The Design of Empirical Research

Lachlan Deer

Social Media and Web Analytics, Spring 2025

Learning Goals

1. Define the term Quantitative Empirical Research
2. Explain the criteria that make a good research question
3. Identify why a given research question is “good”
4. Define Identification in the context of empirical research
5. Explain how using known facts and/or assumptions aids identification in practice

1/ Designing Research

I have a Question

How does the world work?

We'll *never* know everything perfectly

$\Longrightarrow$ There's always scope for new research

$\Longrightarrow$ We'll need to be comfortable with simplifications

A **good** research question is:

1. Well-defined
2. Answerable
3. Understandable

Our goal: Conduct research in a way that's capable of answering the question we asked

Empirical Research

The focus of this class: **quantitative empirical research**

Empirical Research:

- **uses** (structured) **observations from the real world** to attempt to answer questions.

Quantitative:

- **Uses quantitative measurements**
- Usually numbers...
  - Can be hard to measure precisely

The Difficulty of Empirical Research

Potential Problem: Numbers we observe might not tell us what we what to know

- In which case we **cannot** answer the research question
- Why?
  - Observational data often lacks a "what if" or a "what would have been"

Potential Opportunity: Can we figure out how to:

1. Collect the right numbers, or
2. Do the right things to those numbers,

to get an actual answer to our question

$\Longrightarrow$ How can we **design** the right kind of analysis to answer our question

2/ Research Questions

What is a Research Question?

A **good research question** is:

1. **Well-defined**: Clearly identified subject(s), outcomes and an intervention
2. **Answerable**: There exists some evidence that if found, would answer the question
3. **Understandable**: clear details and context that the target audience can understand

What is a Research Question?

Good research questions improve our understanding of how the world works

- Helps improve your "*why*" explanation
- And we are interested in the whys!

Why Start with a Question?

Alternative: Start with some **patterns in the data**

- This is essentially *data mining*

Data mining is **good** at:

- Finding patterns
- Making predictions in stable environments

Data mining is **bad** at:

- Answering causal questions
- Improving our understanding, i.e. the ”why”
- Counterfactual analysis in unseen environments
- Informing theory

Where do Research Questions Come From?

Sources for research questions:

- **Curiosity**: wanting to know how the world works
- **Opportunity**: having access to a dataset, does a research question come to mind

Is the need to make a business decision a form of curiosity?

Have I Got a Good Question?

Criteria for evaluation:

- **Potential Results**: What conclusions can be drawn from your findings?
- **Feasibility**: Is the right data available? (or can it be made available)
- **Scale**: Is there enough resources and time to answer it?
- **Research Design**: can a reasonable research design be found to answer it?
- **Simple, but useful**

Example: Why Do People Contribute Content to Twitter?

Intrinsic vs. Image-Related Utility in Social Media: Why Do People Contribute Content to Twitter?

Olivier Toubia
Columbia Business School, New York, New York 10027,
ot2107@columbia.edu
Andrew T. Stephen
Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, Pennsylvania 15260,
astephen@katz.pitt.edu

We empirically study the motivations of users to contribute content to social media in the context of the popular microblogging site Twitter. We focus on noncommercial users who do not benefit financially from their contributions. Previous literature suggests that there are two main types of utility that motivate these users to post content: intrinsic utility and image-related utility. We leverage the fact that these two types of utility give rise to different predictions as to whether users should increase their contributions when their number of followers increases. To address the issue that the number of followers is endogenous, we conducted a field experiment in which we exogenously added followers (or follow requests, in the case of protected accounts) to a set of users over a period of time and compared their posting activities to those of a control group. We estimated each treated user's utility function using a dynamic discrete choice model. Although our results are consistent with both types of utility being at play, our model suggests that image-related utility is larger for most users. We discuss the implications of our findings for the evolution of Twitter and the type of value firms may derive from such platforms in the future.

Key words: social media; field experiments; dynamic discrete choice models
History: Received: November 17, 2011; accepted: December 29, 2012; Preyas Desai served as the editor-in-chief and Teck Ho served as associate editor for this article. Published online in *Articles in Advance* April 8, 2013.

