

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 implemented 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.

Variable	Description of Modification
subject_id	No change

Variable	Description of Modification
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 occurred in cleaning an observation.

```

# Clean the data to impose uniformity upon the variable encoding.
clean_data <- raw %>%
  mutate(`changed?` = seq(1, 1000) * 0) %>%
  # Fixing messed up hours readings. They start at 60 and upwards and that's 1 hr so I
  # fix based on that.
  mutate(`changed?` = if_else(hours >= 60, `changed?` + 1, `changed?`),
    hours = if_else(hours >= 60, hours / 60, hours)) %>%
  mutate(`changed?` = if_else(community_center %in% c("WOODLAWN",
    "hyde park",
    "woodlawn",
    "Hyd Park",
    "HYDE PARK",
    "Hyde_Park",
    "HYDE_PARK",
    "hyde_park"),
    `changed?` + 1, `changed?`),
    community_center = if_else(community_center %in% c("WOODLAWN", "woodlawn"),
      "Woodlawn", community_center),
    community_center = if_else(community_center %in% c("hyde park",
      "Hyd Park",
      "HYDE PARK",
      "HYDE_PARK",
      "hyde_park",
      "Hyde_Park"),
      "Hyde Park", community_center)) %>%
  # Recoding female variable to be factor categorical variable from 0/1/male/female.
  mutate(`changed?` = if_else(female %in% c("female", "male"),
    `changed?` + 1, `changed?`),
    female = as.double(if_else(female == "female",
      "1", if_else(female == "male", "0", female)))) %>%
  # -99 is missing age data so I reencode it at missing age data.
  # https://cran.r-project.org/web/packages/naniar/vignettes/replace-with-na.html
  mutate(`changed?` = if_else(age == -99, `changed?` + 1, `changed?`),
    age = na_if(age, -99)) %>%
  # Currently I have removed improper values, but I could also justify multiplying by 10.
  mutate(`changed?` = if_else(bmi < 1, `changed?` + 1, `changed?`),
    bmi = ifelse(bmi < 1, NA, bmi)) %>%
  # Fix all BLACK observations to normal capitalization structure.
  mutate(`changed?` = if_else(is.na(race_ethnicity),
    `changed?`, if_else(race_ethnicity == "BLACK",
      `changed?` + 1, `changed?`)),

```

```

race_ethnicity = if_else(is.na(race_ethnicity),
                        race_ethnicity, str_to_sentence(race_ethnicity))) # %>%
# mutate(race_ethnicity = factor(race_ethnicity, c("Black", "Hispanic", "White"))) %>%
# mutate(treatment = factor(treatment, labels = c("No Treatment ($0)", "Treatment ($10)"))) %>%
# mutate(community_center = factor(community_center, labels = c("Woodlawn", "Hyde Park"))) %>%
# mutate(education = factor(education, levels = c("less than high school",
#                                                "high school",
#                                                "higher degree")))
#
#

```

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

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

If there was perfect random assignment, you could randomly pluck 100 people, then pluck 100 more, their distributions should be the same of age BMI Community etc. Systematic differences imply that it's not random. Talk about mean median and stuff but you don't need to do a t-test.

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 4: Distribution of Race/Ethnicities Sampled in Each Group

