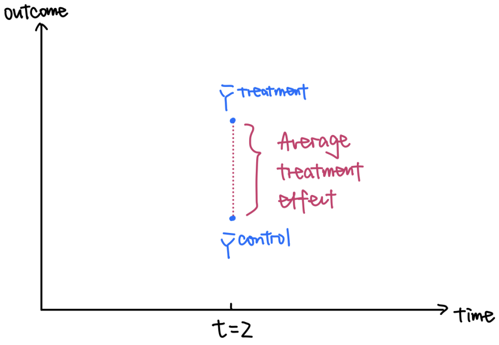
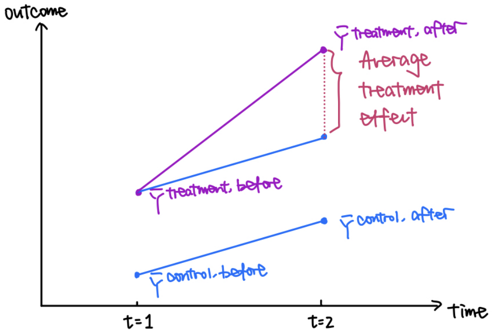
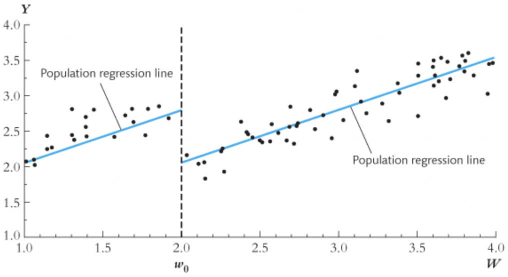
Dis 10: Experiment & Quasi-Experiment; Big Data

1 Experiment & Quasi-Experiment

1.1 Comparison between the two

	Experiment **(Controlled Experiment)**	**Quasi-Experiment** **(Natural Experiment)**
Goal	Obtain causal effect from treatment	
How to conduct	Control all other variations, randomize people into two groups (treated & untreated), so that the only variation between the two groups is whether people received treatment or not.	Use data on real life events, and try to single out the effect of specific treatment. Treatment is administered *as if* it was random.
Costly?	Yes (costly to design an experiment)	No in most cases (since treatment arises naturally)
Example	COVID-19 vaccine trials	Increase of minimum wage in New Jersey
How to obtain causal effects	Differences estimator  outcome $\bar{Y}$ treatment Average treatment effect $\bar{Y}$ control t=2 time	Difference-in-differences estimator  outcome $\bar{Y}$ treatment, after Average treatment effect $\bar{Y}$ treatment, before $\bar{Y}$ control, after $\bar{Y}$ control, before t=1 t=2 time Regression discontinuity 

1.2 One measure of causal effect: Average treatment effect

- Define the following variables:
  - Y_i: outcome for individual i

- X_i: binary variable recording whether treatment is given to individual i
- Y_{1i}: outcome for individual i, if i is treated
- Y_{0i}: outcome for individual i, if i is NOT treated

- Ideally, we'd like to know what $Y_{1i} - Y_{0i}$ is for each individual i. That way, we get to know how effective the treatment is for each person.
- Unfortunately, $Y_{1i} - Y_{0i}$ cannot be obtained ...
  - For individual i, s/he can only receive the treatment or not. There's no such thing as one person both receive the treatment and not receive the treatment at the same time.
  - So we need the next best thing ⇒ **average** the effect of treatment!
- How do we estimate average treatment effect?

Under experiment (specifically, **randomly assigned** controlled experiment):

- Random assignment means that

$$\begin{aligned} E[Y_{1i}] &= E[Y_{1i}|X_i = 1] \\ E[Y_{0i}] &= E[Y_{0i}|X_i = 0] \end{aligned}$$

- Hence, average treatment effect is

$$\begin{aligned} E[Y_{1i} - Y_{0i}] &= E[Y_{1i}] - E[Y_{0i}] \\ &= E[Y_{1i}|X_i = 1] - E[Y_{0i}|X_i = 0] \\ &= E[Y_i|X_i = 1] - E[Y_i|X_i = 0] \\ &= \textbf{Average Treatment Effect} \end{aligned}$$

