HW2

Liancheng Lu

2024-03-05

Github: https://github.com/lianchenglu/BIOSTAT620_HW2

Problem 1: Choose all correct answers

(1a) A clinician chooses to run a case-control study in order to

AB

A. control confounding factors; B. mitigate a low number of cases in the population for the recruitment of participants; C. compare with results found in a cohort study; D. handle missing data.

(1b) Refer to Figure 6 on page 9 of the Lecture Notes, Chapter II. One of conditions required by an instrumental variable Z to control unmeasured confounders is that

C

A. it is correlated with the outcome variable Y ; B. it is correlated with the error term e; C. it is correlated with the exposure variable X; D. none of the above.

(1c) In a clinical study that aims to compare a test drug A to placebo P, a clinician plans to design a fully randomized trial in that patients will be allocated with a treatment by chance. In this way, the research is able to

AD

A. get rid of potential confounding effects; B. maximize the treatment effect of the test drug; C. ensure data privacy; D. assess causal effect of the test drug.

(1d) An epidemiologist plans to design a cohort study to study the influence of PM2.5 concentration on post-transplantation graft survival for ESRD patients in the USA who receive renal replacement therapy. As part of study design, both inclusion and exclusion criteria are created to

CD

A. select only patients exposed to high PM2.5 concentration; B. reduce sample size for cost saving; C. establish a causal effect of exposure to PM2.5 on graft survival; D. define the underlying study population to which analysis results may be applied.

(1e) A public health scientist plans to conduct an observational study to evaluate the efficacy of the covid booster shot on risk of hospitalization using electronic health records from the U-M university hospital. In this study design, this researcher wants to divide the subjects in the database into several age groups, including 0-5, 6-12, 13-18, 19-50, 51-65, and 65+ years old, in order to

AB

A. address potential heterogeneous risk of hospitalization associated with age; B. estimate age-specific efficacy of the covid booster shot on risk of hospitalization; C. have flexibility and convenience in releasing results to the public. D. mitigate the problem of missing data;

Problem 2: Seemingly unrelated regression (SUR) is a widely used method to run multi-outcome regression analysis. This modeling approach may be applied for the analysis of screen activity data collected from your mobile devices. Use the R function systemfit with the method option of SUR to fit your own data that you used previously in Homework #1 with the data freeze date of Jan 26, 2024.

(Due to my data was collected after Jan 14, 2024, I set Z(t) = 1 for day t being January, 17 (the second week of the winter semester))

a

For Total Screen Time:

Intercept: 301.8008

Total.ST.min_lag (Lagged Total Screen Time): -122.7729

is_weekday (Weekday Dummy Variable): -124.9146

IsBeforeJan17 (Dummy for After or Before Jan 17): 79.9354

For Total Social Screen Time:

Intercept: 202.6318

Social.ST.min_lag: -44.2149

is_weekday: -111.6319

IsBeforeJan17: 29.6427

```r
df$Total.ST.min_lag <- c(NA, acf(df[[3]], plot = F)$acf) # total ST
df$Social.ST.min_lag <- c(NA, acf(df[[5]], plot = F)$acf) # total social ST

eq1 <- Total.ST.min ~ Total.ST.min_lag + is_weekday + IsBeforeJan17
eq2 <- Social.ST.min ~ Social.ST.min_lag + is_weekday + IsBeforeJan17

fit <- systemfit(list(eq1=eq1, eq2=eq2), data=df, method="SUR")
summary(fit)
```

```
##
## systemfit results
## method: SUR
##
##         N DF     SSR detRCov   OLS-R2 McElroy-R2
## system 24 16 22172.7  505808 0.726186   0.692455
##
##      N DF     SSR     MSE    RMSE       R2   Adj R2
## eq1 12  8 10912.2 1364.02 36.9327 0.786127 0.705924
## eq2 12  8 11260.6 1407.57 37.5176 0.624092 0.483126
##
## The covariance matrix of the residuals used for estimation
##         eq1     eq2
## eq1 1360.78 1179.10
## eq2 1179.10 1397.24
##
## The covariance matrix of the residuals
##         eq1     eq2
## eq1 1364.02 1189.18
## eq2 1189.18 1407.57
##
## The correlations of the residuals
##          eq1      eq2
## eq1 1.000000 0.858226
## eq2 0.858226 1.000000
##
##
## SUR estimates for 'eq1' (equation 1)
## Model Formula: Total.ST.min ~ Total.ST.min_lag + is_weekday + IsBeforeJan17
##
##                   Estimate Std. Error  t value   Pr(>|t|)
## (Intercept)       301.8008    27.5265 10.96403 4.2527e-06 ***
## Total.ST.min_lag -122.7729    50.9237 -2.41092  0.0424473 *
## is_weekday       -124.9146    29.5315 -4.22988  0.0028768 **
## IsBeforeJan17      79.9354    51.4910  1.55242  0.1591650
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.932658 on 8 degrees of freedom
## Number of observations: 12 Degrees of Freedom: 8
## SSR: 10912.16965 MSE: 1364.021206 Root MSE: 36.932658
## Multiple R-Squared: 0.786127 Adjusted R-Squared: 0.705924
##
##
## SUR estimates for 'eq2' (equation 2)
## Model Formula: Social.ST.min ~ Social.ST.min_lag + is_weekday + IsBeforeJan17
##
##                    Estimate Std. Error  t value   Pr(>|t|)
## (Intercept)        202.6318    26.6426  7.60557 6.2722e-05 ***
## Social.ST.min_lag  -44.2149    40.7589 -1.08479  0.3096196
## is_weekday        -111.6314    30.8561 -3.61780  0.0068042 **
## IsBeforeJan17       29.6427    44.8805  0.66048  0.5275023
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 37.517625 on 8 degrees of freedom
## Number of observations: 12 Degrees of Freedom: 8
## SSR: 11260.577774 MSE: 1407.572222 Root MSE: 37.517625
## Multiple R-Squared: 0.624092 Adjusted R-Squared: 0.483126
```

