STAT 140: Week 3

Evening Help Hours

- Emma:
  - Evening Help: Mondays, 7-9pm (Clapp 420)
  - Tutoring Hours:Wednesdays 7-8 (sign up here)
- Anh:
  - Evening Help: Thursdays, 7-9pm (Clapp 420)
  - Tutoring Hours: Fridays, 4-5pm (sign up here)
- As a reminder, it is MHC policy to wear a mask when in a public area–you are required to wear masks during evening help sessions!!

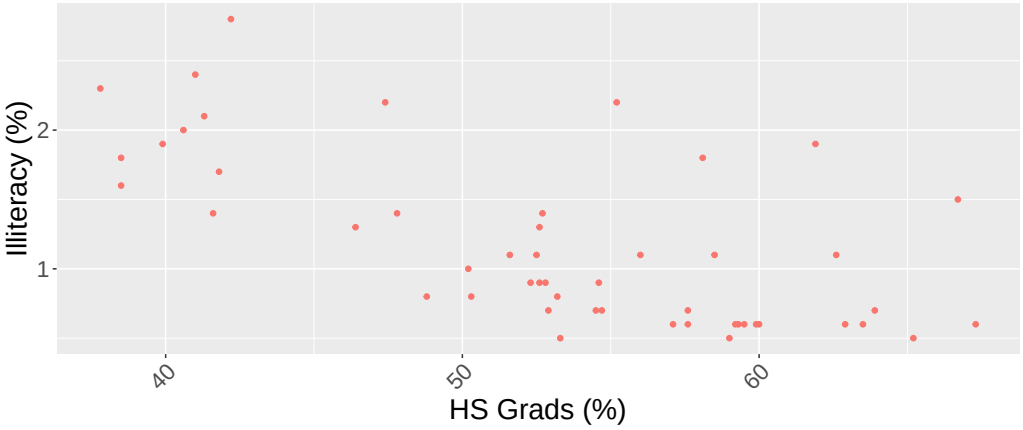
My hope is that after Week 3 (09/13-09/17), you will be able to:

My hope is that after this week, you will be able to:

- Summarize relationships between variables, in graphs and with statistics.
- Interpret statistics for relationships between variables.
- Fit a least squares regression line.
- Determine the appropriateness of a least squares regression line.

Scatterplots 1

- A **scatterplot** provides a case-by-case view of data for two numerical variables.



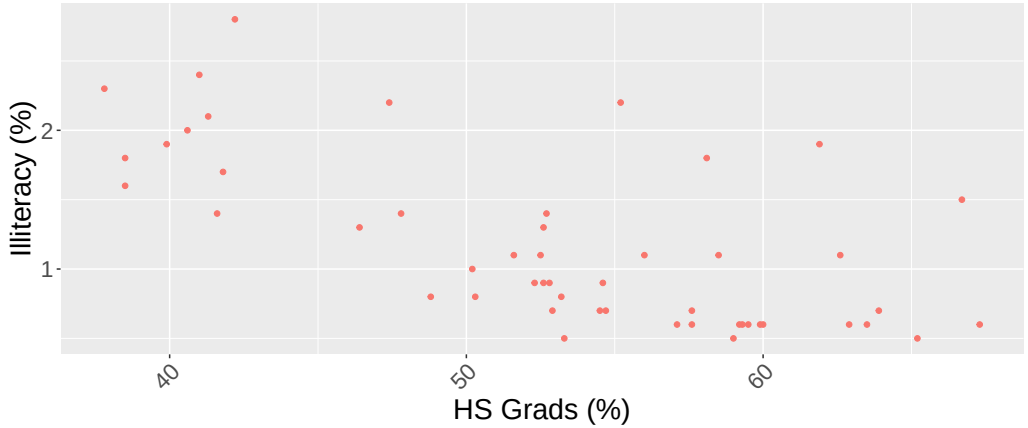
Scatterplots 2

To describe a scatterplot of two numerical variables, you should discuss

- Shape: Does the association look linear or non-linear?
- Direction: Is the association positive or negative?
- Strength: Is the association strong? Do the points look close together, or are they spread out?
- Unusual Values: Are there any points far away from most of the dataset?

