Homework 2: Evaluating Randomized Experiments

Neeraj Sharma

05/02/2020

In 1 and 3, assume randomization worked. For the rest, then you can't assume randomization worked. How to estimate if it's random. Multiple linear regression isn't awesome for prediction but we should use it here for randomness assessment.

Question 1: Estimate the treatment effect and the associated standard error on the raw data given to you, assuming there are no mistakes in or problems with the data.

By definition, the Average Treatment Effect is defined to be $\frac{1}{N}\sum_i y_1(i) - y_0(i)$ where $y_1(i)$ and $y_0(i)$ are the values of the outcome variable (in this case hours exercised) in each treatment scenario. Unfortunately, it is often impractical to utilize this straightforward approach. This formula presumes that one can quantify the outcome of a given individual when treatment is both given and withheld. However, because each individual can only be slotted into one category, this approach cannot be perfectly achieved.

Thus, a random experiment with a control and treatment group can be conducted to smooth out differences among populations in order to isolate the treatment effect specifically. With large enough sample groups, the difference between the mean of the outcome of the treatment group and mean of the outcome of the control group yields the Average Treatment Effect.

Here are the relevant summary statistics of the data set without any modification.

Table 1: Summary Statistics of Hours Exercised across Sample Groups with No Data Cleaning

Treatment	Count	Mean	St Dev	St Err
0	500	20.952	84.02863	3.757875
1	500	27.546	104.36699	4.667434

The mean hours exercised for individuals in the treatment group is 27.546 and the mean hours exercised for individuals in the non-treatment group is 20.952. Assuming randomness was properly implimented in this study, the difference between these two numbers will be the Average Treatment Effect of the treatment based on the analysis I provide above. Thus, the Average Treatment Effect is 6.594 with a standard error of 0.9096.

Question 2: It's always a good idea to check a data set for errors. Clean this data set as you think appropriate. As an answer to this question, note all the kinds of changes you made to the data, a few words explaining your reason for the change, and which observations you changed (noting the observation number included as a variable in the data set for identification purposes). If it is totally obvious which observations you changed and there are a large number in the category (e.g. if you decided to drop all White study participants), you can just note what you did for that change ("I dropped all white participants") and explain why.

I noticed several types of data errors and compiled a representative sample of troubled observations here.

Table 2: Representative sample of unclean observations

subject_id	hours	treatment	community_center	female	age	bmi	education	race_ethnicity
1	2	0	Woodlawn	0	44	28.6634560	high school	BLACK
9	14	0	WOODLAWN	0	19	29.3965780	high school	black
16	180	0	hyde park	1	32	26.9350070	higher degree	BLACK
26	4	0	hyde park	0	-99	33.4841000	higher degree	black
31	13	1	Woodlawn	female	22	22.7878760	higher degree	BLACK
52	1	0	hyde park	male	51	35.4018250	higher degree	white
65	11	1	Woodlawn	0	20	31.7927250	high school	NA
100	7	0	Hyd Park	1	39	34.9566570	higher degree	white
103	8	1	Woodlawn	1	37	0.2565794	high school	BLACK

Given these errors, I perform the following modifying operations to clean the dataset. See appendicies for the specific code I use to accomplish these modifications.

Variable	Description of Modification
subject_id	No change
hours	Hours over 60 are minutes; reformatted to be hours.
treatment	No change.
community_center	No change.
female	Uniformly store data as 0/1, not 0/1/male/female.
age	-99 means missing age data; recoded to be NA.
bmi	BMIs of less than 1 are omitted.
education	No change.
race_ethnicity	Race/Ethnicity BLACK capitalization fixed.
changed?	`changed?` counts changes that occured in cleaning an observation.

It is clear that there are numerous outliers in terms of BMI data. Given that a BMI of less than 18.5 is underweight, having a BMI of ~1 is underweight to the point of impossibility. Thus, I believe these data points were improperly encoded and given the distribution, they appear to have simply misplaced the decimal point two places to the left.

Question 3: With your cleaned data set, re-estimate the treatment effect and estimated standard error, assuming the randomization worked fine.

