Building a Robot Judge: Data Science for Decision-Making

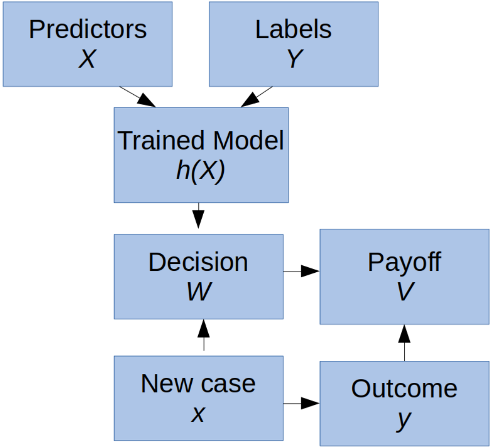
3. Causal Inference Essentials

Instructions before we begin:
(1) Turn on video and set audio to mute
(2) In Participants panel, set zoom name to "Full Name, School / Degree"
(ex: "Leon Smith, ETH Data Science Msc")

Learning Objectives

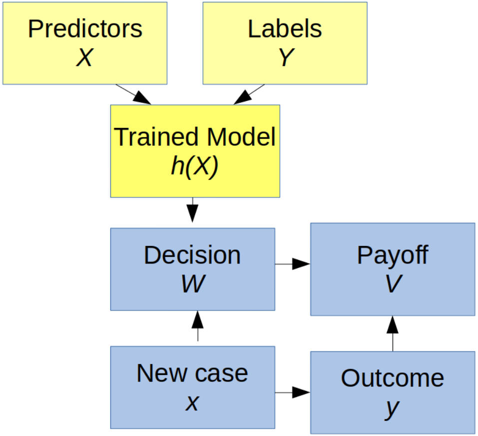
1. Implement and evaluate machine learning pipelines.
2. **Implement and evaluate causal inference designs.**
    - Evaluate (find problems in) causal claims.
    - Apply the standard research designs to produce causal evidence for a given empirical setting – or articulate why it is not possible.
  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

Decision-Making Schema



- A decision-maker observes facts x and makes decision w, which produces payoff $V = u(y, w)$.
- Decision-maker has access to a history of cases with facts X and labels Y, can learn a machine prediction $\hat{y} = h(x)$.

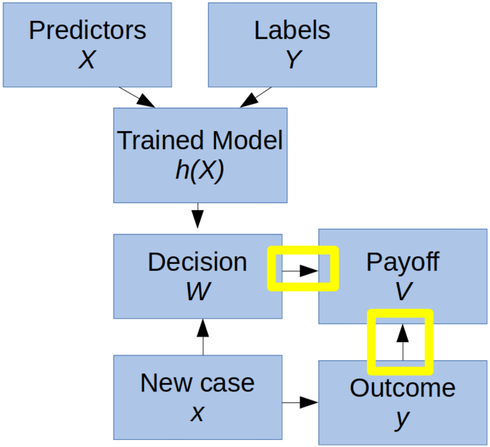
Last Week (Machine Learning)



- A decision-maker observes facts x and makes decision w, which produces payoff $V = u(y, w)$.
- Decision-maker has access to a history of cases with facts X and labels Y, can learn a machine prediction $\hat{y} = h(x)$.

- Good decision-making requires accurate predictions for a relevant outcome (e.g. recidivism) based on observables. We can learn those predictions from data.

This Week (Causal Inference)



- A decision-maker observes facts x and makes decision w, which produces payoff $V = u(y, w)$.
- Decision-maker has access to a history of cases with facts X and labels Y, can learn a machine prediction $\hat{y} = h(x)$.
- In addition to having a good prediction $h(\cdot)$, decision-maker wants to know $u(y, w)$.

- Good decision-making requires accurate *counterfactual* predictions for how changes in decisions impact the payoff-relevant outcome.

