Building a Robot Judge: Data Science for Decision-Making

2. Causal Inference Essentials

Learning Objectives

1. Implement and evaluate machine learning pipelines.
2. **Implement and evaluate causal inference designs.**
   - Evaluate (find problems in) causal claims.
   - Apply the standard research designs to produce causal evidence for a given empirical setting – or articulate why it is not possible.
   - Implement these research designs using Stata regressions.
3. Understand how (not) to use data science tools (ML and CI) to support expert decision-making.

Outline

What is causality?

We say that X causes Y if...

- were we to intervene and change the value of X without changing anything else...
- then Y would also change as a result.

What is causality?

We say that X causes Y if...

- were we to intervene and change the value of X without changing anything else...
- then Y would also change as a result.

Examples of causal questions:

- How does taking this course affect the grade in your master thesis?

What is causality?

We say that X causes Y if...

- were we to intervene and change the value of X without changing anything else...
- then Y would also change as a result.

Examples of causal questions:

- How does taking this course affect the grade in your master thesis?
- If SBB decreased ticket checks, how would that affect ticket sales?

What is causality?

We say that X causes Y if...

- were we to intervene and change the value of X without changing anything else...
- then Y would also change as a result.

Examples of causal questions:

- How does taking this course affect the grade in your master thesis?
- If SBB decreased ticket checks, how would that affect ticket sales?

Non-causal questions are also important:

- can I predict ticket sales next quarter based on all available variables this quarter?

Machine Learning vs Causal Inference

Machine Learning:

- in ML, we already know the truth from the dataset.
- we take the labels as given, we just want to predict them.
- we can always verify our model works using the test set.

Machine Learning vs Causal Inference

Machine Learning:

- in ML, we already know the truth from the dataset.
- we take the labels as given, we just want to predict them.
- we can always verify our model works using the test set.

Causal Inference:

- Causal inference is about what *we don't know yet*.
- how do we know if a new policy will work?
  - for example, wearing masks and disease spread.

Machine Learning vs Causal Inference

Machine Learning:

- in ML, we already know the truth from the dataset.
- we take the labels as given, we just want to predict them.
- we can always verify our model works using the test set.

Causal Inference:

- Causal inference is about what *we don't know yet*.
- how do we know if a new policy will work?
  - for example, wearing masks and disease spread.
- There isn't a machine learning dataset to train a model on.
  - we can't experimentally force people to wear a mask or not.
- How do we solve that?

Causal inference is needed to improve the world

Causal inference is needed to improve the world

Consider another important policy question:

- Should schools encourage/discourage students to use Chat GPT in their studies?

Causal inference is needed to improve the world

Consider another important policy question:

- Should schools encourage/discourage students to use Chat GPT in their studies?
  - no matter what we already know about learning psychology, there will be too much uncertainty about costs/benefits to answer this.
  - We need real-world evidence – ideally a randomized control trial (RCT).
- But what if an RCT is not available?

Causal inference is needed to improve the world

Consider another important policy question:

- Should schools encourage/discourage students to use Chat GPT in their studies?
  - no matter what we already know about learning psychology, there will be too much uncertainty about costs/benefits to answer this.
  - We need real-world evidence – ideally a randomized control trial (RCT).
- But what if an RCT is not available?
- Can use a natural experiment to produce causal estimates:
  - e.g., variation in grades, using differences in the timing of chat GPT adoption (differences-in-differences).

Causal inference is needed to improve the world

Consider another important policy question:

- Should schools encourage/discourage students to use Chat GPT in their studies?
  - no matter what we already know about learning psychology, there will be too much uncertainty about costs/benefits to answer this.
  - We need real-world evidence – ideally a randomized control trial (RCT).
- But what if an RCT is not available?
- Can use a natural experiment to produce causal estimates:
  - e.g., variation in grades, using differences in the timing of chat GPT adoption (differences-in-differences).
- Tech companies understand importance of causality with A/B testing
  - and also with hiring lots of economists, who specialize in causal analysis.
- Social scientists want to use causal inference to understand society and assist public policy.

Causal Statements

- A light switch being flipped turns on the lights.
- Getting a college degree increases career earnings.
- Higher cigarette taxes decrease smoking.
- Higher minimum wages decrease employment.
- Rain dances increase probability of rain

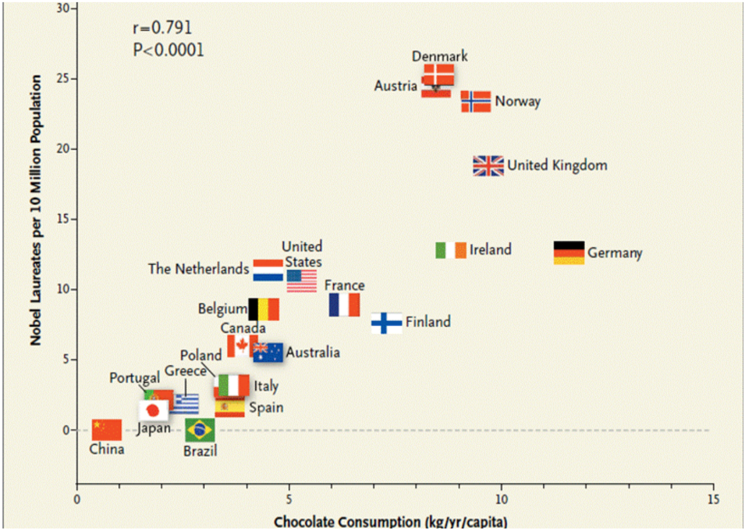
Causal Statements

- A light switch being flipped turns on the lights.
- Getting a college degree increases career earnings.
- Higher cigarette taxes decrease smoking.
- Higher minimum wages decrease employment.
- Rain dances increase probability of rain