Example: Why Do People Contribute Content to Twitter?

Read the paper “Intrinsic vs. Image-Related Utility in Social Media: Why Do People Contribute Content to Twitter?” and answer the following questions:

- What is the research question?
- What is the research design?
- Why is this a “Good” question?

(Ignore Section 5 of the paper, it gets quite mathsy)

What is the Research Question?

What is the Empirical Design?

Why is this a “Good” Question?

Summary

- A good research question is well-defined, answerable and understandable
- Research questions originate from curiosity
- Five additional criteria help the researcher decide if their question is good

3/ Identification

Identification in One Sentence

Suppose that you have:

1. A good research question
2. Some data that may help you answer the research question

Identification is the process of figuring out what part of the variation in your data answers your research question

The Data Generating Process

- One way to think about science generally is that there are regular laws that govern the way the universe works
- These laws are an example of a **data generating process**
  - They work “behind the scenes”
  - … i.e. we **do not observe them** directly
- We **do observe the data resulting from the laws**
  - From which **we try describe or test the which laws are actually at play**
  - … based on whether the data support their predictions
- In social science and business, these laws are less well behaved and more imprecise than the hard sciences
  - But we do believe data comes from somewhat regular laws

Data Generating Processes

- **Two parts** to a data generating process (DGP)
  1. Parts we know
  2. Parts we do not know
     - What we want to learn about
- The parts we know are still important
  - **We don't start from "nothing"** each time we embark on something new
  - It helps us **refine how we think** about what we don't know

A Simple DGP

1. Income is log-normally distributed
2. Being brown-haired gives you a 10% income boost
3. 20% of people are naturally brown-haired
4. Having a college degree gives you a 20% income boost
5. 30% of people have college degrees
6. 40% of people who don't have brown hair or a college degree will choose to dye their hair brown

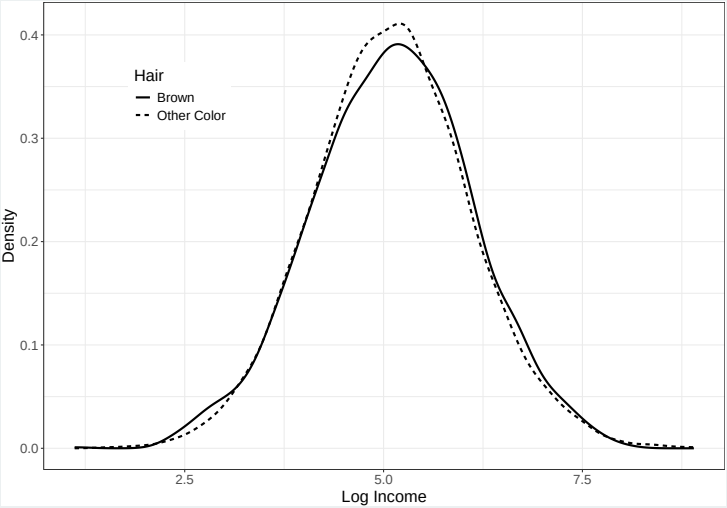
Let's generate data from these laws and view the results!

Simulating Data from the DGP

```
set.seed(987987)

df <-
    tibble(College = runif(5000) < .3) %>%
    mutate(Hair = case_when(
                runif(5000) < .2+.8*.4*(!College) ~ "Brown",
                TRUE ~ "Other Color"
                ),
    logIncome = .1*(Hair == "Brown") +
                .2*College + rnorm(5000) + 5
            )
```

Visualizing Data from the DGP



Can you eye-conometrically see any differences?

Learning About the DGP

Research Question: **What is the effect of being brown-haired on income?**

```
# A tibble: 2 x 2
  Hair        log_income
  <chr>            <dbl>
1 Brown             5.11
2 Other Color       5.10
```

Suggests that brown haired people earn approx. 1% more than people with other colors

- But **our laws say this effect is 10%!**
- Differences in means are not enough in this scenario

Learning About the DGP

Imagine we know everything about the data generating process *except* the effect of brown hair on income

Does this help us get the right answer?

