Lecture 7: Using Indicator Variables

Shuo Liu

UCLA Summer School Econ 103

July 19, 2017

Outline

1. Predictions in Multiple Regression Model
2. Indicator Variables
3. Log-linear Model
4. Treatment Effect

Predictions in Multiple Regression Model

- Consider the model:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e \tag{1}$$

- Given values of independent variables which are outside the sample data x_{20} and x_{30}, how to predict y_0? suppose the true y_0 follows the theoretical model

$$y_0 = \beta_1 + \beta_2 x_{20} + \beta_3 x_{30} + e \tag{2}$$

- Point estimator is:

$$\hat{y}_0 = b_1 + b_2 x_{20} + b_3 x_{30} \tag{3}$$

- Besides point estimator, we need to provide more information about the precision, i.e. we need the variance of the forecasting error.

Predictions in Multiple Regression Model

- The variance of forecasting error $f = y_0 - \hat{y}_0$ is:

$$\begin{aligned} Var(f) &= Var\left((\beta_1 + \beta_2 x_{20} + \beta_3 x_{30} + e) - (b_1 + b_2 x_{20} + b_3 x_{30})\right) \\ &= Var(e - (b_1 + b_2 x_{20} + b_3 x_{30})) \\ &= Var(e) + Var(b_1) + x_{20}^2 Var(b_2) + x_{30}^2 Var(b_3) \\ &+ 2x_{20} Cov(b_1, b_2) + 2x_{30} Cov(b_1, b_3) + 2x_{20} x_{30} Cov(b_2, b_3) \end{aligned}$$

Predictions in Multiple Regression Model

For the SALES example with explanatory variables as $PRICE$, $ADVERT$ and $ADVERT^2$, suppose $PRICE_0 = 6$, $ADVERT_0 = 1.9$ and $ADVERT_0^2 = 3.61$,

$$\widehat{SALES}_0 = 109.72 - 7.64PRICE_0 + 12.15ADVERT_0 - 2.77ADVERT_0^2 = 76.974 \quad (4)$$

our point estimator is \$76974 (unit of SALES is \$1000)

Predictions in Multiple Regression Model

Table 6.3 Covariance Matrix for Andy's Burger Barn Model

	b_1	b_2	b_3	b_4
b_1	46.227019	−6.426113	−11.600960	2.939026
b_2	−6.426113	1.093988	0.300406	−0.085619
b_3	−11.600960	0.300406	12.646302	−3.288746
b_4	2.939026	−0.085619	−3.288746	0.884774

We can use the above table and $\hat{\sigma}^2 = 21.57865$ to calculate the **estimated variance of forecasting error** and then the standard error.

Predictions in Multiple Regression Model

- Suppose the standard error is $Se(f) = 4.7351$;
- Point estimator is $\widehat{SALES}_0 = 76.974$;
- Then the 95% prediction interval for y_0 is (suppose $n = 74$, $K = 4$):

$$[76.974 - t_{(0.975,70)}Se(f), 76.974 + t_{(0.975,71)}Se(f)] = [76.974 \pm 1.9939 * 4.7351] \tag{5}$$
$$= [67.533, 86.415]$$

We predict, with 95% confidence, that the settings for price and advertising expenditure will yield SALES between \$67533 and \$86415.

- **It is useful to distinguish** between **forecasting SALES** and **estimating average sales** given particular settings of PRICE and ADVERT. To forecast SALES, we need to incorporate the random error e_0 when we calculate the variance of forecasting error; While the average SALES already eliminated the effect of random error, so we ignore it when calculating variance of estimation error.

Predictions in Multiple Regression Model

For true values:

$$E(SALES_0) = \beta_1 + \beta_2 PRICE_0 + \beta_3 ADVERT_0 + \beta_4 ADVERT_0^2 \quad (6)$$

$$SALES_0 = \beta_1 + \beta_2 PRICE_0 + \beta_3 ADVERT_0 + \beta_4 ADVERT_0^2 + e_0 \quad (7)$$

Point estimates:

$$E(\widehat{SALES}_0) = b_1 + b_2 PRICE_0 + b_3 ADVERT_0 + b_4 ADVERT_0^2 \quad (8)$$

$$\widehat{SALES}_0 = b_1 + b_2 PRICE_0 + b_3 ADVERT_0 + b_4 ADVERT_0^2 \quad (9)$$

but when calculating variance of forecasting error, for $SALES_0$, it will include $Var(e_0)$; for $E(SALES_0)$, it will not include $Var(e_0)$.

