PPGCC
MESTRADO E DOUTORADO EM CONTABILIDADE
USP | RIBEIRÃO PRETO

RCC4703
Análise de Regressão

Prof. Dr. **Marcelo Botelho** da Costa Moraes
mbotelho@usp.br

Turma: 3º Trimestre / 2020

Agenda – Aula 5/8

- Programação R
- Diferenças em Diferenças
- Efeitos de Moderação
- Propensity Score Matching

Aplicação em Softwares

R Studio – Cross Section e Dados em Painel

- Antes de começar

```
# Instalando Pacotes (só precisa instalar uma única vez)
install.packages("stargazer")
install.packages("dplyr")
install.packages("ggpubr")
install.packages("rstatix")
install.packages("psych")
install.packages("plm")
install.packages("AER")
install.packages("dynpanel")
```

R Studio – Cross Section e Dados em Painel

```
# Carregando Pacotes (tem que carregar todas as vezes)
library(xlsx)
library(stargazer)
library(dplyr)
library(ggpubr)
library(rstatix)
library(psych)
library(stargazer)
library(lmtest)
library(sandwich)
library(car)
library(plm)
library(lmtest)
library(AER)
library(dynpanel)
```

R Studio – Cross Section e Dados em Painel

Importando dados do Excel – ajuste o endereço de onde a planilha está no seu computador

Caixa <- read.xlsx("C:/Users/mbote/Documents/Aula4_SaldoCaixa.xlsx", 1) # importar a primeira planilha

R Studio – Cross Section e Dados em Painel

- Estatística Descritiva
- Comando: **Estat_Descritiva <- summary(Caixa)**

- Teste de Normalidade
- Comando: **shapiro_test(Caixa$CASH)**
- Comando múltiplas variáveis: **Caixa %>% shapiro_test(DIVDUMMY, ALAVANC, TAM, INV, ENDIV, MATDIV, ATIVLIQ, FLCX, CAPEX, GOVDUMMY, FINBNDES, IFRS, CASH)**

- Winsorização
- Comando: **wCaixa <- winsor(Caixa, trim=0.01, na.rm=TRUE)**
- **wCaixa <- cbind(Caixa$EMPRESA, Caixa$ANO, wCaixa$DIVDUMMY, wCaixa$ALAVANC, wCaixa$TAM, wCaixa$INV, wCaixa$ENDIV, wCaixa$MATDIV, wCaixa$ATIVLIQ, wCaixa$FLCX, wCaixa$CAPEX, wCaixa$GOVDUMMY, wCaixa$FINBNDES, wCaixa$IFRS, wCaixa$CASH)**
- **colnames(wCaixa) <- c("EMPRESA", "ANO", "DIVDUMMY", "ALAVANC", "TAM", "INV", "ENDIV", "MATDIV", "ATIVLIQ", "FLCX", "CAPEX", "GOVDUMMY", "FINBNDES", "IFRS", "CASH")**

R Studio – Cross Section e Dados em Painel

- Regressão Linear

Modelo de Regressão Linear

wCaixa <- as.data.frame(wCaixa)

CASH = b0 + betas [DIVDUMMY, ALAVANC, TAM, INV, ENDIV, MATDIV, ATIVLIQ, FLCX, CAPEX, GOVDUMMY, FINBNDES, IFRS]

- **tabela.modeloCash <- wCaixa[c(15, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14)]**
- **formula.Cash <- formula(tabela.modeloCash)**
- **lm.Cash <- lm(data = tabela.modeloCash, formula = formula.Cash)**

R Studio – Cross Section e Dados em Painel

Reportando os Resultados

- Comando: **summary(lm.Cash)**
- Para opção melhor formatada:

stargazer(lm.Cash, type = 'text', font.size = 'small', title = 'Estimativa de Cash por OLS',

dep.var.labels=c('Caixa'),

omit.stat = c('ser'), ci=TRUE, ci.level=0.90, single.row=TRUE)

R Studio – Cross Section e Dados em Painel

- Verificar Heterocedasticidade
- Comando: **bptest(lm.Cash)**
- Se rejeitar H0, deve refazer o modelo com erro padrão
- Caso necessite modelo Robusto:
- Comando: **coeftest(lm.Cash, vcov = vcovHC(lm.Cash, "HC1"))** # HC1 gives us the White standard errors

R Studio – Cross Section e Dados em Painel

- Multicolinearidade
- Comando: **vif(lm.Cash)**

- Gráfico dos Resíduos
- Comando:

```
lm.Cash.res <- resid(lm.Cash)
plot(lm.Cash.res, ylab="Resíduos", xlab="Casos", main="Gráfico dos Resíduos")
```

R Studio – Cross Section e Dados em Painel

- Análise dos Resíduos (Regressão Múltipla e POLS)
- **summary(lm.Cash.res)**
- **shapiro_test(lm.Cash.res)**
- **cor(tabela.modeloCash, lm.Cash.res)**
- **cor(tabela.modeloCash, lm.Cash.res, method = "spearman") # caso de não normalidade**

R Studio – Dados em Painel

- Dados Balanceados?
- Comando: **is.pbalanced(wCaixa)**

- Efeitos Fixos
- Comando:

Cash.fixed <- plm(formula = formula.Cash, data = as.data.frame(wCaixa), index = c("EMPRESA","ANO"), model = "within")

summary(Cash.fixed)

fixef(Cash.fixed) #constante por empresa

pFtest(Cash.fixed, lm.Cash) #teste para efeitos fixos vs. pooled

R Studio – Dados em Painel

- Teste para heterocedasticidade:
- Comando: **bptest(formula.Cash, data = wCaixa, studentize=F)**

- Efeitos Fixos com erro-padrão Robusto
- Comando:

coeftest(Cash.fixed) #coeficientes originais

coeftest(Cash.fixed, vcovHC) #coeficientes robustos

R Studio – Dados em Painel

- Efeitos Aleatórios

Cash.random <- plm(formula = formula.Cash, data = as.data.frame(wCaixa), index = c("EMPRESA","ANO"), model = "random")

summary(Cash.random)