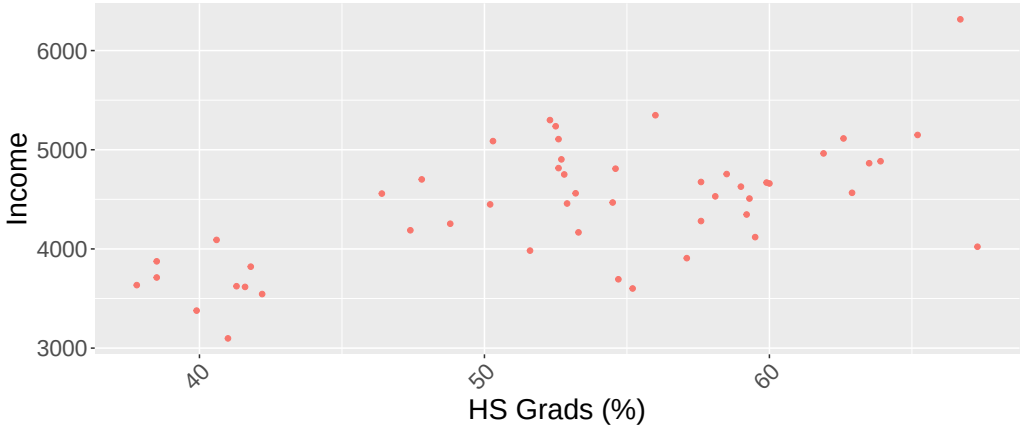
Scatterplots 3

How would you describe the relationship of between high school graduation rate and illiteracy?



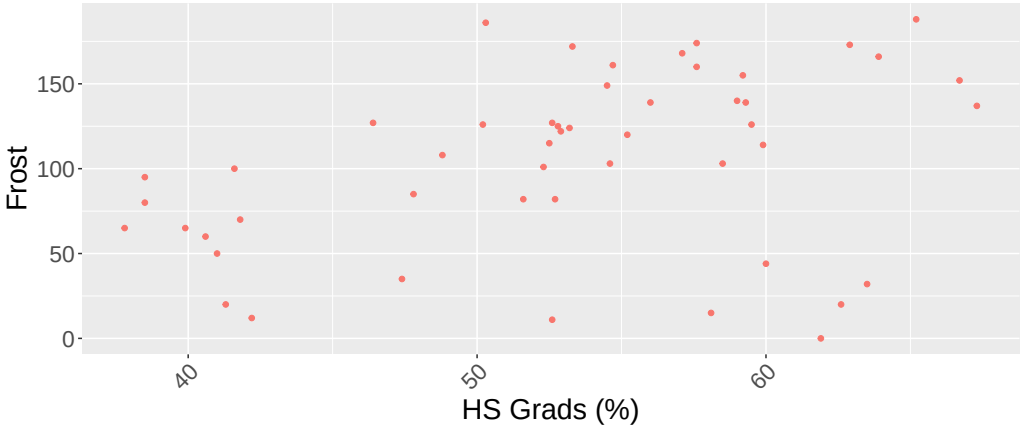
Scatterplots 4

How would you describe the relationship of between high school graduation rate and per capita income?



Scatterplots 5

How would you describe the relationship of between high school graduation rate and mean number of days with minimum temperature below freezing?



Correlation 1

- We could describe a relationship as weak, moderate, or strong, but those are very vague (and probably not super helpful).
- **Correlation**, which always takes values between -1 and 1, formally describes the strength of the linear relationship between two variables. We denote the correlation by R.
  - If X and Y are positively associated, the correlation is positive.
  - If X and Y are negatively associated, the correlation is negative.
  - Strong relationships have correlations close to -1 or 1, and weak relationships have correlations close to 0.

Correlation 2

- Correlation is intended to quantify the strength of a *linear* relationship.
  - When discussing the strength of a scatterplot with a linear relationship, you should report the correlation (if available).
- When possible, look at the correlation *AND* the scatterplot! It is hard to intepret correlation by itself.
  - Under the right circumstances, a weak relationship can have a high correlation (making it look like a strong linear relationship).
  - Nonlinear trends, even when strong, sometimes produce correlations that look weak.
  - http://guessthecorrelation.com/

Correlation 3

- The equation used to calculate correlation can be written as

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

Correlation 4

- To calculate correlation in R...

```
cor(state.x77.df$`HS Grad`, state.x77.df$Illiteracy)
```

```
## [1] -0.6571886
```

```
cor(state.x77.df$`HS Grad`, state.x77.df$Income)
```

```
## [1] 0.6199323
```

```
cor(state.x77.df$`HS Grad`, state.x77.df$Frost)
```

```
## [1] 0.3667797
```

Correlation 5

- The correlation between a state's high school graduation rate and illiteracy is -0.657 (I would call this a moderately strong negative relationship).
- The correlation between a state's high school graduation rate and income is 0.620 (I would call this a moderately strong positive relationship).
- The correlation between a state's high school graduation rate and illiteracy is 0.367 (I would call this a weak positive relationship).

