ECON675 – Assignment 4

Anirudh Yadav

November 9, 2018

Contents

1 Estimating equations

1.1 Moment conditions

The goal of this question is to show that the four given functions are valid moment conditions for the parameter $\theta_t(g)$. That is, we want to show that

$$\mathbb{E}[\psi_{\mathtt{f},t}(\boldsymbol{Z}_i; \theta_t(g))] = 0,$$

for each $\mathtt{f} \in \{\mathtt{IPW}, \mathtt{RI1}, \mathtt{RI2}, \mathtt{DR}\}$. Note that in the derivations below I invoke LIE a lot without specifically mentioning it.

Start with the inverse probability weighting function

$$
\begin{aligned}
\mathbb{E}[\psi_{\mathtt{IPW},t}(\boldsymbol{Z}_i; \theta_t(g))] &= \mathbb{E}\left[\frac{D_i(t) \cdot g(Y_i(t))}{p_t(\boldsymbol{X}_i)}\right] - \theta_t(g) \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{D_i(t) \cdot g(Y_i(t))}{p_t(\boldsymbol{X}_i)} | \boldsymbol{X}_i\right]\right] - \theta_t(g) \\
&= \mathbb{E}\left[\frac{1}{p_t(\boldsymbol{X}_i)}\mathbb{E}\left[D_i(t)|\boldsymbol{X}_i\right]\mathbb{E}\left[g(Y_i(t))|\boldsymbol{X}_i\right]\right] - \theta_t(g)
\end{aligned}
$$

Now,

$$\mathbb{E}\left[D_i(t)|\boldsymbol{X}_i\right] = \Pr[D_i(t) = 1|\boldsymbol{X}_i] = \Pr[T_i = t|\boldsymbol{X}_i] = p_t(\boldsymbol{X}_i).$$

Thus,

$$
\begin{aligned}
\mathbb{E}[\psi_{\mathtt{IPW},t}(\boldsymbol{Z}_i; \theta_t(g))] &= \mathbb{E}\left[\mathbb{E}\left[g(Y_i(t))|\boldsymbol{X}_i\right]\right] - \theta_t(g) \\
&= \mathbb{E}[g(Y_i(t))] - \theta_t(g) \\
&= 0.
\end{aligned}
$$

Next, consider

$$
\begin{aligned}
\mathbb{E}[\psi_{\mathtt{RI1},t}(\boldsymbol{Z}_i; \theta_t(g))] &= \mathbb{E}[e_t(g; \boldsymbol{X}_i)] - \theta_t(g) \\
&= \mathbb{E}[\mathbb{E}[g(Y_i(t)|\boldsymbol{X}_i]] - \theta_t(g) \\
&= \mathbb{E}[g(Y_i(t)] - \theta_t(g) \\
&= 0.
\end{aligned}
$$

And,

$$
\begin{aligned}
\mathbb{E}[\psi_{\mathtt{RI2},t}(\boldsymbol{Z}_i; \theta_t(g))] &= \mathbb{E}\left[\frac{D_i(t) \cdot e_t(g; \boldsymbol{X}_i)}{p_t(\boldsymbol{X}_i)}\right] - \theta_t(g) \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{D_i(t) \cdot e_t(g; \boldsymbol{X}_i)}{p_t(\boldsymbol{X}_i)} | \boldsymbol{X}_i\right]\right] - \theta_t(g) \\
&= \mathbb{E}\left[\mathbb{E}\left[e_t(g; \boldsymbol{X}_i)|\boldsymbol{X}_i\right]\right] - \theta_t(g) \\
&= \mathbb{E}[e_t(g; \boldsymbol{X}_i)] - \theta_t(g) \\
&= 0.
\end{aligned}
$$

Finally, consider the doubly robust function

$$\mathbb{E}[\psi_{\text{DR},t}(\boldsymbol{Z}_i; \theta_t(g))] = \mathbb{E}\left[\frac{D_i(t) \cdot g(Y_i(t))}{p_t(\boldsymbol{X}_i)}\right] - \theta_t(g) - \mathbb{E}\left[\frac{e_t(g; \boldsymbol{X}_i)}{p_t(\boldsymbol{X}_i)}(D_i(t) - p_t(\boldsymbol{X}_i))\right].$$

Using the IPW result above, we know that the first two terms cancel each other out, so that

$$\begin{aligned}
\mathbb{E}[\psi_{\text{DR},t}(\boldsymbol{Z}_i; \theta_t(g))] &= -\mathbb{E}\left[\frac{e_t(g; \boldsymbol{X}_i)}{p_t(\boldsymbol{X}_i)}(D_i(t) - p_t(\boldsymbol{X}_i))\right] \\
&= -\mathbb{E}\left[\frac{e_t(g; \boldsymbol{X}_i)D_i(t)}{p_t(\boldsymbol{X}_i)}\right] + \mathbb{E}[e_t(g; \boldsymbol{X}_i)] \\
&= -\theta_t(g) + \theta_t(g) \\
&= 0.
\end{aligned}$$

So each of the four functions is a valid moment condition for $\theta_t(g)$.