- Efeitos Aleatórios com erro-padrão Robusto
- Comando:

coeftest(Cash.random) #coeficientes originais

coeftest(Cash.random, vcovHC) #coeficientes robustos

R Studio – Dados em Painel

- Teste Breusch-Pagan
- Comando:

Cash.pooled <- plm(formula = formula.Cash, data = as.data.frame(wCaixa), index = c("EMPRESA","ANO"), model = "pooling")

summary(Cash.pooled)

plmtest(Cash.pooled, type=c("bp"))

- Teste Hausman
- Comando: **phtest(Cash.fixed, Cash.random)**

R Studio – Variáveis Instrumentais

- Variáveis Instrumentais
- Comando:

Cash.iv = ivreg(CASH ~ TAM + ENDIV | ENDIV + FLCX + CAPEX, data = wCaixa)

summary(Cash.iv, vcov = sandwich, diagnostics = TRUE)

R Studio – Painel Dinâmico

- Painel Dinâmico
- Comando:

Cash.dyn <-
dpd(formula.Cash,wCaixa,index=c("EMPRESA", "ANO"),1,1) #número de lags e tipo: 0 → IV, 1 → menor número de instrumentos, 2 → IV método 3, 3 → número de momentos crescentes

summary(Cash.dyn)

R Studio – Análise Discriminante

- Amostra = 91 empresas
  - 49 classificadas como Insolventes Concordatárias → CLASS_Y = 0
  - 42 classificadas como Solventes→ CLASS_Y = 1
- Variáveis:
  - LS – Liquidez Seca = (AC-Est)/PC
  - GA – Giro do Ativo = Receita/AT
  - Rep_EST – Representatividade do Estoque = EST/AT
  - Rep_PC – Representatividade do Passivo Circulante = PC/PT
  - EST_CUSTO – Estoque a preço de custo = estoque/custo
  - FORN_VEN – Relação fornecedores x Receita de Vendas = Fornec/Receita

R Studio – Análise Discriminante

- Análise Discriminante

library(MASS)

Comando: **modelo.lda <- lda(CLASS_Y~., data = Aula4_PrevisaoFalencia)**

plot(modelo.lda) #gráfico grupos

predictions <- modelo.lda %>%
predict(Aula4_PrevisaoFalencia) #previsões

mean(predictions$class==Aula4_PrevisaoFalencia$CLASS_Y)
#acerto do modelo

R Studio – Análise Discriminante

classes previstas

head(predictions$class, 6)

previsão de classes posterior

head(predictions$posterior, 6)

modelo discriminante

head(predictions$x, 3)

os números representam a quantidade demonstrada

R Studio – Análise Discriminante

- Matriz de Confusão

install.packages("e1071")

library(pROC)

library(caret)

library(e1071)

lda.pred <- ifelse(predictions$class=="1", "Solvente", "Insolvente")

Class_Y <- ifelse(Aula4_PrevisaoFalencia$CLASS_Y > 0.5, "Solvente", "Insolvente")

CM.lda <- table(lda.pred, Class_Y)

confusionMatrix(CM.lda)

R Studio – Regressão Logística

- Regressão Logística
- Comando:

modelo.log <- glm(CLASS_Y~., data = Aula4_PrevisaoFalencia, family = binomial)

summary(modelo.log)

glm.probs <- predict(modelo.log,type = "response") #previsão

glm.pred <- ifelse(glm.probs > 0.5, “Solvente", “Insolvente") #classificação

table(glm.pred, Aula4_PrevisaoFalencia$CLASS_Y)

R Studio – Regressão Logística

- Matriz de Confusão

Class_Y <- ifelse(Aula4_PrevisaoFalencia$CLASS_Y > 0.5, "Solvente", "Insolvente")

CM.log <- table(glm.pred, Class_Y)

confusionMatrix(CM.log)

R Studio – Regressão Logística

- Curva ROC

roc <- roc(Aula4_PrevisaoFalencia$CLASS_Y, glm.probs)

