PS2 for Econometrics 101, Warwick Econ Ph.D

Exercise 1

Consider the following joint probability distribution for the discrete random variables $X$ and $Y$:

<table>
<thead>
<tr>
<th></th>
<th>Y 10</th>
<th>Y 20</th>
<th>Y 30</th>
</tr>
</thead>
<tbody>
<tr>
<td>X 10</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>X 20</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
</tr>
</tbody>
</table>

1. Compute the marginal probability distribution of $X$, i.e., the probability that $X$ is equal to 10 and 20. Compute the marginal probability distribution of $Y$, i.e., the probability that $Y$ is equal to 10, 20, and 30.

2. Calculate $E(X)$ and $E(Y)$.

3. Calculate $E(Y|X=10)$ and $E(Y|X=20)$. Give an explicit formula for $E(Y|X)$.

Solution

1. Let $P(X = x)$ and $P(Y = y)$ denote marginal distributions for $X$ and $Y$, then

<table>
<thead>
<tr>
<th></th>
<th>Y 10</th>
<th>Y 20</th>
<th>Y 30</th>
<th>$P(X = x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>X 10</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>X 20</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
<td>0.6</td>
</tr>
</tbody>
</table>

| $P(Y = y)$ | 0.5 | 0.3 | 0.2 |

2. $E(X) = 0.4 \times 10 + 0.6 \times 20 = 16$

   $E(Y) = 0.5 \times 10 + 0.3 \times 20 + 0.2 \times 30 = 17$

3. $E(Y|X=10) = 10 \times 0.5 + 20 \times 0.25 + 30 \times 0.25 = 17.5$

   $E(Y|X=20) = 10 \times 1/2 + 20 \times 1/3 + 30 \times 1/6 = 16.666$

   $E(Y|X) = 17.5 \times 1\{X=10\} + 16.666 \times 1\{X=20\}$

Exercise 2

Let $X$ and $Y$ be two discrete random variables taking values $x_i \in S(X) = \{1, 2\}$ and $y_j \in S(Y) = \{1, 2\}$. The joint probability distribution of $X$ and $Y$ is given by

<table>
<thead>
<tr>
<th></th>
<th>Y 1</th>
<th>Y 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>X 1</td>
<td>1/8</td>
<td>2/8</td>
</tr>
<tr>
<td>X 2</td>
<td>2/8</td>
<td>3/8</td>
</tr>
</tbody>
</table>
Find $V(X|Y = 1)$ and $V(X|Y = 2)$. Give a formula for $V(X|Y)$.

**Solution**

Recall the definition of conditional variance for a discrete random variable. For any $y_j \in S(Y)$,

$$V(X|Y = y_j) = \sum_{x_i \in S(X)} [x_i - E(X|Y = y_j)]^2 P(X = x_i|Y = y_j).$$

As for the nonconditional variance, it can also be written as

$$V(X|Y = y_j) = E(X^2|Y = y_j) - [E(X|Y = y_j)]^2,$$

where

$$E(X|Y = y_j) = \sum_{x_i \in S(X)} x_i P(X = x_i|Y = y_j)$$

$$E(X^2|Y = y_j) = \sum_{x_i \in S(X)} x_i^2 P(X = x_i|Y = y_j),$$

and

$$P(X = x_i|Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}.$$

The marginal probabilities of $Y$ can be computed from the joint probabilities, and they are

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$Pr(X = x_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1/8</td>
<td>2/8</td>
</tr>
<tr>
<td>2</td>
<td>2/8</td>
<td>3/8</td>
</tr>
<tr>
<td>$Pr(Y = y_j)$</td>
<td>3/8</td>
<td>5/8</td>
</tr>
</tbody>
</table>

Finally,

$$E(X|Y = 1) = \frac{1}{3} + \frac{2}{3} = \frac{5}{3},$$

$$E(X|Y = 2) = \frac{2}{5} + \frac{3}{5} = \frac{8}{5},$$

$$E(X^2|Y = 1) = \frac{1}{3} + \frac{2}{3} = 3,$$

$$E(X^2|Y = 2) = \frac{2}{5} + \frac{3}{5} = \frac{14}{5},$$

and

$$V(X|Y = 1) = 3 - \frac{25}{9} = \frac{2}{9},$$

$$V(X|Y = 2) = \frac{14}{5} \cdot \frac{6}{25} = \frac{6}{25}.$$

Therefore, $V(X|Y) = 2/9 \times 1\{Y = 1\} + 6/25 \times 1\{Y = 2\}$. 

