

Lecture 2: Idea, Literature, and Data

Yi Chen

ShanghaiTech University

2021

Outline

- 1 Get Prepared Before You Start
- 2 About Research Ideas
- 3 Search and Read Literature
- 4 Publicly Available Data Sets

Outline

- 1 Get Prepared Before You Start
- 2 About Research Ideas
- 3 Search and Read Literature
- 4 Publicly Available Data Sets

ALWAYS GET YOUR WORK BACKED UP!!

You can never afford to lose your hard work!

- Backup in your portable hard disk periodically.
 - Also, don't delete your old do-files/papers. You may revert to an earlier version in the future.
 - Instead, just create a folder named "Backup."
 - We will introduce "version control" more formally in the next lecture.
- It is highly recommended to have a cloud storage service.
 - Three purposes: coordination (with others), synchronization (for yourself), and back-up.
 - Personally, I use Onedrive for personal usage and Dropbox for coordination.
 - "Version History" is a life-saving feature. . .
- You might need VPN.

Figure 1: Example of not backing-up...

9月12日，在南非约翰内斯堡，26岁的Noxolo Ntusi遭遇持枪抢劫，面对歹徒她顽强抵抗，只为保护论文。

Ntusi走在一条小路上，突然有两名劫匪跳下车还挥舞着枪，他们先是抢走了午餐袋，接着抢夺背包时，她冒着生命危险，死死抓住背包不放手。据报道，Ntusi是一名医学研究生，包里的硬盘装有她即将完成的硕士论文，如果论文被抢，就要延期毕业。

她说“我都快写完了，绝不能给他们。劫匪还想把我拖上车，我就蹲下抱成团，最后他们只能放弃。”

事后她也觉得自己的行为不太明智，告诫大家在遇到这类事件的时候，歹徒要什么就给他们，还是保命要紧。目前涉事嫌犯已被逮捕。

Dropbox Snapshot

Dropbox

Q Search



Name ▾	Modified ▾	Members ▾	⋮ ▾
 Coresidence and Intergenerational Transfer	--	2 members	...
 Econ11_Fall2011	--	9 members	...
 education	--	3 members	...
<input type="checkbox"/>  Family Planning ☆	--	2 members	<div>Share</div> <div>...</div>

Upload files

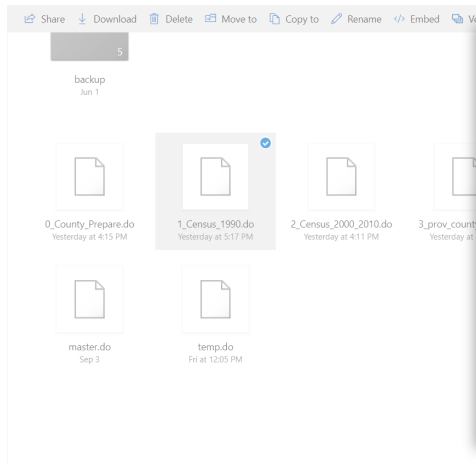
 New shared folder

 New folder

 Show deleted files

Huang Fay Yi CHEN

Dropbox Snapshot



Version History

Modified Date		Modified By	Size
Yesterday at 5:17 PM	...	Yi Chen	12.1 KB
Yesterday at 4:06 PM	...	Yi Chen	12.1 KB
Yesterday at 3:14 PM	...	Yi Chen	11.3 KB
Yesterday at 10:02 AM	...	Yi Chen	14.6 KB
Yesterday at 9:58 AM	...	Yi Chen	14.6 KB
Yesterday at 9:42 AM	...	Yi Chen	14.7 KB
Yesterday at 9:40 AM	...	Yi Chen	16.0 KB
Fri at 12:16 PM	...	Yi Chen	16.4 KB
Fri at 12:06 PM	...	Yi Chen	16.4 KB
Fri at 11:56 AM	...	Yi Chen	16.4 KB



¥16.00

286人付款

dropbox扩容至18GB 慢速扩容dropbox容量到18GB 永久有效 安全

三 wly52017

北京



¥16.00 包邮

287人付款

dropbox升级扩容到18G, 另有全新18G, 19G, 22G现货 稳定永久.

