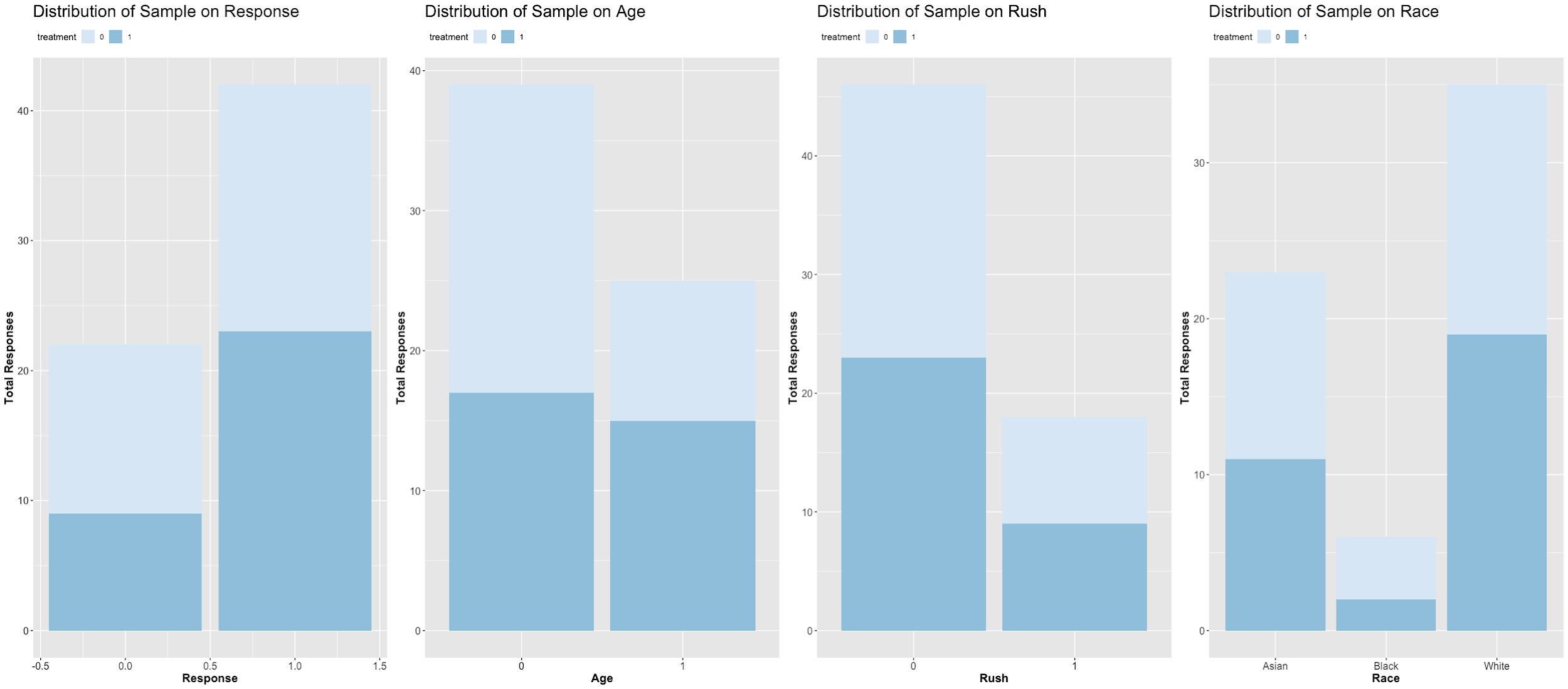
Step 5: Estimate the participant’s age, if he was in a rush or not, observe his race.

Step 6: Repeating steps 1 to 6 again but this time the people who wore makeup removes the makeup and the people who were not wearing makeup runs the experiment with makeup.