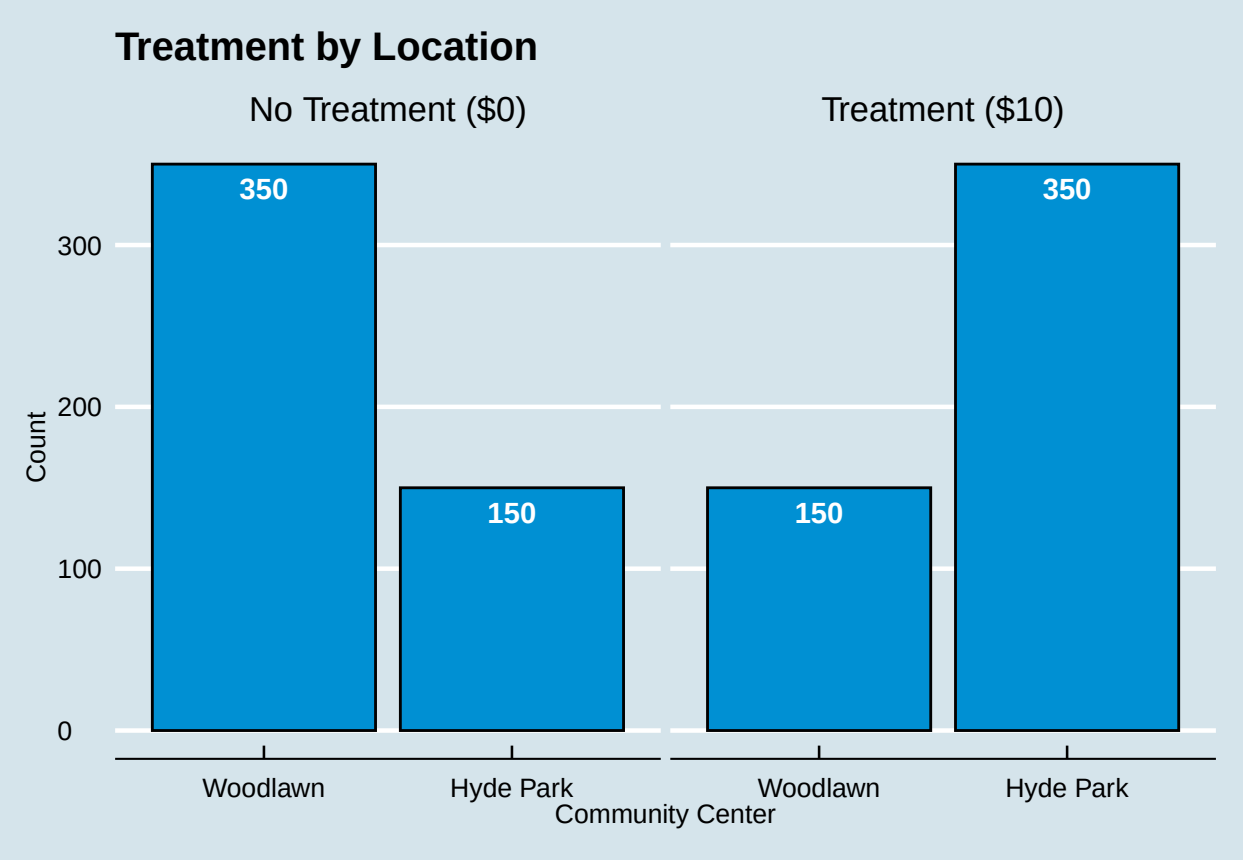
Table 4: Summary Statistics of Hours Exercised across Sample Groups with No Data Cleaning

Treatment	Count	Mean	St Dev	St Err
0	500	5.612	3.387093	0.1514754
1	500	7.486	3.669432	0.1641020

The mean number of hours exercised in both the treatment and control groups has fallen dramatically upon cleaning the data. Numerous entries were coded in minutes instead of hours and those observations were dragging the values up significantly. Those errors have since been corrected. Thus, the Average Treatment Effect upon cleaning the data yet assuming proper randomization is is 1.874 with a standard error of 0.0126.

Question 4: Evaluate whether the randomization appears legitimate. If you don't think the randomization is legitimate, what is your evidence? (Hint: something went wrong.)

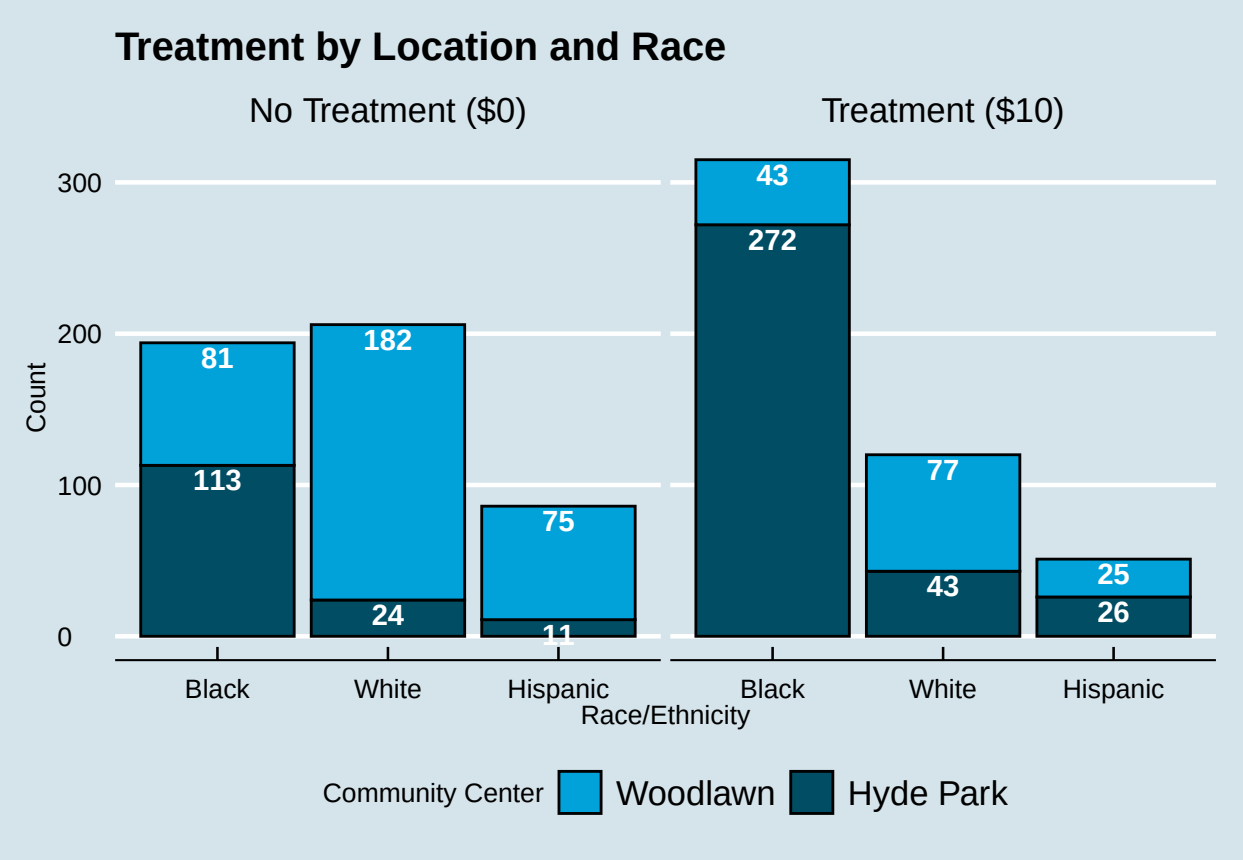
An important piece of insight I gained from Eric's office hours were that given a sufficiently large sample size, perfect random assignment means by definition that one could randomly subdivide the data into equally sized groups of people and the distribution of the covariates should be the same. Thus, if I were to create histograms of some of the specified covariates in this dataset like BMI, age, race, or community, I should observe similar distributions and summary statistics to justify the representativeness and randomness of the allocation of the sample groups. Eric noted that on this problem set specifically is straightforward enough with regards to the identification of systematic differences that imply a non-random allocation that full t-test would not be necessary.



