Introduction to Mendelian Randomization

Zhiguang Huo

Department of Biostatistics
University of Florida

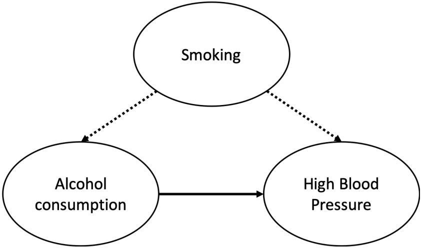
Febuary 26^{th}, 2020

Outline for Mendelian randomization (MR)

- Motivation
- Concepts and goals
- Assumption
- Mathematical theory
- Model diagnostic
- Method
  - One sample MR
  - Two sample MR

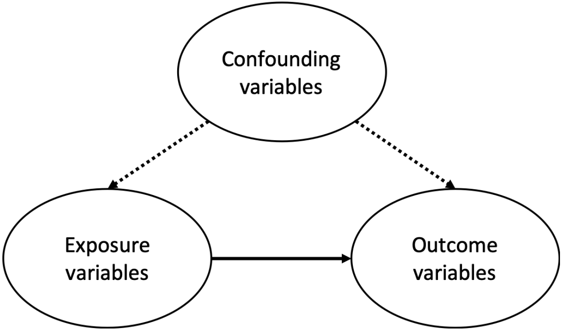
Motivating example

- Goal: investigate the effect of alcohol consumption on blood pressure.
- Observational studies have shown the association between alcohol consumption and blood pressure (Marmot et al., 1994; Fuchs et al., 2001).
- This association could not imply causal effect because of confounders.
  - Smoking increases alcohol assumption.
  - Smoking increases blood pressure.



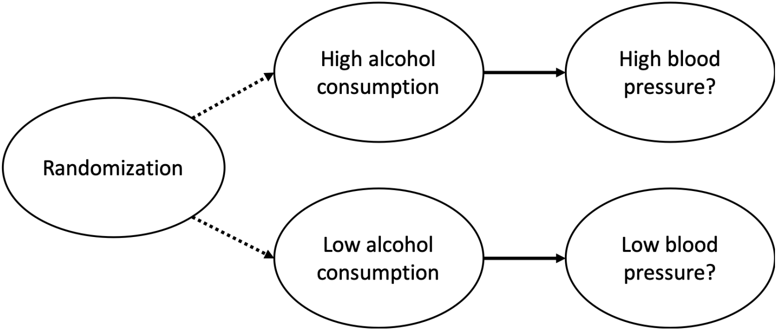
Motivation

- Observational study is very popular in biomedical studies.
- The association from observational study does not imply causality, because of confounding variables.



- For known confounders, we can adjust them as covariates.
- For unknown confounders:
  - Randomized clinical trials (RCT)

Randomized clinical trials (RCT)



- RCT will lead to causal relationships between alcohol consumption and blood pressure.
- Drawbacks:
  - Time consuming
  - Loss of follow-up participants

Idea of Mendelian randomization

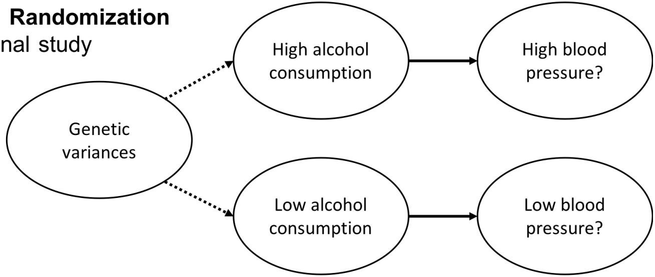
- Alcohol is initially metabolised to acetaldehyde, which can be further eliminated (Davies et al., 2018).
- The major enzyme for this elimination is alcohol dehydrogenase 2 (ALDH2).
- A variant in the ALDH2 gene (Chen et al., 2008) (rs671, reference allele G)
  - Alternative allele A was found in east Asian population
  - Causes a flush response and slows the metabolism of acetaldehyde
- In a study of 4,057 participants (Takagi et al., 2001)
  - Those with two copies of A drank an average of 1.1 g of alcohol.
  - Those with no copies of A allele drank 23.7 g.

