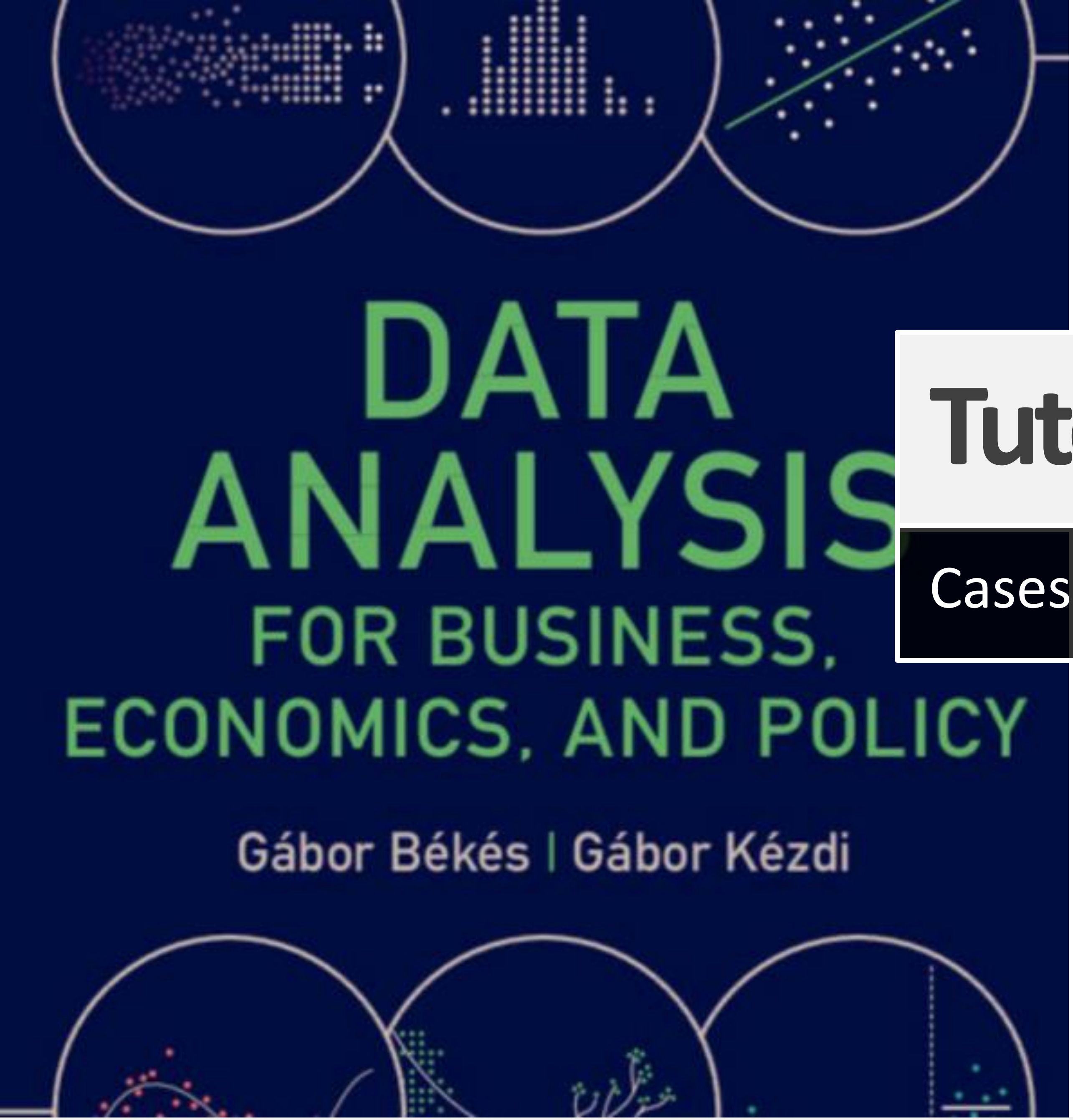


Lecturer: Illés Ferenc

Tutorials

Cases from the book accessible online

In the tutorials we use cases from the book

& we also use real financial data to gain in-depth understanding for applied finance research work, relevant for the industry

Issues/Problems

- If you had any issues with Week 1 health data or with week 3 data, examining the company managerial example efficiency, please ask online as soon as possible or here in the tutorial.

