

FINAL EXAM

This is an open book exam. You may refer to whatever books, notes, etc. you deem helpful, but you may not collude or seek the help of others in any way. The work you submit must be your own.

There are three questions, each with multiple parts. You should answer all of them. There are a total of 100 points, and you have 24 hours to complete the exam. Good luck, and remember: econometrics is fun!

1. Suppose you're conducting a randomized experiment to measure the causal effect of a new weight-loss drug called W8-B-Gon. You randomly assign subjects to treatment and control groups, measure their pre-treatment weight, provide the treatment group with a six-month supply of W8-B-Gon and the control group with a placebo, and then measure their weight again at the conclusion of the six-month trial. Let y_i indicate the change in subject i 's weight over the course of the trial, and let $T_i = 1$ if subject i was part of the treatment group and zero otherwise.

At the conclusion of the trial, you learn that some members of your treatment group didn't follow the rules of your experiment, and sold their W8-B-Gon on the black market instead of taking the treatment. Even worse, some members of your control group bought W8-B-Gon on the black market and ended up taking the treatment! Luckily, your cousin knows someone who knows someone, and is able to find out which of your subjects actually completed the W8-B-Gon treatment, and which ones did not. From this information, you create a new variable, $P_i = 1$ if subject i completed the W8-B-Gon treatment, and zero otherwise.

- a. [10 points] Suppose you estimate the regression $y_i = \alpha + \psi P_i + \eta_i$ via OLS, where η_i is an error term. Does the OLS estimator $\hat{\psi}$ provide a good estimate of the causal effect of W8-B-Gon on weight loss? Explain.
- b. [10 points] Suppose you estimate the regression $y_i = \alpha + \phi T_i + \nu_i$ via OLS instead, where ν_i is an error term. Does the OLS estimator $\hat{\phi}$ provide a good estimate of the causal effect of W8-B-Gon on weight loss? Explain.
- c. [10 points] Is there another way to use these data to obtain a good estimate of the causal effect of W8-B-Gon on weight loss? Explain.

2. The government of Canada recently announced a massive wage subsidy that will be paid to some Canadian employers. Employers that qualify for the subsidy will be paid up to 75 percent of each employee's wage (to a maximum of \$847 per employee per week), for up to 12 weeks between March 15 and June 6 2020. To qualify for the subsidy in a given month, employers must experience a 30% year-over-year decline in monthly revenue.

Suppose you have data on a sample of 10,000 Canadian firms. Your data include each firm i 's total revenue in 2018, 2019, 2020, and 2021 ($R_i^{2018}, R_i^{2019}, R_i^{2020}, R_i^{2021}$), their employment in each of those years ($E_i^{2018}, E_i^{2019}, E_i^{2020}, E_i^{2021}$), the year-over-year percentage change in their monthly revenue between March 15 and June 6 2020 (QR_i), the province where each firm is located (P_i), the industry in which they operate (I_i), and a random lottery number L_i assigned to each firm, where $1 \leq L_i \leq 10$.

- a. [10 points] Is it possible to estimate the causal effect of the wage subsidy on a firm's total revenue for the year 2021 using these data? If yes, clearly explain how. What causal effect are you estimating, what assumptions does your method require, and how would you ensure that you are making valid inferences about the causal effect? If it's not possible to measure the causal effect of the wage subsidy program, explain why not.
- b. [10 points] Now suppose that the government is unable to keep their promise to subsidize all firms that experience a 30% decline in revenue. Instead, *only some of the firms that experience a 30% decline* will receive the subsidy, and the remaining firms will not. Unfortunately, you are unable to observe what factors are used to determine which of the firms experiencing a 30% decline receive the subsidy, and which ones do not. Is it possible to estimate the causal effect of the wage subsidy on a firm's total revenue for the year 2021? If yes, clearly explain how. What causal effect are you estimating, what assumptions does your method require, and how would you ensure that you are making valid inferences about the causal effect? If it's not possible to measure the causal effect of the wage subsidy program, explain why not.
- c. [10 points] Now suppose that in response to intense political pressure, the government eliminates the qualifying requirement that employers must experience a 30% decline in revenue. Instead, all firms that operate in a subset of industries that are deemed to have been most harmed ("harmed industries") will receive the subsidy, and firms in all other industries ("unharmed industries") will not. Is it possible to estimate the causal effect of the wage subsidy on a firm's total revenue for the year 2021? If yes, clearly explain how. What causal effect are you estimating, what assumptions does your method require, and how would you ensure that you are making valid inferences about the causal effect? If it's not possible to measure the causal effect of the wage subsidy program, explain why not.
- d. [10 points] Now suppose that for equity reasons, all firms with $L_i > 5$ will receive the subsidy. Is it possible to estimate the causal effect of the wage subsidy on a firm's total revenue for the year 2021? If yes, clearly explain how. What causal effect are you estimating, what assumptions does your method require, and how would you ensure that you are making valid inferences about the causal effect? If it's not possible to measure the causal effect of the wage subsidy program, explain why not.
- e. [5 points] Which of the methods that you proposed in parts (a)-(d) do you prefer, and why?

