Hausarbeit
Data Analytics II: Empirische Wirtschaftsforschung FS20

Abgabe bis 1. Juni 2020

Sie können maximal 30 Punkte erreichen. Um die volle Punktzahl in einer Teilaufgabe zu erhalten, geben Sie, wo verlangt, kurze und präzise Antworten.
Die Abgabe erfolgt via Canvas und ist bis einschliesslich 1. Juni freigeschalten. Abgaben über andere Kanäle werden *nicht* berücksichtigt.
Eine vollständige Abgabe besteht aus einer "*Nachname_Hausarbeit.pdf*" Datei mit den Resultaten und einer "*Nachname_Code.R*" Datei mit dem kommentierten Code.

1 Oregon Health Insurance Experiment

Gerade in Zeiten von Corona wird in den USA wieder heftig über eine universelle Krankenversicherung diskutiert. Aber was ist überhaupt der Effekt einer Krankenversicherung auf Gesundheit? Diese Frage ist von höchster Relevanz für politische Entscheidungsträger. In diesem Teil verwenden wir echte Daten aus dem "Oregon Health Insurance Experiment" (OHIE), um dieser Frage nachzugehen.

Der Oregon Health Plan (OHP) ist eine stark subventionierte Krankenversicherung (0$ - 20$ Prämie pro Monat) für Niedrigverdiener im US Bundesstaat Oregon. Im Jahr 2008 gab es eine Lotterie, bei der 10,000 Plätze in dieser Versicherung verlost wurden. Allerdings nahmen nicht alle "Gewinner" die Möglichkeit auch an, sich zu versichern oder waren nicht zugangsberechtigt.[1]

Verwenden Sie in diesem Teil den Datensatz "OHIE_data.RData" mit folgenden Variablen:

- *winner*: Dummyvariable. Eins, wenn Person in der Lotterie gewonnen hatte.
- *insured*: Dummyvariable. Eins, wenn Person krankenversichert ist.
- *good_health*: Dummyvariable. Eins, wenn Person angibt bei guter Gesundheit zu sein.
- $female$: Dummyvariable. Eins, wenn Person weiblich ist.
- $hhinc\_pctfpl$: Haushaltseinkommen in Prozent der Armutsgrenze.
- *race*: Kategorievariable, die anzeigt zu welcher ethnischen Gruppe die Person gehört.

[1] Unter diesem Link finden Sie mehr Informationen bei Interesse.

1. Führen Sie eine univariate Regression mit *good_health* als abhängige und *insured* als erklärende Variable durch. Interpretieren Sie die Koeffizienten. Warum schätzt diese Regression höchstwahrscheinlich keinen kausalen Effekt? Nennen und erklären Sie dazu zwei Arten von Endogenität in diesem Kontext. (4 Punkte)

2. Diskutieren Sie kurz, ob die Lotterie des OHIE ein exogenes Instrument für Krankenversicherung generiert. (2 Punkte)

3. Erstellen Sie eine Kreuztabelle zwischen dem randomisierten Angebot eines OHP (*winner*) und dem Versichertenstatus (*insured*) und diskutieren Sie kurz, ob es sich dabei um ein relevantes/starkes Instrument handelt. (2 Punkte)

4. Schätzen Sie den kausalen Effekt von *insured* auf *good_health*, indem Sie *winner* als Instrument verwenden. Interpretieren Sie die Ergebnisse. Wie gross ist der Unterschied zu den OLS Ergebnissen? (2 Punkte)

5. Die Daten basieren auf einer nicht verpflichtenden Umfrage. Die Daten enthalten also nur Personen, die auf die Umfrage geantwortet haben. Diskutieren Sie, inwiefern Nichtbeantwortung der Umfrage ein Problem sein könnte und untersuchen Sie es unter Berücksichtigung der zusätzlichen Kontrollvariablen $female$, $hhinc\_pctfpl$ und $race$. (4 Punkte)

2 Simulation IV

In diesem Teil führen Sie eine Simulation durch, deren DGP an die Anwendung im ersten Teil angelehnt ist. Der Code *Simulation.R* implementiert dafür bereits folgende Schritte:

- Setze Anzahl der Beobachtungen auf $n = 1000$.
- Ziehe zwei korrelierte Fehlerterme u und v aus einer multivariaten Normalverteilung, $u, v \sim N(\mu, \Sigma)$ mit $\mu = [0 \;\; 0]$ und $\mathbf{\Sigma} = \begin{bmatrix} 100 & -7 \\ -7 & 1 \end{bmatrix}$
- Ziehe die Instrumentalvariable als Dummyvariable z mit $z \sim Binomial(1, 0.5)$.
- Generiere die endogene erklärende Variable Krankenversicherung, KV, in zwei Schritten:
  1. Generiere eine temporäre Variable, $temp = -1 + 1 \cdot z + v$
  2. Generiere die endogene Dummyvariable, $KV = 1(temp > 0)$.
- Generiere die abhängige Variable, $G = 50 + 10 \cdot KV + u$. Sie soll einen Gesundheitsindex darstellen, der Werte zwischen 0 und 100 annimmt.

Sie können *Simulation.R* als Grundlage zur Bearbeitung der folgenden Aufgaben verwenden:

6. Analysieren Sie die synthetische Stichprobe, die von *Simulation.R* erstellt wird: (5 Punkte)
   (a) Zeigen Sie geeignete deskriptive Statistiken von G, KV und z.
   (b) Berechnen und berichten Sie die Kovarianz von KV und u ($Cov(KV, u)$), um zu zeigen, dass MLR.4 in diesem DGP verletzt ist.

(c) Zeigen Sie die Ergebnisse der OLS Regression $KV = \pi_0 + \pi_1 z + \epsilon$. Interpretieren Sie π_1 und erklären Sie kurz, warum es sich bei z um ein starkes Instrument handelt.

(d) Zeigen Sie die Ergebnisse der OLS Regression $G = \beta_0 + \beta_1 KV + u$. Wie gross ist die Abweichung vom wahren Wert?

(e) Zeigen Sie die Ergebnisse der IV Regression, in der Sie z als Instrument für KV verwenden. Wie gross ist die Abweichung vom wahren Wert?

7. Führen Sie eine Simulationsstudie durch. Jedem von Ihnen wurde dafür eine Stichprobengrösse zugeteilt. Diese finden Sie neben Ihrer Matrikelnummer in Spalte B des Google Spreadsheets unter diesem Link. (8 Punkte)

   (a) Nutzen Sie einen for-loop, um 1000 Stichproben der Ihnen zugewiesenen Stichprobengrösse aus dem DGP zu ziehen und speichern Sie jeweils die Schätzung von β_1 aus der OLS und der IV Regression.

   (b) Zeigen Sie in zwei separaten oder einem gemeinsamen Histogramm die Verteilung der beiden Schätzverteilungen zusammen mit einer vertikalen Linie auf dem wahren Wert von $\beta_1 = 10$.[2]

   (c) Berechnen und berichten Sie die Verzerrung der beiden Schätzer als die Differenz des Durchschnitts der Schätzverteilung und dem wahren Wert. Tragen Sie die Werte ausserdem im Google Spreadsheet in den Spalten C und D neben Ihrer Matrikelnummer ein.

   (d) Welchen der beiden Schätzer würden Sie in diesem Setting bevorzugen und warum? Erklären Sie kurz.

8. Wiederholen Sie die Aufgaben 7a)-c) mit einem schwachen Instrument. Spezifizieren Sie dazu $temp = -1 + 0.1z + v$ in der Prozedur. Was passiert nun mit der Schätzverteilung? Beschreiben und erklären Sie kurz. Tragen Sie ausserdem die Verzerrung des IV Schätzers im Google Spreadsheet in Spalte E neben Ihrer Matrikelnummer ein. (3 Punkte)

Viel Erfolg!

[2] Falls in einigen Stichproben kein Parameter geschätzt werden konnte, wird R Ihnen einen fehleden Wert (`NA`) ausgeben. Ignorieren Sie diese Stichproben in der Grafik und in allen folgenden Auswertungen.