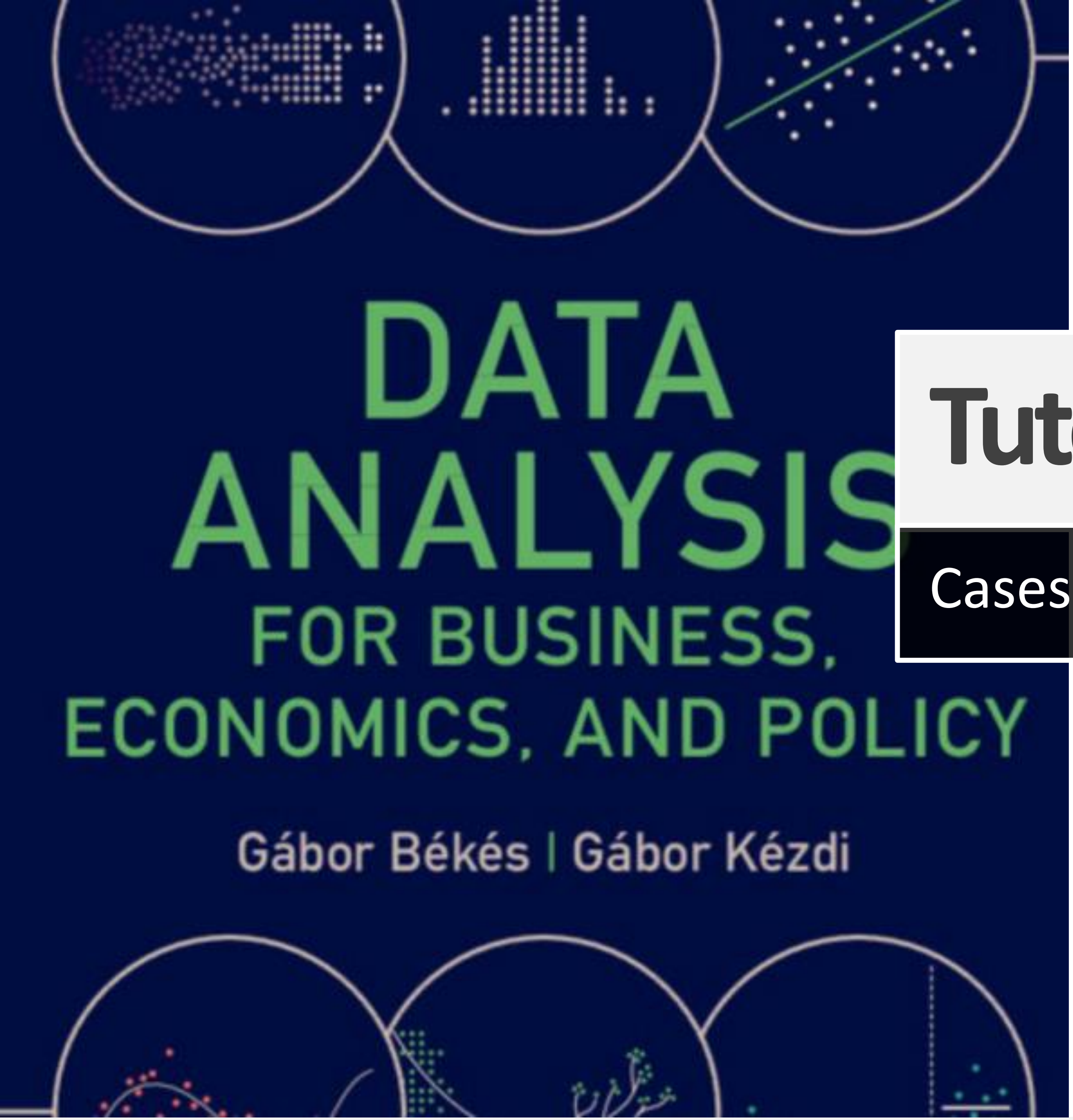


DATA ANALYSIS FOR BUSINESS, ECONOMICS, AND POLICY

Gábor Békés | Gábor Kézdi

Lecturer: Attila Vig

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 are still struggling with last week data, the health data, you need to ask help from your classmates, teammate(s) and/or post question on the Moodle Forum (under week 3, or under the previous week's tutorial, where you got stock.
- This week, we are working with a very complicated Managerial database which we will also use next week, so please understand that working with data takes time, and requires 100%+ attention.
  - You cannot do data analytics, programs while chatting with other people, or watching TV… etc
  - Again back to the rules from Week 1 tutorial: on of the most important thing, is to first understand your data, the variables, the distribution of the variables and delete data only iff strictly necessary.

Tutorial (Week 3)

- Wednesday:
  Family firm data
  Are family firms run better,
  Have better management?
  Incentives may be more aligned?