3. Consider the R output below, about which several questions follow. The output is based on the same school-visit data that you used in Assignments 2, 3, and 4. As a reminder, here are some key variables:

Variable name	Description
schoolid	ID number of the school sector
group	indicates whether school (and student) are in the control group (0), unconditional transfer group (1), or conditional transfer group (2)
benef	indicates whether transfer recipient is Mother or Father
province	province of residence
b_electricity	indicates whether the school had electricity at baseline
b_gender	student's gender at baseline
v1_status	student's enrollment status during the first post-assignment school visit
v2_status	student's enrollment status during the second post-assignment school visit
v3_status	student's enrollment status during the third post-assignment school visit
v4_status	student's enrollment status during the fourth post-assignment school visit
v5_status	student's enrollment status during the fifth post-assignment school visit
v6_status	student's enrollment status during the sixth post-assignment school visit

```
setwd("~/Desktop/moroccoLCT")

library(tidyverse)
library(readxl)
library(sandwich)
library(lmtest)
library(miceadds)

data <- read_excel("a2_school_visits.xls")

data$treat <- 0*(data$group==0) + 1*(data$group==1 & data$benef=="Mother") +
  2*(data$group==1 & data$benef=="Father") + 3*(data$group==2 & data$benef=="Mother") +
  4*(data$group==2 & data$benef=="Father")

data$dropout <- ( (!is.na(data$v1_status) & data$v1_status=="2 dropped out") |
  (!is.na(data$v2_status) & data$v2_status=="2 dropped out") |
  (!is.na(data$v3_status) & data$v3_status=="2 dropped out") |
  (!is.na(data$v4_status) & data$v4_status=="2 dropped out") |
  (!is.na(data$v5_status) & data$v5_status=="2 dropped out") |
  (!is.na(data$v6_status) & data$v6_status=="2 dropped out") ) &
  ( is.na(data$v6_status) |
  ( !is.na(data$v6_status) & data$v6_status != "1 enrolled" &
    data$v6_status != "8 Primary school completed" ) )

data <- data %>% mutate(F = 0 + 1*(b_gender=="F"),
  electricity = 0 + 1*(b_electricity=="1 Yes, in all the school"),
  treat.f = factor(treat),
  treat1 = 0 + 1*(treat==1),
  treat2 = 0 + 1*(treat==2),
  treat3 = 0 + 1*(treat==3),
  treat4 = 0 + 1*(treat==4) )

data <- data[!is.na(data$F) & !is.na(data$treat) & !is.na(data$dropout) & !is.na(data$electricity),]
data %>% group_by(province) %>% summarise(n = n())

## # A tibble: 17 x 2
##   province      n
##   <chr>      <int>
## 1 Al Haouz    1165
## 2 Azilal     2734
```

```

## 3 Chichaoua      2825
## 4 Chtouka Ait Baha 1004
## 5 El Hajeb       416
## 6 El Kelaa Des Sraghna 2397
## 7 Errachidia     1260
## 8 Essaouira      6402
## 9 Ifrane         514
## 10 Jerada        665
## 11 Khenifra      1226
## 12 Meknes        1249
## 13 Nador         1246
## 14 Ouarzazate    5058
## 15 Taourirt      456
## 16 Taroudant     5596
## 17 Tiznit        423

data <- data %>%
  group_by(province) %>%
  mutate(A0 = mean(dropout),
         A1 = mean(treat1),
         A2 = mean(treat2),
         A3 = mean(treat3),
         A4 = mean(treat4),
         A5 = mean(F),
         A6 = mean(electricity) ) %>%
  ungroup()
data <- data %>%
  mutate(B0 = dropout - A0,
         B1 = treat1 - A1,
         B2 = treat2 - A2,
         B3 = treat3 - A3,
         B4 = treat4 - A4,
         B5 = F - A5,
         B6 = electricity - A6 )
data %>% summarise(n = n(), dropout_bar = mean(dropout), treat1_bar = mean(treat1),
                  treat2_bar = mean(treat2), treat3_bar = mean(treat3),
                  treat4_bar = mean(treat4), F_bar = mean(F) )

## # A tibble: 1 x 7
##       n dropout_bar treat1_bar treat2_bar treat3_bar treat4_bar F_bar
##   <int>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl> <dbl>
## 1 34636      0.0406      0.124      0.128      0.279      0.275 0.441

data %>% summarise(electricity_bar = mean(electricity), A_0 = mean(A1) )

## # A tibble: 1 x 2
##   electricity_bar  A_0
##         <dbl> <dbl>
## 1          0.332 0.124

data %>% summarise(A1_bar = mean(A1), A2_bar=mean(A2),
                  A3_bar = mean(A3), A4_bar=mean(A4), A5_bar = mean(A5), A6_bar=mean(A6),
                  B0_bar = mean(B0), B1_bar = mean(B1), B2_bar=mean(B2) )

## # A tibble: 1 x 9
##   A1_bar A2_bar A3_bar A4_bar A5_bar A6_bar  B0_bar  B1_bar  B2_bar
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>   <dbl>   <dbl>
## 1  0.124  0.128  0.279  0.275  0.441  0.332 -4.94e-20 4.70e-18 9.46e-19

data %>% summarise(B3_bar = mean(B3), B4_bar=mean(B4), B5_bar = mean(B5), B6_bar=mean(B6) )

## # A tibble: 1 x 4
##   B3_bar  B4_bar  B5_bar  B6_bar
##   <dbl>   <dbl>   <dbl>   <dbl>
## 1 -1.49e-17 -1.12e-18 -3.63e-18 6.27e-18

model1 <- lm(dropout ~ treat.f + F, data=data)
summary(model1)