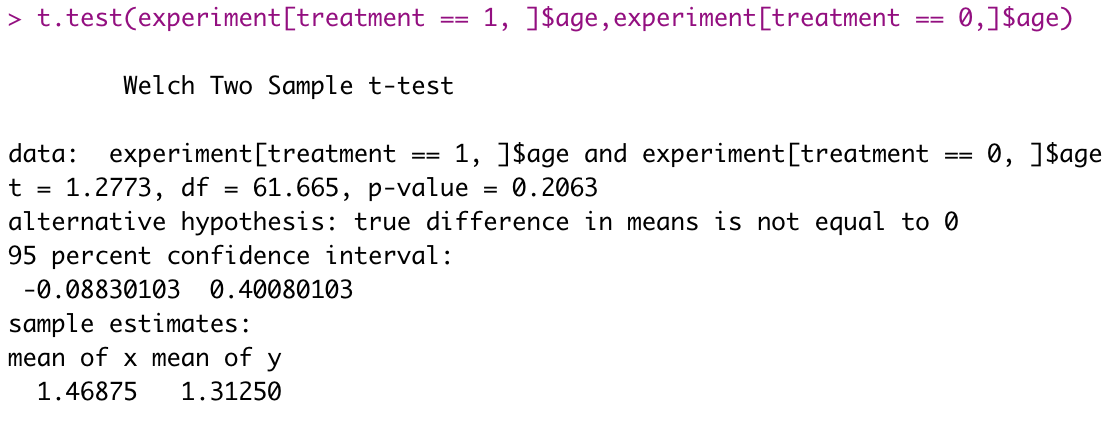
Appendix 3 : Summary Statistic of Data

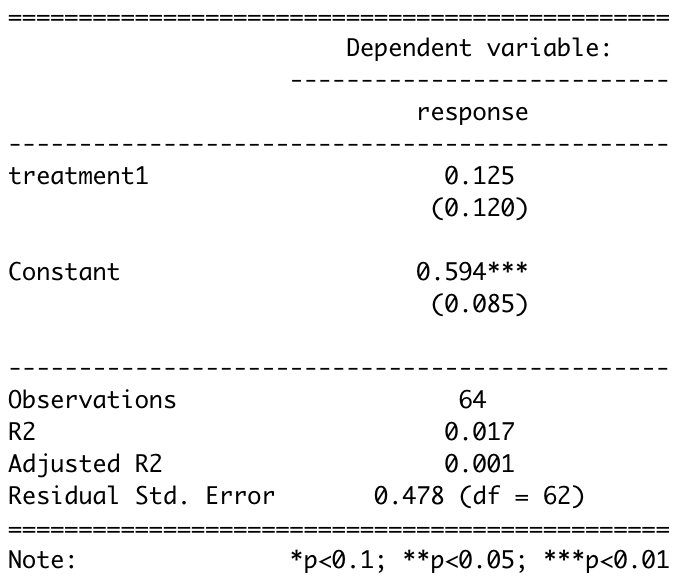




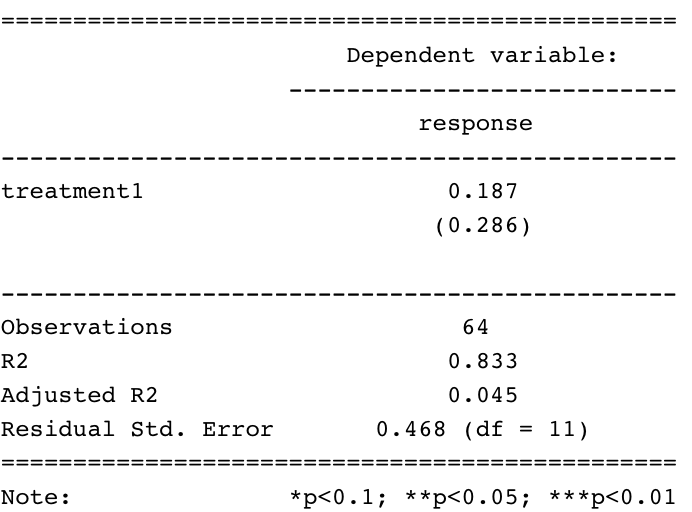


Appendix 4 : Randomization Check

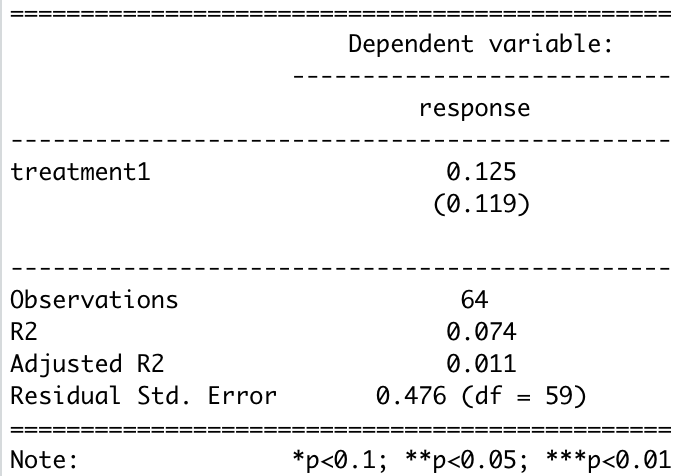
Appendix 4 : Table 1 Treatment on response

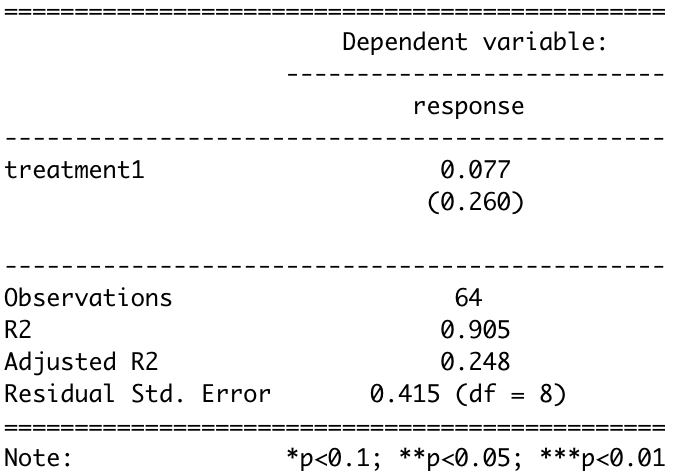


Appendix 4 : Table 2 Treatment on response with time as fixed effect



Appendix 4 : Table 3 Treatment on response with location as fixed effect



Appendix 4 : Table 4 Treatment on response with both time and location as fixed effect

Appendix 4 : Table 5 Treatment on response with both time and location as fixed effect and covariate variables

