GOV 51 Section

Week 4: Introduction to Regression

Pranav Moudgalya
(from James Jolin)

Harvard College

Prediction

- **What is prediction?** Our best guess of a realized outcome.

Prediction

- **What is prediction?** Our best guess of a realized outcome.

$$\textbf{prediction error} = \text{actual outcome} - \text{predicted outcome}$$

Prediction

- **What is prediction?** Our best guess of a realized outcome.

$$\textbf{prediction error} = \text{actual outcome} - \text{predicted outcome}$$

- **How to evaluate prediction errors?**

Prediction

- **What is prediction?** Our best guess of a realized outcome.

$$\textbf{prediction error} = \text{actual outcome} - \text{predicted outcome}$$

- **How to evaluate prediction errors?**
  - **Bias**: take the average of prediction errors.

Prediction

- **What is prediction?** Our best guess of a realized outcome.

$$\textbf{prediction error} = \text{actual outcome} - \text{predicted outcome}$$

- **How to evaluate prediction errors?**
  - **Bias**: take the average of prediction errors.
  - **Root mean-squared error**:average magnitude of the prediction error.

Prediction

- **What is prediction?** Our best guess of a realized outcome.

$$\textbf{prediction error} = \text{actual outcome} - \text{predicted outcome}$$

- **How to evaluate prediction errors?**
  - **Bias**: take the average of prediction errors.
  - **Root mean-squared error**:average magnitude of the prediction error.
- Prediction of binary outcome variable $\rightsquigarrow$ classification problem

Prediction

- **What is prediction?** Our best guess of a realized outcome.

$$\textbf{prediction error} = \text{actual outcome} - \text{predicted outcome}$$

- **How to evaluate prediction errors?**
  - **Bias**: take the average of prediction errors.
  - **Root mean-squared error**:average magnitude of the prediction error.
- Prediction of binary outcome variable $\rightsquigarrow$ classification problem
- Wrong prediction $\rightsquigarrow$ misclassification. Types? Consider the binary outcome variable Y with possible values $\{0, 1\}$.

Prediction

- **What is prediction?** Our best guess of a realized outcome.

$$\textbf{prediction error} = \text{actual outcome} - \text{predicted outcome}$$

- **How to evaluate prediction errors?**
  - **Bias**: take the average of prediction errors.
  - **Root mean-squared error**:average magnitude of the prediction error.
- Prediction of binary outcome variable $\rightsquigarrow$ classification problem
- Wrong prediction $\rightsquigarrow$ misclassification. Types? Consider the binary outcome variable Y with possible values $\{0, 1\}$.
  - **True positive**: predict 1 when actually 1.

Prediction

- **What is prediction?** Our best guess of a realized outcome.

$$\textbf{prediction error} = \text{actual outcome} - \text{predicted outcome}$$

- **How to evaluate prediction errors?**
  - **Bias**: take the average of prediction errors.
  - **Root mean-squared error**:average magnitude of the prediction error.
- Prediction of binary outcome variable $\rightsquigarrow$ classification problem
- Wrong prediction $\rightsquigarrow$ misclassification. Types? Consider the binary outcome variable Y with possible values $\{0, 1\}$.
  - **True positive**: predict 1 when actually 1.
  - **False positive**: predict 1 when actually 0.

Prediction

- **What is prediction?** Our best guess of a realized outcome.

$$\textbf{prediction error} = \text{actual outcome} - \text{predicted outcome}$$

- **How to evaluate prediction errors?**
  - **Bias**: take the average of prediction errors.
  - **Root mean-squared error**:average magnitude of the prediction error.
- Prediction of binary outcome variable $\rightsquigarrow$ classification problem
- Wrong prediction $\rightsquigarrow$ misclassification. Types? Consider the binary outcome variable Y with possible values $\{0, 1\}$.
  - **True positive**: predict 1 when actually 1.
  - **False positive**: predict 1 when actually 0.
  - **True negative**: predict 0 when actually 0.

Prediction

- **What is prediction?** Our best guess of a realized outcome.

$$\textbf{prediction error} = \text{actual outcome} - \text{predicted outcome}$$

- **How to evaluate prediction errors?**
  - **Bias**: take the average of prediction errors.
  - **Root mean-squared error**:average magnitude of the prediction error.
- Prediction of binary outcome variable $\rightsquigarrow$ classification problem
- Wrong prediction $\rightsquigarrow$ misclassification. Types? Consider the binary outcome variable Y with possible values $\{0, 1\}$.
  - **True positive**: predict 1 when actually 1.
  - **False positive**: predict 1 when actually 0.
  - **True negative**: predict 0 when actually 0.
  - **False negative**: predict 0 when actually 1.

