Computational Social Science

MEASUREMENTS & EXPLANATIONS
Measurement

- **Measurement** is the assignment of a number to a characteristic of an object.

- Measuring invisible and complex objects/events (e.g. culture, political leaning, happiness) is tricky
  - People use every day language which is vague
  - We can only measure observable outcomes
  - Many factors may impact the observable outcome and add noise

- How to ensure the quality of measurements?
Quality of Measurements

Reliability: how consistent and stable is our measurement?
• Intra-rater reliability
• Inter-rater reliability
• Test-retest reliability

Validity: do we measure what we want to measure?
• Face validity
• Construct validity
• Criterion-based validity
Cultural Relations

Understanding Similarity

Affinity
Wikipedia

Articles about cuisine/music/literature/... of a country in different language editions

Finnish cuisine

From Wikipedia, the free encyclopedia

Finnish cuisine is notable for generally combining traditional country fare and haute cuisine with contemporary continental style cooking. Fish and meat play a prominent role in traditional Finnish dishes in some parts of the country, while the dishes in others have traditionally included various vegetables and mushrooms. Refugees from Karelia contributed to foods in other parts of Finland.[1] Finnish foods often use wholemeal products (rye, barley, oats) and...
Example: Measurements

Measure cultural similarity between countries/languages via Wikipedia

- Mechanical turk study: which country pair has a more similar cultural? (face validity)

<table>
<thead>
<tr>
<th>Country Pair</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bosnia / Serbia</td>
<td>Which countries are culturally more similar?</td>
</tr>
<tr>
<td>Estonia / Lithuania</td>
<td></td>
</tr>
<tr>
<td>Russia / Ukraine</td>
<td></td>
</tr>
<tr>
<td></td>
<td>England/Hungary or Bosnia/Serbia</td>
</tr>
<tr>
<td>England / Hungary</td>
<td></td>
</tr>
<tr>
<td>Bulgaria / England</td>
<td></td>
</tr>
<tr>
<td>Belgium / Bulgaria</td>
<td></td>
</tr>
</tbody>
</table>
Example: Measurements

– Check if inferred cultural similarity agrees with ESS (criterion-based validity)

– Use different sub-samples (all pages about food, music, literature; use a baselines!) (test-retest reliability)

– Repeat measurements (Wikipedia changes; do our results hold?) (test-retest reliability)

– Theories about relation between similarity and understanding and affinity (construct-based validity)

Paul Laufer, Claudia Wagner, Fabian Flöck and Markus Strohmaier, Mining cross-cultural relations from Wikipedia - A study of 31 European food cultures, ACM Web Science, Oxford, 2015
Problem with Correlations

US spending on science, space, and technology correlates with Suicides by hanging, strangulation and suffocation

Correlation: 99.79% ($r = 0.99789126$)

\[ \rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \]

Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

From [http://tylervigen.com/spurious-correlations](http://tylervigen.com/spurious-correlations)
Problem with Regressions

Do students from elite colleges earn more later in life?

earn ~ b0 + b1 * college + error

– Do people who went to an elite college earn on average more later in life?

Problem “Omitted Variable”

– Being accepted in an elite college correlates with many factors that also impact outcome, e.g. motivation

Idea:

– Apply to elite college correlates with motivation
– earn ~ b0 + b1*college + b2*motivation + error
Statistical Control

Keep all covariates constant except one

What is the average effect of one unit increase in motivation on the earnings of students who went to an elite college?
earn ~ (b0 + b1) + b2*motivation + error

Same for students who did not go to an elite college
earn ~ b0 + b2*motivation + error
\[ Y_i = b_0 + b_1 X_i + \epsilon_i \]

**Bivariate Regression**

Slope estimator \( \hat{b}_1 = \frac{\text{cov}(Y, X)}{\text{var}(X)} = \frac{\sum_i (x_i - \bar{X})(y_i - \bar{Y})}{\sum_i (x_i - \bar{X})} \)

Intercept estimator \( \hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X} \)

\[
SSE = \sum_{i=1}^{N} \left( Y_i - \hat{Y}_i \right)^2
\]

\[
SSE = \sum_{i=1}^{N} \left( Y_i - \left( \hat{b}_0 + \hat{b}_1 X_i \right) \right)^2
\]
Multiple Regression

- \( Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} \)

- Slope of the kth regressor:
  \[ \hat{b}_k = \frac{Cov(Y_i, \bar{x}_{ki})}{Var(\bar{x}_{ki})} \]

- \( \bar{x}_{ki} \) is the residual from the regression of \( x_{ki} \) on all other covariates.

