ODSC 2019 Causal Inference Workshop

Controlled / Fixed Effects Regression

```
# set constants
n <- 10^4
n.time.periods <- 10
n.products <- 5
e1 <- rnorm(n = n)
e2 <- rnorm(n = n)
e3 <- rnorm(n = n)
B <- 2  #set true coefficient on X

# set variables; note X depends on fixed
# effects and other control variables
T.FE <- rep(1:n.time.periods, times = n/n.time.periods)
P.FE <- rep(1:n.products, each = n/n.products)
X <- 0.5 * T.FE + 0.5 * P.FE + rnorm(n = n) +
    e1
C <- rnorm(n = n) + e1 + e2
Y <- B * X + 2 * C + T.FE + P.FE + e3
dat <- data.frame(Y, X, C, P.FE, T.FE)

# explore data
head(dat)
```

```
##           Y        X          C P.FE T.FE
## 1 11.667464 1.799799  2.5309272    1    1
## 2 11.829398 1.593770  2.6739282    1    2
## 3 12.071662 4.559879 -0.1600107    1    3
## 4 13.675831 5.264257 -0.8102698    1    4
## 5  8.397205 1.210317  0.6781977    1    5
## 6 15.220805 2.547363  1.2811877    1    6
```

```
tail(dat)
```

```
##              Y        X          C P.FE T.FE
## 9995  15.25343 4.142010 -2.0418242    5    5
## 9996  24.22399 6.331238  0.1026609    5    6
## 9997  22.45573 3.919652  1.1522290    5    7
## 9998  28.58709 6.625890  0.9953495    5    8
## 9999  32.44717 6.503668  2.1836097    5    9
## 10000 28.80350 8.104927 -0.5756166    5   10
```

```
# controlled reg/fixed effect models
model1 <- lm(Y ~ X, data = dat)
```

```
model2 <- lm(Y ~ X + C, data = dat)
model3 <- lm(Y ~ X + factor(P.FE) + factor(T.FE),
    data = dat)
model4 <- lm(Y ~ X + C + factor(P.FE) + factor(T.FE),
    data = dat)
stargazer(model1, model2, model3, model4, type = "text",
    style = "aer", omit = c("C", "factor"), column.labels = c("Y~X",
        "Y~X+C", "Y~X+FE", "Y~X+C+FE"), dep.var.labels = "Controlled / Fixed Effects Regression",
    omit.stat = c("f", "ser", "rsq", "n"), notes = c("True Coef on X = 2"),
    notes.append = FALSE, add.lines = list(c("Add. Controls",
        "No", "Yes", "No", "Yes"), c("Fixed effects",
        "No", "No", "Yes", "Yes")))
```

```
##
## ===================================================
##                 Controlled / Fixed Effects Regression
##                   Y~X       Y~X+C     Y~X+FE    Y~X+C+FE
##                   (1)        (2)       (3)       (4)
## ---------------------------------------------------
## X               3.581***  3.221***  3.048***  2.003***
##                 (0.016)    (0.011)   (0.023)   (0.008)
##
## Add. Controls     No         Yes       No        Yes
## Fixed effects     No         No        Yes       Yes
## Adjusted R2     0.834      0.932     0.848     0.986
## ---------------------------------------------------
## Notes:         True Coef on X = 2
```

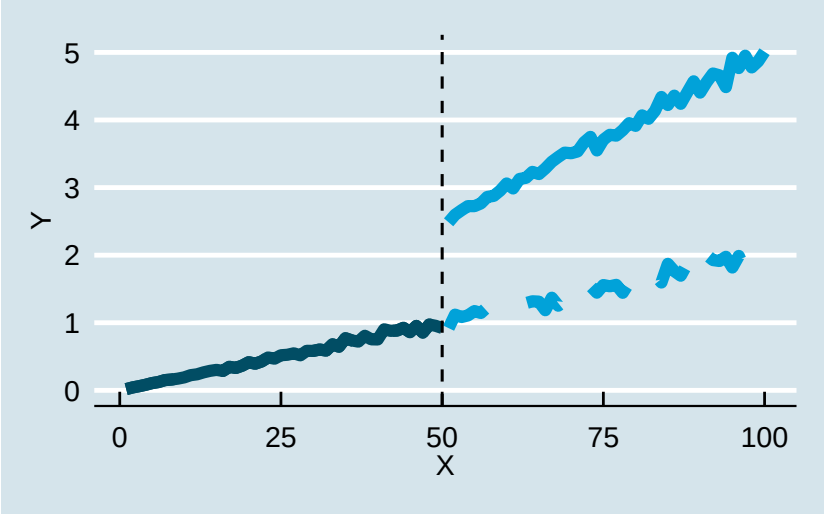
Regression Discontinuity

```
# set constants
n <- 100
mu1 <- 0.02
mu2 <- 0.05
sigma <- 0.001
cutoff <- n/2

# set variables
X <- data.frame(X = 1:n, Y = c((1:cutoff) * rnorm(cutoff,
    mu1, sigma), ((cutoff + 1):n) * rnorm(cutoff,
    mu2, sigma)))
X$counterfactual <- c(X$Y[X$X <= cutoff], c((cutoff +
    1):n) * X$Y[X$X <= cutoff]/c(1:cutoff))
X$cutoff <- X$X <= cutoff
```

```
# explore data
head(X)
```

```
##   X          Y counterfactual cutoff
## 1 1 0.01805270     0.01805270   TRUE
## 2 2 0.04308223     0.04308223   TRUE
## 3 3 0.06141187     0.06141187   TRUE
## 4 4 0.08209949     0.08209949   TRUE
## 5 5 0.10668648     0.10668648   TRUE
## 6 6 0.12265160     0.12265160   TRUE
```

