Econometrics 1: Tutorial IV

Heikki Korpela

December 4, 2023

Contents

1. Practical matters

- An extra tutorial will be held on Friday (Dec 8) at 2.15pm
- My apologies if this is inconvenient for you
  - I had to find a time quickly that seemed like it would have the least clash with lectures for othe core Economics courses
  - Next Monday you will have the last regular lecture with the professor
- You can also email me (as many of you have), and also schedule a meeting if you want to discuss any problems

2. Issues with data

- The dataset for problem 2 is actually quite instructive: it illuminates (depressingly) common features of real data.
- You usually need to read metadata rather carefully to learn what the variables actually mean.
- Technical issues and corrupting data by accident are quite common.
- The original `schooling.txt` had an issue, which could cause it to fail to import.
  - The first row of data had different numbers of whitespace between column names.
  - Software can easily get confused by things like this. There is no universal rule for how to cope with such inconsistencies, so the software often makes a best guess.

How to cope?

- **Look at your data.** Calculate summary statistics like the mean, and check if they make rough sense.
  - Example: a variable holding values like 50 000 is never going to be in logs. (Base R's numeric value can hold values up to about $\exp(709.78)$.)
- If you run into trouble or suspect issues, check out options and help pages for the functions you use to import data.
- Read the metadata. (Generally not required for this course, the assignments should have the descriptions required or refer you to the slides/textbook.)

The journey to metadata 1/6

ILLUSTRATION: ESTIMATING THE RETURNS TO SCHOOLING **159**

In this section we use data on 3010 men taken from the US National Longitudinal Survey of Young Men, also employed in Card (1995). In this panel survey, a group of

The bibliography reference in the textbook description of the data.

The journey to metadata 2/6

Cameron, A. C. and Windmeijer, F. A. G. (1996), *R*-squared Measures for Count Data Regression Models with Applications to Health Care Utilization, *Journal of Business and Economic Statistics*, 14, 209–220.

Cameron, A. C. and Windmeijer, F. A. G. (1997), An *R*-squared Measure of Goodness of Fit for Some Common Nonlinear Regression Models, *Journal of Econometrics*, 77, 329–342.

Campbell, J. Y. and Perron, P. (1991), Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots. In: O. Blanchard and S. Fisher, eds, *NBER Macroeconomics Annual*, 6, 141–201, MIT Press, Cambridge.

Campbell, J. Y. and Shiller, R. J. (1991), Yield Spreads and Interest Rate Movements: A Bird's Eye View, *Review of Economic Studies*, 58, 495–514.

Campbell, J. Y. and Shiller, R. J. (1998), Valuation Ratios and the Long-Run Stock Market Outlook, *Journal of Portfolio Management*, 24, 11–26.

Campbell, J. Y. and Thompson, S. B. (2008), Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?, *Review of Financial Studies*, 21, 1509–1531.

Campbell, J. Y., Lo, A. W. and MacKinlay, A. C. (1997), *The Econometrics of Financial Markets*, Princeton University Press, Princeton, NJ.

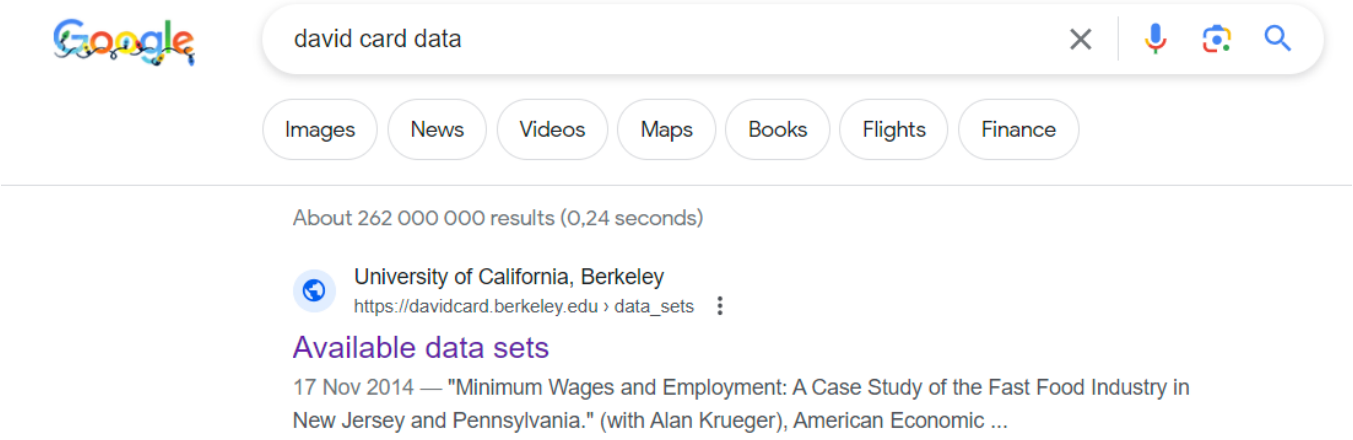
Canova, F. (1995), The Economics of VAR Models. In: K. D. Hoover, ed., *Macroeconometrics: Developments, Tensions and Prospects*, Kluwer Academic Publishers, Boston, MA, 57–97.

