EMPIRISCHES PROJEKT MIT DATENSIMULATIONEN

1 Allgemeines

Bitte geben Sie bis 26.04.2018 die Mitglieder Ihrer Gruppe per E-Mail bekannt.
Die Gruppenarbeit ist mit maximal 4 Personen pro Gruppe durchzuführen.
Der Programmcode und die PDF-Dokumentation sind per E-Mail bis spätestens 31.05.2018 zu übermitteln.
E-Mail Adresse: hugo.bodory@unisg.ch.

2 Projektdesign

Bei diesem Projekt evaluieren Sie ein Schätzverfahren, das die Effektivität eines Beschäftigungsprogramms misst. Die Evaluierung basiert auf der Simulation empirischer Daten. Diese Daten wurden erstmals von LaLonde (1986) verwendet um zu zeigen, ob Arbeitsmarktmassnahmen einen positiven Einfluss auf künftige Löhne haben. Für Ihre Analyse vergleichen Sie die Löhne zwischen Teilnehmern und Nichtteilnehmern des Beschäftigungsprogramms.

3 Daten

Die Daten, die unter anderem auch von Dehejia and Wahba (1999) und Dehejia and Wahba (2002) ausgewertet wurden, können unter `http://users.nber.org/~rdehejia/nswdata2.html` heruntergeladen werden. Verwenden Sie die Textdateien nswre74_treated.txt (185 observations) und psid_controls.txt (2490 observations).

3.1 Variablen

Der gesamte Datensatz besteht aus 2675 Beobachtungen. Die Zielvariable (oder abhängige Variable) ist *RE78*, die anderen Variablen sind die unabhängigen Variablen. Das Arbeitsmarktexperiment untersucht den Effekt des Beschäftigungsprogramms auf das Einkommen im Jahr 1978 unter Berücksichtigung der restlichen unabhängigen Variablen. Die Beschreibung der Variablen finden Sie in Tabelle 1.

Tabelle 1: Beschreibung der Variablen

Variablen	Beschreibung
treatment	Teilnahme am Beschäftigungprogramm (binär)
age	Alter in Jahren
education	Bildung in Jahren
black	Afroamerikaner (binär)
hispanic	Hispanoamerikaner (binär)
married	Verheiratet (binär)
no degree	Kein Highschool Abschluss (binär)
RE74	Einkommen im Jahr 1974 (US-Dollar)
RE75	Einkommen im Jahr 1975 (US-Dollar)
RE78	Einkommen im Jahr 1978 (US-Dollar)

4 Analyse

Diese Projektarbeit kann in vier Bereiche gegliedert werden: (i) Datenaufbereitung, (ii) Datensimulation, (iii) Implementierung des Schätzverfahrens und (iv) dessen Evaluierung.

4.1 Datenaufbereitung

Die Stichprobe für dieses Projekt beschränkt sich auf die Beobachtungen für 780 Afroamerikaner (*black*). Die Daten für die anderen Individuen sind nicht zu berücksichtigen. Neben der Variable *black* werden für die Analyse auch die Variablen *hispanic* und *RE74* nicht verwendet. Somit werden also insgesamt sieben Variablen benötigt.

4.2 Datensimulation

Sie führen eine Monte Carlo-Simulation durch, bei der die empirischen Daten insgesamt fünfzigmal simuliert werden. Simulieren Sie hier nur die sechs unabhängigen Variablen (nicht die Zielvariable *RE78*). Verwenden Sie für die binären Variablen *treatment*, *married* und *nodegree* die empirische Verteilung. Ziehen Sie dann die Variablen *age*, *education* und *RE75* aus einer multivariaten Normalverteilung. Die Mittelwerte und Kovarianzen für die multivariate Normalverteilung berechnen Sie anhand der empirischen Daten innerhalb der acht Gruppen der binären Variablen. Beachten Sie hierbei, dass bei den simulierten Daten die Minima und Maxima der empirischen Daten nicht überschritten werden und dass die simulierten Variablen *age* und *education* in ganzen Zahlen anzugeben sind. Diese (stark vereinfachte) Datensimulation basiert auf den Studien von Busso, DiNardo, and McCrary (2014) und

Bodory, Camponovo, Huber, and Lechner (2018), siehe Kapitel 4.2.

