MBA 753 : Causal Inference Methods for Business Analytics

Dr. Nivedita Bhaktha

Lecture 1

01.08.2024

Agenda

- Housekeeping
- Intro to causal inference
- Intro to regression

Housekeeping

Course Outline

- First course handout has been shared

August

SUN	MON	TUE	WED	THU	FRI	SAT
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

September

SUN	MON	TUE	WED	THU	FRI	SAT
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

Intro to Causal Inference

Why Learn Causal Inference?

Intermittent fasting over two days can help people with Type 2 diabetes

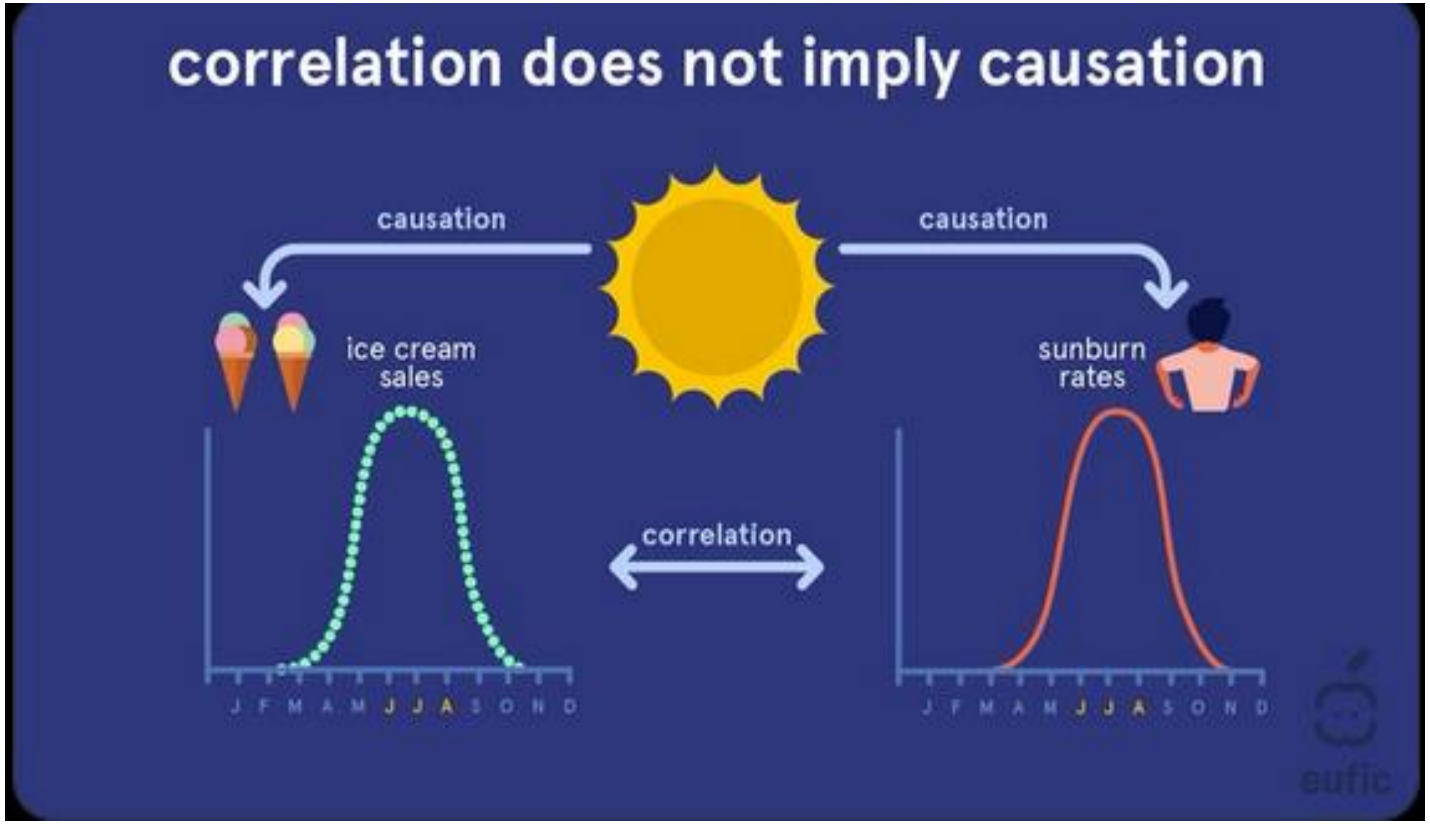
A study found that intermittent fasting had striking metabolic benefits that surpassed the effects of prescription drugs for people with newly diagnosed diabetes.