- In other words, calculating the average of outcome for treated and untreated groups, and then take the difference. This difference represents the average treatment effect.
- Recall from what we learned in Dis 5. Effectively, we have a change of **intercept**, so this can be implemented by regressing outcome Y_i on X_i:

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

  where

$$\begin{aligned} E[Y_i|X_i = 0] &= \beta_0 \\ E[Y_i|X_i = 1] &= \beta_0 + \beta_1 \end{aligned}$$

  meaning that $\beta_1 = E[Y_i|X_i = 1] - E[Y_i|X_i = 0] = \text{Average Treatment Effect}$
  This is why β_1 is called the **differences estimator**.
- Additional variables can be included in the regression to control for omitted variable bias and to reduce the standard error of the regression.

Under quasi-experiment:

- We no longer have random assignment of treatment, since in quasi-experiment, treatment occurs naturally, and is often times selectively implemented (think about public policy).
- This means that when we directly look at the outcome difference between treated and untreated group, this difference captures both **effect from treatment**, and **the underlying difference between the two groups**.
- New idea: Capture **difference in differences**

$$\begin{aligned}
&\underbrace{(E[Y_{1i}|\text{after}] - E[Y_{0i}|\text{after}])}_{\text{average treatment effect + other}} - \underbrace{(E[Y_{1i}|\text{before}] - E[Y_{0i}|\text{before}])}_{\text{other}} \\
&= (E[Y_i|X_i = 1, \text{after}] - E[Y_i|X_i = 0, \text{after}]) - (E[Y_i|X_i = 1, \text{before}] - E[Y_i|X_i = 0, \text{before}]) \\
&= (E[Y_{i,\text{after}}|X_i = 1] - E[Y_{i,\text{after}}|X_i = 0]) - (E[Y_{i,\text{before}}|X_i = 1] - E[Y_{i,\text{before}}|X_i = 0]) \\
&= E[Y_{i,\text{after}} - Y_{i,\text{before}}|X_i = 1] - E[Y_{i,\text{after}} - Y_{i,\text{before}}|X_i = 0] \\
&= \textbf{Average Treatment Effect}
\end{aligned}$$

  where "after" means after treatment, and "before" means before treatment.
- Using regression technique to represent difference-in-differences:

$$\Delta Y_i = \beta_0 + \beta_1 X_i + u_i$$

  where

$$\begin{aligned}
\Delta Y_i &= Y_{i,\text{after}} - Y_{i,\text{before}} \\
E[\Delta Y_i|X_i = 0] &= \beta_0 \\
E[\Delta Y_i|X_i = 1] &= \beta_0 + \beta_1
\end{aligned}$$

  meaning that $\beta_1 = E[\Delta Y_i|X_i = 1] - E[\Delta Y_i|X_i = 0] = \text{Average Treatment Effect}$
  This is why β_1 is called the **difference-in-differences estimator** (now there's an additional time difference).
- Side note: The following regression yields equivalent result compared with the one above

$$Y_{it} = \beta_0 + \beta_1 X_i \times T_t + \beta_2 X_i + \beta_3 T_t + u_{it}$$

  where

  - T_t is a binary variable recording whether time is after the treatment
  - $X_i \times T_t$ is a binary interaction variable. It equals 1 only when i is in treated group, and time is after treatment has been received.

  β_1 here is also the **difference-in-differences estimator**.
- Similar to the experiment environment, additional variables can be included in the regression to control for omitted variable bias and to reduce the standard error of the regression.

1.3 Another measure of causal effect (mainly for quasi-experiment)

- Consider the case when treatment is assigned based on a specific cutoff.

  Example. Only people who have been unemployed for more than 6 months receive job training.