Canova, F. (2007), *Methods for Applied Macroeconomic Research*, Princeton University Press, Princeton, NJ.

Card, D. (1995), Using Geographical Variation in College Proximity to Estimate the Return to Schooling. In: L. N. Christofides, E. K. Grant and R. Swidinsky, eds, *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, University of Toronto Press, Toronto, Canada, 201–222.

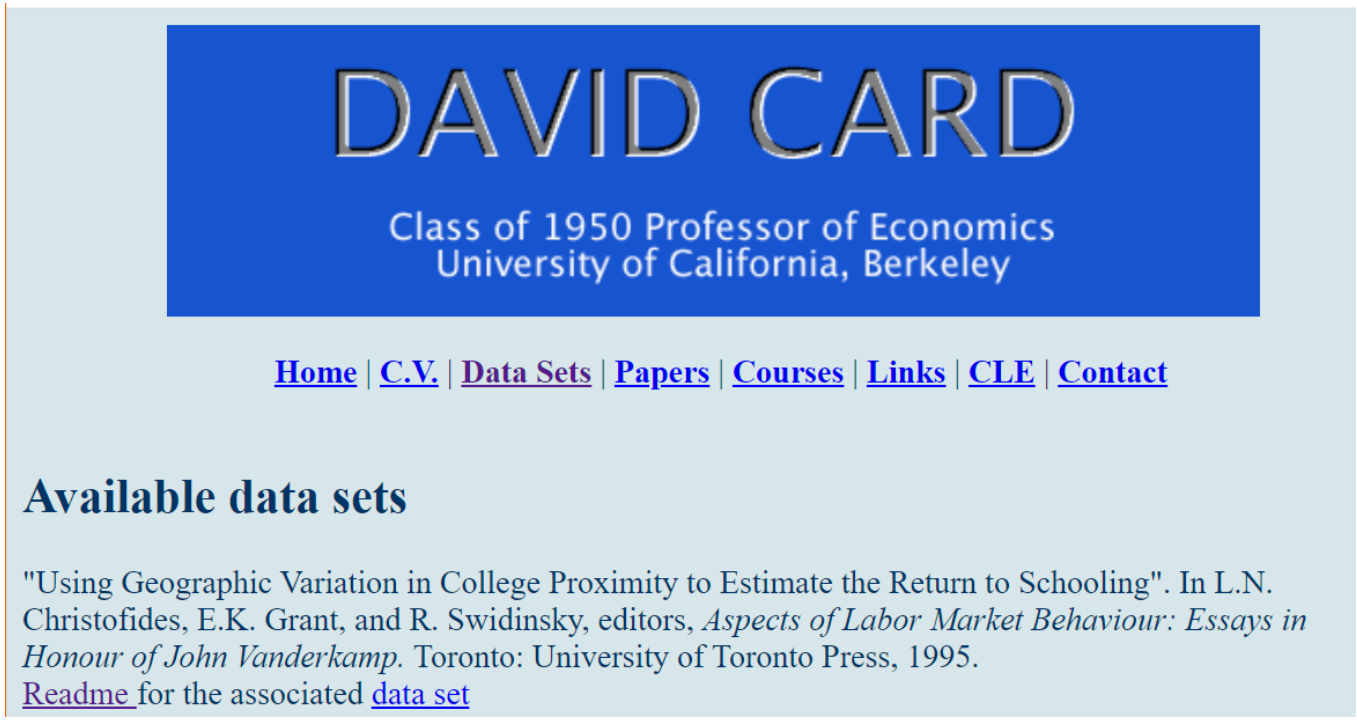
The literature reference in the bibliography.

The journey to metadata 3/6



Quick and dirty: Google search by author's name.