Prediction with a Line

- Simplest possible way to relate two variables: a line, $y = mx + b$.

Prediction with a Line

- Simplest possible way to relate two variables: a line, $y = mx + b$.
- Problem: not everything will fall on the line!
  - Some points will be above the line, some below.
  - Need a way to account for chance variation away from the line.

Prediction with a Line

- Simplest possible way to relate two variables: a line, $y = mx + b$.
- Problem: not everything will fall on the line!
  - Some points will be above the line, some below.
  - Need a way to account for chance variation away from the line.
- Solution to above problem $\rightsquigarrow$ linear model

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

Prediction with a Line

- Simplest possible way to relate two variables: a line, $y = mx + b$.
- Problem: not everything will fall on the line!
  - Some points will be above the line, some below.
  - Need a way to account for chance variation away from the line.
- Solution to above problem $\rightsquigarrow$ linear model

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

$$\underbrace{Y_i}_{\text{Outcome variable for unit } i} = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1}_{\text{Slope}} \cdot \underbrace{X_i}_{\text{Explanatory variable for unit } i} + \underbrace{\epsilon_i}_{\text{Error for unit} i}$$

Linear Model

$$\underbrace{Y_i}_{\text{Outcome variable for unit } i} = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1}_{\text{Slope}} \cdot \underbrace{X_i}_{\text{Explanatory variable for unit } i} + \underbrace{\epsilon_i}_{\text{Error for unit} i}$$

- Notes on the foregoing equation
  - **Coefficients/parameters** (β_0, β_1): true unknown intercept/slope of the line of best fit.

Linear Model

$$\underbrace{Y_i}_{\text{Outcome variable for unit } i} = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1}_{\text{Slope}} \cdot \underbrace{X_i}_{\text{Explanatory variable for unit } i} + \underbrace{\epsilon_i}_{\text{Error for unit} i}$$

- Notes on the foregoing equation
  - **Coefficients/parameters** (β_0, β_1): true unknown intercept/slope of the line of best fit.
  - **Chance error** (ϵ_i): for each unit, accounts for the fact that the line doesn't perfectly fit the data.
    - Each observation allowed to be off the regression line
    - On average, chance errors assumed to be 0.

Linear Model

$$\underbrace{Y_i}_{\text{Outcome variable for unit } i} = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1}_{\text{Slope}} \cdot \underbrace{X_i}_{\text{Explanatory variable for unit } i} + \underbrace{\epsilon_i}_{\text{Error for unit} i}$$

- Notes on the foregoing equation
  - **Coefficients/parameters** (β_0, β_1): true unknown intercept/slope of the line of best fit.
  - **Chance error** (ϵ_i): for each unit, accounts for the fact that the line doesn't perfectly fit the data.
    - Each observation allowed to be off the regression line
    - On average, chance errors assumed to be 0.
  - Useful fiction: this model represents the **data generating process**

Linear Model

$$\underbrace{Y_i}_{\text{Outcome variable for unit } i} = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1}_{\text{Slope}} \cdot \underbrace{X_i}_{\text{Explanatory variable for unit } i} + \underbrace{\epsilon_i}_{\text{Error for unit} i}$$

- Notes on the foregoing equation
  - **Coefficients/parameters** (β_0, β_1): true unknown intercept/slope of the line of best fit.
  - **Chance error** (ϵ_i): for each unit, accounts for the fact that the line doesn't perfectly fit the data.
    - Each observation allowed to be off the regression line
    - On average, chance errors assumed to be 0.
  - Useful fiction: this model represents the **data generating process**
  - **Intercept** (β_0): average value of Y when X is 0

Linear Model

$$\underbrace{Y_i}_{\text{Outcome variable for unit } i} = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1}_{\text{Slope}} \cdot \underbrace{X_i}_{\text{Explanatory variable for unit } i} + \underbrace{\epsilon_i}_{\text{Error for unit} i}$$

- Notes on the foregoing equation
  - **Coefficients/parameters** (β_0, β_1): true unknown intercept/slope of the line of best fit.
  - **Chance error** (ϵ_i): for each unit, accounts for the fact that the line doesn't perfectly fit the data.
    - Each observation allowed to be off the regression line
    - On average, chance errors assumed to be 0.
  - Useful fiction: this model represents the **data generating process**
  - **Intercept** (β_0): average value of Y when X is 0
  - **Slope** (β_1): average change in Y when X increases by one unit.