Counterfactual predictions $\leftrightarrow$ Causal parameters

- Let's say the payoff function $v = u(y, w; \beta)$ has learnable causal parameters β.
  - e.g., the effect of prison sentence w on crime rates v, given recidivism y.
- How to learn β?
  - what we call *empirical* or *econometric* analysis.
  - requires causal inference.
  - this is the focus in applied economics research

Causal effects

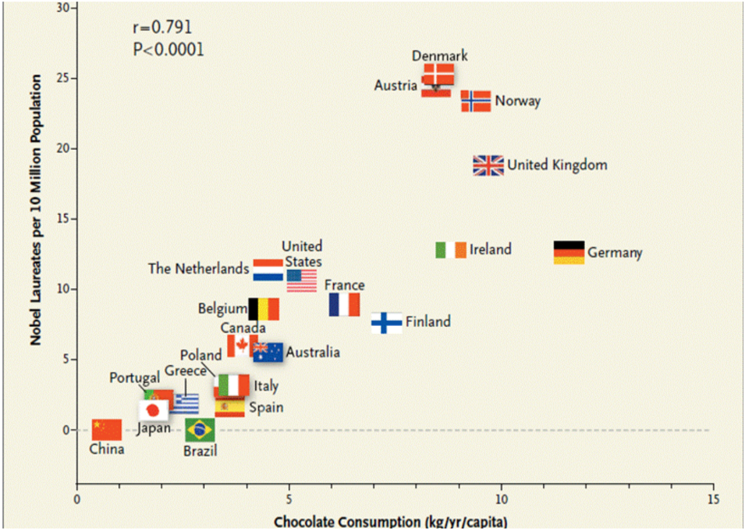
- While economists are often motivated by **why** questions, in research we proceed to address **what if** questions.
- Examples:
  - How does taking this course affect the grade in your master thesis?
    - This is **different** from the **predictive** question: "What is the grade that students taking this course will obtain with their master thesis?"
  - If Zurich imposed a special tax on Uber drivers, how would that effect the supply of Uber rides?
  - etc.

Zoom Chat Activity (2 minutes)

Re-write this "prediction" question as a "what if" question – chat to me privately on Zoom.:

- What is the probability that Ludwig will commit murder if he faces the death penalty?

Correlation does not imply causation



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

Basics

- Represent a **treatment** for row i as a binary random variable $D_i = 0, 1$.
  - $D_i = 1$ if treated, $D_i = 0$ if control
  - e.g., receive a medicine or not (or go to prison or not)
- Define an **outcome** V_i for individual i.
  - e.g., life expectancy.
- Define "**potential outcomes**" (counterfactuals) as:

$$V_i(D_i) = \begin{cases} V_{0i} & \text{if } D_i = 0 \\ V_{1i} & \text{if } D_i = 1 \end{cases}$$

- The **causal effect** of the medicine (treatment) for individal i is $V_{1i} - V_{0i}$.
  - the difference in the outcome between treatment and control.
- **Problem**: For i, we can observe V_{1i} (individual takes medicine) or V_{0i} (no medicine), **but not both.**

Illustration

- Let's take some imaginary data where we can time travel and observe participants Leo and Mia both with/without the medicine:

		Leo	Mia
V_{0i}	life expectancy without medicine	3	5
V_{1i}	life expectancy with medicine	4	5

Illustration: Treatment Effects

- Let's take some imaginary data where we can time travel and observe participants Leo and Mia both with/without the medicine:

		Leo	Mia
V_{0i}	life expectancy without medicine	3	5
V_{1i}	life expectancy with medicine	4	5
$V_{1i} - V_{0i}$	**treatment effect for** i	**1**	**0**

- In this imaginary data, the medicine would work for Leo, but not for Mia.