Tutorial (Week 3)

- Wednesday:

  Family firm data

  Are family firms run better,

  Have better management?

  Incentives may be more aligned?

Data

Code, R

- Friday
  - We go back to using week 1 data
  - Matching based on people' characteristics

Recap: What is Causality

- Lets consider again, the company management example:

- Y = a + b*FamilyfirmDummy + c*company size + e

- We cannot infer here if Y outcome, managerial efficiency or quality is really a result of family firm structure or some other confounding variables, such as profitability. If all or most family firms are more profitable, or active in high profit margin industries, then the management may be better quality because of the “preselection” that the family firms are more profitable.

World management Survey data

- You have observational data for many possible reasons.
- Experiments may be hard, expensive, unethical

Nowadays experiments on people (Human trials have to go through ethic committee approval, and sensitive questions cannot be asked without opt out options)

CH01C Management quality: data collectionPermalink

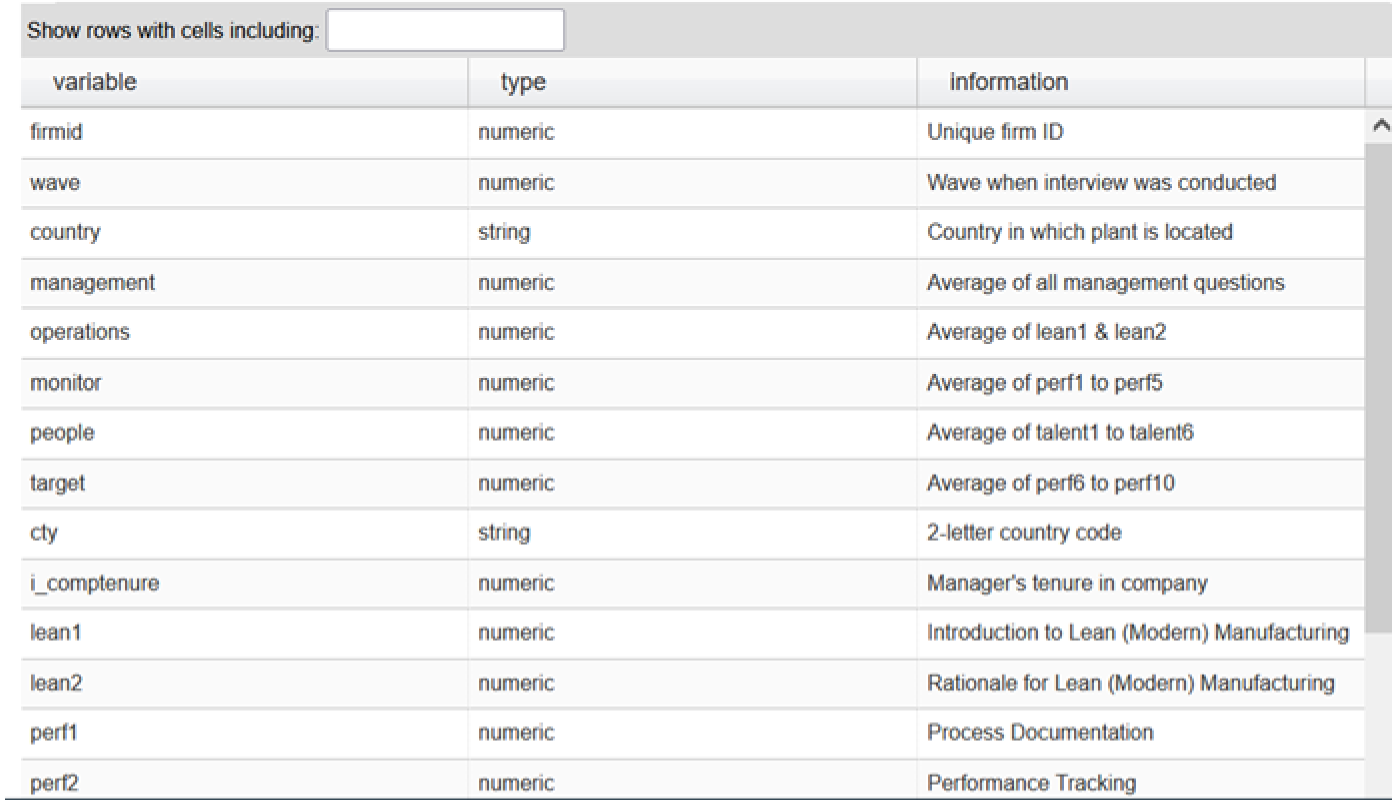
How different are firms and other organizations in the terms of their management practices? Is the quality of management related to how large the firms are? Is it affected by whether the owners are the company founders or their families? To answer these, and many related, questions, we need data on management quality. Such data was collected by the World Management Survey (WMS; https://worldmanagementsurvey.org/), an international research intitative to measure the differences in management practices across organizations and countries.