Ideas behind MR

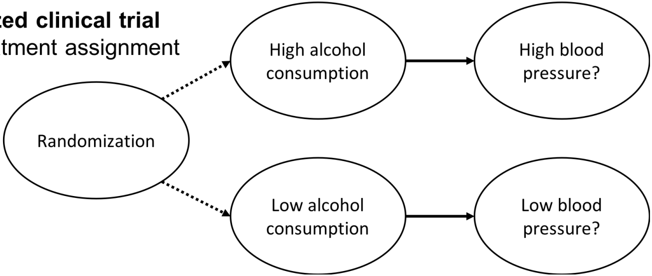
- The genetic variants are inherited from parents
    - This genetic variant is not affected by confounding variables (smoking)
    - This genetic variant is not affected by blood pressure level.
- The genetic variant can define groups of different level of alcohol consumption
    - If allele A non-carriers drank heavy, and had higher blood pressure
    - If allele A carriers drank light, and had lower blood pressure
    - Genetic variants can be thought of random allocation
- Then we can conclude the effect of alcohol consumption on blood pressure is causal.
    - The relationship is not likely to be confounded.
    - It is not likely that blood pressure causes alcohol consumption (reverse causion)

Conceptual analogy between MR and randomized clinical trials (RCT)

Mendelian Randomization
Observational study



Randomized clinical trial
Active treatment assignment



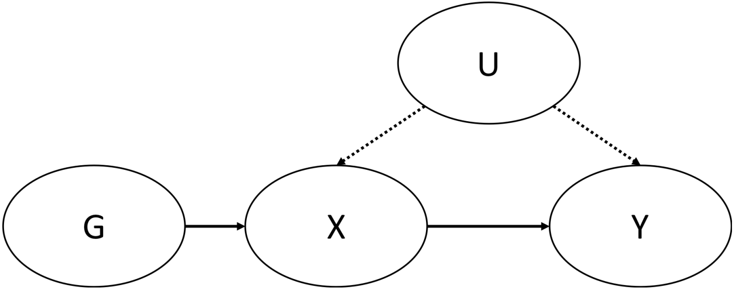
Mendelian randomization (MR)

- Idea: If we cannot randomize the exposure, we can find a randomized instrumental variable to disentangle.
- Genetic variants are also referred as instrumental variables
- The original idea was first proposed as economy field, and MR is also called instrument variable (IV) regression.

Goals of MR studies:

1. Test the causal relationship between the exposure variables and the outcome variable.
2. Estimate the causal effect of the exposure variable on the outcome variable.

Notations



- G: genetic variant
- Y: outcome variable
- X: exposure variable
- U: unknown confounders

Three core assumptions for hypothesis testing

G: genetic variant; Y: outcome variable; X: exposure variable; U: unknown confounders

1. Independence between G and U

$$G \perp U$$

2. Established association between G and X

$$P(X|G) \neq P(X)$$

3. No alternative pathway from G to Y, (exclusion restriction)

$$G \perp Y|X, U$$

- Theorem: testing $G - Y$ association is equivalent to testing causal relationship $Y - X$.

Testing causal relationship (Didelez et al., 2010)

$$\begin{aligned} P(Y,G) &= \int_U \int_X P(Y,X,U,G) \\ &= \int_U \int_X P(Y|X,U)P(X|G,U)P(U)P(G) \\ &= P(G)\int_U P(U) \int_X P(Y|X,U)P(X|G,U) \end{aligned}$$

If $Y \perp X|U$, i.e., $P(Y|X,U) = P(Y|U)$,

$$\begin{aligned} P(Y,G) &= P(G)\int_U P(U)P(Y|U) \int_X P(X|G,U) \\ &= P(G)P(Y) \end{aligned}$$

- Therefore, $Y \perp X|U \rightarrow Y \perp G$
- Under the pre-mentioned assumptions, we only need to test whether Y and G are independent, in order to establish causal relationship between X and Y

Estimating causal effect in linear models

- Two more assumptions for linear regression:
  - The effect of X on Y is linear.
  - No interaction between X and U.
- Suppose data generating models are

$$X = \alpha_0 + \alpha_1 G + \alpha_2 U + \varepsilon_1$$
$$Y = \beta_0 + \beta_1 X + \beta_2 U + \varepsilon_2$$

- We can obtain the following relationship

$$\mathbb{E}(X|G) = \alpha_0 + \alpha_1 G$$
$$\mathbb{E}(Y|G) = \theta_0 + \theta_1 G$$

