SOCIOLOGY 30005

STATISTICAL METHODS OF RESEARCH 2

ASSIGNMENT 1

I. PROBABILITY

Population of interest = US adults in Labor force in 1992.

Relationship of interest = Risk of unemployment and Education attainment

Let, Y be the random variable representing the unemployment status

$$Y = \begin{cases} 1, & \text{if unemployed} \\ 0, & \text{if employed} \end{cases}$$

Let X be the random variable representing education attainment.

$$X = \begin{cases} 1, & \text{if no degree} \\ 2, & \text{if GED} \\ 3, & \text{if High School degree} \\ 4, & \text{if Associates degree} \\ 5, & \text{if Bachelors degree} \\ 6, & \text{if Masters degree or higher} \end{cases}$$

1. Constructing a theoretical contingency table.

	$X=1$	$X=2$	$X=3$	$X=4$	$X=5$	$X=6$	
$Y=1$	π_{11}	π_{12}	π_{13}	π_{14}	π_{15}	π_{16}	$\pi_{1.}$
$Y=0$	π_{01}	π_{02}	π_{03}	π_{04}	π_{05}	π_{06}	$\pi_{0.}$
	$\pi_{.1}$	$\pi_{.2}$	$\pi_{.3}$	$\pi_{.4}$	$\pi_{.5}$	$\pi_{.6}$	π

Let $i=\{1,0\}$ be the indices for Y

$j=\{1,2,3,4,5,6\}$ be the indices for X

π_{ij} = Joint probability of the event when $y=i$ and $x=j$

$\pi_{1.}$ = marginal probability of unemployment

$= P(Y=1)$

$\pi_{0.}$ = marginal probability of employment

$= P(Y=0)$

$\pi_{.1}$ = marginal probability of having no degree

$= P(X=1)$

$\pi_{\cdot 2}$ = marginal probability of having a GED

= $P(x=2)$

$\pi_{\cdot 3}$ = marginal probability of having a high School degree

= $P(x=3)$

$\pi_{\cdot 4}$ = marginal probability of having an Associate's Degree

= $P(x=4)$

$\pi_{\cdot 5}$ = marginal probability of having a Bachelor's Degree

= $P(x=5)$

$\pi_{\cdot 6}$ = marginal probability of having a Master's degree or higher

= $P(x=6)$

Based on the definitions

$$\pi_{1\cdot} = \sum_{j=1}^{6} \pi_{1j}$$

$$\pi_{0\cdot} = \sum_{j=1}^{6} \pi_{0j}$$

Similarly,

$$\pi_{.1} = \sum_{i=0}^{1} \pi_{i1}$$

$$\pi_{.2} = \sum_{i=0}^{1} \pi_{i2}$$

$$\pi_{.3} = \sum_{i=0}^{1} \pi_{i3}$$

$$\pi_{.4} = \sum_{i=0}^{1} \pi_{i4}$$

$$\pi_{.5} = \sum_{i=0}^{1} \pi_{i5}$$

$$\pi_{.6} = \sum_{i=0}^{1} \pi_{i6}$$

2. Marginal probability of unemployment

$$P(Y=1) = ?$$

$$P(Y=1) = \pi_{1.} = \sum_{j=1}^{6} \pi_{ij} = \pi_{11} + \pi_{12} + \pi_{13} + \pi_{14} + \pi_{15} + \pi_{16}$$

$$= P(Y=1, X=1) + P(Y=1, X=2) + P(Y=1, X=3) + P(Y=1, X=4) + P(Y=1, X=5) + P(Y=1, X=6)$$

3. Conditional Probability of unemployment for each possible level of education

(i) $X = 1$

$$P(Y=1 \mid X=1) = \frac{P(Y=1, X=1)}{P(X=1)} = \frac{\pi_{11}}{\pi_{\cdot 1}}$$