Learning About the DGP

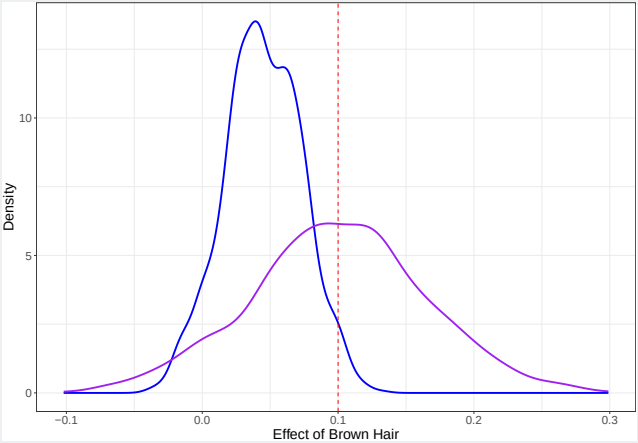
```
# A tibble: 2 x 2
  Hair        `Log Income`
  <chr>              <dbl>
1 Brown               5.34
2 Other Color         5.21
```

Now we see **the effect is 13%** ...

- **Closer to 10%!**
- Difference is due to **randomness**

Re-Running Our Experiment Many Times

What if we re-run the experiment 1000 times?



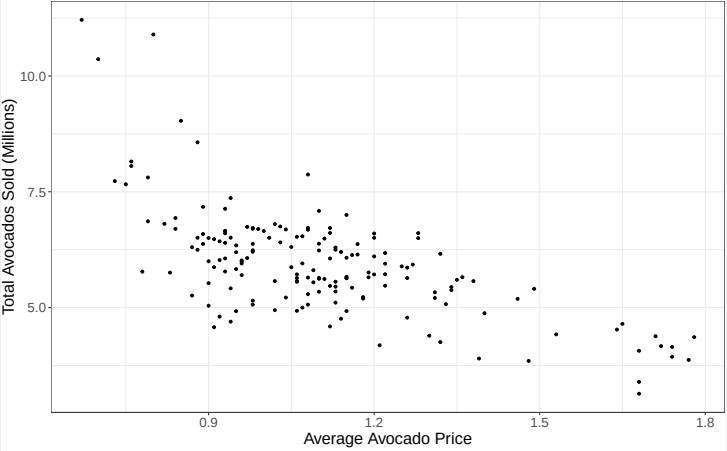
Using what we know can help us get the right answer (on average)!

How the Heck did that Work?

We get the right answer ...

- Or close enough when we had one sample
- On average when we had access to many samples
- **The right answer involved using our knowledge of the DGP**
    - But how exactly?
- **Two ideas** where at play
    - **Looking for variation**
    - **Identification**
- Let's consider these in turn ...

Example: Price & Volume of Avocados



Example: Price & Volume of Avocados

Answer the following three questions:

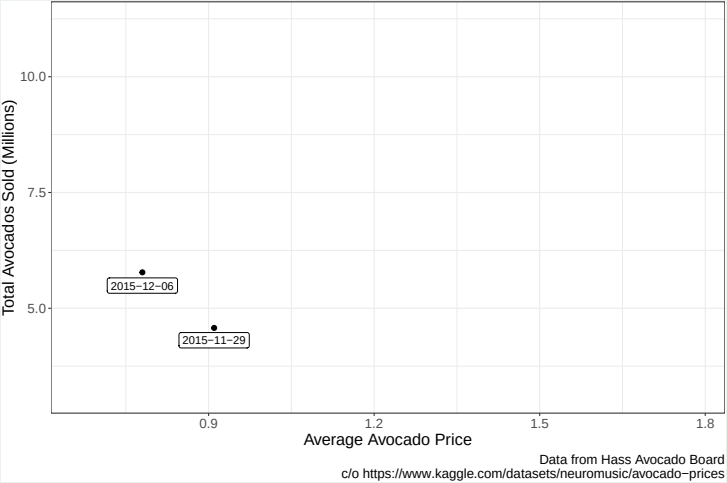
1. What conclusions can you draw from the previous figure?
2. What research question could you answer with this data?
3. Can you answer your question in (2) using the figure?

