

# Regression models

Arvid Sjölander

Department of Medical Epidemiology and Biostatistics  
Karolinska Institutet

A short course on concepts and methods in Causal  
Inference

## So far

- We have considered two methods of confounding adjustment
  - stratification
  - stratification-based standardization
- Under conditional exchangeability, these methods produce
  - conditional (i.e. subpopulation) causal effects, and
  - marginal (i.e. population) causal effects, respectively

## So far, cont'd

- We have considered ideal conditions
  - infinite samples
  - binary variables
- Under these ideal conditions, stratification and stratification-based standardization are ideal methods
  - require no modeling assumptions

## In reality

- In real studies, we often have
  - small/moderate sample sizes
  - continuous and categorical variables
- Under these realistic conditions, stratification and stratification-based standardization may not be feasible
  - few subjects in each stratum
  - unstable estimates (e.g. wide confidence intervals, large p-values)

## Outcome regression models

- To stabilize estimates, statistical models are commonly used
  - e.g. linear regression, logistic regression
- Most common are regression models for the outcome
- Outcome regression models can be viewed as a parametric analog to stratification
  - They estimate conditional (subpopulation) causal effect
- Outcome regression models can also be used to estimate marginal (population) effects, in analog to stratification-based standardization

## Exposure regression model

- Standardization can also be carried out using regression models for the exposure
  - A parametric weighting-based method
- Modeling the exposure is particularly attractive when
  - the mechanisms that bring about the exposure are well understood
  - the exposure is time varying
    - more later

# Outline

## Point exposures

- Conditional effects

- Marginal effects

## Time varying exposures

- Conditional effects

- Marginal effects





# Outline

## Point exposures

Conditional effects

Marginal effects

## Time varying exposures

Conditional effects

Marginal effects

## Motivating example

- Suppose we carry out an observational study to estimate the causal effect of AZT on infection risk for AIDS patients
- 300 subjects enrolled
- Baseline measures:
  - CD4 count ( $L$ ; counts/ $\mu$ l)
  - AZT level ( $A$ ; '0' for 'untreated', '1' for 'treated')
- At end of follow up:
  - infection status ( $Y$ ; '0' for infection, '1' for no infection)

## Data

```
> aids[1:10,]  
      Y A    L  
1    1 1 304  
2    0 1 295  
3    0 0 416  
4    0 1 297  
5    0 0 349  
6    0 0 328  
7    1 1 376  
8    0 0 372  
9    1 1 257  
10   1 1 326
```

## Unadjusted analysis in R

```
> summary(glm(formula=Y~A, family=binomial, data=aids))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-0.4490	0.1652	-2.717	0.006589	**
A	0.8945	0.2368	3.777	0.000159	***

- *Interpretation?*

## The role of CD4 count

- Arguable, CD4 count is an important confounder that we need to adjust for
- But in the data, very few subjects have the same CD4 count
  - Stratification on CD4 count is not feasible

## Statistical assumptions

- Suppose that we are willing to assume that the conditional odds ratio

$$OR(L) = \frac{\Pr(Y = 1 | A = 1, L)}{\Pr(Y = 0 | A = 1, L)} / \frac{\Pr(Y = 1 | A = 0, L)}{\Pr(Y = 0 | A = 0, L)}$$

is constant across levels of  $L$

- We can then ‘borrow information’ across levels of  $L$  to obtain a more stable estimate of  $OR(L)$

## Model formulation

- Statistical assumptions are often formulated through regression models, e.g.

$$\text{logit}\{\Pr(Y = 1|A, L)\} = \alpha + \beta A + \gamma L$$

- *What is the interpretation of  $\alpha$ ,  $\beta$ , and  $\gamma$  in terms of probabilities?*

## Solution

$$\text{logit}\{\Pr(Y = 1|A, L)\} = \alpha + \beta A + \gamma L$$

$$\alpha = \text{logit}\{\Pr(Y = 1|A = 0, L = 0)\}$$

$$\begin{aligned}\beta &= \text{logit}\{\Pr(Y = 1|A = 1, L)\} - \text{logit}\{\Pr(Y = 1|A = 0, L)\} \\ &= \log\{OR(L)\}\end{aligned}$$

$$\gamma = \text{logit}\{\Pr(Y = 1|A, L + 1)\} - \text{logit}\{\Pr(Y = 1|A, L)\}$$



## Underlying assumptions

$$\text{logit}\{\Pr(Y = 1|A, L)\} = \alpha + \beta A + \gamma L$$

- *Which assumptions are encoded in the model?*

## Solution

$$\text{logit}\{\Pr(Y = 1|A, L)\} = \alpha + \beta A + \gamma L$$

- The increase in log odds of being infection free, comparing AZT with no AZT, at a fixed CD4 count  $L$ , does not depend on  $L$
- The increase in log odds for being infection free, comparing CD4 count  $L + 1$  with CD4 count  $L$ , at a given AZT level  $A$ , does not depend on  $A$  or  $L$

## Remember

- **All models are wrong**
  - But if the model is approximately correct, then our conclusions are approximately valid
- Assumptions that we make should be justified by
  - subjects matter knowledge
  - data (e.g. diagnostic tests)

## Fitting the model in $\mathbb{R}$