Morning workouts may be better for weight loss, study finds

Drinking water from plastic bottles causes diabetes?

Why September Babies Are More Successful, According to Science

Does Correlation imply Causation?



Source: European Food Information Council

Spurious Correlation

Source: http://bit.ly/3Yjy3EX

Causal Inference – Terminology

- Causality: X causes Y
  - Causal phrases have direction
  - Changing X results in a change in the distribution of Y
- Most research questions (RQ) are causal in nature
  - To answer such RQ, we are interested in study the data generating process (DGP)
- Causal Diagram: Variables and the causal relationship in the DGP

Terminology II

- Research Question: questions that are answered through research
  - Properties: Well-defined, answerable, understandable
  - From theory to hypothesis
- Empirical Research: describing the density functions of a statistical variable
  - Variable: Any characteristics, property, or quantity that can be measured or counted
    - Scales of measurement: Nominal, Ordinal, Interval, and Ratio
  - Distribution: how often different values occur
    - Density function: function that defines a relationship between a random variable and its probability
    - Summarizing the distribution: produce few numbers that describe the entire distribution

X Y
0 8

−0.6

Terminology III

- Describing Relationships: learning about one variable given the value of another variable
  - Correlation: describes the extent to which two random variables are linearly related

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

    - Covariance: measures the amount of linear dependence between two random variables

    - Conditional distribution: distribution of one variable given the value of another variable