Estimation

- Parameters: β_0, β_1
  - *Unknown* features of the data-generating process.
  - Chance error makes these *impossible* to observe directly.

Estimation

- Parameters: β_0, β_1
  - *Unknown* features of the data-generating process.
  - Chance error makes these *impossible* to observe directly.
- Estimates $\hat{\beta}_0, \hat{\beta}_1$
  - An **estimate** is our best guess about some parameter, given sample.

Estimation

- Parameters: β_0, β_1
  - *Unknown* features of the data-generating process.
  - Chance error makes these *impossible* to observe directly.
- Estimates $\hat{\beta}_0, \hat{\beta}_1$
  - An **estimate** is our best guess about some parameter, given sample.
- **Regression line**: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$
  - Notice how we **NO LONGER** use the notation Y_i or X_i $\rightsquigarrow$ we are now thinking about averages.
  - Average value of Y when X is equal to x.
  - Represents the best guess or **predicted value** of the outcome at x.

Estimation

- Parameters: β_0, β_1
  - *Unknown* features of the data-generating process.
  - Chance error makes these *impossible* to observe directly.
- Estimates $\hat{\beta}_0, \hat{\beta}_1$
  - An **estimate** is our best guess about some parameter, given sample.
- **Regression line**: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$
  - Notice how we **NO LONGER** use the notation Y_i or X_i $\rightsquigarrow$ we are now thinking about averages.
  - Average value of Y when X is equal to x.
  - Represents the best guess or **predicted value** of the outcome at x.

Line of Best Fit is a CONDITIONAL EXPECTATION FUNCTION!

Estimation

- Parameters: β_0, β_1
  - *Unknown* features of the data-generating process.
  - Chance error makes these *impossible* to observe directly.
- Estimates $\hat{\beta}_0, \hat{\beta}_1$
  - An **estimate** is our best guess about some parameter, given sample.
- **Regression line**: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$
  - Notice how we **NO LONGER** use the notation Y_i or X_i $\rightsquigarrow$ we are now thinking about averages.
  - Average value of Y when X is equal to x.
  - Represents the best guess or **predicted value** of the outcome at x.

Line of Best Fit is a CONDITIONAL EXPECTATION FUNCTION!

$$\mathbb{E}[Y \mid x] = \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Regression is not magic

- People love to regress some variable on another, but that doesn't inherently mean anything. If you regress some continuous variable (think income) on some binary variable (think treated vs. not treated) your β_0 and β_1 are just describing the average value within the treated group ($x = 0$) and the difference in the average values in the treated and non-treated group. Why?

Regression is not magic

- People love to regress some variable on another, but that doesn't inherently mean anything. If you regress some continuous variable (think income) on some binary variable (think treated vs. not treated) your β_0 and β_1 are just describing the average value within the treated group ($x = 0$) and the difference in the average values in the treated and non-treated group. Why?

Line of Best Fit is a CONDITIONAL EXPECTATION FUNCTION!

Regression is not magic

- People love to regress some variable on another, but that doesn't inherently mean anything. If you regress some continuous variable (think income) on some binary variable (think treated vs. not treated) your β_0 and β_1 are just describing the average value within the treated group ($x = 0$) and the difference in the average values in the treated and non-treated group. Why?

Line of Best Fit is a CONDITIONAL EXPECTATION FUNCTION!

- Regression becomes useful when you combine it with theory: controlling for potential confounders, modeling DiD or other research designs. Thinking about theory and assumptions must precede regression for its estimates to be useful.

Regression is not magic

- People love to regress some variable on another, but that doesn't inherently mean anything. If you regress some continuous variable (think income) on some binary variable (think treated vs. not treated) your β_0 and β_1 are just describing the average value within the treated group ($x = 0$) and the difference in the average values in the treated and non-treated group. Why?

Line of Best Fit is a CONDITIONAL EXPECTATION FUNCTION!