2
Exercise 3

Let $T$ be the region $T = \{(x, y)^t \in \mathbb{R}^2 : 0 \leq x \leq 2 \text{ and } 0 \leq y \leq x\}$, and let $X$ and $Y$ be random variables with joint density function

$$f_{X,Y}(x, y) = \begin{cases} 
  k x^2 y^2 & \text{for } (x, y)^t \in T \\
  0 & \text{otherwise}
\end{cases}.$$

1. Find the value of $k$.

2. Find the marginal distributions of $X$ and $Y$. Are $X$ and $Y$ independent? Explain your answer.

3. Find $E(X)$ and $E(Y)$.

4. Find $E(XY)$ and $\text{cov}(X, Y)$.

Solution

1. The function $f_{X,Y}(x, y)$ is a proper pdf if

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx \, dy = 1.$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx \, dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} k x^2 y^2 1(0 \leq x \leq 2) 1(0 \leq y \leq x) \, dx \, dy$$

$$= k \int_{-\infty}^{\infty} x^2 1(0 \leq x \leq 2) \left( \int_{-\infty}^{\infty} y^2 1(0 \leq y \leq x) \, dy \right) \, dx$$

$$= k \int_{0}^{2} x^2 \left( \int_{0}^{x} y^2 \, dy \right) \, dx$$

$$= k \int_{0}^{2} x^2 \left( \frac{1}{3} y^3 \right)_{0}^{x} \, dx$$

$$= k \int_{0}^{2} \frac{1}{3} x^5 \, dx = \frac{k}{18} x^6 \bigg|_{0}^{2} = \frac{32}{9} k,$$

that is, $k = \frac{9}{32}$.

2. $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy = \int_{-\infty}^{\infty} \frac{9}{32} x^2 y^2 1(0 \leq x \leq 2) 1(0 \leq y \leq x) \, dy$

$$= \frac{9}{32} x^2 1(0 \leq x \leq 2) \int_{-\infty}^{x} y^2 1(0 \leq y \leq x) \, dy$$

$$= \frac{9}{32} x^2 1(0 \leq x \leq 2) \int_{0}^{x} y^2 \, dy$$

$$= \frac{9}{32} x^2 1(0 \leq x \leq 2) \frac{1}{3} x^3 = \begin{cases} 
  \frac{3}{32} x^5 & \text{if } 0 \leq x \leq 2 \\
  0 & \text{otherwise}
\end{cases}.$$

To find $f_Y(y)$, first notice that the region $\mathcal{T}$ can be also described as $\mathcal{T} = \{(x, y)' \in \mathbb{R}^2: 0 \leq y \leq 2 \text{ and } y \leq x \leq 2\}$ (you can see this by drawing the areas in $\mathbb{R}^2$ for which these inequalities are satisfied). Then

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx = \int_{-\infty}^{\infty} \frac{9}{32} x^2 y^2 1_{y \leq x \leq 2} 1_{0 \leq y \leq 2} dy$$

$$= \frac{9}{32} y^2 1_{0 \leq y \leq 2} \int_{-\infty}^{\infty} x^2 1_{y \leq x \leq 2} dx$$

$$= \frac{9}{32} y^2 1_{0 \leq y \leq 2} \int_{y}^{2} x^2 dx$$

$$= \frac{9}{32} y^2 1_{0 \leq y \leq 2} \frac{1}{3} (8 - y^2) = \begin{cases} 
\frac{24y^2 - 3y^5}{32} & \text{if } 0 \leq y \leq 2 \\
0 & \text{otherwise} 
\end{cases}.$$ 

As $f_{X,Y}(x, y) \neq f_X(x)f_Y(y)$, $X$ and $Y$ are not independent.

3. 

$$E(X) = \int_{-\infty}^{\infty} xf_X(x) dx = \int_{0}^{2} \frac{3}{32} x^6 dx = \frac{12}{7},$$

$$E(Y) = \int_{-\infty}^{\infty} yf_Y(y) dy = \int_{0}^{2} \frac{24y^3 - 3y^6}{32} dy = \frac{9}{7}.$$ 

4. We have

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf_{X,Y}(x, y) dx dy = \frac{9}{4},$$

so the covariance is

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y) = \frac{9}{196}.$$ 

**Exercise 4: Some more properties of conditional expectations and variances**

Let $X$ and $Y$ be discrete random variables taking respectively $n$ and $m$ values denoted $x_1, ..., x_n$ and $y_1, ..., y_m$.

a) Show that

$$E(Y|X = x_k) = \frac{E(Y1\{X = x_k\})}{P(X = x_k)}$$

**Solution**

4
\[
\frac{E(Y\{X = x_k\})}{P(X = x_k)} = \frac{1}{P(X = x_k)} \sum_{1 \leq i \leq n, 1 \leq j \leq m} y_j 1\{x_i = x_k\} P(Y = y_j, X = x_i)
\]

\[
= \frac{1}{P(X = x_k)} \sum_{j=1}^{m} y_j \sum_{i=1}^{n} 1\{x_i = x_k\} P(Y = y_j, X = x_i)
\]

\[
= \frac{1}{P(X = x_k)} \sum_{1 \leq j \leq m} y_j P(Y = y_j, X = x_k)
\]

\[
= \sum_{1 \leq j \leq m} y_j \frac{P(Y = y_j, X = x_k)}{P(X = x_k)}
\]

\[
= \sum_{1 \leq j \leq m} y_j P(Y = y_j | X = x_k)
\]

\[
= E(Y | X = x_k).
\]

b) Show that if \(Y \perp X\), \(E(Y|X) = E(Y)\).