```

```
##
## Call:
## lm(formula = dropout ~ treat.f + F, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07312 -0.03841 -0.03544 -0.02818  0.97356
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.071379   0.002568  27.798 <2e-16 ***
## treat.f1     -0.039540   0.003842 -10.292 <2e-16 ***
## treat.f2     -0.035935   0.003808  -9.438 <2e-16 ***
## treat.f3     -0.044943   0.003123 -14.389 <2e-16 ***
## treat.f4     -0.034715   0.003131 -11.086 <2e-16 ***
## F              0.001741   0.002128   0.818   0.413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1966 on 34630 degrees of freedom
## Multiple R-squared:  0.006598, Adjusted R-squared:  0.006454
## F-statistic: 46 on 5 and 34630 DF, p-value: < 2.2e-16

model2 <- lm(dropout ~ treat.f + F + electricity, data=data)
summary(model2)

##
## Call:
## lm(formula = dropout ~ treat.f + F + electricity, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07608 -0.04172 -0.03582 -0.02785  0.97979
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.074312   0.002626  28.294 < 2e-16 ***
## treat.f1     -0.038494   0.003845 -10.010 < 2e-16 ***
## treat.f2     -0.035782   0.003806  -9.401 < 2e-16 ***
## treat.f3     -0.042002   0.003172 -13.243 < 2e-16 ***
## treat.f4     -0.034364   0.003131 -10.976 < 2e-16 ***
## F              0.001772   0.002128   0.833   0.405
## electricity -0.012096   0.002295  -5.270 1.37e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1966 on 34629 degrees of freedom
## Multiple R-squared:  0.007394, Adjusted R-squared:  0.007222
## F-statistic: 42.99 on 6 and 34629 DF, p-value: < 2.2e-16

model3 <- lm(electricity ~ treat.f + F, data=data)
summary(model3)

##
## Call:
## lm(formula = electricity ~ treat.f + F, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.4881 -0.2740 -0.2552  0.5144  0.7575
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.242518   0.006010  40.353 < 2e-16 ***
## treat.f1     0.086436   0.008992   9.613 < 2e-16 ***
## treat.f2     0.012677   0.008912   1.423   0.155
## treat.f3     0.243088   0.007310  33.252 < 2e-16 ***
## treat.f4     0.028968   0.007329   3.953 7.75e-05 ***
```

```

## F          0.002520  0.004982  0.506  0.613
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4602 on 34630 degrees of freedom
## Multiple R-squared:  0.04466,    Adjusted R-squared:  0.04453
## F-statistic: 323.8 on 5 and 34630 DF,  p-value: < 2.2e-16

model4 <- lm(B0 ~ B1 + B2 + B3 + B4 + B5 + B6 , data=data)
summary(model4)

##
## Call:
## lm(formula = B0 ~ B1 + B2 + B3 + B4 + B5 + B6, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09472 -0.04689 -0.03651 -0.02648  1.00476
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.781e-16  1.055e-03   0.000  1.0000
## B1          -4.197e-02  3.900e-03 -10.761 <2e-16 ***
## B2          -3.787e-02  3.858e-03  -9.818 <2e-16 ***
## B3          -4.403e-02  3.221e-03 -13.670 <2e-16 ***
## B4          -3.737e-02  3.216e-03 -11.620 <2e-16 ***
## B5           2.188e-03  2.127e-03   1.029  0.3037
## B6          -7.771e-03  2.490e-03  -3.121  0.0018 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1963 on 34629 degrees of freedom
## Multiple R-squared:  0.007219,    Adjusted R-squared:  0.007047
## F-statistic: 41.97 on 6 and 34629 DF,  p-value: < 2.2e-16

model5 <- lm(dropout ~ treat.f + F + electricity + treat.f:F, data=data)
summary(model5)

##
## Call:
## lm(formula = dropout ~ treat.f + F + electricity + treat.f:F,
##      data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08097 -0.04264 -0.03566 -0.02711  0.98072
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.070621  0.003225  21.896 < 2e-16 ***
## treat.f1     -0.034963  0.005129  -6.816 9.48e-12 ***
## treat.f2     -0.029454  0.005042  -5.842 5.20e-09 ***
## treat.f3     -0.036066  0.004214  -8.559 < 2e-16 ***
## treat.f4     -0.031376  0.004164  -7.535 4.99e-14 ***
## F            0.010352  0.004840   2.139  0.0325 *
## electricity -0.012134  0.002295  -5.287 1.25e-07 ***
## treat.f1:F    -0.008190  0.007739  -1.058  0.2900
## treat.f2:F    -0.014698  0.007688  -1.912  0.0559 .
## treat.f3:F    -0.013490  0.006290  -2.145  0.0320 *
## treat.f4:F    -0.006956  0.006314  -1.102  0.2706
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1966 on 34625 degrees of freedom
## Multiple R-squared:  0.007562,    Adjusted R-squared:  0.007275
## F-statistic: 26.38 on 10 and 34625 DF,  p-value: < 2.2e-16