This relationship is true due to the property,

$$P(Y=y \mid X=x) = \frac{P(Y=y, X=x)}{P(X=x)}$$

Similarly,

(ii) $X = 2$

$$P(Y=1 \mid X=2) = \frac{P(Y=1, X=2)}{P(X=2)} = \frac{\pi_{12}}{\pi_{\cdot 2}}$$

(iii) $X = 3$

$$P(Y=1 \mid X=3) = \frac{P(Y=1, X=3)}{P(X=3)} = \frac{\pi_{13}}{\pi_{\cdot 3}}$$

(iv) $X = 4$

$$P(Y=1 \mid X=4) = \frac{P(Y=1, X=4)}{P(X=4)} = \frac{\pi_{14}}{\pi_{\cdot 4}}$$

(v) $X = 5$

[redacted] $$P(Y=1 \mid X=5) = \frac{P(Y=1, X=5)}{P(X=5)} = \frac{\pi_{15}}{\pi_{\cdot 5}}$$

(vi) $X = 6$

$$P(Y=1 | X=6) = \frac{P(Y=1, X=6)}{P(X=6)} = \frac{\pi_{16}}{\pi_{.6}}$$

Substantive Interpretation.

$P(Y=1 | X=x) =$ Probability of being unemployed if the education level is at $X = x$.

4. If unemployed, $Y=1$

if no degree, $X=1$

$P(Y=1, X=1) =$ Joint probability of being unemployed and having no degree

$$P(Y=1, X=1) = \underbrace{P(Y=1 | X=1)}_{\text{conditional probability component}} . \underbrace{P(X=1)}_{\text{marginal probability component}}$$

5. Marginal probability of being unemployed /2

$$P(Y=1) = \sum_{x=1}^{6} P(Y=1 \mid X=x)\, P(X=x)$$

USING DATA FROM NALS

/4

6.

$P(Y=1 \mid X=1) = 0.16$

$P(Y=1 \mid X=2) = 0.13$

$P(Y=1 \mid X=3) = 0.88$

$P(Y=1 \mid X=4) = 0.07$

$P(Y=1 \mid X=5) = 0.05$

$P(Y=1 \mid X=6) = 0.03$

we observe that the risk of unemployment decreases as education level increases.

7. Let's assume that education and unemployment are independent,

$$X \perp\!\!\!\perp Y$$

$\therefore \; P(Y=1|X) = P(Y=1) = 0.084$

$\therefore$ 8.4% of individuals in each level of X are expected to be unemployed.

In $X=1$, we have 1579 individuals

$\therefore$ Expected number of unemployed

$= 0.084 * 1579$

≈ 126

But the number of those with no degree and unemployed from the data

Count for $(X=1, Y=1) = 253$

The actual value is greater than that estimated under the assumption of X and Y being independent.

Thus, the assumption may not hold true.

II. EXPECTATION

Variables of interest:

X = Parent years of education

Y = Adult literacy (outcome of interest)

Z = Respondent years of education

1. Theoretical linear model for $Y = f(X, Z)$

True model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$$

$i = \{1, \dots, n\}$ = indice for sample subjects

y_i = adult literacy of i^{th} individual in the sample

β_0 = intercept (no substantive meaning)

x_i = parent years of education for i^{th} individual in the sample

β_1 = For one unit change in parent years of education, holding 'Z' constant, β_1 is the change in adult literacy outcome

Z_i = Years of education of individual (respondent) 'i' in the sample

β_2 = For one unit change in Z (respondent years of education), holding x constant, β_2 is the change in adult literacy outcome.

ε_i = error term
= random deviation of i^{th} individual's outcome from that predicted by the true model.

$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i + \beta_2 Z_i)$

2. (a) $Z_i = \gamma_0 + \gamma_1 X_i + e_i$

(b) $X_i = \delta_0 + \delta_1 Z_i + u_i$

3. (a) $E(Y|Z) = ?$ if we only use 'Z' as a predictor.

$$Y_i = \beta_0 + \beta_2 Z_i + \beta_1 X_i + \varepsilon_i$$

$$E(Y|Z) = \beta_0 + \beta_2 Z + \beta_1 E(X|Z) + E(\varepsilon|Z)$$

(b) Bias due to omission of x.