Case study: Family firms and Quality of Management

- So, we would like to match each family firm with another non family firm and compare the managerial outcome
- **Doing proper Apples to Apples comparison ☺**

Compare apples with apples ☺

Case study: Family firms and Quality of Management

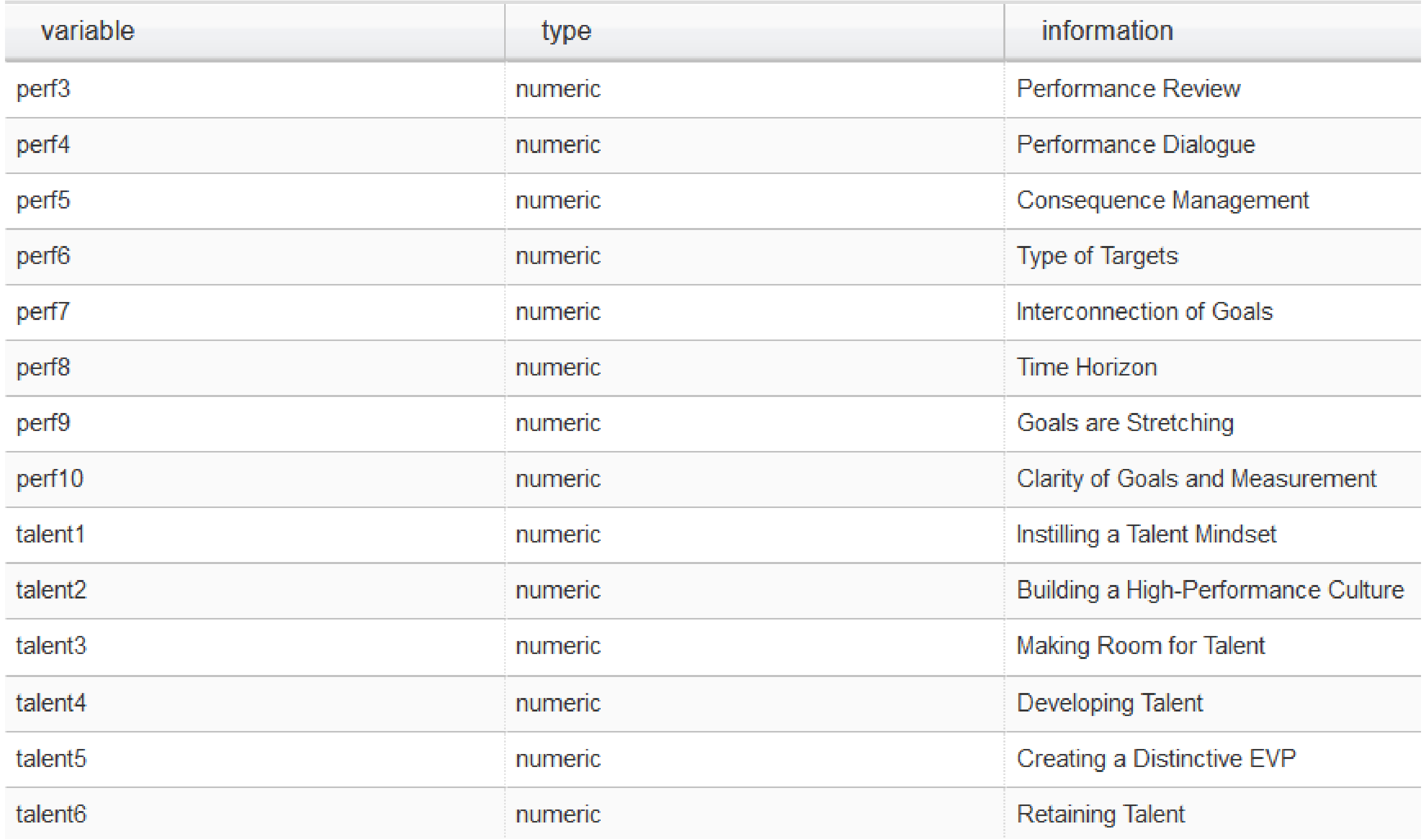


Show rows with cells including:

variable	type	information
firmid	numeric	Unique firm ID
wave	numeric	Wave when interview was conducted
country	string	Country in which plant is located
management	numeric	Average of all management questions
operations	numeric	Average of lean1 & lean2
monitor	numeric	Average of perf1 to perf5
people	numeric	Average of talent1 to talent6
target	numeric	Average of perf6 to perf10
cty	string	2-letter country code
i_comptenure	numeric	Manager's tenure in company
lean1	numeric	Introduction to Lean (Modern) Manufacturing
lean2	numeric	Rationale for Lean (Modern) Manufacturing
perf1	numeric	Process Documentation
perf2	numeric	Performance Tracking

- Variables
- **Great way to learn about coding, how efficiently name variables, so you can recognize them later keep them tights. Never use space in variable names if possible keep them all lower case. Upper / lower case matters in some software solutions.**

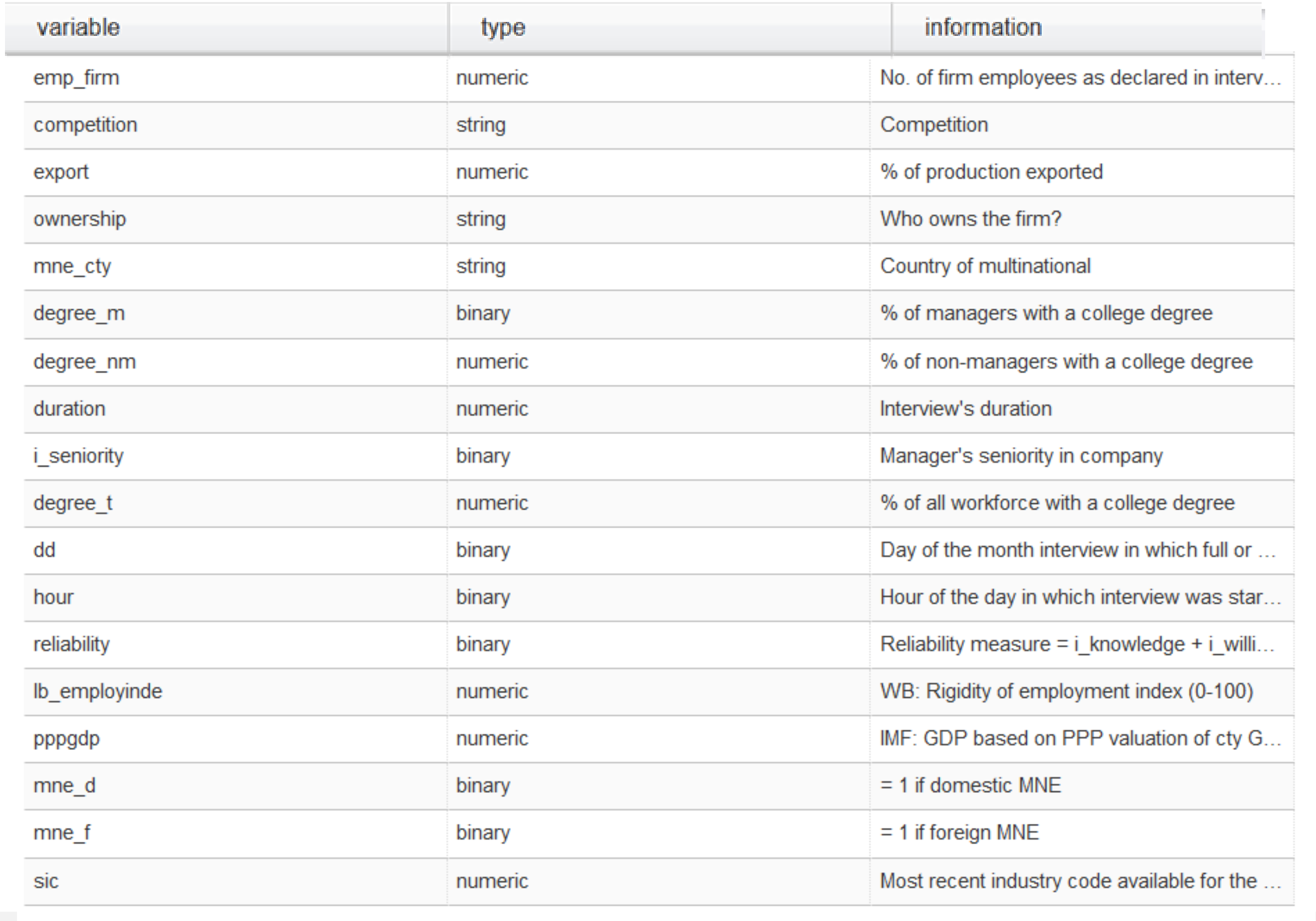
Case study: Family firms and Quality of Management