The journey to metadata 4/6



DAVID CARD

Class of 1950 Professor of Economics
University of California, Berkeley

Home | C.V. | Data Sets | Papers | Courses | Links | CLE | Contact

Available data sets

"Using Geographic Variation in College Proximity to Estimate the Return to Schooling". In L.N. Christofides, E.K. Grant, and R. Swidinsky, editors, *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp.* Toronto: University of Toronto Press, 1995.
Readme for the associated data set

The dataset under the author's homepage.

The journey to metadata 5/6

code_bk.txt

```
Code Book for Ascii File nls.dat

Note: For more information, see the original article:
 David Card,
"Using Geographic Variation in College Proximity to Estimate the Return to Schooling"
 NBER Working Paper 4832, August 1994

This article is published (with the same title and identical tables)
in  Aspects of Labour Market Behaviour: Essays in Honour of John Vandekamp"
edited by Louis N. Christofides, E. Kenneth Grant, and Robert Swidinsky
Toronto: University of Toronto Press, 1995.

The file contains 3613 observations on men in 1976 cross-section
of nls young men (original nls cohort)

Missing value code is .

 Column Locations      Variable Name and Label
-----------------------------------------------------------------------
       1 -  5     id    /*sequential id runs from 1 to 5225 */
       7 -  7     nearc2  /*grew up near 2-yr college*/
      10 - 10     nearc4  /*grew up near 4-yr college*/
      12 - 13     nearc4a /*grew up near 4-yr public college*/
      15 - 16     nearc4b /*grew up near 4-yr priv college*/
      18 - 19     ed76  /*educ in 1976*/
      21 - 22     ed66  /*educ in 1966*/
      24 - 25     age76  /* age in 1976
      27 - 31     daded /*dads education missing=avg*/
      33 - 33     nodaded /* 1 if dad ed imputed*/
      35 - 39     momed /*moms education*/
      41 - 41     nomomed /* 1 if mom ed imputed*/
      43 - 54     weight  /* nls weight for 1976 cross-section */
      56 - 56     momdad14 /*1 if live with mom and dad age 14*/
      58 - 58     sinmom14 /*lived with single mom age 14*/
      60 - 60     step14  /*lived step parent age 14*/
      62 - 62     reg661 /* dummy for region=1 in 1966 */
      64 - 64     reg662  /* dummy for region=2 in 1966 */
      66 - 66     reg663  /* dummy for region=3 in 1966 */
      68 - 68     reg664
      70 - 70     reg665
      72 - 72     reg666
      74 - 74     reg667
      76 - 76     reg668
      78 - 78     reg669  /* dummy for region=9 in 1966 */
      80 - 80     south66 /*lived in south in 1966*/
      82 - 82     work76 /* worked in 1976*/
      84 - 84     work78 /* worked in 1978*/
      86 - 97     lwage76  /*log wage (outliers trimmed) 1976 */
      99 -110     lwage78  /*log wage in 1978 outliers trimmed */
     112 -112     famed /*mom-dad education class 1-9*/
     114 -114     black  /* 1 if black*/
     116 -116     smsa76r /*in smsa in 1976*/
     118 -118     smsa78r /*in smsa in 1978*/
     120 -120     reg76r /*in south in 1976*/
     122 -122     reg78r /*in south in 1978*/
     124 -124     reg80r /* in south in 1980*/
     126 -126     smsa66r /* in smsa in 1966*/
     128 -132     wage76 /*raw wage cents per hour 1976*/
     134 -138     wage78
     140 -144     wage80
     146 -146     noint78 /*1 if noninterview in 78*/
     148 -148     noint80
     150 -150     enroll76 /*1 if enrolled in 76*/
     152 -152     enroll78
     154 -154     enroll80
     156 -157     kww  /*the kww score*/
     159 -161     iq  /* a normed iq score*/
     163 -163     marsta76 /*mar status in 1976 1=married, sp. present *
     165 -165     marsta78
     167 -167     marsta80
     169 -169     libcrd14   /*1 if lib card in home age 14*/
```

Variable descriptions in the dataset.