- Regression becomes useful when you combine it with theory: controlling for potential confounders, modeling DiD or other research designs. Thinking about theory and assumptions must precede regression for its estimates to be useful.
- Main point to takeaway: Regression only yields a **causally identified effect** in the coefficient on a treatment variable if you've controlled for all possible confounding variables in your model such that treatment is effectively random. We call this **unconfoundedness**:

Regression is not magic

- People love to regress some variable on another, but that doesn't inherently mean anything. If you regress some continuous variable (think income) on some binary variable (think treated vs. not treated) your β_0 and β_1 are just describing the average value within the treated group ($x = 0$) and the difference in the average values in the treated and non-treated group. Why?

Line of Best Fit is a CONDITIONAL EXPECTATION FUNCTION!

- Regression becomes useful when you combine it with theory: controlling for potential confounders, modeling DiD or other research designs. Thinking about theory and assumptions must precede regression for its estimates to be useful.
- Main point to takeaway: Regression only yields a **causally identified effect** in the coefficient on a treatment variable if you've controlled for all possible confounding variables in your model such that treatment is effectively random. We call this **unconfoundedness**:

$$T_i \perp \{Y_i(1), Y_i(0)\}_{i=1}^{n} \mid \mathbf{X}_i \forall x_i$$

Least squares

- We get our estimates of β_0, β_1 by the **least squares method.**

Least squares

- We get our estimates of β_0, β_1 by the **least squares method.**
- Minimize the **sum of the squared residuals** (SSR)

$$\text{SSR} = \sum_{i=1}^{n} (\text{prediction error}_i)^2 = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

Least squares

- We get our estimates of β_0, β_1 by the **least squares method.**
- Minimize the **sum of the squared residuals** (SSR)

$$\text{SSR} = \sum_{i=1}^{n} (\text{prediction error}_i)^2 = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\tilde{\beta}_0, \tilde{\beta}_1}{\arg\min} \sum_{i=1}^{n} (Y_i - (\tilde{\beta}_0 + \tilde{\beta}_1 X_i))^2$$

Least squares

- We get our estimates of β_0, β_1 by the **least squares method.**
- Minimize the **sum of the squared residuals** (SSR)

$$\text{SSR} = \sum_{i=1}^{n} (\text{prediction error}_i)^2 = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\tilde{\beta}_0, \tilde{\beta}_1}{\arg\min} \sum_{i=1}^{n} (Y_i - (\tilde{\beta}_0 + \tilde{\beta}_1 X_i))^2$$

- Finds the line that minimizes the magnitude of the prediction errors!
- Least squares line **always** goes through $(\bar{X}, \bar{Y})$!

More on Regression & Model fit

- Estimated slope is related to correlation:

$$\hat{\beta} = (\text{correlation of } x \text{ and } y) \times \frac{\sigma_y}{\sigma_x}$$

More on Regression & Model fit

- Estimated slope is related to correlation:

$$\hat{\beta} = (\text{correlation of } x \text{ and } y) \times \frac{\sigma_y}{\sigma_x}$$

- Prediction error for regression: **Sum of squared residuals**

$$\text{SSR} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (\hat{\alpha} + \hat{\beta} x_i))^2$$

More on Regression & Model fit

- Estimated slope is related to correlation:

$$\hat{\beta} = (\text{correlation of } x \text{ and } y) \times \frac{\sigma_y}{\sigma_x}$$

- Prediction error for regression: **Sum of squared residuals**

$$\text{SSR} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (\hat{\alpha} + \hat{\beta} x_i))^2$$

- Benchmarking our predictions using the **proportional reduction in error**:

$$\frac{\text{reduction in prediction error using model}}{\text{baseline prediction error}}$$

More on Model fit

- Baseline prediction error without a regression is using the mean of Y to predict. This is called the **Total sum of squares**:

$$\text{TSS} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

More on Model fit

- Baseline prediction error without a regression is using the mean of Y to predict. This is called the **Total sum of squares**:

$$\text{TSS} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

- Combining SSR and TSS from earlier we get the **coefficient of determination**, or R^2:

$$R^2 = \frac{\text{TSS} - \text{SSR}}{\text{TSS}} = \frac{\text{how much smaller LS prediction errors are vs mean}}{\text{prediction error using the mean}}$$

- **Can be very misleading!** Does not guarantee linear model is ideal for given data.