This plot clearly shows a bias in the group allocations based on survey site. Were this data to be randomly created, we would expect to see an even distribution of individuals in the treatment and non-treatment groups across Woodlawn and Hyde Park. Instead, we find a disproportionate amount of the control group was assigned at Hyde Park and conversely, a disproportionate number of people in the treatment group were assigned in Woodlawn. A random experiment would not exhibit this property.

Further analysis on other covariates supports the conclusion that race is not controlled between treatment groups.

```
## Warning: Factor `race_ethnicity` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```



In general, plots of most other covariates look fairly decently normal. For brevity, I only provide summary statistics to articulate this point:

```
## # A tibble: 2 x 7
##   treatment `Mean Female` `Sd Dev Female` `Mean Age` `Sd Dev Age`
##   <fct>             <dbl>           <dbl>      <dbl>        <dbl>
## 1 No Treat~         0.636           0.482       32.9        10.7
## 2 Treatmen~         0.582           0.494       31.7         9.69
## # ... with 2 more variables: `Mean BMI` <dbl>, `Sd Dev BMI` <dbl>
```

```
## # A tibble: 2 x 4
## # Groups:   treatment [2]
##   treatment         `less than high school` `high school` `higher degree`
##   <fct>                               <int>         <int>           <int>
## 1 No Treatment ($0)                      54           208             238
## 2 Treatment ($10)                        45           253             202
```

```
## Warning: Factor `race_ethnicity` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
## # A tibble: 2 x 5
## # Groups:   treatment [2]
##   treatment         Black Hispanic White  `NA`
##   <fct>             <int>    <int> <int> <int>
## 1 No Treatment ($0)   194       86   206    14
## 2 Treatment ($10)     315       51   120    14
```

How to tell if you aren't random. Clues that it is the case is done through a balance table. Compare means of each observable metric between treatment and control groups. You do a t-test to compare them. THIS IS

NOT GREAT, but still do it to confirm you are balanced on your observables.

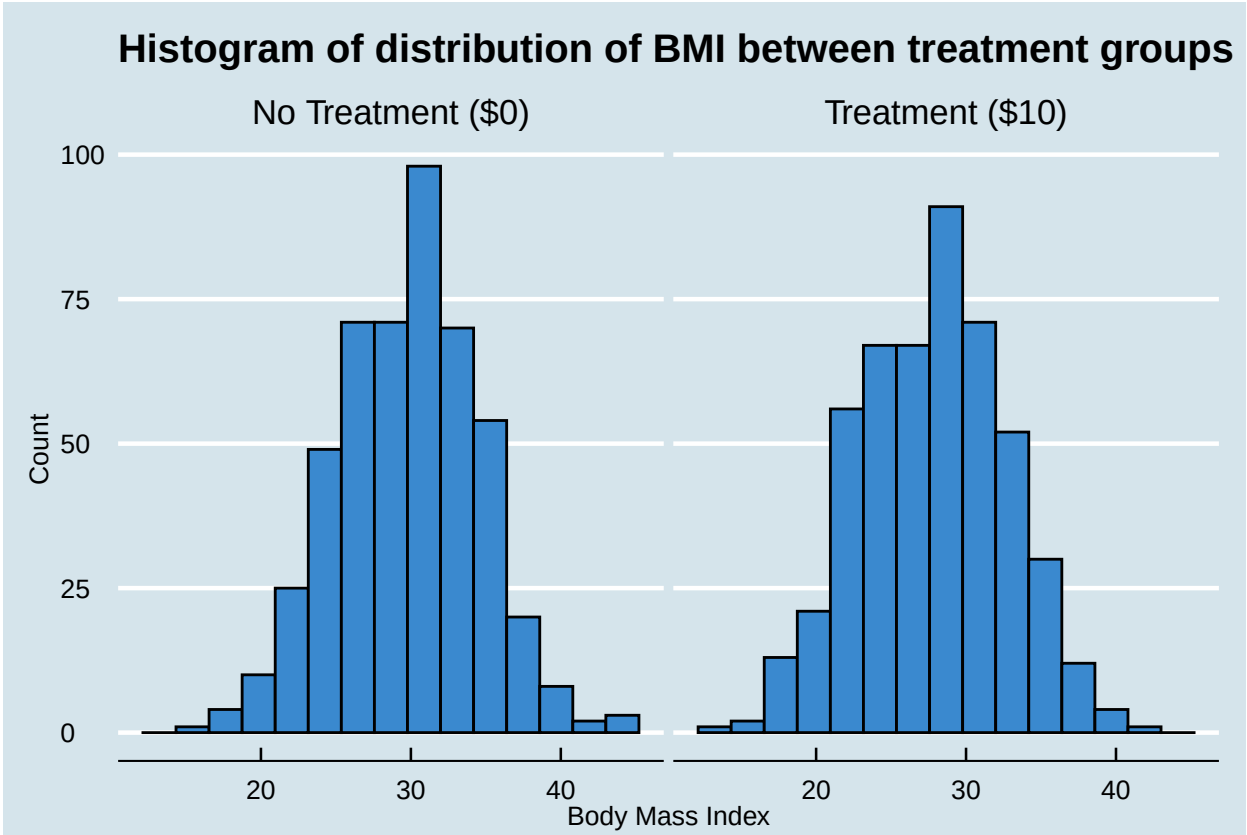
```
## Warning in mean.default(community_center): argument is not numeric or
## logical: returning NA