Solution

\[
E(Y|X) = \sum_{i=1}^{n} 1\{X = x_i\} E(Y|X = x_i).
\]

Because \(Y \perp X\),

\[
E(Y|X = x_i) = \sum_{j=1}^{m} y_j P(Y = y_j | X = x_i)
\]

\[
= \sum_{j=1}^{m} y_j P(Y = y_j)
\]

\[
= E(Y).
\]

Therefore,

\[
E(Y|X) = \sum_{i=1}^{n} 1\{X = x_i\} E(Y)
\]

\[
= E(Y) \sum_{i=1}^{n} 1\{X = x_i\}
\]

\[
= E(Y).
\]

c) Let \(e\) be a random variable such that \(E(e) = 0\). Show that if \(e \perp X\), \(E(e|X) = 0\).

Then, show that if \(E(e|X) = 0\), \(cov(e, X) = 0\). What is the strongest assumption: \(e \perp X\), \(E(e|X) = 0\), or \(cov(e, X) = 0\)? What is the most interpretable assumption?

Solution

If \(e \perp X\), it follow from question b) that \(E(e|X) = E(e) = 0\). This proves the first statement.

To prove the second one, it suffices to notice that \(cov(e, X) = E(eX) = E(XE(e|X)) = 0\).
This proves that $e \perp X \Rightarrow E(e|X) = 0 \Rightarrow \text{cov}(e,X) = 0$. The strongest assumption is therefore $e \perp X$, while the weakest one is $\text{cov}(e,X) = 0$. But the most interpretable is $e \perp X$.

d) Prove Theorem 3.1.10.

Solution

\[
V(f(X)Y + g(X)|X) = E((f(X)Y + g(X))^2|X) - (E(f(X)Y + g(X)|X))^2 \\
= E(f(X)^2Y^2 + 2f(X)g(X)Y + g(X)^2|X) - (f(X)E(Y|X) + g(X))^2 \\
= f(X)^2E(Y^2|X) + 2f(X)g(X)E(Y|X) + g(X)^2 \\
- f(X)^2E(Y|X)^2 - 2f(X)g(X)E(Y|X) - g(X)^2 \\
= f(X)^2E(Y^2|X) - f(X)^2E(Y|X)^2 \\
= f(X)^2V(Y|X).
\]

Exercise 5: Measurement error is an issue on the rhs, not on the lhs

Assume you would like to know the coefficient of the univariate regression of $Y$ on $X$ (irrespective of whether it can receive a causal interpretation or not).

a) Unfortunately, $Y$ is not well measured in your data, and you only observe $Y^* = Y + u$. $u$ is just some noise which we call a measurement error. For instance, when asked about their total household income, people often fail to answer precisely, say because they do not remember the exact wage of their partner. So the answer they give you, $Y^*$ can be written as $Y$, their true household income, plus the mistake they make in evaluating it, $u$. If $u$ has mean 0 and is uncorrelated with $X$ and $Y$ (random measurement error), show that the population coefficient of the regression of $Y^*$ on $X$ is actually equal to the population coefficient of the regression of $Y$ on $X$.

Solution

The coefficient of the regression of $Y^*$ on $X$ is

\[
\frac{\text{cov}(Y^*,X)}{V(X)} = \frac{\text{cov}(Y,X) + \text{cov}(u,X)}{V(X)} \\
= \frac{\text{cov}(Y,X)}{V(X)},
\]

which is the coefficient of the regression of $Y$ on $X$.

b) Now assume that $Y$ is properly measured, but not $X$: we only get to observe $X^* = X + v$ where $v$ is some noise. If $v$ has mean 0 and is uncorrelated with $X$ and $Y$ (random measurement error), show that the population coefficient of the regression of $Y$ on $X^*$ is not equal to the population coefficient of the regression of $Y$ on $X$.

Solution
The coefficient of the regression of $Y$ on $X^*$ is

$$\frac{\text{cov}(Y, X^*)}{V(X^*)} = \frac{\text{cov}(Y, X) + \text{cov}(Y, v)}{V(X) + V(v)} = \frac{\text{cov}(Y, X)}{V(X) + V(v)},$$

which is lower (in absolute value) than the coefficient of the regression of $Y$ on $X$.

(c) Assume you regress the number of children in an household on its total income. If there is measurement error for total income in your data, will the absolute value of your regression coefficient be lower or lower than the absolute value of the true regression coefficient?

**Solution**

The absolute value of the regression coefficient will be lower than that of the regression coefficient of $Y$ on $X$ (cf. question b)). Measurement error for explanatory variables biases their coefficients towards zero.

