Section 26

Lecture 9

Plan

- Estimation with a point treatment.
  - Standardisation
  - Propensity methods
- Marginal structural models
- Uncertainty quantification

Some fundamentals (as a reminder for you)

Slides 249-257 describe some fundamentals about statistical modelling. All the details will not be covered in the lectures. The idea is that you might find this background information useful.
Slides labelled with an asterisk (*) are, in particular, additional details that we will not study in depth in class.

Reminder: Maximum Likelihood Estimation (MLE)

Consider a vector $\theta = [\theta_1, \theta_2, \ldots, \theta_k]^\mathsf{T}$ of parameters that indexes the distribution $\{f(\cdot\,;\theta) \mid \theta \in \Theta\}$, where Θ is a parameter space.
We evaluate the observed data sample $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$, which gives us the likelihood,

$$L_n(\theta) = L_n(\theta; \boldsymbol{Y}) = f_n(\boldsymbol{Y}; \theta),$$

where $f_n(\boldsymbol{Y}; \theta)$ is a product of n density functions evaluated at $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$.
MLE maximises the likelihood, i.e.

$$\theta = \arg\max_{\theta \in \Theta} L_n(\theta\,;\boldsymbol{Y}).$$

The logarithm is a monotone function, and thus it is more convenient to maximise the log-likelihood: $\ell(\theta\,;\boldsymbol{Y}) = \log L_n(\theta\,;\boldsymbol{Y})$. If $\ell(\theta\,;\boldsymbol{Y})$ is differentiable in θ, we solve $M(\boldsymbol{Y};\theta) = \frac{\delta \ell(\theta\,;\boldsymbol{Y})}{\delta\theta}$, , i.e. the score equations (also called likelihood equations)

$$p_1 \equiv \frac{\partial \ell}{\partial \theta_1} = 0, \quad \frac{\partial \ell}{\partial \theta_2} = 0, \quad \ldots, \quad \frac{\partial \ell}{\partial \theta_k} = 0.$$

.

*Reminder: Likelihood continues

We need local concavity. Thus, the Hessian matrix

$$\mathbf{H}\left(\widehat{\theta}\right) = \begin{bmatrix} \left.\frac{\partial^2 \ell}{\partial \theta_1^2}\right|_{\theta=\widehat{\theta}} & \left.\frac{\partial^2 \ell}{\partial \theta_1 \, \partial \theta_2}\right|_{\theta=\widehat{\theta}} & \cdots & \left.\frac{\partial^2 \ell}{\partial \theta_1 \, \partial \theta_k}\right|_{\theta=\widehat{\theta}} \\ \left.\frac{\partial^2 \ell}{\partial \theta_2 \, \partial \theta_1}\right|_{\theta=\widehat{\theta}} & \left.\frac{\partial^2 \ell}{\partial \theta_2^2}\right|_{\theta=\widehat{\theta}} & \cdots & \left.\frac{\partial^2 \ell}{\partial \theta_2 \, \partial \theta_k}\right|_{\theta=\widehat{\theta}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial^2 \ell}{\partial \theta_k \, \partial \theta_1}\right|_{\theta=\widehat{\theta}} & \left.\frac{\partial^2 \ell}{\partial \theta_k \, \partial \theta_2}\right|_{\theta=\widehat{\theta}} & \cdots & \left.\frac{\partial^2 \ell}{\partial \theta_k^2}\right|_{\theta=\widehat{\theta}} \end{bmatrix},$$

is negative semi-definite at $\widehat{\theta}$. The Fisher information matrix is defined as $\mathcal{I}(\theta) = \mathbb{E}\left[\mathbf{H}\left(\widehat{\theta}\right)\right]$.