```
> summary(glm(formula=Y~A+L, family=binomial, data=aids))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-3.135377	0.911839	-3.439	0.000585	***
A	1.259045	0.273873	4.597	4.28e-06	***
L	0.007229	0.002396	3.017	0.002554	**

- Interpretation?
- Why is the estimated adjusted log odds ratio bigger than the estimated unadjusted log odds ratio (=0.89)?

# Outline

## Point exposures

Conditional effects

Marginal effects

## Time varying exposures

Conditional effects

Marginal effects

# Standardization

- The standardization formula

$$\Pr(Y_a = 1) = \sum_L \Pr(Y = 1|A = a, L)\Pr(L)$$

holds under conditional exchangeability, given  $L$

- Once we have fitted the model for  $\Pr(Y = 1|A, L)$ , we can estimate the right handside, for  $a = 0$  and  $a = 1$  separately, by
  - replacing the factual exposure level with  $a$ , for each individual
  - estimating  $\Pr(Y = 1|A = a, L)$  for each individual (i.e. for each observed value of  $L$ )
  - averaging these estimates over all individuals (i.e. over all observed values of  $L$ )

## In R

```
> aids0=aids
> aids0["A"]=rep(0,nrow(aids0))
> f=glm(formula=Y~A+L,family=binomial,data=aids)
> p0=predict(object=f,newdata=aids0,type="respons")
> mean(p0)
[1] 0.3533108

> aids1=aids
> aids1["A"]=rep(1,nrow(aids1))
> f=glm(formula=Y~A+L,family=binomial,data=aids)
> p1=predict(object=f,newdata=aids1,type="respons")
> mean(p1)
[1] 0.6471369
```

## Standardization, cont'd

$$\text{logit}\{\Pr(Y = 1|A, L)\} = \alpha + \beta A + \gamma L$$

$$\hat{\Pr}(Y_1 = 1) = 0.65 \quad \hat{\Pr}(Y_0 = 1) = 0.35$$

- We can use the estimates of  $\Pr(Y_1 = 1)$  and  $\Pr(Y_0 = 1)$  to construct an estimate of the causal log odds ratio

$$\begin{aligned} \log\{\hat{C\hat{O}R}\} &= \log\left\{\frac{\hat{\Pr}(Y_1 = 1)}{1 - \hat{\Pr}(Y_1 = 1)} / \frac{\hat{\Pr}(Y_0 = 1)}{1 - \hat{\Pr}(Y_0 = 1)}\right\} \\ &= \log\left\{\frac{0.65}{1 - 0.65} / \frac{0.35}{1 - 0.35}\right\} = 1.24 \end{aligned}$$