IV estimators are essentially ratio estimators

- Since

$$\begin{aligned}\theta_1 &= \mathbb{E}[Y|G = g+1] - \mathbb{E}[Y|G = g] \\ &= \beta_1(\mathbb{E}[X|g+1] - \mathbb{E}[X|g]) + \beta_2(\mathbb{E}[U|g+1] - \mathbb{E}[U|g]) \\ &= \beta_1\alpha_1\end{aligned}$$

- Therefore, $\beta_1 = \theta_1/\alpha_1$
- When $X \in \mathbb{R}$, $G \in \mathbb{R}$, the IV estimator can be written as the ratio of two OLS estimator

$$\hat{\beta}_{IV} = \frac{\hat{\theta}_1}{\hat{\alpha}_1}$$

- The se of $\hat{\beta}_{IV}$ can be determined by delta method (Wald, 1940).

Instrumental Variable estimation in linear models

- Suppose $\boldsymbol{G} \in \mathbb{R}^{n \times l}$ and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ have same dimension (i.e., $p = l$, both may contain intercept), and confounder $\boldsymbol{U}$ is absorbed in the error ε

$$Y = \boldsymbol{X}^\top \boldsymbol{\beta} + \varepsilon$$

- The usual OLS does not give unbiased estimation for unconfounded effect, because $\boldsymbol{X}$ and ε are correlated.

$$\boldsymbol{X}^\top Y = \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{X}^\top \varepsilon$$

- If the instrument G is independent of error ε

$$\boldsymbol{G}^\top Y = \boldsymbol{G}^\top \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{G}^\top \varepsilon$$
$$\hat{\boldsymbol{\beta}}_{IV} = (\boldsymbol{G}^\top \boldsymbol{X})^{-1} \boldsymbol{G}^\top Y$$
$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}) \sim N(0, \sigma^2 Q_{GX}^{-1} Q_{GG} Q_{XG}^{-1}),$$

where $Q_{GX} = \lim_{n \to \infty} \frac{\boldsymbol{G}^\top \boldsymbol{X}}{n}$, $Q_{GG} = \lim_{n \to \infty} \frac{\boldsymbol{G}^\top \boldsymbol{G}}{n}$

Connection with the ratio estimator

Suppose $\boldsymbol{X} = (1, X) \in \mathbb{R}^{n\times 2}$, $\boldsymbol{G} = (1, g) \in \mathbb{R}^{n\times 2}$

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{IV} &= (\hat{\beta}_0, \hat{\beta}_{IV})^\top \\ &= (\boldsymbol{G}^\top \boldsymbol{X})^{-1} \boldsymbol{G}^\top Y \\ &= (\boldsymbol{G}^\top \boldsymbol{X})^{-1} (\boldsymbol{G}^\top \boldsymbol{G})(\boldsymbol{G}^\top \boldsymbol{G})^{-1} \boldsymbol{G}^\top Y \\ &= \{(\boldsymbol{G}^\top \boldsymbol{G})^{-1} (\boldsymbol{G}^\top \boldsymbol{X})\}^{-1} \{(\boldsymbol{G}^\top \boldsymbol{G})^{-1} \boldsymbol{G}^\top Y\}\end{aligned}$$

It can be verified that

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_1}{\hat{\alpha}_1},$$

where β_1 is the slope of regressing Y on g, α_1 is the slope of regressing X on g.

Generalized methods of moment

What if $\boldsymbol{G} \in \mathbb{R}^l$ has more dimension than $\boldsymbol{X} \in \mathbb{R}^p$ (i.e., $l > p$), more equations than the number of parameters.

$$g_n(\boldsymbol{\beta}) = \frac{1}{n}\boldsymbol{G}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

- If $l = p$, we could obtain an estimate of $\boldsymbol{\beta}$ by setting $g_n(\boldsymbol{\beta}) = 0$
- More generally, for some matrix $\boldsymbol{W} \in \mathbb{R}^{l\times l}$, let

$$J_n(\boldsymbol{\beta}) = n g_n(\boldsymbol{\beta})^\top \boldsymbol{W}_n g_n(\boldsymbol{\beta})$$

- The goal is to set $J_n(\boldsymbol{\beta})$ close to zero.
- $$\begin{aligned}\boldsymbol{\beta}_{GMM} &= \arg\min J_n(\boldsymbol{\beta}) \\ &= \{(\boldsymbol{X}^\top \boldsymbol{G})\boldsymbol{W}_n(\boldsymbol{G}^\top \boldsymbol{X})\}^{-1}(\boldsymbol{X}^\top \boldsymbol{G})\boldsymbol{W}_n(\boldsymbol{G}^\top \boldsymbol{Y})\end{aligned}$$
- The scale of $\boldsymbol{W}_n$ does not change $\boldsymbol{\beta}_{GMM}$

