BIOSTAT620 HW2

Zihao Han

2024-03-07

PROBLEM 1

ALL THE CODE IS IN THE github, $https://github.com/ZihaoHanGitHub/biostat620_hw2$

1a:B

1b:A

1c:AD

1d:D

1e:AB

PROBLEM 2

(a)

```
## Equation 1: SUR for Total Screen Time:
```

```
##
## SUR estimates for 'eq1' (equation 1)
## Model Formula: Total.ST.min ~ Y1_lag1 + X + Z
##
##                Estimate  Std. Error t value   Pr(>|t|)
## (Intercept) 352.0822019  93.7943895 3.75377 0.00077732 ***
## Y1_lag1       0.0759624   0.1474990 0.51500 0.61045444
## X           113.7532152  64.1235419 1.77397 0.08656981 .
## Z            11.6993908  55.0486027 0.21253 0.83318166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 158.330817 on 29 degrees of freedom
## Number of observations: 33 Degrees of Freedom: 29
## SSR: 726990.782164 MSE: 25068.647661 Root MSE: 158.330817
## Multiple R-Squared: 0.088851 Adjusted R-Squared: -0.005406
```

```
##
## Equation 2: SUR for Social Screen Time:
```

```
##
## SUR estimates for 'eq2' (equation 2)
## Model Formula: Social.ST.min ~ Y2_lag1 + X + Z
##
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  98.266204  36.251233  2.71070 0.011161 *
## Y2_lag1       0.196162   0.151791  1.29232 0.206450
## X            34.623299  27.181768  1.27377 0.212860
## Z           -25.896557  23.394815 -1.10694 0.277419
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.588376 on 29 degrees of freedom
## Number of observations: 33 Degrees of Freedom: 29
## SSR: 128586.344076 MSE: 4434.011865 Root MSE: 66.588376
## Multiple R-Squared: 0.11593 Adjusted R-Squared: 0.024475
```

(b)

Identify covariates in each model that are statistically significant at level $\alpha = 0.05$. Explain.

From the table, in the model of Total Screen Time, we get The intercept is statistically significant at the 0.05 level (with $pvalue = 0.00077732 < 0,05$), suggesting that the average value of the response variable is significantly different from zero. For other variables, Y1_lag1 with the $pvalue = 0.61045444 > 0.05$, X with the $pvalue = 0.08656981 > 0.05$, Z with the $pvalue = 0.83318166 > 0.05$, indicating these three variables does not have a significant effect on Total Screen Time.

From the table, in the model of Social Screen Time, we get The intercept is statistically significant at the 0.05 level (with $pvalue = 0.011161 < 0,05$), suggesting that the average value of the response variable is significantly different from zero. For other variables, Y2_lag1 with the $pvalue = 0.206450 > 0.05$, X with the $pvalue = 0.212860 > 0.05$, Z with the $pvalue = 0.277419 > 0.05$, indicating these three variables does not have a significant effect on Total Screen Time.

(c)

Test the null hypothesis $\beta_3 = \gamma_3 = 0$, that is, $Z(t)$ is not an important predictor in BOTH screen time outcomes. Draw conclusion at $\alpha = 0.05$ level.

```
## Do not have sufficient evidence to reject the null hypothesis,
##       with the pvalue 0.8438972
```

```
## This test use the Wald test, with the wald test statistics 0.03877393
```

PROBLEM 3

(a)

Explain why X_i and ϵ_i are independent.

X_i denote the i^{th} patients using A or B drugs, and ϵ_i denoted the error term in this model. Since in this model, the dataset is collected from a randomized clinical trail, it means that the drug selection in the clinical trial is random during the treatment, each individual has the equal chance to use A or B drugs, therefore,

X_i would be independent on each other factors. More than that, in SLR assumption, the independence assumption is one of the basic assumption in SLR, therefore, the error term ϵ_i is always independent on other variables. Hence, X_i and ϵ_i are independent.

(b)

In model (1), explain which parameter represents the treatment effect of drug A, and explain which parameter represents the treatment effect of drug B

Since the individual who receive drug A (coded by $X_1 = 1$), and who are randomized to receive drug B (coded by $X_i = -1$). The effect parameter of drug A is β_1, and the effect parameter of drug B is $-\beta_1$.

(c)

Show that the treatment effects identified in part (b) are invariant for the inclusion of any confounding covariate Z into the model (1)

Since we plug in the confounding covariate Z into the model, the model equation is become $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \epsilon_i, i = 1, ..., 2n$. However, the effect of drug A and drug B do not change, The effect parameter of drug A is still β_1, and the effect parameter of drug B is still $-\beta_1$. Plug in a confounding covariate Z does not change the explanation of effect parameter, meaning that the plug in the Z does not change the interpretation of effects of drug A or drug B. Hence, the treatment effects are invariant for the inclusion of any confounding covariates.

(d)

Give the estimate of the causal effect (i.e. ATE) when drug B is a placebo.

Since drug B is a placebo, ATE is defined as the average causal effect of the treatment across all subjects. Given that

$ATE = E[Y|X_i = 1] - E[Y|X_i = -1]$, defined as the difference in the expected values of Y between those who receive drug A and those who receive drug B (placebo). Therefore, $ATE = \beta_1 - (-\beta_1) = 2\beta_1$.