- Then their estimated variance has a difference, which is the estimator of $Var(e_0) = \hat{\sigma}^2$

$$\hat{Var}(E(\widehat{SALES}_0)) = \hat{Var}(\widehat{SALES}_0) - \hat{\sigma}^2 = \hat{Var}(f) - \hat{\sigma}^2 \quad (10)$$

Predictions in Multiple Regression Model

- Then the standard error for $E(\widehat{SALES}_0)$ is:

$$Se(E(\widehat{SALES}_0)) = \sqrt{\hat{Var}(f) - \hat{\sigma}^2} = 0.9177 \tag{11}$$

Then the 95% interval estimate for $E(SALES_0)$ is:

$$[E(\widehat{SALES}_0) \pm t_{0.975,n-K} * Se(E(\widehat{SALES}_0))] = [75.144, 78.804] \tag{12}$$

which is narrower than that for $SALES_0$: $[67.533, 86.415]$.

Indicator Variables

- Indicator variables are used to account for **qualitative** factors in econometric models, they are often called dummy, binary or dichotomous variables, because they just take **two values**.
- Usually the values are one or zero, to indicate the presence or absence of a certain characteristic.

$$D = \begin{cases} 1 & \text{if characteristic is present} \\ 0 & \text{if characteristic is not present} \end{cases} \tag{13}$$

Indicator Variables

Consider a model to predict the value of a house as a function of its characteristics: size, location, number of bedrooms and age. Add one indicator variable to the model,

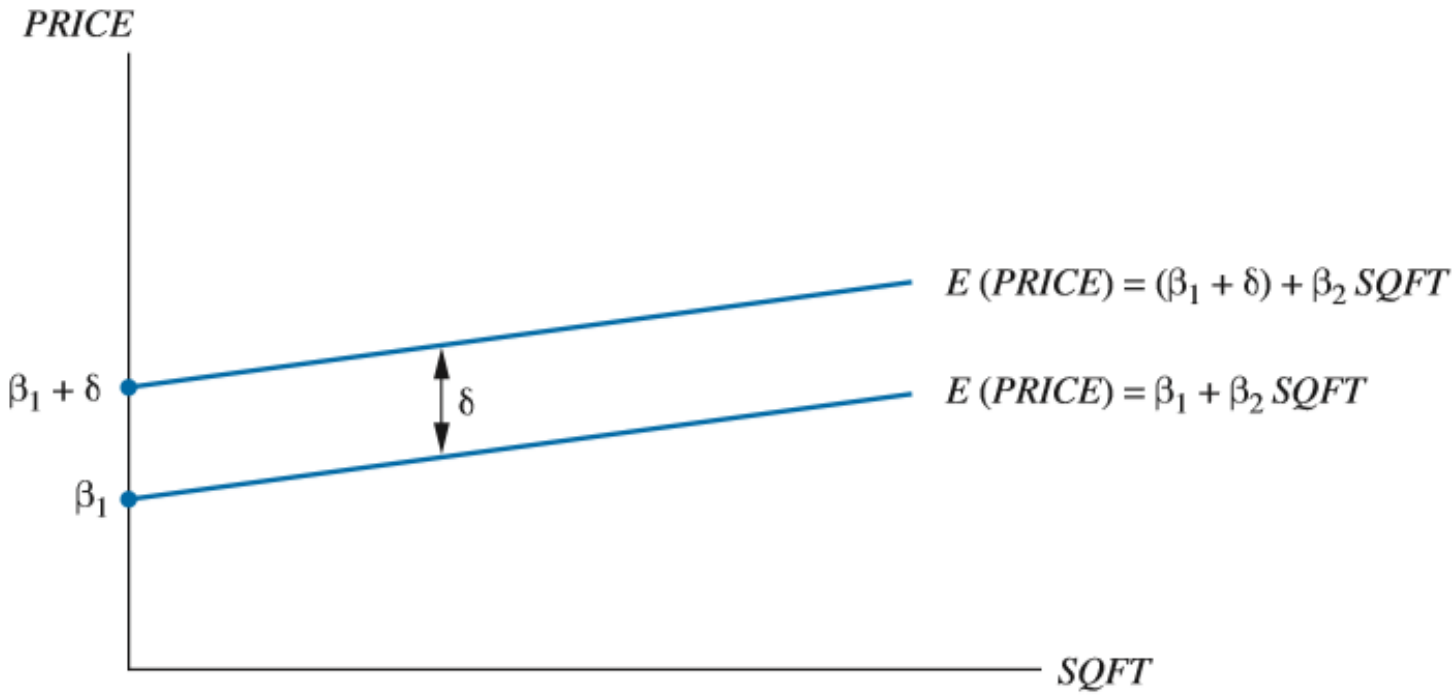
$$D = \begin{cases} 1 & \text{if property is in a desirable neighborhood} \\ 0 & \text{if property is NOT in a desirable neighborhood} \end{cases} \tag{14}$$

$$PRICE = \beta_1 + \delta D + \beta_2 SQFT + e \tag{15}$$

then our model implies that:

- $$E(PRICE) = \begin{cases} (\beta_1 + \delta) + \beta_2 SQFT & \text{when } D = 1 \\ \beta_1 + \beta_2 SQFT & \text{when } D = 0 \end{cases} \tag{16}$$
- Adding an indicator variable causes a parallel shift in the expected price by the amount δ.
- The properties of OLS estimators are not affected by the fact there exist indicator variables as explanatory variables.