- Interpretation?*



## Standard errors

$$\log\{\hat{C\hat{O}R}\} = 1.24$$

- As for all estimates, we wish to quantify the uncertainty due to sampling variability
- Standard errors can be obtained with some additional programming
  - Sandwich formula
  - Bootstrap

## Outcome models vs exposure models

- So far, we have considered models for the outcome, e.g

$$\text{logit}\{\Pr(Y = 1|A, L)\} = \alpha + \beta A + \gamma L$$

- Sometimes, it may be more natural to use a model for the exposure, e.g.

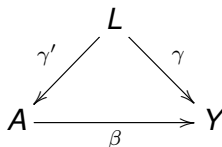
$$\text{logit}\{\Pr(A = 1|L)\} = \alpha + \gamma' L$$

- For instance, we may know more about the guidelines for AZT administration, than we know about the biological mechanisms underlying infection
- Exposure models can also be used for standardization
  - With an estimation technique called 'Inverse Probability Weighting' (IPW)

# In a DAG

$$\text{logit}\{\Pr(Y = 1|A, L)\} = \alpha + \beta A + \gamma L$$

$$\text{logit}\{\Pr(A = 1|L)\} = \alpha + \gamma' L$$



## IPW - the recipe

- **Step 1:** Fit a regression model for the exposure
- **Step 2:** For each subject, use the fitted exposure model to estimate a subject-specific weight

$$W = 1/\hat{\text{Pr}}(A|L)$$

- For instance, suppose that  $\hat{\Pr}(A = 1 | L = 305) = 0.2$
- Subjects with  $(A = 1, L = 305)$  then get the weight  $1/0.2 = 5$ , and subjects with  $(A = 0, L = 305)$  get the weight  $1/(1 - 0.2) = 1.25$
- **Step 3:** Fit an unadjusted **weighted** regression model for the outcome, e.g.

$$\text{logit}\{\Pr(Y = 1|A)\} = \alpha + \beta A$$

- Due to the weighting, the coefficient for the exposure is standardized
  - e.g.  $\beta = \log\{\hat{C\hat{O}R}\}$  (under conditional exchangeability, given  $L$ )

# In R

```
> library(ipw)
> t=ipwpoint(exposure=A,family="binomial",
             link="logit",denominator=~L,data=aids)
> w=t$ipw.weights
> summary(glm(formula=Y~A,family=binomial,
              data=aids,weights=w))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-0.6496	0.1206	-5.386	7.19e-08	***
A	1.2143	0.1714	7.084	1.40e-12	***

- *Why is the exposure model estimate different from the outcome model estimate (=1.24)?*

## Standard errors

- When using weights, Stata gives by default robust (Sandwich) standard errors
  - If the weights are estimated (in opposite to known *a priori*) these standard errors are incorrect
- Correct standard errors can be obtained with some additional programming
  - Sandwich formula
  - Bootstrap

## Why IPW works

- The unadjusted regression model

$$\text{logit}\{\Pr(Y = 1|A)\} = \alpha + \beta A$$

produces the marginal (population) causal effect of  $A$  on  $Y$   
if exposed and unexposed are exchangeable

- This is in fact true in the weighted population
  - Assuming that conditional exchangeability, given  $L$ , holds in the unweighted population
  - The weighting procedure eliminates all confounding by  $L$

# Outline

## Point exposures

Conditional effects

Marginal effects

## Time varying exposures

Conditional effects

Marginal effects



# Outline

## Point exposures

Conditional effects

Marginal effects

## Time varying exposures

Conditional effects

Marginal effects

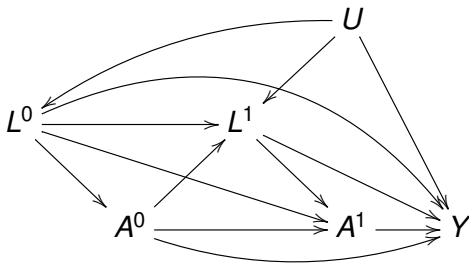
## Motivating example

- Suppose we carry out an observational study to estimate the causal effect of AZT on infection risk for AIDS patients
- 300 subjects enrolled
- At  $t = 0$  (baseline) and  $t = 1$  we measure:
  - CD4 count ( $L^t$ ; counts/ $\mu$ l)
  - AZT level ( $A^t$ ; '0' for 'untreated at  $t$ ', '1' for 'treated at  $t = 1$ ')
- At end of follow up:
  - infection status ( $Y$ ; '0' for infection, '1' for no infection)

## Data