## Warning in mean.default(community_center): argument is not numeric or
## logical: returning NA

## Warning in mean.default(education): argument is not numeric or logical:
## returning NA

## Warning in mean.default(education): argument is not numeric or logical:
## returning NA

## Warning in mean.default(race_ethnicity): argument is not numeric or
## logical: returning NA

## Warning in mean.default(race_ethnicity): argument is not numeric or
## logical: returning NA
```

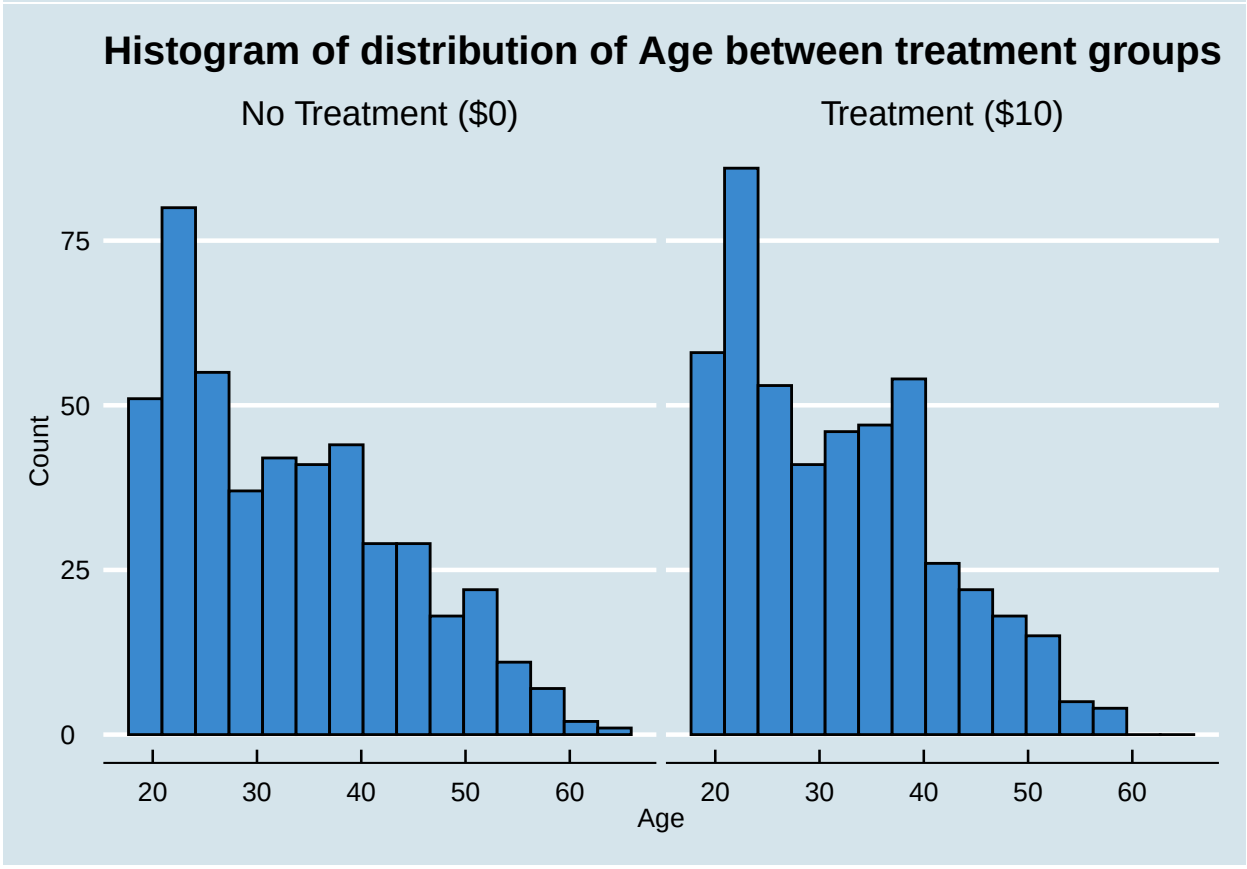
treatment	subject_id	hours	community_center	female	age	bmi	education	race_ethnicity	changed?
0	517.356	5.612	NA	0.636	NA	NA	NA	NA	0.536
1	483.644	7.486	NA	0.582	NA	NA	NA	NA	0.580