The journey to metadata 6/6

```
List of means, min/max  Note some vars are missing for some observations.
Missing value code is .



 Variable     N          Mean       Std Dev       Minimum       Maximum
 ---------------------------------------------------------------------------
 ID        3613       2609.78       1498.51     2.0000000       5225.00
 NEARC2    3613     0.4317741     0.4953919             0     1.0000000
 NEARC4    3613     0.6781068     0.4672669             0     1.0000000
 NEARC4A   3613     0.4921118     0.5000070             0     1.0000000
 NEARC4B   3613     0.1859950     0.3891565             0     1.0000000
 ED76      3613    13.2252975     2.7497411             0    18.0000000
 ED66      3613    10.7428730     2.4590854             0    18.0000000
 AGE76     3613    28.1752007     3.1718104    24.0000000    34.0000000
 DADED     3613    10.0028785     3.2960212             0    18.0000000
 NODADED   3613     0.2241904     0.4171058             0     1.0000000
 MOMED     3613    10.3421672     3.0293785             0    18.0000000
 NOMOMED   3613     0.1143094     0.3182308             0     1.0000000
 WEIGHT    3613     320318.35     168006.76      75607.00    1752340.00
 MOMDAD14  3613     0.7921395     0.4058326             0     1.0000000
 SINMOM14  3613     0.1001937     0.3002997             0     1.0000000
 STEP14    3613     0.0384722     0.1923599             0     1.0000000
 REG661    3613     0.0445613     0.2063671             0     1.0000000
 REG662    3613     0.1549958     0.3619508             0     1.0000000
 REG663    3613     0.1940216     0.3955003             0     1.0000000
 REG664    3613     0.0691946     0.2538199             0     1.0000000
 REG665    3613     0.2095212     0.4070232             0     1.0000000
 REG666    3613     0.0929975     0.2904691             0     1.0000000
 REG667    3613     0.1101578     0.3131296             0     1.0000000
 REG668    3613     0.0309992     0.1733394             0     1.0000000
 REG669    3613     0.0935511     0.2912434             0     1.0000000
 SOUTH66   3613     0.4126764     0.4923837             0     1.0000000
 WORK76    3613     0.8350401     0.3711957             0     1.0000000
 WORK78    3613     0.7351232     0.4413287             0     1.0000000
 LWAGE76   3010     6.2618319     0.4437977     4.6051702     7.7848893
 LWAGE78   2639     6.3291080     0.4442450     4.6965200     8.2409240
 FAMED     3613     5.9128148     2.6504318     1.0000000     9.0000000
 BLACK     3613     0.2300028     0.4208925             0     1.0000000
 SMSA76R   3613     0.6947135     0.4605924             0     1.0000000
 SMSA78R   3319     0.6929798     0.4613273             0     1.0000000
 REG76R    3613     0.3996679     0.4898978             0     1.0000000
 REG78R    3319     0.3968063     0.4893089             0     1.0000000
 REG80R    3227     0.4028509     0.4905473             0     1.0000000
 SMSA66R   3613     0.6426792     0.4792768             0     1.0000000
 WAGE76    3017   576.0888300   263.8199090    25.0000000       2404.00
 WAGE78    2656   724.5591114   526.1991520    17.0000000      17628.00
 WAGE80    2520   869.8940476   492.1729068    27.0000000      13857.00
 NOINT78   3613     0.0813728     0.2734447             0     1.0000000
 NOINT80   3613     0.1068364     0.3089479             0     1.0000000
 ENROLL76  3613     0.0946582     0.2927827             0     1.0000000
 ENROLL78  3317     0.0654206     0.2473038             0     1.0000000
 ENROLL80  3220     0.0583851     0.2345066             0     1.0000000
 KWW       3543    33.4891335     8.6918079             0    56.0000000
 IQ        2470   102.5878543    15.4450703    50.0000000   156.0000000
 MARSTA76  3604     2.3571032     2.1096377     1.0000000     6.0000000
 MARSTA78  3319     2.2136186     2.0058342     1.0000000     6.0000000
 MARSTA80  3227     2.1041215     1.9088835     1.0000000     6.0000000
 LIBCRD14  3598     0.6717621     0.4696372             0     1.0000000
 ---------------------------------------------------------------------------
```

Variable means in the dataset.

Missing values

- Missing values in data do not inherently mean the data is corrupted.
- Example: what is the annual wage for a person who is not employed? Is zero a meaningful value for their wage?
- By default, R's `lm` ignores any rows where any regressors (or the dependent variable) have missing values.

Missing values

- There are at least three ways to cope with missing data. The appropriate method depends on the context.
  - Dropping the rows with missing data (you can use this during this course; it's what R does by default)
  - Multiple imputation (replace missing values with predictions); not covered in this course
  - Discretizing continuous values and including missingess as an additional category; not covered in this course

Dropping missing values in R

```
schooling <- read.table(
        'schooling.txt', header = T)
schooling <- subset(schooling, !is.na(iqscore))
```

3. Instrumental variables

- Instrumental variables are one way of dealing with cases where the regular OLS assumptions don't hold
- It is one (of many) causal identification strategies; others include
- Other popular strategies:
  - Differences-in-differences (DiD or DD)
  - Regression discontinuity (RD)
  - Randomized controlled trials (RCT)
  - Adjustment/unconfoundedness (see also: matching and weighting)