三 zhuexuyan1986

北京



¥15.00 包邮

95人付款

dropbox扩容至18GB 已经扩容好的18GB 22GB账号 永久有效容量

三 eason_will

广东 广州



¥38.00 包邮

20人付款

★1G Dropbox Coupon code促销码 永久升级扩容 安全可靠

三 wanjin823

美国



¥150.00 包邮

8人付款

Dropbox 专业版plus 扩容1T 年费 信用卡通

三 苹果itunes

美国



¥4.99 包邮

26人付款

OneDrive 5T 5120GB 云存储 大容量 dropbox box 扩容 永久

三 crashhades

浙江 杭州



¥63.00

5人付款

【现货全新】Dropbox 22GB 可改登录邮箱 升级扩容 永久促销

三 caobulue

北京



¥15.00

14人付款

官方推荐方式升级扩容指定dropbox到18G 安全稳定可靠永久容量

三 caobulue

北京

Keep Your Working Folder Organized

A complete project involves numerous files of different types.

- First of all, it makes researchers feel better. . . Just like a tidy room.
- More importantly, you don't wish to waste time in searching for a file from nowhere (or find a wrong file. . .)
- Tip: try to avoid using Chinese in the path. Some software may not support Chinese!

Keep Your Working Folder Organized



Outline

- 1 Get Prepared Before You Start
- 2 About Research Ideas**
- 3 Search and Read Literature
- 4 Publicly Available Data Sets

You need an idea first

An idea can either be,

- Question-driven
- Data-driven
- Method-driven

Question driven—"From Life, Beyond Life"

- Literature is a good source of learning. But this is not necessarily the case for great ideas. Your mind might be bounded by previous studies.
- Many, if not most, influential and pioneering works seem to come from nowhere. Some examples,
 - Qian (2008): Tea and sex imbalance
 - Chen et al. (2013): Huai river, air pollution, and life expectancy
 - Bai and Jia (2016): Keju, social mobility, and violence
 - Chen et al. (2021): Notching R&D
- Staying in the ivory tower is not good for economic research. It is still mostly a social science.
 - Read news, hot social issues, blogs, articles on WeChat et al.
 - But keep your mind open and listen to different opinions.

Not only empirical research, the origins of many theoretical or econometric studies are also from life.

- Becker's Economics of Family
- Heckman Two-Step
- Shapley's study on matching stability
 - Boston mechanism versus deferred acceptance
- Akerlof's lemons market and Spence's signaling model

Data Driven Ideas?

- I used to frame data-driven ideas as “thinking about your comparative advantage”
- Then I realize question-driven is **the only way** of generating ideas.
- By why so many studies appear to be data-driven?
 - Research emphasizes “marginal contribution.”
 - If a question is important and can be answered with any data, then it probably has been answered already.
 - Novel data allows you to answer some question that has not been answered before—**but you are still answering a question!**

A Negative Example of Method/Data-Driven Research

- Oster (2005): hepatitis B can account for around 75 percent of the “missing women” in China.
- In China, the sex ratio at birth for the first birth was in the normal. The abnormally high sex ratio observed at birth was basically due to higher birth orders.
- Lin and Luoh (2008): three million births from the Hepatitis B Mass Immunization (HBMI) national databank
 - Marginal increase in the probability of having a male birth for HBV mothers relative to non-HBV mothers is only 0.0025, at most.
 - This estimate does not vary with birth order or sex composition of previous children.
- Oster et al. (2010): “Hepatitis B does not explain male-biased sex ratios in China”

Big Data

There is no doubt that big data is changing our life and research.

- Two important features for researchers: large size, new information.