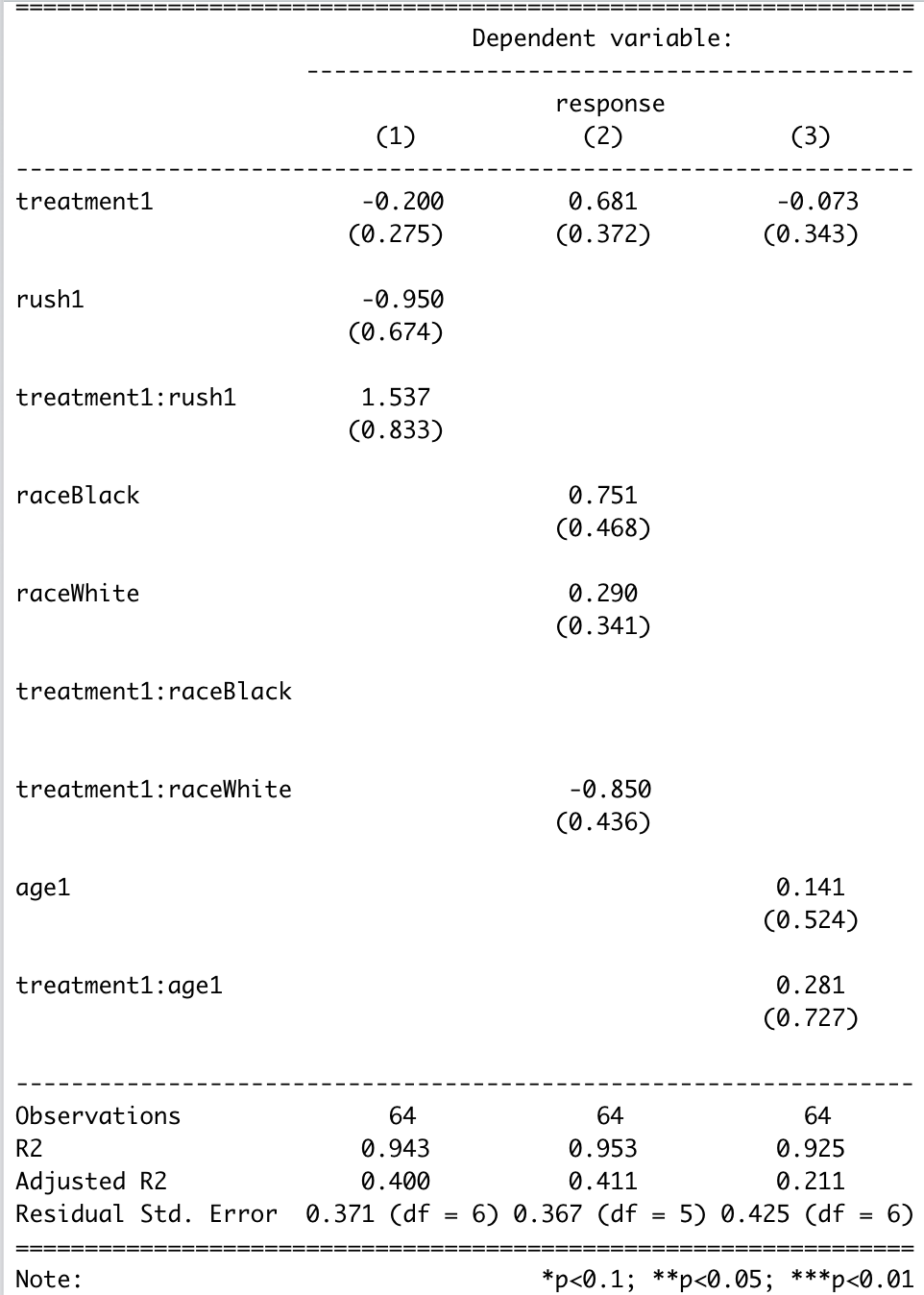
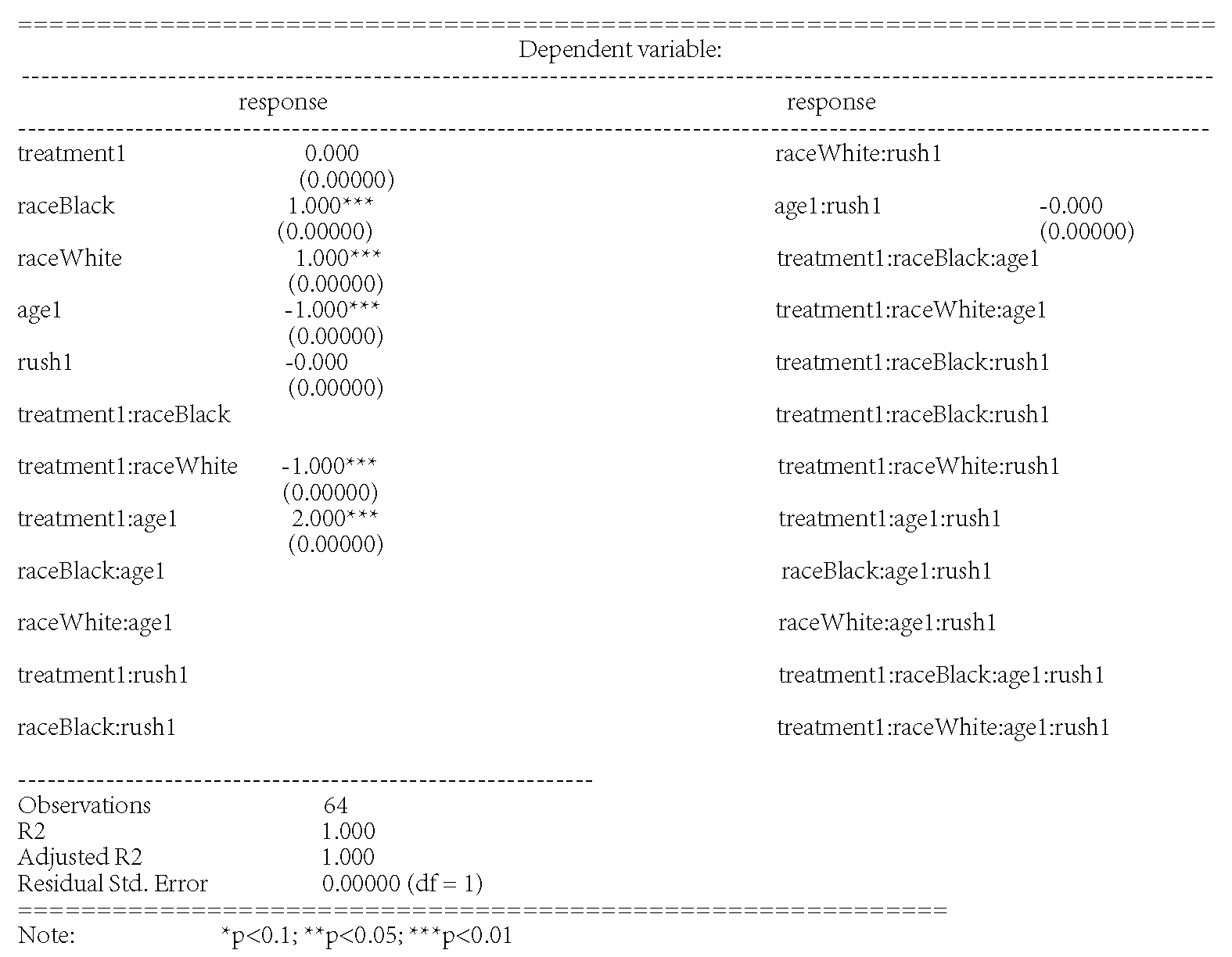


Table 6



Appendix 5: Reference articles that inspired our works

|  |  |  |
| --- | --- | --- |
|  | URL | Brief Description (Key Words) |
| 1 | <https://www.nytimes.com/2011/10/13/fashion/makeup-makes-women-appear-more-competent-study.html> | Makeup effect on the perception of a woman’s competency |
| 2 | <https://www.teenvogue.com/story/how-makeup-affects-depression-anxiety> | How Makeup Can Affect People Dealing With Anxiety and Depression |
| 3 | <https://www.theguardian.com/commentisfree/2015/oct/21/why-do-girls-wear-makeup-google-answer> | Why do girls wear makeup? You asked Google - here’s the answer |
| 4 | <https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00226/full> | Faces with Light Makeup Are Better Recognized than Faces with Heavy Makeup |
| 5 | <https://www.psychologytoday.com/us/blog/meet-catch-and-keep/201502/5-research-backed-reasons-we-wear-makeup> | 5 Research-Backed Reasons We Wear Makeup |
| 6 | <https://www.seniorlivinghelp.co/what-do-old-people-like/> | What Do Old People Like? |