Logistic regression

Suppose $Y \in \{0, 1\}$. Define $\beta = [\beta_1, \beta_2, \ldots, \beta_k]^\mathsf{T}$ as a vector of k parameter and consider a k dimensional covariate $\mathbf{X}$. Then the logistic model is defined as

$$\text{logit}(\mathbb{E}[Y_i \mid \mathbf{X}_i]) = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{\beta}^T \mathbf{X}_i,$$

or, equivalently, we can write that that Y follows a Bernoulli distribution,

$$P(Y_i = y \mid \mathbf{X}_i) = {p_i}^y (1-p_i)^{1-y} = \left(\frac{e^{\boldsymbol{\beta}^T \mathbf{X}_i}}{1+e^{\boldsymbol{\beta}^T \mathbf{X}_i}}\right)^y \left(1 - \frac{e^{\boldsymbol{\beta}^T \mathbf{X}_i}}{1+e^{\boldsymbol{\beta}^T \mathbf{X}_i}}\right)^{1-y}$$

$$= \frac{e^{\boldsymbol{\beta}^T \mathbf{X}_i \cdot y}}{1+e^{\boldsymbol{\beta}^T \mathbf{X}_i}}.$$

Thus the likelihood is $\mathcal{L}(\beta) = \prod_{i=1}^n {p_i}^{Y_i}(1-p_i)^{1-Y_i}$, which can be solved numerically, e.g. solving the score equations (you can derive this from the log-likelihood, take derivatives wrt. β).

$$\sum_{i=1}^n \begin{pmatrix} 1 \\ X_i \end{pmatrix} \left(Y_i - \frac{\exp(\beta^T X_i)}{1+\exp(\beta^T X_i)}\right) = 0.$$

M-estimation, preliminaries

You only need to know the basics of M-estimation. Some of the slides on M-estimation, labelled with *, are additional readings that you do not need to study in detail.

Consider a generic *statistical* model, and suppose we have i.i.d. random vectors $Z_1, \ldots, Z_n$ where $Z \sim \mathbb{P}_Z(z)$ from this model. Let θ be a k dimensional parameter. If θ fully characterizes $\mathbb{P}_Z(z)$, then we write $\mathbb{P}_Z(z;\theta)$. Let θ_0 denote the true value of θ. It follows that if θ fully characterizes $\mathbb{P}_Z(z)$, then the true density is $\mathbb{P}_Z(z;\theta_0)$.
We are considering the (classical) statistical problem of deriving an estimator for θ.

Definition of an M-estimator

Definition (M-estimator)

An M-estimator for θ is the solution $\hat{\theta}$ (assuming that it exists and is well defined) to the $(k \times 1)$ system of estimating equations

$$\sum_{i=1}^{n} M(Z_i; \hat{\theta}) = 0,$$

We say that $M(z; \theta) = \{M_1(z; \theta), \ldots, M_k(z; \theta)\}^T$ is an *unbiased estimating function for* $\mathbb{E}_\theta(M(Z_i; \theta)) = 0$. The expectation is taken wrt. to the distribution of Z at θ. From now on, we will suppress the subscript when we evaluate the expectation in the true value θ_0, i.e. $\mathbb{E}(M(Z_i; \theta)) \equiv \mathbb{E}_{\theta_0}(M(Z_i; \theta))$.

MLE is an M-estimator

Consider a fully parametric model $\mathbb{P}_Z(z;\theta)$. Define,

$$M(z;\theta) = \frac{\delta \log(\mathbb{P}_Z(z;\theta))}{\delta\theta},$$

where the right hand side is a k dimensional vector of derivatives. Solving an estimating equation with this $M(z;\theta)$ yields a maximum likelihood estimator (MLE), and thus the MLE is an M-estimator.

Methods of moment estimators are M-estimators

Consider a fully parametric model $\mathbb{P}_Z(z;\theta)$. Define,

$$M_m(Z_i;\theta) = Z_i^m - \mathbb{E}_\theta(Z_i^m),$$

where $m = 1, \ldots, k$, i.e. k is the dimension of θ.

Overview of properties of M-estimators

`This is for your information, not something we will go through in detail`

Theorem (M-estimator)

Under suitable regularity conditions, $\hat{\theta}$ is a consistent and asymptotically normal estimator,

$$\hat{\theta} \xrightarrow{P} \theta_0$$

and

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} \mathcal{N}(0, \Sigma),$$

where Σ is a covariance matrix.

*Sufficient regularity conditions for M-estimators

`This is for your information, not something we will go through in detail`

Suppose that the following regularity conditions hold.

