Doubly robust estimation in Generalized Linear Models with Stata

Arvid Sjölander
Nicola Orsini

Background

- Common aim in epi: estimate the “effect” of a particular exposure on a particular outcome, adjusted for covariates

- Common statistical method: Generalized Linear Model (GLM) for the outcome

  → Estimates through Maximum Likelihood (ML)

Background, cont'd

- Sometimes, the outcome model may be difficult to well specify
  → High dimensional covariates
  → Weak knowledge about outcome mechanism

- The exposure effect can also be estimated by using regression model for the exposure
  → Convenient when the exposure mechanisms are well understood
  → Estimates through G-estimation

Background, cont'd

- Often, no clear preference for either strategy
- Both models can be combined to construct a Doubly Roubust (DR) estimator for the exposure effect
  - → Unbiased if either model is correct, not neccesarily both
  - → Two chances in one to obtain unbiased estimates
- Theory well developed, but little is implemented.

Outline

- Brief review of Generalized Linear Models (GLMs)
- A brief introduction to DR theory for GLMs
- Our Stata implementation: **drglm**

A simple GLM

- A = exposure, Y = outcome, L = covariates

$$g\{E(Y \mid A, L; \beta, \gamma)\} = \beta A + \gamma^T L$$

- g() is a suitable link function, e.g.
  - Continuous Y – identity link
  - ”Count” (e.g. 0,1,2,3,…) Y – log link
  - Binary (0/1) Y – logit link

- The model is completed with a suitable distribution for Y, e.g.
  - Continuous Y – Normal
  - ”Count” Y – Poisson
  - Binary Y – Bernoulli

Interpretation

- $$g\{E(Y \mid A, L; \beta, \gamma)\} = \beta A + \gamma^T L$$
- β is the conditional (given L) mean difference in Y, when A is increased with 1 unit:

$$\beta = g\{E(Y \mid A = a+1, L)\} - g\{E(Y \mid A = a, L)\}$$

→ Losely: the "effect" of A on Y, adjusted for L

- The link function determines the scale
  - Identity link – mean difference
  - Log link – (log) mean ratio
  - Logit link – (log) odds ratio

A closer look at the model

$$g\{E(Y \mid A, L; \beta, \gamma)\} = \beta A + \gamma^T L$$

- The main model

$$g\{E(Y \mid A, L)\} - g\{E(Y \mid A = 0, L)\} = \beta A$$

quantifies the effect of A on Y

- The outcome nuisance model

$$g\{E(Y \mid A = 0, L; \gamma)\} = \gamma^T L$$

is primarily included to adjust for L

Estimation

- The ML estimator of (β, γ) (`glm` in Stata) can be expressed as the solution to the score equation system

$$\sum_{i=1}^{n} \begin{pmatrix} A_i \\ L_i \end{pmatrix} \left\{ Y_i - E(Y \mid A_i, L_i; \beta, \gamma) \right\} = 0$$

Impact of model misspecification

- Unbiasedness estimate of β even if distributional assumption for Y is incorrect (but efficiency loss).

- Biased estimate of β if outcome nuisance model $g\{E(Y|A=0,L)\} = \gamma^T L$ is incorrect.

- Sometimes, the outcome model may be difficult to well specify
  → High dimensional covariates
  → Weak knowledge about outcome mechanism

- Sometimes easier to specify a model for the exposure
  → E.g. when the exposure mechanism is well understood

G-estimation

- We can estimate β with a nuisance model for the exposure.

- The main idea (for identity link):

  → If β was known, we could construct residuals $Y_i - \beta A_i$
  → $Y_i - \beta A_i$ is an unbiased prediction of $E(Y|A = 0, L_i)$
  → Given L_i, $E(Y|A = 0, L_i)$ is a constant, uncorrelated with A_i

- This suggests the following estimation strategy:

  → find the value of β for which $Y_i - \beta A_i$ becomes conditionally uncorrelated with A_i, given L_i, in the sample.

In terms of estimating equations

- Find the value of β which solves the equation

$$\sum_{i=1}^{n} \{A_i - E(A \mid L_i)\}(Y_i - \beta A_i) = 0$$

- We need a model for $E(A|L)$, e.g.

$$h\{E(A \mid L; \alpha)\} = \alpha^T L$$

which can be fitted using `glm` in Stata

The two equations (identity outcome link)

- ML score equation:

$$\sum_{i=1}^{n} A_i \left\{ Y_i - \beta A_i - E\left(Y \mid A = 0, L_i; \gamma\right) \right\} = 0$$

- G-estimating equation:

$$\sum_{i=1}^{n} \left\{ A_i - E(A \mid L_i; \alpha) \right\} \left(Y_i - \beta A_i \right) = 0$$

- Combined:

$$\sum_{i=1}^{n} \left\{ A_i - E(A \mid L_i; \alpha) \right\} \left\{ Y_i - \beta A_i - E\left(Y \mid A = 0, L_i; \gamma\right) \right\} = 0$$

Doubly robust property

- It can be shown that the combined estimating equation

$$\sum_{i=1}^{n} \{A_i - E(A \mid L_i; \alpha)\}\{Y_i - \beta A_i - E(Y \mid A = 0, L_i; \gamma)\} = 0$$

produces an estimate of β that is unbiased if either

→ outcome nuisance model correct, or

→ exposure nuisance model correct

- It is doubly robust!