**Exercise 6: Monte-Carlo simulations**

Assume that $X$, $u$ and $Y$ are 3 random variables such that $X$ follows a uniform distribution on $(0, 20)$, $u$ follows a $N(0, 50^2)$ distribution and is independent of $X$, and $Y = 2X + u$.

(a) What is the value of $\beta$, the population coefficient of a regression of $Y$ on $X$?

**Solution**

$$\beta = \frac{\text{cov}(Y, X)}{V(X)} = \frac{\text{cov}(2X + u, X)}{V(X)} = 2,$$

because $u \perp X \Rightarrow \text{cov}(e, X) = 0$ (cf. exercise 4.c).

Generate a data set containing 1000 realizations of random variables $X_i$, $u_i$, and $Y_i$ following the same distribution as $X$, $u$, and $Y$, respectively.

(b) Regress $Y$ on $X$. Is the estimator you obtain close to the true value of the coefficient?

(c) Generate a new data set containing 10 realizations of $X_i$, $u_i$, and $Y_i$. Is the estimator you obtain close to the true value of the coefficient?

(d) Generate a new data set containing 100 000 realizations of $X_i$, $u_i$, and $Y_i$. Is the estimator you obtain close to the true value of the coefficient?

(e) Explain why b), c), and d) are a good illustration of Theorem ?? in the notes.

**Solution**

Theorem ?? says that when the sample size gets large, $\hat{\beta}$ gets close to $\beta$ with a very large probability. Here, with only 10 observations $\hat{\beta}$ is pretty far from 2. With 100 observations things are much better, and with 100 000 observations the estimator is essentially equal to the true parameter up to a very small margin of error.
f) Redo once again the exercise with 1000 observations, but with \( u_i \) following a \( N(0, 50^2) \) distribution. Is the estimator you obtain closer or further from 2 than in question b)? Explain why this illustrates Theorem 5.2.2.

**Solution**

As \( u \) is independent of \( X \), it follows from exercise 4.c) that

\[
E(Y|X) = 2X + E(u|X) = 2X + E(u) = 2X.
\]

Similarly,

\[
V(u|X) = E(u^2|X) - E(u|X)^2 = E(u^2) - E(u)^2 = V(u).
\]

The last but one equality follows from the fact that \( u \) is independent of \( X \) (which in turn implies \( u^2 \) is independent of \( X \)) and from exercise 4.c). Therefore the regression decomposition of \( Y \) with respect to \( X \) satisfies all the conditions under which Theorem ?? applies (linearity of the CEF, homoscedasticity). This theorem says that the asymptotic variance of a univariate regression coefficient is a linear function of the variance of the error term of the regression. When changing the distribution of \( u_i \) from \( N(0, 50^2) \) to \( N(0, 500^2) \), we multiply the variance of the error term by 100. We therefore also multiply the asymptotic variance of the regression coefficient by 100, which means we divide by 100 the statistical precision of our estimator.

That is why the regression coefficient is further away from 2 in f) than in b).

g) Let’s get back to \( u_i \) following a \( N(0, 50^2) \) distribution. Generate a data set with 600 triples of 1000 realizations of variables \((X^j_i, e^j_i, Y^j_i)\) \(1 \leq j \leq 600\). Run the 600 regressions of \( Y^j_i \) on \( X^j_i \), compute \( \sqrt{n} (\hat{\beta} - 2) \) for each regression, and use the kdensity command to plot a graph of the density of the 600 realizations of \( \sqrt{n}(\hat{\beta} - 2) \) and the density of a normal distribution. Explain why this illustrates Theorem ??.

**Solution**

The density (histogram) of the 600 \( \sqrt{n}(\hat{\beta} - 2) \) is bell-shaped, and looks like the density of a normal distribution. This is consistent with Theorem ??, which says that \( \sqrt{n}(\hat{\beta} - 2) \rightarrow N(0, \sigma^2) \).

**Exercise 7: To proxy or not to proxy?**

You are interested in the effect of education (\( S_i \)) on wages (\( Y_i \)). Assume the LCEM assumption is verified, CIA holds after conditioning for IQ at age 6 (\( I_i \)), and \( E(\eta_i|I_i) = \lambda I_i \). Under those three assumptions, you can follow the same steps as for the proof of Theorem ?? in the notes to show that

\[
Y_i = \alpha + \rho S_i + \lambda I_i + \varepsilon_i,
\]

with \( \text{cov}(S_i, \varepsilon_i) = \text{cov}(I_i, \varepsilon_i) = 0 \). Theorem ?? also shows that if you were to regress \( Y_i \) on \( S_i \) and \( I_i \), the coefficient of \( S_i \) would pick up \( \rho \), the causal effect of education on wages. Unfortunately, in your data you only observe IQ at age 25 (\( I^*_i \)). It might sound reasonable to
include $I^*_i$ in the regression, as a proxy for $I_i$. To simplify, assume that $I^*_i$ is a deterministic linear function of $I_i$ and $S_i$:

$$I^*_i = \pi_0 + \pi_1 I_i + \pi_2 S_i. \quad (0.0.3)$$

$\pi_2$ represents the causal effect of schooling on IQ at age 25.

a) Should we expect $\pi_2$ to be positive or negative?