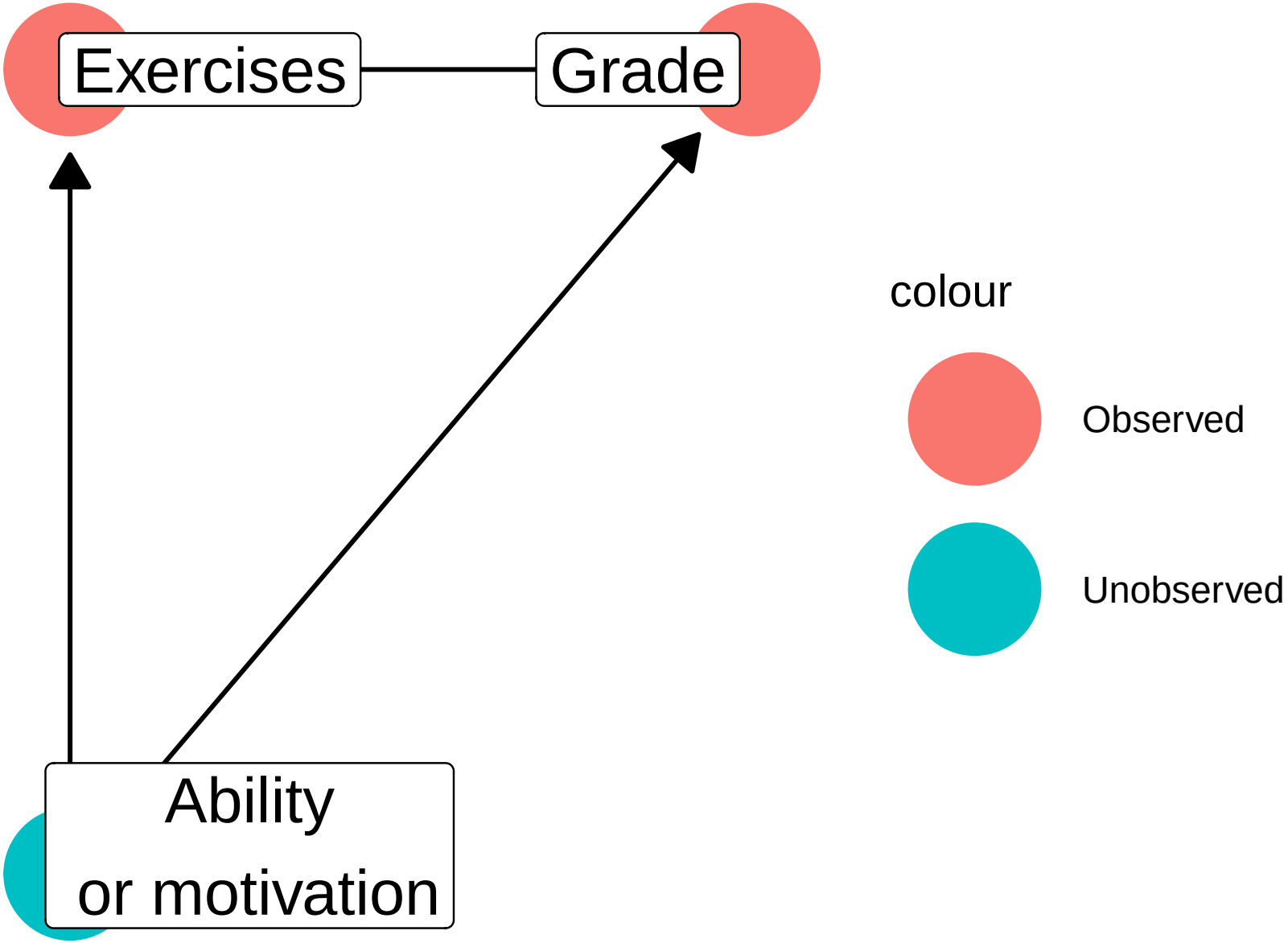
Further stuff on causality

- The books: Angrist and Pischke: Mostly Harmless Econometrics, or Mastering Metrics (less technical)
- Complementary reading
  - Pearl: The Book of Why
  - Imbens 2020: Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics
- Courses
  - Applied Microeconometrics I and II (Aalto University) cover empirical research using micro-level data
  - Applied Macroeconometrics I and II (UoH) cover time series using macroeconomic data
  - Applied courses are very helpful if you want to do or at least cite and understand causal empirical work in your thesis (you probably will)