Diferenças em Diferenças

Differences-in-Differences

Diff-in-Diff

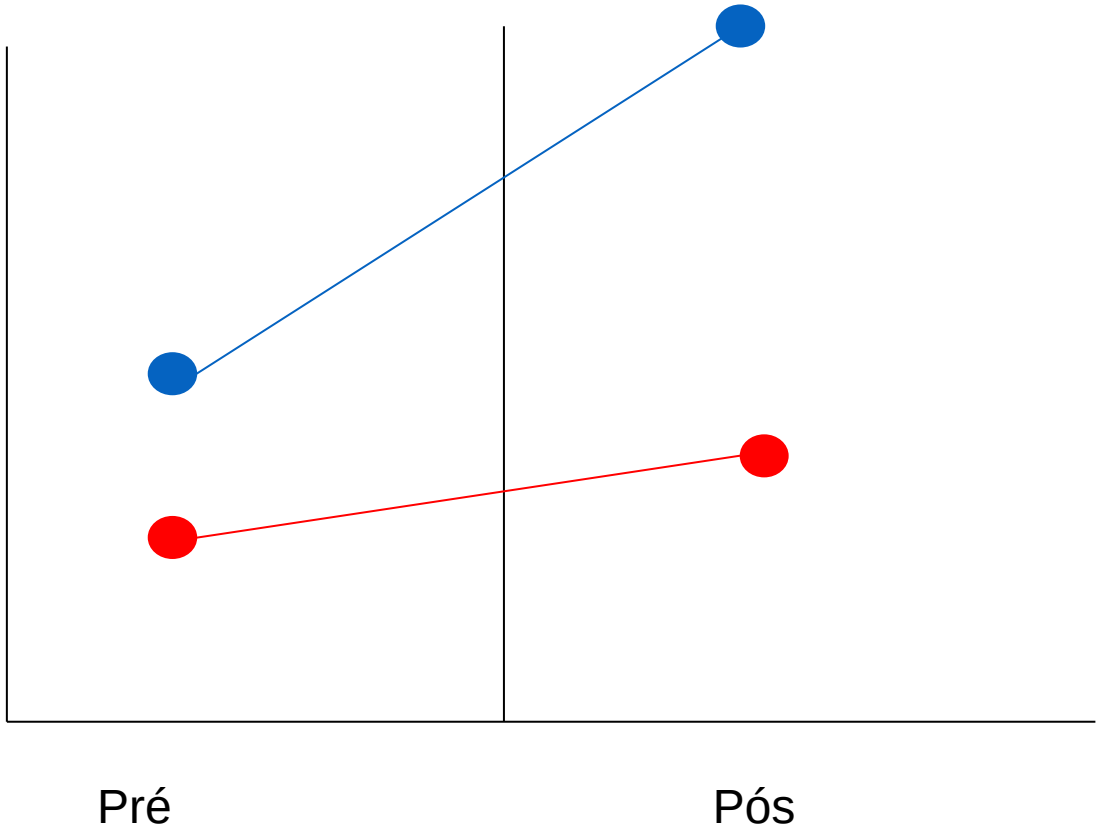
Differences-in-Differences

Pressuposto

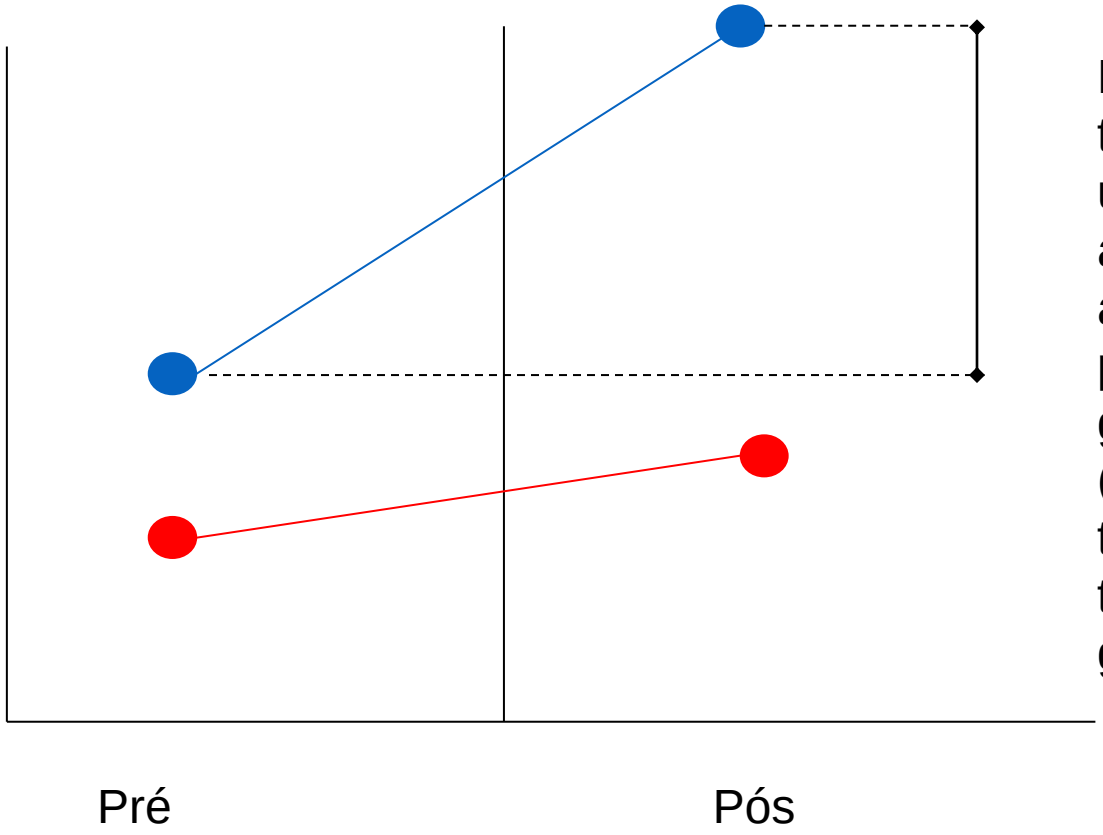
- O que tenha acontecido com o grupo de controle ao longo do tempo é o que teria acontecido com o grupo de tratamento na ausência da variação (choque exógeno)

- Talvez uma das estratégias de identificação mais popular na pesquisa aplicada atual
- Tentativas de imitar a atribuição aleatória com tratamento e amostra de "comparação" (controle)
- Aplicação do modelo de efeitos fixos de duas vias (*two-way fixed effects model*)