But don't treat big data as panacea ([Reading Material 2.1](#)). Actually, it is surprisingly difficult to do research with big data.

- Hard to link different pieces of information (too few variables).
- Hard to link across individuals.
- Miss those who need the most help.
- In practice, most research with big data is aggregated at the regional level.

Proprietary Data

Economists are increasingly relying on “unique” data that are hard for others to get, especially for top journals.

- There is nothing wrong in it. “Unique” data often include new information and allow for new perspectives.

Several issues to keep in mind,

- Increasing usage of proprietary data raises greater concerns about replicability in empirical research.
- Proprietary data may make your research “easier,” but it does not (and should not) make your research “easy.”
- Conflict of interest.

The Economic Journal Data Policy

<http://www.res.org.uk/view/datapolicyEconomic.html>

*a request for an exemption based on the grounds that the data are from **a proprietary data source** that is not accessible to other researchers; papers using such data are **discouraged** but will be considered on an individual basis by the Editor; the exemption must be requested at the time of submission; if the paper is accepted a file (README.pdf) that describes how the data was collected and used to obtain the results must be provided.*

Outline

- 1 Get Prepared Before You Start
- 2 About Research Ideas
- 3 Search and Read Literature**
- 4 Publicly Available Data Sets



NEVER Underestimate the Importance of Searching & Reading Literature


- Two purposes: **reading for learning** and for **a specific research idea**
- General reading—our students are reading way too few papers!
- Specific reading—after you come up with an “idea,” I highly recommend you do some brainstorming **before reading the literature**.
- It feels bad if you search the literature and find out someone has been working on the same idea.
 - Don't give up immediately!
 - “Precisely” state your marginal contribution, no smaller, no bigger.
 - Anyway, it is way better than finding out the study after one year's work.
- Again, research emphasizes “marginal contribution,” which cannot be defined if you don't know what has been done already.

Where to Search Literature?

- JSTOR/ScienceDirect/Wiley Online Library
 - Not so recommended nowadays. You can only find publications there.
 - Latest research is working paper, which can take years to turn into publication.
- Google (Scholar)
- IDEAS
- CNKI/国家哲学社会科学文献中心
 - Don't overlook Chinese publications! Some papers are really good.
 - They are also good sources of finding most recent studies.
- Two powerful tools when you cannot directly download the paper:
 - **Sci-hub**
 - **"Versions"** in Google scholar

Google Scholar—All Versions

 Landersø Effects of School Starting Age on the Family 

 Articles

[Any time](#)
[Since 2019](#)
[Since 2018](#)
[Since 2015](#)
[Custom range...](#)


[Sort by relevance](#)
Sort by date

☒ include patents
☒ include citations

Effects of school starting age on the family

[RK Landersø](#), [HS Nielsen](#), ... - Journal of Human ..., 2019 - jhr.uwpress.org

This paper investigates intra-family spillovers from the focal child's timing of school start. We first show how school starting age affects the timing of subsequent educational transitions. Exploiting quasi-random variation in school starting age induced by date of birth, we then document effects on parental outcomes. At child age seven, for example, being one year older at school start increases maternal employment with four percentage points; at child age 15, it increases the likelihood that parents still cohabit with eight percentage points. Our ...

☆  Cited by 1 [Related articles](#) [All 3 versions](#)

[\[PDF\] uwpress.org](#)

Showing the best result for this search. [See all results](#)

Google Scholar—All Versions

Google Scholar



Articles

3 results (0.01 sec)

All versions

Effects of school starting age on the family

[RK Landersø](#), [HS Nielsen](#)... - Journal of Human ..., 2019 - jhr.uwpress.org

This paper investigates intra-family spillovers from the focal child's timing of school start. We first show how school starting age affects the timing of subsequent educational transitions. Exploiting quasi-random variation in school starting age induced by date of birth, we then ...