variable	type	information
perf3	numeric	Performance Review
perf4	numeric	Performance Dialogue
perf5	numeric	Consequence Management
perf6	numeric	Type of Targets
perf7	numeric	Interconnection of Goals
perf8	numeric	Time Horizon
perf9	numeric	Goals are Stretching
perf10	numeric	Clarity of Goals and Measurement
talent1	numeric	Instilling a Talent Mindset
talent2	numeric	Building a High-Performance Culture
talent3	numeric	Making Room for Talent
talent4	numeric	Developing Talent
talent5	numeric	Creating a Distinctive EVP
talent6	numeric	Retaining Talent

- Variables 3
- **Take note all the variables, you need to be aware of the variables for your projects/ work, to know what you can work with.**
- **And ultimately, you also have to have an idea of what variables you are missing**

Case study: Family firms and Quality of Management



variable	type	information
emp_firm	numeric	No. of firm employees as declared in interv…
competition	string	Competition
export	numeric	% of production exported
ownership	string	Who owns the firm?
mne_cty	string	Country of multinational
degree_m	binary	% of managers with a college degree
degree_nm	numeric	% of non-managers with a college degree
duration	numeric	Interview's duration
i_seniority	binary	Manager's seniority in company
degree_t	numeric	% of all workforce with a college degree
dd	binary	Day of the month interview in which full or …
hour	binary	Hour of the day in which interview was star…
reliability	binary	Reliability measure = i_knowledge + i_willi…
lb_employinde	numeric	WB: Rigidity of employment index (0-100)
pppgdp	numeric	IMF: GDP based on PPP valuation of cty G…
mne_d	binary	= 1 if domestic MNE
mne_f	binary	= 1 if foreign MNE
sic	numeric	Most recent industry code available for the …

- Variables 3
- **You may not find the expected results, may not be able to “nail down” causality which could be partly due to inappropriate controls, or because of “overspecification” , putting in too many controls. (some of which could be highly correlated measure the same thing)**

Exact Matching of Family firms

- **What are the variables we cant to consider for exact matching?**

```
Hmisc::describe(data$management)
data <- data %>%
    mutate(
        empbin5 = cut(emp_firm, quantile(emp_firm, seq(0,1,1/5)), include.lowest = TRUE, right = FALSE),
      agecat = (age_young == TRUE) + 2*(age_mid == TRUE) + 3*(age_old == TRUE) + 4*(age_unknown == TRUE))

data_agg <- data %>%
    group_by(degree_nm_bins, agecat, competition, empbin5, industry, countrycode) %>%
 dplyr::summarise(
        n = n(), n0 = sum(1-foundfam_owned), n1 = sum(foundfam_owned),
        y0 = sum(management*(foundfam_owned == 0))/sum(1-foundfam_owned),
        y1 = sum(management*(foundfam_owned == 1))/sum(foundfam_owned)
    ) %>%
    ungroup()
```

Exact Matching of Family firms

- **What are the variables we cant to consider for exact matching?**

```
    group_by(degree_nm_bins, agecat, competition, empbin5, industry, countrycode) %>%
dplyr::summarise(
            n = n(), n0 = sum(1-foundfam_owned), n1 = sum(foundfam_owned),
            y0 = sum(management*(foundfam_owned == 0))/sum(1-foundfam_owned)
```

So matching firm identified in bins., in the same firm age category, roughly same firms size based on same employee bin, industry competition, industry and country code

Are we satisfied matching based on these variables?