  ⇒ treatment $= 1$ if duration of unemployment > 6 months

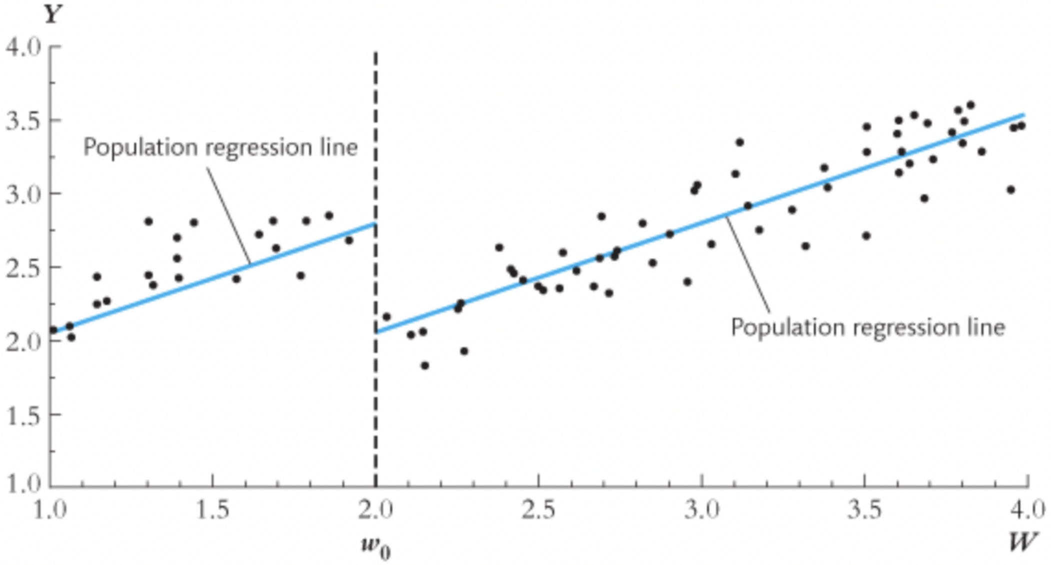
  treatment $= 0$ if duration of unemployment ≤ 6 months

  Example. Only counties with poverty rate less than certain number C receive federal aid.

  ⇒ treatment $= 1$ if county poverty rate $< C$.

  treatment $= 0$ if county poverty rate $\geq C$.

- One might be interested to know if the same regression is run among the treated group, and then run again among the untreated group, if there's going to be any difference.
  - More specifically, if there's a change in intercept right around the cutoff value, then the amount of change in intercept represents the causal effect from treatment!
  - Illustrating graphically:



  - This is the idea of **regression discontinuity** (discontinuity at the cutoff value that determines whether someone gets treated or not).

- How to estimate the causal effect using regression discontinuity?
  - Define the following variables:
    - ∗ Y_i: outcome for individual i
    - ∗ W_i: variable that will determine whether someone receives treatment
    - ∗ X_i: binary variable recording whether someone receives treatment
      - · $X_i = 1$ if $W_i > C$, where C is the cutoff value
      - · $X_i = 0$ if $W_i \leq C$
  - Consider the following model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

Here,

$$\lim_{W_i \downarrow C} \hat{Y}_i = \lim_{W_i \downarrow C} \left(\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 W_i\right) = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 C \qquad \text{(Approach from the right side of C)}$$
$$\lim_{W_i \uparrow C} \hat{Y}_i = \lim_{W_i \uparrow C} \left(\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 W_i\right) = \hat{\beta}_0 + 0 + \hat{\beta}_2 C \qquad \text{(Approach from the left side of C)}$$

so the causal effect $= \lim_{W_i \downarrow C} \hat{Y}_i - \lim_{W_i \uparrow C} \hat{Y}_i = \hat{\beta}_1$

- Side note: What we have been talking about is called **sharp regression discontinuity**. "Sharp" here means that the decision on receiving the treatment is determined solely through a cutoff rule.

  In reality, there are cases where the decision on receiving the treatment isn't determined solely on a cutoff. For example, someone who has been unemployed more than 6 months is qualified for job training, but they might not participate due to the job training site being too far away from home. Another scenario is that someone who has been unemployed for roughly 6 months gets job training, making the cutoff less precise.

  These cases are handled by **fuzzy regression discontinuity** models. We won't go too much into the theory part of the analysis, but you should know when the fuzzy design is more appropriate than the sharp design.

