# Test 04: Numeric Y vs Numeric X

Richard

1/30/2022

# Load required packages

```
library(tidyverse) # if you're using macOS, you can run: l
```

```
## -- Attaching packages ---------------------------------
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
library(skimr)
library(ggplot2)
```

# Prepare Data

Please read the intro about data at here

```
Hsb <- within(
  read.csv("https://stats.idre.ucla.edu/stat/data/hsb2.csv"
    race <- as.factor(race)
    schtyp <- as.factor(schtyp)
    prog <- as.factor(prog)
})
```

# Numeric variables

- In general, we deal with numeric variables all the time
  - e.g., temperature, rain volume, salary, ...
- It is rich value and contains more information than categorical variables
- Sometimes, we want to find the **relation** between two numeric variables e.g., "temperature and our electricity bills", "how far you live to the downtown and your income", ...
- A **relation** does not mean a **causality**

# Relation/Correlation vs Causality

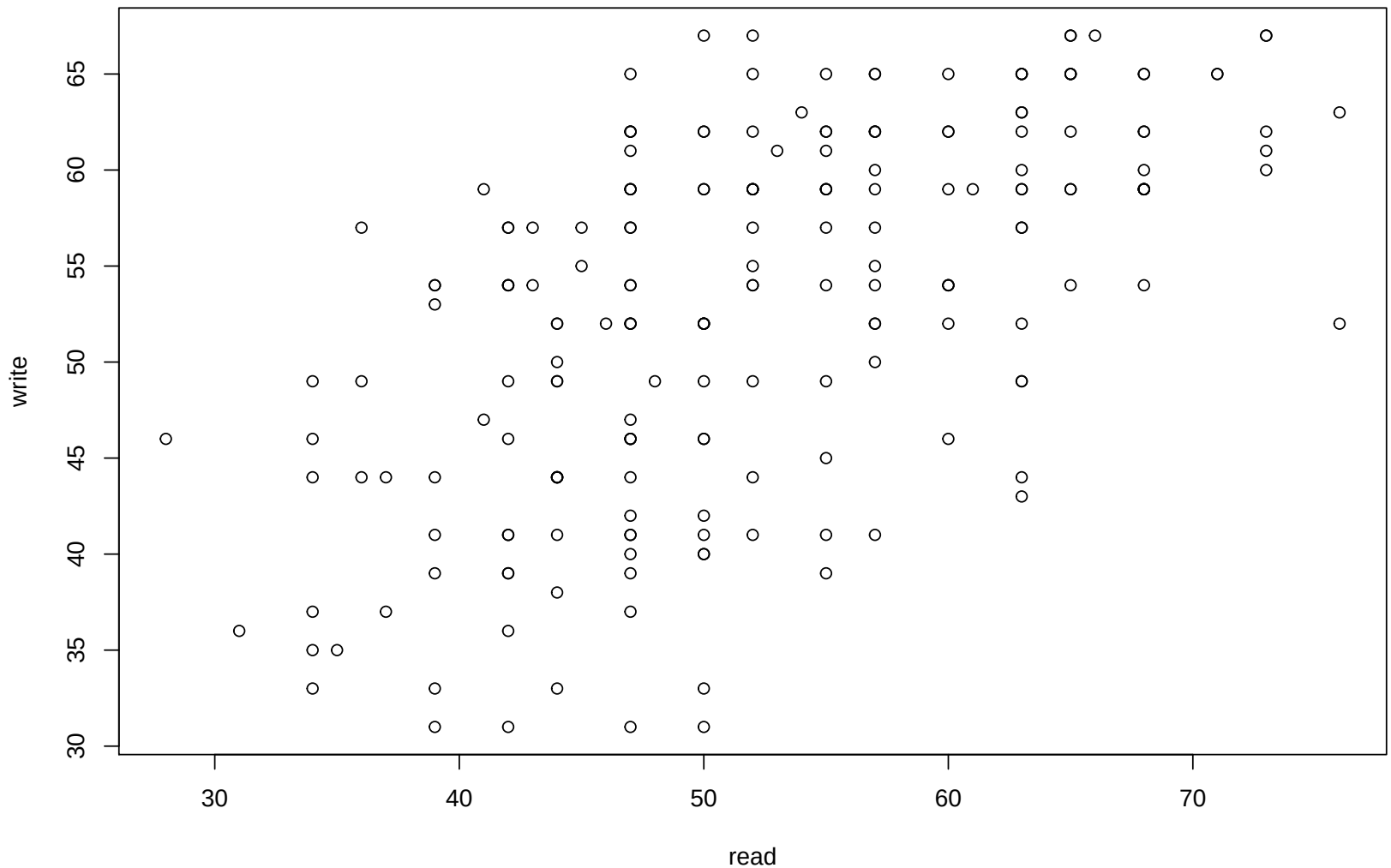
- Relation/correlation shows that if X increases, Y will increase or decrease, depending on the relation is positive or negative correlated
- Causality means that because of X, so we have Y
  - we know which variable happens first, then we have the outcome
- For example, return to our proposed relation: *"how far you live to the downtown and your income"*
  - it is difficult to know which variable causes which variable
  - e.g., maybe you're rich so you live in downtown; or because you're living in downtown so you find a better job; or because you was born in a high-class family so you are not only live in downtown but also have a good-paid job
- In this class, we focus on correlation, not causality, which requires a more sophisticated class in the future (Econometrics class)