```
# plot
X %>% ggplot(aes(X, Y, color = cutoff)) + geom_line(lwd = 2) +
    geom_line(aes(X, counterfactual), lty = 2,
        lwd = 2) + geom_vline(xintercept = n/2,
    lty = 2) + xlab("X") + ylab("Y") + theme_economist() +
    scale_color_economist() + theme(legend.position = "none")
```




```
# regression discontinuity model
model1 <- lm(Y ~ X + I(X > 50) + X * I(X > 50),
    data = X)
stargazer(model1, type = "text", style = "aer",
    column.labels = c("Y~X+I(X>Cutoff)+X*I(X>Cutoff)"),
    dep.var.labels = "Regression Discontinuity",
    omit.stat = c("f", "ser", "rsq", "n", "adj.rsq"),
    notes = c("Causal Impact = 1.5"), notes.append = FALSE,
    intercept.bottom = F)
```

```
##
## ========================================
```

```
##               Regression Discontinuity
##             Y~X+I(X>Cutoff)+X*I(X>Cutoff)
## ----------------------------------------
## Constant                0.001
##                        (0.017)
##
## X                     0.020***
##                        (0.001)
##
## I(X > 50)               0.005
##                        (0.046)
##
## X:I(X > 50)           0.030***
##                        (0.001)
##
## ----------------------------------------
## Notes:      Causal Impact = 1.5
```

```
print(paste("Causal Impact ~ -0.049+50*0.031 =",
    round(coef(model1)["I(X > 50)TRUE"] + coef(model1)["X:I(X > 50)TRUE"] *
        50, 2)))
```

```
## [1] "Causal Impact ~ -0.049+50*0.031 = 1.5"
```

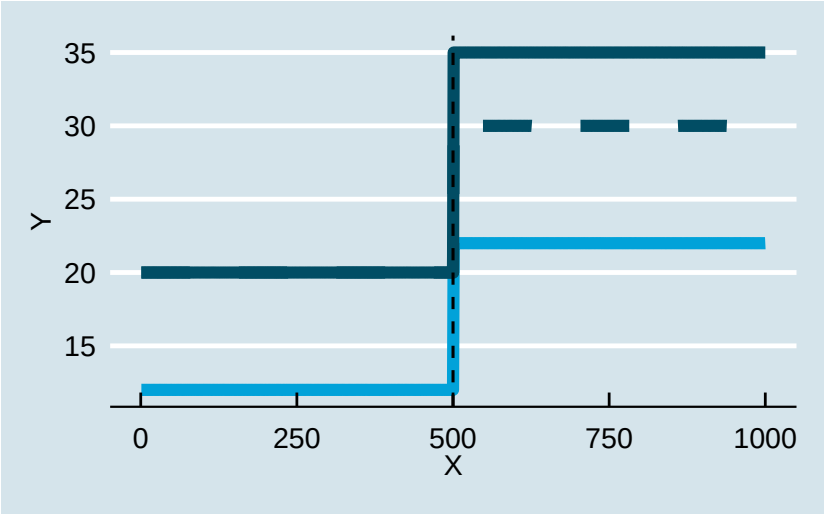
Difference in Difference

```
# set constants
n.per.group <- 500
time.periods <- 1000
mu1 <- 12
mu2 <- 20
delta <- 10
causal.effect <- 5
sigma <- 0.001
cutoff <- n.per.group/2

# set variables
X1.pre <- rnorm(time.periods/2, mu1, sigma)
X1.post <- rnorm(time.periods/2, mu1 + delta,
    sigma)
X2.pre <- rnorm(time.periods/2, mu2, sigma)
X2.post <- rnorm(time.periods/2, mu2 + delta +
    causal.effect, sigma)
X <- data.frame(time = rep(1:time.periods, times = 2),
    Post = rep(c(0, 1, 0, 1), each = time.periods/2),
```

```
    G = rep(c(0, 1), each = time.periods), Y = c(X1.pre,
        X1.post, X2.pre, X2.post))
X$counterfactual[X$G == 0] <- NA
X$counterfactual[X$G == 1] <- X$Y[X$G == 1] -
    X$Post[X$G == 1] * causal.effect

# explore data
head(X)
```

```
##   time Post G        Y counterfactual
## 1    1    0 0 12.00022             NA
## 2    2    0 0 12.00075             NA
## 3    3    0 0 12.00012             NA
## 4    4    0 0 11.99910             NA
## 5    5    0 0 12.00126             NA
## 6    6    0 0 12.00058             NA
```

```
# plot
X %>% ggplot(aes(time, Y, color = factor(G))) +
    geom_line(lwd = 2) + geom_line(aes(time, counterfactual),
    lty = 2, lwd = 2) + geom_vline(xintercept = 500,
    lty = 2) + xlab("X") + ylab("Y") + theme_economist() +
    scale_color_economist() + theme(legend.position = "none")
```




```
# difference in difference model
model1 <- lm(Y ~ Post + G + Post * G, data = X)
stargazer(model1, type = "text", style = "aer",
    column.labels = c("Y~Post+G+Post*G"), dep.var.labels = "Difference in Difference",
    omit.stat = c("f", "ser", "rsq", "n", "adj.rsq"),
    notes = c("Causal Impact = 5"), notes.append = FALSE,
    intercept.bottom = F)
```

```