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

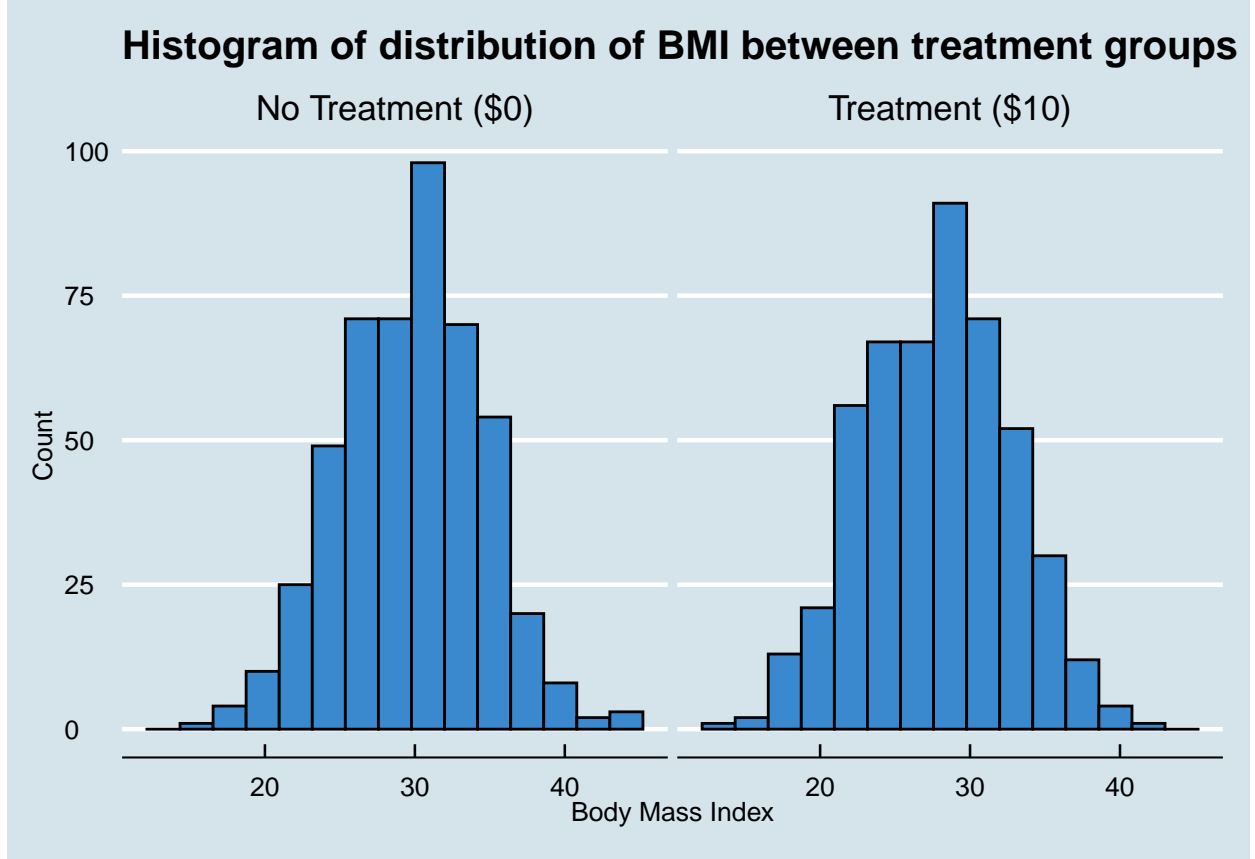
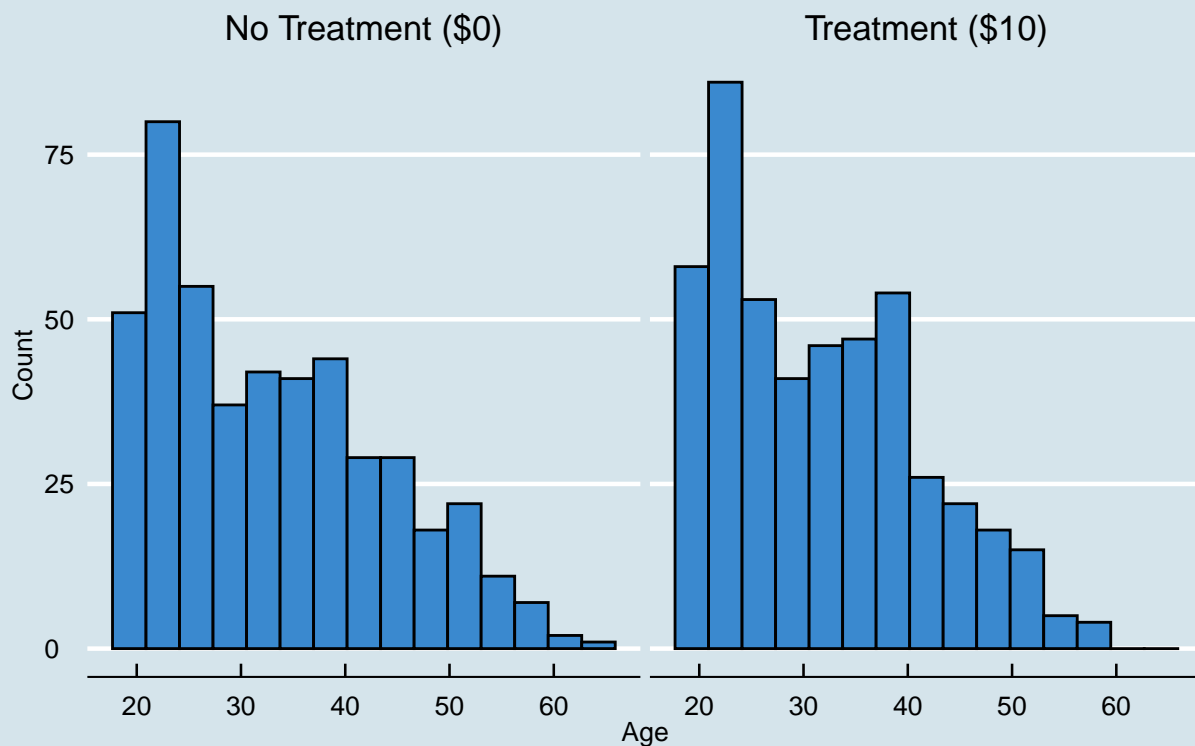