Optimal $\boldsymbol{W}_n$

- It can be proved that when $\boldsymbol{W}_n = (\frac{1}{n}\boldsymbol{G}^\top \boldsymbol{G}\hat{\sigma}^2)^{-1}$, $\boldsymbol{\beta}_{GMM}$ is optimal.
- $$\boldsymbol{\beta}_{GMM} = \{(\boldsymbol{X}^\top \boldsymbol{G})(\boldsymbol{G}^\top \boldsymbol{G})^\top (\boldsymbol{G}^\top \boldsymbol{X})\}^{-1}(\boldsymbol{X}^\top \boldsymbol{G})(\boldsymbol{G}^\top \boldsymbol{G})^\top (\boldsymbol{G}^\top \boldsymbol{Y})$$
- The asymptotic distribution
$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{GMM} - \boldsymbol{\beta}) \sim N(0, \sigma^2 Q_{GX}^{-1} Q_{GG} Q_{XG}^{-1}),$$
- In the economics literature, this is also referred as two-stage least squares (2SLS) estimator, or instrumental variable estimator (IV)
$$\boldsymbol{\beta}_{IV} = \{(\boldsymbol{X}^\top \boldsymbol{G})(\boldsymbol{G}^\top \boldsymbol{G})^\top (\boldsymbol{G}^\top \boldsymbol{X})\}^{-1}(\boldsymbol{X}^\top \boldsymbol{G})(\boldsymbol{G}^\top \boldsymbol{G})^\top (\boldsymbol{G}^\top \boldsymbol{Y})$$

Two-stage least squares (2SLS) estimator

- 2SLS estimator

$$\boldsymbol{\beta}_{IV} = \{(\boldsymbol{X}^\top \boldsymbol{G})(\boldsymbol{G}^\top \boldsymbol{G})^\top (\boldsymbol{G}^\top \boldsymbol{X})\}^{-1} (\boldsymbol{X}^\top \boldsymbol{G})(\boldsymbol{G}^\top \boldsymbol{G})^\top (\boldsymbol{G}^\top \boldsymbol{Y})$$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}) \sim N(0, \sigma^2 Q_{GX}^{-1} Q_{GG} Q_{XG}^{-1}),$$

- computationally simple and stable
  1. Compute $\hat{\boldsymbol{X}}$ (i.e., regress $\boldsymbol{X}$ on $\boldsymbol{G}$, obtain fitted value)

$$\hat{\boldsymbol{X}} = \boldsymbol{G}(\boldsymbol{G}^\top \boldsymbol{G})^{-1} \boldsymbol{G} \boldsymbol{X}$$

  2. Then regress $\boldsymbol{Y}$ on $\hat{\boldsymbol{X}}$

$$\begin{aligned} \boldsymbol{\beta}_{IV} &= (\hat{\boldsymbol{X}}^\top \hat{\boldsymbol{X}})^{-1} \hat{\boldsymbol{X}}^\top \boldsymbol{Y} \\ &= \{(\boldsymbol{X}^\top \boldsymbol{G})(\boldsymbol{G}^\top \boldsymbol{G})^\top (\boldsymbol{G}^\top \boldsymbol{X})\}^{-1} (\boldsymbol{X}^\top \boldsymbol{G})(\boldsymbol{G}^\top \boldsymbol{G})^\top (\boldsymbol{G}^\top \boldsymbol{Y}) \end{aligned}$$

Intuition behind 2SLS

- Use instrumental variables (genetic variants) to exact the variation of the exposure variable (X) that is independent of confounding variables

$$\hat{\boldsymbol{X}} = \boldsymbol{G}(\boldsymbol{G}^\top \boldsymbol{G})^{-1}\boldsymbol{G}\boldsymbol{X}$$

- Use this part of variation to estimate the causal effect.