```
> aids[1:10,]  
   id t A   L Y  sA  ssA  sL  
1    1 0 1 342 0   1    0 342  
2    1 1 0 328 0   1    1 670  
3    2 0 0 442 0   0    0 442  
4    2 1 1 322 0   1    0 764  
5    3 0 0 367 0   0    0 367  
6    3 1 1 217 0   1    0 584  
7    4 0 0 393 0   0    0 393  
8    4 1 1 435 0   1    0 828  
9    5 0 0 346 0   0    0 346  
10   5 1 1 277 0   1    0 623
```

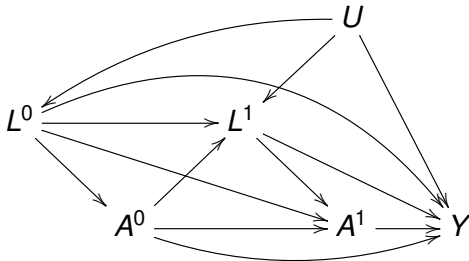
## DAG



- We assume sequential exchangeability, but make no further assumptions about the underlying causal structure



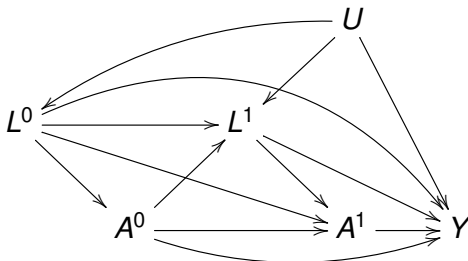
## Problem with the unadjusted analysis



- CD4 count is a confounder
  - Without adjusting for  $(L^0, L^1)$ , an observed association cannot be given a causal interpretation



## Problem with the standard adjusted analysis



- By adjusting for  $L^1$  we open the path non-causal  $A^0 \rightarrow L^1 \leftarrow U \rightarrow Y$ 
  - The association between  $A^0$  and  $Y$  cannot be given a causal interpretation
- In general (i.e. with  $T$  time points), adjusting for  $(L^0, \dots, L^T)$  open non-causal paths to  $(A^0, \dots, A^{T-1})$



## Sequential adjustment

- One regression model for each time point  $t$ , adjusted for the whole observed past up to  $t$ , e.g.

$$\text{logit}\{\Pr(Y = 1 | L^0, A^0, \dots, L^t, A^t)\} = \alpha + \underbrace{\psi^t A^t}_{\text{causal effect}} + \underbrace{\beta^t s A^{t-1} + \gamma^t s L^t}_{\text{confounding adjustment}}$$

# In R