Differences-in-Differences

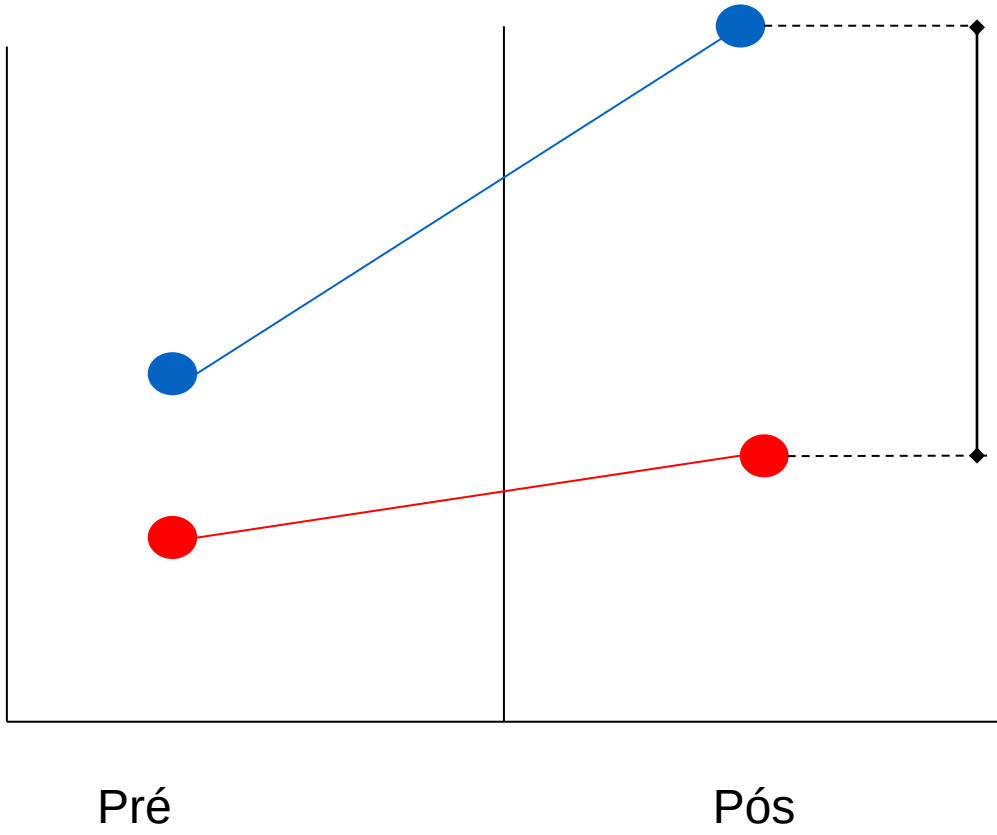


Differences-in-Differences



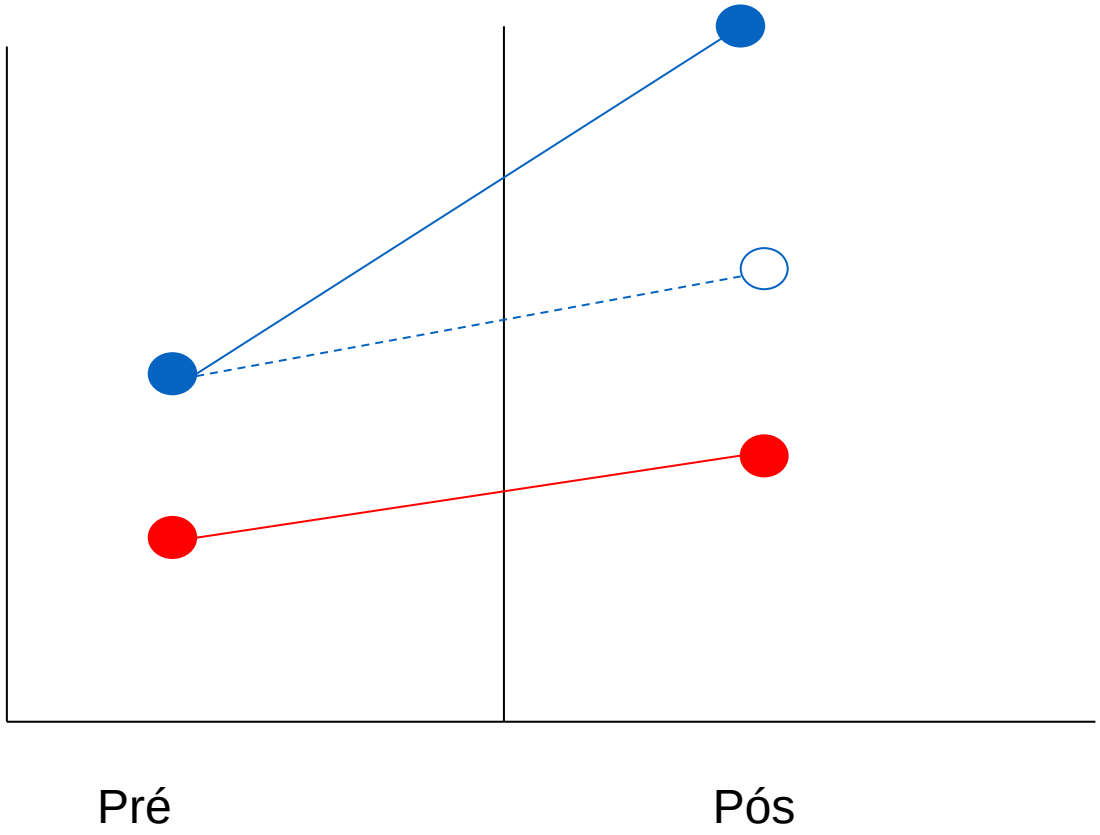
Efeito do tratamento usando apenas dados anteriores e posteriores do grupo T (ignorando a tendência temporal geral).