Overfitting

- **In-sample fit**: how well your model predicts the data used to estimate it.
  - R^2 is one measure of in-sample fit

Overfitting

- **In-sample fit**: how well your model predicts the data used to estimate it.
  - R^2 is one measure of in-sample fit
- **Out-of-sample fit**: how well your model predicts new data.

Overfitting

- **In-sample fit**: how well your model predicts the data used to estimate it.
  - R^2 is one measure of in-sample fit
- **Out-of-sample fit**: how well your model predicts new data.
- **Overfitting**: OLS optimizes in-sample fit; may do poorly out of sample.

Multiple predictors

$$Y_i = \alpha + \beta_1 X_i + \boldsymbol{\gamma} \mathbf{Z}_i + \epsilon_i$$

Multiple predictors

$$Y_i = \alpha + \beta_1 X_i + \boldsymbol{\gamma} \mathbf{Z}_i + \epsilon_i$$

- Motivation for multiple regression:
  - Better predictions (at least in-sample).

Multiple predictors

$$Y_i = \alpha + \beta_1 X_i + \boldsymbol{\gamma} \mathbf{Z}_i + \epsilon_i$$

- Motivation for multiple regression:
  - Better predictions (at least in-sample).
  - Better interpretation as **ceteris paribus** relationships:
    - Consider above expression
    - β_1 is the change in Y on average for a one unit increase in X, holding constant $\mathbf{Z}$ (i.e.,ceteris paribus)

Multiple predictors

$$Y_i = \alpha + \beta_1 X_i + \boldsymbol{\gamma} \mathbf{Z}_i + \epsilon_i$$

- Motivation for multiple regression:
  - Better predictions (at least in-sample).
  - Better interpretation as **ceteris paribus** relationships:
    - Consider above expression
    - β_1 is the change in Y on average for a one unit increase in X, holding constant $\mathbf{Z}$ (i.e.,ceteris paribus)
  - **Statistical control** in a cross-sectional study.

Multiple predictors

$$Y_i = \alpha + \beta_1 X_i + \boldsymbol{\gamma} \mathbf{Z}_i + \epsilon_i$$

- Motivation for multiple regression:
  - Better predictions (at least in-sample).
  - Better interpretation as **ceteris paribus** relationships:
    - Consider above expression
    - β_1 is the change in Y on average for a one unit increase in X, holding constant $\mathbf{Z}$ (i.e.,ceteris paribus)
  - **Statistical control** in a cross-sectional study.
- How do we estimate the coefficients? $\rightsquigarrow$ The same way as before! Minimize SSR.

$$\text{SSR} = \sum_{i=1}^{n} (\epsilon_i)^2 = \sum_{i=1}^{n} (Y_i - (\hat{\alpha} + \hat{\beta} X_i + \hat{\boldsymbol{\gamma}} \mathbf{Z}_i))^2$$

$$(\hat{\alpha}, \hat{\beta}, \hat{\boldsymbol{\gamma}}) = \underset{\tilde{\alpha}, \tilde{\beta}, \tilde{\boldsymbol{\gamma}}}{\arg\min} \sum_{i=1}^{n} (Y_i - (\tilde{\alpha} + \tilde{\beta} X_i + \tilde{\boldsymbol{\gamma}} \mathbf{Z}_i))^2$$

Model fit with multiple predictors

- R^2 mechanically increases when you add a variables to the regression.
  - But this could be overfitting!!

Model fit with multiple predictors

- R^2 mechanically increases when you add a variables to the regression.
  - But this could be overfitting!!
- Solution: penalize regression models with more variables.
  - Occam's razor: simpler models are preferred

Model fit with multiple predictors

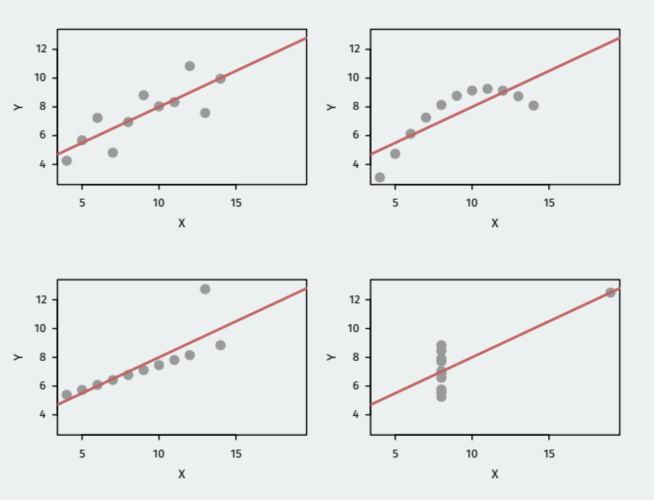
- R^2 mechanically increases when you add a variables to the regression.
  - But this could be overfitting!!
- Solution: penalize regression models with more variables.
  - Occam's razor: simpler models are preferred
- Adjusted R^2: lowers regular R^2 for each additional covariate.
  - If the added covariates doesn't help predict, adjusted R^2 goes down!