Illustration: Selection Bias

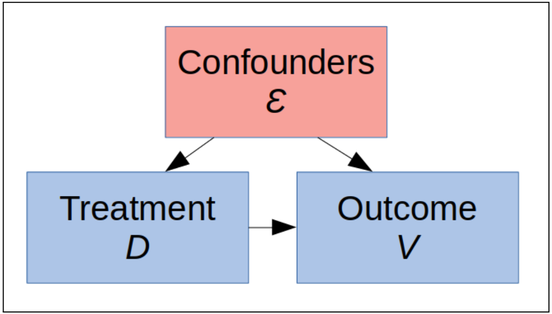
Let's say that in reality, Leo gets the medicine ($D_{\text{Leo}} = 1$) and Mia does not ($D_{\text{Mia}} = 0$):

		Leo	Mia
V_{0i}	life expectancy without medicine	3	5
V_{1i}	life expectancy with medicine	4	5
$V_{1i} - V_{0i}$	treatment effect	1	0
D_i	actual treatment assignment	1	0
V_i	**actual health outcome**	**4**	**5**

- Note that $V_{\text{Leo}} < V_{\text{Mia}}$:
  - based on these outcomes, one would be led to believe that the medicine actually harms the patient!
  - This is **selection bias** or **confounding**.

Selection Bias due to Confounders

- Leo has a pre-existing tendency in life expectancy, that is correlated with treatment assignment.
  - this tendency is a **confounder** or **omitted variable**
  - if we could observe this tendency, we could control or adjust for it.
  - but if unobserved, resulting analysis will be biased.



- $\rightarrow$ Observational studies of medicines don't work well, because relatively sick individuals will be more likely to take the medicine.

Formalizing Selection Bias or Confounding

The difference in observed outcomes between treatment group and control group is:

$$\underbrace{\mathbb{E}[V_{1i}|D_i = 1]}_{\text{avg outcome for treatment}} - \underbrace{\mathbb{E}[V_{0i}|D_i = 0]}_{\text{avg outcome for control}}$$

subtract $\mathbb{E}[V_{0i}|D_i = 1]$ (*not observed*) from first term, add to second term:

$$\rightarrow \underbrace{\mathbb{E}[V_{1i}|D_i = 1] - \mathbb{E}[V_{0i}|D_i = 1]}_{\text{Treatment Effect on Treated}} + \underbrace{\mathbb{E}[V_{0i}|D_i = 1] - \mathbb{E}[V_{0i}|D_i = 0]}_{\text{"Selection Bias"}}$$

- When does the difference in observed outcomes capture the **average treatment effect** (on the treated)?
  - only if there is no selection bias:

$$\mathbb{E}[V_{0i}|D_i = 1] = \mathbb{E}[V_{0i}|D_i = 0]$$

(equivalent to saying their are no confounders).

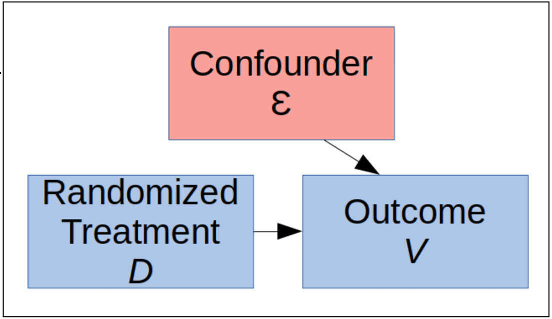
Questions: Answer by Private Zoom Chat (2 minutes)

- If last name starts with A-M:
  - what are likely confounders for the effect of education on income?
- If last name starts with N-Z:
  - Why is selection bias not a problem in a lab experiment?

Random Assignment

Random assignment $\rightarrow$ D_i independent of potential outcomes:

$$\mathbb{E}[V_{1i}|D_i = 1] = \mathbb{E}[V_{1i}|D_i = 0] = E[V_{1i}]$$
$$\mathbb{E}[V_{0i}|D_i = 1] = \mathbb{E}[V_{0i}|D_i = 0] = E[V_{0i}]$$
$$\rightarrow \text{selection bias} = 0.$$



Therefore, the difference in observed outcomes

$$\mathbb{E}[V_{1i}|D_i = 1] - \mathbb{E}[V_{0i}|D_i = 0]$$

captures the average treatment effect:

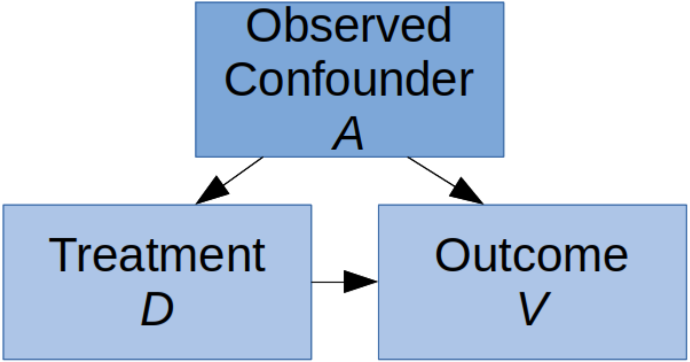
$$\mathbb{E}[V_{1i} - V_{0i}|D_i = 1] = \mathbb{E}[V_{1i} - V_{0i}|D_i = 0] = \mathbb{E}[V_{1i} - V_{0i}]$$

and provides a **counterfactual prediction** for effect of taking treatment.

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
  - Differences-in-differences
- Week 5:
  - Adjusting $\times$ machine learning: Double ML
  - Diffs-in-diffs $\times$ machine learning: Synthetic control
- Week 7:
  - (Deep) Instrumental variables
  - Regression discontinuity design

Adjusting (controlling) for observables



- What if the treated group and the non-treated group differ only by a set of observable characteristics?
- This is the case of observed confounders.
  - also called "selection on observables" or "conditional independence"
  - justifies causal interpretation of regression estimates

Example

- Effect of going to school $D_i \in \{0,1\}$ on lifetime income $V_i \geq 0$.
  - Say that we observe an IQ test, A_i, for each individual.
- The difference in outcomes, conditional on characteristics, is

$$\mathbb{E}[V_{1i}|A_i, D_i = 1] - \mathbb{E}[V_{0i}|A_i, D_i = 0]$$

$$= \underbrace{\mathbb{E}[V_{1i}|A_i, D_i = 1] - \mathbb{E}[V_{0i}|A_i, D_i = 1]}_{\text{Treatment Effect}} + \underbrace{\mathbb{E}[V_{0i}|A_i, D_i = 1] - \mathbb{E}[V_{0i}|A_i, D_i = 0]}_{\text{Selection Bias}}$$

- Conditional Independence holds when

$$\mathbb{E}[V_{0i}|A_i, D_i = 1] = \mathbb{E}[V_{0i}|A_i, D_i = 0]$$

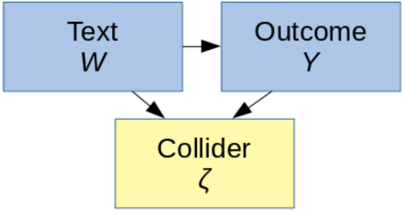
that is, selection bias is zero conditional on observables.

When is confounding relevant?

- Four possible types of potential confounders:
  1. observed confounders
     - not a problem; just include in the regression
  2. unobserved variables that are not correlated with the outcome:
     - also not a problem.
  3. unobserved variables that are not correlated with treatment
     - also not a problem
  4. unobserved variables correlated with about treatment and outcome.
     - **this is the problem.**
     - often way to know whether all confounders are observed.

Is adding controls always a good idea?

- The short answer is no.
  - With random assignment or a good identification strategy (natural experiment), you don't need controls.



- "Bad controls" (colliders or mediators) are variables that are jointly determined along with the outcome.
  - for example, controlling for occupation in the effect of education on income: education affects both occupation and income.
  - Adjusting for these variables could add bias.

Poll 3.1 – Colliders (3 minutes)

- We want to run a regression for the effect of number of policemen per capita on number of prisoners per capita.
- Which of the following would *not* be a collider in this regression?