1.2 Plug-in estimators

The plug-in IPW estimator is

$$\hat{\theta}_{\text{IPW},t}(g) = \frac{1}{n}\sum_{i=1}^{n}\frac{D_i(t)g(Y_i)}{\hat{p}_t(\boldsymbol{X}_i)},$$

where $\hat{p}_t(\boldsymbol{X}_i)$ is the estimated propensity score. Note that since there are multiple treatment levels, the estimated propensity score would have to be computed using a suitable discrete choice model. For instance, $\hat{p}_t(\boldsymbol{X}_i)$ could be estimated using a multinomial logit model.

The plug-in projection (or regression imputation) estimator is

$$\begin{aligned}
\hat{\theta}_{\text{RI1},t}(g) = \hat{\mathbb{E}}[e_t(g; \boldsymbol{X}_i)] = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbb{E}}[g(Y_i(t))|\boldsymbol{X}_i] &= \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbb{E}}[g(Y_i(t))|\boldsymbol{X}_i, D_i(t) = 1] \\
&= \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbb{E}}[g(Y_i)|\boldsymbol{X}_i, D_i(t) = 1],
\end{aligned}$$

where the second last equality uses the ignorability assumption. We need to make a choice about how to estimate the conditional expectation term. I think we could use NLS, or possibly a non-parametric method like kernel regression. To ease notation, let $\hat{\mu}_t(\boldsymbol{X}_i)$ be the parametric or nonparametric estimate of $\mathbb{E}[g(Y_i)|\boldsymbol{X}_i, D_i(t) = 1]$. Then, the projection estimator is

$$\hat{\theta}_{\text{RI1},t}(g) = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_t(\boldsymbol{X}_i)$$

The plug-in 'hybrid' imputation estimator

$$\hat{\theta}_{\text{RI2},t}(g) = \frac{1}{n}\sum_{i=1}^{n}\frac{D_i(t)\hat{\mu}_t(\boldsymbol{X}_i)}{\hat{p}_t(\boldsymbol{X}_i)}.$$

3

Finally, the plug-in doubly robust estimator is given by

$$\hat{\theta}_{\text{DR},t}(g) = \frac{1}{n}\sum_{i=1}^{n}\frac{D_i(t)g(Y_i)}{\hat{p}_t(\boldsymbol{X}_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\mu}_t(\boldsymbol{X}_i)}{\hat{p}_t(\boldsymbol{X}_i)}(D_i(t) - \hat{p}_t(\boldsymbol{X}_i))$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i(t)(g(Y_i) - \widehat{\mu}_t(\boldsymbol{X}_i))}{\hat{p}_t(\boldsymbol{X}_i)} + \widehat{\mu}_t(\boldsymbol{X}_i)\right).$$

As discussed in Abadie and Catteneo (2018), the relative performance of the above estimators depends on the features of the data generating process. In finite samples, IPW estimators become unstable when the propensity score approaches zero or one and regression imputation estimators may suffer from extrapolation biases. Doubly robust estimators include safeguards against bias caused by misspecification but impose additional specification choices that may affect the resulting estimate.

1.3 Estimating the variance of potential outcomes

2   Estimating average treatment effects

Estimation results are overleaf. Estimates of ATEs are very different across the Lalonde and PSID control samples (for all estimators), which suggests that there are large and systematic differences between the two control groups . For instance, ATEs are generally positive when using the Lalonde control, while many ATEs are large and negative when using the PSID control group. This makes sense because the NSW program was targeted at individuals who had faced economic and social problems prior to enrolment (i.e. these respondents, on average, had low incomes), whereas the PSID sample is more representative of the entire US population.

Most estimates of ATTs are broadly similar across the samples. The ATTs using matching estimators are quite similar across both samples. Again, this makes sense, because matching estimators specifically ensure that there is a decent degree of covariate balance between the treatment and control samples (i.e. the matching estimators would pick respondents from the PSID sample who have similar characteristics to the respondents in the Lalonde experiment).

