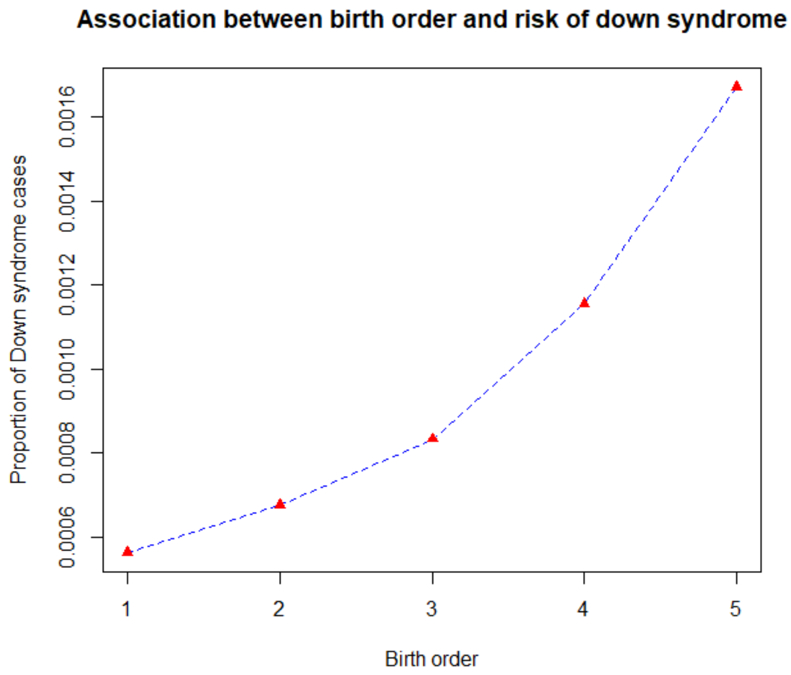
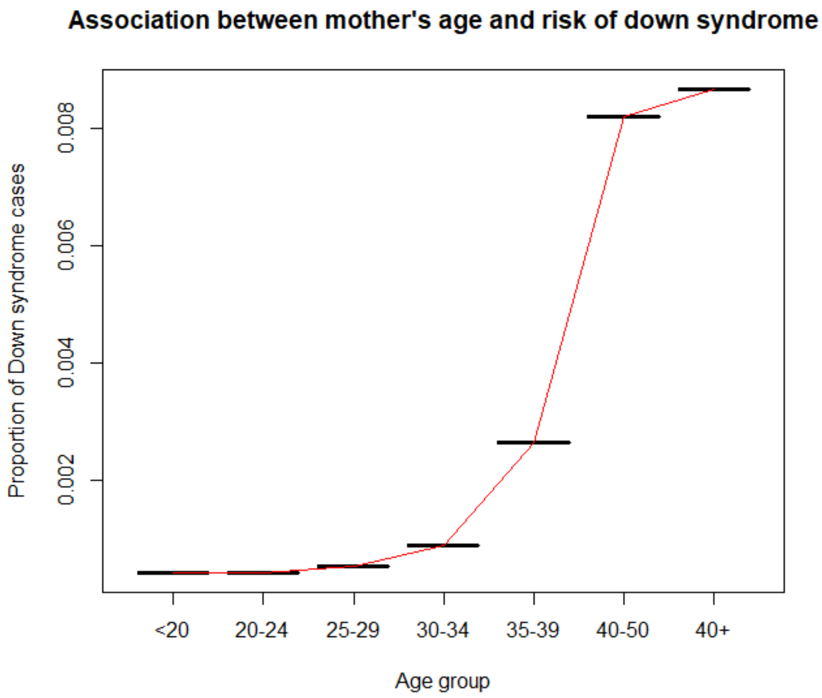
Matching Methods for Causal Inference

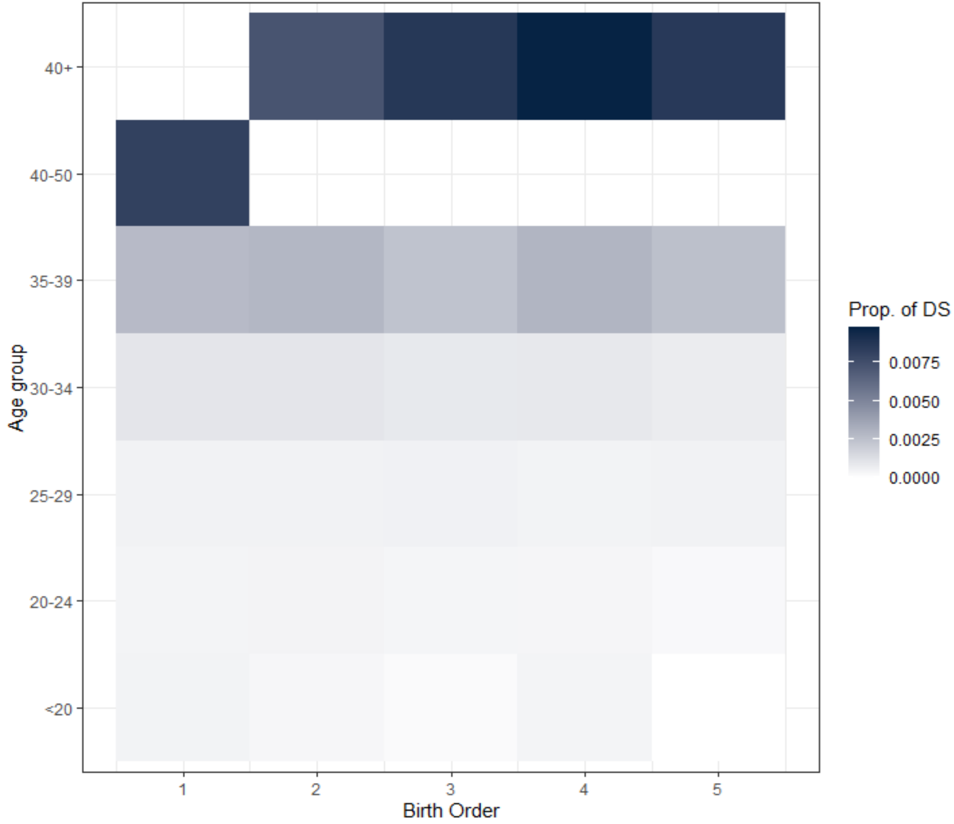
HPA 540/STAT 507: Applied Epi Research Methods