Compare to:

- When people carry umbrellas, there is increased probability of rain
- When ice cream trucks are out, people wear shorts more often.
- Colds tend to clear up after taking cold medicine.

Correlation does not imply causation



More here: `http://www.tylervigen.com/spurious-correlations`

Important Notes

- "X causes Y":
  - does not mean that X is the only thing that causes Y
  - does not mean that all Y must be X
- For example, using a light switch causes the light to go on:
  - But not if the bulb is burned out (no Y, despite X), or if the light was already on (Y without X)
  - We would still say that using the switch causes the light.
  - The important thing is that X changes the probability that Y happens, not that it necessarily makes it happen for certain.

The Problem of Causal Inference

- If we have a correlation, how can we tell if it is causal or not?

The Problem of Causal Inference

- If we have a correlation, how can we tell if it is causal or not?
- "Does X cause Y?" can be rephrased as "If we manipulated X, would Y change as a result?"

The Problem of Causal Inference

- If we have a correlation, how can we tell if it is causal or not?
- "Does X cause Y?" can be rephrased as "If we manipulated X, would Y change as a result?"
- Example:
  - $X = 0$ or 1 for getting a vaccine or not
  - Y $=$ 0 or 1, for catching flu or not
  - Take one person – Angela – set her X to zero and check Y, then set her X to one and check Y.
  - If Y's are different, then X causes Y.

The Problem of Causal Inference

- If we have a correlation, how can we tell if it is causal or not?
- "Does X cause Y?" can be rephrased as "If we manipulated X, would Y change as a result?"
- Example:
  - $X = 0$ or 1 for getting a vaccine or not
  - Y $=$ 0 or 1, for catching flu or not
  - Take one person – Angela – set her X to zero and check Y, then set her X to one and check Y.
  - If Y's are different, then X causes Y.
- Problem:
  - Angela can't be in two places at once. either she got the vaccine or not.

Problem: Angela can't be in two places at once.

Problem: Angela can't be in two places at once.

- Solution 1:
  - compare Angela, who doesn't have the vaccine, to Beatrice, who does have the vaccine.

Problem: Angela can't be in two places at once.

- Solution 1:
  - compare Angela, who doesn't have the vaccine, to Beatrice, who does have the vaccine.
  - Problem:
    - Angela and Beatrice are different – there are lots of other factors/reasons contributing to the chance of catching the flu.
    - this is called "selection bias" or "confounding"

Problem: Angela can't be in two places at once.

- Solution 1:
  - compare Angela, who doesn't have the vaccine, to Beatrice, who does have the vaccine.
  - Problem:
    - Angela and Beatrice are different – there are lots of other factors/reasons contributing to the chance of catching the flu.
    - this is called "selection bias" or "confounding"
- Solution 2:
  - compare Angela's chances of getting the flu before and after getting the vaccine
    - (this is the longitudinal or panel data approach, focus of Week 4)

Problem: Angela can't be in two places at once.

- Solution 1:
  - compare Angela, who doesn't have the vaccine, to Beatrice, who does have the vaccine.
  - Problem:
    - Angela and Beatrice are different – there are lots of other factors/reasons contributing to the chance of catching the flu.
    - this is called "selection bias" or "confounding"
- Solution 2:
  - compare Angela's chances of getting the flu before and after getting the vaccine
    - (this is the longitudinal or panel data approach, focus of Week 4)
  - Problem (time-varying confounders):
    - other things are changing in Angela's life that affect her chances of catching the flu.

The Goal of Causal Inference

The Goal of Causal Inference

- The goal of causal inference is making as good a guess as possible as to what Y would have been if X had been different.
  - that "would have been" is called a **counterfactual**
- Put differently: We would like to get close to having two people that are exactly the same except that one has X=0 and one has X=1

The Goal of Causal Inference

- The goal of causal inference is making as good a guess as possible as to what Y would have been if X had been different.
  - that "would have been" is called a **counterfactual**
- Put differently: We would like to get close to having two people that are exactly the same except that one has X=0 and one has X=1
- In many scientific fields, you get causal variation with **experiments**.
  - If X is a randomly assigned **treatment** in a large sample, we know that the people in each **treatment group** are identical on average.
  - but in many contexts – especially in social science – experiments are not possible to do.

