Economics 675: Applied Microeconometrics – Fall 2018

Assignment 3 – Due date: Mon 22-Oct

Last updated: June 4, 2018

Contents

Guidelines:

- You may work in (small) groups while solving this assignment.
- Submit individual solutions via `http://canvas.umich.edu` in one PDF file collecting everything (e.g., derivations, figures, tables, computer code).
- Computer code should be done in both `Stata` and `R`. If the numerical results do not agree across the two statistical software platforms, you must explain why that is the case.
- Start each question on a separate page. Always add a reference section if you cite other sources.
- Clearly label all tables and figures, and always include a brief footnote with useful information.
- Always attach your computer code as an appendix, with annotations/comments as appropriate.
- Please provide as much detail as possible in your answers, both analytical and empirical.

1 Question 1: Non-linear Least Squares

Let $\{(y_i, \mathbf{x}_i') : 1 \leq i \leq n\}$ be a random sample with $y_i \in \mathbb{R}$ and $\mathbf{x}_i \in \mathbb{R}^d$. Suppose that

$$\mathbb{E}[y_i|\mathbf{x}_i] = \mu(\mathbf{x}_i'\boldsymbol{\beta}_0) \qquad \text{and} \qquad \mathbb{V}[y_i|\mathbf{x}_i] = \sigma^2(\mathbf{x}_i),$$

where the (link) function $\mu : \mathbb{R} \mapsto \mathbb{R}$ is known, but the (heteroskedasticity) function $\sigma^2 : \mathbb{R}^d \mapsto \mathbb{R}_{++}$ and the parameter $\boldsymbol{\beta}_0 \in \mathbb{R}^d$ are unknown. Consider the M-estimator:

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu(\mathbf{x}_i'\boldsymbol{\beta}))^2,$$

under appropriate regularity conditions required below. For example, you may assume as many integrability and differentiability conditions (e.g., on $\mu(\cdot)$) as needed.

1. Give sufficient conditions so that $\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \mathbb{E}[(y_i - \mu(\mathbf{x}_i'\boldsymbol{\beta}))^2]$ is identified.
   Are there conditions under which $\boldsymbol{\beta}_0$ can be written in closed form?

2. Give sufficient conditions so that $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \to_d \mathcal{N}(\mathbf{0}, \mathbf{V}_0)$, and identify precisely the form of the asymptotic variance $\mathbf{V}_0 \in \mathbb{R}^{d\times d}$. (Hint: you could start from the first order condition and rewrite the problem into a Z-estimation.)

3. Propose an estimator $\hat{\mathbf{V}}_n^{\mathtt{HC}}$ of $\mathbf{V}_0$, and give sufficient conditions so that $\hat{\mathbf{V}}_n^{\mathtt{HC}} \to_p \mathbf{V}_0$.
   Construct an asymptotic 95% confidence interval for $\|\boldsymbol{\beta}_0\|^2$, with $\|\cdot\|$ the Euclidean norm.

4. Suppose that $\sigma^2(\mathbf{x}_i) = \sigma^2$, an unknown constant. Show that $\mathbf{V}_0$ simplifies.
   Propose an estimator $\hat{\mathbf{V}}_n^{\mathtt{HO}}$ of $\mathbf{V}_0$, and give sufficient conditions so that $\hat{\mathbf{V}}_n^{\mathtt{HO}} \to_p \mathbf{V}_0$.
   Construct an asymptotic 95% confidence interval for $\|\boldsymbol{\beta}_0\|^2$, with $\|\cdot\|$ the Euclidean norm.

5. Suppose that $y_i|\mathbf{x}_i \sim \mathcal{N}(\mu(\mathbf{x}_i'\boldsymbol{\beta}_0), \sigma^2)$. Show that $\hat{\boldsymbol{\beta}}_{\mathtt{ML}} = \hat{\boldsymbol{\beta}}_n$, where $\hat{\boldsymbol{\beta}}_{\mathtt{ML}}$ is the MLE of $\boldsymbol{\beta}_0$.
   Derive the MLE of σ^2. Does it coincide with the one proposed in part 4?