Data

Code, R

- Friday
  - IV, and regression discontinuity example
  - We go back to using week 1 data

Recap: What is Causality

- Causality – is about interpretation
- You see a pattern in the data – revealed by regression analysis
- Then, you interpret it....
- unless...
  - I you get to design your own experiment
  - in that case you have a causal effect in mind and you induce controlled variation a variable
  - if all goes fine you know how to interpret patterns

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)

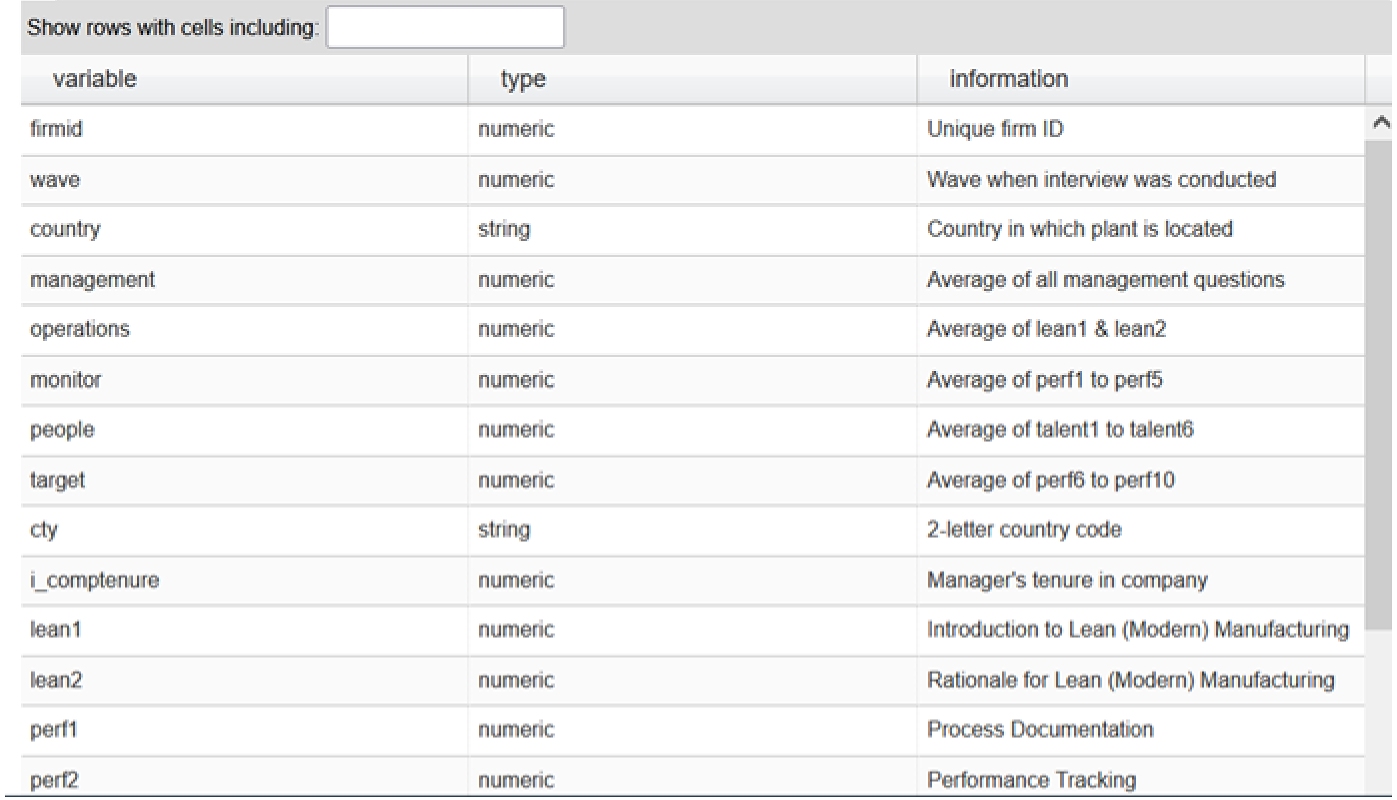
CH01C Management quality: data collectionPermalink

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 initiative to measure the differences in management practices across organizations and countries.

Case study: Family firms and Quality of Management

- Though experiment
- We investigate whether the fact that a company is owned by its founder, or their family members, has an effect on the quality of management.
- Whether founder/family owned companies are better or worse managed than other firms, on average because of their ownership.
- This is a causal question: we are after an effect.
- Great way to understand what the intervention and the counterfactuals are.