Regression Analysis

Intro to Regression

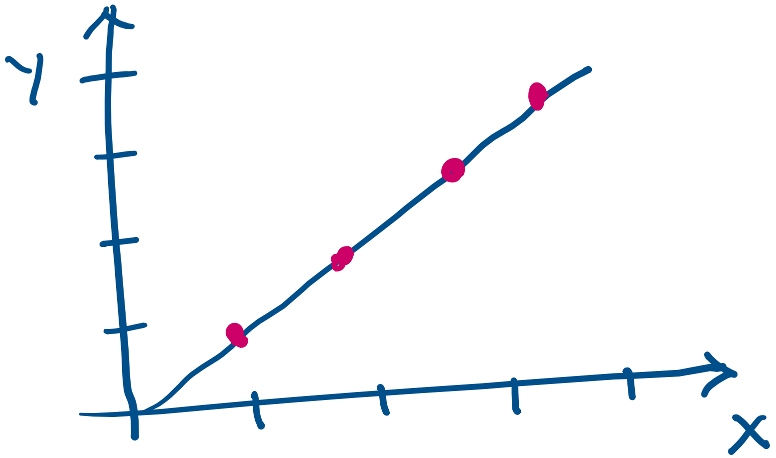
Regression Analysis

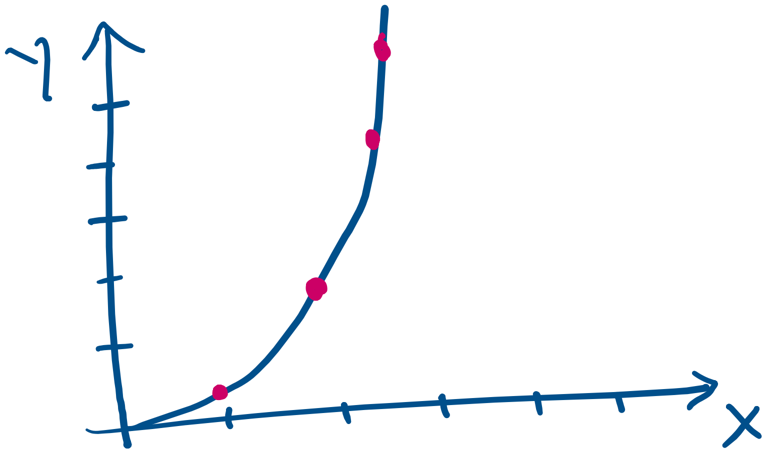
- Analysis of the relationship between two or more variables
- Function: a mathematical relationship to predict values of one variable (**Y**) corresponding to given values of another variable (**X**)
- **Y**: is referred to as the dependent variable, the response variable or the predicted variable
- **X**: is referred to as the independent variable, the explanatory variable or the predictor variable

$$Y = f(X)$$

Functional Relationship

- Relationship can be expressed by a mathematical formula $Y = f(X)$
  - All observations fall directly on the line/curve





- In most empirical studies, observations might not follow such a perfect functional relationship

Regression Analysis - Examples

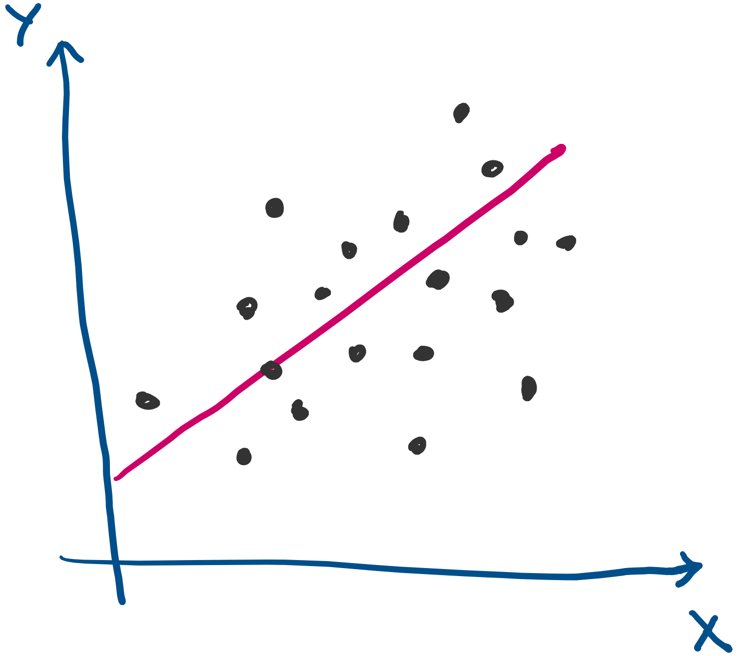
- Y is a function of X

Y	X
The time needed to fill a soft drink vending machine	The number of cases needed to fill the machine
Maintenance cost of cars	The age of cars
Yield of wheat per acre	Fertilizers used
Grades obtained in Causal Inference class	The number of hours spent studying

Scatter Plots

- Represents relationship between two variables
- Observations don't perfectly lie on the curve/line
- So named because of scattering of points around the line/curve