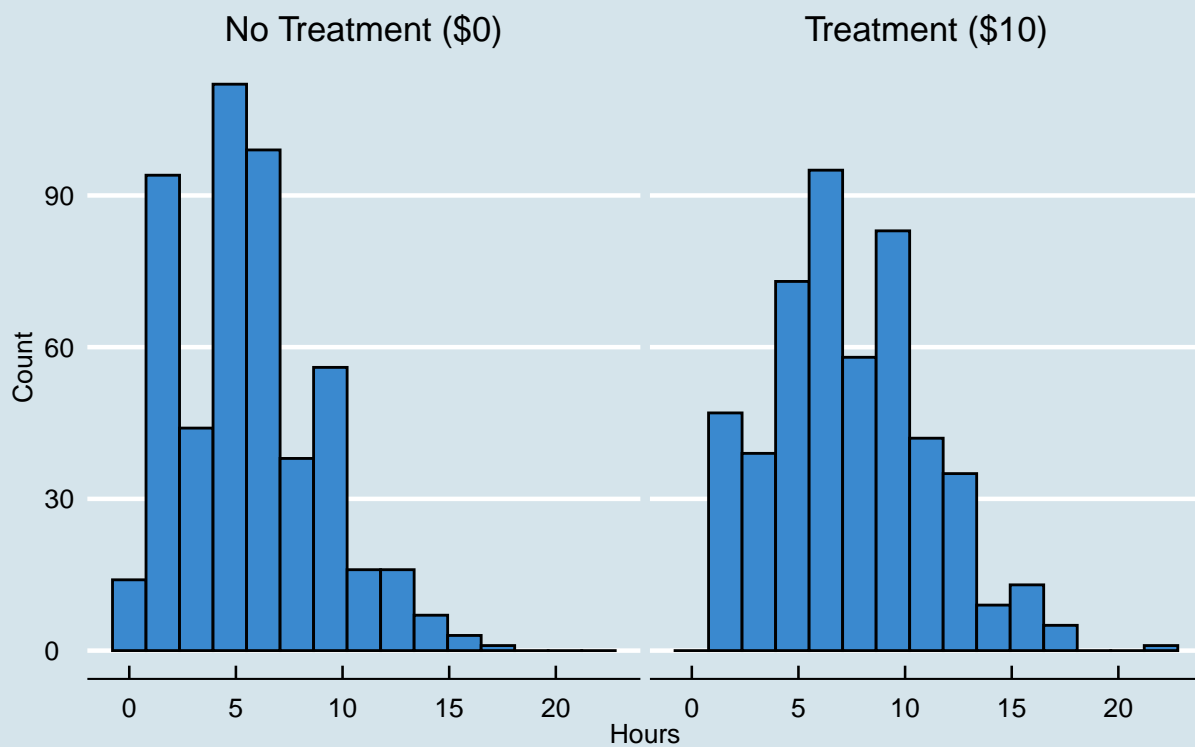
Table 5: Summary Statistics of BMI across Sample Groups

