Estimating adjusted absolute risks in a cross-sectional register-based study with **logit** and **margins**

Anna Johansson

Karolinska Institutet, Stockholm, Sweden

Cancer Registry of Norway, Oslo

2022-10-12

Stata Conference, Oslo

Outline

- Background and motivation
  - Absolute and relative effects
  - Matched cohort design

- Using **margins** to obtain adjusted absolute risks
  - For which populations are these absolute risks valid

- *THIS IS WORK IN PROGRESS! We may have done something incorrect or non-sensible…*

Background and motivation

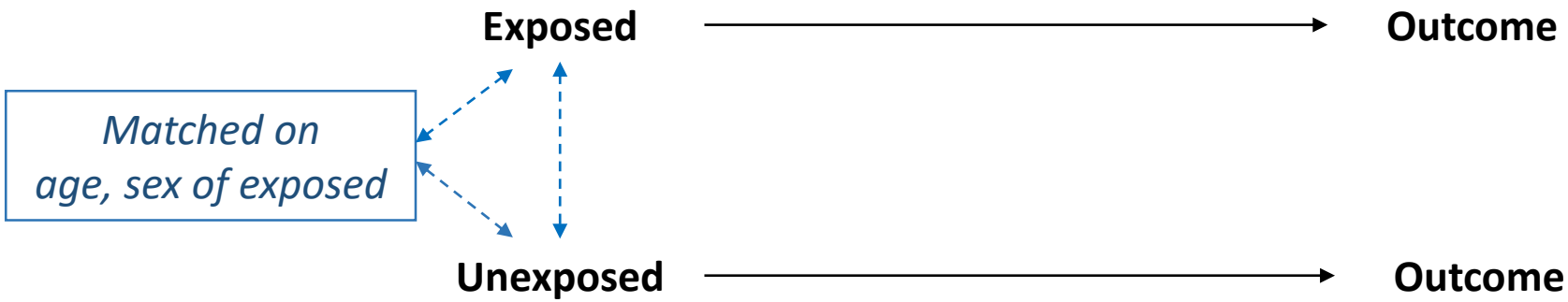
- In epidemiology we are often interested in quantifying effects of exposures not only on the relative scale (e.g. risk ratios and odds ratios), but also on the absolute scale (e.g. risks and risk differences).
- Relative measures typically gives the strength of the exposure-outcome association, whereas absolute risks give the impact on the population
  - i.e. how many persons will be affected, risk × population-at-risk

- Relative risks for an exposure can be estimated while adjusting for other factors (confounders). Such estimates are average adjusted effects, unless we allow for interactions (effect modification).
- However, absolute risks are often crude (averaged in the full population) or estimated in subgroups (for a given covariate pattern), e.g. for women, aged 45 and diagnosed 1990.

- We wanted to explore different ways to obtain **adjusted absolute risks from observational data**, using the **margins** postestimation command in Stata
- We wanted to understand how this postestimation could be used for matched data
- We wanted to understand which populations such estimates are averaged over

Matched cohort studies and motivating example

- Matched cohort design involves matching exposed persons to unexposed persons so that they are similar on matching factors
- Typically, unexposed are matched so that they have similar distribution of the matching factors (e.g. age, sex) as the exposed.
- Such matching will create a sample where exposed and unexposed are similar, and thus confounding by matching factors is eliminated.
- Crude estimates of risks and relative risks are therefore controlled for confounding automatically (as opposed to case-control matching, where the matching needs to be adjusted/controlled for in the analysis in order to produce valid confounder-adjusted estimates).
  - Crude estimates are only valid for a special population, i.e. population with same distr of age, sex as the exposed



Matched cohort studies and motivating example

- The motivating research question is from reproductive/cancer epidemiology
- **Research question**: Is breast cancer during pregnancy associated with higher risks for adverse pregnancy outcomes, e.g. caeserean section (CS)

- Dataset included 4.5 million singleton births from Swedish Medical Birth Register 1973-2017
- Assessed for CS at delivery (binary outcome CS, cross-sectional analysis)
- Births exposed to maternal cancer was matched 1:10 to healthy births (unexposed)
  - Maternal age of delivery
  - Year of delivery
  - Birth order (1, 2, 3+)

- N=203 exposed, 2030 unexposed (Total: 2233 obs)

- We estimated odds ratios using logistic regression

Matched cohort studies and motivating example

	CS (yes/no) N/N	Model 1 Cond log OR (95% CI)					
Healthy (unexp)	381/1,639	1.00					
Cancer (exp)	96/107	4.56 (3.32-6.28)					

```
clogit cesarean i.cancer, group(matchstrata)                    // Model 1
```