☆ Cited by 1 [Related articles](#)

[CITATION] Effects of School Starting Age on the Family

[RK Landersø](#), [HS Nielsen](#)... - Journal of Human ..., 2018 - forskningsdatabasen.dk

[PDF] EFFECTS OF SCHOOL STARTING AGE ON THE FAMILY

[RK LANDERSØ](#), [HS NIELSEN](#), [M SIMONSEN](#) - 2019 - rockwoolfonden.dk

This paper investigates intra-family spillovers from the focal child's timing of school start. We first show how school starting age affects the timing of subsequent educational transitions. Exploiting quasi-random variation in school starting age induced by date of birth, we then ...

[\[PDF\] rockwoolfonden.dk](#)


How to Search Literature Efficiently?

- Related literature review is a good start
 - Handbook
 - Journal of Economic Literature
 - Journal of Economic Perspective
- But literature review does not necessarily cover the latest research.
- It is very useful & important to find out the latest literature.
 - Then you can track earlier research from the Reference.



How to Find Out the Latest Research?

- After reading the literature review, you should have an idea about who are the leading experts in this area. Check their personal website to find out their latest related research.
- Google scholar is a very powerful tool. Through “cited by” and “related articles”, one can easily pin down the latest research.

Google Citations



The Career Costs of Children



All News Images Videos More Settings Tools

About 56,800,000 results (0.62 seconds)

Scholarly articles for **The Career Costs of Children**

The **career costs** of **children** - **Adda** - Cited by 114

Diverting **children** from a life of crime: Measuring **costs** ... - **Greenwood** - Cited by 462

... How women's schooling and **career** affect the process ... - **Blossfeld** - Cited by 995

The Career Costs of Children: Journal of Political Economy: Vol 0, No 0

www.journals.uchicago.edu/doi/abs/10.1086/690952

by J Adda - 2017 - Cited by 113 - Related articles

Mar 8, 2017 - We quantify the life cycle **career costs** associated with **children**, how they decompose into loss of skills during interruptions, lost earnings ...

[PDF] The Career Costs of Children - UCL

www.ucl.ac.uk/~uctpb21/Cpapers/fertility_resubmitJPE.pdf ▼

by J Adda - 2015 - Cited by 113 - Related articles

Jan 1, 2015 - Our objective is to develop an estimable life-cycle model to assess the **career costs** of **children**. ... Third, depending on ability and expected fertility, women may sort into occupations that minimize the expected **career costs** of **children**.

The Career Costs of Children - IDEAS/RePEc

<https://ideas.repec.org/p/iza/izadps/dp6201.html> ▼

by J Adda - 2011 - Cited by 113 - Related articles

J1 - Labor and Demographic Economics -- Demographic Economics. J2 - Labor and Demographic Economics -- Demand and Supply of Labor. J31 - Labor and Demographic Economics -- Wages, Compensation, and Labor **Costs** --- Wage Level and Structure; Wage Differentials.

[Abstract](#) · [Bibliographic info](#) · [Related research](#) · [References](#)

Google Citations

Scholar

About 25 results (0.02 sec)

All citations

Articles

Case law

My library

Any time

Since 2017

Since 2016

Since 2013

Custom range...

Sort by relevance

Sort by date

☒ include patents

☒ include citations

☒ Create alert

The career costs of children

☐ Search within citing articles

[PDF] Hours, Occupations, and Gender Differences in Labor Market Outcomes

A Erosa, L Fuster, [G Kambojov](#), R Rogerson - 2017 - [sites.google.com](#)

Abstract We document a robust negative relationship between the log of mean annual hours in an occupation and the standard deviation of log annual hours within that occupation. We develop a unified model of occupational choice and labor supply that features heterogeneity

[All 6 versions](#) [Cite](#) [Save](#) [More](#)

[PDF] [google.com](#)

[PDF] Can Women Have Children and a Career? IV Evidence from IVF Treatments

[P Lundborg](#), [E Plug](#), [AW Rasmussen](#) - American Economic Review, 2017 - [economists.nl](#)

Abstract This paper introduces a new IV strategy based on IVF (in vitro fertilization) induced fertility variation among childless women to estimate the causal effect of having children on their career. For this purpose, we use administrative data on IVF treated women in Denmark.