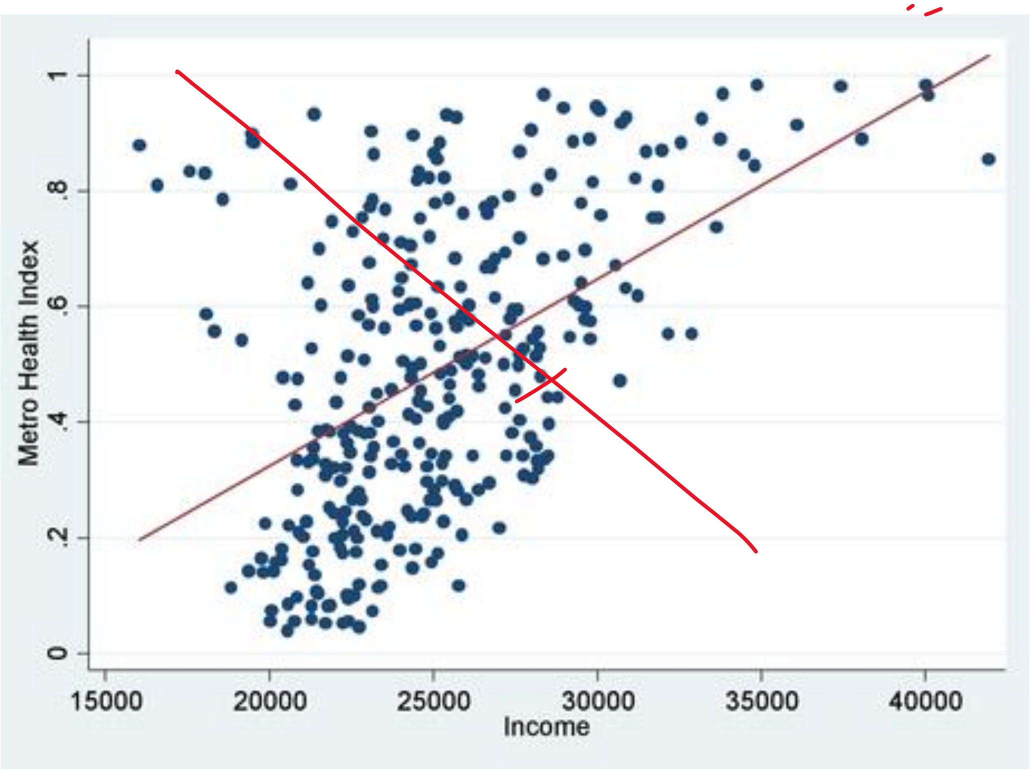
- Scattering represents variations in Y that is not associated with X and is random



$$Y = f(X) + \varepsilon$$

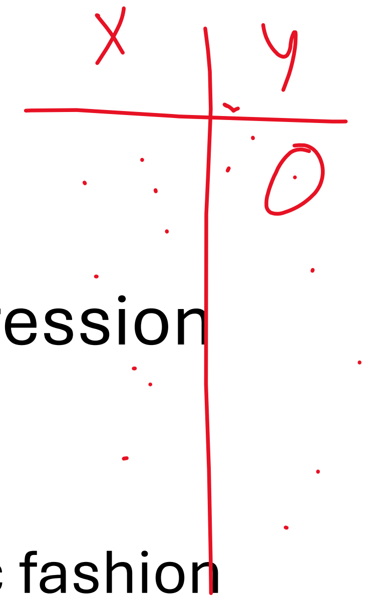
Scatter Plots

- What are X and Y?
- Describe the relationship between X and Y.
- Is the slope positive or negative?