##
## ================================
##          Difference in Difference
##              Y~Post+G+Post*G
## --------------------------------
## Constant        12.000***
##                  (0.000)
##
## Post            10.000***
##                  (0.000)
##
## G                8.000***
##                  (0.000)
##
## Post:G           5.000***
##                  (0.000)
##
## --------------------------------
## Notes:   Causal Impact = 5
```

Instrumental Variable

```
# set constants
n <- 1000
mu <- 0
sigma1 <- 2
sigma2 <- 0.5
beta <- 1
bias <- 2

# set variables
set.seed(19)
X <- NULL
X$e1 <- rnorm(n = n, mean = mu, sd = sigma1)
X$e2 <- rnorm(n = n, mean = mu, sd = sigma2)
X$Z <- rnorm(n = n, mean = mu, sd = sigma2)
X$X <- X$Z + X$e1
X$Y <- beta * X$X - bias * X$e1 + X$e2
X <- data.frame(X)

# explore data
head(X)
```

```
##           e1          e2           Z
```

```
## 1 -2.3789075  0.07544523 -0.04687593
## 2  0.7771625  0.02083529 -0.30982941
## 3 -0.6886667  0.39221778 -0.37480744
## 4 -1.0957923 -0.94253986  0.90961395
## 5  1.9613244  0.24031022 -0.66300937
## 6 -0.4732920 -0.37496661  0.22992028
##            X          Y
## 1 -2.4257834  2.4074768
## 2  0.4673330 -1.0661566
## 3 -1.0634741  0.7060770
## 4 -0.1861783  1.0628664
## 5  1.2983151 -2.3840236
## 6 -0.2433717  0.3282457
```

```
# IV model
model1 <- lm(Y ~ X, data = X)
model2 <- lm(X ~ Z, data = X)
model3 <- lm(Y ~ predict(model2), data = X)
model4 <- ivreg(Y ~ X | Z, data = X)
stargazer(model1, model2, model3, model4, type = "text",
    style = "aer", column.labels = c("Y~X", "Stage 1: X~Z",
        "Stage 2: Y~Xh", "IV"), dep.var.labels = c("",
        "", ""), covariate.labels = c("Constant",
        "X", "Z", "Xhat"), model.names = F, omit.stat = c("ser",
        "rsq", "n", "adj.rsq"), notes = c("Causal Impact = 1"),
    notes.append = FALSE, intercept.bottom = F)
```

```
##
## ===============================================================================
##
##                                Y~X      Stage 1: X~Z Stage 2: Y~Xh    IV
##                                (1)          (2)           (3)        (4)
## -------------------------------------------------------------------------------
## Constant                      -0.023       -0.014        0.028      0.028
##                              (0.033)      (0.062)       (0.064)    (0.165)
##
## X                           -0.919***                               1.649***
##                              (0.017)                                (0.464)
##
## Z                                         0.747***
##                                           (0.130)
##
## Xhat                                                   1.649***
##                                                         (0.180)
##
```

```
## F Statistic (df = 1; 998) 2,995.970***  32.969***     83.583***
## ---------------------------------------------------------------------------
## Notes:                    Causal Impact = 1
```

Double Selection

```
# set constants
N <- 10^3
N.Coeff <- 5 * 10^2
beta <- 2
C.mu <- rep(0, N.Coeff)
C.rho <- 0.5
beta.C.mu.sigma <- 10
beta.C.n.zero <- 25

# set variables
set.seed(19)
C.var <- rnorm(N.Coeff, mean = 1, sd = 1)^2
C <- as.data.frame.matrix(genCorGen(n = N, nvars = N.Coeff,
    params1 = C.mu, params2 = C.var, dist = "normal",
    rho = C.rho, corstr = "ar1", wide = "True"))[,
    -1]
betaC <- rnorm(N.Coeff, mean = beta.C.mu.sigma,