Case study: Family firms and Quality of Management

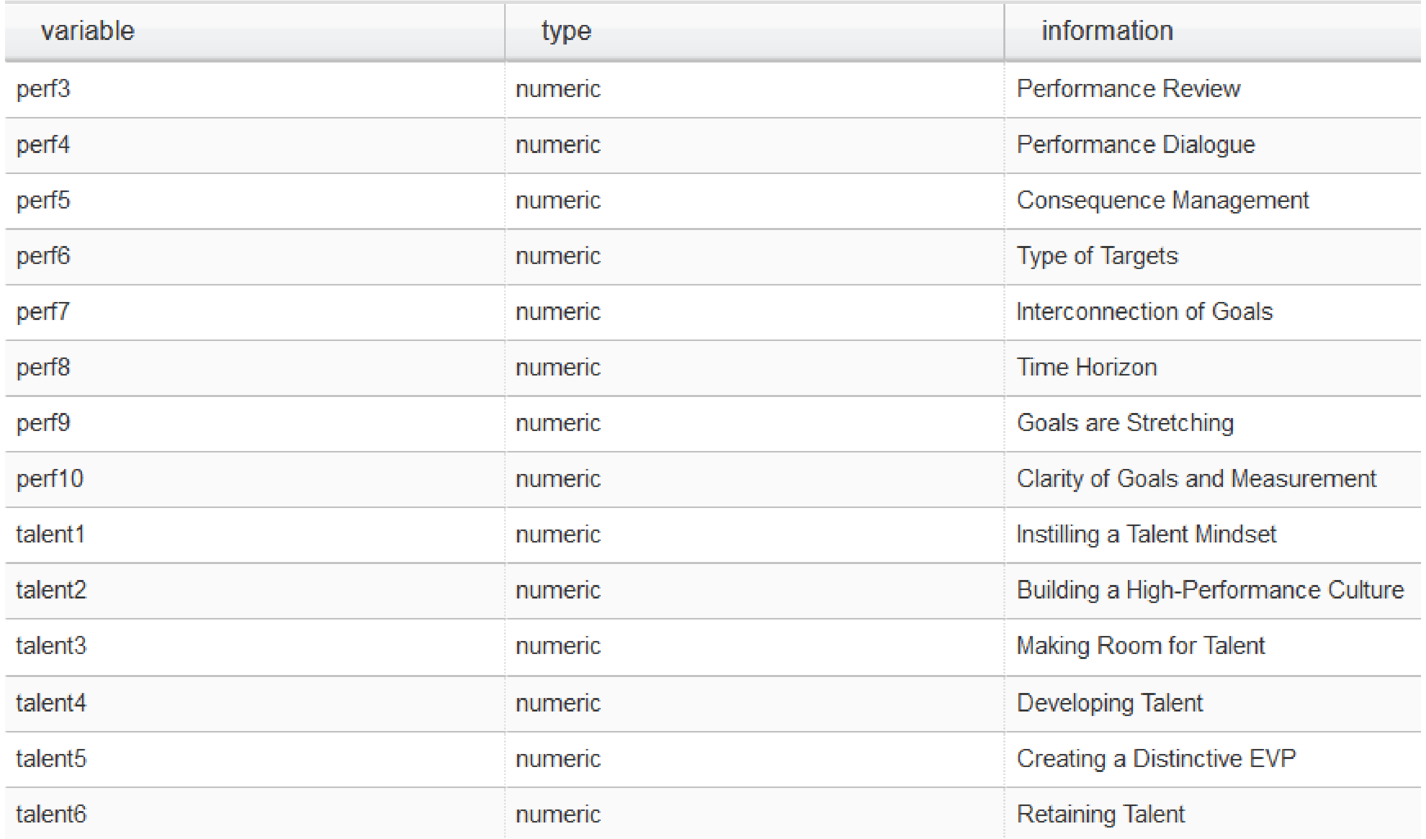


Show rows with cells including:

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.**

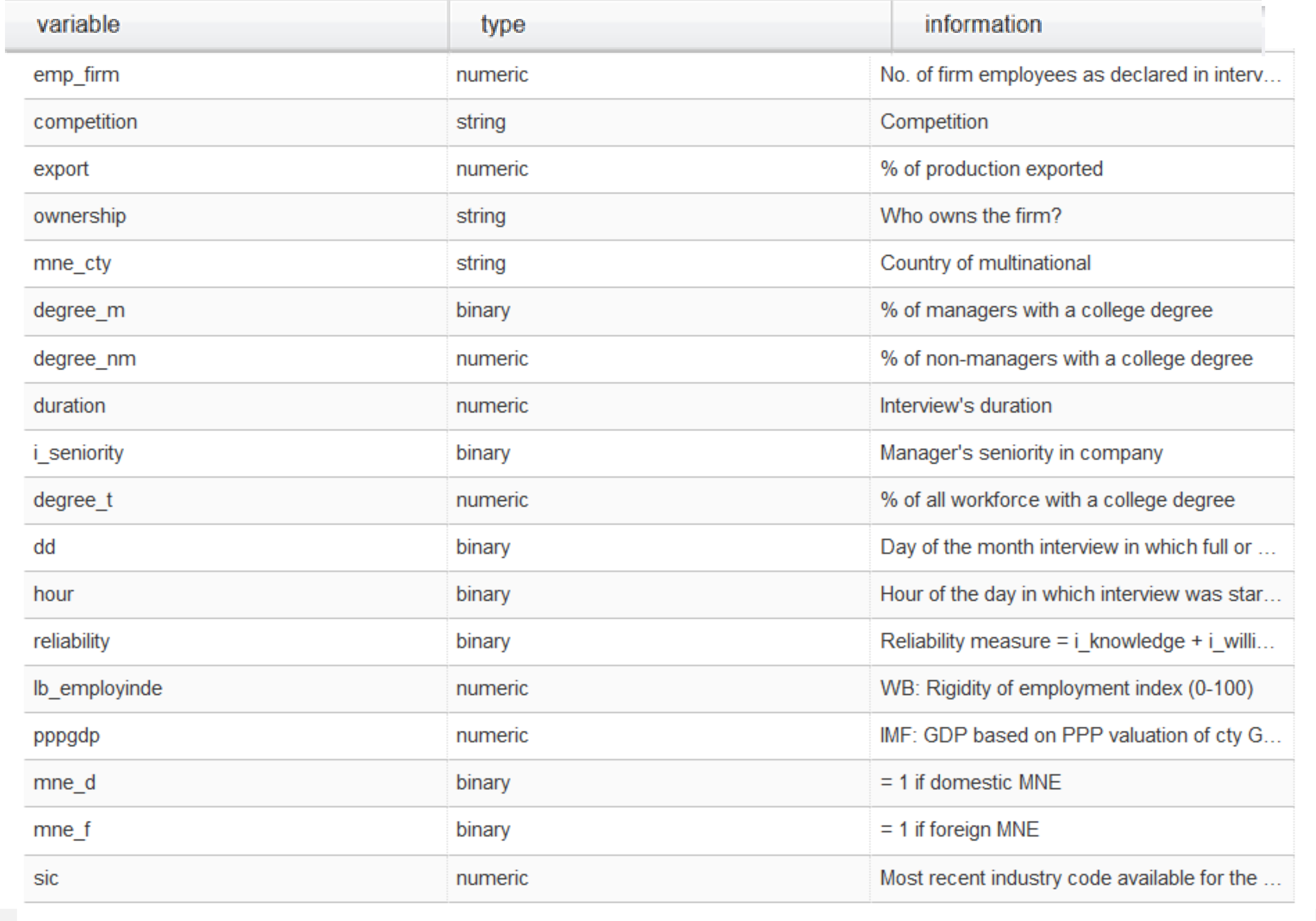
Case study: Family firms and Quality of Management