1. $\sup_{\theta \in \Theta} |M_n(\theta) - M_0(\theta)| \xrightarrow{P} 0$.
2. For all $\epsilon > 0$, $\inf\{|M_0(\theta)| : d(\theta, \theta_0) \geq \epsilon\} > 0 = |M_0(\theta_0)|$.
   For this condition it is sufficient that there exists a unique solution, Θ is compact and M is continuous.
3. $M_n(\hat{\theta}_n) = o_P(1)$.

where $M_n(\theta) = \mathbb{E}_n(M(Z;\theta))$ is the expectation over the empirical distribution and $M_0(\theta) = \mathbb{E}(M(Z;\theta))$ over the true data generating law.

*Proof that the conditions above are sufficient for the consistency of M-estimators

Proof.

From the 2nd condition, for all $\epsilon > 0$ there is a $\delta > 0$ such that

$$
\begin{aligned}
& P(d(\hat{\theta}_n, \theta_0) \geq \epsilon) \\
\leq & P(|M_0(\hat{\theta}_n)| - |M_0(\theta_0)| \geq \delta) \\
= & P(|M_0(\hat{\theta}_n)| - |M_n(\hat{\theta}_n)| + |M_n(\hat{\theta}_n)| - |M_n(\theta_0)| + |M_n(\theta_0)| - |M_0(\theta_0)| \geq \delta) \\
\leq & P(|M_0(\hat{\theta}_n)| - |M_n(\hat{\theta}_n)| \geq \frac{\delta}{3}) + P(|M_n(\hat{\theta}_n)| - |M_n(\theta_0)| \geq \frac{\delta}{3}) + \\
& P(|M_n(\theta_0)| - |M_0(\theta_0)| \geq \frac{\delta}{3}).
\end{aligned}
$$

Condition 1 implies that the first and third probabilities go to zero.
Condition 3 implies that the second goes to zero. □

Example: Smoking Cessation A on weight gain Y.

1566 cigarette smokers aged 25-74 years. The outcome weight gain measured after 10 years.

Mean baseline characteristics	A = 1	A = 0
Age, years	46.2	42.8
Men, %	54.6	46.6
White, %	91.1	85.4
University, %	15.4	9.9
Weight, kg	72.4	70.3
Cigarettes/day	18.6	21.2
Years smoking	26.0	24.1
Little exercise, %	40.7	37.9
Inactive life, %	11.2	8.9

Miguel A Hernan and James M Robins. *Causal inference: What if?* CRC Boca Raton, FL:, 2018.

On estimation of causal effects

From slide 73, remember that from an experiment where A is randomised conditional on L, or more generally when consistency, positivity and exchangeability ($Y^a \perp\!\!\!\perp A \mid L$) hold, we have that

$$\begin{aligned}\mathbb{E}(Y^a) &= \sum_l \mathbb{E}(Y \mid L = l, A = a) \Pr(L = l) \\ &= \mathbb{E}\left[\frac{I(A = a)}{\pi(A \mid L)} Y\right].\end{aligned}$$

where $\pi(a \mid l) = P(A = a \mid L = l)$.
This equality motivates different estimators.

Regression estimator

We can also write

$$\begin{aligned}\mathbb{E}(Y^a) &= \sum_l \mathbb{E}(Y \mid L = l, A = a) \Pr(L = l) \\ &= \mathbb{E}(\mathbb{E}(Y \mid L, A = a)),\end{aligned}$$

where you should note that the outer expectation in the second line is with respect to the marginal of L. Denote

$$\mathbb{E}(Y \mid L = l, A = a) = Q(l, a).$$

$Q(l, a)$ is usually unknown, even in an experiment.

Regression estimator

Consider a parametric regression model $Q(I, a; \beta)$ of $Q(I, a)$; that is a linear or nonlinear function of (I, a) and the finite-dimensional parameter β.
We estimate β from the observed data. For example, we could in our conditional randomised trial pose a simple linear model

$$Q(I, a; \beta) = \beta_1 + \beta_2 a + \beta_3^T I,$$

which can be fitted with least squares methods.
If the outcome is binary ($Y \in \{0, 1\}$), we could fit a logistic regression model such as

$$\text{logit}\{Q(I, a; \beta)\} = \beta_1 + \beta_2 a + \beta_3^T I.$$

We can fit the logistic regression models with maximum likelihood estimators.