# Research question

- If `write` and `read` scores are correlated?

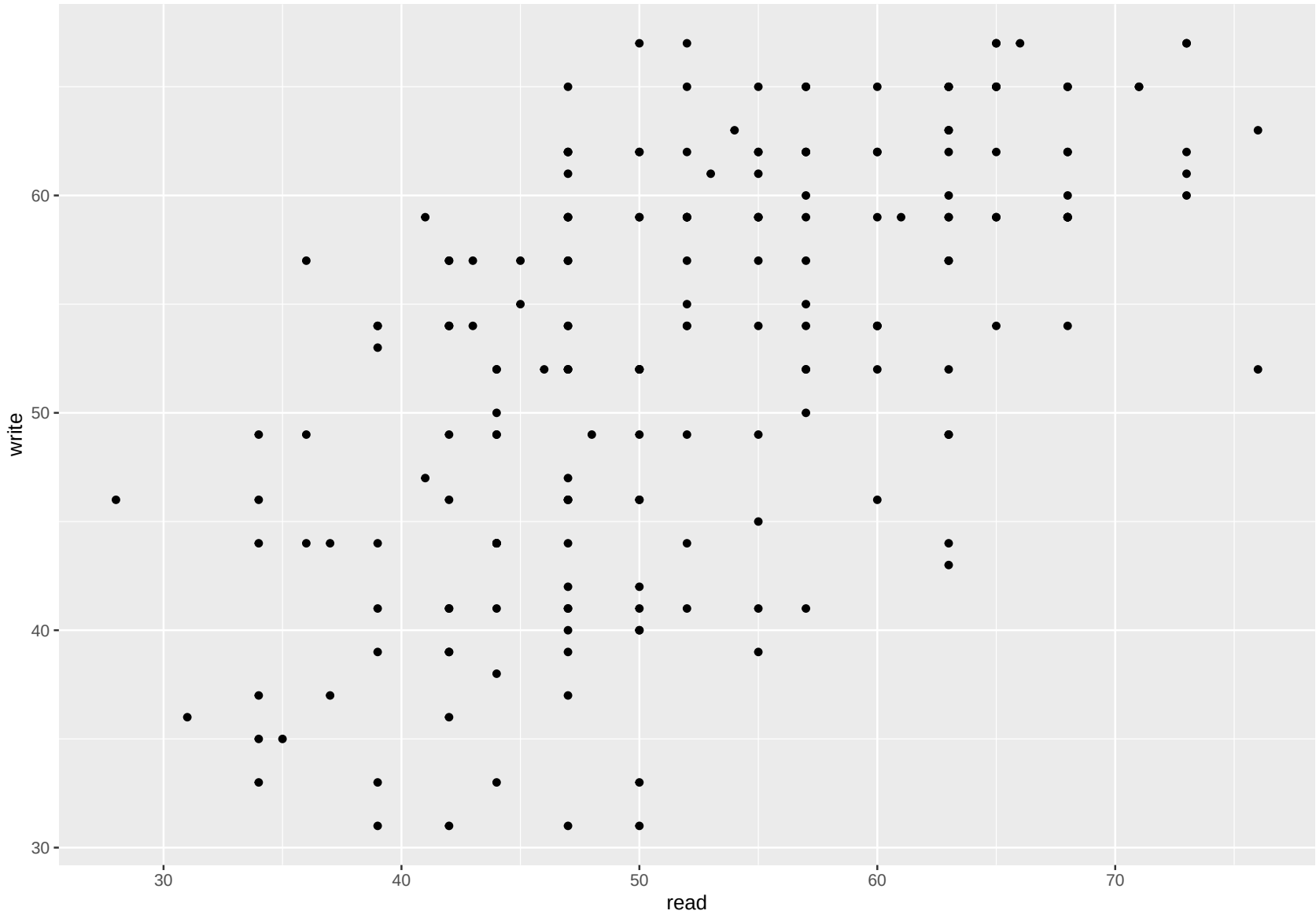
# Scatter plot

```
attach(Hsb)
plot(x = read, y = write)
```