Resume Audit Study

Bertrand and Mullainathan (2004)

- 5,000 resumes sent to help-wanted ads in Boston and Chicago
- Randomized otherwise equivalent resumes to have African-American or White sounding names:
  - Emily Walsh or Greg Baker relative to Lakisha Washington or Jamal Jones

Resume Audit Study

Bertrand and Mullainathan (2004)

- 5,000 resumes sent to help-wanted ads in Boston and Chicago
- Randomized otherwise equivalent resumes to have African-American or White sounding names:
  - Emily Walsh or Greg Baker relative to Lakisha Washington or Jamal Jones
- Results:
  - 50% gap in callback rate for black-sounding names

Resume Audit Study

Bertrand and Mullainathan (2004)

- 5,000 resumes sent to help-wanted ads in Boston and Chicago
- Randomized otherwise equivalent resumes to have African-American or White sounding names:
  - Emily Walsh or Greg Baker relative to Lakisha Washington or Jamal Jones
- Results:
  - 50% gap in callback rate for black-sounding names
- Caveats:
  - “Lakisha” or “Jamal” might signal non-racial factors, e.g. socioeconomic status.
  - Fryer and Levitt (2004) find no long-term life outcome differences for people with more black-sounding names, adjusting for other background factors.

Limitations of Experiments (2 minutes)

- Last Names A-L:
  - think of a social science setting where an experiment would be impossible or unethical.
- Last Names M-Z:
  - think of a natural science setting where an experiment would be impossible or unethical.

Causality without experiments

Causality without experiments

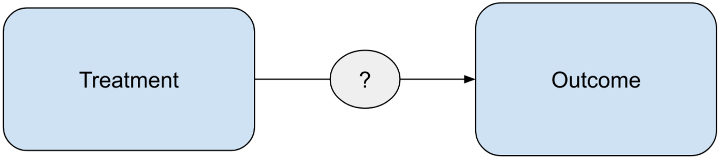
- The **research design**, **identification strategy**, or **empirical strategy** is the approach used with observational data (i.e. data not generated by a randomized trial) to approximate a randomized experiment.

Causality without experiments

- The **research design**, **identification strategy**, or **empirical strategy** is the approach used with observational data (i.e. data not generated by a randomized trial) to approximate a randomized experiment.
- Today:
  - Adjusting (controlling) for observed confounders
- Week 4:
  - Regression discontinuity design
  - Differences-in-differences
- Week 6:
  - Adjusting $\times$ machine learning: Double ML
- Week 7:
  - Instrumental variables

Outline

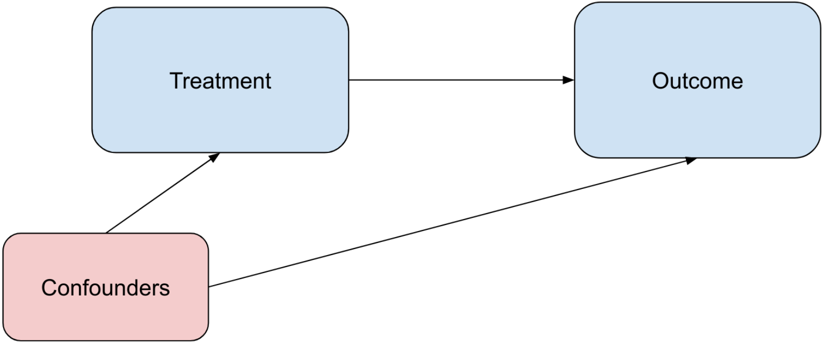
Causal Graphs



- We are interested in determining whether a significant correlation between "treatment" and "outcome" indicates a causal link.

Confounders

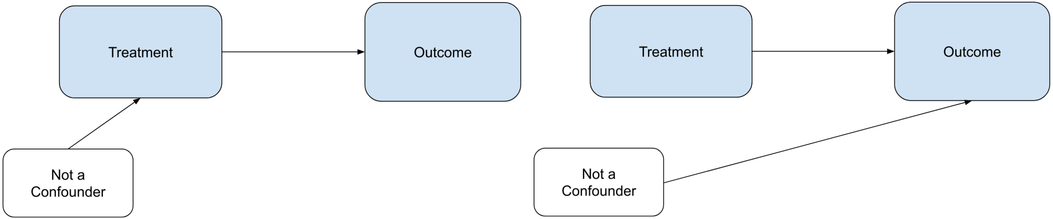
- Confounders affect both the treatment and the outcome:



- **In the presence of confounders, a correlation between the treatment and the outcome does not indicate a causal link.**
  - Example: eating ice cream causes heat stroke.

Not Confounders

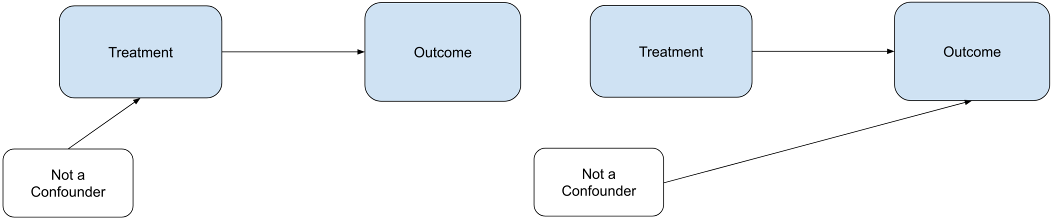
- Variables that affect just the treatment, or just the outcome, are not confounders.