Solution

$\pi_2$ is probably positive. We expect education to increase IQ at age 25, conditional on IQ at 6.

b) Use (0.0.2) and (0.0.3) to show that

$$Y_i = \alpha - \pi_1 \frac{\pi_0}{\pi_1} + \left( \rho - \pi_1 \frac{\pi_2}{\pi_1} \right) S_i + \frac{\lambda}{\pi_1} I^*_i + \epsilon_i \quad (0.0.4)$$

Solution

(0.0.3) is equivalent to

$$I_i = \frac{1}{\pi_1} I^*_i - \frac{\pi_0}{\pi_1} - \frac{\pi_2}{\pi_1} S_i.$$ 

Plugging this into (0.0.2) yields the result.

c) Use (0.0.4) to show that $\beta$, the population coefficient of $S_i$ in a regression of $Y_i$ on $S_i$ and $I^*_i$ is equal to $\rho - \pi_1 \frac{\pi_2}{\pi_1}$.

Solution

$$\beta = \frac{\text{cov}(Y_i, \tilde{S}_i)}{V(\tilde{S}_i)} = \frac{\text{cov}(\alpha - \frac{\pi_0}{\pi_1}, \tilde{S}_i) + \left( \rho - \pi_1 \frac{\pi_2}{\pi_1} \right) S_i + \frac{\lambda}{\pi_1} I^*_i + \epsilon_i, \tilde{S}_i)}{V(\tilde{S}_i)} = \rho - \pi_1 \frac{\pi_2}{\pi_1}.$$ 

The first equality follows from the Frisch-Vaugh theorem. The second one follows from Equation (0.0.4). The third one follows from the three following facts. $\text{cov}(\alpha - \frac{\pi_0}{\pi_1}, \tilde{S}_i) = 0$ because $\alpha - \frac{\pi_0}{\pi_1}$ is a constant. Then, let $S_i = \delta_0 + \delta_1 I^*_i + \tilde{S}_i$ be the regression decomposition of $S_i$ on $I^*_i$ and a constant. $\text{cov}(S_i, \tilde{S}_i) = \text{cov}(\delta_0 + \delta_1 I^*_i + \tilde{S}_i, \tilde{S}_i) = V(\tilde{S}_i)$, because $I^*_i$ and $\tilde{S}_i$ are uncorrelated by construction. For the same reason, $\text{cov}(I^*_i, \tilde{S}_i) = 0$. Finally, one can follow the same steps as in the proof of Theorem ?? in the notes to show that $\text{cov}(\epsilon_i, \tilde{S}_i) = 0$.

d) Will this regression under or overestimate $\rho$?

Solution

It seems likely that IQ at age 6 is positively correlated with wages ($\lambda > 0$), that IQ at age 6 is positively correlated with IQ at age 25 ($\pi_1 > 0$), and that education has a positive impact on IQ at age 25 ($\pi_2 > 0$). Therefore this regression will probably underestimate $\rho$.

e) Explain intuitively why something goes wrong when we include $I^*_i$ in the regression.
Solution

IQ at age 25 is a function of both IQ at age 6 and schooling. So when we include it in the regression, it allows controlling for the effect of IQ at age 6, which is good, but it also captures part of the effect attributable to schooling. As a rule, you should try not to include as a control in a regression a variable which could be an outcome. IQ at age 25 can be seen as a function of education, and it could actually be an interesting question to measure the effect of $S_i$ on $I^*_i$, so $I^*_i$ could be an outcome of the regression. Variables which can be included as controls are variables which are unlikely to be affected by the treatment. Here, gender, race, or parents occupation could be included as controls, because it is (almost) impossible that education has an effect on race and gender, while it seems unlikely that children education affects parents occupation.

f) Is the curse worse than the disease? In other words, is this coefficient likely to be further away from $\rho$ than the coefficient from an univariate regression of $Y_i$ on $S_i$?

Solution

Following the 2nd omitted variable formula, the coefficient of the univariate regression is $\rho + \lambda \mu$ where $\mu$ is the coefficient of a regression of $I_i$ on $S_i$. It sounds reasonable to assume that $\pi_2$ is small relative to $\pi_1$: the effect of IQ at age 6 on IQ at age 25 is probably larger than the effect of education. It also sounds reasonable to assume that $I_i$ and $S_i$ are fairly strongly correlated, which implies $\mu$ is probably large. As a result, we expect the second coefficient to be closer to $\rho$ than the first one. The cure it not perfect, but it still better than the disease. So here, as an exception, we might still want to include a variable which could be the outcome of our regression as a control, knowing that this will not be perfect, but still better than what we would get without it.