- For example for estimating \( b_2 \)
  \( X_{2i} = b_0 + b_1 X_{1i} + \bar{x}_{2i} \)

The part of X2 that is independent of X1
Goal: Causality

Basic Idea goes back to John Stuart Mill (1806-1873)

“Thus, if a person eats of a particular dish, and dies in consequence, that is, would not have died if he had not eaten of it, people would be apt to say that eating of that dish was the cause of his death.”

Counterfactual model of Causal Inference (also known as Rubin Causal Model or Potential Outcomes Model of Causal Inference)

Does going to an elite college make Yi earn more later? If Yi would have chosen a normal college, would Yi have earned less?
Did going to an elite college impact future earnings of Yi?
Causal Effect

We cannot treat and not treat the same person at the same time. So we compute:

\[ Y_i(T) - Y_j(C) \]

Average Treatment Effect:

\[ ATE = E[Y(T)] - E[Y(C)] \]

How can we make sure that the treatment effect is not caused by other differences between i and j?
Experiment

Gold Standard Method for causal inference!

Assignment of subjects to treatment or control group is random and groups are large enough to wash out differences in other covariates.

Manipulation of treatment is under control of researcher. Different levels of treatment should lead to different levels of effect.
Limitations of Experiments

- Expensive
- Not all treatments are possible and ethical
- Internal validity is high, but external validity is often limited.
- Non-interference assumption is often violated in social science field experiments
Among given “organic” data (e.g. human trace data), can we find a subset that looks like generated by an experiment?

Balance pre-treatment covariates $X$

matching == pruning
Does Special Training Help Job Promotion?

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. “
Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal
Correcting for education, the treated group has higher positions.

\[ pos = b0 + b1 \times edu + b2 \times is\_treated \]

\[ \gamma = \text{estimated treatment effect} \]

(binary variable)
Quadratic Regression

\[ \text{pos} = b_0 + b_1^*e + b_2^*edu^2 + \gamma^*\text{is\_treated} \]

Model Dependence

Reason: Imbalance of covariates

Correcting for education, the treated group has lower positions.
Regression After Pruning

1) Preprocessing (matching)

2) Estimation of effects (regression models)

Matching has reduced model dependence!
Matching Approximates Randomized Experiment

Completely randomized:
- Flip a coin for each patient. Heads -> “T”, tails -> “C”.
- Could get unlucky: all men assigned “T”

Fully blocked experiment:
- First pair up similar patients, same gender, age, …
- Then flip a coin for each pair. One gets “T”, one “C”.
- Balances the known covariates.

Both balance unknown covariates.

**Fully blocked experiment dominates complete randomization!**
Mahalanobis Distance Matching

Approximates **fully blocked experiment**

Many Variations:
- Optimal match, greedy match,
- match 1:1 or 1:many, and so on
- Prune bad matched with distance > threshold (“caliper”)

Gary King, "Why Propensity Scores Should Not Be Used for Matching“, Methods Colloquium, 2015, https://www.youtube.com/watch?v=rBv39pK1iEs
Euclidean distance

Euclidean distance doesn’t make sense when different dimensions are on different scales.

– For example: yearly income, age, gender, body weight

Problem: distance is dominated by largest values

**Def.** The euclidian distance between two points $x = (x_1, \ldots, x_p)^t$ and $y = (y_1, \ldots, y_p)^t$ in the p-dimensional space $\mathbb{R}^p$ is defined as

$$d_E(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_p - y_p)^2} = \sqrt{(x - y)^t(x - y)}$$
Mahalanobis Distance

Conceptual fix: first rescale each dimension to N(0,1)

\[ d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})}. \]

Inverse of variance-covariance matrix

In practice: could use “expert scaling” and Euclidean distance.
Propensity Score Matching

Approximates **complete randomization**

\[ \pi_i = \Pr(T_i = 1|X) = \frac{1}{1 + e^{-x_i\beta}} \]
Propensity Score Matching

Approximates **complete randomization**

Education level not balanced
⇒ High education, low propensity score

Worst case: \( p(T=1) = 0.5 \)  ⇒ random pruning
Summary Matching Methods

• Matching methods are powerful and help to approximate causality

• Problems
  – Researchers have lots of freedom when decide how to match
  – We remove data → we need to specify for which group the causal effect holds
  – Most matching methods have been developed for low number of covariates
  – Worst case: random pruning → increases imbalance → increases bias and model dependence

• Compare results from different matching methods, different dimensionality reduction methods, different models
  – Avoid model dependence and method dependence!
Natural Experiment

Like in experiments but...