Matched cohort studies and motivating example

	CS (yes/no) N/N	Model 1 Cond log OR (95% CI)	Model 2 Log adj OR (95% CI)				
Healthy (unexp)	381/1,639	1.00	1.00				
Cancer (exp)	96/107	4.56 (3.32-6.28)	4.31 (3.16-5.89)				

```
clogit cesarean i.cancer, group(matchstrata)                    // Model 1

logit cesarean i.cancer agespline* yearspline* i.birthorder, or  // Model 2
```

Matched cohort studies and motivating example

	CS (yes/no) N/N	Model 1 Cond log OR (95% CI)	Model 2 Log adj OR (95% CI)				
Healthy (unexp)	381/1,639	1.00	1.00				
Cancer (exp)	96/107	4.56 (3.32-6.28)	4.31 (3.16-5.89)				
		N=1924	N=2233				

```
clogit cesarean i.cancer, group(matchstrata)                    // Model 1

logit cesarean i.cancer agespline* yearspline* i.birthorder, or  // Model 2
```

- Both model 1 and model 2 are adjusted for the matching factors. Both take matching into account.
- In model 1 the ORs are conditional on the matching factors, much similar to the estimates we get from a model adjusted for covariates (model 2), where ORs are conditional on levels of adjusted variables.
- This is inference for the whole population (both the exposed and unexposed, regardless of their distribution of age, year, birthorder).
- The causal estimand is the conditional causal effect within age-year-birthorder strata of the population.
- *Estimates differ in model 1 and 2 – why? Lost 28 strata (N=309 obs) with no events; or non-collapsibility; or not adjusting sufficiently for matching strata via age-year-birthorder covariates in model 2.*

Matched cohort studies and motivating example

	CS (yes/no) N/N	Model 1 Cond log OR (95% CI)	Model 2 Log adj OR (95% CI)				
Healthy (unexp)	381/1,639	1.00	1.00				
Cancer (exp)	96/107	4.56 (3.32-6.28)	4.31 (3.16-5.89)				
		N=1924	N=2233				

```
clogit cesarean i.cancer, group(matchstrata)                    // Model 1

logit cesarean i.cancer agespline* yearspline* i.birthorder, or  // Model 2
```

Matched cohort studies and motivating example

	CS (yes/no) N/N	Model 1 Cond log OR (95% CI)	Model 2 Log adj OR (95% CI)		Crude risk N/total		
Healthy (unexp)	381/1,639	1.00	1.00		0.1886		
Cancer (exp)	96/107	4.56 (3.32-6.28)	4.31 (3.16-5.89)		0.4729		
		N=1924	N=2233		RD: 0.2843		

```
clogit cesarean i.cancer, group(matchstrata)                    // Model 1
```

```
logit cesarean i.cancer agespline* yearspline* i.birthorder, or  // Model 2
```

- The crude risks are the risks in a special population that has the age-year-birthorder distributions as the exposed group (due to the matching).

Matched cohort studies and motivating example

	CS (yes/no) N/N	Model 1 Cond log OR (95% CI)	Model 2 Log adj OR (95% CI)		Crude risk N/total	Model 2 Risk adj (margins)	
Healthy (unexp)	381/1,639	1.00	1.00		0.1886	0.1885	
Cancer (exp)	96/107	4.56 (3.32-6.28)	4.31 (3.16-5.89)		0.4729	0.4744	
		N=1924	N=2233		RD: 0.2843	RD: 0.2859	

```
clogit cesarean i.cancer, group(matchstrata)                    // Model 1

logit cesarean i.cancer agespline* yearspline* i.birthorder, or  // Model 2
margins i.cancer
```

- Margins will predict the probability of outcome
  - If all in sample was exposed (exp=1), and then take average
  - If all in sample was unexposed (exp=0), and then take average
- Averaging over the sample, i.e. an estimated risk for a hypothetical population with same age, year and birth order distribution as the exposed (because exposed and unexposed are matched).
  - However, based on adjusted conditional effects, i.e. applicable for the whole population.
- The adjusted risks are very similar to the crude risks – as we would expect. Crude risks are only calculated from the group, while the adjusted risks are calculated for the whole sample. However, similar matching

Matched cohort studies and motivating example

	CS (yes/no) N/N	Model 1 Cond log OR (95% CI)	Model 2 Log adj OR (95% CI)	Model 3 Log unadj OR (95% CI)	Crude risk N/total	Model 2 Risk adj (margins)	Model 3 Risk unadj (margins)
Healthy (unexp)	381/1,639	1.00	1.00	1.00	0.1886	0.1885	0.1886
Cancer (exp)	96/107	4.56 (3.32-6.28)	4.31 (3.16-5.89)	3.86 (2.87-5.20)	0.4729	0.4744	0.4729
		N=1924	N=2233	N=2233	RD: 0.2843	RD: 0.2859	RD: 0.2843