```

```
model6 <- lm.cluster(dropout ~ treat.f + F + electricity + treat.f:F, data=data, cluster="province")
summary(model6)
```

```
## R^2= 0.00756
##
##              Estimate Std. Error   t value    Pr(>|t|)
## (Intercept)  0.070620853 0.005512989 12.8099016 1.443175e-37
## treat.f1     -0.034963467 0.008297383 -4.2137944 2.511158e-05
## treat.f2     -0.029454432 0.008160177 -3.6095336 3.067480e-04
## treat.f3     -0.036065991 0.006897874 -5.2285661 1.708298e-07
## treat.f4     -0.031376430 0.008731188 -3.5936035 3.261360e-04
## F            0.010352068 0.009234452  1.1210269 2.622764e-01
## electricity -0.012134072 0.004825219 -2.5147194 1.191272e-02
## treat.f1:F   -0.008189870 0.009159680 -0.8941218 3.712567e-01
## treat.f2:F   -0.014698138 0.012845914 -1.1441878 2.525458e-01
## treat.f3:F   -0.013489606 0.009074656 -1.4865143 1.371431e-01
## treat.f4:F   -0.006956247 0.010135062 -0.6863546 4.924895e-01
```

- [2 points] What does the following command do?

```
data <- data[!is.na(data$F) & !is.na(data$treat) & !is.na(data$dropout) & !is.na(data$electricity),]
```
- [2 points] The coefficient on `treat.f1` in `model11` is `-0.039540`. Interpret the value of this coefficient.
- [3 points] The coefficient on `treat.f1` in `model12` is `-0.038494`. Interpret the value of this coefficient. Why is it different than the coefficient on `treat.f1` in `model11`?
- [4 points] The coefficient on `treat.f1` in `model13` is `0.086436`. How is this related to the coefficients on `treat.f1` in `model11` and `model12`? Explain.
- [3 points] The coefficient on `B1` in `model14` is `-4.197e-02`. How is this related to the coefficient on `treat.f1` in `model12`? Explain.
- [2 points] The coefficient on `treat.f1` in `model15` is `-0.034963`. Interpret the value of this coefficient.
- [4 points] The coefficients on `treat.f1` in `model15` and `model16` are the same. What's the relationship between these two models? Which do you prefer, and why?
- [5 points] Do the reported estimates indicate that the effect of treatment is the same for boys and girls? Explain.