[Cited by 3](#) [Related articles](#) [All 9 versions](#) [Cite](#) [Save](#) [More](#)

[PDF] [economists.nl](#)

Top Earnings Inequality and the Gender Pay Gap: Canada, Sweden, and the United Kingdom

[NM Fortin](#), [B Bell](#), [M Böhm](#) - Labour Economics, 2017 - Elsevier

Abstract This paper explores the consequences of the under-representation of women in top jobs for the overall gender pay gap. Using administrative annual earnings data from Canada, Sweden, and the United Kingdom, it applies the approach used in the analysis of

[Related articles](#) [All 6 versions](#) [Cite](#) [Save](#)

[PDF] [iza.org](#)

Parenthood and productivity of highly skilled labor: evidence from the groves of academe

[M Krapf](#), [HW Ursprung](#), C Zimmermann - Journal of Economic Behavior & ..., 2017 - Elsevier

Abstract We examine the effect of parenthood on the research productivity of academic economists. Combining the survey responses of nearly 10,000 economists with their publication records as documented in their RePEc accounts, we do not find that motherhood

[Cited by 18](#) [Related articles](#) [All 33 versions](#) [Cite](#) [Save](#)

[PDF] [econstor.eu](#)

General Reading in Addition to Specific Reading

- Bad research claims $x \rightarrow y$ while actually showing $x \leftrightarrow y$.
- Good research correctly identifies $x \rightarrow y$.
- Better research speaks to $x \rightarrow w \rightarrow y$.
- A good understanding of w maybe related to an even broader scope of literature.
 - Sometimes you don't know how to write because you don't have the knowledge on what to write.
 - Though the payoff of general reading is not immediate to the current project, the long-run returns are pretty high.

Example

Chen et al. (2020): Sent-down-youth and Rural Education in China

- Cultural Revolution
- Great Famine
- “Third-line” Construction
- Rural Education expansion

How to Follow the Latest Literature?

You can “subscribe” to the latest issues of

- NBER working paper (highly recommended)
- IZA working paper
- Top-tier journals/Top field journals

NBER Subscription

The National Bureau of Economic Research | Sign up for NBER News | www.nber.org/prefs/notify

NBER Programs

- ☒ AG -- Aging
- ☐ AP -- Asset Pricing
- ☐ CF -- Corporate Finance
- ☒ CH -- Children
- ☐ DAE -- Development of the American Economy
- ☒ DEV -- Development Economics
- ☒ ED -- Economics of Education
- ☐ EEE -- Environment and Energy Economics
- ☐ EFG -- Economic Fluctuations and Growth
- ☒ HC -- Health Care
- ☒ HE -- Health Economics
- ☐ IFM -- International Finance and Macroeconomics
- ☐ IO -- Industrial Organization
- ☐ ITT -- International Trade and Investment
- ☐ LE -- Law and Economics
- ☒ LS -- Labor Studies
- ☐ ME -- Monetary Economics
- ☒ PE -- Public Economics
- ☐ POL -- Political Economy
- ☐ PR -- Productivity, Innovation, and Entrepreneurship
- ☐ TWP -- Technical Working Papers

JEL Classes

- ☐ A -- General Economics and Teaching
- ☐ B -- History of Economic Thought, Methodology, and Heterodox Approaches
- ☐ C -- Mathematical and Quantitative Methods
- ☐ D -- Microeconomics
- ☐ E -- Macroeconomics and Monetary Economics
- ☐ F -- International Economics
- ☐ G -- Financial Economics
- ☒ H -- Public Economics
- ☒ I -- Health, Education, and Welfare
- ☒ J -- Labor and Demographic Economics
- ☐ K -- Law and Economics
- ☐ L -- Industrial Organization
- ☐ M -- Business Administration and Business Economics • Marketing • Accounting • Personnel Economics
- ☐ N -- Economic History
- ☐ O -- Economic Development, Innovation, Technological Change, and Growth
- ☐ P -- Economic Systems
- ☐ Q -- Agricultural and Natural Resource Economics • Environmental and Ecological Economics
- ☐ R -- Urban, Rural, Regional, Real Estate, and Transportation Economics
- ☐ Y -- Miscellaneous Categories
- ☐ Z -- Other Special Topics