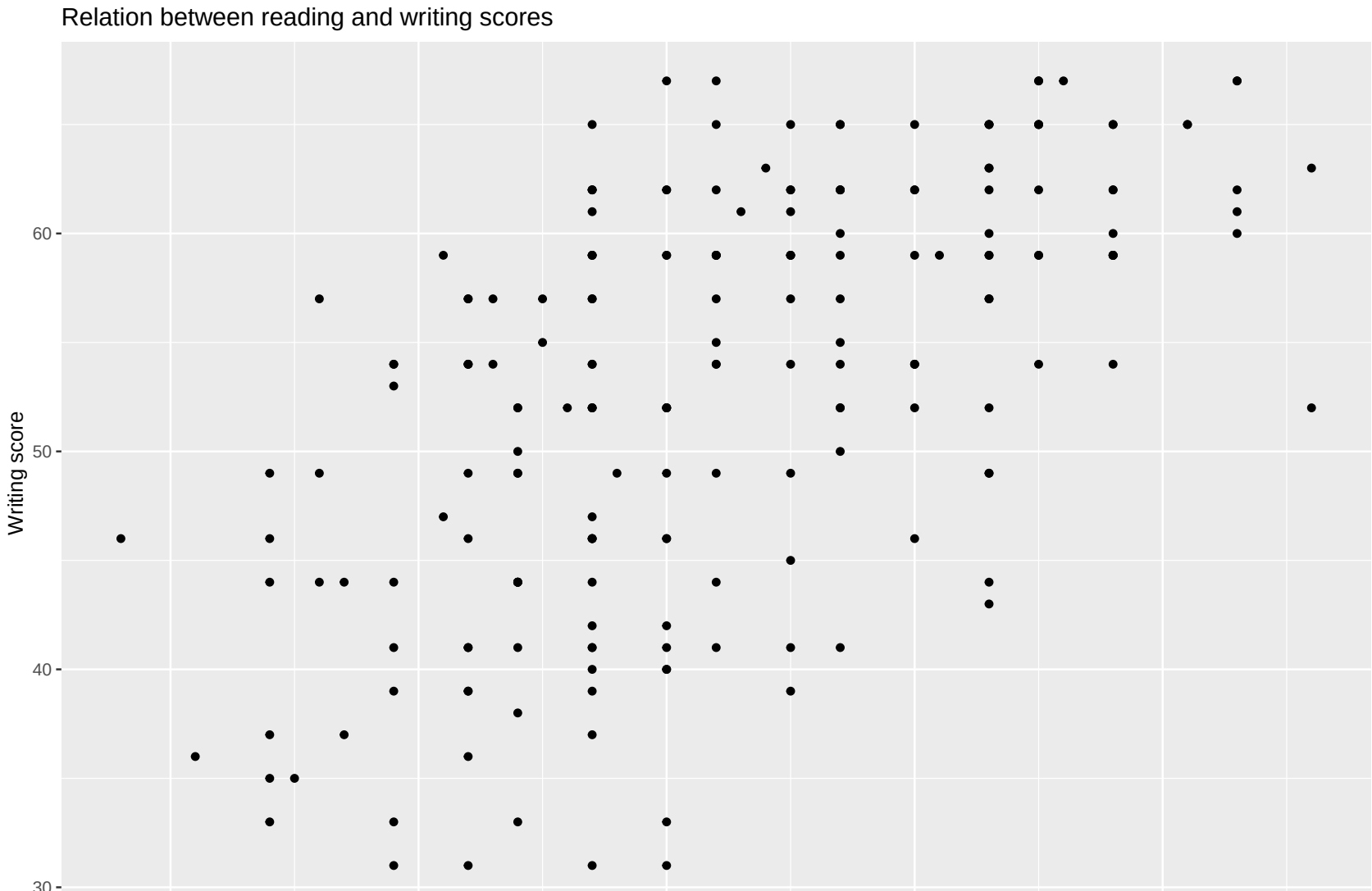
# More beautiful plot: using ggplot

```
ggplot(Hsb, aes(x=read, y=write)) + geom_point()
```