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 corre[illegible]d measure the same thing)**

21. Regression … Data Analysis by Bekes and Kezdi

Case study: Family firms and Quality of Management

- Next data preparation
- **In an R code (link under URL icon), the authors show how they identify a firm as being family firms, and create some crucial variables such as competition for the industry.**

```
data <- data %>%
        mutate(
                compet_weak = factor(competition == "0 competitors" | competition == "1-4 competitors"),
                compet_moder = factor(competition == "5-9 competitors"),
                compet_strong = factor( competition == "10+ competitors")
        )
data %>%
  group_by(competition) %>%
  summarise(weak = max(compet_weak == TRUE),
           moder = max(compet_moder == TRUE),
           strong = max(compet_strong == TRUE))
```

Case study: Family firms and Quality of Management

- Next data preparation, creating dummy variables

```
# age
data <- data %>%
  mutate(age_young = factor(firmage<30 & !is.na(firmage)),
         age_old = factor(firmage>80 & !is.na(firmage)),
         age_unknown = factor(is.na(firmage)),
         age_mid = factor(age_young == FALSE & age_old == FALSE & age_unknown == FALSE))
```

You really need to think through the variables when you create them whether they make sense.

- NOTE The authors do not use firm age as liner value, because ex ante we do not expect that a 30 year old firm management is better in comparison with a 10 year old firm, in the same way as a 80 years and 60 years old

Case study: Family firms and Quality of Management

- Next data preparation,

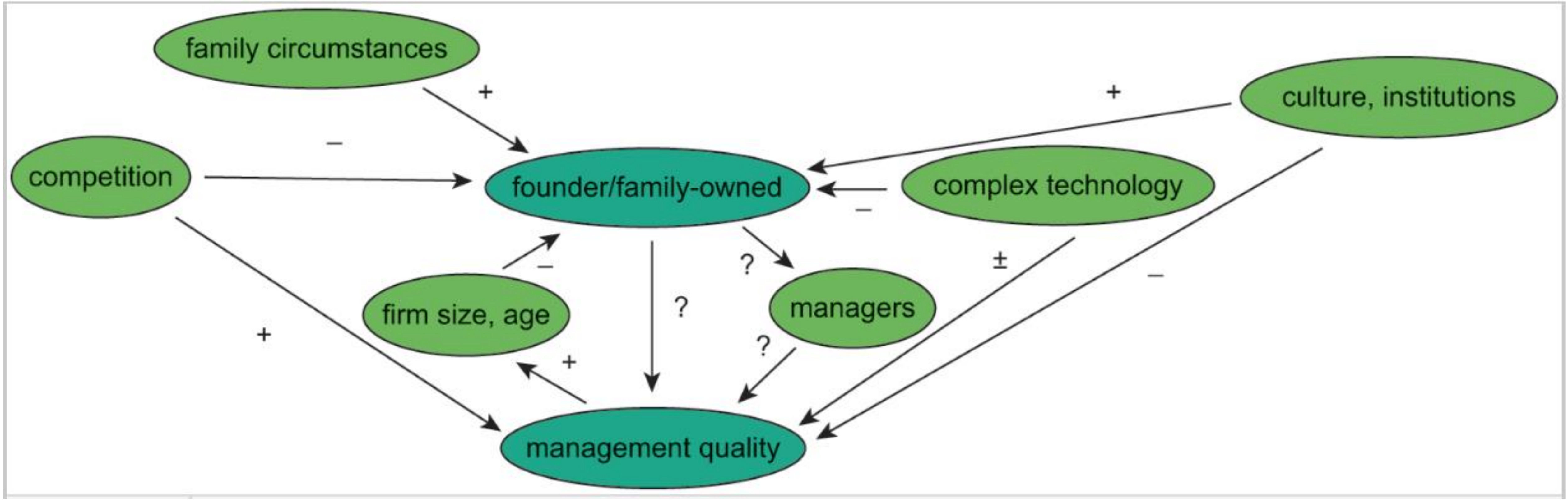
```
# Drop tiny and large firms
data %>%
  filter(emp_firm<50)  %>%
  summarise(n = n())
data %>%
  filter(emp_firm>5000)  %>%
  summarise(n = n())
data <- data %>%
  filter (!(emp_firm<50 | emp_firm>5000))
# Save workfile ----------------------------------------------------
write_csv(data, paste0(data_out, "wms_da_textbook-work.csv"))
# N=8439
```