Definition (Correctly specified model)

A model is correctly specified if there exists a value β_0 such that $Q(I, a; \beta)$ evaluated at β_0 yields the true function $Q(I, a)$.

PS: As in any regression setting, the models we have posited may or may not be correctly specified.

Example continues

- We can estimate the conditional sample mean $\hat{\mathbb{E}}(Y \mid A = 1) = 4.5$ in quitters and $\hat{\mathbb{E}}(Y \mid A = 0) = 2.0$ in non-quitters. More specifically, the difference is

$$\hat{\mathbb{E}}(Y \mid A = 1) - \hat{\mathbb{E}}(Y \mid A = 0) = 2.5 \text{ (95\% CI : } 1.7, 3.4),$$

  but we will not assign a causal interpretation to the estimates.
- Let L include the baseline variables sex (0: male, 1: female), age (in years), race (0: white, 1: other), education (5 categories), intensity and duration of smoking (number of cigarettes per day and years of smoking), physical activity in daily life (3 categories), recreational exercise (3 categories), and weight (in kg).
- Suppose $A \perp\!\!\!\perp Y^a \mid L$.

Standardization: A natural way of estimating counterfactual outcomes

If we knew $Q(l, a)$, a natural way of estimating $\mathbb{E}(Y^a)$ is by the empirical average

$$\frac{1}{n}\sum_{i=1}^{n} Q(L_i, a),$$

motivated by the identification formula expression $\mathbb{E}(\mathbb{E}(Y \mid L, A = a))$. When we do not know $Q(l, a)$, but we assume that our model $Q(L_i, a; \beta)$ is correctly specified, we can use the outcome regression estimator to get the estimator

$$\hat{\mu}_{REG}(a) = \frac{1}{n}\sum_{i=1}^{n} Q(L_i, a; \hat{\beta}).$$

For example, using the linear estimator from the previous slide, we can estimate $\mathbb{E}(Y^{a=1})$ - $\mathbb{E}(Y^{a=0})$ by

$$\frac{1}{n}\sum_{i=1}^{n} Q(L_i, 1; \hat{\beta}) - \frac{1}{n}\sum_{i=1}^{n} Q(L_i, 0; \hat{\beta}) = \hat{\beta}_2,$$

that is, the regression parameter is the causal effect.

More broadly, our causal effects are not equal to regression coefficients

- Whereas the causal effect turned out to be equal to a regression coefficient in the previous slide, regression coefficients are not necessarily equal to our causal effect of interest.
- For example, the coefficients in the logistic regression model

$$\text{logit}\{Q(I, a; \beta)\} = \beta_1 + \beta_2 a + \beta_3^T I.$$

do not necessarily translate to a causal effect of interest.

Standardization (G-computation)

We say that standardization is a plug-in g-formula estimator because it simply replaces the conditional mean outcome in the g-formula by its estimates.

Section 27

Propensity score methods

Matching on the propensity score (intuitive motivation)

- In a homework you will see that, for all a,

$$Y^a \perp\!\!\!\perp A \mid L \implies Y^a \perp\!\!\!\perp A \mid \pi(a \mid L).$$

- We could, for each treated individual (i.e. individual with $A = 1$), match this individual with an untreated individual with *similar* propensity score.
- Then crudely compare the mean in the two groups.
- This crude comparison should be fine, but...
- Potential problems with matching (but not weighting, as we will se next)
  - What does similar propensity score mean? A conservative approach means that we "waste" data, but a loose approach mean that we compare people with different propensity scores...
  - How many matches should we choose?
  - Do we really get the average treatment effect?