Differences-in-Differences

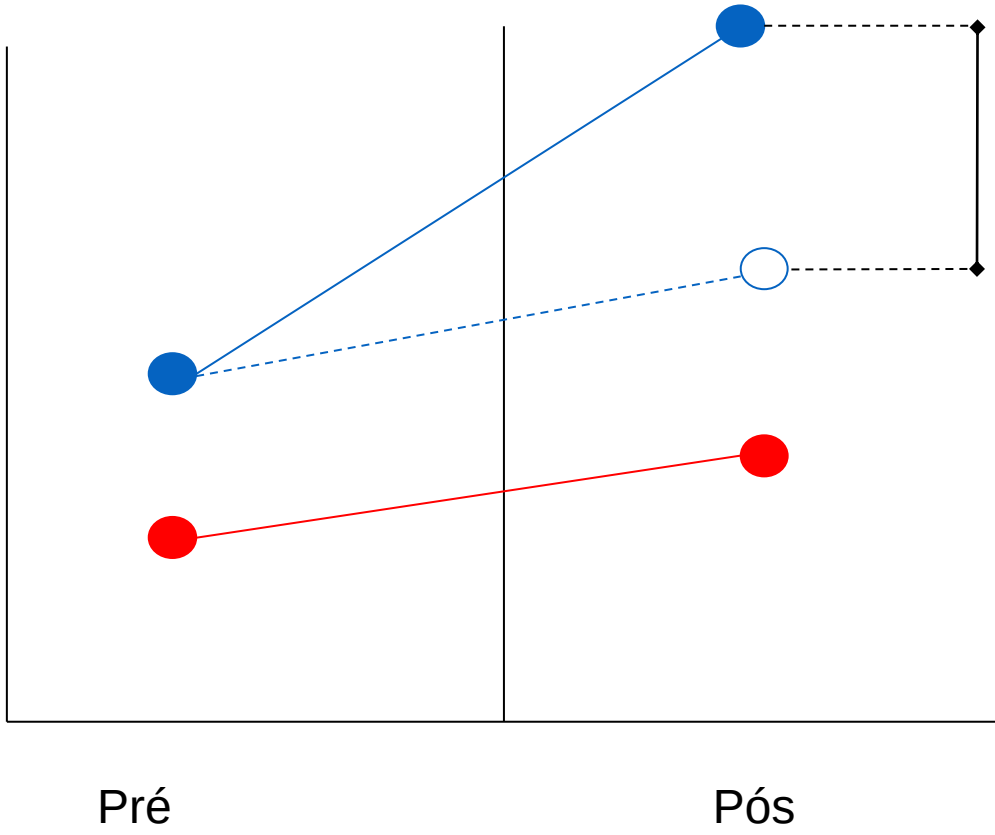


Efeito do tratamento usando apenas comparação de T & C da pós-intervenção (ignorando as diferenças pré-existentes entre os grupos T & C).

Differences-in-Differences



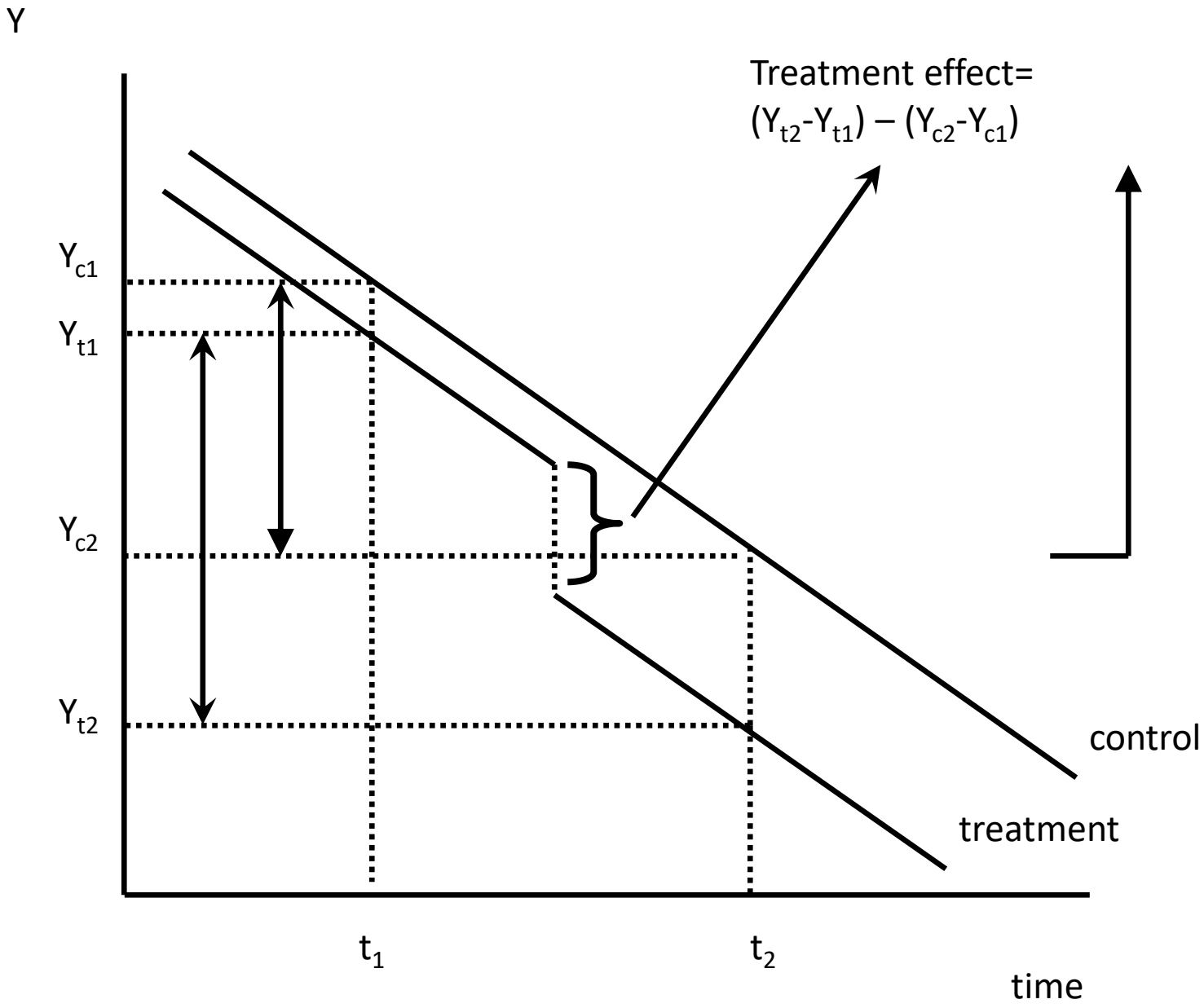
Differences-in-Differences



Efeito da diferença-na-diferença do tratamento (levando em consideração diferenças pré-existentes entre T & C e tendência geral do tempo).