    sd = beta.C.mu.sigma)
betaC[beta.C.n.zero:N.Coeff] <- 0
Treatment <- rep(0, N)
Treatment[0:(N/2)] <- 1
Treatment <- sample(Treatment)
e <- rnorm(N)
Y <- beta * Treatment + data.matrix(C) %*% betaC +
    e
X <- data.frame(Y, Treatment, C)

# explore data
head(X[, 1:5])
```

```
##           Y Treatment          V1
## 1  54.19154         1 -0.21468905
## 2 -19.55717         0  0.25325331
## 3 153.91293         0 -0.20771878
## 4 132.56959         0 -0.05069902
## 5 -95.73945         0  0.01174809
## 6  73.51079         1 -0.03666658
##            V2          V3
```

```
## 1 -1.26013819  0.61072150
## 2  1.88221311  0.09154465
## 3 -1.33291039 -0.46438715
## 4 -0.50129271 -0.37716702
## 5 -0.02347128  0.65687552
## 6  0.53461290  0.10374449
```

```
# double selection
C <- data.matrix(X[, -which(colnames(X) %in% c("Y",
    "Treatment"))])
glmnet.model1 <- cv.glmnet(C, X$Y, alpha = 1)
Y.on.X <- colnames(C)[unlist(predict.cv.glmnet(glmnet.model1,
    s = "lambda.1se", type = "nonzero"))]
glmnet.model2 <- cv.glmnet(C, X$Treatment, alpha = 1)
T.on.X <- colnames(C)[unlist(predict.cv.glmnet(glmnet.model2,
    s = "lambda.1se", type = "nonzero"))]
var.union <- unique(c(Y.on.X, T.on.X))
length(var.union)
```

```
## [1] 22
```

```
lm.formula <- paste("Y~Treatment+", paste(var.union,
    collapse = "+"), sep = "")
model1 <- lm(Y ~ Treatment, data = X)
model2 <- lm(Y ~ ., data = X)
model3 <- lm(lm.formula, data = X)
stargazer(model1, model2, model3, type = "text",
    style = "aer", column.labels = c("No Controls",
        "All Controls", "Double Selection"), dep.var.labels = c("",
        "", ""), covariate.labels = c("Treatment"),
    omit = c("V", "Constant"), model.names = F,
    omit.stat = c("ser", "rsq", "n", "adj.rsq"),
    notes = c("Causal Impact = 2"), notes.append = FALSE)
```

```
##
## ============================================================================================
##
##                  No Controls            All Controls               Double Selection
##                      (1)                    (2)                         (3)
## --------------------------------------------------------------------------------------------
## Treatment          -1.751                 1.995***                    2.018***
##                    (6.099)                (0.091)                     (0.063)
##
## F Statistic 0.082 (df = 1; 998) 18,626.830*** (df = 501; 498) 415,864.300*** (df = 23; 976)
## --------------------------------------------------------------------------------------------
## Notes:      Causal Impact = 2
```

Causal Forests

```
# set constants
N <- 5 * 10^3
N.Coeff <- 5
N.groups <- 4
beta <- rep(c(1:N.groups), each = N/N.groups)
var.group <- beta
C.mu <- rep(0, N.Coeff)
C.rho <- 0.5
C.var <- rnorm(N.Coeff, mean = 1, sd = 1)^2
beta.C.mu.sigma <- 5

# set variables
set.seed(19)
C <- as.data.frame.matrix(genCorGen(n = N, nvars = N.Coeff,
    params1 = C.mu, params2 = C.var, dist = "normal",
    rho = C.rho, corstr = "ar1", wide = "True"))[,
    -1]
betaC <- rnorm(N.Coeff, mean = beta.C.mu.sigma,
    sd = beta.C.mu.sigma)
Treatment <- rep(0, N)
Treatment[0:(N/2)] <- 1
Treatment <- sample(Treatment)
e <- rnorm(N)
Y <- beta * Treatment + data.matrix(C) %*% betaC +
    e
X <- data.frame(Y, Treatment, C, Group = as.character(var.group))

# explore data
head(X[, c(1:5, ncol(X))])
```

```