Indicator Variables



Indicator Variables

- The value $D = 0$ defines the **reference group** or the **base group**;
- We could pick any group as the base group. For example, if we define:

$$LD = \begin{cases} 1 & \text{if property is NOT in a desirable neighborhood} \\ 0 & \text{if property is in a desirable neighborhood} \end{cases} \tag{17}$$

then, correspondingly, the model would be:

$$PRICE = \beta_1 + \lambda LD + \beta_2 SQFT + e \tag{18}$$

- What would happen if we include both D and LD in the regression? Now the model is:

$$PRICE = \beta_1 + \delta D + \lambda LD + \beta_2 SQFT + e \tag{19}$$

where $D + LD = 1$ applies for every data point, this is the problem of "exact collinearity".

- **So for two-type characteristic,** we can only include one indicator variable into the model.

Indicator Variables

- Now consider the model with interaction variable:

$$PRICE = \beta_1 + \beta_2 SQFT + \gamma(SQFT \times D) + e \tag{20}$$

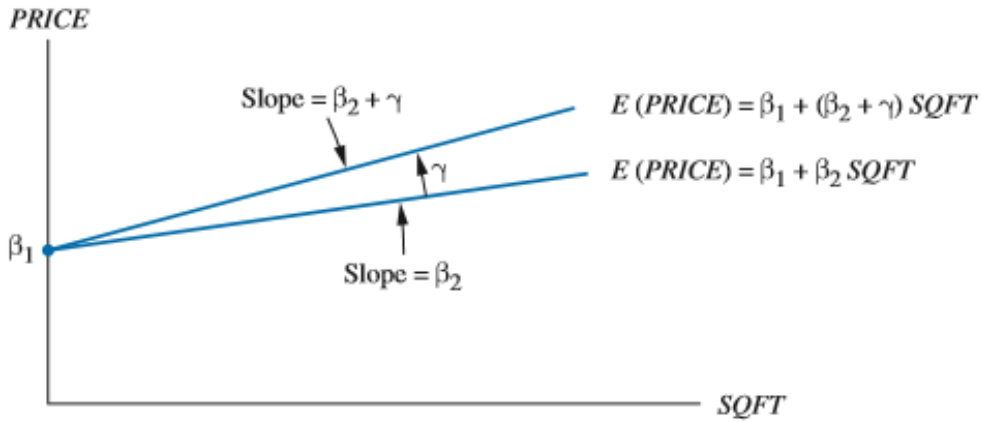
- The new variable $(SQFT \times D)$ is the product of house size and the indicator variable and it captures the interaction effect of location and size on house price.
- Now the expected price is as follows:

$$\begin{aligned} E(PRICE) &= \beta_1 + \beta_2 SQFT + \gamma(SQFT \times D) \\ &= \begin{cases} \beta_1 + (\beta_2 + \gamma)SQFT & \text{when } D = 1 \\ \beta_1 + \beta_2 SQFT & \text{when } D = 0 \end{cases} \end{aligned}$$

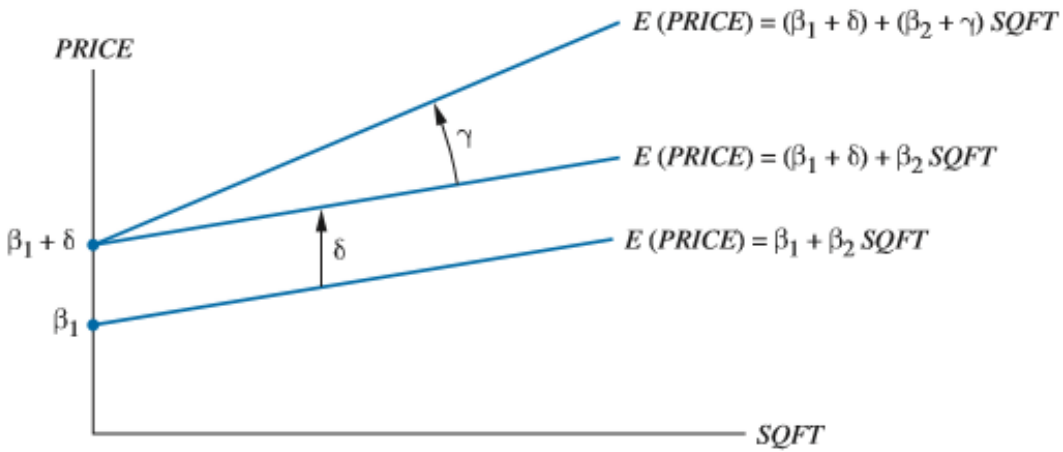
- The slope can be expressed as:

$$\frac{\partial E(PRICE)}{\partial SQFT} = \begin{cases} \beta_2 + \gamma & \text{when } D = 1 \\ \beta_2 & \text{when } D = 0 \end{cases} \tag{21}$$

Indicator Variables



(a)



(b)

FIGURE 7.2 (a) A slope-indicator variable. (b) Slope- and intercept-indicator variables.

