Predict V – Recall

$$V = \bar{V} + e$$

$$\hat{V} = \bar{V} + s$$

$$V = \hat{V} - s + e$$

This is your error. You do not observe s and e separately. You do not know whether you are making error s or e

- s = systematic error
  - depends on biases or logical fallacies
- e = unsystematic (random) error
  - it is literally unpredictable, but can be reduced if you have a larger sample size
- $\hat{V}$ is your prediction, while $\bar{V}$ is the «perfect» prediction that you can make if you had no bias (no systematic error)

Goal

$$V = \hat{V} - s + e$$

- Goal is to predict V
- Your theory tells you that variables X that you observe predict V

$$V = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + error$$

$\hat{V}$

Regression

- **Regression analysis** is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain

Independent variables: the variables used to explain the dependent variable

Regression analysis

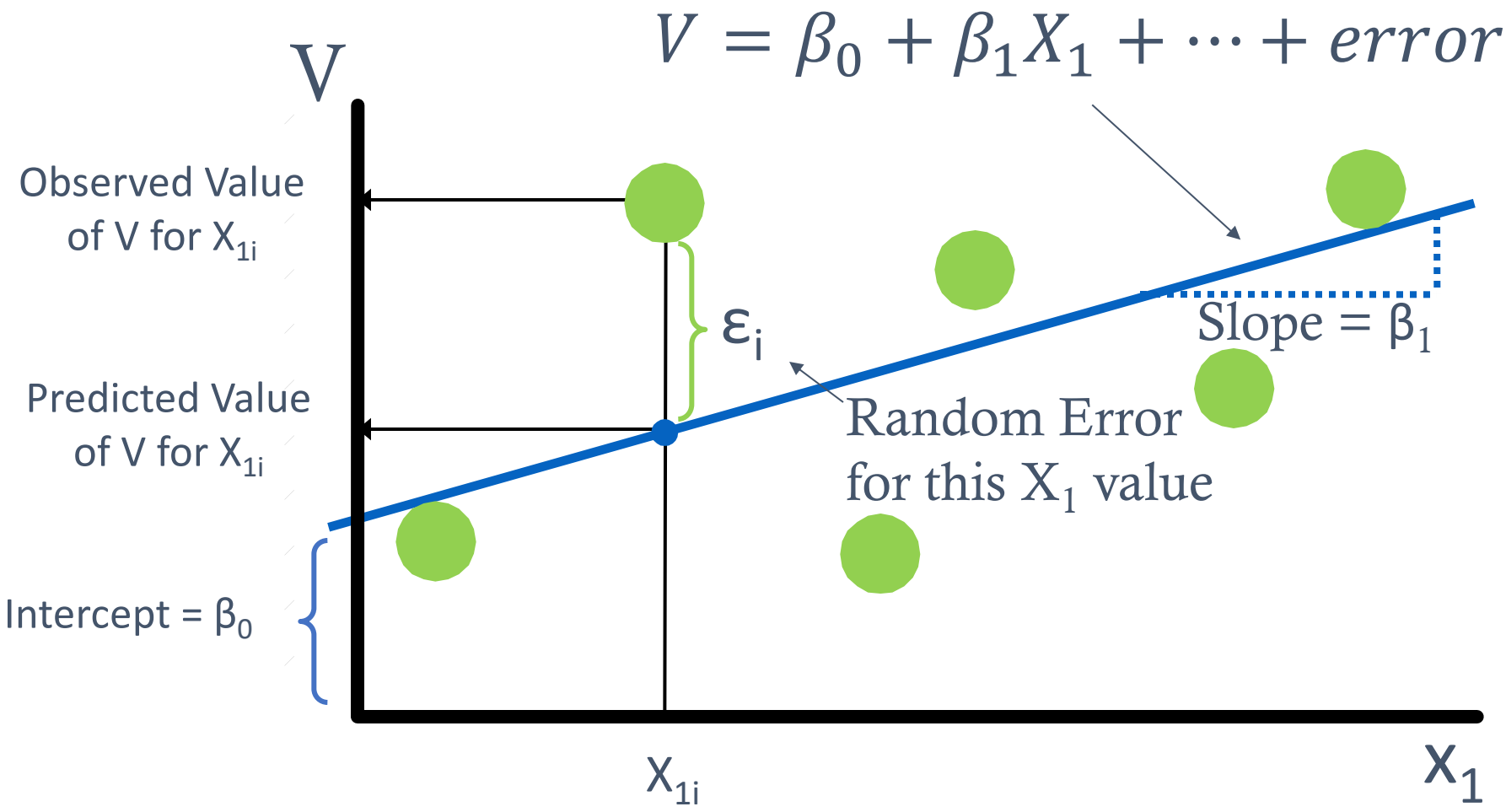
$$V = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + error$$

- Relationship between V and Xs is described by a linear function
- Regression analysis measures correlations, unless the research design is structured in such a way as to nail down causality (see later and next week)
- Interpretation of the coefficients: β_1 impact of X_1 keeping constant X_2, X_3, X_4, ...