# Add labels

```
ggplot(Hsb, aes(x=read, y=write)) + geom_point() +
  xlab("Reading score") + ylab("Writing score") +
  labs(title = "Relation between reading and writing scores
```



# Change the theme of the plot

```
ggplot(Hsb, aes(x=read, y=write)) + geom_point() +
  xlab("Reading score") + ylab("Writing score") +
  labs(title = "Relation between reading and writing scores
  theme_minimal()
```

# To save the plot to file

```
ggsave("path_to_file.png") # I WILL NOT RUN, WILL DEMO IN C
```

# Correlation

```
cor(read, write)
```

```
## [1] 0.5967765
```

Test significance of the correlation:

```
cor.test(read, write)
```

```
##
##  Pearson's product-moment correlation
##
## data:  read and write
## t = 10.465, df = 198, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to
## 95 percent confidence interval:
##  0.4993831 0.6792753
## sample estimates:
##       cor
## 0.5967765
```

# Regression

- In addition to correlation, we can run a regression between `X` and `Y`
- What is regression?
  - We use OLS to draw a line that show the relation between X and Y
  - There are so many possible lines that can draw thru the scatter plot
  - OLS method chooses the line that minimize the squared errors (a bit technical here, let me explain more!)

# Fit regression

```
ols_reg_fit = lm(formula = write ~ read, data = Hsb)
summary(ols_reg_fit)
```

```
##
## Call:
## lm(formula = write ~ read, data = Hsb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.5447  -5.1225   0.6451   6.3259  15.4553
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.95944    2.80574   8.539 3.55e-15 ***
## read         0.55171    0.05272  10.465  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
```

# Discussion

- The coefficient of `read` is close to the correlation coefficient, but not exactly the same
- It has 3 stars!
  - More stars indicate more significant –> good news, we discover something!
- How to understand this `0.55` coefficient?
  - When `read` score increase 1 point, what will happen to `write` score?
- What is `Intercept` coefficient?
  - What if `read` score = 0?

# Extension 1: More independent variables

- More variables in the right-hand side:
  - Why we put more variables to the regression?
  - E.g., Does gender affect the write score? Why we don't put it to consideration?

# R code example

```
ols_reg_fit = lm(formula = write ~ read + female, data = Hs
summary(ols_reg_fit)
```

```
##
## Call:
## lm(formula = write ~ read + female, data = Hsb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.523  -5.658   0.168   5.043  15.175
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.22837    2.71376   7.454 2.80e-12 ***
## read         0.56589    0.04938  11.459  < 2e-16 ***
## female       5.48689    1.01426   5.410 1.82e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

# Discuss

- `read` coefficient changed!
  - larger or smaller?
  - why? because we controlled for gender!
- Similarly, we can control as nearly all variables we think/find it important
- Which variables we need to control?
  - Ask an expert
  - Read the literature

# Extension 2: transformed variables in regression

- Sometimes, we want to transform variables before putting them to the regression
- For example, we may want to take log of scores before regressions

```
# transform
write_log = log(Hsb$write)
read_log = log(Hsb$read)

# fit
ols_reg_fit = lm(formula = write_log ~ read_log)
summary(ols_reg_fit)
```

```
##
## Call:
## lm(formula = write_log ~ read_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

# A quicker method: using I()

```
ols_reg_fit = lm(formula = I(log(write)) ~ I(log(read)))
summary(ols_reg_fit)
```

```
##
## Call:
## lm(formula = I(log(write)) ~ I(log(read)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50072 -0.08877  0.02909  0.11874  0.29463
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.71628    0.21963   7.814 3.18e-13 ***
## I(log(read))  0.56708    0.05573  10.176  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
```

# Another quiz

- Can you run a regression between `write` and `read` and the squares of `read`?

# Extension 3: Interaction between X variables

- For example: we want to regress `write` on `read`, `female`, and the interaction between these two
- We can do manually, or using the below code:

```
ols_reg_fit = lm(formula = write ~ read*female)
summary(ols_reg_fit)
```

```
##
## Call:
## lm(formula = write ~ read * female)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3247  -5.1255  -0.1181   4.9666  15.5834
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.52388    3.84511   4.297 2.72e-05 ***
## read         0.63602    0.07141   8.907 3.59e-16 ***
```

# Last words for this lecture

- Oops, you may be too tired at this step
- But not yet finished, we need to learn more about assumption diagnostics
- We also learn how to `tidy` the regression results in the next lecture