```
## Warning: `cols` is now required.
## Please use `cols = c(Black, Hispanic, White, `NA`)`
```

Table 6: Distribution of Race/Ethnicities Sampled in Each Group

treatment	Black	Hispanic	White	NA
0	194	86	206	14
1	315	51	120	14


Histogram of distribution of BMI between treatment groups
No Treatment ($0)
Treatment ($10)
Count
100
75
50
25
0
20
30
40
Body Mass Index


Histogram of distribution of Age between treatment groups
No Treatment ($0)
Treatment ($10)
Count
75
50
25
0
20
30
40
50
60
Age

Question 5: Offer your best hypothesis/hypotheses as to what went wrong with the randomization? What evidence do you have to support your hypothesis(es)? For each of these hypotheses, describe your best strategy for estimating a plausible treatment effect, in spite of the bad randomization. (But don't actually estimate that treatment effect.)

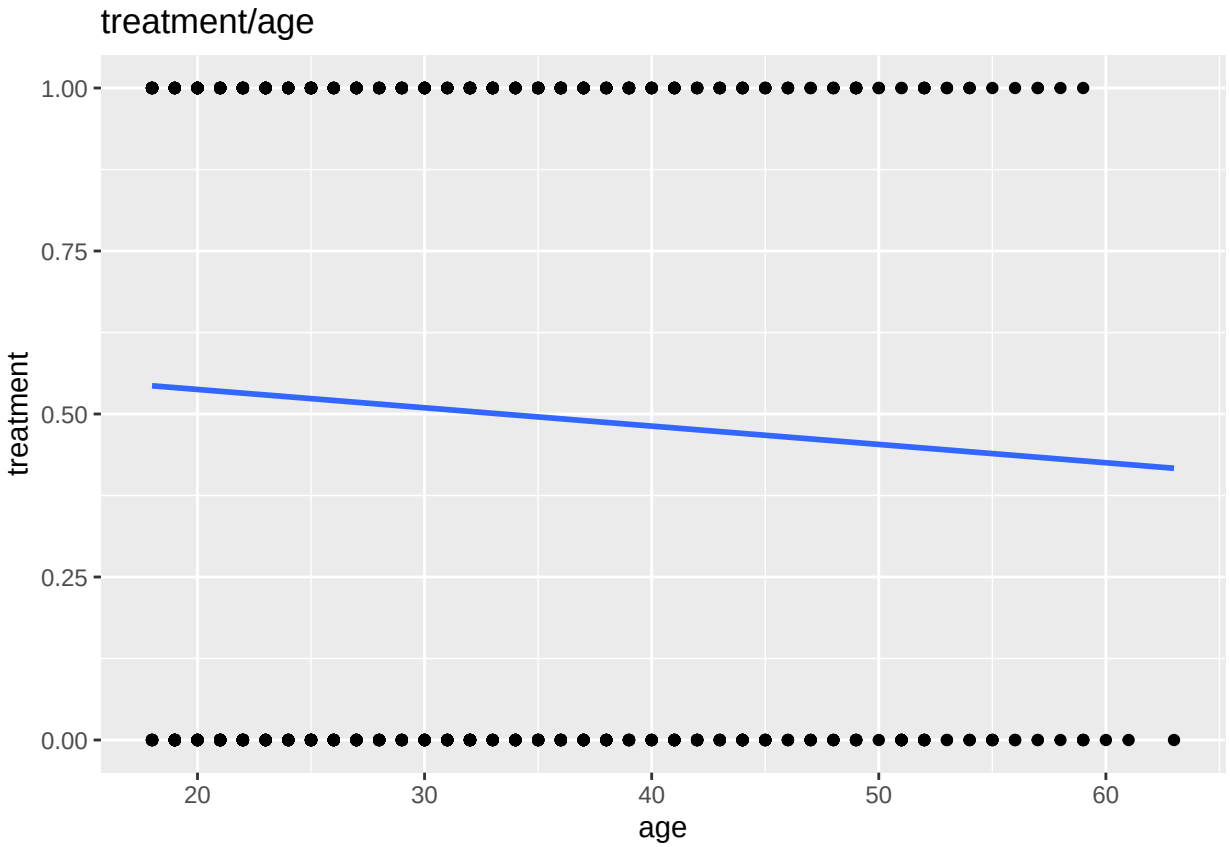
Multiple Linear Regression Example

fit <- lm(y ~ x1 + x2 + x3, data=mydata) summary(fit) # show results anova function can be used to compare different linear models

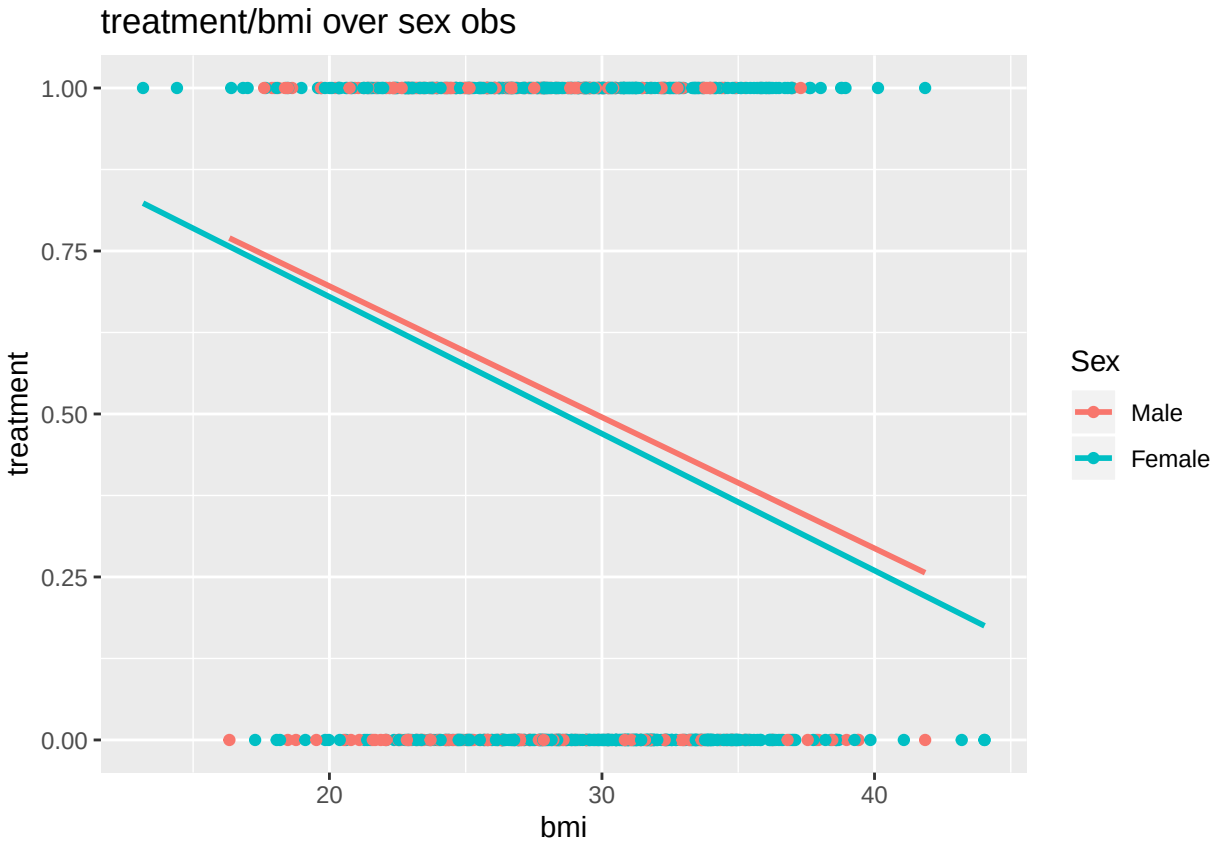
```
##
## Call:
## glm(formula = treatment ~ bmi, family = binomial, data = clean_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6614  -1.1313   0.6906   1.1426   1.6819
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.51458    0.40548   6.202 5.59e-10 ***