Exact Matching of Family firms

- **Compare the results with matching.**
- **Are the management of family firms, relative to matched control firms better or worse?**

```
# ATE/ATET
data_agg %>%
        filter(n0>0 & n1>0) %>%
        summarise(ATE = weighted.mean(y1-y0, n), ATET =
weighted.mean(y1-y0, n1))
```

PSA Matching of Family firms

```
# *************************************************************************
# * Matching on the propensity score
# *************************************************************************

# NOTE: the R code calculates ATET with the estimand=="ATT" option

# Function only works with non-missing values and factor variables
data_pscore <- data %>%
  dplyr::select(all_of(c(y_var, x_var, control_vars, control_vars_to_interact))) %>%
  na.omit() %>% mutate( industry = factor( industry ),
                        countrycode = factor( countrycode ) )

# with all control vars ---------------------------------------------------
```

PSA Matching of Family firms

```
# Step 1 - Matching
formula_pscore1 <- as.formula(paste0(x_var, " ~ ",
             paste(c(control_vars, control_vars_to_interact), collapse = " + ")))
mod_match <- matchit(formula_pscore1,
               data = data_pscore,
               method = 'nearest', distance = 'logit', replace=TRUE, estimand="ATT")
summary(mod_match)
```

```
# Step 2 - restrict data to matched
data_match <- match.data(mod_match)

# Please note that nhe "number of matched observations" calculated by
# this code varies marginally from the one on p607 in the textbook.
dim(data_match)
```

PSA Matching of Family firms

```
# Step 3 - Estimate treatment effects
# NOTE: We use weights here,to account for control observations that were matchet to
multiple treated osb
#      This is different from weights used to estimate ATE!
reg_match <- feols(management ~ foundfam_owned,
                data = data_match,
               weights = data_match$weights
                )

out1 <- summary(reg_match)

ATET_PSME1 <- out1$coefficients[2]
ATET_PSME1_SE <- out1$se[2]
```

PSA Matching of Family firms

```
# with all controls + interactions ---------------------------------------------

# Step 1 - Matching
formula_pscore2 <- as.formula(paste(x_var, " ~ " ,
        paste(control_vars_to_interact, collapse = ":"),
        " + (", paste(control_vars, collapse = "+"),")*(",
        paste(control_vars_to_interact, collapse = "+"),")",sep=""))

mod_match2 <- matchit(formula_pscore2,
                   data = data_pscore,
                   method = 'nearest', distance = 'logit', replace=TRUE, estimand="ATT")

summary(mod_match2)
```

PSA Matching of Family firms

```
# Step 2 - restrict data to matched
data_match2 <- match.data(mod_match2)
# Please note that nhe "number of matched observations" calculated by
# this code varies marginally from the one on p607 in the textbook.
dim(data_match2)
# Step 3 - Estimate treatment effects
# NOTE: We use weights here,to account for control observations that were matchet to multiple treated osb
#       This is different from weights used to estimate ATE!
reg_match2 <- feols(management ~ foundfam_owned,
            data = data_match2, weights = data_match2$weights)

out2 <- summary(reg_match2)
ATET_PSME2 <- out2$coefficients[2]
ATET_PSME2_SE <- out2$se[2]
```

PSA Matching of Family firms – Checking Comm Support

```
# ****************************************************************
# * CHECK common support
# ****************************************************************

# Country, cometition, industry
c1 <- CrossTable(data$foundfam_owned, data$compet_moder, na.rm=T )
c2 <- CrossTable(data$foundfam_owned, data$compet_strong, na.rm=T)

i <- CrossTable(data$foundfam_owned, data$industry, na.rm=T)
c <- CrossTable(data$foundfam_owned, data$countrycode, na.rm=T)


cbind(c1$prop.row, c2$prop.row, i$prop.row, c$prop.row)
```