4.3 Implementierung des Schätzverfahrens

Sie können ein Algorithmus Ihrer Wahl implementieren. Die im Folgenden beschriebene Schätzmethode ist nur eine von vielen Möglichkeiten.

4.3.1 Algorithmus zur Schätzung des Effekts der Teilnahme am Beschäftigungsprogramm (inklusive Standardfehler)

Der Effekt $\hat{\theta}$ kann anhand der empirischen Daten folgendermassen geschätzt werden:

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} \big(\mu(1) - \mu(0)\big) \tag{1}$$

$$\mu(1) = Xb_1 \tag{2}$$

$$\mu(0) = Xb_0 \tag{3}$$

$$b_1 = (X_1'X_1)^{-1}X_1'y \tag{4}$$

$$b_0 = (X_0'X_0)^{-1}X_0'y. \tag{5}$$

Die Stichprobengrösse N ist 780. Die Matrix X hat die Dimension 780x6 (d.h. 780 Zeilen und 6 Spalten). Die erste Spalte beinhaltet einen Vektor mit Einsen, die Spalten 2-6 sind die unabhängigen Variablen *age*, *education*, *married*, *nodegree* und *RE75*. Die Variablen b_1 und b_0 sind Vektoren mit den OLS-Steigungskoeffizienten, wobei die Matrizen X_1 bzw. X_0 Teilmatrizen von X für die Teilnehmer bzw. Nichtteilnehmer des Beschäftigungsprogramms sind. Die Variable y steht für *RE78*.

Der Standardfehler von $\hat{\theta}$ kann mit der Standard-Bootstrapmethode ermittelt werden. Führen Sie 19 Bootstrap Replikationen ($B = 19$) für die Berechnung des Standardfehlers $\hat{\sigma}(\hat{\theta})$ durch, wobei bei jeder Replikation der Bootstrap Effekt $\hat{\theta}^b$ zu ermitteln ist.

$$\hat{\sigma}(\hat{\theta}) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^b - \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^b \right)^2}. \tag{6}$$

4.3.2 Algorithmus zur Schätzung simulierten Effekte inklusive Standardfehler

Die simulierten Effekte basieren auf den simulierten Daten. Die Formeln für die Berechnung lauten:

$$\hat{\theta}_{sim} = \frac{1}{N}\sum_{i=1}^{N}\big(\mu_{sim}(1) - \mu_{sim}(0)\big) \tag{7}$$

$$\mu_{sim}(1) = X_{sim}b_1 + u\sqrt{\frac{1}{N_1 - K}\sum_{i=1}^{N_1}(X_1 b_1 - y_1)^2} \tag{8}$$

$$\mu_{sim}(0) = X_{sim}b_0 + v\sqrt{\frac{1}{N_0 - K}\sum_{i=1}^{N_0}(X_0 b_0 - y_0)^2}. \tag{9}$$

X_{sim} sind die Matrizen X, die in jeder Simulation generiert werden. Die Vektoren u und v sind standardnormalverteilte Zufallsvariablen mit N Elementen. N_1 und N_0 sind Variablen für die Anzahl der Teilnehmer und Nichtteilnehmer des Beschäftigungsprogramms. Die Variable K gibt die Anzahl der Spalten von X wieder. y_1 und y_0 stehen für die Zielvariable $RE78$ mit den Beobachtungen der Teilnehmer und Nichtteilnehmer des Programms.

Berechnen Sie für jeden simulierten Effekt $\hat{\theta}_{sim}$ den Standardfehler $\hat{\sigma}(\hat{\theta}_{sim})$. Verwendung Sie hierfür die Formel von Gleichung 6.