More Statistics for Linear Relationships 1

- Explanatory Variable: A variable that *might* cause a change in another variable.
- Response Variable: A variable that *might* be changed due to changes in another variable.
- Perfect Relationships: If we know the exact value of Y, the response variable, just by knowing the value of X, the explanatory variable, X and Y have a perfect relationship.
- If you can describe the relationship between X and Y using a line, X and Y have a linear relationship.

More Statistics for Linear Relationships 2

- A linear relationship can be written as

$$y = \beta_0 + \beta_1 x$$

- When we use X to predict Y, we usually call X the explanatory or **predictor** variable and Y the response.
- The values β_0 and β_1 represent the model's intercept and slope (also known as model parameters or coefficients).

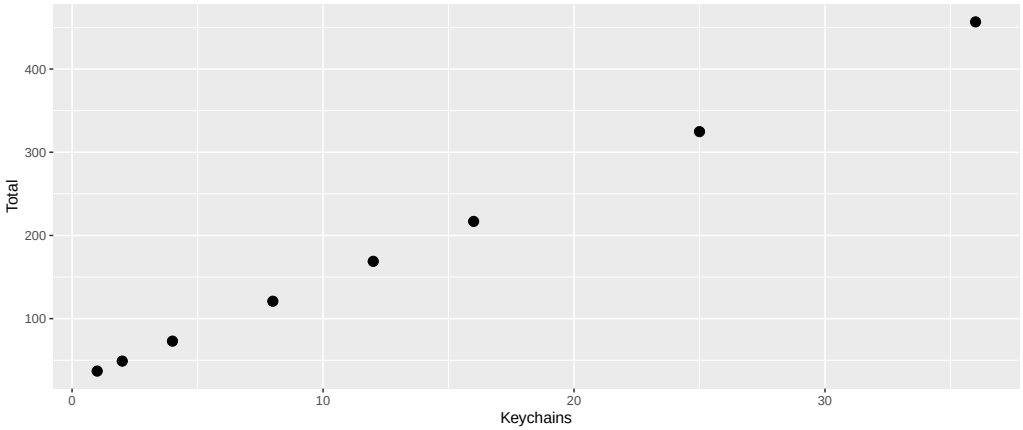
Perfect Linear Relationship 1

- Example:
  - Suppose an Etsy seller makes custom keychains. She charges a \$25 fee to design the keychain and sells the keychains for \$11.99 apiece.
  - If we know the number of keychains we want to buy (X), we can perfectly predict the total amount of money we will be spending (Y).

$$y = 25.00 + 11.99x$$

This is an example of a perfect relationship! We know exactly how much we are planning on paying if we know how many keychains we want to buy.

Perfect Linear Relationship 2



Linear Regression

- It is rare for all of the data to fall perfectly on a straight line, but X and Y can have a linear relationship EVEN IF the relationship is not perfect!
- **Linear regression** is the statistical method of fitting a line to data, where the relationship between two variables, X and Y, can be modeled by a straight line.

$$y = b_0 + b_1 x$$

- b_0 is a statistic estimating the intercept parameter, β_0.
- b_1 is a statistic estimating the slope parameter, β_1.