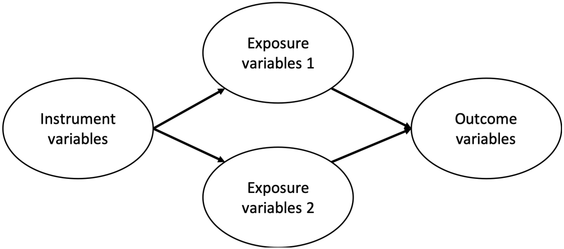
$$\boldsymbol{\beta}_{\boldsymbol{IV}} = (\hat{\boldsymbol{X}}^\top \hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}^\top \boldsymbol{Y}$$

Caution about assumptions

1. Independence between G and U, (usually untestable)
    - This is the assumption that introduce the “randomization”
2. Known association between G and X (testable)
    - Weak genetic instrument can lead to poor estimation of causal effect
    - Intuition: large variability in $\hat{\alpha}_1$ will lead to large variability in $\hat{\beta}_{IV}$.

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_1}{\hat{\alpha}_1},$$

3. No other pathway from G to Y other than through X (exclusion restriction)
    - Pleiotropy



Relaxed assumptions: adjust for known confounders

Suppose there is a set of known confounders W (population stratification, demographic/behavioral/socio-economical factor), denote U to be unknown confounders

1. $G \perp U | W$
2. G correlates with $X | W$
3. $G \perp Y | X, U, W$

- Testing $Y \perp X | W, U$ is equivalent to testing $Y \perp G | W$.
- In linear models, $\beta_1 = \theta_1 / \alpha_1$ still holds

$$\begin{aligned}
\mathbb{E}(Y|X, W, U) &= \beta_0 + \beta_1 X + \beta_2 W + \beta_3 U \\
\mathbb{E}(X|G, W) &= \alpha_0 + \alpha_1 G + \alpha_2 W \\
\mathbb{E}(Y|G, W) &= \theta_0 + \theta_1 G + \theta_2 W
\end{aligned}$$

All the previous math works!

Model diagnostic

- Independence between G and U
  - None
- Validity of the instruments
  - F-statistics (> 10) is the rule of thumb
- Pleitropy
  - Sargan's test
  - J-statistics
- Equavelence between $\boldsymbol{\beta}_{IV}$ and $\boldsymbol{\beta}_{OLS}$.
  - Durbin-Wu-Hausman test

Weak instrument variable

- Evaluate the validity of the instrument variable by fitting the following:

$$\hat{\boldsymbol{X}} = \boldsymbol{G}(\boldsymbol{G}^\top \boldsymbol{G})^{-1}\boldsymbol{G}\boldsymbol{X}$$

  - Goodness of modeling fitting is assessed by F-statistics
  - F-statistics > 10 indicates strong instrument variable
  - F-statistics < 10 indicates weak instrument variable

Overidentifying restrictions and Sargan's test

We can detect pleiotropy and the validity of IV if

- The number of IVs (I) is more than the number of causal effects (p) to be estimated; not all I equations can be exactly zero
- The null hypothesis is $\boldsymbol{G} \perp (\boldsymbol{Y} - \boldsymbol{X}\beta)$
  - Instrument is orthogonal to the error term
  - There is no direct effect left once conditional on $\boldsymbol{X}$
- Sargan's test (Sargan, 1958; Small, 2007) for 2SLS for I instrumental variables and 1 causal effect :

$$\{\boldsymbol{G}(Y - \hat{\boldsymbol{\theta}}_{2SLS}\boldsymbol{X})\}^\top\{\hat{\sigma}^2\boldsymbol{G}^\top\boldsymbol{G}\}^{-1}\{\boldsymbol{G}(Y - \hat{\boldsymbol{\theta}}_{2SLS}\boldsymbol{X})\} \to \chi^2(I-1)$$

under the null that all instruments are valid.

J-statistics

Hansen (1982) gave general results

$$J_n(\boldsymbol{\beta}) = n g_n(\boldsymbol{\beta})^\top \hat{\boldsymbol{W}}_n g_n(\boldsymbol{\beta}) \to \chi^2(l - p)$$

as long as $\hat{\boldsymbol{W}}_n$ converges to the optimal $\boldsymbol{W}_0$ and $\boldsymbol{\beta}$ is efficient GMM estimator.