Assumptions

- **Linearity** (but you can adjust the model for non-linear relationship, see next)
- **Homoskedasticity: Constance variance for all level of X** (important for testing)
- **X and error not correlated** (Crucial→Causality)

Linear Regression



Least Squares Criterion

- Estimates obtained by finding the values of the β's that **minimize the sum of the squared residuals**

$$\underset{\boldsymbol{\beta}}{\text{Min}} \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (V_i - \hat{V}_i)^2 =$$

$$\sum_{i=1}^{n} (V_i - \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots)^2$$

OLS estimator

- In the simple linear regression (one independent variable) $V = \beta_0 + \beta_1 X_1 + error$ the formula for the OLS estimator of β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(V_i - \bar{V})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

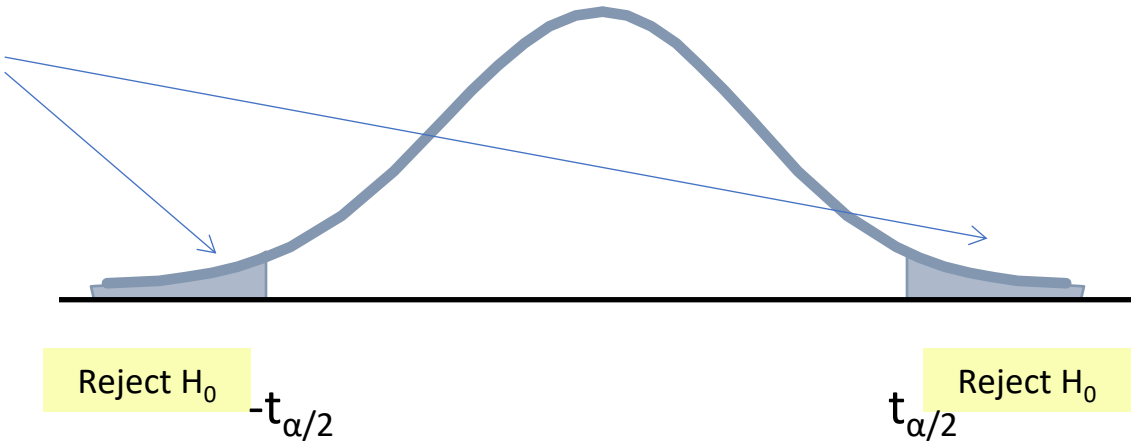
- In general for $V = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + error$

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}(X'V)$$

Statistical significance of $\hat{\beta}_1$

- Null and alternative hypotheses
  - H_0: $\beta_1 = 0$ (no linear relationship)
  - H_1: $\beta_1 \neq 0$ (linear relationship does exist)
- Test statistic

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_\beta}$$



Properties of OLS estimator

- **B**est: in terms of efficiency
- **L**inear
- **U**nbiased: the expected value of the estimator $\hat{\beta}$ is equal to the true value of the parameter
- **E**stimator

- Estimator is also **consistent** … converges to true parameter as sample size gets larger and larger
- It is crucial in $V = \hat{V} - s + e$ …
  - *The error e reduces as you increase sample, but not s … then if there is s you do not get to the right prediction of V even if you increase your sample size as much as you can*

Interpretation of the slope and the Intercept

- β_0 is the estimated average value of V when the value of the X's is zero

- β_1 is the estimated change in the average value of V as a result of a one-unit change in X_1 (and similarly for the other X's)