Differences-in-Differences

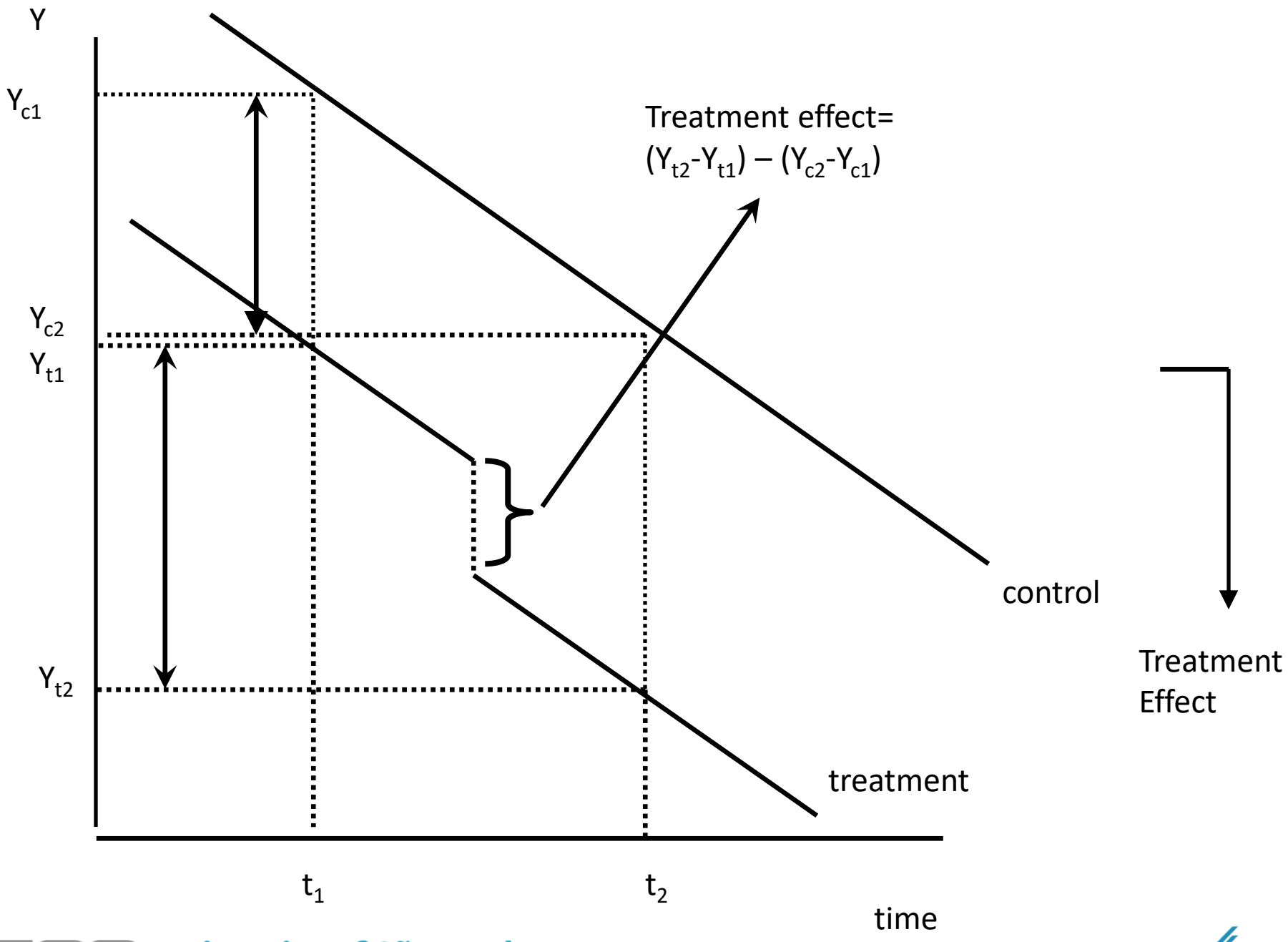
	Antes Tratamento	Depois Tratamento	Diferença
Grupo 1 (Treat)	Y_{t1}	Y_{t2}	$\Delta Y_t = Y_{t2}-Y_{t1}$
Grupo 2 (Control)	Y_{c1}	Y_{c2}	$\Delta Y_c = Y_{c2}-Y_{c1}$
Diferença			$\Delta\Delta Y$ $\Delta Y_t - \Delta Y_c$


Y
Treatment effect=
$(Y_{t2}-Y_{t1}) - (Y_{c2}-Y_{c1})$
Y_{c1}
Y_{t1}
Y_{c2}
Y_{t2}
control
treatment
t_1
t_2
time