1.4 Validity concerns

- See textbook chapter 13.2 for threat to internal and external validity of experiments
- See textbook chapter 13.5 for threat to internal and external validity of quasi-experiments

2 Big Data

- So far, our main objective of econometric analysis is to derive the casual relationship between two variables: we are interested in the coefficient on some variables when running a regression, and the coefficient could be interpreted causally if we can establish causal relationship.
- Another objective of econometric analysis is to predict values:
  - When prediction is our main objective, we often have more predictors than the number of observations available to us. When this happens, we say that we are handling **big data**.
  - Too many predictors make OLS an inappropriate fitting technique, as OLS will try to overfit the pattern of the data.
  - So we need alternatives.
- Model for prediction:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$$

  where all variables have been standardized.

  When I say "standardized", what I mean is, say that the true observations of Y and X are denoted as Y^* and X^*, then the "standardized" variables that end up entering the model are

$$Y_i = Y_i^* - \overline{Y_i^*} \qquad\qquad X_{ki} = \frac{X_{ki}^* - \overline{X_{ki}^*}}{sd(X_{ki}^*)}$$

- How to test whether a model is good at prediction?
  - Use some part of the sample to estimate our model, and then use the rest to test prediction.
  - For the rest of the data used for prediction, we know all the X values, and we know the true Y value. So we can predict Y based on these X, and obtain $\hat{Y}$.
  - Error of prediction for specific observation $i = Y_i - \hat{Y}_i = Y_i^* - \hat{Y}_i^*$
  - Mean Squared Prediction Error (→ average error) $= MSPE = E[(Y_i - \hat{Y}_i)^2] = E[(Y_i^* - \hat{Y}_i^*)^2]$
- Idea #1: **Trade estimators being unbiased for estimators with smaller standard errors**
  - **Shrinkage estimator**: An estimator that "shrinks" the OLS estimator toward a specific number, and thereby reduces the variance of the estimator.
  - Two types of shrinkage estimator: Ridge and Lasso

$$\hat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\beta_1,\ldots,\beta_k} \left[\sum_{i=1}^{n}(Y_i - \beta_1 X_{1i} - \ldots - \beta_k X_{ki})^2\right]$$

$$\hat{\boldsymbol{\beta}}_{Ridge} = \arg\min_{\beta_1,\ldots,\beta_k} \left[\sum_{i=1}^{n}(Y_i - \beta_1 X_{1i} - \ldots - \beta_k X_{ki})^2\right] + \underbrace{\lambda_{Ridge}\sum_{j=1}^{k}\beta_j^2}_{\text{penalizing term}}$$

$$\hat{\boldsymbol{\beta}}_{Lasso} = \arg\min_{\beta_1,\ldots,\beta_k} \left[\sum_{i=1}^{n}(Y_i - \beta_1 X_{1i} - \ldots - \beta_k X_{ki})^2\right] + \underbrace{\lambda_{Lasso}\sum_{j=1}^{k}|\beta_j|}_{\text{penalizing term}}$$

  Both Ridge and Lasso regressors shrink $\hat{\beta}$ towards 0.

- Idea #2: **Don't fully abandon OLS, but be selective on the regressors**
  - **Principal Component**: Replace the X variables as linear combination of them, denoted as

$$PC_j = w_1 X_1 + w_2 X_2 + \ldots + w_k X_k$$

    - These PC_j need to satisfy the following conditions:
      1. $w_1^2 + w_2^2 + \ldots + w_k^2 = 1$;
      2. PC_1 maximizes the variance of its linear combination;
      3. PC_2 maximizes the variance of its linear combination, while uncorrelated with PC_1;
      4. And so on
    - This way, instead of having a bunch of X variables with only some of them having meaningful variation that would explain Y, we now have principal components that captures important X variations, and the degree of variation being captured is descending (the first PC captures more than the second, the second captures more than the third, and so on).
  - **Stepwise selection**: Algorithm for selecting which subset of regressors to include in the model.
    - Forward selection: Start off with 0 regressor included, then add one at a time.
    - Backward selection: Start off with all regressors included, then take out one at a time.
    - Evaluation criterion: adjusted R^2 (bigger ⇔ better); AIC and BIC (smaller ⇔ better)
      - In Stata, use `vselect X1 X2 ...` to get adjusted R^2, AIC, and BIC for subset of variables.