Source: Minnesota Department of Health

Regression Function

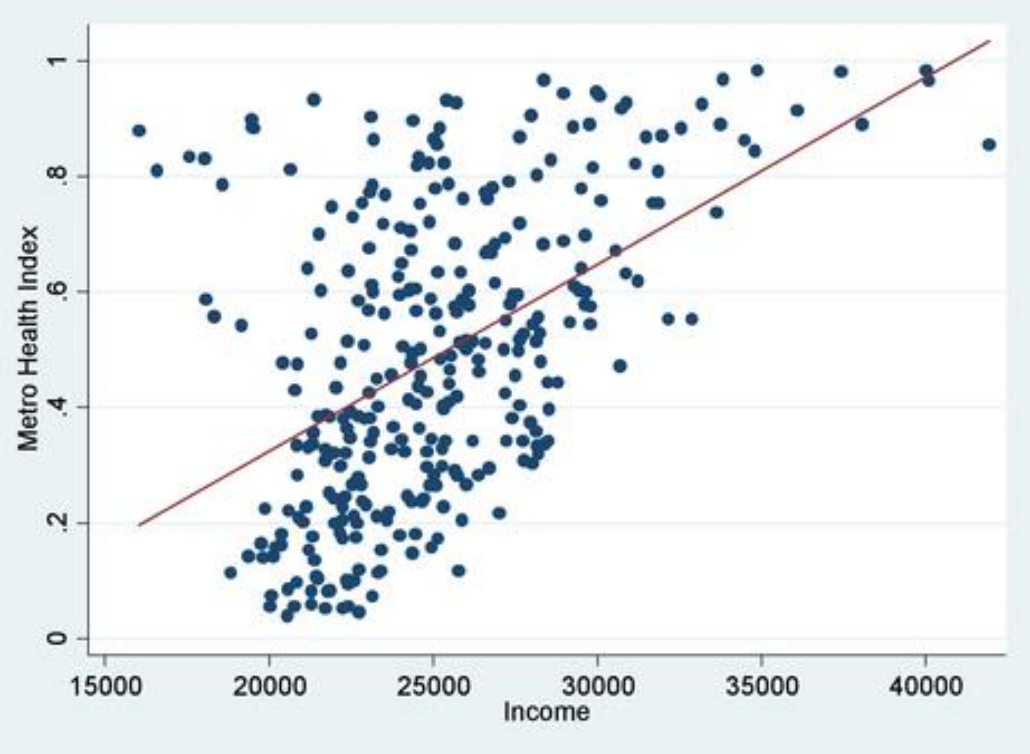


- Regression analysis is a formal means of expressing a regression function: $Y = f(X) + \varepsilon \rightarrow$ random
  sys
  - Probability distribution of Y for each level of X $\rightarrow$ $P(Y|X)$
  - Means of these probability distributions vary in some systematic fashion with X

- $f(X)$ is the systematic regression function
  - It is not known in advance and must be determined

Determining Regression Function

- The red line is the line of best fit
  - Minimizes the error between the predicted values and the observed values
- The line passes through conditional mean
  - Conditional mean: Mean of Y given X
- The approach used for finding such a line is called Ordinary Least Squares (OLS)