Introduction to Regression

- How does schooling affect income?
- Assume a linear model
$$V_i = \alpha + \beta s_i + \epsilon_i$$
- V_i is wages as a function of s_i, years of education
- α, the "intercept" or "constant", gives the expected income with no schooling ($s_i = 0$)
  - assume $\alpha = 0$ going forward.
- ϵ_i includes all other factors affecting income besides schooling, including randomness
- $\beta =$ the slope parameter summarizing how wages vary with schooling.

OLS Estimator

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

- The Ordinary Least Squares (OLS) Estimator is the workhorse of applied microeconometrics.
- Assume that s_i is de-meaned. Then the OLS estimator is given by

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

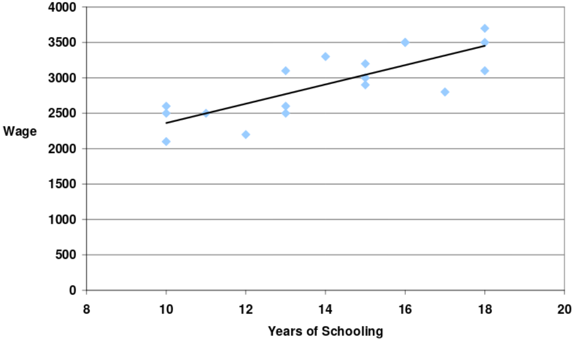
In stata:

```
reg V s

      Source |       SS           df       MS      Number of obs   =        51
-------------+----------------------------------   F(1, 49)        =     13.61
       Model |  48708.3001         1  48708.3001   Prob > F        =    0.0006
    Residual |   175306.21        49  3577.67775   R-squared       =    0.2174
-------------+----------------------------------   Adj R-squared   =    0.2015
       Total |   224014.51        50  4480.2902    Root MSE        =    59.814

------------------------------------------------------------------------------
           V |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |  -.0222756   .0060371    -3.69   0.001    -.0344077   -.0101436
       _cons |   1060.732    32.7009    32.44   0.000     995.0175    1126.447
------------------------------------------------------------------------------
```

Interpreting OLS Coefficients



- $\hat{\beta}$ gives the predicted change in the outcome variable V in response to increasing the explanatory variable s by 1.
  - In this case, 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

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

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

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

- 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.
- 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

$$V_i = \beta s_i + \gamma a_i + \eta_i \tag{1}$$

  where η_i is random (exogenous), but we cannot measure ability a_i.
- Now we have

$$\begin{aligned}\hat{\beta} = \frac{\sum_{i=1}^n s_i V_i}{\sum_{i=1}^n s_i^2} &= \frac{\sum_{i=1}^n s_i(\beta s_i + \gamma a_i + \eta_i)}{\sum_{i=1}^n s_i^2} \\ &= \beta + \frac{\sum_{i=1}^n s_i(\gamma a_i)}{\sum_{i=1}^n s_i^2} + \frac{\sum_{i=1}^n s_i \eta_i}{\sum_{i=1}^n s_i^2}\end{aligned}$$

- 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, $\hat{\beta}$ is a biased estimate for β.

Understanding omitted variable bias

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

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

- **Poll** 3.2: How does the example of ability/schooling/income fit in this table?

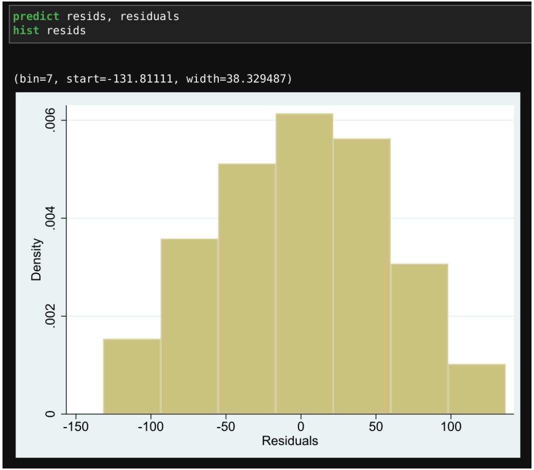
Statistical Significance

- 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 haveo 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 a ***p*-value** for the hypothesis that $\beta \neq 0$.