- Large J-statistic will reject null hypothesis so that at least one instrument might be invalid

Test the equality of IV estimator and OLS estimator

The null hypothesis is OLS is consistent and fully efficient

- If there is no unmeasured confounders, OLS estimator will be consistent and efficient; IV is consistent under null or alternative
- Large discrepancy between $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{IV}$ suggests that there is confounding and OLS cannot be trusted.
- Durbin-Wu-Hausman test (Hausman, 1978)

$$(\hat{\boldsymbol{\beta}}_{IV} - \hat{\boldsymbol{\beta}}_{OLS})^\top D^{-1}(\hat{\boldsymbol{\beta}}_{IV} - \hat{\boldsymbol{\beta}}_{OLS}) \to \chi^2(p),$$

where $D = Var(\hat{\boldsymbol{\beta}}_{IV}) - Var(\hat{\boldsymbol{\beta}}_{OLS})$

One sample MR

- If we have everything in the same study (so-called one sample)
  - Instrument variable (genetic variants)
  - Exposure variable
  - Outcome variable
- We could apply 2SLS to examine the causal effect of the exposure variable on the outcome variable
- 2SLS has been implemented in the *ivreg* function in R package *AER*

Two sample MR

- One sample MR with 2SLS works great.
- Sometime, it is hard to have everything
  - Instrument variable (genetic variants)
  - Exposure variable
  - Outcome variable
- We could apply two sample MR method.

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_1}{\hat{\alpha}_1}$$

  - As long as we know the association between
    1. Instrument variable and the Exposure variable
    2. Instrument variable and the Outcome variable
  - The causal effect could be estimated

Resources for instrumental variables

- Summary statistics of genome-wide association study (GWAS):
  - GWAS catalog: `https://www.ebi.ac.uk/gwas/`
  - UK Biobank: `https://docs.google.com/spreadsheets/d/1kvPoupSzsSFBNSztMzl04xMoSC3Kcx3CrjVf4yBmESU`
- Web application for two sample MR
  - MR-base `http://app.mrbase.org/`

Two sample MR

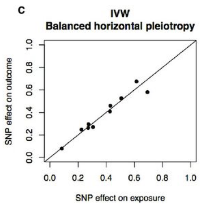
Methods:

- Single instrument variable: Wald method.

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_1}{\hat{\alpha}_1}$$

  - Single SNP
  - Ploygeneic risk score (PRS): summarize of multiple SNPs

- Multiple instrument variables: inverse-variance weighted (IVW) linear regression



See Hemani et al. (2018) for more details.

Summary

- MR is an effect way to establish causal relationship.
- Goals:
  1. Test existence of causal effect.
  2. Estimate the strength of the causal effect.
- Three core assumptions:
  1. Independence between G and U
  2. Established association between G and X
  3. No alternative pathway from G to Y, (exclusion restriction)
- Methods:
  1. One sample MR
  2. Two sample MR
- Model diagnostics.

Reference

- Major reference:
  - `https://research.fhcrc.org/content/dam/stripe/hsu/files/IV_Mendelian_lecture_1.pdf`
- Other references:
  - See next page

Chen, L., Smith, G. D., Harbord, R. M., and Lewis, S. J. (2008). Alcohol intake and blood pressure: a systematic review implementing a mendelian randomization approach. *PLoS medicine*, 5(3).

Davies, N. M., Holmes, M. V., and Smith, G. D. (2018). Reading mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *Bmj*, 362:k601.

Didelez, V., Meng, S., Sheehan, N. A., et al. (2010). Assumptions of iv methods for observational epidemiology. *Statistical Science*, 25(1):22–40.

Fuchs, F. D., Chambless, L. E., Whelton, P. K., Nieto, F. J., and Heiss, G. (2001). Alcohol consumption and the incidence of hypertension: The atherosclerosis risk in communities study. *Hypertension*, 37(5):1242–1250.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054.

Hausman, J. A. (1978). Specification tests in econometrics.

Econometrica: Journal of the econometric society, pages 1251–1271.

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., et al. (2018). The mr-base platform supports systematic causal inference across the human phenome. *Elife*, 7:e34408.

Marmot, M. G., Elliott, P., Shipley, M. J., Dyer, A. R., Ueshima, H., Beevers, D. G., Stamler, R., Kesteloot, H., Rose, G., and Stamler, J. (1994). Alcohol and blood pressure: the intersalt study. *Bmj*, 308(6939):1263–1267.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society*, pages 393–415.

Small, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479):1049–1058.

Takagi, S., Baba, S., Iwai, N., Fukuda, M., Katsuya, T., Higaki, J., Mannami, T., Ogata, J., Goto, Y., and Ogihara, T. (2001). The aldehyde dehydrogenase 2 gene is a risk factor for hypertension in japanese but does not alter the sensitivity to pressor effects of alcohol: the suita study. *Hypertension Research*, 24(4):365–370.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, 11(3):284–300.