6. Show that if the (link) function $\mu(\cdot)$ is unknown, then $\boldsymbol{\beta}_0$ is not identifiable.
   Can you give conditions so that identifiability of $\boldsymbol{\beta}_0$ is restored?

7. Suppose now that $y_i = \mathbf{1}(\mathbf{x}_i'\boldsymbol{\beta}_0 - \varepsilon_i \geq 0)$, with $\varepsilon_i|\mathbf{x}_i \sim F(\cdot)$, where

   $$F(u) = \frac{1}{1 + e^{-u/s_0}}, \qquad u \in \mathbb{R},\ s_0 > 0.$$

   denotes the logistic c.d.f. symmetric around zero. Apply the results above and derive the exact formulas for $\mu(\mathbf{x}_i'\boldsymbol{\beta}_0)$, $\sigma^2(\mathbf{x}_i)$ and $\mathbf{V}_0$.

8. Instead of using the nonlinear least squares, consider the maximum likelihood estimation of $\boldsymbol{\beta}_0$. Does the MLE give the same result?

9. Consider the `pisofirme.csv` data described in Section 4. Using the results above, estimate a Logistic regression model for missing outcome data using the missing indicator $\mathtt{dmissing}_i$. Specifically, consider the simple model with the binary outcome variable $s_i = 1 - \mathtt{dmissing}_i$ and the covariates

   - `S_age`,
   - `S_HHpeople`,

- log(`S_incomepc`+1).

(a) Compute $\hat{\boldsymbol{\beta}}_n$, $\hat{\mathbf{V}}_n^{\texttt{HC}}$, as well as t-test statistic, asymptotic p-value, and asymptotic 95% confidence interval for each coefficient in $\boldsymbol{\beta}_0$.

(b) Compute the p-value and 95% confidence interval for each coefficient in $\boldsymbol{\beta}_0$ using the nonparametric bootstrap to approximate the finite sample distribution of the t-test statistic.

(c) Predict for each observation the probability of being observed, given by $\mu(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_n)$, which is sometimes called the propensity score. Plot a kernel density estimate of $\mu(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_n)$.

2 Question 2: Semiparametric GMM with Missing Data

Consider again the `pisofirme.csv` data described in Section 4. This question will estimate the treatment effect of installing cement floors in poor households on the incidence of anemia in children 0-5 years old, using a parametric discrete choice model but accounting for missing data semiparametrically.

Let $(y_i, t_i, s_i, \mathbf{x}_i)'$, $i = 1, 2, \cdots, n$, be a random sample with $y_i \in \{0, 1\}$ denoting the *observed* outcome of interest ($\texttt{danemia}_i$), $t_i \in \{0, 1\}$ denoting the treatment effect indicator ($\texttt{dpisofirme}_i$), $s_i \in \{0, 1\}$ denoting the missing outcome indicator ($s_i = 1 - \texttt{dmissing}_i$), and $\mathbf{x}_i \in \mathbb{R}^d_i$ denoting a vector of observed covariates. Notice that the observed outcome satisfies: $y_i = s_i \cdot y_i^*$, where y_i^* is the true outcome not always observed, and sometimes called latent or potential outcome. Suppose the following conditional moment condition holds:

$$\mathbf{0} = \mathbb{E}[m(y_i^*, t_i, \mathbf{x}_i; \boldsymbol{\beta}_0)|t_i, \mathbf{x}_i], \qquad m(y_i^*, t_i, \mathbf{x}_i; \boldsymbol{\beta}) = y_i^* - F(t_i \cdot \theta + \mathbf{x}_i'\boldsymbol{\gamma}),$$

where $F(\cdot)$ is known (e.g., logit or probit link function), $\boldsymbol{\beta}_0 = (\theta_0, \boldsymbol{\gamma}_0')'$, and $\boldsymbol{\gamma}_0 \in \mathbb{R}^d$.