- E.g.:
  - presence of ice cream truck affects probability of eating ice cream, but not probability of heat stroke.
  - old age increases probability of heat stroke, but not probability of eating ice cream

Not Confounders

- Variables that affect just the treatment, or just the outcome, are not confounders.



- E.g.:
  - presence of ice cream truck affects probability of eating ice cream, but not probability of heat stroke.
  - old age increases probability of heat stroke, but not probability of eating ice cream

Note: Randomized experiments knock out the arrow from all potential confounders to the treatment (which is randomly determined by construction).

Identification with Observed Confounders

- Another example: Effect of a person's income D on committing crimes Y.
  - what is a potential confounder A that might affect income D and crime choices Y?
  - That is, the estimated correlation between D and Y is **biased** by the presence of A.

Identification with Observed Confounders

- Another example: Effect of a person's income D on committing crimes Y.
  - what is a potential confounder A that might affect income D and crime choices Y?
  - That is, the estimated correlation between D and Y is **biased** by the presence of A.
- Assume that:
  - A is education, affecting both income and crime.
  - we can measure A.
  - A is the only confounder.

Identification with Observed Confounders

- Another example: Effect of a person's income D on committing crimes Y.
  - what is a potential confounder A that might affect income D and crime choices Y?
  - That is, the estimated correlation between D and Y is **biased** by the presence of A.
- Assume that:
  - A is education, affecting both income and crime.
  - we can measure A.
  - A is the only confounder.
- Under these assumptions, we can **identify** the effect of D on Y by netting out the components of D and Y that are driven by A.
  - this is called "adjusting for" or "controlling for" A

Adjusting (controlling) for observables

1. learn the function $\hat{D}(A)$, compute residual $\tilde{D} = D - \hat{D}$
2. learn the function $\hat{Y}(A)$, compute residual $\tilde{Y} = Y - \hat{Y}$
3. $\rightarrow$ the relationship between $\tilde{D}$ and $\tilde{Y}$ is causal.

Adjusting (controlling) for observables

1. learn the function $\hat{D}(A)$, compute residual $\tilde{D} = D - \hat{D}$
2. learn the function $\hat{Y}(A)$, compute residual $\tilde{Y} = Y - \hat{Y}$
3. $\rightarrow$ the relationship between $\tilde{D}$ and $\tilde{Y}$ is causal.

- In standard econometrics, one would assume linearity, e.g.

$$D(A) = \beta A, Y(A) = \gamma A$$

  - learn $\hat{\beta}$ and $\hat{\gamma}$ with linear regression (ordinary least squares)
  - then $\tilde{D} = D - \hat{\beta}A$ and $\tilde{Y} = Y - \hat{\gamma}A$

Adjusting (controlling) for observables

1. learn the function $\hat{D}(A)$, compute residual $\tilde{D} = D - \hat{D}$
2. learn the function $\hat{Y}(A)$, compute residual $\tilde{Y} = Y - \hat{Y}$
3. $\rightarrow$ the relationship between $\tilde{D}$ and $\tilde{Y}$ is causal.

- In standard econometrics, one would assume linearity, e.g.

$$D(A) = \beta A, Y(A) = \gamma A$$

  - learn $\hat{\beta}$ and $\hat{\gamma}$ with linear regression (ordinary least squares)
  - then $\tilde{D} = D - \hat{\beta}A$ and $\tilde{Y} = Y - \hat{\gamma}A$

- Notes:
  - A can be multivariate, e.g. $D(\boldsymbol{A}) = \boldsymbol{A}'\beta$
  - with newer approaches using machine learning for causal inference, can have arbitrary functional relationships for $D(\boldsymbol{A})$ and $Y(\boldsymbol{A})$.

Adjusting for observables: Intuition

- We are removing differences in Y and D that are predicted by A.
- Intuitively, we are comparing individuals as if they had the same value for A.
  - this is why we can say, "showing effect of D on Y, holding A constant."

When does confounding preclude causal inference?

1. observed confounders
   - not a problem; can control for them

When does confounding preclude causal inference?

1. observed confounders
   - not a problem; can control for them
2. unobserved variables that do not affect the outcome, or do not affect the treatment:
   - also not a problem

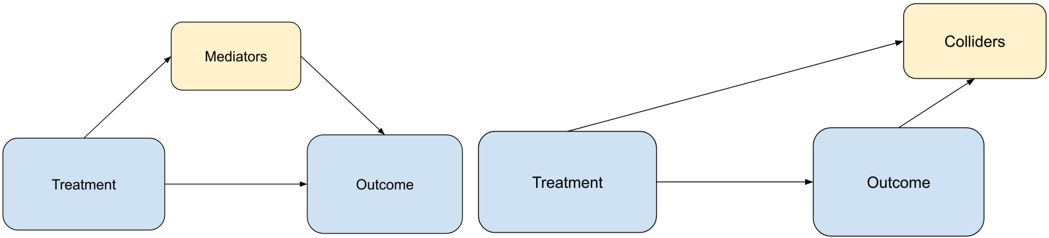
When does confounding preclude causal inference?