Illustrating R-squared's Deficiencies

- What do you notice?

Illustrating R-squared's Deficiencies

- What do you notice? all these graphs have the same R^2



Modeling Difference-in-Differences with regression

- We can develop a linear model that will yield the Difference-in-difference ATT estimator that we encountered in Week 1.

Modeling Difference-in-Differences with regression

- We can develop a linear model that will yield the Difference-in-difference ATT estimator that we encountered in Week 1.

$$Y_i = \beta_0 + \beta_1 P_i + \beta_2 T_i + \beta_3 P_i \times T_i + \epsilon_i \text{ (Population Model)}$$
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 P_i + \hat{\beta}_2 T_i + \hat{\beta}_3 P_i \times T_i \text{ (Estimated)}$$

Modeling Difference-in-Differences with regression

- We can develop a linear model that will yield the Difference-in-difference ATT estimator that we encountered in Week 1.

$$Y_i = \beta_0 + \beta_1 P_i + \beta_2 T_i + \beta_3 P_i \times T_i + \epsilon_i \text{ (Population Model)}$$
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 P_i + \hat{\beta}_2 T_i + \hat{\beta}_3 P_i \times T_i \text{ (Estimated)}$$

- Note P_i is an indicator of period, and T_i is an indicator of whether the unit belongs to a group that is ever treated.

Modeling Difference-in-Differences with regression

- We can develop a linear model that will yield the Difference-in-difference ATT estimator that we encountered in Week 1.

$$Y_i = \beta_0 + \beta_1 P_i + \beta_2 T_i + \beta_3 P_i \times T_i + \epsilon_i \text{ (Population Model)}$$
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 P_i + \hat{\beta}_2 T_i + \hat{\beta}_3 P_i \times T_i \text{ (Estimated)}$$

- Note P_i is an indicator of period, and T_i is an indicator of whether the unit belongs to a group that is ever treated.
- Recall the DiD estimator takes the difference in the outcome among the treated group ($T_i = 1$) between $P_i = 1$ and $P_i = 0$ and subtracts the difference in control group ($T_i = 0$) between $P_i = 1$ and $P_i = 0$.

DiD Estimator Proof

Treated group ($T_i = 1$) at $P_i = 1$ |

DiD Estimator Proof

Treated group ($T_i = 1$) at $P_i = 1$ $\Big|$ $Y_i = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \epsilon_i$

DiD Estimator Proof

Treated group ($T_i = 1$) at $P_i = 1$	$Y_i = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \epsilon_i$
Treated group ($T_i = 1$) at $P_i = 0$	

DiD Estimator Proof

Treated group ($T_i = 1$) at $P_i = 1$	$Y_i = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \epsilon_i$
Treated group ($T_i = 1$) at $P_i = 0$	$Y_i = \beta_0 + \beta_2 + \epsilon_i$

DiD Estimator Proof

Treated group ($T_i = 1$) at $P_i = 1$	$Y_i = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \epsilon_i$
Treated group ($T_i = 1$) at $P_i = 0$	$Y_i = \beta_0 + \beta_2 + \epsilon_i$
Control group ($T_i = 0$) at $P_i = 1$	

DiD Estimator Proof

Treated group ($T_i = 1$) at $P_i = 1$	$Y_i = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \epsilon_i$
Treated group ($T_i = 1$) at $P_i = 0$	$Y_i = \beta_0 + \beta_2 + \epsilon_i$
Control group ($T_i = 0$) at $P_i = 1$	$Y_i = \beta_0 + \beta_1 + \epsilon_i$

DiD Estimator Proof

Treated group ($T_i = 1$) at $P_i = 1$	$Y_i = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \epsilon_i$
Treated group ($T_i = 1$) at $P_i = 0$	$Y_i = \beta_0 + \beta_2 + \epsilon_i$
Control group ($T_i = 0$) at $P_i = 1$	$Y_i = \beta_0 + \beta_1 + \epsilon_i$
Control group ($T_i = 0$) at $P_i = 0$	