1. Under appropriate regularity conditions, show that setting

$$\mathbf{g}_0(t_i, \mathbf{x}_i) = \frac{f(t_i \cdot \theta_0 + \mathbf{x}_i'\boldsymbol{\gamma}_0)}{F(t_i \cdot \theta_0 + \mathbf{x}_i\boldsymbol{\gamma}_0)(1 - F(t_i \cdot \theta_0 + \mathbf{x}_i'\boldsymbol{\gamma}_{\mathbf{0}})} \begin{bmatrix} t_i \\ \mathbf{x}_i' \end{bmatrix}$$

   leads to an optimal unconditional moment condition: $\mathbf{0} = \mathbb{E}[\mathbf{g}_0(t_i, \mathbf{x}_i) m(y_i^*, t_i, \mathbf{x}_i; \boldsymbol{\beta}_0)]$. (Hint: refer to Section 14.4.3 in Woodridge (2002).) Show further that if $F(\cdot)$ is the logistic cdf, then the form of the instruments simplifies to $\mathbf{g}_0(t_i, \mathbf{x}_i) = [t_i, \mathbf{x}_i']'$.

2. Assume $s_i \perp\!\!\!\perp (y_i^*, t_i, \mathbf{x}_i)$, which is sometimes known as *Missing Completely at Random* (MCAR).

   (a) Imposing regularity conditions, show that dropping the observations with missing outcomes leads to valid inference. That is, show that the estimator $\hat{\boldsymbol{\beta}}_{\texttt{MCAR}}$ solving the moment condition

$$\mathbf{0} \approx \hat{\mathbb{E}}[\mathbf{g}_0(t_i, \mathbf{x}_i) m(y_i, t_i, \mathbf{x}_i; \hat{\boldsymbol{\beta}}_{\texttt{MCAR}})|s_i = 1]$$

   is consistent (for $\boldsymbol{\beta}_0$) and asymptotically normal (under appropriate centering and scaling). Propose a feasible estimator that is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{\texttt{MCAR}}$.

   (b) Using the `pisofirme.csv` data implement the feasible estimator. Report the point estimator and 95% confidence interval, using the nonparametric bootstrap, for each element of the population parameter vector $\boldsymbol{\beta}_0$. (For this question assume F is the logistic cdf and use the same covariates `S_age`, `S_HHpeople`, log(`S_incomepc`+1) as well as the treatment indicator `dpisofirme`.)

3. Assume $s_i \perp\!\!\!\perp y_i^* \mid (t_i, \mathbf{x}_i)$, which is sometimes known as *Missing at Random* (MAR). Now consider the conditional moment restriction

$$\mathbb{E}[s_i m(y_i^*, t_i, \mathbf{x}_i)|t_i, \mathbf{x}_i] = 0.$$

   (a) Show that the optimal instruments remain the same as $\mathbf{g}_0(t_i, \mathbf{x}_i)$.

   (b) Under regularity conditions, it can be shown that $\tilde{\boldsymbol{\beta}}_{\texttt{MAR}}$, which solves the moment condition (here we are using seemingly arbitrary instruments $\mathbf{g}_0(t_i, \mathbf{x}_i)/p_0(t_i, \mathbf{x}_i)$, which are valid but not optimal)

$$\mathbf{0} \approx \hat{\mathbb{E}}\left[\frac{s_i}{p_0(t_i, \mathbf{x}_i)} \mathbf{g}_0(t_i, \mathbf{x}_i) m(y_i, t_i, \mathbf{x}_i; \tilde{\boldsymbol{\beta}}_{\texttt{MAR}})\right],$$

where $p_0(t_i, \mathbf{x}_i) = \mathbb{E}[s_i|t_i, \mathbf{x}_i] = \mathbb{P}[\text{not missing}|t_i, \mathbf{x}_i]$, satisfies $\tilde{\boldsymbol{\beta}}_{\texttt{MAR}} \to_p \boldsymbol{\beta}_0$ and that $\sqrt{n}(\tilde{\boldsymbol{\beta}}_{\texttt{MAR}} - \boldsymbol{\beta}_0) \to_d \mathcal{N}(0, \mathbf{V}_0)$ for some positive-definite variance-covariance matrix $\mathbf{V}_0$.