Indicator Variables

What if the house location affects both the intercept and the slope?

- We can incorporate both effects into the model:

$$PRICE = \beta_1 + \delta D + \beta_2 SQFT + \gamma(SQFT \times D) + e \tag{22}$$

- Now the expected price is as follows:

$$E(PRICE) = \begin{cases} (\beta_1 + \delta) + (\beta_2 + \gamma)SQFT & \text{when } D = 1 \\ \beta_1 + \beta_2 SQFT & \text{when } D = 0 \end{cases} \tag{23}$$

- Suppose we use UTOWN as a specific example of the indicator D, and specifies a regression equation for house prices:

$$PRICE = \beta_1 + \delta_1 UTOWN + \beta_2 SQFT + \gamma(SQFT \times UTOWN) + \beta_3 AGE \tag{24}$$
$$+\delta_2 POOL + \delta_3 FPLACE + e$$

Indicator Variables

Table 7.1 Representative Real Estate Data Values

PRICE	*SQFT*	*AGE*	*UTOWN*	*POOL*	*FPLACE*
205.452	23.46	6	0	0	1
185.328	20.03	5	0	0	1
248.422	27.77	6	0	0	0
287.339	23.67	28	1	1	0
255.325	21.30	0	1	1	1
301.037	29.87	6	1	0	1

Indicator Variables

Table 7.2 House Price Equation Estimates

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	24.5000	6.1917	3.9569	0.0001
UTOWN	27.4530	8.4226	3.2594	0.0012
SQFT	7.6122	0.2452	31.0478	0.0000
SQFT×*UTOWN*	1.2994	0.3320	3.9133	0.0001
AGE	−0.1901	0.0512	−3.7123	0.0002
POOL	4.3772	1.1967	3.6577	0.0003
FPLACE	1.6492	0.9720	1.6968	0.0901

$R^2 = 0.8706$ $SSE = 230184.4$

- The estimated regression model for a house near the university (UTOWN=1) is:

$$\begin{aligned}\widehat{PRICE} &= (24.5 + 27.453) + (7.6122 + 1.2994)SQFT - 0.1901AGE \\ &\quad + 4.3772POOL + 1.6492FPLACE \\ &= 51.953 + 8.9116SQFT - 0.1901AGE \\ &\quad + 4.3772POOL + 1.6492FPLACE\end{aligned}$$

Indicator Variables

- For a house in another area (outside university town) we have:

$$\widehat{PRICE} = 24.5 + 7.6122SQFT - 0.1901AGE + 4.3772POOL + 1.6492FPLACE$$

- The results indicate that:
  1. The **location premium** for house near the university is \$27453;
  2. The change in expected price per additional square foot is \$89.12 for house near the university and \$76.12 for houses in other areas;
  3. House depreciates \$190.1 per year;
  4. A pool increases the value of a house by \$4377.2;
  5. A replace increases the value of a house by \$1649.2.

Indicator Variables

Another WAGE example for applying indicator variables:

- Consider the wage equation:

$$WAGE = \beta_1 + \beta_2 EDUC + \delta_1 BLACK + \delta_2 FEMALE \qquad (25)$$
$$+\gamma(BLACK \times FEMALE) + e$$

- Consequently, the expected value is:

$$E(WAGE) = \begin{cases} \beta_1 + \beta_2 EDUC & \text{for WHITE-MALE} \\ (\beta_1 + \delta_1) + \beta_2 EDUC & \text{for BLACK-MALE} \\ (\beta_1 + \delta_2) + \beta_2 EDUC & \text{for WHITE-FEMALE} \\ (\beta_1 + \delta_1 + \delta_2 + \gamma) + \beta_2 EDUC & \text{for BLACK-FEMALE} \end{cases} \qquad (26)$$

Indicator Variables

Table 7.3 **Wage Equation with Race and Gender**

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	−5.2812	1.9005	−2.7789	0.0056
EDUC	2.0704	0.1349	15.3501	0.0000
BLACK	−4.1691	1.7747	−2.3492	0.0190
FEMALE	−4.7846	0.7734	−6.1863	0.0000
BLACK × *FEMALE*	3.8443	2.3277	1.6516	0.0989

$R^2 = 0.2089$ $SSE = 130194.7$

- What if we want to check whether race and/or gender affect significantly the wage equation?
- Then we come to the following joint hypothesis testing:

$$H_0 : \delta_1 = \delta_2 = \gamma = 0 \tag{27}$$

$$H_1 : \textbf{at least one of } \delta_1, \delta_2, \gamma \text{ not equal to zero} \tag{28}$$