## bmi         -0.08721    0.01390  -6.275 3.50e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1350.2  on 973  degrees of freedom
## Residual deviance: 1308.5  on 972  degrees of freedom
##   (26 observations deleted due to missingness)
## AIC: 1312.5
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] 0.4281314
```

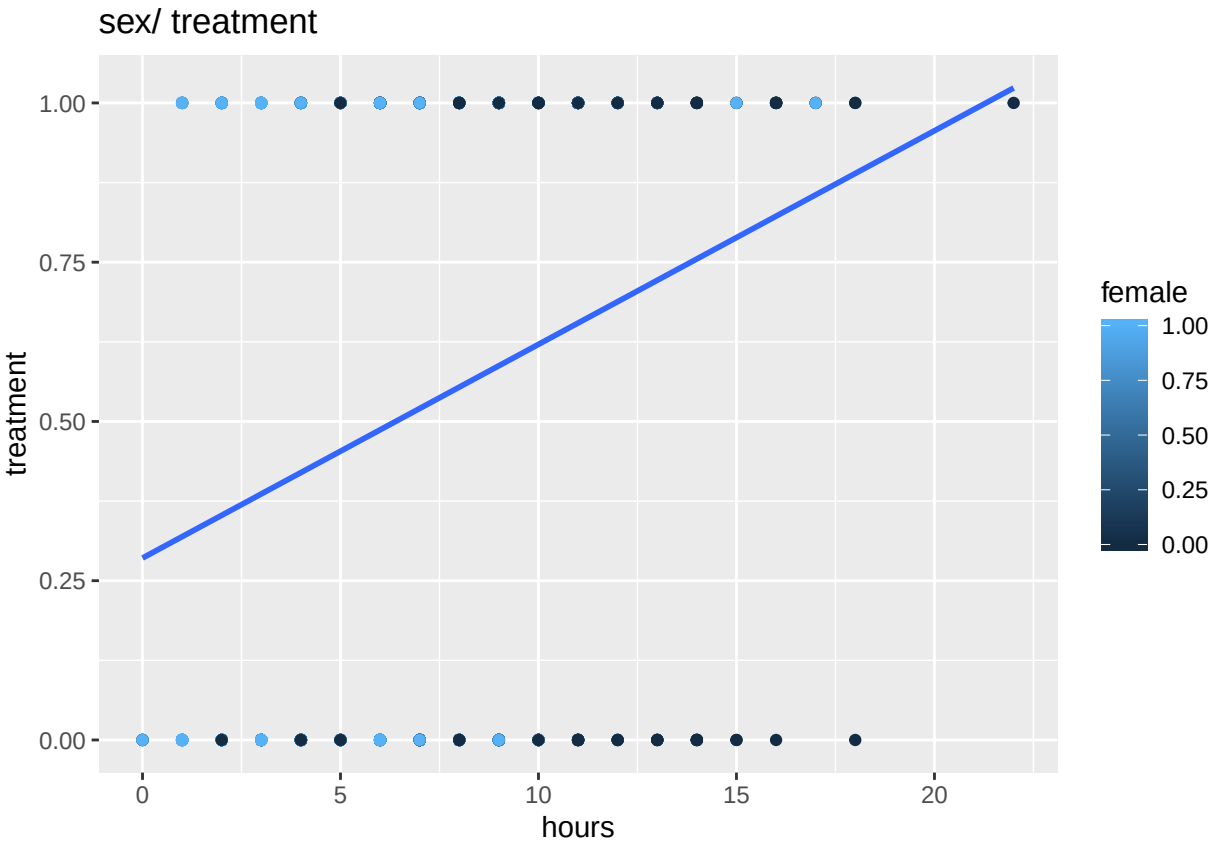
```
## Warning: In lm.wfit(x, y, w, offset = offset, singular.ok = singular.ok,
##     ...) :
##  extra argument 'family' will be disregarded
```



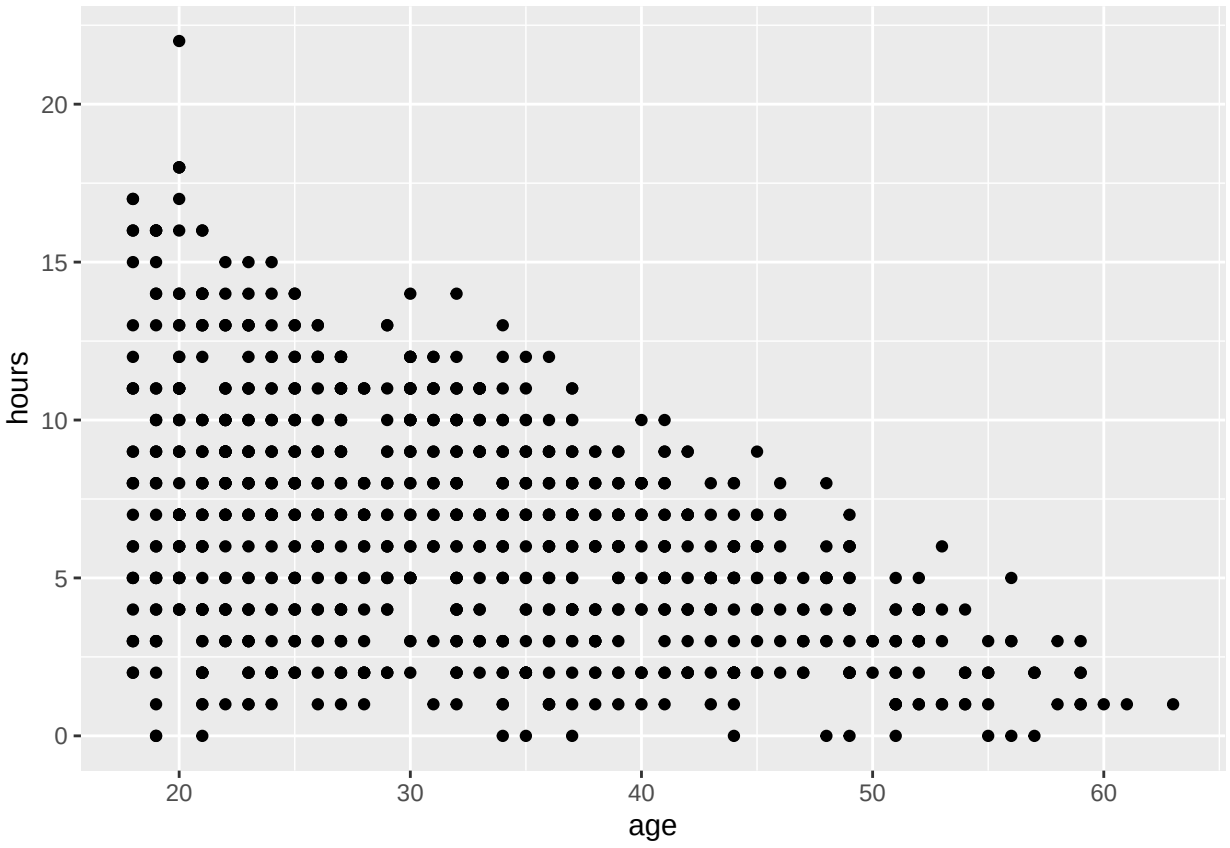