- Assignment is only “as-if-random”

- Researcher does not control the intervention or treatment
What is the impact of receiving a scholarship on future performance?

Min score  | Test score threshold | Max score

No scholarship | scholarship

Min score  | Max score
Regression Discontinuity (RD)

Coarsen variable S. Try different bin width!

\[ O = \beta_0 + \beta_1 S + \beta_2 I(x > 11) + \varepsilon \]

Future performance  Coarsend test score  Dummy Var: received scholarship or not

\( \beta_2 \) estimates the effect of receiving scholarship for people with similar test-scores.
RD Robustness Checks

– Can individuals control if they are above or below the threshold?
– Jumps at placebo points versus at threshold?
– Do other variables jump at the threshold?
– Are the results robust to the usage of different bin widths?
– Distinguish between discontinuity and non-linearity
  • Are the results robust against including higher order polynomials?
A. Linear $E[Y_0 | X]$ 

B. Nonlinear $E[Y_0 | X]$ 

C. Nonlinearity mistaken for discontinuity

From *Mostly Harmless Econometrics*
Fuzzy Discontinuity

Threshold can be fuzzy

– e.g. if individuals above the threshold are much more likely to receive a scholarship

– In that case the test score is an instrument for the causal relationship between scholarship and performance.
Another Type of Natural Experiments: Instrumental Variables

Study by Angrist 1990:
- What is the effect of military service (M) on lifetime earnings (E)?

\[ E = \alpha + \beta M + \epsilon \]

Correlated, because “earning potential” is omitted

Beta is biased and inconsistent!

Diagram:
- M (military service)
- E (lifetime earnings)
- P (earning potential, unobserved)
Instrumental variable should be strongly correlated with the included endogenous regressors (M), but not affect outcome variable directly (E).

L is the draft-lottery (can be 0 or 1).

L is an instrument for the causal effect of M on E.

How to estimate the causal effect via instrumental variable $L$?

$$\hat{\beta}_{Wald} = \frac{E[E|L=1] - E[E|L=0]}{E[M|L=1] - E[M|L=0]} \quad \text{“Wald Estimator” if } L \text{ is binary}$$

Plausibility check: what if $L$ is perfect instrument?

Example: winning the lottery $\Rightarrow$ 90% chance of joining military & 10% joining without invitation & $2,200$ per year difference

$$\frac{2,200}{0.8} = 2,750 \quad \text{(what if 0.00001?)}$$

Angrist found that military service decreases earnings about $2,741$ dollar per year
Instrumental Variable

2-Stage Least Square Estimation
We can have multiple instrumental variables

1st stage:

\[ M = \alpha + \beta L + \varepsilon \]

If \( \hat{\beta} \) is significant (t-test), then \( L \) is a good instrument!

\[ \hat{M} = \hat{\alpha} + \hat{\beta} L \]

2st stage:

\[ E = \gamma + \delta \hat{M} + \varepsilon \]

\( \delta \) is a consistent (= asymptotically unbiased) estimator and estimates the causal effect of \( M \) on \( E \)
Example: Emotional Contagion

User’s expression \rightarrow \text{Friends’ expression}

- Social influence? Homophily? Common exposure?

- Use meteorological data as an instrument

No manipulation of user experience!
Emotion on Facebook


Slides provided by Lorenzo Coviello. Thanks! Later partially modified.