Indicator Variables

- Recall that the F statistic for a joint hypothesis is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(n-K)} \tag{29}$$

  Here $J = 3$, we use $SSE_U = 130194.7$ from Table 7.3.
- The SSE_R comes from the fitted restricted model:

$$\widehat{WAGE} = -6.7103 + 1.9803EDUC \tag{30}$$

  for which $SSE_R = 135771.1$.
- Then we calculate $F = 14.21$. The 1% critical value is $F_{(0.99,3,995)} = 3.8$. Since $F > F_c$, we reject the H_0 and accept H_1, that is, race and/or gender affect significantly the wage equation.

Indicator Variables

- We can incorporate both effects into the model:

$$PRICE = \beta_1 + \delta D + \beta_2 SQFT + \gamma(SQFT \times D) + e \tag{31}$$

- Now the expected price is as follows:

$$E(PRICE) = \begin{cases} (\beta_1 + \delta) + (\beta_2 + \gamma)SQFT & \text{when } D = 1 \\ \beta_1 + \beta_2 SQFT & \text{when } D = 0 \end{cases} \tag{32}$$

- **Comments:** by introducing both intercept and slope-indicator variables we have essentially assumed that the regressions in the two neighborhoods (near the university or far away from university) are completely different. We can also estimate (31) by estimating two separate equations using two subsets of data.

Indicator Variables

Remarks on the F-test:

- The usual F-test of a joint hypothesis relies on the assumptions MR1-MR6 of the multiple regression model;
- Of particular relevance for testing the equivalence of two regressions is the **assumption MR3 (homoskedasticity)**, that the variance of the error term, $Var(e_i) = \sigma^2$, is the same for all observations i.
- If we are considering possibly different slopes and intercepts for parts of the data, it might also be true that the error variances are different in the two parts of the data, i.e. assumption MR3 is violated.
- It is important to note that in such a case, the usual F-test is not valid.

Indicator Variables

Remarks on indicator variable: for depicting an "M-type" characteristic, usually we need $M-1$ indicator variables.

For example,

- Seasonal factor: $D_1 = 1(Spring), D_1 \neq 0$(not Spring); $D_2 = 1(Summer), D_2$ 0(not Summer); $D_3 = 1(Fall), D_3 \neq 0$(not Fall).
- What if we have $D_4 = 1(Winter), D_4 \neq 0$(not Winter)? Then by incorporating $D_1 - D_4$ into the model, we have the problem of exact collinearity, since $D_1 + D_2 + D_3 + D_4 = 1$ for each data point.

Log-linear Model

- Consider the wage equation in log-linear form:

$$ln(WAGE) = \beta_1 + \beta_2 EDUC + \delta FEMALE + e \tag{33}$$

- What is the interpretation of δ?
- Expanding the model, we have:

$$E(ln(WAGE)) = \begin{cases} \beta_1 + \beta_2 EDUC & \text{for male } (FEMALE = 0) \\ (\beta_1 + \delta) + \beta_2 EDUC & \text{for female } (FEMALE = 1) \end{cases} \tag{34}$$

- So that the difference between females and males is

$$E(ln(WAGE))|_{FEMALE} - E(ln(WAGE))|_{MALE} = \delta \tag{35}$$

- That is, δ is approximately the percentage difference between females and males.

Log-linear Model

- The estimated model is:

$$\widehat{ln(WAGE)} = 1.6539 + 0.0962EDUC - 0.2432FEMALE \tag{36}$$

- So we estimate that there is a 24.32% differential between male and female wages.
- For a better calculation, the wage difference is:

$$E(ln(WAGE))|_{FEMALE} - E(ln(WAGE))|_{MALE} \tag{37}$$

$$= E\left(ln(\frac{WAGE|_{FEMALE}}{WAGE|_{MALE}})\right) = \delta$$

- Then we have:

$$\frac{WAGE|_{FEMALE}}{WAGE|_{MALE}} = e^{\delta} \tag{38}$$

Log-linear Model

- Subtracting 1 from both sides:

$$\frac{WAGE|_{FEMALE}}{WAGE|_{MALE}} - \frac{WAGE|_{MALE}}{WAGE|_{MALE}} = \frac{WAGE|_{FEMALE} - WAGE|_{MALE}}{WAGE|_{MALE}}$$
$$= e^{\delta} - 1$$

- The percentage difference between wages of females and males is $100(e^{\delta} - 1)\%$.
- Plugging in the estimated $\hat{\delta} = -0.2432$, we estimate the wage differential between males and females to be:

$$100(e^{\delta} - 1)\% = 100(e^{-0.2432} - 1)\% = -21.59\% \tag{39}$$

Treatment Effect

If we want to analyze the effect of one "treatment", we must perform a **randomized controlled experiment**.