Residuals

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

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



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.

```
: reg V s

      Source |       SS           df       MS      Number of obs   =        51
-------------+----------------------------------   F(1, 49)        =     13.61
       Model |  48708.3001         1  48708.3001   Prob > F        =    0.0006
    Residual |   175306.21        49  3577.67775   R-squared       =    0.2174
-------------+----------------------------------   Adj R-squared   =    0.2015
       Total |   224014.51        50  4480.2902    Root MSE        =    59.814

------------------------------------------------------------------------------
           V |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |  -.0222756   .0060371    -3.69   0.001    -.0344077   -.0101436
       _cons |   1060.732    32.7009    32.44   0.000     995.0175    1126.447
------------------------------------------------------------------------------
```

- 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 right beneath the point estimate.

t-statistics and p-values

- 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
  - $t \in [-2, 2] \rightarrow$ no effect
- A high t (in absolute value) is associated with a small p-value (e.g., $t = 1.96 \rightarrow p = .05$).

```
reg V s

      Source |       SS           df       MS      Number of obs   =        51
-------------+----------------------------------   F(1, 49)        =     13.61
       Model |  48708.3001         1  48708.3001   Prob > F        =    0.0006
    Residual |   175306.21        49  3577.67775   R-squared       =    0.2174
-------------+----------------------------------   Adj R-squared   =    0.2015
       Total |   224014.51        50   4480.2902   Root MSE        =    59.814

------------------------------------------------------------------------------
           V |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |  -.0222756   .0060371    -3.69   0.001    -.0344077   -.0101436
       _cons |   1060.732    32.7009    32.44   0.000     995.0175    1126.447
------------------------------------------------------------------------------
```

Small p-values are often indicated on regression tables with stars to indicate statistical significance.

Multivariate Regression

- Setup: n_D observations with n_x explanatory variables.
  - Let V be the $n_D \times 1$ vector for the outcome variable (also called dependent variable).
  - Let X be the $n_D \times n_x$ matrix of explanatory variables (also called independent variables or predictors)
- The $n_x \times 1$ vector of OLS coefficients (one for reach explanatory variable) is

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

with standard errors given by the diagonal entries of

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

```
reg V s1 s2
```

	(1)	(2)	(3)	(4)
		Log Positive Cites		
Judge Age (Years)	-0.00797**	-0.00790**	-0.00702**	0.0351*
	(0.00140)	(0.00114)	(0.00127)	(0.0133)
Age Squared				-0.000356**
				(0.000118)
Court-Year FE	X	X	X	X
First-Year Baseline		X	X	X
Cohort FE / Trends			X	X
N	13655	13655	13655	13655
R-sq	0.674	0.694	0.701	0.702

Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01

Table 8 from Ash and MacLeod (2020). Ordinary least squares regression results for

$$y_{ist} = \alpha + \gamma_1 A_{ist} + \gamma_2 A_{ist}^2 + \epsilon_{ist}$$

- y_{ist} = citations to cases by judge i working in court s at year t:
- A_{ist} = age (in years) for judge i in court s at t.
- A_{ist}^2= age squared

Poll 3.3: Select all true statements describing the table (4 minutes).

Differences-in-Differences

- Example: taxes raised in canton A, but **not** in canton B
  - what is the effect on prices in canton A?
- **differences-in-differences** (DD) **estimator is**

$$[Y_{A1} - Y_{A0}] - [Y_{B1} - Y_{B0}]$$

= **price change in treated canton, relative to price change in comparison canton.**

- Identification assumption: **"parallel trends"**
  - Absent tax change, trend in prices would have been the same in cantons A and B.