DiD Estimator Proof

Treated group ($T_i = 1$) at $P_i = 1$	$Y_i = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \epsilon_i$
Treated group ($T_i = 1$) at $P_i = 0$	$Y_i = \beta_0 + \beta_2 + \epsilon_i$
Control group ($T_i = 0$) at $P_i = 1$	$Y_i = \beta_0 + \beta_1 + \epsilon_i$
Control group ($T_i = 0$) at $P_i = 0$	$Y_i = \beta_0 + \epsilon_i$

DiD Estimator Proof

Treated group ($T_i = 1$) at $P_i = 1$	$Y_i = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \epsilon_i$
Treated group ($T_i = 1$) at $P_i = 0$	$Y_i = \beta_0 + \beta_2 + \epsilon_i$
Control group ($T_i = 0$) at $P_i = 1$	$Y_i = \beta_0 + \beta_1 + \epsilon_i$
Control group ($T_i = 0$) at $P_i = 0$	$Y_i = \beta_0 + \epsilon_i$

$$\text{(Change in Treated Group)} - \text{(Change in Control Group)} =$$
$$(\beta_0 + \beta_1 + \beta_2 + \beta_3 + \epsilon_i - (\beta_0 + \beta_2 + \epsilon_i)) - (\beta_0 + \beta_1 + \epsilon_i - (\beta_0 + \epsilon_i)) = \beta_3$$

DiD Estimator Proof

Treated group ($T_i = 1$) at $P_i = 1$	$Y_i = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \epsilon_i$
Treated group ($T_i = 1$) at $P_i = 0$	$Y_i = \beta_0 + \beta_2 + \epsilon_i$
Control group ($T_i = 0$) at $P_i = 1$	$Y_i = \beta_0 + \beta_1 + \epsilon_i$
Control group ($T_i = 0$) at $P_i = 0$	$Y_i = \beta_0 + \epsilon_i$

$$\text{(Change in Treated Group)} - \text{(Change in Control Group)} =$$
$$(\beta_0 + \beta_1 + \beta_2 + \beta_3 + \epsilon_i - (\beta_0 + \beta_2 + \epsilon_i)) - (\beta_0 + \beta_1 + \epsilon_i - (\beta_0 + \epsilon_i)) = \beta_3$$

- Upshot: β_3 (coefficient on interaction of treatment group indicator and period indicator) is the DiD ATT estimator!

Fixed Effects

- In regression, we can use something called **fixed effects** to control for unobserved characteristics such as ability level in studies of educational policy.

Fixed Effects

- In regression, we can use something called **fixed effects** to control for unobserved characteristics such as ability level in studies of educational policy.
- We often include **time** and **unit** fixed effects to account for time-specific, but unit invariant fixed effects and unit-specific, but time invariant fixed effects, respectively.

Fixed Effects

- In regression, we can use something called **fixed effects** to control for unobserved characteristics such as ability level in studies of educational policy.
- We often include **time** and **unit** fixed effects to account for time-specific, but unit invariant fixed effects and unit-specific, but time invariant fixed effects, respectively.
- Operationally, this means just including a factor variable in your regression that uniquely represents each time period or unit.

Fixed Effects

- In regression, we can use something called **fixed effects** to control for unobserved characteristics such as ability level in studies of educational policy.
- We often include **time** and **unit** fixed effects to account for time-specific, but unit invariant fixed effects and unit-specific, but time invariant fixed effects, respectively.
- Operationally, this means just including a factor variable in your regression that uniquely represents each time period or unit.
- Great way to account for some unobserved potential confounding variables, but often not sufficient!

Fixed Effects

- In regression, we can use something called **fixed effects** to control for unobserved characteristics such as ability level in studies of educational policy.
- We often include **time** and **unit** fixed effects to account for time-specific, but unit invariant fixed effects and unit-specific, but time invariant fixed effects, respectively.
- Operationally, this means just including a factor variable in your regression that uniquely represents each time period or unit.
- Great way to account for some unobserved potential confounding variables, but often not sufficient!

```
lm(y ~ x + as.factor(unit_indicator))

OR

lm(y ~ x + as.factor(time_indicator))
```