Source: Minnesota Department of Health

OLS Approach

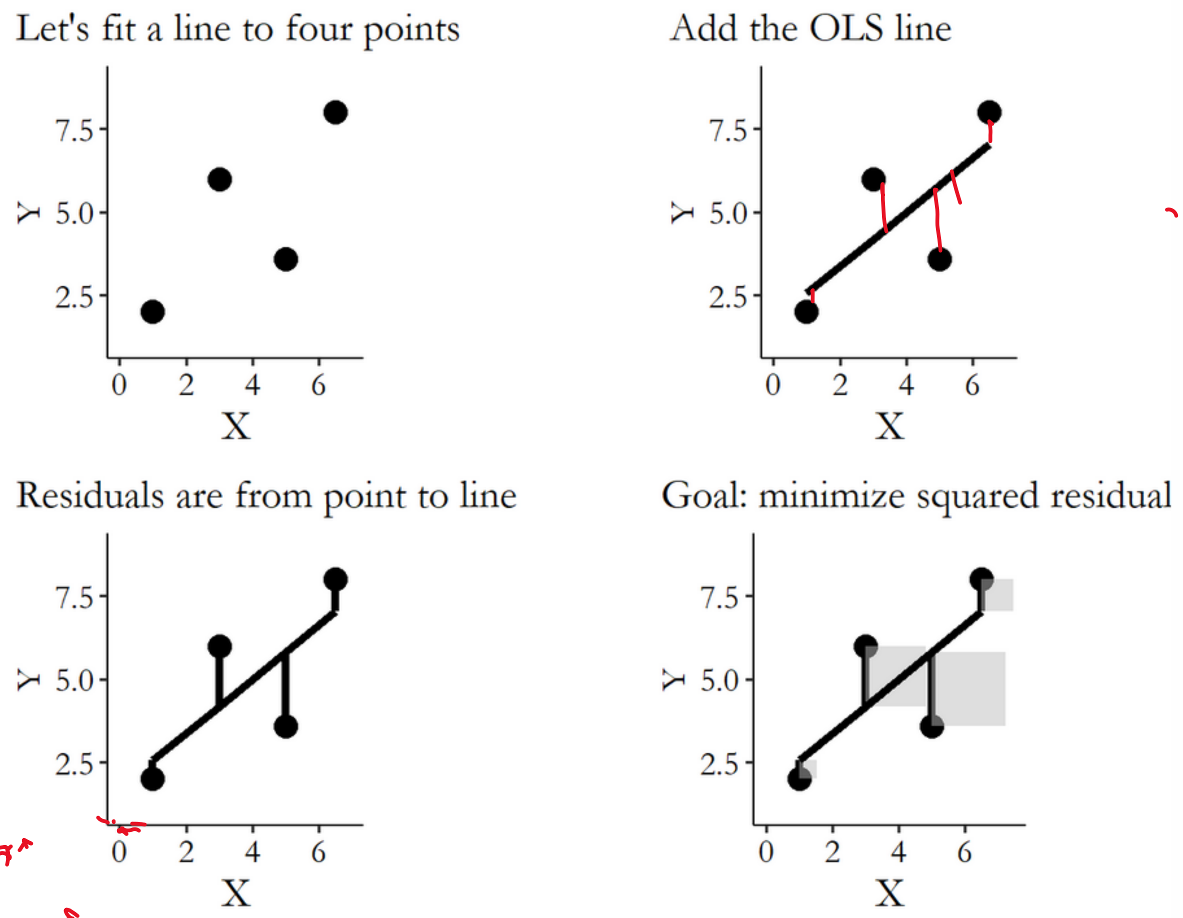
- OLS picks the line that minimizes the sum of squared residuals
- **Residual:** difference between an observation's actual value and the predicted value
  - Predicted value: $\widehat{Y}_i = b_0 + b_1 X_i$
  - Residual: $\boldsymbol{\varepsilon_i = Y_i - \widehat{Y}_i}$

$Y = \beta_0 + \beta_1 X$

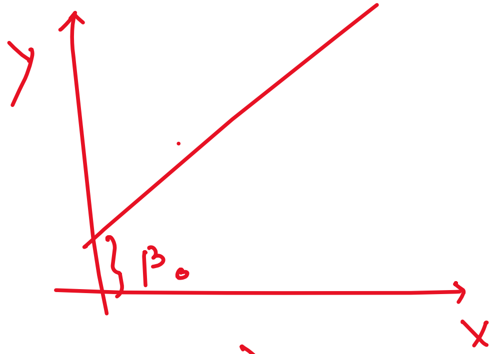
Y_i, X_i

collected (known)

$\sum_{i=1}^{n} |e_i|$



OLS Estimation



- Minimize $\sum_{i=1}^{n} \varepsilon_i^2$ using calculus
  - slope → $\hat{\beta}_1 = r \frac{S_y}{S_x}$ SD(Y) SD(X) cor
  - Intercept → $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 * \bar{x}$ → mean x

  mean Y

$Y = \beta_0 + \beta_1 X$

$\hat{\beta}_0 \quad \hat{\beta}_1$ → estimates

$\hat{Y}_i$

- Predicted values: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

- This model is linear in parameters and linear in predictors
- The conditional mean is $E(Y|X) = \beta_0 + \beta_1 X$