Treatment	Count	Mean	St Dev	St Err
No Treatment (\$0)	486	29.78517	4.728381	0.2144837
Treatment (\$10)	488	27.78196	4.848491	0.2194809

Histogram of distribution of Age between treatment groups



Histogram of distribution of Hours between treatment groups

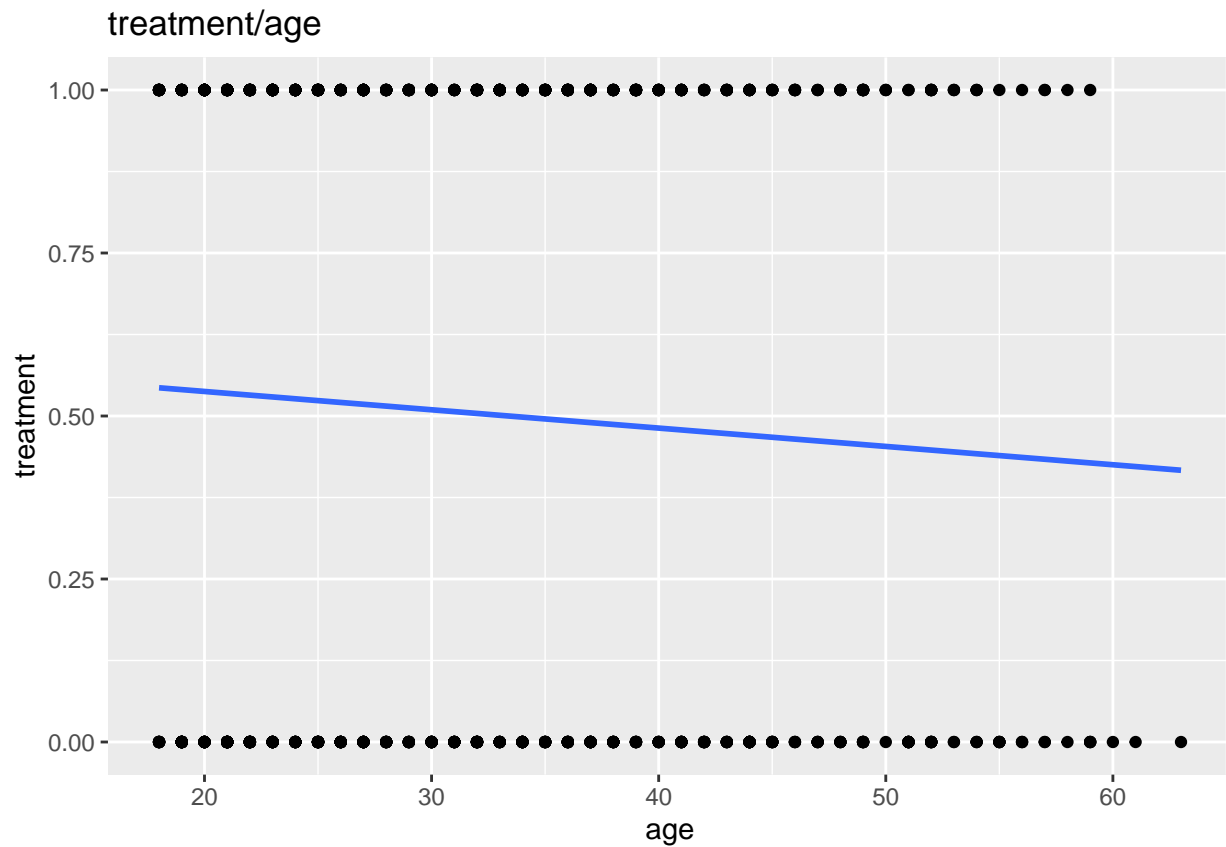


Question 5: Offer your best hypothesis/hypotheses as to what went wrong with the randomization? What evidence do you have to support your hypothesis(es)? For each of these hypotheses, describe your best strategy for estimating a plausible treatment effect, in spite of the bad randomization. (But don't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