PSA Matching of Family firms – Checking Comm Support

```
# ************************************************************************
# * CHECK common support  CONTINUED here
# ************************************************************************


…
# College Degree
data %>%
  group_by(foundfam_owned) %>%
  summarise(min = min(degree_nm , na.rm=T),
            max = max(degree_nm , na.rm=T),
            p1 = quantile(degree_nm , probs = 0.01, na.rm=T),
            p5 = quantile(degree_nm , probs = 0.05, na.rm=T),
            p95 = quantile(degree_nm , probs = 0.95, na.rm=T),
            q99 = quantile(degree_nm, probs = 0.99, na.rm=T),
            n = n())
```

PSA Matching of Family firms – Checking Comm Support

```
# ***************************************************************
# * CHECK common support  CONTINUED here
# ***************************************************************
…
# Employment
data %>%
  group_by(foundfam_owned) %>%
  summarise(min = min(emp_firm , na.rm=T),
            max = max(emp_firm , na.rm=T),
            p1 = quantile(emp_firm , probs = 0.01, na.rm=T),
            p5 = quantile(emp_firm, probs = 0.05, na.rm=T),
            p95 = quantile(emp_firm, probs = 0.95, na.rm=T),
            q99 = quantile(emp_firm, probs = 0.99, na.rm=T),
            n = n())
# * common support check passed
```

Run regression with matched data

Now,

Implement the matching regressions.

Calculate the difference in price of airline ticket in routes effected operated by the effected airlines versus unaffected airlines.

Reflect on Matching

Which matching do you find convincing?

Would you do matching?

Is it better then systematically control for characteristics,

- What are the gains
- What are the costs?

- Reflect with your “study mates” or online in the Moodle discussion forum

21. Regression … Data Analysis by Bekes and Kezdi

End

NOTE:

We are now in the last stretch of the semester, so it is a good time to reflect.

Do you understand the difference between correlation and Causality?

Think of a business example where you can measure causality and share.

Tutorial (Week 4)

- Wednesday:

  Family firm data

  Are family firms run better,

  Have better management?

  Incentives may be more aligned?

- Friday
  - IV, and regression discontinuity example
  - We go back to using week 1 data
  - Try to match in this data, based on personal charactheristcs

Data

Code, R

Health data – last week

Last week, we used again the health data, and try to create a regression discontinuity analysis, where we focused on the “break” at 65

Specifically, we tested whether retirement or enrolment in medicare and/or retirement has a positive effect in blood pressure.

Make sure, you were able to create a scatter plot logBP against logage, or also BP against Age, and perhaps zoom in to the 65 year range.

Include other controls : hh_income, Sdummy_rave, Sdummy_edu

Health data – regression discontinuity

So last week the dependent variable was blood pressure of the "focal" person and tried to explain his/her blood pressure with food consumption, and age effectively. We have done the following regressions:

a) Log(BP) = a+ b*Dummy$_{65}$ + β1*goodvegi + β2*Reallygoodvegi + c1*65_GV+ c2*65_RGV + u

b) Log(BP) = a+ b*Dummy$_{65}$ + β1*goodvegi + β2*Reallygoodvegi + c1*lnAge + c2*lnAge2+ + c1*65_GV+ c2*65_RGV + d1*Dwoman + u

c) Log(BP) = a+ b*Dummy$_{65}$ + β1*goodvegi + β2*Reallygoodvegi + c1*lnAge + c2*lnAge2+ + c1*65_GV+ c2*65_RGV + d1*Dwoman + d2*Wo_GV + d3*Wo_RealGV + u

Include other controls : hh_income, Sdummy_rave, Sdummy_edu

Health data – matching, why?

— Let’s compare women and man, women tend to eat more vegi (our data supports that).

— As mentioned last week, we need to address the concern that there are differences across man and women, and across different age groups.

—It would be cool, if we could have identical twins (also called monozygotic twins who share the same genomes and are always of the same sex.

- □ And we could run an experiment, where one twin would be eating healthy , lots of vegetables for 2 years, while the other twin had very little vegetables
- □ ,…. But these kind of human experiments hard to do, and most of the case, prohibited because it would be constituting to human experimentations. The twin who is not eating lots of vegetables could be hurt in the trial.