Interpretation of β_1

$$V = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \ldots + error$$

$$\beta_1 = \beta_0 + \beta_1(\mathbf{X_1 + 1}) + \beta_2 \mathrm{x}_2 + \ldots error - (\beta_0 + \beta_1 \mathbf{X_1} + \beta_2 \mathrm{X}_2 + error)$$

$$\mathbf{d}V = \beta_1 \mathbf{dX_1}$$

Interpretation of β_1

- Unit of measures and specifications are important!!
- $V = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + error$
  - $\mathbf{d}V = \beta_1 \mathbf{d}X_1 \rightarrow \beta_1 =$ # units of increase of V following a 1 unit increase in X_1 (be sure you understand what the units of measure of X_1 and V are)
- $lnV = \beta_0 + \beta_1 lnX_1 + \beta_2 lnX_2 + \ldots + error$
  - $\mathbf{d}lnV = \beta_1 \mathbf{d}lnX_1 \rightarrow \beta_1 =$ % increase in V following a 1% increase in X_1
  - $\mathbf{d}lnV/\mathbf{d}lnX_1 = (dV/V)/(dX_1/X_1)$
- $lnV = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + error$
  - $\mathbf{d}lnV = \beta_1 \mathbf{d}X_1 \rightarrow \beta_1 =$ % increase in V following a 1 unit increase in X_1
  - $\mathbf{d}lnV/\mathbf{d}X_1 = (dV/V)/dX_1$

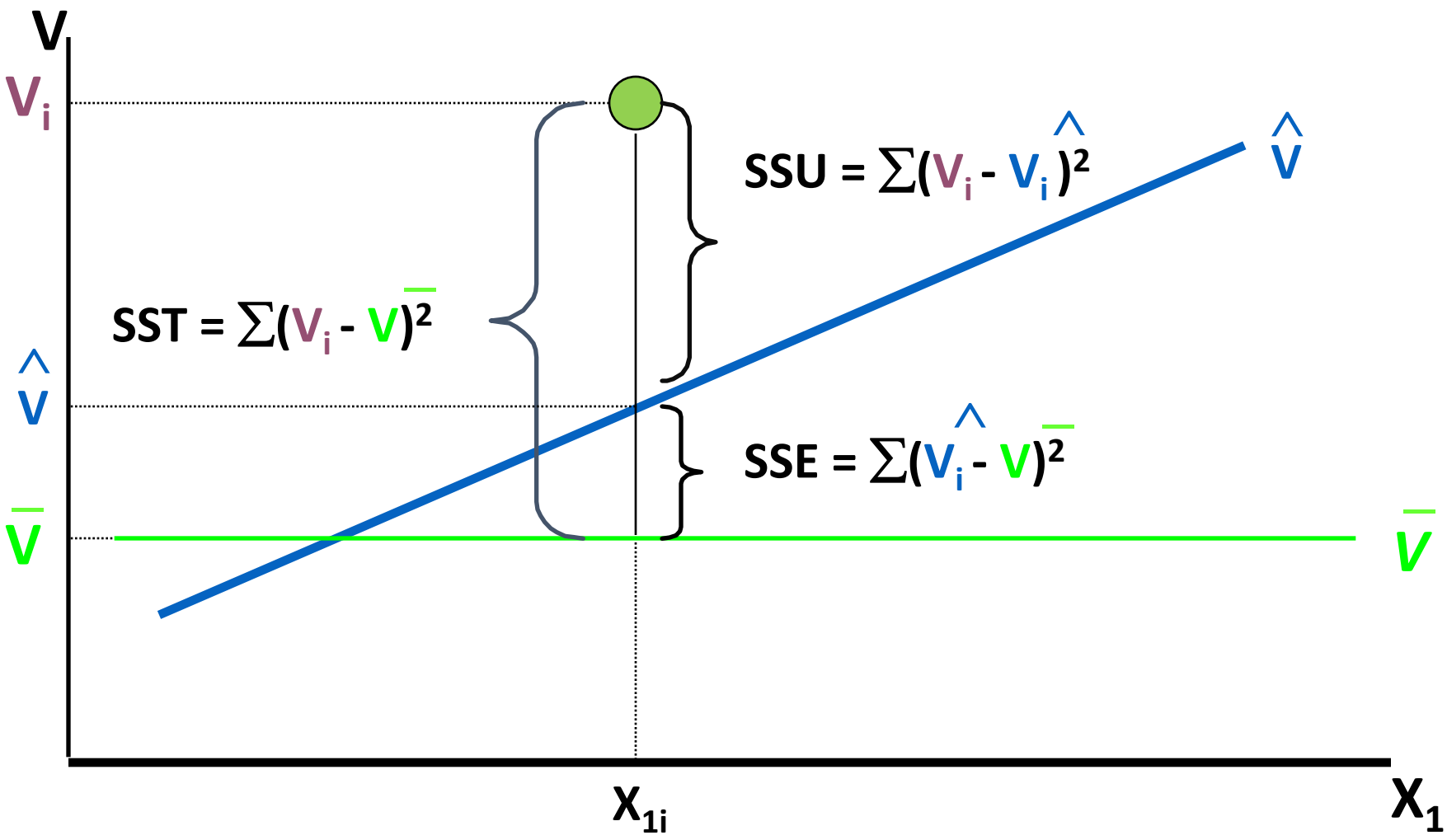
Coefficient of Determination, R^2

- The **coefficient of determination** is the portion of the total variation in the dependent variable that is explained by variation in the independent variable