If we estimate the true model,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$$

then from OLS estimates, $\{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\}$

$$E(\hat{\beta}_0) = \beta_0 \;;\; E(\hat{\beta}_1) = \beta_1 \;;\; E(\hat{\beta}_2) = \beta_2$$

But if we instead believe that,

$$y_i = \beta_0^* + \beta_2^* z_i + \varepsilon_i^*$$

From what we studied in class for a single predictor linear model,

$$\hat{\beta}_2^* = \frac{\hat{cov}(Y, Z)}{\hat{var}(Z)} = \frac{\hat{cov}(\beta_0 + \beta_1 X + \beta_2 Z + \varepsilon, Z)}{\hat{var}(Z)}$$

$$= \frac{\cancel{\hat{cov}(Z, \beta_0)}^{\,0} + \hat{cov}(Z, \beta_1 X) + cov(Z, \beta_2 Z) + \cancel{cov(Z, \varepsilon)}^{\,0}}{\hat{var}(Z)}$$

$$= \frac{\beta_1 \hat{cov}(Z, X) + \beta_2 \hat{var}(Z)}{\hat{var}(Z)}$$

$$= \beta_2 + \beta_1 \cdot \frac{\hat{cov}(Z, X)}{\hat{var}(Z)}$$

$$\text{Bias} = \frac{\beta_1 \cdot \hat{cov}(z,x)}{\hat{var}(z)}$$

Bias = 0 if,

(i) $\beta_1 = 0$, x does not belong in the true model, so its omission has no effect on Y.

(ii) $\hat{cov}(z,x) = 0$, if z and x are not correlated then omitting one doesn't affect the other.

4. $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i \quad - (1)$

$x_i = \delta_0 + \delta_1 z_i + u_i \quad - (2)$

$y_i = \theta_0 + \theta_1 x_i + v_i \quad - (3)$

$z_i = \gamma_0 + \gamma_1 x_i + e_i \quad - (4)$

a) θ_1 = Total effect of x on y

$$E(y|x) = \beta_0 + \beta_1 x + \beta_2 E(z|x) + \varepsilon$$

$$\simeq \theta_0 + \theta_1 x + u_i$$

s.t. $U_i = \beta_2 E(z|x) + \varepsilon_i$

b) Based on (1)

Direct effect of x on y = β_1

c) $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$

$$= \beta_0 + \beta_1 x_i + \beta_2(\gamma_0 + \gamma_1 x_i) + \varepsilon_i + \beta_2 e_i$$

$$= \beta_0 + (\beta_1 + \beta_2 \gamma_1) x_i + \varepsilon_i + \beta_2 e_i$$

earlier we saw that

Direct effect of x on y = β_1

Indirect effect = $\beta_2 \gamma_1$

(d) if

$$y_i = \beta_0 + (\beta_1 + \beta_2 \gamma_1) x_i + \varepsilon_i + \beta_2 e_i$$

and

$$y_i = \theta_0 + \theta_1 x_i + v_i$$

we see that,

$$\theta_1 = \underbrace{\beta_1}_{\substack{\text{Direct} \\ \text{effect}}} + \underbrace{\beta_2 \gamma_1}_{\substack{\text{Indirect} \\ \text{effect}}}$$

$$\begin{tikzcd}
 & Z \arrow[dr, "\beta_2"] & \\
X \arrow[ur, "\gamma_1"] \arrow[rr, "\beta_1"'] & & Y
\end{tikzcd}
\qquad\qquad
X \xrightarrow{\theta_1} Y$$

5. Using OLS in R.

Direct effect $= \hat{\beta}_1 = 3.99$

Indirect effect $= \hat{\beta}_2 \cdot \hat{\gamma}_1 = 4.75$

Total effect $= \hat{\theta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\gamma}_1$

$= 3.99 + 4.75$

$= 8.74$