- Classify semantic content of status updates using LIWC
- Emotion: fraction of posts with positive/negative words
Individual-Level Model

\[ y_{jt} = \theta_t + f_j + \lambda \frac{1}{\delta_{jt}} \sum_i a_{ijt} y_{it} + \epsilon_{jt} \]

- \( y_{jt} \) = user \( j \)'s happiness at time \( t \), fraction of posts with positive/negative words
- \( j \) = user whose emotion we’re predicting
- \( i \) = a friend of user \( j \)
- \( t \) = time window of interest
- \( \Theta_t \) = time-related fixed effect (there are “happy times”)
- \( f_j \) = user-related fixed effect (there are “happy users”)
- \( \delta_{jt} \) = degree of user \( j \) at time \( t \) (friends come and go)
- \( a_{ijt} \) = strength of relationship at time \( t \) between \( i \) and \( j \)

Cumulative effect of a user on their friends

Computationally demanding. One observation per (user, time) pair
City-Level Model

\[ \bar{y}_{gt} = \theta_t + \bar{f}_g + \lambda \sum_i n_g \sum_{j \in S_a} \frac{1}{\delta_{jt}} a_{ijt} + \epsilon_{gt} \]

- \( \bar{y}_{gt} \): average emotion in city \( g \) at time \( t \)
- \( \theta_t \): time-related fixed effect (there are “happy times”)
- \( \bar{f}_g \): city-related fixed effect (there are “happy cities”)
- \( \lambda \): average strength of relationship between \( i \) and an individual in city \( g \)
- \( n_g \): number of users in city \( g \)
- \( \epsilon_{gt} \): residual error

\( Y_{gt} \) = average emotional influence on an individual in city \( g \)

- \( g \): city whose aggregate emotion we’re predicting
- \( \Theta_t \): time-related fixed effect (there are “happy times”)
- \( f_g \): city-related fixed effect (there are “happy cities”)
- \( n_g \): number of users in city \( g \)
Instrument $z$ should not affect $\bar{y}_{gt}$ when $Y_{gt}$ is held constant.

But weather in $g$ and in friends’ cities could be correlated. => Weather directly influences emotions in $g$!

Break correlation between friends’ rain and user’s rain

- Restrict data to (city,day) **WITHOUT** rain
- Restrict data to (city,day) **WITH** rain
Rainfall as Instrumental Variable

Explanatory variable (social contagion)

Instrumental variable $z_{gt}$

- RAIN: 1 if it rained, 0 otherwise

Two-stage regression

Friends’ rain $z_{gt}$ → Friends’ emotion $Y_{gt}$ → Users’ emotion $\bar{y}_{gt}$
Results

rain decreases positive posts by 1% 

\( \beta \)

a positive post leads to another 1.7
Results

They found evidence for emotional contagion

But does rain really impact happiness?

What are we measuring?
- People complain about rain on Twitter, ok. Does that make them unhappy?

Just conversational dynamics?
- User A: All the rain is making me depressed.
- User B: Poor guys who have to suffer in the rain.
Natural Experiments
Summary

Natural experiments can be powerful alternatives to experiments

Watch out for cutoffs that cannot be directly manipulated by individuals (e.g. size- or age-based thresholds, eligibility criteria, location, …)

Find variables that are highly correlated with your regressor but not with your outcome
Difference in Differences

- Control and treatment groups are not randomly assigned and start at different “levels” before treatment.

- Sometimes control and treatment group outcomes can be assumed to move in parallel.
Difference in Differences

- P: treatment group
- S: control group
  
  Start from different levels
  Not randomized experiment

- Difference between $P_2$ (observed outcome of treatment group) and $Q$ (expected outcome of treatment group) is the treatment effect

"Illustration of Difference in Differences" by Danni Ruthvan - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons
Twitter for Migration Studies

E. Zagheni, K. Garimella, I. Weber, B. State: Inferring International and Internal Migration Patterns from Twitter Data. WWW’14

- Use streaming API filter for geo-tagged tweets from OECD countries
- Pick 3,000 users per country, get their tweets
- Estimate out-migration and oversample “static” countries
- Get data for ~500K users
- After activity thresholding left with ~15K
Estimated Out-Migration Rates

- average OECD
- Mexico
- USA
- Germany
- Japan
Difference-in-Differences

Out-migration rates clearly an overestimate
Main reason: Non-representative user set
Selection bias is changing over time in unknown manner
Assumption: changing “in lockstep” for different countries
Focus on between-country differences

\[ \hat{\delta}_c^t = (m_c^t - m_{oeecd}^t) - (m_c^{t-1} - m_{oeecd}^{t-1}) \]

= the out-migration rate from country c to all other countries
... = # users identified as migrants from c / # users in c
Relative changes 2011 -> 2012

Limited by data sparsity: migration is a rare event