Hence the estimate of causal effect ATE is $2\beta_1$ when B is a placebo.

PROBLEM 4

(a)

What is the variance of the error $\tilde{\epsilon} = \beta_1 \epsilon + e$ under the assumption that the two errors e and ϵ are independent?

Since two errors e and ϵ are independent,

$Var[\tilde{\epsilon}] = Var[\beta_1 \epsilon + e] = Var[\beta_1 \epsilon] + Var[e]$

Therefore, $Var[\tilde{\epsilon}] = \beta_1^2 Var[\epsilon] + Var[e] = \beta_1^2 \sigma_\epsilon^2 + \sigma_e^2$

(b)

What is the variance of the unbiased estimator of $\tilde{\beta}_1 = \alpha_1\beta_1$ denoted by $\hat{\tilde{\beta}}_1$, when a random sample of n observations (Z_i, X_i, Y_i),$i = 1, ..., n$, are collected from a biomedical study?

Since $Y = \tilde{\beta}_0 + \tilde{\beta}_1 Z + \tilde{\varepsilon}$, where $\tilde{\beta}_1 = \alpha_1\beta_1$.

The unbiased estimator of $\tilde{\beta}_1$ is $\tilde{\beta}_1 = \frac{\sum_{i=1}^n (Z_i - Z)(Y_i - \bar{Y})}{\sum_{i=1}^n (Z_i - \bar{Z})^2}$

With the Variance $Var(\tilde{\beta}_1) = \frac{Var[\tilde{\epsilon}]}{SSZ}$, where $SSZ = \sum_{i=1}^n (Z_i - \bar{Z})^2$

Hence $Var(\tilde{\beta}_1) = \frac{\beta_1^2\sigma_\epsilon^2 + \sigma_e^2}{SSZ}$

(c)

What is the variance of the unbiased estimator α_1, denoted by $\hat{\alpha_1}$, with the random sample of n observations (Z_i, X_i, Y_i), $i = 1, ..., n$,?

Since the unbiased estimator of α_1 is $\alpha_1 = SSXZ/SSZ$, therefore it given the variance

$Var(\hat{\alpha_1}) = \frac{Var[e]}{SSZ} = \frac{\sigma_e^2}{SSZ}$, where $SSZ = \sum_{i=1}^n (Z_i - \bar{Z})^2$.

(d)

Deriving the variance of the IV estimator $\hat{\beta}_1 = \frac{\hat{\tilde{\beta}}_1}{\hat{\alpha}_1}$ seems to be analytically challenging. One may invoke the method of bootstrap to numerically evaluate this variance with the random sample of n observations (Z_i, X_i, Y_i), $i = 1, ..., n$,. Describe the major steps and pseudo code that you design to implement the bootstrap method in the calculation of the variance of the IV estimator.

Since the $\hat{\tilde{\beta}}_1$ is the effect of Z on Y, and $\hat{\alpha}_1$ is the effect of Z on X, therefore, they are independent of each other.

Therefore $Var(\hat{\tilde{\beta}}_1) = Var(\frac{\hat{\tilde{\beta}}_1}{\hat{\alpha}_1}) = \frac{Var(\hat{\tilde{\beta}}_1)}{Var(\hat{\alpha}_1)}$

Hence, $Var(\hat{\tilde{\beta}}_1) = \frac{\beta_1^2\sigma_\epsilon^2 + \sigma_e^2}{SSZ} * \frac{SSZ}{\sigma_e^2} = \frac{\beta_1^2\sigma_\epsilon^2 + \sigma_e^2}{\sigma_e^2}$

Pseudo Code for implement the Bootstrap method for calculating the variance of the IV estimator:

Loading the Data

\$ data = read.csv(“.”), colname = c(“Y”,“X”,“Z”)\$

Define the Bootstrap

iteration $= n$ for (i in 1:iteration) { # sampling from the original dataset # calculate the variance of VI estimator }

Calculate the Variance of VI estiamtor by bootstrap

PROBLEM 5

We denote that x_{i2} as an indicator of gender, $x_{i2} = 0$ for male, and $x_{i2} = 1$ for male.

Therefore,

$y_i = \beta_0^F + \beta_1^F z_i + \beta_2^F x_{i1} + \varepsilon_i^F$

$y_i = \beta_0^M + \beta_1^M z_i + \beta_2^M x_{i1} + \varepsilon_i^M$

$y_i = \beta_0 + \beta_1 z_i + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4(z_i \times x_{i2}) + \beta_5(x_{i1} \times x_{i2}) + \varepsilon_i$, when $x_{i2} = 0$ is becomes to

$y_i = \beta_0 + \beta_1 z_i + \beta_2 x_{i1} + \epsilon_i$, therefore, $\beta_0^M = \beta_0, \beta_1^M = \beta_1, \beta_2^M = \beta_2$

When $x_{i2} = 1$, the linear model is given by:

$y_i = \beta_0 + \beta_1 z_i + \beta_2 x_{i1} + \beta_3 + \beta_4 z_i + \beta_5 x_{i1}$, therefore, $\beta_0^F = \beta_0 + \beta_3, \beta_1^F = \beta_1 + \beta_4, \beta_2^F = \beta_2 + \beta_5$