[Subscribe](#) [Unsubscribe](#)

Comments, questions, or suggestions should be sent to Jean Roth at jroth@nber.org

NBER Subscription

The Latest NBER Research (2021-08-16) ★

ntw+nwph2021-08-1630 代表 NBER

发给 chenyeicon

2021-08-16 12:31 隐藏信息

发件人: ntw+nwph2021-08-1630<ntw+nwph2021-08-1630@nber.org> 代表 NBER<bulletin@nber.org>

收件人: chenyeicon<chenyeicon@163.com>

时间: 2021年8月16日 (周一) 12:31

大小: 44 KB

NBER

August 16, 2021



New This Week: The Latest NBER Working Papers

The National Bureau of Economic Research circulates research by its affiliated economists as working papers intended for professional and public discussion and comment. The papers have not been peer reviewed.

1. [Corrective Regulation with Imperfect Instruments](#)
Eduardo Dávila and Ansgar Walther #29160
2. [Cigarette Taxes, Smoking, and Health in the Long-Run](#)
Andrew I. Friedson, Moyan Li, Katherine Meckel, Daniel I. Rees, and Daniel W. Sacks #29145
3. [Contract Labor and Firm Growth in India](#)
Marianne Bertrand, Chang-Tai Hsieh, and Nick Tsivanidis #29151

How to Read Literature Efficiently?

Economic papers are often... very long... sometimes stupidly long... It would be too time-consuming to read every paper carefully word-by-word.

My suggestion is: understand your purpose of reading each paper first.

- Source of learning
 - institutional details; which data to use; empirical strategy; conclusion
- Source of reference
 - overview of other related studies
- Source of supporting evidence
- Source of idea
 - Can you improve this study?
 - Does the author mention some unsolved problems or direction of future research?
 - Can the idea/empirical strategy be linked to different research?

Economic Papers are Getting Ridiculously Long. . .

The Econometric Society

An International Society for the Advancement of Economic Theory
in its Relation to Statistics and Mathematics

September 5, 2019

www.econometricsociety.org

Statement from the editors of *Econometrica*, *Quantitative Economics* and *Theoretical Economics*














Published papers in many of the economics journals have gotten much longer over the years. The Econometric Society journals *Econometrica*, *Theoretical Economics*, and *Quantitative Economics*, are no exception. The average length of a paper in *Econometrica* was about 12 pages in the 1970s, and is now about 36 pages, not including online appendices, with many papers over 50 pages. Some of this may be due to changes in technology, such as the increased ability to estimate complex models using large data sets, or the greater technical difficulty in some areas in theory or econometrics, but probably not all of it.

- Literature that is directly related to your research, or the base of your “marginal contribution”
 - You definitely need to read them very carefully. . . for multiple times.
 - Motivation/related literature/data/empirical setting/discussion.
 - You can also learn the way how to organize and write the paper.
 - These authors have high chances of reviewing your article. So you don't wish to offend them by misinterpreting their study and exaggerating your own marginal contribution.
- For other literature, in most cases, it is enough to know their “existence.” So you can come back to read them more carefully if necessary.
 - Instead of the details, knowing what the paper does is more important.
 - Read the abstract & introduction → glimpse the data/empirical setting/figures & tables → and jump to the conclusion. This would be sufficient for 80% of such kind of reading.