```
> summary(glm(formula=Y~A+L,family=binomial,  
              data=aids,subset=(t==0)))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-2.9830428	1.1184606	-2.667	0.00765	**
A	1.1034001	0.4159269	2.653	0.00798	**
L	0.0009101	0.0029366	0.310	0.75662	

```
> summary(glm(formula=Y~A+ssA+sL,family=binomial,  
              data=aids,subset=(t==1)))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-3.0383520	1.6352535	-1.858	0.06316	.
A	-0.0975989	0.4378070	-0.223	0.82359	
ssA	1.0847198	0.4200912	2.582	0.00982	**
sL	0.0006448	0.0021127	0.305	0.76020	

- *Interpretation?*



## Sequential standardization

- Sequential adjustment gives the effect of  $A^0, A^1, \dots, A^T$ , separately, for different subpopulations
- We may want to estimate the joint effect of  $(A^0, \dots, A^T)$  for the whole population
- This can be done with sequential standardization
  - The G-formula

## Outcome models vs exposure models

- In practice, we must use regression models to carry out sequential standardization
  - Just like for ordinary (e.g. point exposure) standardization
- And just like for ordinary standardization, we can use either outcome models or exposure models
  - However, for time varying exposures, there are some serious disadvantages of using outcome models

- \_\_\_\_\_

$$\Pr(Y_{a^0 \dots a^T} = 1) = \left\{ \sum_{L^0 \dots L^T} \Pr(Y = 1 | L^0, A^0 = a^0, \dots, L^T, A^T = a^T) \right. \\ \left. \prod_{t=0}^T \Pr(L^t | L^0, A^0 = a^0, \dots, L^{t-1}, A^{t-1} = a^{t-1}) \right\}$$

- We need one model for the outcome, given the observed past

$$\Pr(Y = 1|L^0, A^0, \dots, L^T, A^T)$$

and one model for the covariate, at each time point, given the observed past

$$\Pr(L^t | L^0, A^0, \dots, L^{t-1}, A^{t-1})$$

## Example

$$\text{logit}\{\Pr(Y = 1|L^0, A^0, \dots, L^T, A^T)\} = \alpha + \beta \mathbf{s}A^T + \gamma \mathbf{s}L^T$$

$$L^t | (L^0, A^0, \dots, L^{t-1}, A^{t-1}) \sim N\left(\alpha' + \beta' \mathbf{s}A^{t-1} + \gamma' \mathbf{s}L^{t-1}, \sigma^2\right)$$

- Fit the models, plug the estimates into

$$\Pr(Y_{a^0 \dots a^T} = 1) = \left[ \sum_{L^0 \dots L^T} \frac{\overbrace{e^{\alpha + \beta \mathbf{s}A^T + \gamma \mathbf{s}L^T}}^{\Pr(Y=1|L^0, A^0=a^0, \dots, L^T, A^T=a^T)}}{1 + e^{\alpha + \beta \mathbf{s}A^T + \gamma \mathbf{s}L^T}} \prod_{t=0}^T \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-\{L^t - (\alpha' + \beta' \mathbf{s}A^{t-1} + \gamma' \mathbf{s}L^{t-1})\}^2}{2\sigma^2}}}_{\Pr(L^t|L^0, A^0=a^0, \dots, L^{t-1}, A^{t-1}=a^{t-1})} \right]$$

- This approach is problematic for two reasons

## Computational difficulties

$$\Pr(Y_{a^0 \dots a^T} = 1) = \left[ \sum_{L^0 \dots L^T} \frac{e^{\alpha + \beta s a^T + \gamma s L^T}}{1 + e^{\alpha + \beta s a^T + \gamma s L^T}} \prod_{t=0}^T \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-\{L^t - (\alpha' + \beta' s a^{t-1} + \gamma' s L^{t-1})\}^2}{2\sigma^2}}}_{\Pr(L^t | L^0, A^0 = a^0, \dots, L^{t-1}, A^{t-1} = a^{t-1})} \right]$$

- To calculate the right handside we need to solve a  $T + 1$  dimensional integral
  - impossible analytically
  - awkward numerically



## Interpretational difficulties

$$\Pr(Y_{a^0 \dots a^T} = 1) = \left[ \sum_{L^0 \dots L^T} \frac{\overbrace{\Pr(Y=1|L^0, A^0=a^0, \dots, L^T, A^T=a^T)}^{e^{\alpha+\beta sa^T + \gamma sL^T}}}{1 + e^{\alpha+\beta sa^T + \gamma sL^T}} \prod_{t=0}^T \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-\{L^t - (\alpha' + \beta' sa^{t-1} + \gamma' sL^{t-1})\}^2}{2\sigma^2}}}_{\Pr(L^t|L^0, A^0=a^0, \dots, L^{t-1}, A^{t-1}=a^{t-1})} \right]$$

- There is no simple interpretation of the right handside parameters in terms of the left handside
  - To calculate the left handside we need to calculate the right handside for each of the  $2^{T+1}$  combinations of  $(a_0, a_1)$
  - Feasible when  $T$  is small (e.g. = 1), but awkward for larger numbers



## Fitting of MSMs

- MSMs are fitted by using regression models for the exposure, together with IPW
- Under sequential exchangeability, the following recipe produces consistent estimates of the (causal) model parameters

## The recipe

- **Step 1:** Fit a regression model for  $A^t$ , given the observed past up to  $t$ , e.g.

$$\text{logit}\{\Pr(A^t = 1|L^0, A^0, \dots, L^{t-1}, A^{t-1}, L^t)\} = \alpha' + \beta' \mathbf{s}A^{t-1} + \gamma' \mathbf{s}L^t$$

- **Step 2:** For each subject, use the fitted exposure model to estimate a subject-specific weight

$$W = 1/\prod_{t=0}^T \hat{\text{Pr}}(A^t = 1|L^0, A^0, ..., L^{t-1}, A^{t-1}, L^t)$$

- **Step 3:** Fit the MSM using **weighted** regression, as if it would have been a model for  $\Pr(Y = 1|A^0, \dots, A^T)$ , e.g.

$$\text{logit}\{\Pr(Y = 1|A^0, \dots, A^T)\} = \alpha + \beta \mathbf{s}A^T$$

In R