Answers to the questions

1. Avocado sales tend to be lower in weeks where the price of avocados is high
2. What is the effect of a price increase on the number of avocados people buy
3. No, covariation is not enough!

Covariation is Not Enough

Consider these datapoints from two consecutive weeks:



Covariation is Not Enough

Why did price drop and quantity rise from January to February that year?

- Is it because a drop in price made people buy more?
- Is it because the market was flooded with avocados so people wouldn't pay as much for them?
- Is it because the high price in January made suppliers bring way more avocados to market in February?

It's **probably a little bit of all of these reasons**

- Which means its going to be **tough** to answer our RQ

Where's Your Variation?

How can we find the variation in the data that answers our question?

- We have to ask **"What is the variation that we want to find?"**
- We want **variation in people buying avocados** (rather than people selling them) that is **driven by changes in the price**

We need to use what we know about the data generating process to learn a little more

- Or, what we are comfortable **assuming**

“Useful” Variation & Assumptions

Assumption: At the beginning of each month, avocado **suppliers make a plan** for what **avocado prices** will be **each week in that month**, and **never change their plans** until the next month

- Avocado **sellers cannot respond to unexpected jumps in demand** week-to-week within a month

$\Longrightarrow$ **Variation in price and quantity from week to week in the same month** will **isolate variation in** people **buying** avocados that can only be **driven by changes in the price**

- We would only want to **use week-to-week variation with months** to answer our research question
- Lurking question: How do we do it?!

Identification

Identification is the process of **figuring out what part of the variation in your data answers your research question** and **isolating it**

- Ensuring that our **calculation** identifies a **single** theoretical **mechanism of interest**
- A research question takes us from theory to hypothesis
- Identification takes us from hypothesis to the data
    - **Making sure** that we have **a way of testing that hypothesis in the data**
    - And not accidentally testing some other hypothesis instead.

Identification and *This* Course

This course: Variation based on (quasi-) experiments in the field

- Why? Less assumptions needed on how consumers and firms behave
  - (That's the claim ...)
  - (It's more subtle than that if one wants to think deeper)
  - This is not cost-free ... can limit scope of what we can study

Example: Why Do People Contribute Content to Twitter?

Read the paper “Intrinsic vs. Image-Related Utility in Social Media: Why Do People Contribute Content to Twitter?” and answer the following questions:

- Explain the two reasons a Twitter user wants to post according to the authors.
- What variation do they use to answer their question?
- Why couldn’t they have used an existing dataset that contains usernames, follower counts and post behaviour to answer their research question

(Ignore Section 5 of the paper)

Summary

- Identification is the process of figuring out what part of the variation in your data answers your research question and isolating it
- Researchers use existing knowledge and assumptions to help overcome the identification challenge when tackling new questions

4/ Wrap Up

Context and Omniscience

Understanding context is incredibly important

- Enables you to block alternative explanations and identify the answer to your question
- Thoughts of Two Nobel Laureates Joshua Angrist and Alan Krueger (2001):

"Here the challenges are not primarily technical in the sense of requiring new theorems or estimators. Rather, progress comes from detailed institutional knowledge and the careful investigation and quantification of the forces at work in a particular setting. Of course, such endeavors are not really new. They have always been at the heart of good empirical research."

What should you do this week?

- Enroll in a Lab Section
- Install R if you haven't done so already
- Prepare your answers to Lab Assignment 1, they'll be discussed in Week 2's Lab Section.
- Read assigned readings for the next lecture.

Acknowledgements

This lecture borrows heavily from the book “The Effect” by Nicholas Huntington-Klein

- In particular, I borrow from Chapters 1, 2, and 5.

License & Citation

Suggested Citation:

```
@misc{smwa2025_design,
      title={"Social Media and Web Analytics: The Design of
             Empirical Research"},
      author={Lachlan Deer},
      year={2025},
      url = "https://tisem-digital-marketing.github.io/2024-smwa"
}
```

This course adheres to the principles of the Open Science Community of Tilburg University. This initiative advocates for transparency and accessibility in research and teaching to all levels of society and thus creating more accountability and impact.