- **random sample**: formed by **randomly assigning** subjects to treatment and control (not receive the treatment) groups.

For example:

- Do hospitals make people healthier/less healthy? : we use sample data containing people not going to hospitals (control group) and people going to the hospitals (experiment group), **but this data exhibits a selection bias**, because the treatment group is not formed randomly, but is determined by (the subjects') choice.
- How to perform a random experiment in this case?

Treatment Effect

Other Examples:

- How much does an additional year of education increase the wages of married women? (sample data on working women's wages is determined by women's choice to join the labor force)
- How much does participation in a job-training program increase wages? (is the treatment group formed randomly or non-randomly?)
- How much does a dietary supplement contribute to weight loss? (how to perform a randomized controlled experiment in this case?)

Treatment Effect

- In general, selection bias interferes with a straightforward examination of the data, and makes more difficult for us to measure a causal effect. (**Remember correlation does not imply causation**)
- We would like to randomly assign subjects to a treatment group, with others being treated as a control group. (treatment group/control group is not determined by subjects' choices)
- However, the ability to perform randomized controlled experiments in economics is limited because the subjects are people, and their economic well-being is at stake.

Treatment Effect

1. Difference Estimator: consider a simple regression model in which the explanatory variable is an indicator variable, indicating whether a particular individual is in the treatment or control group.

- Define the indicator variable d_i as:

$$d_i = \begin{cases} 1 & \text{individual in treatment group} \\ 0 & \text{individual in control group} \end{cases} \tag{40}$$

- The model is then:

$$y_i = \beta_1 + \beta_2 d_i + e_i, \quad i = 1, 2, \cdots, n \tag{41}$$

where e_i represents collection of other factors affecting the outcome.

- And the regression functions are:

$$E(y_i) = \begin{cases} \beta_1 + \beta_2 & \text{individual in treatment group} \\ \beta_1 & \text{individual in control group} \end{cases} \tag{42}$$

Treatment Effect

- The least squares estimator for β_2, the treatment effect, is:

$$b_2 = \frac{\sum_{i=1}^{n}(d_i - \bar{d})(y_i - \bar{y})}{\sum_{i=1}^{n}(d_i - \bar{d})^2} = \bar{y}_1 - \bar{y}_0 \tag{43}$$

  where

$$\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i, \quad \bar{y}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} y_i \tag{44}$$

- The estimator b_2 is called the difference estimator, because it is the difference between the sample means of the treatment and control groups.

Treatment Effect

- The difference estimator can be rewritten as:

$$b_2 = \beta_2 + \frac{\sum_{i=1}^{n}(d_i - \bar{d})(e_i - \bar{e})}{\sum_{i=1}^{n}(d_i - \bar{d})^2} = \beta_2 + \bar{e}_1 - \bar{e}_0 \tag{45}$$

- For the treatment effect to be unbiased we must have:

$$E(b_2) - \beta_2 = E(\bar{e}_1 - \bar{e}_0) = E(\bar{e}_1) - E(\bar{e}_0) = 0 \tag{46}$$

- If we allow individuals to self-select themselves into the treatment and control groups, then:

$$E(\bar{e}_1) - E(\bar{e}_0) \neq 0 \tag{47}$$

  is the selection bias in the estimation of the treatment effect
- We can eliminate the self-selection bias by randomly assigning individuals to the treatment and control groups
- Doing so would guarantee that there are no systematic differences between the groups, except for the treatment itself

Treatment Effect

Application of Difference Estimator: (need to guarantee a randomized controlled experiment)

How does the class size affect the student performance?

Table 7.6a **Summary Statistics for Regular-Sized Classes**

Variable	Mean	Std. Dev.	Min	Max
TOTALSCORE	918.0429	73.1380	635	1229
SMALL	0.0000	0.0000	0	0
TCHEXPER	9.0683	5.7244	0	24
BOY	0.5132	0.4999	0	1
FREELUNCH	0.4738	0.4994	0	1
WHITE_ASIAN	0.6813	0.4661	0	1
TCHWHITE	0.7980	0.4016	0	1
TCHMASTERS	0.3651	0.4816	0	1
SCHURBAN	0.3012	0.4589	0	1
SCHRURAL	0.4998	0.5001	0	1

$N = 2005$

Treatment Effect

Table 7.6b Summary Statistics for Small Classes

Variable	Mean	Std. Dev.	Min	Max
TOTALSCORE	931.9419	76.3586	747	1253
SMALL	1.0000	0.0000	1	1
TCHEXPER	8.9954	5.7316	0	27
BOY	0.5150	0.4999	0	1
FREELUNCH	0.4718	0.4993	0	1
WHITE_ASIAN	0.6847	0.4648	0	1
TCHWHITE	0.8625	0.3445	0	1
TCHMASTERS	0.3176	0.4657	0	1
SCHURBAN	0.3061	0.4610	0	1
SCHRURAL	0.4626	0.4987	0	1