Sometimes you may want to truncate the data to exclude the extreme outliers, that is what the authors are doing

Case study: Family firms and Quality of Management data, and understanding a thought experiment

- Observational cross-sectional data
- World Management Survey = cross-section of many firms in manufacturing from 21 countries.
- The outcome variable is the management score.
- The causal variable is founder/family ownership.
- Several tasks before running regressions
  - Think about and identify sources of variation in ownership,
  - Draw a causal map
  - Decide on observable variables to condition on

Case study: Family firms and Quality of Management
Causal Directed Acyclic Graphs (DAGs)



Case study: Family firms and Quality of Management thought experiment : Variation in Ownership (X)

- Let us look for variation in x, ownership. Think + identify + decide.

- So we want to test whether the variation in Ownership affects (cause) better management

- Let's recreate the book example
- *OPEN the R code for analysis from the github.*
- *https://github.com/gabors-data-analysis/da_case_studies/blob/master/ch21-ownership-management-quality/ch21-wms-02-analysis.R*

Case study: Family firms and Quality of Management
thought experiment : Variation in Ownership (X)

- *Prepare for the analysis, for formatting, pull in the data,*

```
# set data dir, load theme and functions
source("ch00-tech-prep/theme_bg.R")
source("ch00-tech-prep/da_helper_functions.R")

# data used
source("set-data-directory.R") #data_dir must be first defined #

use_case_dir <- file.path("ch21-ownership-management-quality/")

data_in <- use_case_dir
data_out <- use_case_dir
output <- paste0(use_case_dir,"output/")
create_output_if_doesnt_exist(output)
```

Case study: Family firms and Quality of Management

thought experiment : Variation in Ownership (X)

- *Prepare for the analysis, for formatting, pull in the data,*

```
# Read in data -----------------------------------------------------------
data <- read_csv(paste0(data_out, "wms_da_textbook-work.csv"))
data %>%
  group_by(foundfam_owned) %>%
  summarise (mean(management))
# Set variables to use --------------------------------------------------
y_var <- "management"
x_var <- "foundfam_owned"
control_vars <- c("degree_nm", "degree_nm_sq", "compet_moder", "compet_strong",
                  "lnemp", "age_young", "age_old", "age_unknown")
control_vars_to_interact <- c("industry", "countrycode")
data %>%
        dplyr::select(all_of(c(control_vars, control_vars_to_interact))) %>%
        summary()
```

Case study: Family firms and Quality of Management thought experiment : Variation in Ownership (X)

- *Analysis*

- Regression of managerial quality without controls on X

```
# OLS with no control vars. -------------------------------------------------------
formula1 <- as.formula(paste0(y_var, " ~ ",x_var))
ols1 <- feols(formula1, data=data)

# OLS with all control vars -------------------------------------------------------

formula2 <- as.formula(paste0(y_var, " ~ ",x_var," + ",
              paste(c(control_vars, control_vars_to_interact), collapse = " + ")))

ols2 <- feols(formula2, data=data)
```

Case study: Family firms and Quality of Management
thought experiment : Variation in Ownership (X)

- *Analysis*

- Regression of managerial quality without controls on X

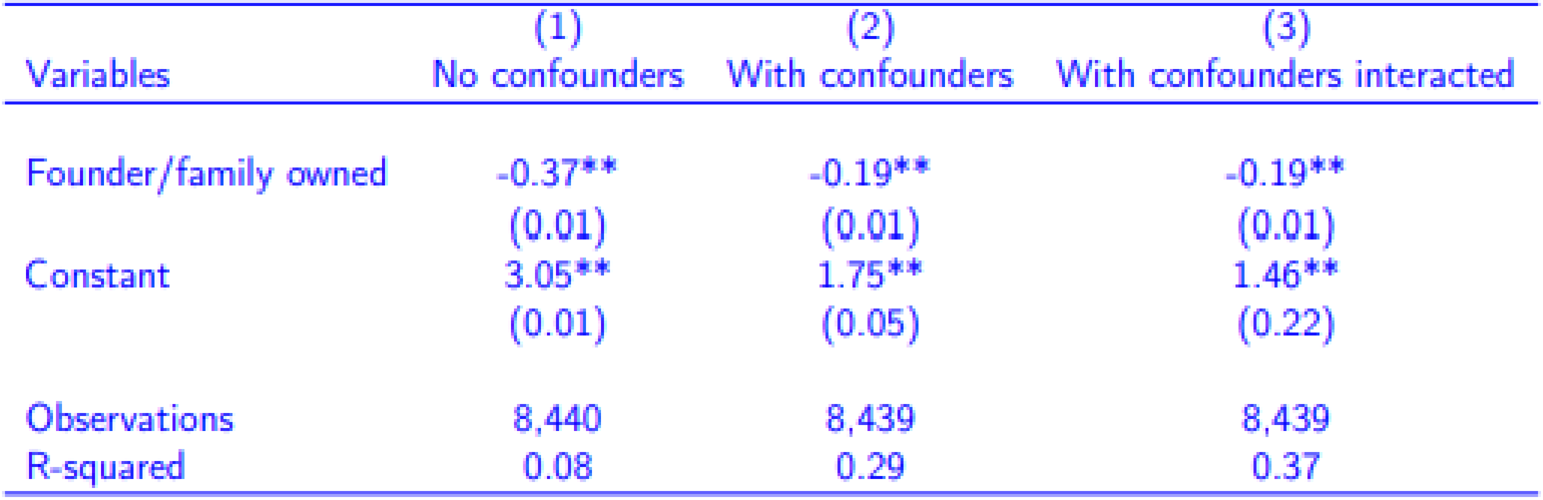
```
# OLS with no control vars. --------------------------------------------------
formula1 <- as.formula(paste0(y_var, " ~ ",x_var))
ols1 <- feols(formula1, data=data)

# OLS with all control vars --------------------------------------------------

formula2 <- as.formula(paste0(y_var, " ~ ",x_var," + ",
                paste(c(control_vars, control_vars_to_interact), collapse = " + ")))

ols2 <- feols(formula2, data=data)
```