Exercise 8: Why you should never use outcomes as controls

Assume you are interested in the effect of a training program for unemployed on the probability they find a job in less than 6 months. On that purpose you run a randomized experiment in which some unemployed are randomly offered to participate in the training, while others are randomly excluded from it. Let $Y_i$ be a dummy for whether an unemployed had found a job in less than 6 months, let $S_i$ be a dummy for whether she had applied for more than 10 jobs, before she found a job for those who found a job, and before the end of the 6 months for those who did not. Finally, let $D_i$ be a dummy for whether the unemployed was assigned to the treatment or to the control group.

1) To measure the effect of the treatment on the probability to find a job in less than 6 months, you can run a regression of $Y_i$ on a constant and $D_i$.

a) Give the formula for the population coefficient of $D_i$ in this regression (as the model is saturated, you have to do better than writing this coefficient as a function of the covariance between $Y_i$ and $D_i$)
b) Show that because treatment was randomly allocated, this coefficient captures the average treatment effect.

c) Assume the estimator of this coefficient is 0.1. Interpret this value.

**Solution**

a) \( \beta = E(Y_i|D_i = 1) - E(Y_i|D_i = 0) \).

b) \( E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = E(Y_{i1}|D_i = 1) - E(Y_{i0}|D_i = 0) = E(Y_{i1}) - E(Y_{i0}) = E(Y_{i1} - Y_{i0}) \).

c) The program increases the share of unemployed who find a job in less than 6 months by 10 percentage points.

2) Assume you also want to measure the effect of the program on the probability that unemployed apply to more than 10 jobs.

a) Which extremely simple regression should you run?

b) The treatment increases the share of unemployed who apply to 10 jobs or more by 20 percentage points.

3) After seeing these results, you might form the following hypothesis: maybe the reason why the training helps unemployed people to find a job faster is not because they learn something useful during the training which makes them more employable (human capital channel), but because the training increases their motivation to find a job and therefore make them apply to more jobs (search effort channel). If the second hypothesis is the correct one, as a policy maker you could save a lot of money by replacing these long and complicated training programs by simple coaching programs intended at motivating unemployed so they increase their search effort.

To test this hypothesis, a simple idea would be to regress \( Y_i \) on a constant, \( D_i \), and \( S_i \), and to see if including \( S_i \) kills the effect of \( D_i \) you found in the first regression without \( S_i \). This could be a good indication that the effect of \( D_i \) goes through the fact it increases search effort, and once you account for search effort there is no effect anymore.

The purpose of this question is to show that doing this would actually be a very bad idea. It follows from the “regression as matching” Theorem that the population coefficient of \( D_i \) in this regression is equal to

\[
w(E(Y_i|D_i = 1, S_i = 0) - E(Y_i|D_i = 0, S_i = 0)) + (1-w)(E(Y_i|D_i = 1, S_i = 1) - E(Y_i|D_i = 0, S_i = 1)),
\]
where \( w \) is a weight included between 0 and 1. We are going to show that this coefficient could be equal to 0 even if the training has an effect on unemployment which does not go through the search effort channel.

On that purpose, we are gonna make an unrealistic assumption, just to simplify the discussion: we could reach to the same conclusions than the one we will obtain under more realistic assumptions. We are going to assume that applying to 10 jobs versus 9 only increases your probability of finding a job, but applying to 11, 12,... versus 10 jobs has no effect on your probability of finding a job, and that applying to 0, 1,...,8 versus 9 jobs also does not affect your probability to find a job. Your search effort impacts your probability of finding a job only if you cross the threshold of 10 applications.

It follows from the previous question that receiving the treatment increases the share of people who apply to at least 10 jobs. Under the assumption that someone who applies for at least 10 jobs without receiving the training would also apply to at least 10 jobs if she receives the training, we can partition the population into three types: the unmotivated guys (\( U \)), who will apply to less than 10 jobs even if they receive the training, the normally motivated guys (\( NM \)), who will apply to less than 10 jobs if they do not receive the training, but will apply to more than 10 jobs if they receive it, and the motivated guys (\( M \)) who will apply to more than 10 jobs irrespective of whether they receive the training or not.

a) Show that

\[
E(Y_i|D_i = 1, S_i = 0) - E(Y_i|D_i = 0, S_i = 0) = E(Y_{i1} - Y_{i0}|U) - \frac{P(NM)}{P(U) + P(NM)}(E(Y_{i0}|NM) - E(Y_{i0}|U))
\]

and that

\[
E(Y_i|D_i = 1, S_i = 1) - E(Y_i|D_i = 0, S_i = 1) = E(Y_{i1} - Y_{i0}|M) - \frac{P(NM)}{P(M) + P(NM)}(E(Y_{i1}|M) - E(Y_{i1}|NM)).
\]