$N = 1738$

Treatment Effect

Models of Interest (starting from (49), we add additional controls/explan variables):

$$TOTALSCORE = \beta_1 + \beta_2 SMALL + e \tag{48}$$

$$TOTALSCORE = \beta_1 + \beta_2 SMALL + \beta_3 TCHEXPER + e \tag{49}$$

Table 7.7 Project STAR: Kindergarden

	(1)	(2)	(3)	(4)
C	918.0429***	907.5643***	917.0684***	908.7865***
	(1.6672)	(2.5424)	(1.4948)	(2.5323)
SMALL	13.8990***	13.9833***	15.9978***	16.0656***
	(2.4466)	(2.4373)	(2.2228)	(2.2183)
TCHEXPER		1.1555***		0.9132***
		(0.2123)		(0.2256)
SCHOOL EFFECTS	*No*	*No*	*Yes*	*Yes*
N	3743	3743	3743	3743
adj. R^2	0.008	0.016	0.221	0.225
SSE	20847551	20683680	16028908	15957534

Standard errors in parentheses
Two-tail *p*-values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Treatment Effect

Further, add controls of school fixed effects:

- The students in our sample are enrolled in 79 different schools;
- One way to account for school effects is to include an indicator variable for **each school** (in all 78 indicator variables);
- That is, we can introduce 78 new indicators:

$$SCHOOL_j = \begin{cases} 1 & \text{if student is in school j} \\ 0 & \text{Otherwise} \end{cases} \tag{50}$$

- The model becomes:

$$TOTALSCORE = \beta_1 + \beta_2 SMALL + \beta_3 TCHEXPER + \sum_{j=2}^{79} \delta_j SCHOOL_j + e \tag{51}$$

- The regression function for a student i in school j is:

$$E(TOTALSCORE_i) = \tag{52}$$

$$\begin{cases} (\beta_1 + \delta_j) + \beta_3 TCHEXPER_i & \text{if student is in regular class} \\ (\beta_1 + \delta_j + \beta_2) + \beta_3 TCHEXPER_i & \text{if student is in small class} \end{cases}$$

Treatment Effect

- The results for the models with school fixed effects are in columns (3) and (4) of the Table 7.7

Table 7.7 Project STAR: Kindergarden

	(1)	(2)	(3)	(4)
C	918.0429***	907.5643***	917.0684***	908.7865***
	(1.6672)	(2.5424)	(1.4948)	(2.5323)
SMALL	13.8990***	13.9833***	15.9978***	16.0656***
	(2.4466)	(2.4373)	(2.2228)	(2.2183)
TCHEXPER		1.1555***		0.9132***
		(0.2123)		(0.2256)
SCHOOL EFFECTS	*No*	*No*	*Yes*	*Yes*
N	3743	3743	3743	3743
adj. R^2	0.008	0.016	0.221	0.225
SSE	20847551	20683680	16028908	15957534

Standard errors in parentheses
Two-tail p-values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Treatment Effect

- A way to check for random assignment is to **regress SMALL on the available characteristics** and check for any significant coefficients, or an overall significant relationship;
- If there is random assignment, we should not find any significant relationships;
- Because SMALL is an indicator variable, we use the linear probability model.