Some Tips

- Keep your literature folder organized by topics
- Mark most important literature
- Respect the authors. . . by citing the latest version
- Keep notes, don't trust your memory too much (you may return to a specific paper several times during your research)

Keep Your Literature Folder Organized

名称	修改日期	类型	大小
 China	2017/6/21 10:56	文件夹	
 Discrimination	2017/5/20 6:50	文件夹	
 Family	2017/7/15 9:32	文件夹	
 Firm	2016/12/10 17:34	文件夹	
 Other Countries	2017/9/11 9:44	文件夹	
 Review	2017/9/2 15:21	文件夹	
 中文期刊	2017/1/10 10:02	文件夹	
 Alesina (2013) Women and Plough.pdf	2017/5/7 19:09	Adobe Acrobat ...	2,053 KB
 Barth (2017) Dynamics of Gender Earning Differentials.pdf	2017/5/8 19:33	Adobe Acrobat ...	831 KB
 Bertrand (2011) Handbook New Perspectives on Gender.pdf	2016/8/29 16:48	Adobe Acrobat ...	1,273 KB
 Chapter 48 Race and gender in the labor market.pdf	2015/3/7 9:25	Adobe Acrobat ...	7,372 KB
 Chen (2010) CDC.pdf	2017/5/21 17:08	Adobe Acrobat ...	534 KB
 Tyrowicz (2015) Gender Wage Gap Over The Life Cycle.pdf	2017/5/7 19:17	Adobe Acrobat ...	840 KB

Outline

- 1 Get Prepared Before You Start
- 2 About Research Ideas
- 3 Search and Read Literature
- 4 Publicly Available Data Sets

Why “Publicly Available”

- Articles in top journals are using increasingly crazy data. For example
 - Lundborg et al. (2017): Can Women Have Children and a Career?
 - Denmark administrative registration data + IVF (in vitro fertilization) register/treatment/outcome
- So why we still learn the data that everyone has access to?
 - Most practically, not everyone has crazy data sets. . . and not everyone needs to publish in top journals. . .
 - Even if yes, publicly available data sets can serve as nice complements.
 - Replicability is important for credible research.
 - Practice! In case you have access to crazy data sets in the future!
 - My personal experience is—“unique” data sets are a lot harder to use.
 - They are not ready for public use. Data can be messy and not well documented.

A Couple of Notes about Survey Data

I will skip the detailed introduction of each data set (**Reading Material 2.2** for details). Instead, I will discuss several common issues when using survey data.

- ① Coresidence pitfall
- ② Sample attrition
- ③ National representativeness
- ④ Misreport
- ⑤ Data imputation
- ⑥ Variable comparability

1. Coresidency Pitfall

- Prior to more modernized household surveys (such as CHALRS and CFPS), information is mostly conditional on within-household.
 - e.g., if you need to construct a variable named “mother’s age,” you can only define such variable for those who coreside with their mothers.
 - Therefore, whenever you use this piece of information, you are imposing a coresidency restriction implicitly.
- This creates problems for many studies, for example
 - Use twin strategy to estimate returns to education.
 - Burden of elderly care on women’s labor supply.
- Later surveys contain “basic” information about close relatives regardless of coresidency status.
 - But you should still be cautious whether using a certain variable brings unintended restrictions.

2. Sample Attrition

- Sample attrition is a common threat to all panel surveys.
- It is tremendously difficult to trace the same person several years later, especially for migrants.
 - RuMic is only able to track about 60% of the migrants one year later.
- The observations in each wave looks about the same because new sample will be added.
 - But when your analysis introduces panel information, sample attrition issue arises.

3. National Representativeness

- It is surprisingly difficult to represent a general population with a random sample. Details about sample weight will be discussed in later lectures.
- What you need to keep in mind at this moment is:
 - Not all surveys can represent a nation (e.g., CHNS)
 - Be very careful when you want to represent a subpopulation (e.g., big & small provinces in CFPS)

4. Misreport

- Surveys are report by persons, therefore, misreport is inevitable—either intentionally or unintentionally.
- One classic example is the child-under-report in population censuses.
 - Related to China's biased sex ratio
 - Not so important in 1982, but becomes especially prevalent in 1990 and 2000.
- Misreport is essentially measurement error. Recall the econometric class, when does it matter?