I had some issues with the estimation of both ATEs and ATTs:

- I had problem getting RI/IPW/DR estimates for PSID covariate set A in `R` (using the `CausalGAM` package) so I used `STATA` only in this case.

- IPW and DR ATT estimates are the same in `STATA`; I'm not sure why.

- Many of the logit regressions for estimating propensity scores did not converge, so I generally set the maximum number of iterations to 25.

Table 1: Estimation and Inference on ATE and ATT

(a) ATE

	Experimental Data			PSID Control			
	$\hat{\tau}$	s.e.	C.I.		$\hat{\tau}$	s.e.	C.I.	
Mean Diff.								
	1794.34	670.82	479.53	3109.16	−15204.78	655.91	−16490.37	−13919.19
OLS								
a	1582.17	658.74	291.04	2873.30	6302.40	1209.32	3932.13	8672.66
b	1506.90	657.12	218.95	2794.85	4699.26	1027.12	2686.11	6712.41
c	1501.37	662.91	202.07	2800.68	4284.34	1031.34	2262.91	6305.77
Reg. Impute								
a	1462.27	629.75	227.97	2696.57	−11195.04	1741.33	−14608.04	−7782.04
b	1454.13	630.67	218.02	2690.24	−10398.22	3549.31	−17354.86	−3441.58
c	1427.53	641.70	169.80	2685.26	−11920.18	3497.68	−18775.64	−5064.72
IPW								
a	1537.40	629.76	303.07	2771.72	−13507.18	2800.20	−18995.57	−8018.79
b	1469.62	630.67	233.51	2705.72	−7246.13	3549.83	−14203.79	−288.46
c	1468.10	641.70	210.37	2725.84	−7486.53	3498.73	−14344.05	−629.01
D. Robust								
a	1473.17	629.75	238.87	2707.48	−13507.18	2800.20	−18995.57	−8018.79
b	1451.05	630.67	214.95	2687.16	−11419.47	3549.36	−18376.22	−4462.72
c	1423.88	641.70	166.15	2681.61	−12504.50	3497.70	−19359.99	−5649.00
N1 Match								
a	1829.80	779.60	301.78	3357.81	−15619.49	1153.32	−17879.99	−13358.99
b	1875.83	734.95	435.32	3316.34	−9349.56	3974.77	−17140.11	−1559.01
c	1671.74	726.06	248.66	3094.82	−9561.97	4033.54	−17467.71	−1656.23
p Match								
a	1541.57	646.32	274.79	2808.35	−15858.78	6749.93	−29088.64	−2628.92
b	1488.63	765.40	−11.56	2988.83	8646.26	15055.98	−20863.46	38155.98
c	1256.81	677.13	−70.37	2583.98	−9561.97	4033.54	−17467.71	−1656.23

(b) ATT

|	Experimental Data			PSID Control			
	$\hat{\tau}$	s.e.	C.I.		$\hat{\tau}$	s.e.	C.I.	
Mean Diff.								
	1794.34	670.82	479.53	3109.16	−15204.78	655.91	−16490.37	−13919.19
OLS								
a	1582.17	658.74	291.04	2873.30	6302.40	1209.32	3932.13	8672.66
b	1506.90	657.12	218.95	2794.85	4699.26	1027.12	2686.11	6712.41
c	1501.37	662.91	202.07	2800.68	4284.34	1031.34	2262.91	6305.77
Reg. Impute								
a	1726.60	688.76	376.65	3076.55	8543.16	1233.32	6125.85	10960.48
b	1809.70	693.87	449.72	3169.68	4932.65	1088.73	2798.75	7066.56
c	1844.61	694.82	482.76	3206.46	−11920.18	3834.63	−19436.06	−4404.30
IPW								
a	1765.86	698.04	397.70	3134.02	2750.93	886.00	1014.36	4487.49
b	1741.49	701.89	365.78	3117.20	2204.32	940.61	360.73	4047.91
c	1774.86	702.34	398.27	3151.45	2417.65	973.18	510.22	4325.09
D. Robust								
a	1765.86	698.04	397.70	3134.02	2750.93	886.00	1014.36	4487.49
b	1741.49	701.89	365.78	3117.20	2204.32	940.61	360.73	4047.91
c	1774.86	702.34	398.27	3151.45	2417.65	973.18	510.22	4325.09
N1 Match								
a	1558.16	776.73	35.76	3080.55	1671.33	1226.74	−733.07	4075.74
b	1731.61	732.36	296.18	3167.04	2043.34	1006.33	70.92	4015.75
c	1137.42	813.44	−456.91	2731.76	2003.80	935.02	171.16	3836.43
p Match								
a	1688.12	786.82	145.95	3230.29	2460.73	825.48	842.79	4078.68
b	1858.18	803.62	283.09	3433.27	2105.43	1129.41	−108.21	4319.07
c	1606.61	809.77	19.46	3193.76	2018.73	1405.34	−735.72	4773.19