Hint: which types (\( U \), \( NM \), and \( M \)) are in population satisfying \( \{D_i = 1, S_i = 0\} \)? Same question for the populations satisfying \( \{D_i = 0, S_i = 0\} \), \( \{D_i = 1, S_i = 1\} \), and \( \{D_i = 1, S_i = 1\} \).

b) Explain why \( E(Y_{i1} - Y_{i0}|U) \) and \( E(Y_{i1} - Y_{i0}|M) \) are measures of the effect of the training on employment which cannot be driven by search effort.

c) Use the two equations derived in the previous question to explain why both \( E(Y_i|D_i = 1, S_i = 0) - E(Y_i|D_i = 0, S_i = 0) \) and \( E(Y_i|D_i = 1, S_i = 1) - E(Y_i|D_i = 0, S_i = 1) \) could be equal to 0, even if both \( E(Y_{i1} - Y_{i0}|U) \) and \( E(Y_{i1} - Y_{i0}|M) \) are strictly positive.

d) Send me a 10 lines email explaining why you should not use outcome variables as controls in a regression.

**Solution**

a) \( E(Y_i|D_i = 1, S_i = 0) - E(Y_i|D_i = 0, S_i = 0) \) compares the share of unemployed who had found a job in less than 6 months, among those whose received the treatment and applied to
less than 10 jobs, and among those who did not receive the treatment and applied to less than
10 jobs. The first group one only includes unmotivated guys, while the second one includes
unmotivated guys and normally motivated guys. Therefore, one can rewrite
\[
E(Y_i|D_i = 1, S_i = 0) - E(Y_i|D_i = 0, S_i = 0)
\]
\[
= E(Y_{i1}|U) - E(Y_{i0}|NM or U)
\]
\[
= E(Y_{i1}|U) - \frac{P(NM)}{P(U) + P(NM)}E(Y_{i0}|NM) - \frac{P(U)}{P(U) + P(NM)}E(Y_{i0}|U)
\]
\[
= E(Y_{i1} - Y_{i0}|U) - \frac{P(NM)}{P(U) + P(NM)}(E(Y_{i0}|NM) - E(Y_{i0}|U)).
\]
\[
E(Y_i|D_i = 1, S_i = 1) - E(Y_i|D_i = 0, S_i = 1)
\]
\[
= E(Y_{i1}|M or NM) - E(Y_{i0}|M)
\]
\[
= \frac{P(M)}{P(M) + P(NM)}E(Y_{i1}|M) + \frac{P(NM)}{P(M) + P(NM)}E(Y_{i1}|NM) - E(Y_{i0}|M)
\]
\[
= E(Y_{i1} - Y_{i0}|M) - \frac{P(NM)}{P(M) + P(NM)}(E(Y_{i1}|M) - E(Y_{i1}|NM)).
\]
b) Irrespective of whether they receive the treatment or not, the motivated guys will always
apply to more than 10 jobs. Therefore, under our unrealistic assumption, \(E(Y_{i1} - Y_{i0}|M)\)
cannot come from the fact that treatment increases their search effort. The same reasoning
applies to the unmotivated guys.
c) \(E(Y_{i0}|NM) - E(Y_{i0}|U)\) is likely to be positive: normally motivated guys are more motivated
than the unmotivated guys, so if both groups do not receive the treatment, normally motivated
guys are more likely to find a job. Therefore, \(E(Y_i|D_i = 1, S_i = 0) - E(Y_i|D_i = 0, S_i = 0)\)
could be equal to 0, even if \(E(Y_{i1} - Y_{i0}|U)\) is strictly positive. Similarly, \(E(Y_i|D_i = 1, S_i = 1) - E(Y_i|D_i = 0, S_i = 1)\)
could be equal to 0 even if \(E(Y_{i1} - Y_{i0}|M)\) is strictly positive, because \(E(Y_{i1}|M) - E(Y_{i1}|NM)\) is likely to be positive.

**Exercise 9: Modelling the probability of arrest**

Type:
```
scc install bcuse
```
```
bcuse grogger
```
in Stata to download the grogger data set. This data contains information on 2725 american
males observed in 1986. In the data, the variable narr86 is equal to the number of times each
of these men has been arrested by the police in 1986. Use it to create a dummy equal to to 1
if someone was arrested at least once by the police at some point in 1986, which we will call arr86. This is the variable we will try to model. As explanatory variables, we will use their income in 1986 (inc86), an age dummy equal to 1 for those born before 1960 (born60), and one dummy for black people (black).

a) Our dependent variable is binary. Does the homoscedasticity assumption make sense? Which standard errors should we use for our regression coefficients?