→ Two chances in one to obtain valid inference

Standard errors

- Standard errors must take estimation of all three models into account.

  → Sandwich estimator, or

  → Bootstrap

Other link functions

- The DR estimating equation can be adapted for log links with a minor modification.

- For logit links, special techniques are required (Tchetgen Tchetgen et al, 2010).

Simulation

- 1000 samples of 100 observations each, from

$$L = (L_1, L_2) \qquad L_1 \perp L_2 \qquad L_1, L_2 \sim N(0,1)$$

$$A \mid L \sim N\left(\underbrace{\alpha_0 + \alpha_1 L_1 + \alpha_2 L_2 + \alpha_{12} L_1 L_2}_{\text{Exposure nuisance model}}, 1 \right)$$

$$\alpha_0 = \alpha_1 = \alpha_2 = 1, \quad \alpha_{12} = -1.5$$

$$Y \mid A, L \sim N\left(\underbrace{\beta_0 A + \beta_1 A L_1}_{\text{Main model}} + \underbrace{\gamma_0 + \gamma_1 L_1 + \gamma_2 L_2 + \gamma_{12} L_1 L_2}_{\text{Outcome nuisance model}}, 1 \right)$$

$$\beta_0 = 1.5, \quad \beta_1 = 1, \qquad \gamma_0 = \gamma_1 = \gamma_2 = -1, \quad \gamma_{12} = 1.5$$

Stata analysis

```
drglm Y A, main(L1) outcome(L1 L2 L1L2)
```

- main model:

$$E(Y \mid A, L) - E(Y \mid A = 0, L) = \beta_0 A + \beta_1 A L_1$$

- outcome nuisance model:

$$E(Y \mid A = 0, L; \gamma) = \gamma_0 + \gamma_1 L_1 + \gamma_2 L_2 + \gamma_{12} L_1 L_2$$

- exposure nuisance model:

$$E(A \mid L; \alpha) = \alpha_0 + \alpha_1 L_1 + \alpha_2 L_2 + \alpha_{12} L_1 L_2$$

→ Default: exposure model = outcome model

Output

$$E(Y \mid A, L) - E(Y \mid A = 0, L) = \beta_0 A + \beta_1 A L_1$$

$$\beta_0 = 1.5, \quad \beta_1 = 1$$

Double Robust Estimator — Number of obs = **100**
Outcome link = **identity**
Exposure link = **identity**

Y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
main						
AL1	**1.046265**	**.1733402**	**6.04**	**0.000**	**.7065244**	**1.386005**
A	**1.515002**	**.1160573**	**13.05**	**0.000**	**1.287534**	**1.74247**

Simulation results

- Analyses:
  - → I: both E(Y|A=0,L) and E(A|L) correct.
  - → II: E(A|L) misspecified - α_{12} omitted
  - → III: E(Y|A=0,L) misspecified - γ_{12} omitted

- Mean estimates:

	β_0= 1.5		
	I	II	III
ML	1.50	1.50	0.82
G	1.50	0.84	1.50
DR	1.50	1.50	1.50

	β_1 = 1		
	I	II	III
ML	1.00	1.00	1.01
G	1.00	1.07	1.00
DR	1.00	1.00	1.00

Additional features/options

- The **drglm** command supports the ”xi-notation”
- The most common link functions
  - → Identity link
  - → Log link
  - → Logit link
- Sandwich standard errors or bootstrap standard errors
- Works with most common post-estimation commands

Testing for no association

$$E(Y \mid A, L) - E(Y \mid A = 0, L) = \beta_0 A + \beta_1 A L_1$$

$$H_0 : \beta_0 = \beta_1 = 0$$

```
. testparm A AL1

 ( 1)  [main]AL1 = 0
 ( 2)  [main]A = 0

           chi2(  2) =  261.17
         Prob > chi2 =    0.0000
```

Summary

- Doubly robust estimators are highly attractive, since they give the resarcher two chances of obtaining unbiased estimates.

- With the new Stata command **`drglm`**, DR estimation in GLMs is easy and convenient

References

- Robins JM. (2000).Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science* 1999, pp. 6-10.

- Tchetgen Tchetgen E, Robins J, Rotnitzky A. (2010). On doubly robust estimation in a semiparametric odds ratio model. Biometrika 97, 171-180.

- Orsini N, Sjölander A. (2011). Doubly robust estimation in GLMs using Stata. *In preparation for Stata Journal.*