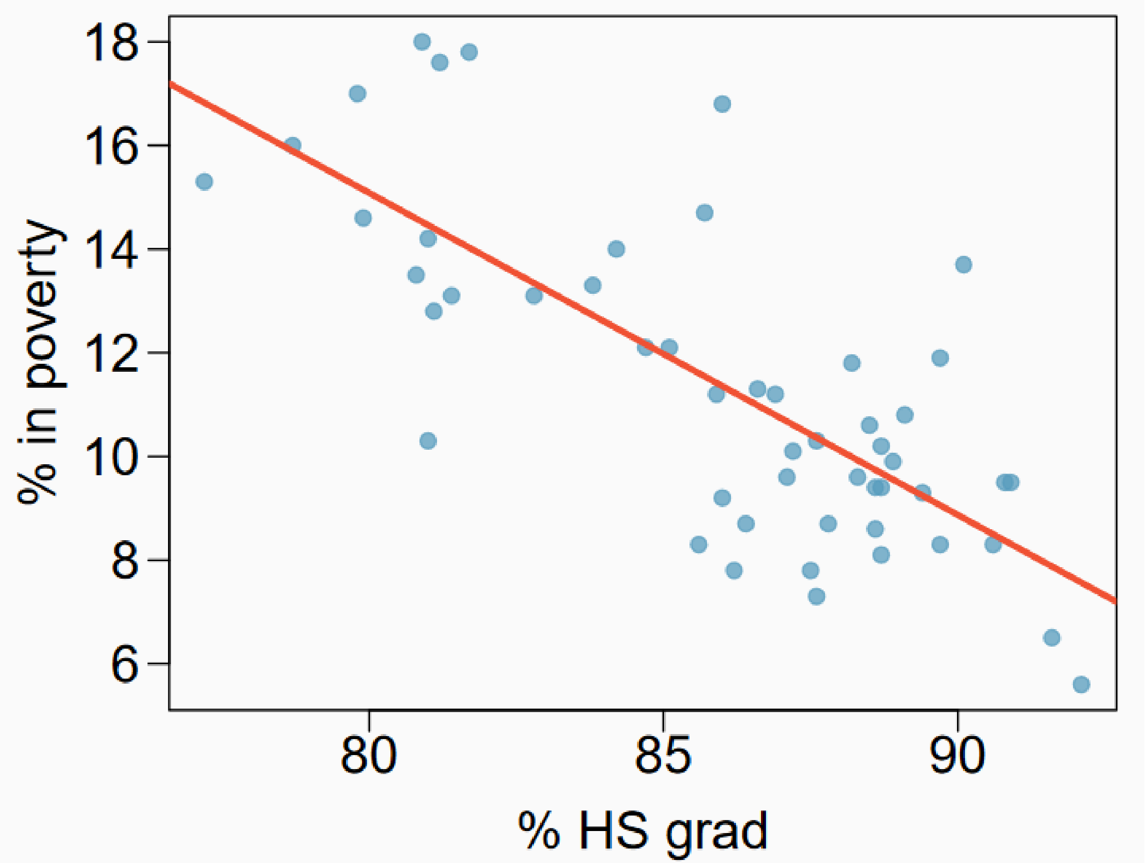
Expected value

Example - Estimation

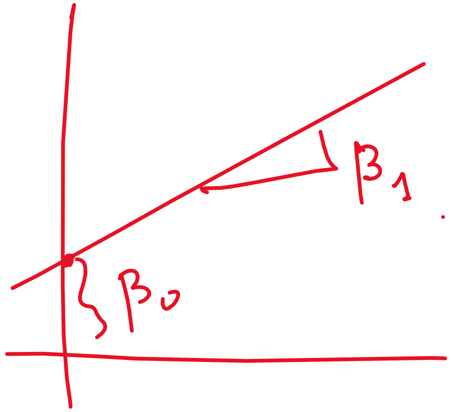
	% HS Grad (X)	% in Poverty (Y)
Mean	86.01	11.35
SD	3.73	3.1
r	-0.75	

$$\hat{\beta}_1 = r \cdot \frac{s_y}{s_x} = -0.62$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} = 64.67$$



Interpretation of Parameters



$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- β_0 and β_1 are parameters or regression coefficients
- ε_i is the random error

$E(Y|X) = \beta_0 + \beta_1 X = \beta_0$

$E(Y|X=0) = \beta_0$

- β_0 is the Y intercept of the regression line
  - When the scope of the model includes X = 0, β_0 is the mean of the probability distribution of Y at X = 0
- β_1 is the slope of the regression line
  - It gives the change in mean of the probability distribution of Y per unit increase in X

Example - Interpretation

$$\widehat{\% \, in \, poverty} = 64.68 - 0.62 * \% \, HS \, grad$$

β_0: 64.68

0

Increase/ decrease

Are the Interpretations Causal?