##            Y Treatment          V1
## 1  33.587271         1  0.58919899
## 2  49.299055         0  0.24934014
## 3  -2.434526         0 -0.43294964
## 4 -43.642276         1  0.25653971
## 5  21.765675         0 -0.02637059
## 6 -53.671680         1 -2.02725419
##            V2         V3 Group
## 1  1.08505932  3.0956992     1
## 2 -0.02567679  0.2031037     1
## 3  1.07740595  0.7587372     1
## 4 -0.30862664 -3.5337931     1
## 5 -0.12183912  0.1029569     1
```

```
## 6 -1.24164849 -2.3248514     1

# regular OLS
C <- data.matrix(X[, -which(colnames(X) %in% c("Y",
    "Treatment"))])
model1 <- lm(Y ~ ., data = X[, -which(colnames(X) ==
    "Group")])
model2 <- lm(Y ~ . + Treatment * Group, data = X)
stargazer(model1, model2, type = "text", style = "aer",
    column.labels = c("All Controls", "All Controls + Group Interactions"),
    dep.var.labels = c("", "", ""), covariate.labels = c("Treatment"),
    omit = c("V", "Constant", "^Group"), model.names = F,
    omit.stat = c("ser", "rsq", "n", "adj.rsq"),
    notes = c("Average Treatment Effect = 2.5"),
    notes.append = FALSE)

##
## ================================================================================
##
##                            All Controls          All Controls + Group Interactions
##                                 (1)                              (2)
## --------------------------------------------------------------------------------
## Treatment                     2.503***                         0.986***
##                                (0.036)                          (0.057)
##
## Treatment:Group2                                               1.022***
##                                                                 (0.080)
##
## Treatment:Group3                                               2.103***
##                                                                 (0.080)
##
## Treatment:Group4                                               2.999***
##                                                                 (0.080)
##
## F Statistic      291,479.600*** (df = 6; 4993)  237,077.000*** (df = 12; 4987)
## --------------------------------------------------------------------------------
## Notes:           Average Treatment Effect = 2.5

# causal forest
cf <- causal_forest(X = model.matrix(~., data = X[,
    -which(colnames(X) %in% c("Y", "Treatment"))]),
    Y = X$Y, W = X$Treatment, honesty = T, honesty.fraction = 0.5)
pred <- predict(cf)$predictions
cf %>% variable_importance() %>% as.data.frame() %>%
    mutate(variable = colnames(model.matrix(~.,
```

```
        data = X[, -which(colnames(X) %in% c("Y",
            "Treatment"))]))) %>% arrange(desc(V1))
```

```
##           V1    variable
## 1 0.60036797      Group4
## 2 0.10246498      Group3
## 3 0.06692440          V5
## 4 0.06678192          V1
## 5 0.05738828          V3
## 6 0.05036280          V4
## 7 0.04017082          V2
## 8 0.01553884      Group2
## 9 0.00000000 (Intercept)
```

```
tapply(pred, X$Group, mean)
```

```
##        1        2        3        4
## 1.130463 1.909108 3.165226 3.896715
```

```
data.frame(true = beta, est = pred, Group = X$Group) %>%
    ggplot(aes(true, est, color = Group)) + geom_point() +
    xlab("True Treatment") + ylab("Estimated Treatment") +
    theme_economist() + scale_color_economist() +
    theme(legend.position = "none")
```