1. observed confounders
   - not a problem; can control for them
2. unobserved variables that do not affect the outcome, or do not affect the treatment:
   - also not a problem
3. unobserved variables that affect both the treatment and outcome.
   - **this is the problem – unobserved confounders or "omitted variable bias".**
   - in general, there is no way to know for sure whether all confounders are observed.

Why not control for everything? Colliders and Mediators

Why not control for everything? Colliders and Mediators

- **Mediators** are intermediate outcomes / mechanisms – affected by the treatment, but then they affect the outcome.
  - e.g., controlling for occupation when looking at the effect of education on income.
- **Colliders** are affected by both the treatment and the outcome.
  - e.g., controlling for marital status when looking at the effect of education on income.



- The presence of mediators and colliders does not produce omitted variable bias.
- Actually, **adjusting for them will induce bias**.
  - $\rightarrow$ have to be careful about what variables to adjust for.

Reverse Causation or Joint Causation

- **Reverse causation**: "Outcome" affects "Treatment".
  Joint causation: there is bidirectional causation.

- e.g., effect of policing on crime rates.
- In this case, cannot recover a causal relationship, even if adjusting for observables.
  - have to use RCTs or natural experiments (weeks 4, 7)

Activity on Confounders

Consider the **effect of education on income**:

- If last name starts with A-H:
  - what are likely **confounders** for the effect of education on income?
- If last name starts with I-P:
  - what are likely **mediators** for the effect of education on income?
- If last name starts with Q-Z:
  - what are likely **colliders** for the effect of education on income?

Outline

Outline

Linear Regression Models

- How does schooling affect income?
- Assume a linear model

$$Y_i = \alpha + \beta s_i + \epsilon_i$$

  - Y_i = the income of person i ("outcome variable")
  - s_i = his/her years of education ("treatment variable" or "explanatory variable")

Linear Regression Models

- How does schooling affect income?
- Assume a linear model
$$Y_i = \alpha + \beta s_i + \epsilon_i$$
  - Y_i = the income of person i ("outcome variable")
  - s_i = his/her years of education ("treatment variable" or "explanatory variable")
- α, the "intercept" or "constant", gives the expected income with no schooling ($s_i = 0$)
  - normalize $\alpha = 0$ going forward.

Linear Regression Models

- How does schooling affect income?
- Assume a linear model

$$Y_i = \alpha + \beta s_i + \epsilon_i$$

  - Y_i = the income of person i ("outcome variable")
  - s_i = his/her years of education ("treatment variable" or "explanatory variable")
- α, the "intercept" or "constant", gives the expected income with no schooling ($s_i = 0$)
  - normalize $\alpha = 0$ going forward.
- ϵ_i includes all other factors affecting income besides schooling, including randomness

Linear Regression Models

- How does schooling affect income?
- Assume a linear model
$$Y_i = \alpha + \beta s_i + \epsilon_i$$
    - Y_i = the income of person i ("outcome variable")
    - s_i = his/her years of education ("treatment variable" or "explanatory variable")
- α, the "intercept" or "constant", gives the expected income with no schooling ($s_i = 0$)
    - normalize $\alpha = 0$ going forward.
- ϵ_i includes all other factors affecting income besides schooling, including randomness
- β = the slope parameter summarizing how wages vary with schooling.

Ordinary Least Squares (OLS) Estimator

$$Y_i = \alpha + \beta s_i + \epsilon_i$$

Ordinary Least Squares (OLS) Estimator

$$Y_i = \alpha + \beta s_i + \epsilon_i$$

- Assume that Y_i and s_i are de-meaned. Then the OLS estimator is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n} s_i Y_i}{\sum_{i=1}^{n} s_i^2} = \frac{\mathrm{Cov}[Y_i, s_i]}{\mathrm{Var}[s_i]}$$

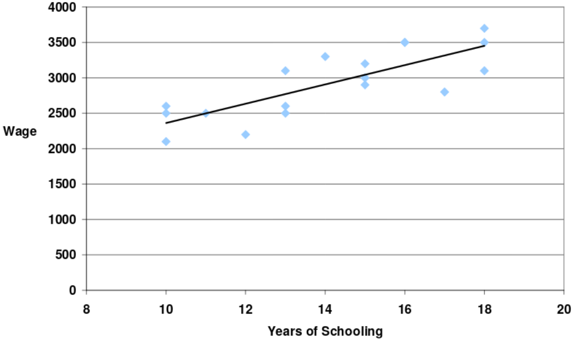
```
import statsmodels.formula.api as smf
ols = smf.ols(formula='price ~ CRIM', data=df).fit()
ols.summary()
```

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.203
Model:	OLS	**Adj. R-squared:**	0.201
Method:	Least Squares	**F-statistic:**	124.0
Date:	Sat, 02 Oct 2021	**Prob (F-statistic):**	8.11e-26
Time:	17:17:08	**Log-Likelihood:**	-1649.9
No. Observations:	490	**AIC:**	3304.
Df Residuals:	488	**BIC:**	3312.
Df Model:	1		
Covariance Type:	nonrobust		