Differences-in-Differences

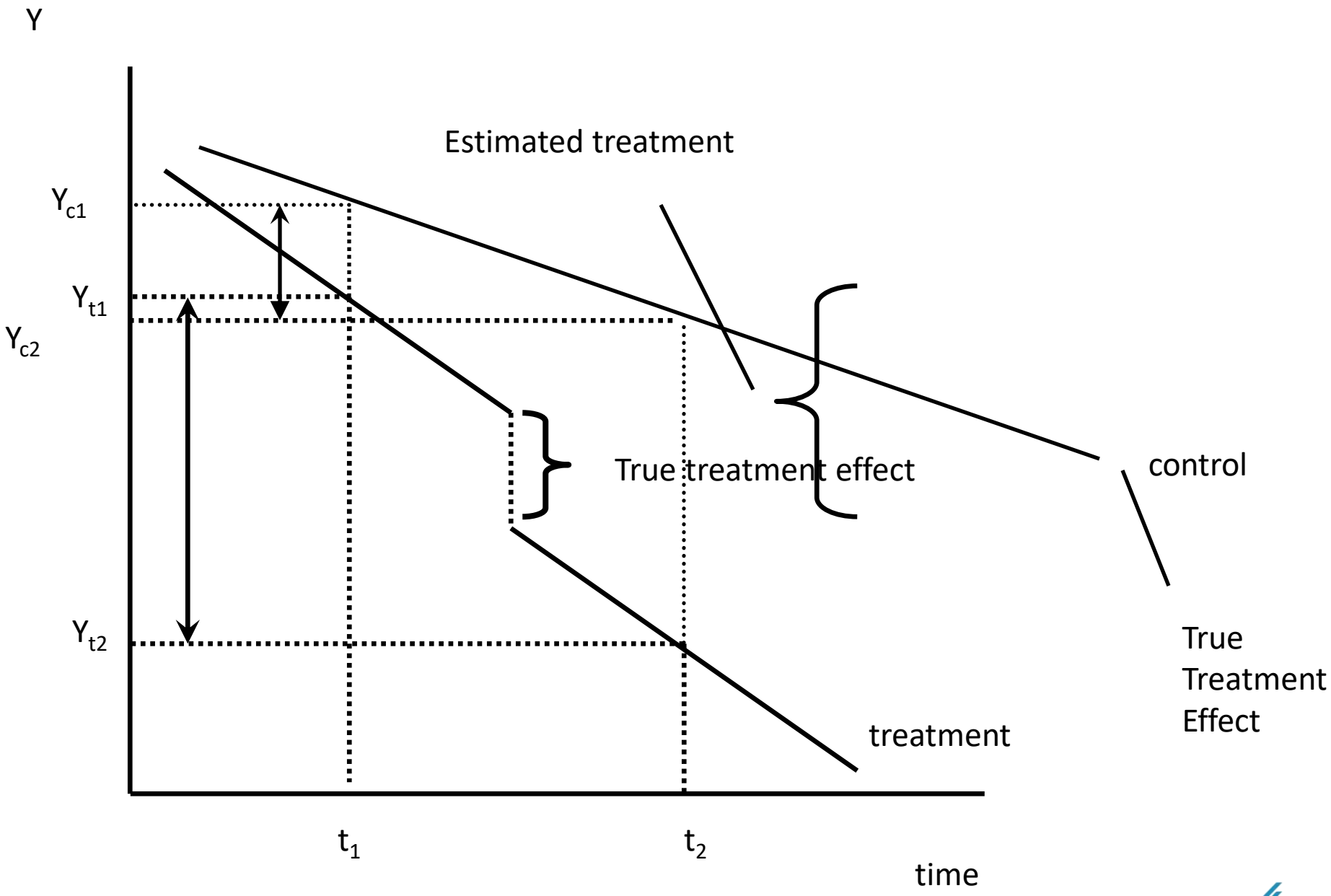
Pressupostos chave

- Grupo de controle identifica o caminho do tempo dos resultados que teriam ocorrido na ausência do tratamento
- Neste exemplo, Y cai por Y_{c2}-Y_{c1} mesmo sem o tratamento
- Observe que os "níveis" subjacentes de resultados não são importantes


Y
Y_{c1}
Y_{c2}
Y_{t1}
Y_{t2}
Treatment effect=
$(Y_{t2}-Y_{t1}) - (Y_{c2}-Y_{c1})$
control
treatment
Treatment
Effect
t_1
t_2
time

Differences-in-Differences

- Em contraste, o importante é que as tendências de tempo na ausência do tratamento são as mesmas em ambos os grupos
- Se o tratamento ocorrer em uma área com uma tendência diferente, então se subestima / supera o efeito do tratamento
- Neste exemplo, suponha que a intervenção ocorra em área com queda mais rápida


Y
Estimated treatment
Y_{c1}
Y_{t1}
Y_{c2}
True treatment effect
control
Y_{t2}
True
Treatment
Effect
treatment
t_1
t_2
time

Modelo Econométrico Básico

- Dados variam por
  - indivíduo (i) → firma
  - tempo (t)
  - Resultado é Y_{it}
- Somente dois períodos
- Tratamento vai ocorrer apenas em um grupo de observações (ex. firmas, etc.)

Modelo Econométrico

- Três variáveis
  - T_{it} =1 se obs i pertence ao grupo que será eventualmente tratado
  - A_{it} =1 no período em que o tratamento ocorre
  - $T_{it}A_{it}$ -- termo de interação, tratamento após a intervenção
- $Y_{it} = \beta_0 + \beta_1 T_{it} + \beta_2 A_{it} + \beta_3 T_{it} A_{it} + \varepsilon_{it}$

$$Y_{it} = \beta_0 + \beta_1 T_{it} + \beta_2 A_{it} + \beta_3 T_{it} A_{it} + \varepsilon_{it}$$

	Antes Tratam.	Depois Tratam.	Diferença
Grupo 1 (Treat)	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\Delta Y_t = \beta_2 + \beta_3$
Grupo 2 (Control)	β_0	$\beta_0 + \beta_2$	$\Delta Y_c = \beta_2$
Diferença			$\Delta\Delta Y = \beta_3$

Modelo Econométrico Geral

- Dados variam por
  - indivíduo (i) → firma
  - tempo (t)
  - Resultado é Y_{it}
- Muitos períodos
- A intervenção ocorrerá em um grupo de firma, mas também em uma variedade de tempo

Modelo Econométrico Geral

- u_i é um efeito da firma
- v_t é um conjunto complete do efeito do ano (tempo)
- Modelo de análise de covariância

- $Y_{it} = \beta_0 + \beta_3 T_{it}A_{it} + u_i + \lambda_t + \varepsilon_{it}$

Vantagem do Diff-in-Diff

- Suponha que os tratamentos não sejam aleatórios, mas sistemáticas
  - Ocorre em firmas com maior ou menor média Y
  - Ocorre em períodos de tempo com diferentes Y's
- Isso é capturado pela inclusão dos efeitos firmas / tempo - permite a covariância entre
  - u_i and $T_{it}A_{it}$
  - λ_t and $T_{it}A_{it}$

Vantagem do Diff-in-Diff

- Efeitos de grupo (firma)
  - Captura as diferenças entre os grupos que são constantes ao longo do tempo
- Efeitos do ano
  - Captura as diferenças ao longo do tempo que são comuns a todos os grupos (firmas)

Differences-in-Differences

Verificação de Robustez

- Se possível, use dados em vários períodos de pré-tratamento para mostrar que a diferença entre tratamento e controle é estável
  - Não é necessário que as tendências sejam paralelas, apenas para saber a função de cada uma

- Se possível, use dados em vários períodos pós-tratamento para mostrar que a diferença incomum entre tratamento e controle ocorre apenas concomitantemente ao programa.