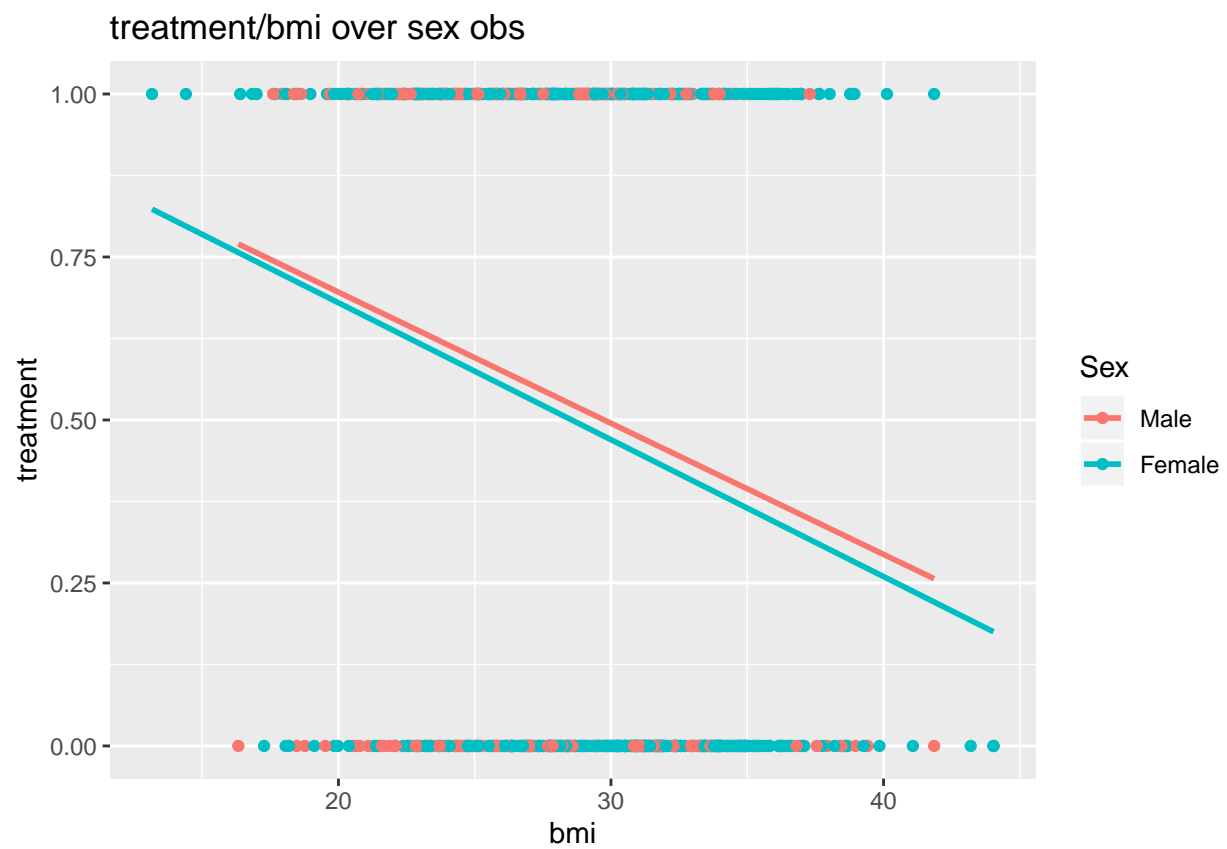
Multiple Linear Regression Example

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

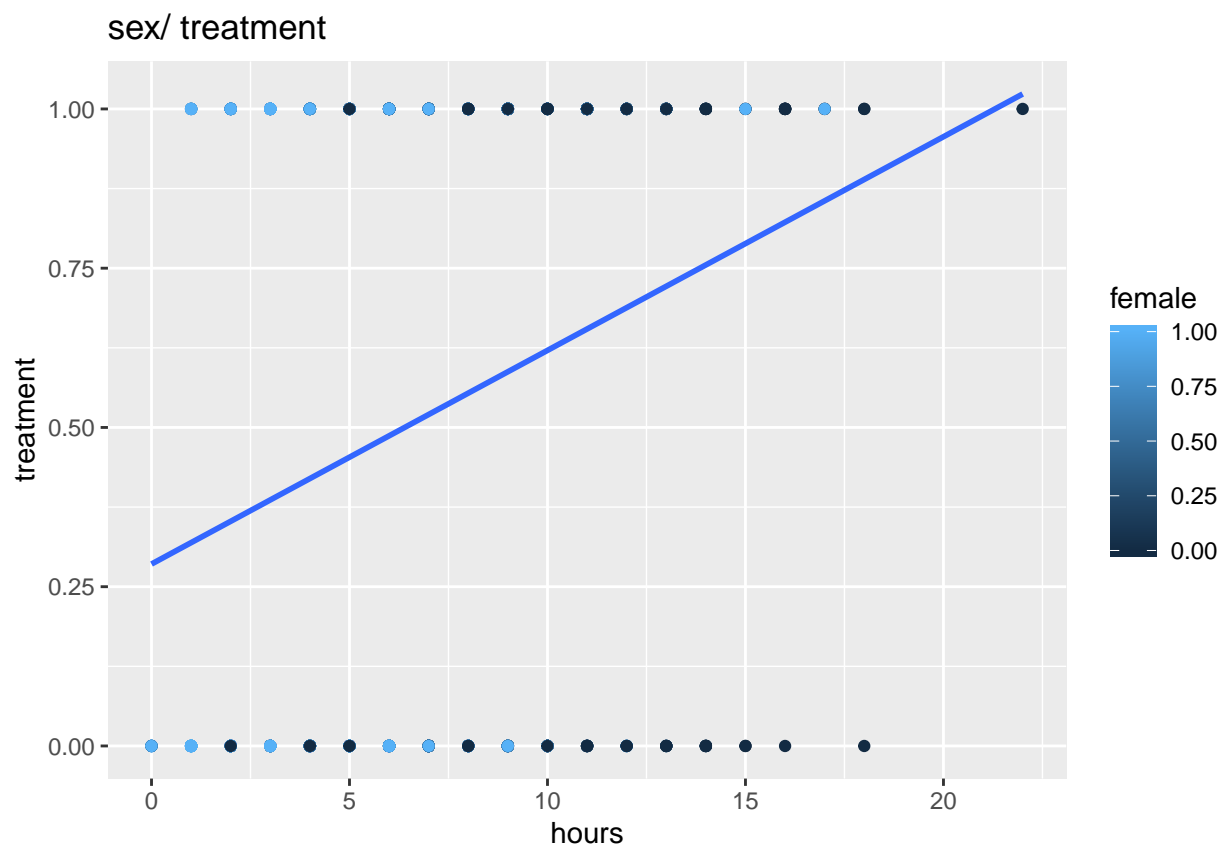
```
##
## Call:
## glm(formula = treatment ~ bmi, family = binomial, data = clean_data)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6614  -1.1313   0.6906   1.1426   1.6819
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.51458    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).