5. Data Imputation

- For ease of usage, the survey often provides some imputed value (e.g., disposable income).
- Keep in mind the implicit assumptions required by the imputation.
- However, in some data, you even don't know the imputation process (e.g., many data from NBS).

6. Question Comparability

Question on self-rated health in CFPS 2010:

P3 您认为自己的健康状况如何？

1. 健康 2. 一般 3. 比较不健康 4. 不健康 5. 非常不健康

Self-rated health in CFPS 2012:

P201 您认为自己的健康状况如何？

- 1.非常健康 2.很健康 3.比较健康
4.一般【不读出】 5.不健康

Data Sets Other than Household Survey

Household survey is only one among many types of data

- Firm-level data (e.g., Chinese Industrial Enterprises Database)
- Customs data
- Patent data

Principles of Choosing Data Sets

How to choose an appropriate data set?

- Well... Sometimes you just don't have a choice. The data has to contain all the information you need for the analysis.
 - CHNS cannot be used to study consumption.
 - Some studies require panel information.
 - You need information in certain years to exploit a policy shock.
- In some empirical design, you can extract difference pieces of information from different data sets
 - Synthetic cohorts analysis
 - Two-sample instrumental variables
- Some data sets natural fit some topics
 - UHS fits for the study of consumption and saving
 - CHARLS is suitable for studies related to the elderly

If there are multiple data sets fit the criteria

- National representativeness
 - Census > nationally representative surveys > other
- Sample size
- Quality control
 - More recent surveys usually do better
- Reputation
 - Avoid using a “unique” data set to do the research which can be done with a publicly available data set.
- Use multiple data for robustness purpose is not a bad idea.
- Alternatively, you can combine multiple data sets to increase the sample size.
 - If doing so, be very careful about the comparability across different surveys!

Reference

- Bai, Y. and R. Jia (2016). Elite Recruitment and Political Stability: The Impact of the Abolition of China's Civil Service Exam. *Econometrica* 84(2), 677–733.
- Chen, Y., A. Ebenstein, M. Greenstone, and H. Li (2013). Evidence on the Impact of Sustained Exposure to Air Pollution on Life Expectancy from China's Huai River policy. *Proceedings of the National Academy of Sciences* 110(32), 12936–12941.
- Chen, Y., Z. Fan, X. Gu, and L.-A. Zhou (2020). Arrival of Young Talent: The Send-down Movement and Rural Education in China. *American Economic Review* 110(11), 3393–3430.
- Chen, Z., Z. Liu, J. C. Suárez Serrato, and D. Y. Xu (2021). Notching R&D Investment with Corporate Income Tax Cuts in China. *American Economic Review* 111(7), 2065–2100.
- Lin, M.-J. and M.-C. Luoh (2008). Can Hepatitis B Mothers Account for the Number of Missing Women? Evidence from Three Million Newborns in Taiwan. *American Economic Review* 98(5), 2259–73.
- Lundborg, P., E. Plug, and A. W. Rasmussen (2017, June). Can Women Have Children and a Career? IV Evidence from IVF Treatments. *American Economic Review* 107(6), 1611–37.
- Oster, E. (2005). Hepatitis B and the Case of the Missing Women. *Journal of Political Economy* 113(6), 1163–1216.
- Oster, E., G. Chen, X. Yu, and W. Lin (2010). Hepatitis B Does Not Explain Male-biased Sex Ratios in China. *Economics Letters* 107(2), 142–144.
- Qian, N. (2008). Missing Women and the Price of Tea in China: The Effect of Sex-Specific Earnings on Sex Imbalance. *The Quarterly Journal of Economics* 123(3), 1251–1285.