3 Problems

1. (Prediction) Consider the following prediction model:

$$Y_i = \beta_1 X_i + u_i$$

The model is estimated to be the following:

$$\hat{Y}_i = \underset{(2)}{10} X_i$$

We also know that $\overline{Y_i^*} = 5, \overline{X_i^*} = 50$, and $sd(X_i^*) = 6$.

(a) Construct 90% confidence interval for the effect of three units increase of X_i onto Y_i.

Recall that a confidence interval is constructed as

$$[\text{point estimate} - z_{1-\frac{\alpha}{2}} \times se(\text{point estimate}), \quad \text{point estimate} + z_{1-\frac{\alpha}{2}} \times se(\text{point estimate})]$$

where $\alpha =$ significance level $= 1-$ confidence level.

In this case, since we are looking at the effect of three units increase of X_i onto Y_i,

$$\text{point estimate} = 3\hat{\beta}_1 = 3 \times 10 = 30$$

And the standard error of the point estimate is

$$se(\text{point estimate}) = se(3\hat{\beta}_1) = \sqrt{Var(3\hat{\beta}_1)} = \sqrt{3^2 Var(\hat{\beta}_1)} = 3 \times se(\hat{\beta}_1) = 3 \times 2 = 6$$

Since we are constructing a 90% confidence interval, the significance level $= 1-$ confidence level $= .1$, which means $\alpha = .1$. This tells us that the z-score we should look for is $z_{1-\frac{.1}{2}=.95}$, which is the level of z that yields $\Phi(z) = .95$.

Looking it up on z-table, we find that $z_{.95} = 1.65$.

Thus, with all pieces calculated, the 90% confidence interval for the effect of three units increase of X_i onto Y_i is

$$[30 - 1.65 \times 6, \quad 30 + 1.65 \times 6] \quad \Leftrightarrow \quad [20.1, 39.9]$$

(b) A new entry of X^* has $X_i^* = 20$. What's the predicted true value of Y (that is, what is $\hat{Y}_i^*$)?

A tricky thing to adjust for in prediction model is that the Y and X entering the model are already demeaned / standardized. Therefore, to obtain $\hat{Y}_i^*$, we need to first find $\hat{Y}_i$ by standardizing X:

$$X_i = \frac{X_i^* - \overline{X_i^*}}{sd(X_i^*)} = \frac{20 - 50}{6} = -5$$

Now with the standardized X_i, let's plug it into the prediction model:

$$\begin{aligned} \hat{Y}_i &= 10 \times X_i \\ \hat{Y}_i^* - \overline{Y_i^*} &= 10 \times (-5) \\ \hat{Y}_i^* &= 5 - 50 = -45 \end{aligned}$$

Thus, the predicted true value of Y is -45.

(c) The true value for this new entry of X_i^* is $Y_i^* = -50$. Additionally, we have another predicted $\hat{Y}_i^* = 40$ where the true $Y_i^* = 20$ instead. Suppose that $\overline{Y_i^*} = 5$ is held constant. What's the MSPE (Mean Squared Prediction Error)?

We have two predictions with their true values, so

$$\begin{aligned} MSPE &= E[(Y_i - \hat{Y}_i)^2] = E[(Y_i^* - \hat{Y}_i^*)^2] \\ &= \frac{1}{2}\left[(Y_1^* - \hat{Y}_1^*)^2 + (Y_2^* - \hat{Y}_2^*)^2\right] \\ &= \frac{1}{2}\left[(-50 - (-45))^2 + (20 - 40)^2\right] = 212.5 \end{aligned}$$

The MSPE is 212.5 here.

(d) Write the prediction model to be in terms of true data points, instead of the demeaned and standardized values.