Case study: Family firms and Quality of Management thought experiment : Variation in Ownership (X)



	(1)	(2)	(3)
Variables	No confounders	With confounders	With confounders interacted
Founder/family owned	-0.37**	-0.19**	-0.19**
	(0.01)	(0.01)	(0.01)
Constant	3.05**	1.75**	1.46**
	(0.01)	(0.05)	(0.22)
Observations	8,440	8,439	8,439
R-squared	0.08	0.29	0.37

Note: Outcome variable: management quality score. Robust standard error estimates in parentheses.** p <0.01 and * means p<0.05.

This significance notation is a bit strange, please stick to convention *** if p<0.01, ** if p<0.05 and * if p<0.1. To clarify * means significance at 10% level, *** means significance at 1% level

Case study: Family firms and Quality of Management

thought experiment : Variation in Ownership (X)

- Let us look for variation in x, ownership. Think + identify + decide.
- Cultural and institutional factors, norms in a society. Affect cost of starting business, FDI. How about y?
- Likely endogenous source, culture, norms correlated with management, too.
- How about family features. Children of founders, their interests, skills. Clearly affects if ownership may be passed on. How about y?
- Likely exogenous - gender/number of kids not related to management quality
- This is the variation we need but not use as control!

Case study: Family firms and Quality of Management

Sources of Variation in Ownership

- Family circumstances – exogenous variation in x
- Competition – common cause confounder
- Culture and institutions – common cause confounder
- Technology, product type – common cause confounder
- Firm size, firm age – hard – may be mechanisms of reverse causality
- Feature of managers (their age, experience) – mechanism I which ones to control on?

Case study: Family firms and Quality of Management

thought experiment : Variation in Ownership (X)

- *Analysis, step 2, regressions with controls*

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

  formula2 <- as.formula(paste0(y_var, " ~ ",x_var," + ",
                 paste(c(control_vars, control_vars_to_interact), collapse = " + ")))

  ols2 <- feols(formula2, data=data)

  # OLS with all controls + interactions -----------------------------------------------------
  formula3 <- as.formula(paste(y_var, " ~ ",x_var," + ",
        paste(control_vars_to_interact, collapse = ":"),
        " + (", paste(control_vars, collapse = "+"),")*(",
        paste(control_vars_to_interact, collapse = "+"),")",sep=""))
  ols3 <- feols(formula3, data=data)
```

Case study: Family firms and Quality of Management
thought experiment : Variation in Ownership (X)

- *Analysis, step 2, regressions with controls*

- Recall how did we have controls, and interaction variables, the authors of the book created an array for the controls and the interactions :

```
y_var <- "management"
x_var <- "foundfam_owned"

control_vars <- c("degree_nm", "degree_nm_sq", "compet_moder", "compet_strong",
                  "lnemp", "age_young", "age_old", "age_unknown")
control_vars_to_interact <- c("industry", "countrycode")