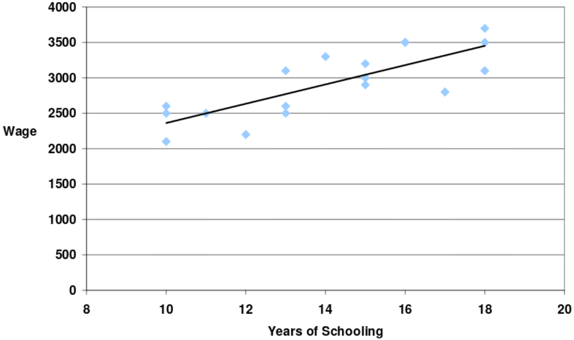
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	23.1147	0.344	67.143	0.000	22.438	23.791
CRIM	-0.4059	0.036	-11.135	0.000	-0.478	-0.334

Interpreting OLS Coefficients



- $\hat{\beta} = \frac{\partial Y}{\partial s}$, the predicted change in the outcome variable Y in response to increasing the treatment variable s by 1.
  - In this example, the average increase in income for taking one more year of school.

Interpreting OLS Coefficients



- $\hat{\beta} = \frac{\partial Y}{\partial s}$, the predicted change in the outcome variable Y in response to increasing the treatment variable s by 1.
  - In this example, the average increase in income for taking one more year of school.
- Using the estimated constant $\hat{\alpha}$ and estimated slope coefficient $\hat{\beta}$, we obtain a predicted income $\hat{Y}$ for any level of schooling s as

$$\hat{Y}(s) = \hat{\alpha} + \hat{\beta} s$$

Multivariate OLS

- OLS models can be generalized to multiple variables:

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$$

Multivariate OLS

- OLS models can be generalized to multiple variables:

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$$

- Or

$$y_i = \alpha + \boldsymbol{x}_i' \beta + \epsilon_i$$

where $\boldsymbol{x}_i$ is a vector of n_x explanatory variables.

Multivariate OLS

- OLS models can be generalized to multiple variables:

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$$

- Or

$$y_i = \alpha + \boldsymbol{x}_i' \beta + \epsilon_i$$

  where $\boldsymbol{x}_i$ is a vector of n_x explanatory variables.
- For n_D observations and n_x explanatory variables, with $n_x < n_D$
  - Let Y be the $n_D \times 1$ vector for the outcome variable.
  - Let $\boldsymbol{X}$ be the $n_D \times n_x$ matrix of explanatory variables
    - none of the variables can by collinear (that is, a linear transformation of another variable).

Multivariate OLS

- OLS models can be generalized to multiple variables:

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$$

- Or

$$y_i = \alpha + \boldsymbol{x}_i' \beta + \epsilon_i$$

  where $\boldsymbol{x}_i$ is a vector of n_x explanatory variables.
- For n_D observations and n_x explanatory variables, with $n_x < n_D$
  - Let Y be the $n_D \times 1$ vector for the outcome variable.
  - Let $\boldsymbol{X}$ be the $n_D \times n_x$ matrix of explanatory variables
    - none of the variables can by collinear (that is, a linear transformation of another variable).
- The $n_x \times 1$ vector of OLS coefficients (one for reach explanatory variable) is

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Y$$

Outline

- The **OLS exogeneity assumption** is $\text{Cov}[s_i, \epsilon_i] = 0$
  - (treatment is uncorrelated with error; equivalent to no confounders).

- The **OLS exogeneity assumption** is $\text{Cov}[s_i, \epsilon_i] = 0$
  - (treatment is uncorrelated with error; equivalent to no confounders).
- We have

$$
\begin{aligned}
\hat{\beta} = \frac{\sum_{i=1}^{n} s_i Y_i}{\sum_{i=1}^{n} s_i^2} &= \frac{\sum_{i=1}^{n} s_i(\beta s_i + \epsilon_i)}{\sum_{i=1}^{n} s_i^2} \\
&= \left(\frac{\sum_{i=1}^{n} s_i^2}{\sum_{i=1}^{n} s_i^2}\right)\beta + \frac{\sum_{i=1}^{n} s_i(\epsilon_i)}{\sum_{i=1}^{n} s_i^2} \\
&= \beta + \frac{\sum_{i=1}^{n} s_i \epsilon_i}{\sum_{i=1}^{n} s_i^2}
\end{aligned}
$$

- The **OLS exogeneity assumption** is $\text{Cov}[s_i, \epsilon_i] = 0$
  - (treatment is uncorrelated with error; equivalent to no confounders).
- We have

$$
\begin{aligned}
\hat{\beta} = \frac{\sum_{i=1}^n s_i Y_i}{\sum_{i=1}^n s_i^2} &= \frac{\sum_{i=1}^n s_i(\beta s_i + \epsilon_i)}{\sum_{i=1}^n s_i^2} \\
&= (\frac{\sum_{i=1}^n s_i^2}{\sum_{i=1}^n s_i^2})\beta + \frac{\sum_{i=1}^n s_i(\epsilon_i)}{\sum_{i=1}^n s_i^2} \\
&= \beta + \frac{\sum_{i=1}^n s_i \epsilon_i}{\sum_{i=1}^n s_i^2}
\end{aligned}
$$