Tutorial (Week 4) - Matching in health data

- Let’s check what characteristics we can match on, what “buckets” we can use.
- Try to match based on age buckets, income, education using the code from the previous class.
- Create first a running number, id in the original file, when we match, we do not want to match the person with him/hers so we have to exclude those matches

Match based on:

- Gender , column WC in the excel file if you need to find ☺
- Race and education, columns “WE” and “WF”
- age_cat, column “XL”
- Married dummy variable, Column “XM”
- Income_cat, column “XQ”

Health data – Matching step 1, create match file

When we match, we should save a matchbase file, where we should keep the following data:

- Id for the person, we created
- blood_pressure, bmi, heart_risk, weight, height
- veggies fruits veggies_n_fruits veggies_gr fruits_gr veggies_n_fruits_gr
- And the controls we use: gender, race, education, age_cat, married, income_cat

In the matchbase.xls file, generate new variables

Mch_bp= blood_pressure (and drop the blood pressure variable, not to override)

Mch_bmi= bmi

Mch_heartrisk=heart_risk….

And so on

Health data – Matching 2 – exact match

Base data

ID	Bp	Age_cat	Income_cat	Married
1	202	aged 30-39	low	1
2	210	aged 60-69	low	1
3	199	aged 18-29	low	0
4	250	aged 40-49	mid	0

Match data

ID	MchAge_cat	MchIncome	MchMarried
10	aged 30-39	High	1
22	aged 18-29	Mid	1
35	aged 30-39	Mid	1
120	aged 30-39	low	1

Only this is match based on all criteria we considered here

Health data – Matching 2 – exact match

Base data

ID	Bp	Age_cat	Income_cat	Married
1	202	aged 30-39	low	1
2	210	aged 60-69	low	1
3	199	aged 18-29	low	0
4	250	aged 40-49	mid	0

Match data

ID	MchAge_cat	MchIncome	MchMarried
10	aged 30-39	High	1
22	aged 18-29	Mid	1
35	aged 30-39	low	1
120	aged 30-39	low	1

What happens if there are multiple match to the same person:

A- maybe you can consider more "finer" matching, matching on more stricter age group for example, or more stricter income category

B- you take the averages of the match person values, and use that as a reference

Health data – difference regression - matching

Once, we matched we have to create difference variables:

Bpdiff = BP of focal person - BP of matched person, new dependent variable.

If there are multiple matches, then

Bpdiff= BP of the focal person (the observation in the original file) - average of the BP of the matched people

We also want to create the difference variables for controls:

- blood_pressure, bmi, heart_risk, weight, height
- veggies fruits veggies_n_fruits veggies_gr fruits_gr veggies_n_fruits_gr
- And the controls we use: gender, race, education, age_cat, married, income_cat

Health data – difference regression - matching

Rerun regressions,

(BP) = a+ b*Dummy$_{65}$ + β1*goodvegi + β2*Reallygoodvegi +u

we did before but now with differences

Please do the following regressions:

a) BPdiff = a+ b*DiffinVegigr + e
   - □ Where diffinvegigr is the difference in the grams of vegetables consumed by the person we examine in comparison with a matched person

b) BPdiff = a+ b*DiffinVegigr +c* Dummy65+ e
   - □ Since last class we show that there is a regression discontinuity around 65, perpaphs is is good to control for that since the age category covers the whole decade and a 62 year old maybe be still working, stressed, while a 66 year old is retired.

Health data – difference regression - matching

Are we more confident, that Vegetables consumption has positive benefits.

Perhaps, we can “nail down” the results with using alternative potential output (PO) measures, such as the risk of heart problems, and the person bmi idex which are available in the file.

In empirical analysis, using secondary data, when we cannot clearly establish causality, one of the potential technique or tool to test different outcome variables which could provide corroborative evidence.

Health data – difference regression - matching

- In Week 5, we learn about the Difference in differences technique.
- Here, we already have seen that there is difference across Man and Women. In general Man tend to have higher blood pressure even if they live a healthy life style.
- Test:
  - Bmi (=Y) = a+ b*Vegetable consumption + c controls
  - BMI difference (BMI of the person – BMI of the matched firm = a+ b*difference in Vegetable consumption + c*…

Health data – regression with DID

- In Week 5, we learn about the Difference in differences technique.
- Here, we already have seen that there is difference across Man and Women, and matched for it. Next week, instead of matching, which has lots of potential problems. Could be computational intensive, matching may not be possible for half the sample etc…
- So next week, we go back to the drawing board and consider testing for causality using the difference in differences approach.
-