data %>%
        dplyr::select(all_of(c(control_vars, control_vars_to_interact))) %>%
        summary()
```

Case study: Family firms and Quality of Management

Conditioning on Confounders by Regression

- Regression of Y on X with conditioning on observable confounder variables (z1, z2, ...):

$$y^E = \beta_0 + \beta_{1x} + \beta_2 z_1 + \beta_3 z_2 + \ldots (1)$$

Advice, since normally you have a long list of control, you can just include an array for controls in notation

- Note: β_1 always = estimate of average difference in y between observations that are different in x but have the same values for z1, z2, ... Even if not causal.
- If the z_1, z_2, ... variables capture all endogenous sources of variation, x is exogenous in the regression.
  - Conditional on z1, z2, ... , variation in x is exogenous.
  - OLS estimate of β1 is a good estimate of ATE of x on y.

Case study: Family firms and Quality of Management

to do and practice

Create at least 5 regression model output in a table as on slide 20, showing all controls with the exception of the country and industry controls and interactions of those. Normally, we would consider them as fixed effects and do not display them in a table.

1) *Model 1 : Management quality (Y) = a+ b*Family firm dummy (Dummy$_{ff}$) + e*
2) *Model 2 : (Y) = a+ b* Dummy$_{ff}$+ c* industry competition (IC) measure + e*
    - *See slide 10 for the industry competition measures.*
3) *Model 2 : (Y) = a+ b* Dummy$_{ff}$ + c1* IC+ c2** age_old + *c3* age_mid + *c4* age_young + …
4) *Create your own 2 models, please do not forget to add in the "so called fixed effects"*
5) …

Practice outputting a nicely formatted table

Case study: Family firms and Quality of Management
to do and practice

- *Practice some interaction… - this is lead up for future class on difference in differences.*

- *So we think there is some difference in management quality across firms which are family firms and which are not.*
- *But then, we also see that family management quality may also vary across industries.*
- *Now what if family firms are concentrated in certain industries, have we controlled for that ? How to deal with that?*

Case study: Family firms and Quality of Management

to do and practice

- *Practice some interaction… - this is lead up for future class on difference in differences.*

- *We can create Family firms dummy and industry dummies.*
- *We had only industry competition in the analysis, and industry dummies as controls,*
- *We can create industry family interactions: …*

- *Experiment more on your own, and discuss it on Moodle why this makes sense. We come back to the topic in 1—2 weeks* ☺

End

NOTE:

We are getting close to the middle of the course, 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 3)

- 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

Data

Code, R

Week3_subsampleofWeek1data

Tutorial (Week 3)

- First lets compare women and Man

- Maybe we are struggling, nailing down the vegetable consumption, but perhaps, we can examine the beneficial effect of retirement to access to health.

- We also need to address the concern that there are differences across man and Women. … we will get back to that next week again.

Regression Discontinuity

- Well, I probably should have looked for better data but I tried to just make do with what we have…. And I am getting a bit ahead of ourselves

	Male		Female	
	mean	median	mean	median
age	47.35032	47	47.02732	46
Vegigr	117.1391	37.625	127.3436	79
Fruitgr	125.5223	37.625	138.2346	90
65+ (cl W)	0.199806	0	0.192323	0
Income (cl L)	7.831634	7	7.382882	7

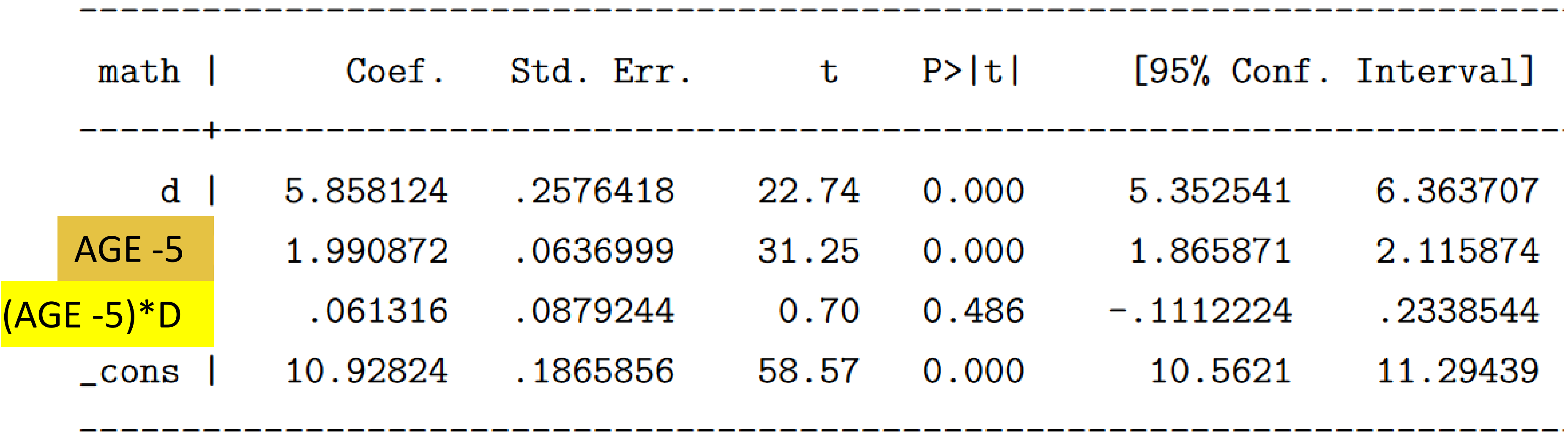
Health data – regression discontinuity

- Is retirement, help to reduce bloodpressure ? ☺
- At Age 65, most people retire in the USA, gain access to healthcare
- With age maybe bloodpressure also decline… so we need to control that

RECALL

RD *Example: - Preschool effect not from the book*