Solution

No, homoscedasticity does not make sense in this context. Even assuming that $E(Y_i|X_i)$ is linear, which implies that $E(e_i|X_i) = 0$,

$$E(e_i^2|X_i) = V(e_i|X_i) = V(Y_i - X_i'\beta|X_i) = V(Y_i|X_i) = P(Y_i = 1|X_i) \cdot (1 - P(Y_i = 1|X_i)).$$

The first equality holds because $E(e_i|X_i) = 0$, the third because of Theorem 3.1.11 in the notes, the fourth because $Y_i$ is binary. So $E(e_i^2|X_i)$ cannot be constant unless $P(Y_i = 1|X_i)$ is constant, in which case $X_i$ would not be a very interesting regressor. You should therefore use robust standard errors.

b) Regress arr86 on inc86. How can you interpret the coefficient? Is it significantly different from 0 at conventional levels? Interpret the p-value of the coefficient. Does 0 lie within the 95% confidence interval of the coefficient? How does this relate to its p-value? Compute the minimum detectable difference of this coefficient. Do you think this coefficient can receive a causal interpretation? How can you interpret the constant in that regression?

Solution

The coefficient is -0.0012. The dependent variable is binary and inc86 is expressed in hundreds of dollars. So the coefficient means that a 100 dollars increase in income is associated with a decrease of the average probability of arrest of 0.0012, i.e. of 0.12 percentage points.

This coefficient is significantly different from 0 at any conventional level (its p-value is below 0.001), and as a result 0 does not lie within 95% confidence interval of the coefficient (if it did, this would contradict the fact that the p-value is strictly below 0.05).

As explained during the lectures, for $\alpha = 0.05$ and $\lambda = 0.8$, the MDD is equal to the standard error of the coefficient multiplied by 1.96 + 0.85. Here this is equal to 0.0003. We can distinguish the regression coefficient from 0 with a 80% probability provided it is greater in absolute value than 0.0003. Here that’s not a very relevant computation to make because we reject the null (0 does not lie within the 95% CI), but that’s just to show you how to compute the MDD when your analysis leads you to accept the null.

The coefficient can clearly not receive a causal interpretation. There are many other determinants of the probability of arrest which are not included in this regression and which are probably correlated with income. If you have in mind a simple Becker type model of crime, people decide or not to engage into criminal activity after comparing costs and benefits. The
benefit of crime is the utility you derive from it, say the monetary gain you derive from theft. The cost is the perceived disutility of arrest, multiplied by the perceived probability of being caught. One could for instance argue that the perceived probability of being caught is correlated with education which is correlated with income, which is not included in the regression.

In a regression, the constant is the mean of the outcome for the reference category, i.e. for the subgroup with $X_i = 0$ for all the covariates included in the regression. Here the only covariate is income. So the constant is the average probability of arrest among people whose income is equal to 0.

c) Regress arr86 on inc86 and black. Interpret the inc86 coefficient again. How did it change from the first to the second regression? Use the second omitted variable formula to infer from this the sign of the correlation between inc86 and black. Interpret the coefficient of the black variable.

**Solution**

The coefficient is now -0.0011, which means that a 100 dollars increase in income is associated with a decrease of the average probability of arrest of 0.0011, i.e. of 0.11 percentage points.

The coefficient has slightly increased.

Using the sample version of the second omitted variable formula, this implies that $\hat{\mu} \times \hat{\lambda}$ is negative. $\hat{\lambda}$ is the coefficient of black in that regression, which is positive, from which we can conclude that $\hat{\mu}$ is negative: income is negatively correlated with the black variable.

The coefficient of black in this regression is 0.14. This means that for equal levels of income, black have a probability of being arrested 14 percentage points higher than non black.

d) Now, generate a born60*black variable, and regress arr86 on black, born60, and born60*black. Interpret the coefficient of the constant and the three variables of that regression. Is the black/non black difference in the probability of arrest different for older than for younger people?

**Solution**

The answer to this question can be derived from the third example of the “saturated models” section of the lectures. The subgroup for which the three variables in the regression are 0 are non black men born after 1960. So this is the reference category. The constant tells us that the percentage of them who got arrested in 1986 is 25.6%. The black coefficient captures the difference in arrest rate across black and non black men born after 1960. This implies that the percentage of men arrested in 1986 was 17.3 percentage points higher among black men born after 1960 than among non black men born after 1960. The born60 coefficient captures the difference in arrest rate across non black men born before 1960 and non black men born after 1960. This implies that the percentage of men arrested in 1986 was 1.9 percentage points lower among non black men born before 1960 than among non black men born after 1960.

---

1This is strictly speaking true only when the model is saturated, but let us omit this.
The difference is not significantly different from 0. Finally, the born60*black coefficient is the DID coefficient which compares the black-non black differences in arrest rate across men born before and after 1960. This DID is very small and not significantly different from 0. So the black/ non black difference in the probability of arrest does not seem to vary a lot across age groups.


This paper uses concepts we have not studied yet, so do not worry if you do not understand all of it, just try to get the essential of the message. You should not spend too much time reading it (1 hour and a half max).