Finding the Slope

- The slope of regression line can be estimated by

$$b_1 = \frac{s_y}{s_x} R$$

Finding the Intercept

- If $\bar{x}$ is the sample mean of the explanatory variable and $\bar{y}$ is the sample mean of the vertical variable, then the point $(\bar{x}, \bar{y})$ is always on the line.
- We can use the **point slope** form of a line,

$$y - y_0 = slope \times (x - x_0)$$

- Values can be substituted and this can be rearranged for the formula for the intercept,

$$b_0 = \bar{y} - b_1 \times (\bar{x})$$

Identifying the Line

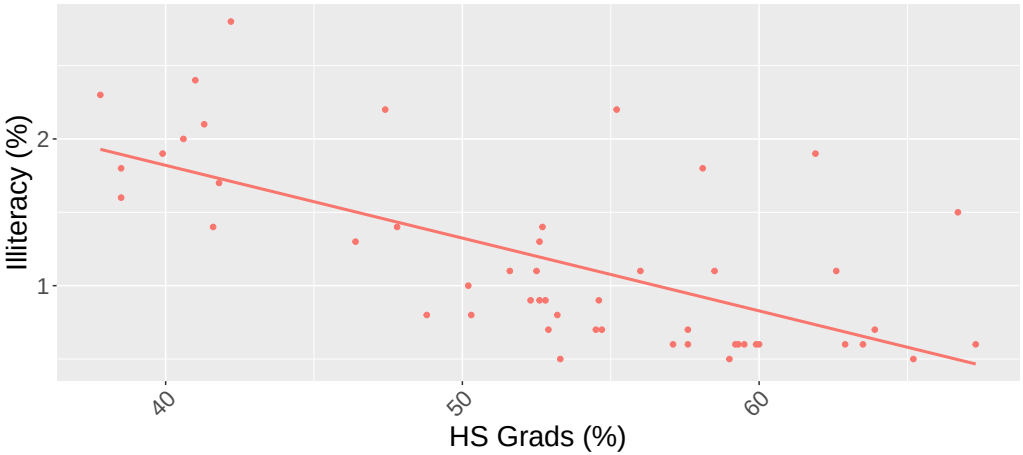
To identify the line from summary statistics:

- Estimate the slope parameter, $b_1 = (s_y/s_x)R$
- Knowing that the point $(\bar{x}, \bar{y})$ is on the line, use $x_0 = \bar{x}$ and $y_0 = \bar{y}$ with the point-slope equation:

$$y - \bar{y} = b_1(x - \bar{x})$$

- Simplify the equation, which would reveal that $b_0 = \bar{y} - b_1\bar{x}$.

Linear Relationships: Example 1



Linear Relationships: Example 2

- Calculate the equation for predicting a state's illiteracy from its high school graduation rate.
- First, we calculate the sample means.

```
x_bar <- mean(state.x77.df$`HS Grad`)
y_bar <- mean(state.x77.df$Illiteracy)

x_bar
```

```
## [1] 53.108
```

```
y_bar
```

```
## [1] 1.17
```

Linear Relationships: Example 3

- Next, we calculate the sample standard deviations.

```
s_x <- sd(state.x77.df$`HS Grad`)
s_y <- sd(state.x77.df$Illiteracy)

s_x
```

```
## [1] 8.076998
```

```
s_y
```

```
## [1] 0.6095331
```

Linear Relationships: Example 4

- The last "ingredient" we need is the correlation.

```
r <- cor(state.x77.df$Illiteracy, state.x77.df$`HS Grad`)
```

- Now, we can calculate the slope:

```
b_1 <- r*s_y/s_x
b_1
```

```
## [1] -0.04959494
```

Linear Relationships: Example 5

- Finally, we can calculate the intercept:

```
b_0 <- y_bar - b_1*x_bar
b_0
```

```
## [1] 3.803888
```

- We can describe the linear relationship between a state's high school graduation rate and illiteracy using the equation:

$$\textit{Illiteracy} = 3.804 + -0.050 \times \textit{HS Grad}$$

Linear Relationships: Example 6

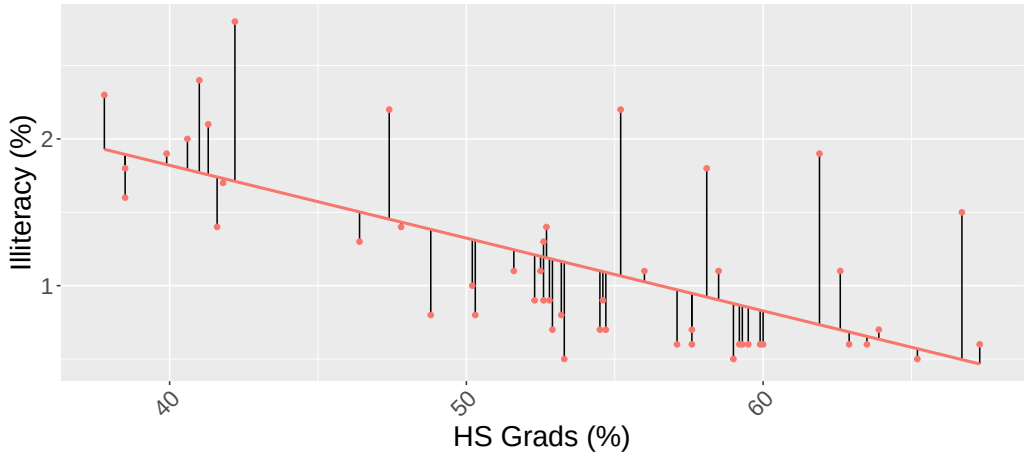

```
model <- lm(data = state.x77.df, formula = Illiteracy ~ `HS Grad`)
model
```

```
##
## Call:
## lm(formula = Illiteracy ~ `HS Grad`, data = state.x77.df)
##
## Coefficients:
## (Intercept)    `HS Grad`
##     3.80389     -0.04959
```

Interpreting Linear Relationships 1

- The slope of a line can be interpreted as the expected increase or decrease in the Y variable that would occur if the explanatory variable X increased by one unit.
  - For example, for every one percent increase in high school graduation rate, a state's illiteracy rate is expected to decrease by 0.050%.
- The intercept of a line can be interpreted as the expected value of the Y variable if the explanatory variable X were 0.
  - For example, when a state has a high school graduation rate of 0%, we expect their illiteracy rate to be 3.80%.

Interpreting Linear Relationships 2



Predicting Y 1

- It is rare for all of the data to fall perfectly on a straight line. It's more common for the data to appear as a *cloud of points*.
- This is the same as saying there is often some error associated with predicting Y based on X. We express this as:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Here, ε represents the error.

Predicting y 2

- We can describe the linear relationship between illiteracy and high school graduation rate using the equation:

$$\textit{Illiteracy} = 3.804 + -0.050 \times \textit{HS Grad}$$

- What percent illiteracy do we predict for Massachusetts (58.5% HS Grad)? For Illinois (52.6%)? Texas (47.4%)?

Predicting y 3

- A "hat" on Illiteracy indicates that we are estimating the illiteracy rate.
- For Massachusetts...

$$\widehat{Illiteracy} = 3.804 + -0.050 \times 58.5$$

- We predict a 0.879% illiteracy rate for a state with a high school graduation rate of 58.5%.

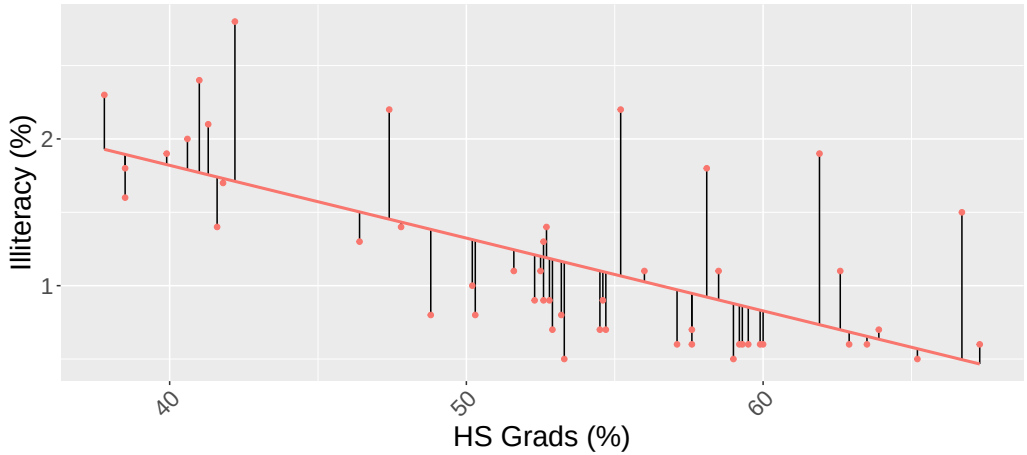
Residuals 1

- **Residuals** are the leftover variation in the data after accounting for model fit:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Each observation will have a residual.
- You can also think of the residual as being the vertical distance from an observation to the line (model) describing the data.
- If the observation is above the regression line, then its residual is positive.
- If the observation is below the regression line, then its residual is negative.