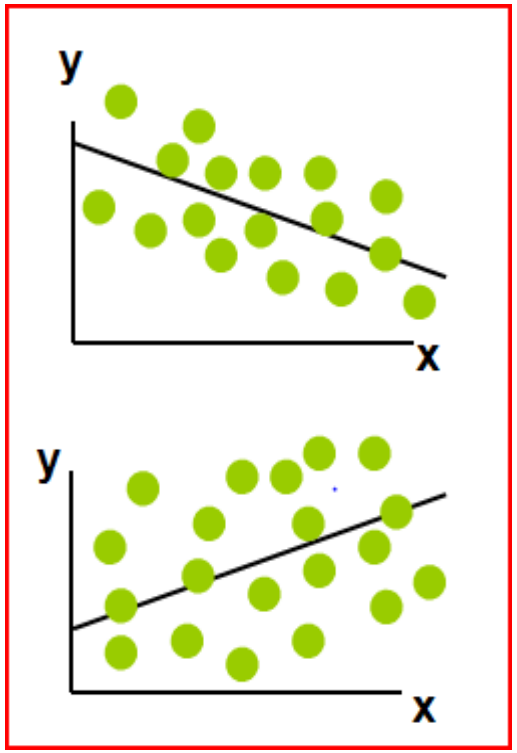
- The coefficient of determination is also called **R-squared** and is denoted as R^2

$$R^2 = \frac{SSE}{SST}$$ where $0 \leq R^2 \leq 1$

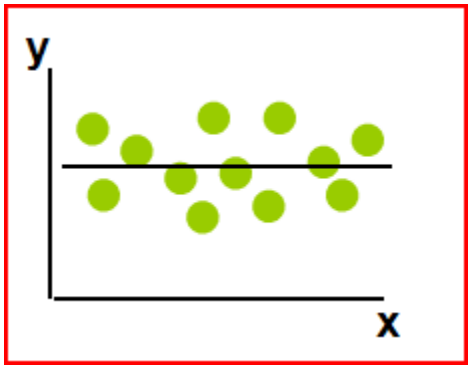
Explained and Unexplained Variation



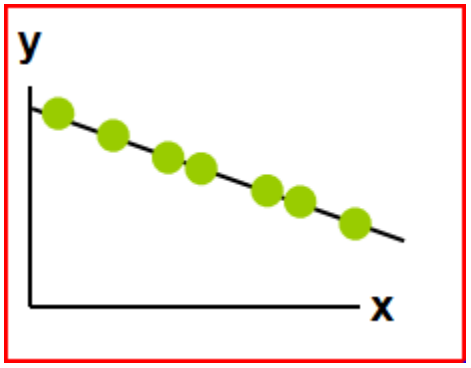
Examples of R^2



$0 < R^2 < 1$



$R^2 = 0$



$R^2 = 1$

Overall statistical significance of the regression

- The F test:

$$F = \frac{R^2}{1-R^2}\frac{n-p}{p-1}$$

Correlation vs causality

- Examples – correlation or causality?
  - *Younger entrepreneurs make more profits*
  - *Start-ups with more qualified employees make more profits*

Correlation vs causality

- There are two main reasons why there can be correlation without causality
  - A third variable is «causing» both variables (*omitted variable*)
  - There is *reverse causality*
- Young people have higher propensity to take risky ventures, which are on average more profitable (omitted variable)
- A more profitable company pays higher salaries which attracts more qualified employees (reverse causality)

Correlation vs causality

- Why do we care?
- With correlations we can make wrong decisions because we confuse them for causal relationships (a sort of bias once again)
- E.g.
  - You hire a young entrepreneur as your partner, but he has a family with three kids and is less likely to take risky ventures, which makes the start-up, on average, less profitable
  - A start-up hires qualified employees who are costly and do not help the company to become more profitable

Correlation vs causality

- However, correlations are not useless:

1) If you know that two variable are correlated, you can still make predictions because they move together. This is what «big data» people often claim today. Also, they claim that they can control for many contingencies

2) You can falsify theories (recall: *theory = series of logical steps moving from antecedents to conclusions*):
   - Theory: *younger people take more risk → more profits*
   - You find that younger people do not make more profits → if you controlled for all possible contingencies, your theory is wrong (falsified)
   - However, you still have no clue for why you find no correlation or even a negative one (because this is a correlation not a causal link)