The problem to be solved

- Suppose we want to know how forcing everyone to do exercises affects learning
  - We also assume grades are a good measure of learning
- But those scoring well on exercises probably were more motivated or able to start with
  - They might have done well on the exam without the exercises
  - Motivation/ability is usually (partially) unobserved
- Thus, the regression might show positive correlation between exercises and course grade even if exercises do <u>nothing</u> for learning (more commonly, the regression might under- overestimate the effect)

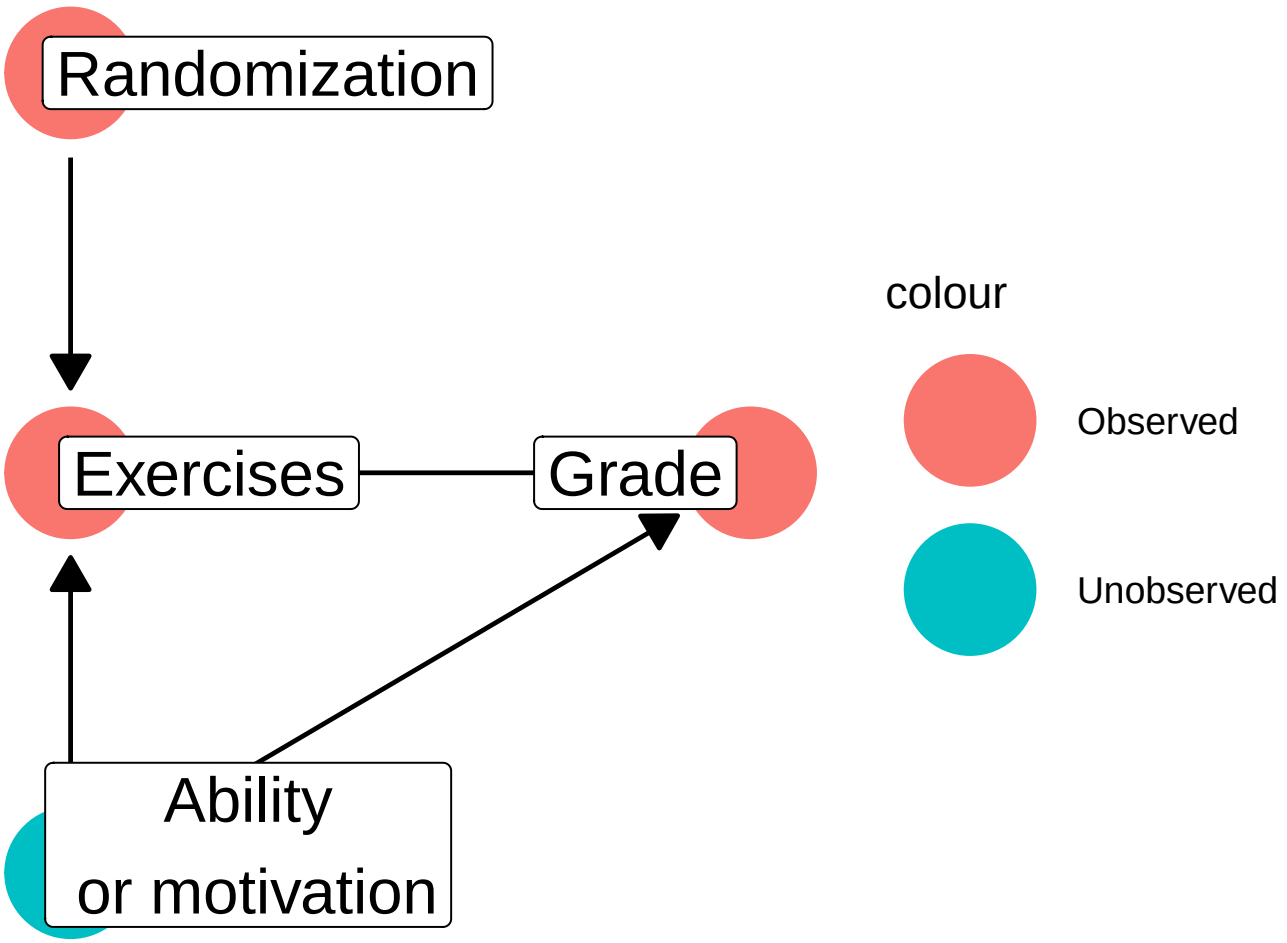
The problem to be solved



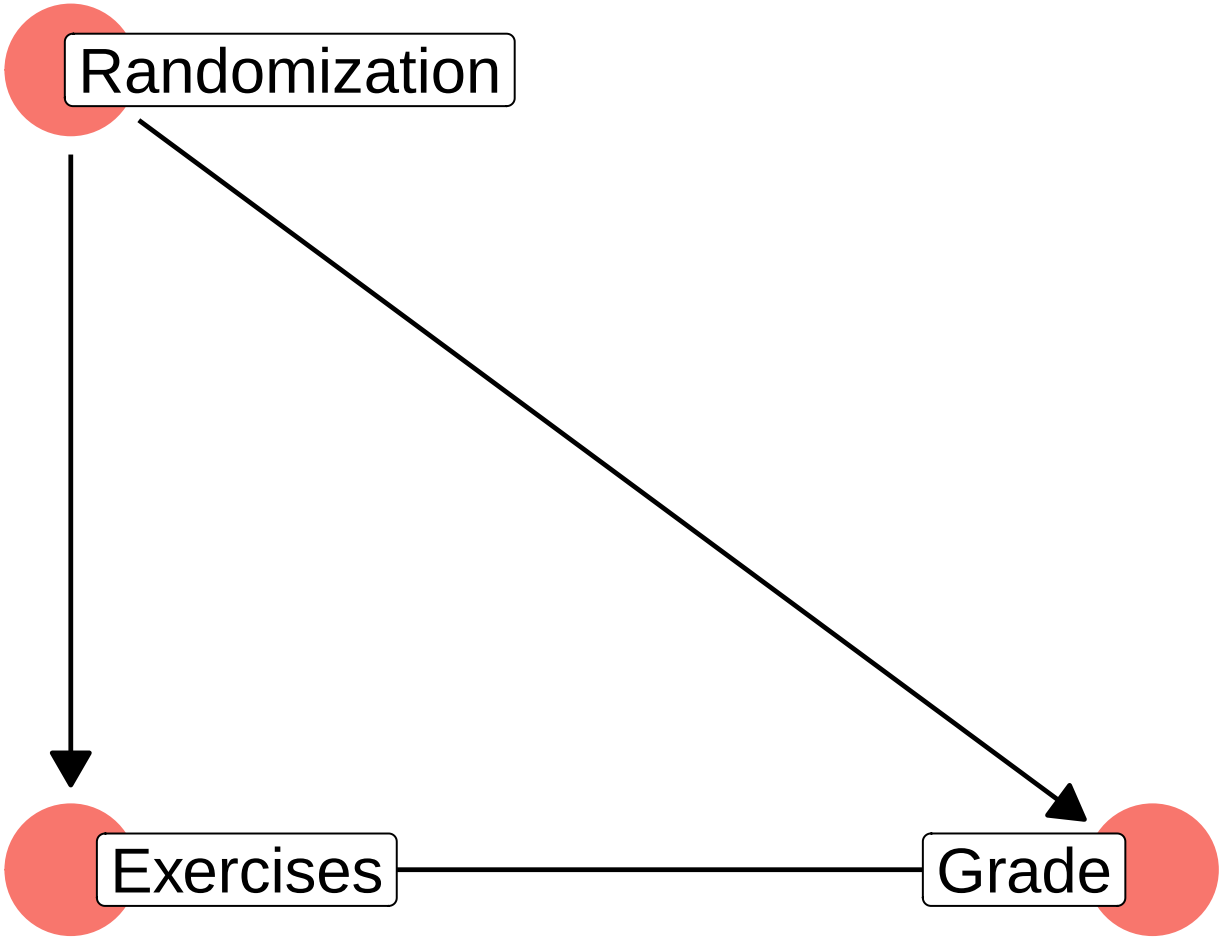
The solution

- One solution is a randomized controlled trial (RCT)
- A randomly selected half has to do the exercises to pass the course
- Successful random assignment is, by construction, independent of grades, ability, and whatever else you might think of.
- However, the RCT might have imperfect takeup: not everyone assigned to do exercises does them
- In this case, we can (under certain conditions) still use random assignment as an instrument for the exercises