$X = \alpha_0 + \alpha_1 Y + \varepsilon$

- No cause-and-effect pattern is necessarily implied by the regression model

% HS grad reducing % poverty
affecting
leading

$\beta_0 \quad \beta_1$

$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{cov(Y,X)}{\sigma_X \sigma_Y}$

$\hat{\alpha}_1 = r \frac{S_X}{S_Y}$

$\hat{\beta}_1 = r \frac{S_Y}{S_X}$

$cov(X,Y) = cov(Y,X)$

Recap

Summary

- Correlation does not imply causation
- We answer causal research questions through empirical research
  - Empirical research involves describing relationships among two or more variables.
- Regression analysis provides functional relationship between two or more variables
- OLS is used to estimate the parameters of a regression equation

- Objectives achieved:
  - Can differentiate between causal and associative relationships
  - Can interpret intercepts and slopes

References

- Joshua D. Angrist and Jörn-Steffen Pischke, Mostly Harmless Econometrics, Princeton University Press.
- Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. Applied linear statistical models. McGraw Hill Education.
- Scott Cunningham, Causal Inference: The Mix Tape, Yale University Press.

Thank You ☺

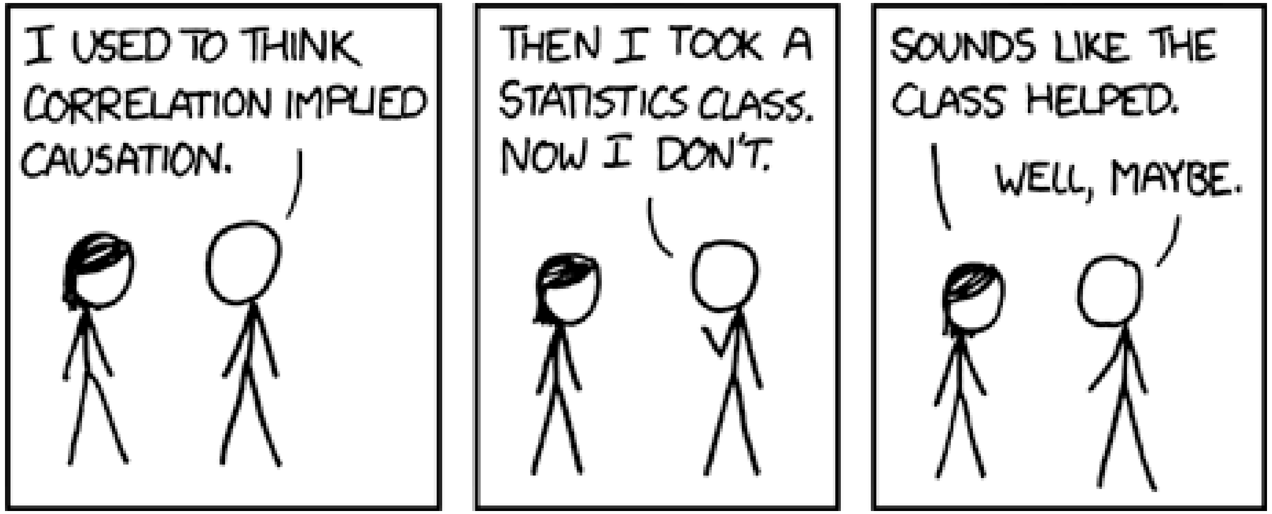


Image Source: XKCD