The endogeneity problem

$$V = \beta_0 + \beta_1 \mathbf{X_1} + \beta_2 X_2 + \ldots + error$$

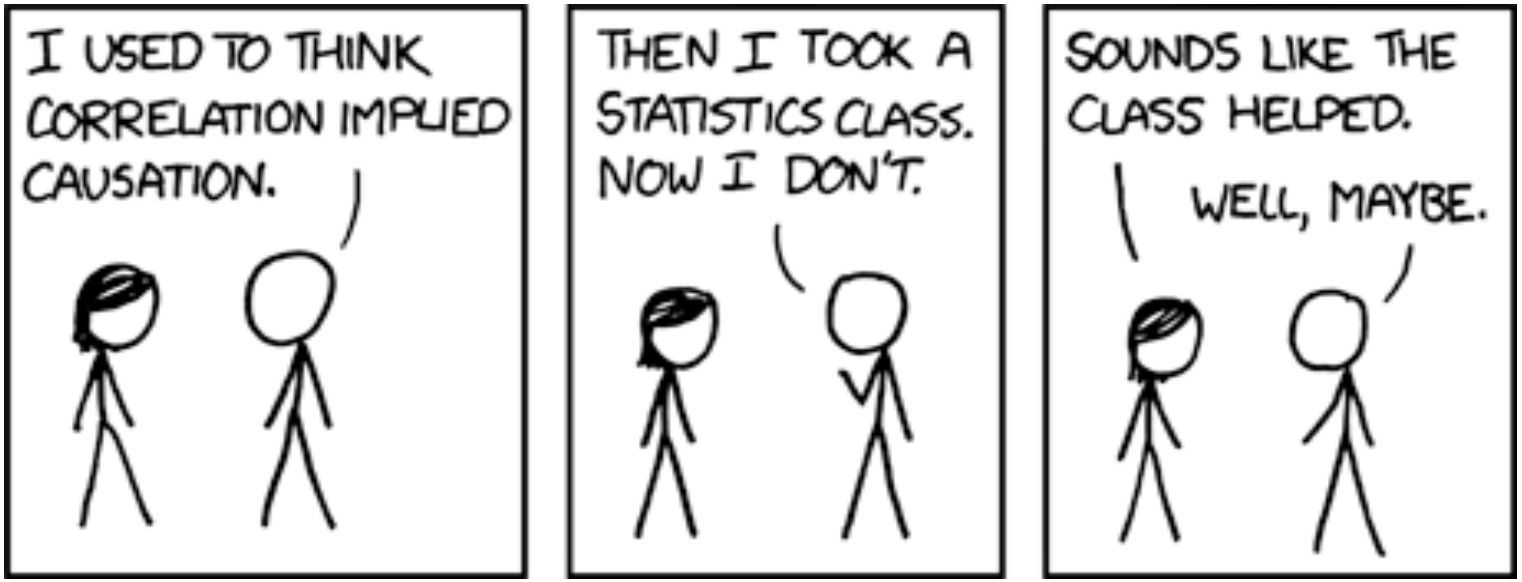
- Problem: $\mathbf{X_{1i}}$ correlated with $error_i$
- Why?
  - Endogeneity (reverse causality) … $\mathbf{X_1}$ causes **V** but **V** also causes $\mathbf{X_1}$ (simultaneity bias)
  - Omitted variables
  - (Measurement error)
- Consequence?
  - Estimated β_1 biased and inconsistent:
    - The expectd value of the estimator is not equal to the true parameter
    - As you increase sample size you do not converge to the true parameter
- It is one source of the **systematic error** s!!
- If correlation b/w ($\mathbf{X_1}$, *error*) is
  - **Positive** → OLS overestimates
  - **Negative** → OLS underestimates

Solutions to the endogeneity problem

- Matching (*this class*)
- Instrumental variables (*next class*)
- Difference-in-difference, experiments (*next class*)

Matching

- Find observations similar along many different characteistics (e.g. variables X) but the variable of interest
- Example: *value of patents*
  - You want to understand how man-month → value of patents
  - But man-month is *endogenous* (why?)
  - Match patents that are identical along many dimensions (e.g. same country, technological class, education, age of inventor, …) but man-months (the dimension of interest)
- Essentially you create groups and define dummy variables that take value 1 if you belong to that group and zero otherwise, and take expected values of each group
- The advantage is that you can weight each group
- Limitations: it is matching on observables


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

References

- Jeffrey Wooldridge, Introductory Econometrics: A Modern Approach, 4th ed. Thomson, chapters 1 and 2