- Taking expectations:

$$
\begin{aligned}
\mathbb{E}[\hat{\beta}] &= \beta + \mathbb{E}[\frac{\sum_{i=1}^n s_i \epsilon_i}{\sum_{i=1}^n s_i^2}] \\
&= \beta + \frac{\text{Cov}[s_i, \epsilon_i]}{\text{Var}[s_i]} \\
&= \beta
\end{aligned}
$$

Endogeneity

- When conditional independence is not satisfied, we say that "s is endogenous":
  - That is, an explanatory variable s_i is said to be **endogenous** if it is correlated with unobserved factors (confounders) that are also correlated with the outcome variable.

Endogeneity

- When conditional independence is not satisfied, we say that "s is endogenous":
  - That is, an explanatory variable s_i is said to be **endogenous** if it is correlated with unobserved factors (confounders) that are also correlated with the outcome variable.
- Since the error term ϵ_i includes all unobserved factors affecting the outcome, we can define **endogeneity** as correlation between an explanatory variable and the error term:

$$\text{Cov}[s_i, \epsilon_i] \neq 0$$

Formalizing omitted variable bias

- Assume that the "true" model is
$$Y_i = \beta s_i + \gamma a_i + \eta_i \tag{1}$$
where η_i is exogenous by assumption ($\text{Cov}[s_i, \eta_i] = 0$), but we cannot measure ability a_i.

Formalizing omitted variable bias

- Assume that the "true" model is
$$Y_i = \beta s_i + \gamma a_i + \eta_i \tag{1}$$
where η_i is exogenous by assumption ($\text{Cov}[s_i, \eta_i] = 0$), but we cannot measure ability a_i.
- Now we have

$$\hat{\beta} = \frac{\sum_{i=1}^n s_i Y_i}{\sum_{i=1}^n s_i^2} = \frac{\sum_{i=1}^n s_i(\beta s_i + \gamma a_i + \eta_i)}{\sum_{i=1}^n s_i^2}$$
$$= \beta + \frac{\sum_{i=1}^n s_i(\gamma a_i)}{\sum_{i=1}^n s_i^2} + \frac{\sum_{i=1}^n s_i \eta_i}{\sum_{i=1}^n s_i^2}$$

Formalizing omitted variable bias

- Assume that the "true" model is
$$Y_i = \beta s_i + \gamma a_i + \eta_i \tag{1}$$
where η_i is exogenous by assumption ($\text{Cov}[s_i, \eta_i] = 0$), but we cannot measure ability a_i.
- Now we have
$$\hat{\beta} = \frac{\sum_{i=1}^n s_i Y_i}{\sum_{i=1}^n s_i^2} = \frac{\sum_{i=1}^n s_i(\beta s_i + \gamma a_i + \eta_i)}{\sum_{i=1}^n s_i^2}$$
$$= \beta + \frac{\sum_{i=1}^n s_i(\gamma a_i)}{\sum_{i=1}^n s_i^2} + \frac{\sum_{i=1}^n s_i \eta_i}{\sum_{i=1}^n s_i^2}$$
- Taking expectations gives
$$\mathbb{E}[\hat{\beta}] = \beta + \underbrace{\gamma \frac{\text{Cov}[s_i, a_i]}{\text{Var}[s_i]}}_{\text{Omitted variable bias}} + \underbrace{\frac{\text{Cov}[s_i, \eta_i]}{\text{Var}[s_i]}}_{=0 \text{ by assumption}}$$

$\rightarrow$ if ability is correlated with schooling ($\text{Cov}[s_i, a_i] \neq 0$), $\hat{\beta}$ is a biased estimate for β.

Understanding omitted variable bias

$$\mathbb{E}[\hat{\beta}] = \beta + \underbrace{\gamma \frac{\text{Cov}[s, a]}{\text{Var}[s]}}_{\text{Omitted variable bias}}$$

		Correlation of omitted variable with explanatory variable	
		$\text{Cov}[s, a] > 0$	$\text{Cov}[s, a] < 0$
Correlation of omitted variable with outcome	$\gamma > 0$	$\hat{\beta} > \beta$	$\hat{\beta} < \beta$
	$\gamma < 0$	$\hat{\beta} < \beta$	$\hat{\beta} > \beta$

- Check for understanding:
  - which of the four cells (top left, top right, bottom left, bottom right) are we in, for the case where y = income, s = education, and a = ability.

Adjusting for confounders with multivariate regression

$$Y_i = \beta s_i + \gamma a_i + \eta_i$$

- What if we can observe both schooling s_i and ability a_i (e.g., from an IQ test)?
- Then we can adjust for ability and obtain an unbiased causal estimate for β, simply by adding a_i to the OLS regression.
- e.g.:

```
ols = smf.ols(formula="income ~ educ + test_score", data=df).fit()
```

Outline

Statistical Significance

- The value for β provides a prediction for the effect of the explanatory variable on the outcome.
  - But if this prediction is very noisy, then it might not be useful for policy analysis.