- Next, we can add the interaction term to allow for both shift in intercepts and shift in slopes:
  $math = \beta 0 + \beta 1D + \beta 2(age - 5) + \beta 3D * (age - 5) + u$



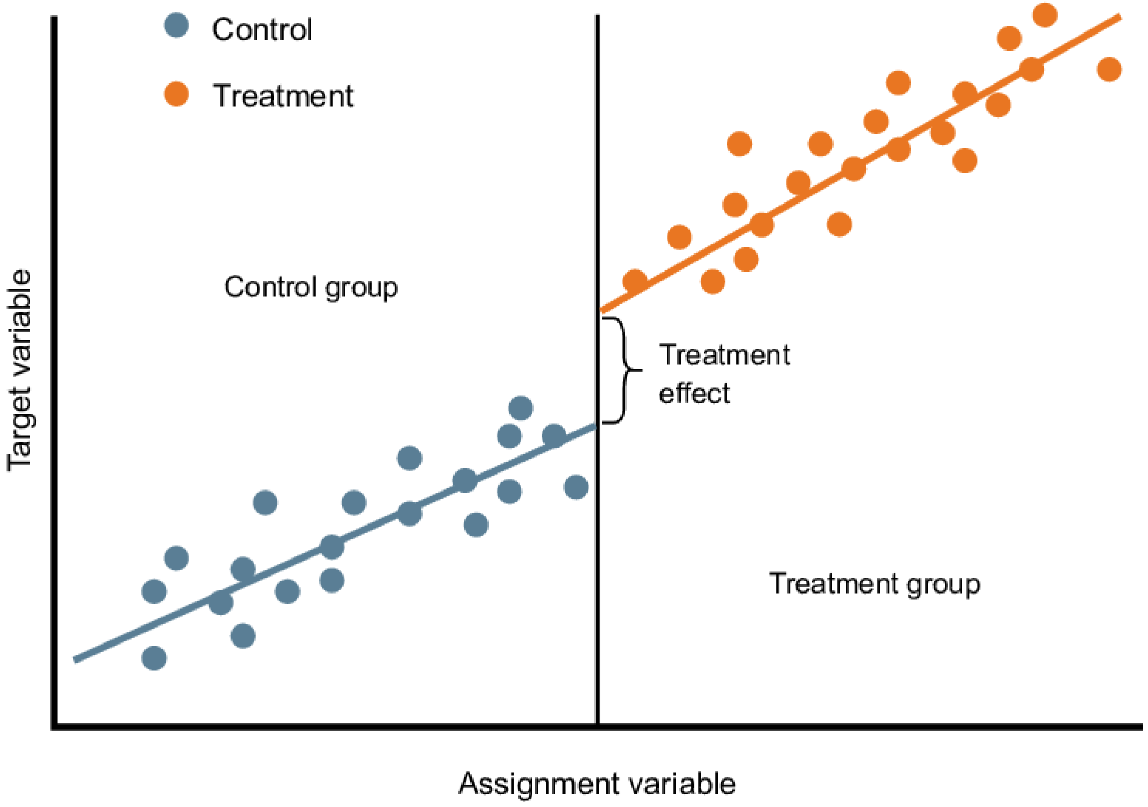
```
------------------------------------------------------------------------------
    math |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------+-----------------------------------------------------------------------
       d |   5.858124   .2576418    22.74   0.000     5.352541    6.363707
  AGE -5     1.990872   .0636999    31.25   0.000     1.865871    2.115874
(AGE -5)*D    .061316   .0879244     0.70   0.486    -.1112224    .2338544
   _cons |   10.92824   .1865856    58.57   0.000      10.5621    11.29439
------------------------------------------------------------------------------
```

The change in slopes before and after the cutoff value is captured by β3. Here we find no evidence for different slopes since its t value = 0.70 is small.

Health data – regression discontinuity

logBP	Natural logarithm of the blood pressure numbers added up
goodvegi	Dummy variable, takes on the value of one, if vegatable consumption is >200, zero otherwise
RgoodVegi	Dummy variable, takes on the value of one, if vegatable consumption is >300, zero otherwise
65cut	Dummy variable takes on the vakue of 1 if the respondent age is >65
LnAge	Natural logarithm of the respondent age in years
LnAge2	Squared term of the natural logarithm of the respondent age in years
hh_income	Incume stepwise dummy from the original data file, availabel from the book's authors
Dwoman	Dummy variable takes on the value of one if the respondent is women, zero otherwise
Sdummy_race	Stepwise race dummy from the original file
Sdummy__edu	Stepwise education dummy from original file
65_GV	Interaction variable of 65cut and goodvegi variables
Wo_GV	Interaction variable of Dwoman and goodvegi variables
65_RealGV	Interaction variable of 65cut and Rgoodvegi variables
Wo_RealGV	Interaction variable of Dwoman and RgoodVegi variables

We are testing treatment – “Medicare”



- Hopefully the treatment “Medicare enrollment or Retirement reduces blood pressure instead of increasing, and indeed that is what you should find.

- The assignment variable : Age
- Target variable: Blood pressure
- With Age, blood pressure increases, but there is a break

Health data – regression discontinuity

— Lets examine Log(BP) = a+ b*Dummy$_{65}$ + c1*lnAge + c2*lnAge2 …

Please do the following regressions:

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

b) Log(BP) = a+ b*Dummy$_{65}$ + β1*goodvegi + β2*Reallygoodvegi + c1*lnAge + c2*lnAge2 …

c) Log(BP) = a+ b*Dummy$_{65}$ + β1*goodvegi + β2*Reallygoodvegi + c1*lnAge + c2*lnAge2 …

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

Include other controls : hh_income, Sdummy_rave, Sdummy_edu

Health data – regression discontinuity

It may be useful to interact vegetable consumption with the 65 cutoff and woman dummy

Please do 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 – regression discontinuity

Create also a graph for blood pressure and age, and see whether it is possible to see visually the structural break at 65.

It may not be possible, as there are many confounding effect, change in income, etc.

But good idea to practice

Just 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

- What is the conclusion…

- Is there an impact of age65….?
  - Can we conclude…causality?

Health data – regression with DID

- We will continue this topic later n week 5 with difference in differences….

- 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.
- Women also tend to eat more vegetables, more likely to be vegetarian.
  - “Approximately 5% of people in the United States are vegetarians (Gallup, 2018), the majority of whom are women (Rosenfeld, 2018, Ruby, 2012). Accordingly, gender has played a central role in psychological investigations of vegetarianism (Rosenfeld, 2018, Ruby, 2012).”