3 Post-model selection inference

3.1 Summary statistics and density plots

The summary statistics for each estimator are presented in Table 2 and the kernel density plots are just below. As expected, the empirical distribution of $\hat{\beta}$ is approximately normal with a mean very close to the true parameter, $\beta_0 = 0.5$. The empirical distribution of $\tilde{\beta}$ looks fairly normal, but clearly $\tilde{\beta}$ is inconsistent for β_0. As expected, the empirical distribution of $\check{\beta}$ is clearly not normal, and the mean is a bit than the true parameter.

Table 2: **Summary Statistics for each Estimator**

Estimator	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
(i) $\hat{\beta}$	-0.50	0.32	0.52	0.51	0.71	1.32
(ii) $\tilde{\beta}$	0.77	1.26	1.36	1.36	1.47	1.88
(iii) $\check{\beta}$	-0.50	0.32	0.52	0.54	0.72	1.66
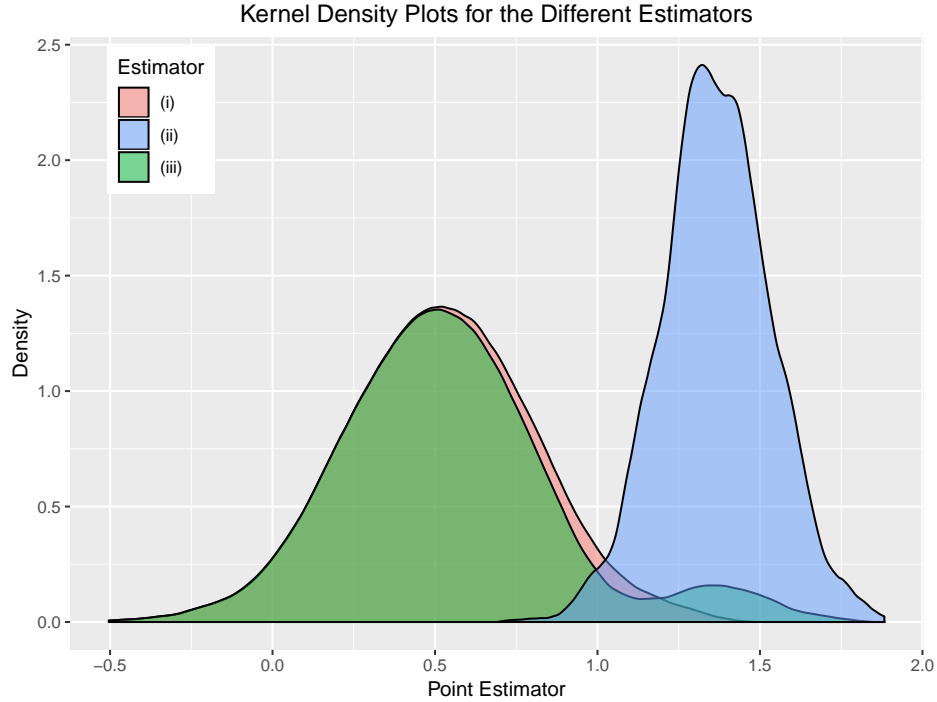
3.2 Coverage rates

To compute the coverage rates, I first create an indicator variable which is equal to 1 if, for a given simulation, the 95% asymptotic CI contains the true parameter $\beta_0 = 0.5$; I do this for each estimator and then take the mean of the indicator variable to get the coverage rate. Note that the asymptotic CI using $\check{\beta}$ is just equal to the CI using $\hat{\beta}$ whenever $\check{\beta} = \hat{\beta}$, and equal to the CI using $\tilde{\beta}$ otherwise.

The $\hat{\beta}$'s coverage rate is 0.936, which is very close to 0.95, as expected. In contrast, $\tilde{\beta}$'s coverage rate is only 0.001, which is unsurprising since $\tilde{\beta}$ is inconsistent for β_0. Finally, $\check{\beta}$'s coverage rate is 0.900, which is because $\check{\beta}$ is also inconsistent for β_0.