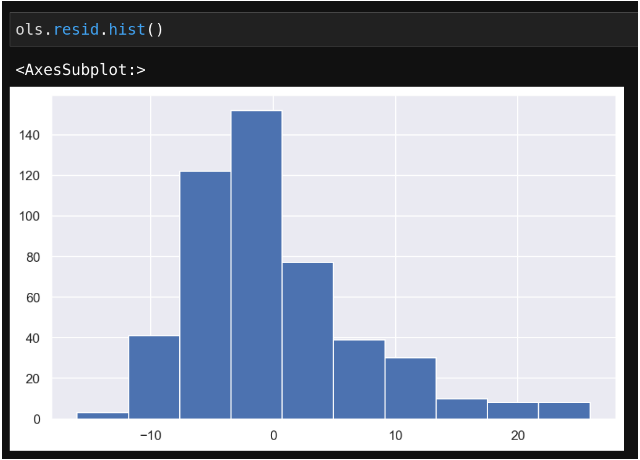
Statistical Significance

- The value for β provides a prediction for the effect of the explanatory variable on the outcome.
  - But if this prediction is very noisy, then it might not be useful for policy analysis.
- To do causal *inference*, we have to determine whether the effect is statistically significant.
  - This is generally achieved by computing a **standard error** for each coefficient, and then using the standard error to compute **confidence intervals** and a ***p*-value** for the hypothesis that $\beta \neq 0$.

Residuals

- The **residuals** or **errors** from an OLS regression are defined as

$$\begin{aligned}\tilde{\epsilon}_i &= Y_i - \hat{Y}_i \\ &= Y_i - \hat{\alpha} - \hat{\beta} s_i\end{aligned}$$



Standard Errors

- The **standard error** (SE) for the OLS estimate $\hat{\beta}$ is

$$\hat{\sigma}_\beta = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\tilde{\epsilon}_i^2},$$

  the square root of the average of the squared residuals.

- SE provides information about the precision of the estimate: a lower standard error is a more precise estimate.
- On regression tables, usually reported in parentheses beneath the point estimate.

```
ols.summary()
```

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.203
Model:	OLS	**Adj. R-squared:**	0.201
Method:	Least Squares	**F-statistic:**	124.0
Date:	Sat, 02 Oct 2021	**Prob (F-statistic):**	8.11e-26
Time:	17:17:08	**Log-Likelihood:**	-1649.9
No. Observations:	490	**AIC:**	3304.
Df Residuals:	488	**BIC:**	3312.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	23.1147	0.344	67.143	0.000	22.438	23.791
CRIM	-0.4059	0.036	-11.135	0.000	-0.478	-0.334

Standard Errors

- The **standard error** (SE) for the OLS estimate $\hat{\beta}$ is

$$\hat{\sigma}_\beta = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\tilde{\epsilon}_i^2},$$

  the square root of the average of the squared residuals.

- SE provides information about the precision of the estimate: a lower standard error is a more precise estimate.
- On regression tables, usually reported in parentheses beneath the point estimate.

```
ols.summary()
```

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.203
Model:	OLS	**Adj. R-squared:**	0.201
Method:	Least Squares	**F-statistic:**	124.0
Date:	Sat, 02 Oct 2021	**Prob (F-statistic):**	8.11e-26
Time:	17:17:08	**Log-Likelihood:**	-1649.9
No. Observations:	490	**AIC:**	3304.
Df Residuals:	488	**BIC:**	3312.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	23.1147	0.344	67.143	0.000	22.438	23.791
CRIM	-0.4059	0.036	-11.135	0.000	-0.478	-0.334

- In multivariate OLS with predictor matrix $\boldsymbol{X}$, there is a separate standard error for the coefficient on each predictor, given by diagonal entries of the $n_x \times n_x$ matrix

$$\hat{\sigma}_\beta\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}}$$

t-statistics, p-values, and confidence intervals

- A rule of thumb for statistical significance is to compute the ***t*-statistic**:

$$t = \frac{\hat{\beta}}{\hat{\sigma}_\beta}$$

  - $t > 2 \rightarrow$ statistically significant positive effect, $t < 2 \rightarrow$ statistically significant negative effect
- A high t (in absolute value) is associated with a small ***p*-value** (e.g., $t = \pm 1.96 \rightarrow p = .05$).
  - Small p-values are often indicated on regression tables with stars to indicate statistical significance.
- **95% confidence intervals** indicate (roughly) that the coefficient is 95% likely to reside within that interval.

```
ols.summary()
```

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.203
Model:	OLS	**Adj. R-squared:**	0.201
Method:	Least Squares	**F-statistic:**	124.0
Date:	Sat, 02 Oct 2021	**Prob (F-statistic):**	8.11e-26
Time:	17:17:08	**Log-Likelihood:**	-1649.9
No. Observations:	490	**AIC:**	3304.
Df Residuals:	488	**BIC:**	3312.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	23.1147	0.344	67.143	0.000	22.438	23.791
CRIM	-0.4059	0.036	-11.135	0.000	-0.478	-0.334