Residuals 2



Residuals 3

The residual of the i^{th} observation (x_i, y_i) is the difference of the observed response (y_i) and the response we would predict based on the model fit $(\hat{y}_i)$:

$$e_i = y_i - \hat{y}_i$$

- If we predicted an illiteracy rate of 0.879% for Massachusetts and the illiteracy rate was actually 1.1%, the residual is $e_{MA} = 1.1 - 0.879 = 0.221$.
- If we predicted an illiteracy rate of 1.120% for Illinois and the illiteracy rate was actually 0.9%, the residual is $e_{IL} = 0.9 - 1.120 = -0.295$.
- The size of a residual is usually discussed in terms of its absolute value. In this case, a residual of -0.295 is actually bigger than a residual of 0.221, since it is further away from the line.

Least Squares Regression 1

- We want the "best" fit for a linear model.
- Fitting linear models by eye can be subjective and does not always result in the line of best fit.
- How do we fit the line of best fit? What does best fit even mean?
  - One goal in picking the right linear model is for these residuals to be as small as possible.

Least Squares Regression 2

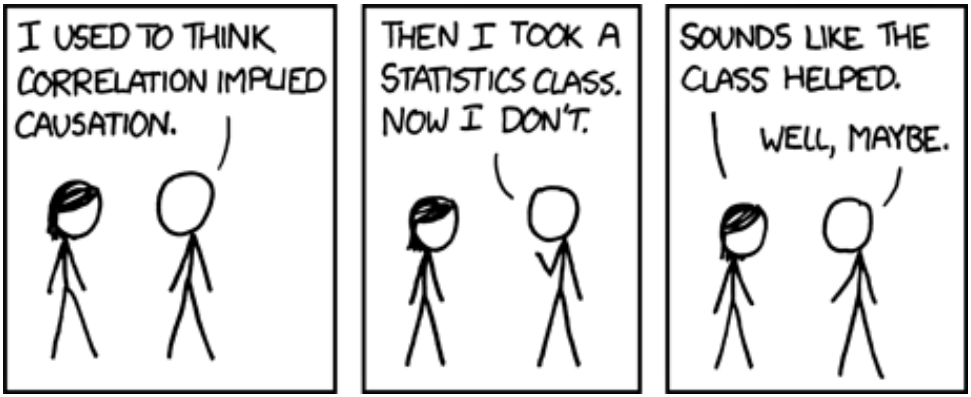
- In statistics, one common way of defining the "best" fit is by minimizing the residuals (a.k.a., minimize the distances of the points to the line).
- One common way to do this is by minimizing the sum of the squared residuals.
- The resulting line is called the **least squares line**, and the formulas for estimating the least squares line are

$$b_1 = \frac{s_y}{s_x} R$$

and

$$b_0 = \bar{y} - b_1 \times (\bar{x})$$

Correlation and Causality



https://xkcd.com/552/

Establishing Causality 1

If you are going to establish a causal relationship, you must use an experiment. However, experiments are not always enough. You also need:

1. Theoretical Justification: Using theoretical knowledge and context clues, there must be an explanation of **WHY** X might cause Y.

2. Temporal Sequencing: X **MUST** come before Y to show X causes Y.

3. Evidence of Association: There must be empirical evidence of an association between X and Y to show X causes Y.

- Empirical — coming from data.
- This is where we talk about things like correlation and scatterplots!

Establishing Causality 2

4. Elimination of Alternate Causes: There must be no other reasonable explanation for why X causes Y.

- This includes elimination of confounding variables!
- If, and only if, you have shown all four of these things, you can imply causality.

Establishing Causality 3

Do observational studies have value?

- Establishing causation is not always the goal!
  - Sometimes it is enough to know that there are differences in two groups without needing to know why those differences exist.
- Establishing causation can be done in other ways (RARELY).
  - Sometimes there is a theoretical reason that confounding factors cannot be responsible for changes in the response variable.
  - Sometimes it is unethical to run a randomized experiment.
  - Example: Does smoking cause lung cancer?
- Analysis of observational data may lend evidence toward causal theories and suggest the direction of future research.
  - Example: Association between smoking and lung cancer.