Diff-in-Diff Regression

- Can estimate the diff-in-diff effect using

$$Y_{jt} = \alpha + \gamma \text{Treat}_{jt} + \lambda \text{After}_{jt} + \rho \text{Treat*After}_{jt} + \varepsilon_{jt}$$

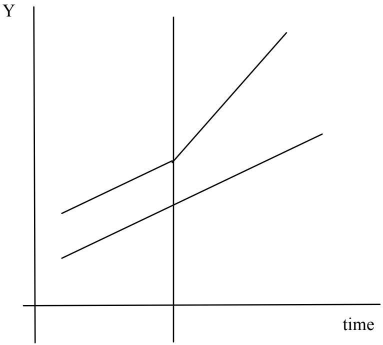
  - canton j, period t
  - Treat $=$ 1 for the reform canton
  - After $=$ 1 for the post-reform period.

```
reg price treat after c.treat#c.after
```

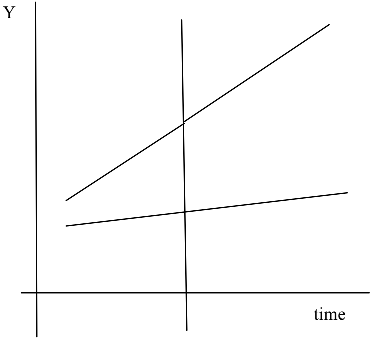
- Interpreting coefficients:
  - α, average in non-treated group, pre-treatment
  - γ, difference between treated and non-treated in pre-treatment period
  - λ, change in the control group after reform
  - ρ, the diff-in-diff treatment effect estimate (change in treatment group, relative to change in control group).

Diff-in-diff: Parallel trends assumption

Good



Not Good



Fixed-Effects Regression

- **Fixed-effects regression** generalizes diffs-in-diffs to > 2 groups and > 2 periods
  - Requires panel (longitudinal) data
  - identification assumption is the same: parallel trends.

$$Y_{jt} = \delta_j + \gamma_t + \beta T_{jt} + \varepsilon_{jt}$$

- δ_j = canton fixed effects
  - categorical variables equaling one for canton j's observations, zero otherwise
- γ_t = year fixed effects
  - categorical variables equaling one for year t's observations, zero otherwise

```
reghdfe price treat_post, absorb(canton year)
```

FE regression is an empirical workhorse

- At any given time, taxes and prices across cantons could be correlated for many confounding reasons.
- Diffs-in-diffs holds constant many of the most important confounders:
  - time-invariant canton-level factors
  - nationwide time-varying factors
- Potential confounders must
  - vary over time by canton
  - correlated with outcome variable
  - correlated with the timing of treatment/reforms

Threats to validity for FE regression

- Can check that treatment cantons evolved similarly to comparison cantons before reform.
  - can also add canton-specific trends.
- Skeptical questions to ask:
  - Why did the treatment group adopt the policy, and not the control group?
  - Were other policies adopted at the same time that might also affect the outcome?
  - Could the treatment spill over into the comparison cantons?

Activity: Private Zoom Chat (3 minutes)

- Imagine that cantons Zurich and Zug each enact a tax cut and you estimate a negative effect on local employment using fixed effects regression. What are some potential confounding factors that would bias this estimate?
  - chat answers to me privately by zoom.

A note on standard errors

- Consider the regression for cantonal tax cuts and employment. We have 26 cantons.
  - the default standard errors formula for OLS assume that all observations are independent realizations.
- Compare the following analyses:
  - including the 10 years before and after the reform ($N = 260$)
  - including the 20 years before and after ($N = 520$)
- Using the default SE's, the second analysis would give much more precise estimate, even though the data contain nearly equivalent information.

Solution: Clustering Standard Errors

Cluster standard errors:

- statistically acknowledges how many independent sources of information there are in the data.
- the standard approach is to cluster at the unit where treatment is assigned.
  - in this example, by canton.

```
reghdfe employment treat_post, absorb(canton year) cluster(canton)
```

- for city-level reforms cluster by city, etc.

Breakout Rooms: Fixed-Effects Regression in Stata

- See link to stata DO file template in zoom chat.
  - recommended: collaborate using atom teletype portal
  - alternative 1: work together on a google doc (make a copy)
  - alternative 2: one person codes and share screen
- Will post solved DO file after lecture.