```
> library(ipw)
> t=ipwtm(exposure=A,family="binomial",
  link="logit",denominator=~ssA+sL,id=id,
  tstart=0,timevar=t,type="all",data=aids)
> w=t$ipw.weights
> summary(glm(formula=Y~sA,family=binomial,
  data=aids,weights=w,subset=(t==1)))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-2.4924	0.1820	-13.697	< 2e-16	***
sA	0.4680	0.1342	3.486	0.00049	***

## Standard errors

- When using weights, Stata gives by default robust (Sandwich) standard errors
  - If the weights are estimated (in opposite to known *a priori*) these standard errors are incorrect
- Correct standard errors can be obtained with some additional programming
  - Sandwich formula
  - Bootstrap

## Why IPW works

- The unadjusted regression model

$$\text{logit}\{\Pr(Y = 1|A^0, \dots, A^T)\} = \alpha + \beta \mathbf{s}A^T$$

produces the marginal (population) causal effect of  $(A^0, \dots, A^T)$  on  $Y$  if exposed and unexposed are exchangeable

- This is in fact true in the weighted population
  - Assuming that sequential exchangeability holds in the unweighted population
  - The weighting procedure eliminates all confounding by  $(L^0, \dots, L^T)$





## More complex situations

- In real studies
  - outcomes are often measured repeatedly
  - the survival time (often censored) is often the main target of analysis
- MSMs can be used for repeated outcomes and survival outcomes
  - analysis and interpretation get more complex
  - beyond the scope of this course

## Doubly robust estimation

- To standardize, we can use either an
  - outcome model, or an
  - exposure model
- It is possible to combine both models into a **doubly robust** estimator
  - Consistent if either model is correct, not necessarily both
  - Two chances of valid inference instead of only one
  - Beyond the scope of this course

## Summary

- ‘Ordinary’ regression models are models for the outcome
- By default, these produce conditional (subpopulation) effects
- Marginal (population) effects can be obtained by standardization
  - either with an outcome model, or with an exposure model