```
## Warning: In lm.wfit(x, y, w, offset = offset, singular.ok = singular.ok,
##     ...) :
##  extra argument 'family' will be disregarded

## Warning: In lm.wfit(x, y, w, offset = offset, singular.ok = singular.ok,
##     ...) :
##  extra argument 'family' will be disregarded
```




```
## Warning: In lm.wfit(x, y, w, offset = offset, singular.ok = singular.ok,
##     ...) :
##  extra argument 'family' will be disregarded
```




```
## Warning: Removed 56 rows containing missing values (geom_point).
```

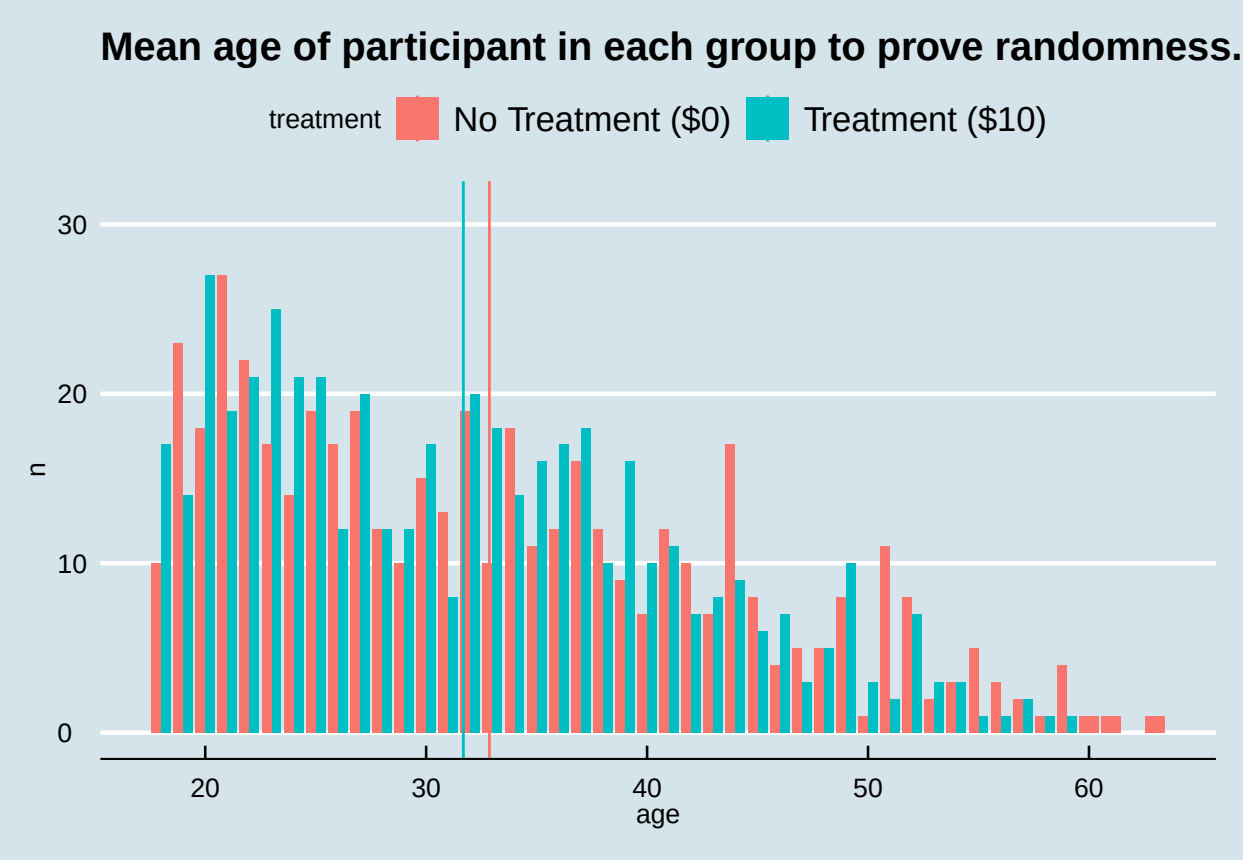


Appendicies

Code to clean data in question 2

```
## # A tibble: 2 x 2
##   treatment         `mean(age, na.rm = TRUE)`
##   <fct>                                 <dbl>
## 1 No Treatment ($0)                      32.9
## 2 Treatment ($10)                        31.7
```

```
## Warning: Removed 2 rows containing missing values (geom_col).
```


Mean age of participant in each group to prove randomness.
treatment
No Treatment ($0)
Treatment ($10)
30
20
10
0
n
20
30
40
50
60
age