The solution

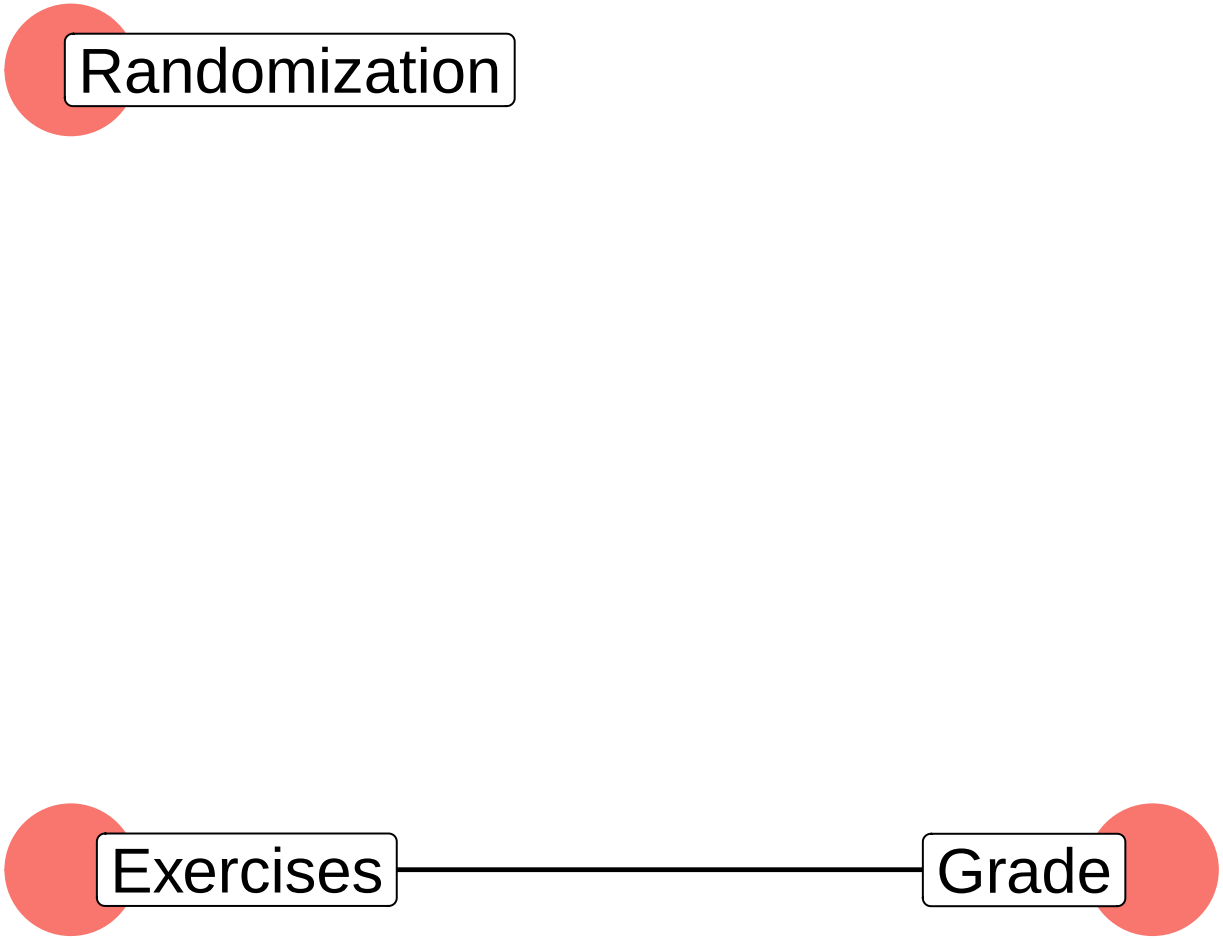


Bad instrument: endogenous



The instrument should not be correlated with the outcomes.

Bad instrument: irrelevant



The instrument should be correlated with the regressor.

Instruments: only one strategy

- Instrument relevance is easy to test for, but a difficult problem to solve when it occurs (and also sadly common)
- An even bigger problem is that the endogeneity assumption is only partly testable
- The Sargan test **assumes that at least some instruments are valid**.
  - See e.g. the textbook p. 168
- IV results are local: they are informative about how instrument-induced changes in X affect Y
- Any identification strategy (other than the idealized RCT) will include some untestable assumptions; generally, the question is about the plausibility of the assumptions

Instruments: only one strategy

- Applied work based on IV alone is arguably more prone to criticism than, for example, DiD and RD (which have their own issues)
- Coming up with such criticism is arguably the easiest job in economics
- The extreme (silly) version of a skeptical take is known as "Friends don't let Friends do IV"
- A more sane approach comes under the heading "Friends *Do* Let Friends Do IV":

  > *If you are going to use an observational IV, you do need to think very carefully. – –*
  > *Identifying causal effects is hard. Willingly limiting yourself to a subset of methods and declaring one method off-limits is like a football coach saying he doesn't want his quarterback to ever try to pass the ball.*