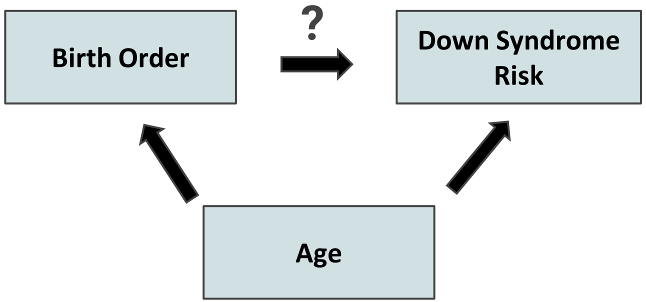
Confounding: A problem with observational studies





Confounding: A problem with observational studies





Association between birth order on Down syndrome risk

- Based on the heatmap, association between birth order and Down syndrome is confounded by age. Infact, birth order has **no effect** on risk of Down syndrome
- Let us see what happens if we ignore the confounder "Age".

$$H_0 : p_1 = p_2 = \ldots = p_5 \quad \text{vs} \quad H_1 : p_i \neq p_j \text{ for at least one } i \neq j$$

P-value ≈ 0 i.e. we have strong evidence that birth order is associated with risk of DS.

- **Incorrect conclusion!**

Birth Order 1	Birth Order 2	Birth Order 3	Birth Order 4	Birth Order 5
0.000563	0.000676	0.000833	0.001155	0.001671

Table1 : Proportion of Down Syndrome cases grouped by birth order

Association between birth order on Down syndrome risk

- We arrived at an incorrect conclusion because we ignored the confounder i.e. Age.
- If we would have conducted the hypothesis tests after stratifying the data by birth order, we would have arrived at the correct conclusion.

What did we do?

- We grouped subjects together based on age to alleviate the effect of the confounder.
- This is an example of **Matching.**

Matching methods

- One of the ways of dealing with confounders in the data.
- Matching methods aim to “balance” the distribution of covariates across treatment levels.
- In our example, we want to “balance” the distribution of age across mothers corresponding to different birth orders.

Steps involved:

  - STEP 1 : Define “closeness”.
  - STEP 2 : Match the subjects based on “closeness” measure.
  - STEP 3 : Estimate causal effect using the matched samples.

STEP 1: Define “closeness”

- Select the covariates you want to use to match the subjects.
- Combine the covariates to obtain a distance measure.

- ❑ In our example, subjects A and subject B will be “close” A and B fall under the **same age group** - ***Exact Matching!***

Subject A	35-39 years	Birth order = 3
Subject B	35-39 years	Birth order = 2

STEP 1: Define “closeness”

- Assume we have information about an additional confounder - estimate of daily exposure to a set of harmful chemicals in ppm.
- “Closeness” based on a continuous covariate:

$$D_{AB} \propto (C_A - C_B)^2$$

Subject A	20-24 years	124 ppm
Subject B	35-39 years	147 ppm
Subject C	<20 years	200 ppm

STEP 2: Match the subjects based on “closeness” measure.

- Exact matching based on the covariate age:
  { A , C } , { B , D, E }
- 1:1 Nearest neighbor matching based on chemical exposure
  { A , D } is a match since D is the “nearest neighbor” of A based on D_{AB}

Subject A	35-39 years	124 ppm
Subject B	<20 years	200 ppm
Subject C	35-39 years	147 ppm
Subject D	<20 years	110 ppm
Subject E	<20 years	210 ppm

Popular measures of closeness

- Assume, for the sake of convenience, a general set up:
  - Treatment (T = 0,1), Covariates (X) and Outcome (Y = 0,1)
- Two summaries of “closeness” as a function of X:

$$\text{Propensity score} = P(T = 1 \mid X) \;\; ; \quad D_{AB} = \mid e_A - e_B \mid$$

$$\text{Disease Risk score} = P(Y = 1 \mid X) \;\; ; \quad D_{AB} = \mid d_A - d_B \mid$$

- These summary measures are used in subsequent statistical analysis. E.g. used as predictors in regression models.

Disease Risk Score (DRS) vs Propensity Score (PS) matching

- Use of PS is preferred in the presence of high correlations between covariates and exposure since use of DRS exaggerates statistical significance in such situations.
- In scenarios with relatively low correlations between covariates and exposure, use of DRS, PS and traditional regression methods - all have comparable performance.
- PS and DRS based models performed better than traditional regression methods when the model is misspecified.
- DRS is useful in scenarios where PS is inappropriate or performs poorly i.e. exposures that are rare or have large number of categories.

References

- **Stuart EA**. *Matching Methods for Causal Inference: A Review and a Look Forward.* Statist. Sci. 2010;25(1). DOI: 10.1214/09-STS313
- **Arbogast PG, Ray WA.** *Performance of Disease Risk Scores, Propensity Scores, and Traditional Multivariable Outcome Regression in the Presence of Multiple Confounders.* American Journal of Epidemiology. 2011;174(5):613–620. DOI: 10.1093/aje/kwr143

- Data taken from the R package “`dsrTest`”