Note that the above estimator is not feasible since the propensity score $p_0(t_i, \mathbf{x}_i)$ is unknown. Propose a feasible estimator denoted by $\hat{\boldsymbol{\beta}}_{\texttt{MAR}}$. Is it true, in general, that $\hat{\boldsymbol{\beta}}_{\texttt{MAR}}$ and $\tilde{\boldsymbol{\beta}}_{\texttt{MAR}}$ are asymptotically equivalent?

(c) Using the `pisofirme.csv` data implement the feasible estimator. Report the point estimator and 95% confidence interval, using the nonparametric bootstrap, for each element of the population parameter vector $\boldsymbol{\beta}_0$. (For simplicity assume both F and p_0 are the logistic cdf.)

(d) In real application when the above estimator is used, researchers worry about situations where the estimated $\hat{p}(\cdot)$ is very close to zero. Redo the above exercise, but use only observations that $\hat{p}(t_i, \mathbf{x}_i) \geq 0.1$ (this is called trimming). Does the result change a lot?

3 Question 3: When Bootstrap Fails

Here we consider a very simple model in which the nonparametric bootstrap fails. Assume you have a random sample $\{x_i\}$ from $\mathsf{Uniform}[0, \theta_0]$, and recall that the maximum likelihood estimator of θ_0 is $\max_i x_i$. Without loss of generality let $\theta_0 = 1$, and simple calculation implies

$$n \cdot \left(1 - \max_i x_i\right) \to_d \mathsf{Exponential}(1).$$

In the simulation below, use $n = 1000$.

1. Set the seed to be 123, so that your result is replicable. Simulate a random sample from the $\mathsf{Uniform}[0, 1]$ distribution with $n = 1000$, and then implement the nonparametric bootstrap 599 times, and plot the distribution of the following bootstrap statistic:

   $$n \cdot \left(\max_i x_i - \max_i x_i^*\right), \qquad x_i^* \sim_{iid} \mathsf{Discrete.Uniform}\{x_1, x_2, \cdots, x_n\},$$

   where $\{x_i^*\}$ is the bootstrap sample (here we are using the nonparametric bootstrap hence $\{x_i^*\}$ is obtained by sampling from $\{x_i\}$ with replacement). Does it coincide with the theoretical $\mathsf{Exponential}(1)$ distribution?

2. Now we consider the parametric bootstrap. Again set the seed to be 123 and simulate a random sample from the $\mathsf{Uniform}[0, 1]$ distribution with $n = 1000$. Here we consider the parametric bootstrap with 599 repetitions. To be more specific, for each bootstrap you simulate a bootstrap sample $\{x_i^*\}$ from $\mathsf{Uniform}[0, \max_i x_i]$ distribution (this is why it is called the parametric bootstrap, since you use some parametric assumption), and then compute the bootstrap statistic

   $$n \cdot \left(\max_i x_i - \max_i x_i^*\right), \qquad x_i^* \sim_{iid} \mathsf{Uniform}[0,\ \max_i x_i].$$

   Plot the distribution of the bootstrap statistic. Does it coincide with the theoretical $\mathsf{Exponential}(1)$ distribution?

3. Give an intuitive reason why the nonparametric bootstrap fails in this example. (Hint: what is $\mathbb{P}[\max_i x_i^* = \max_i x_i]$ converging to in the nonparametric bootstrap?)