4.4 Evaluierung des Schätzverfahrens

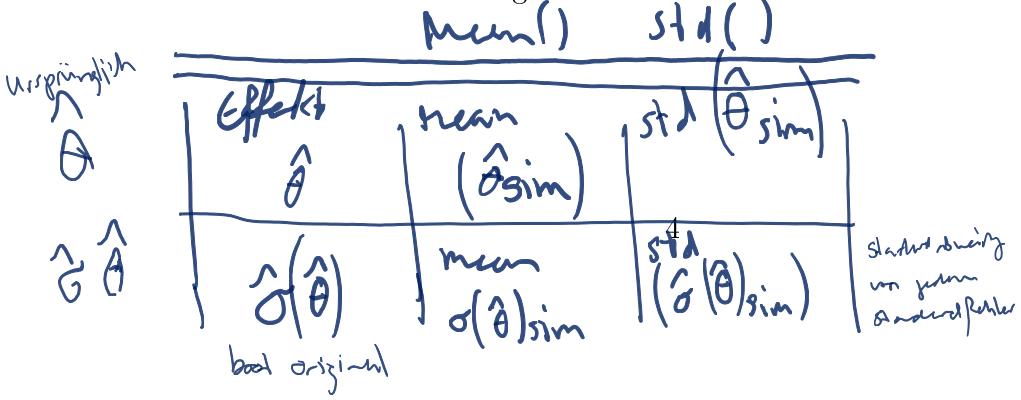
Erstellen Sie eine Tabelle mit den Mittelwerten und Standardabweichungen der Variablen *age*, *education*, *married*, *nodegree*, *RE75* und *RE78*, getrennt nach Programmteilnehmern und Nichtteilnehmern.

Generieren Sie eine Tabelle mit zwei Zeilen, die jeweils drei Statistiken je Zeile beinhaltet. In der ersten Zeile sind (i) der Effekt $\hat{\theta}$, (ii) der Mittelwert der simulierten Effekte $\hat{\theta}_{sim}$ und (iii) deren Standardabweichung anzuführen. In die zweite Zeile der Tabelle sind (i) der Standardfehler $\hat{\sigma}(\hat{\theta})$, (ii) der Mittelwert der simulierten Standardfehler $\hat{\sigma}(\hat{\theta}_{sim})$ sowie (iii) deren Standardabweichung einzutragen.

Berechnen Sie abschliessend die Überdeckungswahrscheinlichkeit (coverage probability), d.h. den Anteil der Simulationen, bei denen der Effekt $\hat{\theta}$ innerhalb der 95% Konfidenzintervalle um die simulierten Effekte $\hat{\theta}_{sim}$ liegt. Die Konfidenzintervalle CI werden folgendermassen berechnet:

$$CI = [\hat{\theta}_{sim} - 1.959964 * \hat{\sigma}(\hat{\theta}_{sim}); \hat{\theta}_{sim} + 1.959964 * \hat{\sigma}(\hat{\theta}_{sim})] \tag{10}$$

Kommentieren Sie Ihre Ergebnisse.



5 Programmierung

Der Programmcode soll alle wesentlichen Themenblöcke der Vorlesungen beinhalten. Dies schliesst die Erstellung von Kontrollstrukturen, Funktionen, grafischen Darstellungen, Tabellen, Zufallsvariablen, Statistiken und Algorithmen von Schätzverfahren ein. Versuchen Sie auch, neben der prozeduralen Programmierung (Funktionen) die objektorientierte Programmierung (Klassen) in Ihren Code einzubauen.

6 Dokumentation

Um den Programmcode lesbar zu machen, ist dieser mit Kommentaren zu versehen. Erstellen Sie bitte zusätzlich ein PDF-Dokument, das die einzelnen Programmabschnitte, Grafiken und Tabellen erläutert. Geben Sie in diesem PDF-Dokument auch die Aufgabenaufteilung innerhalb Ihrer Gruppe bekannt.

Literatur

Bodory, H., L. Camponovo, M. Huber, and M. Lechner (2018): "The Finite Sample Performance of Inference Methods for Propensity Score Matching and Weighting Estimators," *mimeo*.

Busso, M., J. DiNardo, and J. McCrary (2014): "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators," *Review of Economics and Statistics*, 96, 885–897.

Dehejia, R. H., and S. Wahba (1999): "Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programmes," *Journal of American Statistical Association*, 94, 1053–1062.

——— (2002): "Propensity Score Matching Methods for Non-Experimental Causal Studies," *Review of Economics and Statistics*, 84, 151–161.

LaLonde, R. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76, 604–620.