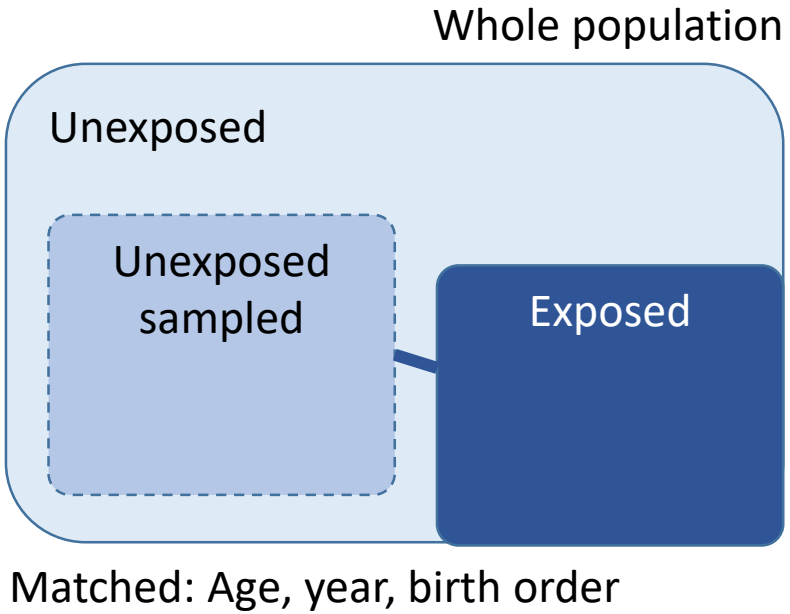
```
clogit cesarean i.cancer, group(matchstrata)                    // Model 1

logit cesarean i.cancer agespline* yearspline* i.birthorder, or  // Model 2
margins i.cancer

logit cesarean i.cancer, or                                     // Model 3
margins i.cancer
```

- For comparison, we also fit an unadjusted model (model 3) and apply margins.
- This will return the crude risks.
- I.e. the effect in the very special population of age-sex-birthorder similar to exposed.
- The causal estimand is the average (marginal) causal effect among the exposed.

Which population is the risks valid for

- If no other confounding than the matching factors, then crude estimates of risks are adjusted for confounding from matching factors by design
  - However, crude risks are valid for a very special population: A population with the same distribution of the matching factors as the exposed, i.e. a population similar to the births with maternal cancer (wrt age, year, birth order)
- Unadjusted odds ratios are only valid for the special population in the sample, as are the following postestimated risks from **margins**
- Adjusted (conditional) odds ratios are valid for the whole population, yet the postestimated adjusted risks from **margins** are only applied to the special population in the sample

- **Issue 1:** What if we have additional confounding factors? E.g. maternal education
- **Issue 2:** What if we want to also draw conclusions about the risks in the whole population?

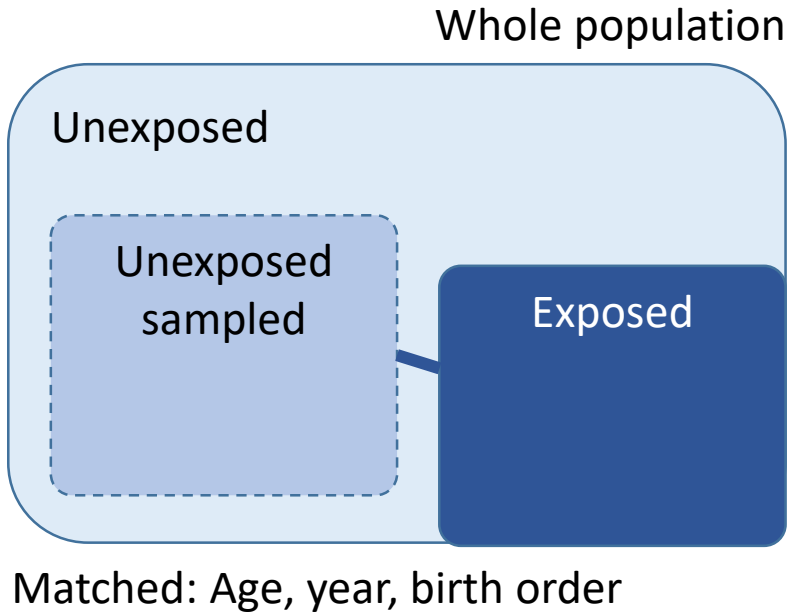


Issue 1: Additional confounding – maternal education

- The matched unexposed are similar to exposed wrt age, year and birthorder. But not education.