4 Appendix: `pisofirme.csv` Data Description

Piso Firme, which means "firm floor," is a program designed to replace dirt floors with cement floors in the homes of low-income families in Mexico, and improve child health and development. Eligible households were offered up to 50 square meters (538 square feet) of cement valued at about 1,500 Mexican pesos (approximately $150 US).

The program began as a state program in Coahuila since 2000 and provided cement floors to more than 34,000 households by the survey time (2005). Later the program was adopted by the Federal Government and gradually scaled up to other states. One state that did not fully scale up Piso Firme by 2005 was the neighboring State of Durango.

The analysis of Cattaneo, Galiani, Gertler, Martinez, and Titiunik (2009) is based on families residing in the twin cities of Gómez Palacios and Lerdo (control) and Torreón (treatment) that straddle the border of the States of Durango and Coahuila, respectively. These cities are split administratively between the two states, but are effectively a single urban area in socioeconomic terms. Therefore the homogeneity of the households in the twin cities as well as the timing differences in adopting the Piso Firme program provides the identification of the treatment effect.

The file `pisofirme.csv` contains the data used in Cattaneo, Galiani, Gertler, Martinez, and Titiunik (2009) (number of observations: 4,092).

Variable	obs	Mean	St.Dev	Min	Max	Description
danemia	3758	.3861	.4869	0	1	Anemia
dmissing	4092	.0816	.2738	0	1	anemia missing indicator
dpisofirme	4092	.4839	.4998	0	1	Treatment Indicator (=1 Torreon; =0 Durango)
idcluster	4052	–	–	–	–	ID Census Block
S_age	4092	2.613	1.713	0	5	Age
S_gender	4092	.5044	.5000	0	1	Male (=1)
S_childma	4092	.9653	.1830	0	1	Mother of at least one child in household present (=1)
S_childmaage	3890	27.43	6.385	14	66	Mother's age (if present)
S_childmaeduc	3888	6.968	2.689	0	16	Mother's years of schooling (if present)
S_childpa	4092	.7803	.4140	0	1	Father of at least one child in household present (=1)
S_childpaage	3037	30.49	7.712	13	77	Father's age (if present)
S_childpaeduc	3027	6.990	3.082	0	16	Father's years of schooling (if present)
S_HHpeople	4092	5.746	2.227	0	19	Number of household members
S_rooms	4092	2.045	1.101	1	10	Number of rooms
S_waterland	4092	.9716	.1659	0	1	Water connection (=1)
S_waterhouse	4092	.5232	.4995	0	1	Water connection inside the house (=1)
S_electricity	4092	.9890	.1043	0	1	Electricity (=1)
S_milkprogram	4092	.0694	.2541	0	1	Beneficiary of government milk supplement program (=1)
S_foodprogram	4092	.0298	.1701	0	1	Beneficiary of government food program (=1)
S_seguropopular	4092	.0156	.1241	0	1	Beneficiary of seguro popular (=1)
S_hasanimals	4092	.4934	.5000	0	1	Has animals on land (=1)
S_animalsinside	4092	.1928	.3946	0	1	Animals allowed to enter the house (=1)
S_garbage	4092	.8174	.3863	0	1	Uses garbage collection service (=1)
S_washhands	4092	3.725	1.500	0	10	Number of times respondent washed hands the day before
S_incomepc	4089	993.1	3344	0	127266.7	Total household income per capita
S_cashtransfers	4089	12.73	34.49	0	950	Transfers per capita from government programs
S_assetspc	4090	20946	7226	5573.627	64594.2	Total value of household assets per capita

References

Cattaneo, M. D., S. Galiani, P. J. Gertler, S. Martinez, and R. Titiunik (2009): “Housing, Health, and Happiness,” *American Economic Journal: Economic Policy*, 1, 75–105.

Woodridge, J. (2002): *Econometric analysis of cross sectional data and panel data.* Cambridge and London: MIT press.