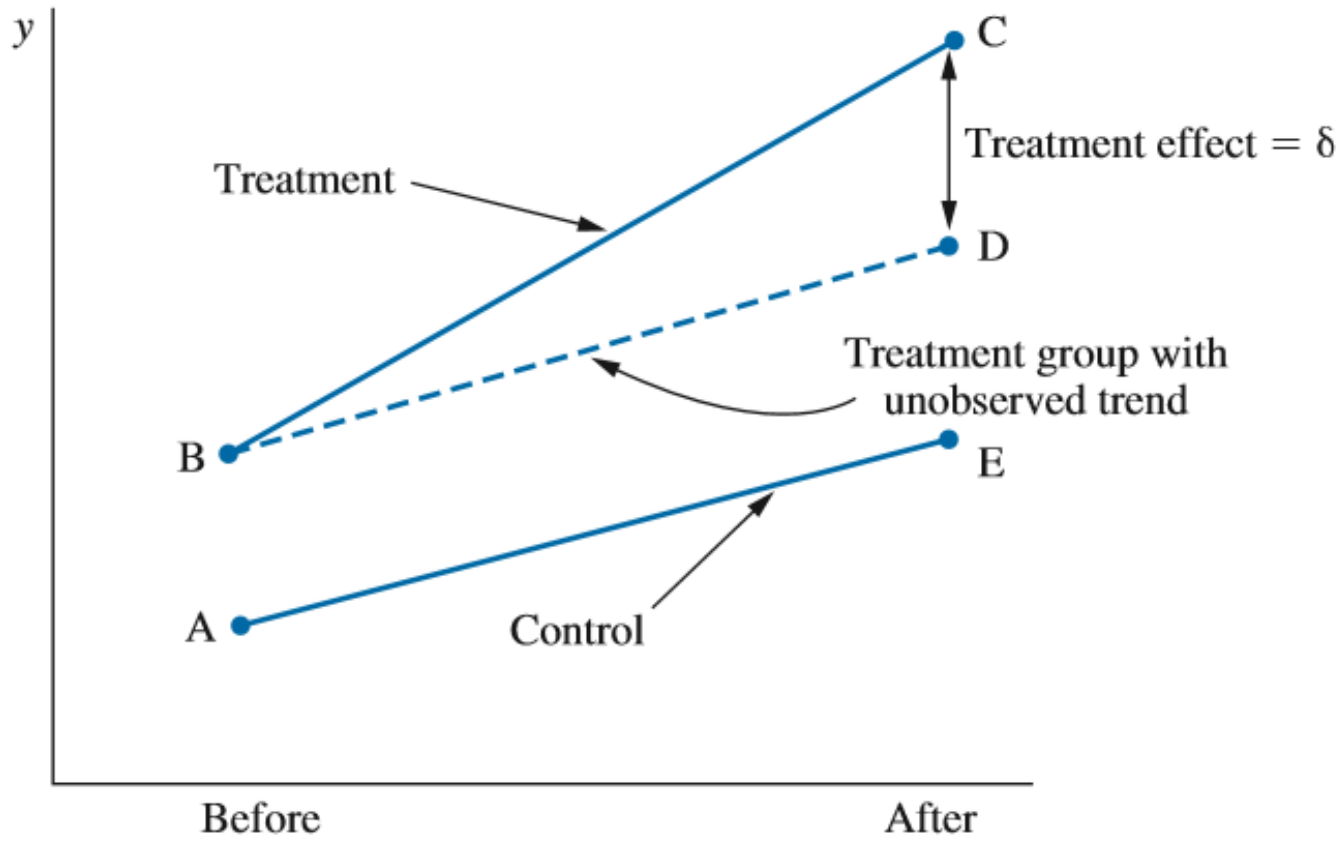
$$SMALL = \alpha_1 + \alpha_2 BOY + \alpha_3 WHITEASIAN + \alpha_4 TCHEXPER \quad (53)$$
$$+\alpha_5 FREELUNCH + e$$

2. Difference-in-Difference Estimator

- Randomized controlled experiments are rare in economics because they are expensive and involve human subjects
- Natural experiments, also called quasi-experiments, rely on observing real-world conditions that approximate what would happen in a randomized controlled experiment
- In these cases, treatment appears as if it were randomly assigned
- **We consider estimating the treatment effect using "before and after" data.**

Treatment Effect

We observe two groups before and after a policy change, with the **treatment group** being affected by the policy and the **control group** being unaffected by the policy.



Treatment Effect

- Estimation of the treatment effect is based on data averages for the two groups in the two periods

$$\hat{\delta} = (\hat{C} - \hat{E}) - (\hat{B} - \hat{A}) \tag{54}$$

$$= (\bar{y}_{Treatment,After} - \bar{y}_{Control,After}) - (\bar{y}_{Treatment,Before} - \bar{y}_{Control,Before})$$

- The estimator $\hat{\delta}$ is called a difference-in-difference (abbreviated as DD) estimator of the treatment effect.

Treatment Effect

More formally, consider the model:

$$y_{it} = \beta_1 + \beta_2 TREAT_i + \beta_3 AFTER_t + \beta_4 (TREAT_i \times AFTER_t) + e_{it} \quad (55)$$

- The regression function is:

$$E(y_{it}) = \begin{cases} \beta_1 & \text{TREAT=0,AFTER=0,Group A} \\ \beta_1 + \beta_2 & \text{TREAT=1,AFTER=0,Group B} \\ \beta_1 + \beta_3 & \text{TREAT=0,AFTER=1,Group E} \\ \beta_1 + \beta_2 + \beta_3 + \beta_4 & \text{TREAT=1,AFTER=1,Group C} \end{cases}$$

Treatment Effect

- Using the points in the figure:

$$\delta = (C - E) - (B - A) \tag{56}$$

$$= [(\beta_1 + \beta_2 + \beta_3 + \beta_4) - (\beta_1 + \beta_3)] - [(\beta_1 + \beta_2) - (\beta_1)]$$

- Using the least squares estimates, we have:

$$\hat{\delta} = [(b_1 + b_2 + b_3 + b_4) - (b_1 + b_3)] - [(b_1 + b_2) - (b_1)] \tag{57}$$

$$= (\bar{y}_{Treatment,After} - \bar{y}_{Control,After}) - (\bar{y}_{Treatment,Before} - \bar{y}_{Control,Before})$$

Read the sections 7.56 and 7.57