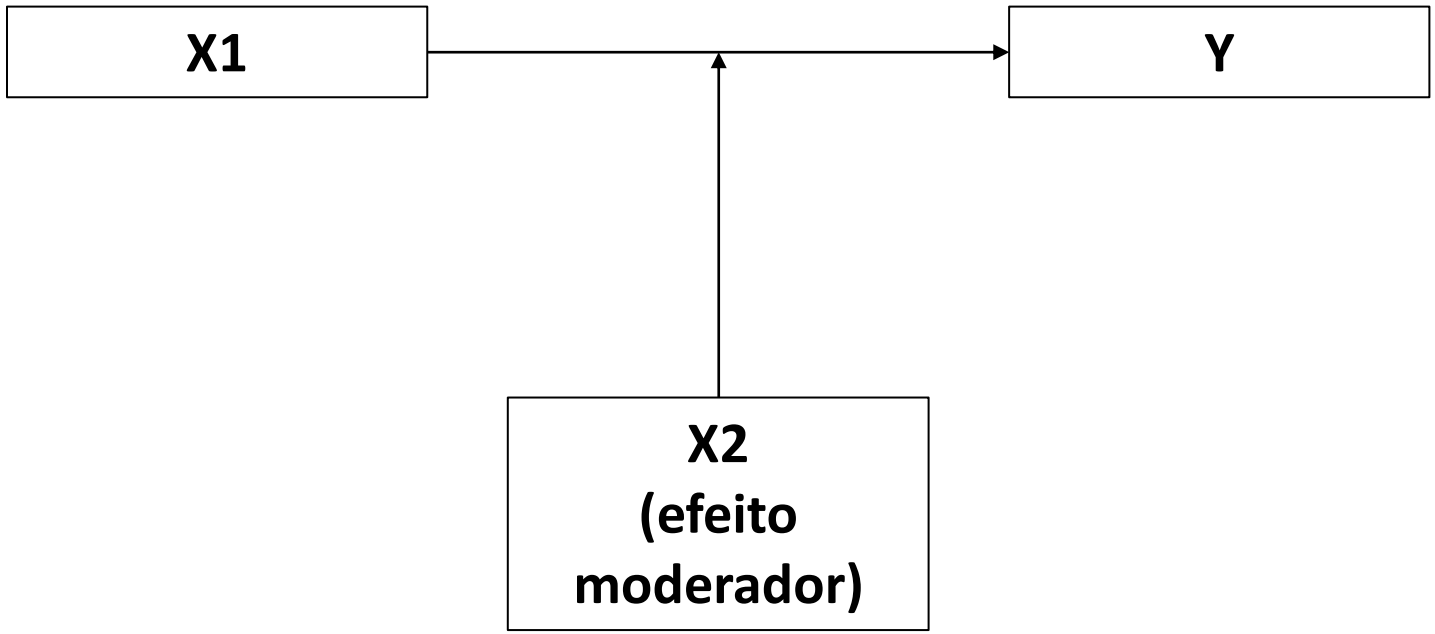
- Alternativamente, use dados em vários indicadores para mostrar que a resposta ao programa é apenas manifestada para aqueles que esperamos que seja (por exemplo, a estimativa *diff-in-diff* do impacto da adoção das IFRS nas receitas deve ser zero)

Variáveis de Interação

Variáveis de Interação

- A interação é um conceito de três variáveis. Uma delas é a variável dependente (Y) e as outras duas são variáveis explicativas (X1 e X2)
- Existe uma interação entre X1 e X2 se o impacto de um aumento em X2 em Y depender do nível de X1
- Para incorporar a interação no modelo de regressão múltipla, adicionamos a variável explicativa $(X_1 - \bar{X}_1) \times (X_2 - \bar{X}_2)$. Há evidência de uma interação se o coeficiente de $(X_1 - \bar{X}_1) \times (X_2 - \bar{X}_2)$ for significativo (teste *t* com p-valor ≤ 0,05)

Variáveis de Interação



Propensity Score Matching

PSM

Propensity Score Matching

- Viés na regressão múltipla
  - Endogeneidade
  - Má especificação da forma funcional dos modelos

- Efeito de tratamento médio (avarage treatment effect – ATE)
  - $W_i = \beta_0 + \beta_1 D_1 + \varepsilon_i$ ➔ Experimento não plausível
  - $W_i = \beta_0 + \beta_1 D_1 + \beta X_i + \varepsilon_i$

Propensity Score Matching

- O PSM
  - $D_i = \beta_0 + \beta X_i + \varepsilon_i$

  - Grupo de Tratamento → $D_i = 1$
  - Grupo de Controle → $D_i = 0$

Decisões no PSM

Escolhas Primárias na estimativa do Propensity Score

- Identificação dos grupos de tratamento e controle

- Especificação do modelo preditivo

Decisões no PSM

Escolhas Primárias na formação da Amostra Pareada (Matching)

- Matching com e sem reposição
- Distância do Calibrador
- Correspondência ‘‘ Um-para-Um ’’ e ‘‘ Um-para-Muitos ’’

Para a Próxima Aula

- Artigo: Performance matched discretionary accrual measures. Kothari, Leone & Wasley. JAE 2005.

Obrigado pela Atenção!!!

Até a próxima aula

 /mbotelhocm

 /mbotelhocm

 /in/mbotelhocm