Motivation for inverse probability weighting (IPW)

- We would like to adjust for confounding: imbalance between L's among those who are treated and untreated.
- Suppose that we find a treated subject i, who due to her confounders was *unlikely* to be treated. That is, $\pi(1 \mid L_i)$ is small.
- We *upweight* her, so that she represents herself but also the others like herself (in terms of L) who were unexposed.
- Similarly, we upweight untreated individuals with a small value of $\pi(0 \mid L_i)$.
- Heuristically, we can think about the weighted sample as a pseudopopulation where we observe each individual for each exposure level. In particular, $\pi^*(0 \mid L_i) = \pi^*(1 \mid L_i)$ for all i in the weighted population (which we indicate by the $*$).
- In this pseudopopulation, confounders are balanced between treatment groups, and a crude comparison estimates a causal effect (Intuitively, we get a new DAG for this pseudopopulation, where the arrow from L to A is omitted).

Motivating example

Suppose the counterfactual data are:

Group:	A			B			C		
Response Y^1:	1	1	1	2	2	2	3	3	3
Response Y^0:	0	0	0	1	1	1	2	2	2

and the average treatment effect $\mathbb{E}(Y^{a=1}) - \mathbb{E}(Y^{a=0}) = 1$.
but we observe:

Group:	A			B			C		
Response Y^1:	1	1	?	?	2	?	3	?	?
Response Y^0:	?	?	0	1	?	1	?	2	2

The naive contrast $\mathbb{E}(Y \mid A = 1) - \mathbb{E}(Y \mid A = 0) = \frac{7}{4} - \frac{6}{5} = 0.55$.

Example from Oliver Dukes.

Example continues

- However, from the table we see that,

$$\hat{\pi}(1, \text{group A}) = \frac{2}{3}$$
$$\hat{\pi}(1, \text{group B}) = \frac{1}{3}$$
$$\hat{\pi}(1, \text{group C}) = \frac{1}{3}$$

- Let us estimate $\mathbb{E}(Y^{a=1})$ by a weighted average, where each observation is weighted by $\frac{1}{\hat{\pi}(1,\text{group X})}$,Group X $\in \{\text{Group A}, \text{Group B}, \text{Group C}\}$,

$$\frac{(1+1)\frac{3}{2} + 2\frac{3}{1} + 3\frac{3}{1}}{\frac{3}{2} + \frac{3}{2} + \frac{3}{1} + \frac{3}{1}} = 2$$

and estimate $\mathbb{E}(Y^{a=0})$ by weighting each observation by $\frac{1}{\hat{\pi}(0,\text{Group X})}$, Group X $\in \{\text{Group A}, \text{Group B}, \text{Group C}\}$,

$$\frac{0\frac{3}{1} + (1+1)\frac{3}{2} + (2+2)\frac{3}{2}}{\frac{3}{1} + \frac{3}{2} + \frac{3}{2} + \frac{3}{2} + \frac{3}{2}} = 1.$$

Estimation when the propensity score is known

When $\pi(a \mid l)$ is a known function, the estimator of $\mathbb{E}(Y^a)$ is

$$\hat{\mu}_{IPW}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i = a) Y_i}{\pi(A_i \mid L_i)}.$$

The propensity score $\pi(a \mid l)$, unlike the function $Q(l, a)$, is known in randomised experiments (it is determined by the investigator). However, in most observational data settings, it is unknown.
PS: This estimator has been known for a long time and is often called the Horvitz Thompson estimator in survey sampling[38].

[38] Daniel G Horvitz and Donovan J Thompson. "A generalization of sampling without replacement from a finite universe". In: *Journal of the American statistical Association* 47.260 (1952), pp. 663–685.

Estimation when the propensity score is unknown

More generally, we can propose a regression model $\pi(A \mid L; \gamma)$ for $\pi(A \mid L)$, and we can consider the estimator

$$\hat{\mu}_{IPW}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i = a) Y_i}{\pi(A_i \mid L_i; \gamma)}.$$

For example, suppose that we fit a logistic regression model and find the MLE $\hat{\gamma}$ of γ, which is the solution to the estimating equation (See slide 250)

$$\sum_{i=1}^{n} \begin{pmatrix} 1 \\ L_i \end{pmatrix} \left(A_i - \frac{\exp(\gamma_1 + \gamma_2^T L_i)}{1 + \exp(\gamma_1 + \gamma_2^T L_i)} \right) = 0.$$