To do so, we need to write down the expression of the demeaned and the standardized variables:

$$\begin{aligned} Y_i &= \beta_1 X_i + u_i \\ Y_i^* - \overline{Y_i^*} &= \beta_1 \frac{X_i^* - \overline{X_i^*}}{sd(X_i^*)} + u_i \\ Y_i^* &= \underbrace{\overline{Y_i^*} - \frac{\beta_1 \overline{X_i^*}}{sd(X_i^*)}}_{\gamma_0} + \underbrace{\frac{\beta_1}{sd(X_i^*)}}_{\gamma_1} X_i^* + u_i \end{aligned}$$

This implies that the estimated new parameters are

$$\begin{aligned} \hat{\gamma}_0 &= \overline{Y_i^*} - \frac{\hat{\beta}_1 \overline{X_i^*}}{sd(X_i^*)} = 5 - \frac{10 \times 50}{6} = -78.33 \\ \hat{\gamma}_1 &= \frac{\hat{\beta}_1}{sd(X_i^*)} = \frac{10}{6} = 1.67 \end{aligned}$$

Therefore, the estimated model in terms of true data points is

$$\hat{Y}_i^* = -78.33 + 1.67 X_i^*$$

Side note: Try plugging in $X_i^* = 20$ that we used in part (b). $\hat{Y}_i^*$ using the true data points is predicted to be -44.93, which is basically -45 that we found in part (b) accounting for rounding errors.

2. A company has a voluntary training course offered to its salespeople since 2010, and it wants to evaluate how effective the program is to help decide whether the program will continue going forward. To do this, the company has class registration information on who participated in this training course, along with recorded sales for their salespeople in 2005 and in 2015.

   (a) What is the treatment here? Is this an experiment or a quasi-experiment?

   The treatment here is participating in the training course. This is a quasi-experiment, since this is not a carefully designed randomized trial where people are randomized into taking the training course, but we are rather taking the dataset on the treatment and consider the trial to be "as if" randomly given to people.

   (b) To study the average treatment effect, use the following model:

$$\text{sales}_{it} = \beta_0 + \beta_1 X_i + \beta_2 T_t + \beta_3 X_i \times T_t + u_{it}$$

   where

   - sales_{it} records the sales of individual i in year t
   - X_i is a binary variable. $X_i = 1$ if the individual participated in the training course.
   - T_t is a binary variable. $T_t = 1$ if the year is 2015. $T_t = 0$ if the year is 2005.

   which coefficient records average treatment effect? Explain why.

   β_3 records average treatment effect, since it is the difference-in-differences estimator. To see why,

$$\begin{aligned}
E[\text{sales}_{it}|X_i = 0, T_t = 0] &= \beta_0 \\
E[\text{sales}_{it}|X_i = 1, T_t = 0] &= \beta_0 + \beta_1 \\
E[\text{sales}_{it}|X_i = 0, T_t = 1] &= \beta_0 + \beta_2 \\
E[\text{sales}_{it}|X_i = 1, T_t = 1] &= \beta_0 + \beta_1 + \beta_2 + \beta_3
\end{aligned}$$

   Difference-in-differences estimate

$$\begin{aligned}
&(E[\text{sales}_{it}|X_i = 1, T_t = 1] - E[\text{sales}_{it}|X_i = 0, T_t = 1]) \\
&\quad - (E[\text{sales}_{it}|X_i = 1, T_t = 0] - E[\text{sales}_{it}|X_i = 0, T_t = 0]) \\
&= (\beta_1 + \beta_3) - (\beta_1) = \beta_3
\end{aligned}$$

   Thus, β_3 represents the average treatment effect.

   (c) Suppose that all salespeople whose pay rate at the company is above $100k chose to not participate in the training course. Is there concern regarding internal validity?

   Yes, there is concern regarding internal validity. This is a classic example of attrition: people self select towards receiving the treatment or not, which causes selection bias.

   More explicitly, we can think about this as people making more than $100k has higher ability, which is part of the unobserved error. This starts to become a concern of omitted variable bias if

   - higher ability is correlated with the β_3 coefficient of interest, which means higher ability is correlated with the decision of receiving treatment (whether $X_i = 1$), and
   - higher ability affects sales

   The second bullet point most definitely holds. With people self selecting, the first bullet point also holds, which is why this self selection raises concern of internal validity.