- The education distribution will differ between exposed and unexposed.
- The whole sample will have a distribution of education that is different from the exposed.
- The whole sample will have a distribution of education that is different from the whole population.
- The more controls (unexposed; 1:1, 1:2, 1:3, …) the more similar will the education distribution in the sample be to the education distribution in the whole population.

- Wrt education, the sample does not represent the exposed or the whole population. It is a very odd population to average over.



Issue 1: Additional confounding – maternal education

	CS (yes/no) N/N	Model 4 Log adj OR (95% CI)			
Healthy (unexp)	381/1,639	1.00			
Cancer (exp)	96/107	4.46 (3.26-6.11)			

```
logit cesarean i.cancer agespline* yearspline* i.birthorder i.educ, or  // Model 4
```

Issue 1: Additional confounding – maternal education

	CS (yes/no) N/N	Model 4 Log adj OR (95% CI)	Crude risk N/total	Model 4 Risk (margins)	Model 4 Risk (margins) cancer=1
Healthy (unexp)	381/1,639	1.00	Not valid	0.1882	0.1837
Cancer (exp)	96/107	4.46 (3.26-6.11)	Not valid	0.4799	0.4729
				RD: 0.2918	RD: 0.2892

```
logit cesarean i.cancer agespline* yearspline* i.birthorder i.educ, or  // Model 4
margins i.cancer
margins i.cancer, subpop(if cancer==1)
```

- If we do not specify a population in **margins**, then the averaged risk will be for an odd population that is a mix of the matched and whole population, wrt education
- However, we can specify that we wish to only average over the exposed population, i.e. also wrt education distribution among the exposed.
- This yield risks for a population that is similar to the exposed wrt age, year, birth order, education

Issue 2: Obtain estimates for the whole population

- In order to get back to the whole population, we can upweight the matched sample to represent the whole population
  - Need to know the sampling fractions of unexposed

- Recall, sampling of matched cohort:
  - We have included all exposed → weight=1
  - We have included a sample of unexposed → weight=1/prob of being sampled as unexposed

- In each matching stratum *i*:
  - Weight = 1/sampling fraction = $1/(n_i/N_i)$
  - E.g. If 5% unexposed sampled, then each sampled should represent 20 in the whole pop

N_i = eligible unexposed in whole pop
n_i = sampled unexposed

- Upweighting of matched sample can be included in margins via the [pw] option

```
gen wt=1/(n_stratum/N_stratum)
margins i.exposed [pw=wt]
```

Issue 2: Obtain estimates for the whole population

	CS (yes/no) N/N	Model 4 Log adj OR (95% CI)	Crude risk N/total	Model 4 Risk (margins)	Model 4 Risk (margins) cancer=1	Model 4 Risk (margins) Whole pop
Healthy (unexp)	381/1,639	1.00	Not valid	0.1882	0.1837	0.1474
Cancer (exp)	96/107	4.46 (3.26-6.10)	Not valid	0.4799	0.4729	0.4135
				RD: 0.2918	RD: 0.2892	RD: 0.2661

```
logit cesarean i.cancer agespline* yearspline* i.birthorder i.educ, or // Model 4

margins i.cancer,
margins i.cancer, subpop(if cancer==1)
margins i.cancer [pw=wt]
```

- The ”whole pop” estimates are risks for a population that have the same distribution of age, year, birth order and education as in the **whole population**
- This may be of interest if we e.g. want to draw conclusions on the impact of cancer on the whole population, given how it is distributed wrt age, year, birth order and education.

In summary

- We can (we think!) use **margins** to obtain absolute risks from matched cohort data via postestimation
- Important to understand 1) for which population the odds ratios are valid, and 2) which population we are marginalising over
- We have shown that **margins** can be used to estimate risks in a population similar to the exposed population
- If sampling fractions of unexposed are known, we can obtain estimates of risk in the whole population via upweighting using the **[pw]** option

- We have shown this for cross-sectionally analysed data
- These results should also carry over to standard cohort data (time-to-event)
  - An additional aspect; matching is at start of follow-up; hence, matching in a cohort study with follow-up does not account for confounding during follow-up, or for informative censoring or competing risks

- We believe that it is important to not only report relative risks but also absolute risks in many epidemiological research studies, also in prescence of confounding

Acknowledgements

- Frida Lundberg for the **example and analysis**
- Esa Läärä, Arvid Sjölander for **matched cohort methods**
- Paul Lambert for **margins**