b

Two covariates in model eq1 (Total screen use time), total.st.min_lag (total screen use time on a lag) and is_weekday (whether it was a working day or not), were statistically significant at the significance level of alpha = 0.05. In model eq2 (social screen time), is_weekday was the only significant covariate. However, IsBeforeJan17 was not significant in either equation. This means that lagged screen use and whether it was a workday had a significant effect on total screen use, while whether it was a workday also significantly affected social screen use.

c

The coefficient on Z(t) in both models is not statistically significant at the alpha = 0.05 level, so we cannot reject the null hypothesis. In other words, we do not have sufficient evidence that Z(t) is a significant predictor of either screen-time outcome.

Problem 3: Consider a linear model that is used to estimate the treatment effect based on a dataset collected from a randomized clinical trial.

a. Explain why Xi and epsilon are independent.

The randomization process ensured that each subject had an equal chance of receiving either treatment. This ensured that the two groups of subjects were similar in all known or unknown factors that might influence the outcome prior to drug administration.

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.

The parameter beta1 represents the therapeutic effect of drug A. When Xi = 1, the patient received drug A. So beta1 reflects the difference in effect between receiving drug A and not receiving treatment.

The therapeutic effect of drug B is expressed by $-\beta_1$. When $Xi = -1$, $-\beta_1$ represents the change in effect of receiving drug B relative to beta0.

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

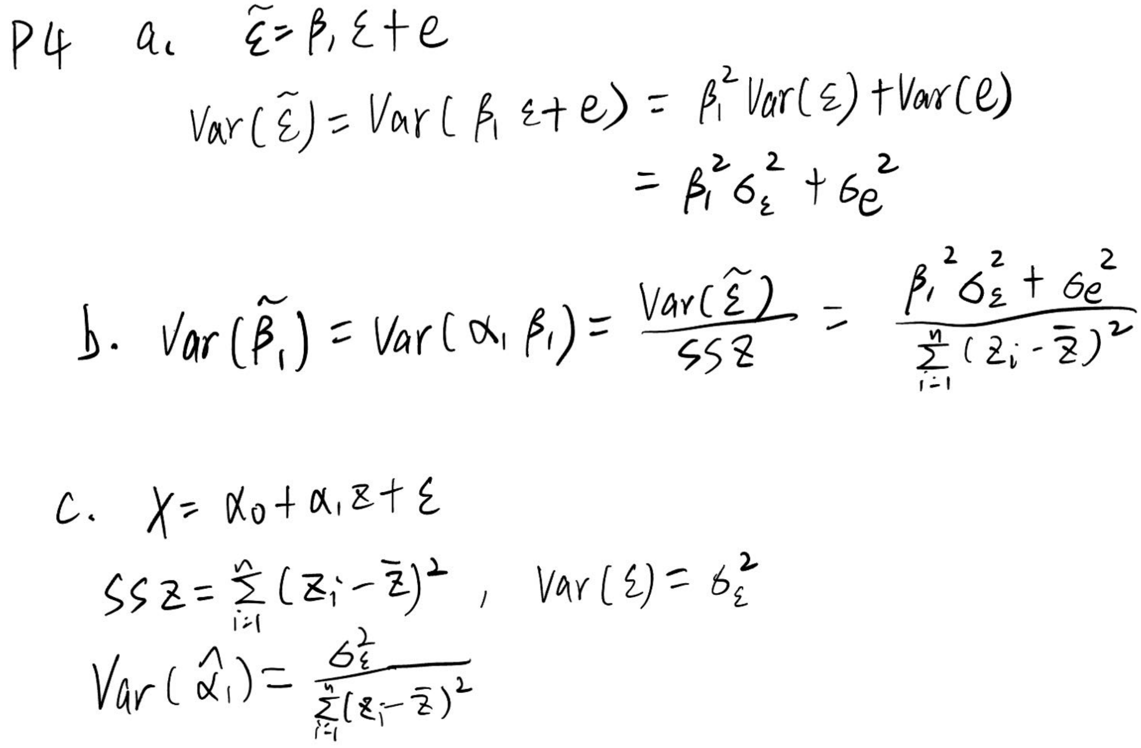
Since treatment group assignment was randomized, any potential confounding variable Z should be independent of treatment assignment Xi. Random assignment ensures that the distribution of Z is similar in the treatment and control groups if Z is any confounding variable related to Yi. This indicates that the influence of Z on Yi has been equally distributed at random across all treatment groups; hence, the estimated effect of Xi on Yi remains unchanged when Z is included in the model, which means

$$\mathbb{E}[Y_i|X_i = 1, Z] - \mathbb{E}[Y_i|X_i = -1, Z] = \mathbb{E}[Y_i|X_i = 1] - \mathbb{E}[Y_i|X_i = -1]$$

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

When drug B is placebo, the mean treatment effect (ATE) of drug A is the average difference in effect between the population receiving drug A and placebo in the randomized controlled trial, which means $(\beta_0 + \beta_1) - (\beta_0 - \beta_1) = 2\beta_1$.

Problem 4:



P4 a. $\tilde{\varepsilon} = \beta_1 \varepsilon + e$

$Var(\tilde{\varepsilon}) = Var(\beta_1 \varepsilon + e) = \beta_1^2 Var(\varepsilon) + Var(e)$

$= \beta_1^2 \sigma_\varepsilon^2 + \sigma_e^2$

b. $Var(\tilde{\beta}_1) = Var(\alpha_1 \beta_1) = \frac{Var(\tilde{\varepsilon})}{SSZ} = \frac{\beta_1^2 \sigma_\varepsilon^2 + \sigma_e^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2}$

c. $X = \alpha_0 + \alpha_1 z + \varepsilon$

$SSZ = \sum_{i=1}^{n} (z_i - \bar{z})^2$, $Var(\varepsilon) = \sigma_\varepsilon^2$

$Var(\hat{\alpha}_1) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2}$

Figure 1: Problem 4

d.

1. Randomly draw a sample of observations from the original dataset with replacement to create a bootstrap sample. 2. Using the bootstrap sample, re-estimate the parameters $\hat{\alpha_1}$ and $\hat{\beta}_1$.
2. Calculate the ratio $\frac{\hat{\beta}_1}{\hat{\alpha_1}}$ using the re-estimated parameters from the bootstrap sample.
3. Repeat steps 1-3.
4. Calculate the sample variance of the bootstrap estimates.

pseudo

```
## Assume original_data is an array of tuples (Z, X, Y)
original_data = [(Z1, X1, Y1), ..., (Zn, Xn, Yn)]
num_bootstraps = 1000
bootstrap_estimates = []

for i in 1 to num_bootstraps do:
    bootstrap_sample = sample_with_replacement(original_data, size=n)
    alpha_hat = regress(X ~ Z using bootstrap_sample)
    beta_hat = regress(Y ~ Z using bootstrap_sample)
    beta_iv_star = beta_hat / alpha_hat
    append(bootstrap_estimates, beta_iv_star)

variance_beta_iv = variance(bootstrap_estimates)
return variance_beta_iv
```

Problem 5:

Model (5) for females: $y_i = \beta_{0F} + \beta_{1F} z_i + \beta_{2F} x_{i1} + \varepsilon_{Fi}$

Model (6) for males: $y_i = \beta_{0M} + \beta_{1M} z_i + \beta_{2M} x_{i1} + \varepsilon_{Mi}$

Model (7): $y_i = \beta_0 + \beta_1 z_i + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 (z_i \times x_{i2}) + \beta_5 (x_{i1} \times x_{i2}) + \varepsilon_i$

$$\begin{aligned}
\beta_{0F} &\text{ corresponds to } \beta_0 \text{ when } x_{i2} = 0 \text{ (for females)},\\
\beta_{0M} &\text{ corresponds to } \beta_0 + \beta_3 \text{ when } x_{i2} = 1 \text{ (for males)},\\
\beta_{1F} &\text{ corresponds to } \beta_1 \text{ when } x_{i2} = 0 \text{ (for females)},\\
\beta_{1M} &\text{ corresponds to } \beta_1 + \beta_4 \text{ when } x_{i2} = 1 \text{ (for males)},\\
\beta_{2F} &\text{ corresponds to } \beta_2 \text{ when } x_{i2} = 0 \text{ (for females)},\\
\beta_{2M} &\text{ corresponds to } \beta_2 + \beta_5 \text{ when } x_{i2} = 1 \text{ (for males)}.
\end{aligned}$$