BIOSTAT620_HW2

Bulun Te

2024-03-05

Problem 1

1(a)

B

1(b)

C

1(c)

AD

1(d)

D

1(e)

AB

Problem 2

```r
data = data %>% mutate(
  Total.ST.min_lag1 = lag(Total.ST.min,1),
  Social.ST.min_lag1 = lag(Social.ST.min,1)
) %>% filter(!is.na(Total.ST.min_lag1))

data %>% head()
```

```
##         Date Total.ST Total.ST.min Social.ST Social.ST.min Pickups Pickup.1st
## 1 2024-01-01    30min           30     24min            24      13       4:02
## 2 2024-01-02  1h43min          103   1h25min            85      50       7:09
## 3 2024-01-03   1h9min           69      9min             9      39       7:40
## 4 2024-01-04    57min           57     34min            34      35       9:00
```

```
## 5 2024-01-05   1h9min           69      3min              3      54       9:00
## 6 2024-01-06    55min           55     20min             20      49       8:01
##   proportion duration_per_use weekday semester semester_weekday
## 1 0.80000000         2.307692       1        0                0
## 2 0.82524272         2.060000       1        0                0
## 3 0.13043478         1.769231       1        0                0
## 4 0.59649123         1.628571       1        0                0
## 5 0.04347826         1.277778       1        0                0
## 6 0.36363636         1.122449       0        0                0
##   Total.ST.min_lag1 Social.ST.min_lag1
## 1                29                 18
## 2                30                 24
## 3               103                 85
## 4                69                  9
## 5                57                 34
## 6                69                  3
```

(2a)

```r
library(systemfit)
```

```
## Warning: package 'systemfit' was built under R version 4.3.3

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

## Loading required package: lmtest

## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Please cite the 'systemfit' package as:
## Arne Henningsen and Jeff D. Hamann (2007). systemfit: A Package for Estimating Systems of Simultaneou
##
## If you have questions, suggestions, or comments regarding the 'systemfit' package, please use a forur
## https://r-forge.r-project.org/projects/systemfit/
```

```
eq1 = Total.ST.min ~ 1 + Total.ST.min_lag1 + weekday + semester
eq2 = Social.ST.min ~ 1 + Social.ST.min_lag1 + weekday + semester
eqs = list(eq1, eq2)

fit_2a = systemfit(eqs,method="SUR",data=data)

fit_2a %>% summary()
```

```
##
## systemfit results
## method: SUR
##
##         N DF    SSR detRCov   OLS-R2 McElroy-R2
## system 52 44 152612 2415359 0.184328   0.332455
##
##      N DF      SSR     MSE    RMSE       R2   Adj R2
## eq1 26 22 120639.8 5483.63 74.0515 0.150640 0.034818
## eq2 26 22  31972.4 1453.29 38.1221 0.290509 0.193760
##
## The covariance matrix of the residuals used for estimation
##         eq1     eq2
## eq1 5362.07 2264.60
## eq2 2264.60 1432.59
##
## The covariance matrix of the residuals
##         eq1     eq2
## eq1 5483.63 2356.68
## eq2 2356.68 1453.29
##
## The correlations of the residuals
##          eq1      eq2
## eq1 1.000000 0.834816
## eq2 0.834816 1.000000
##
##
## SUR estimates for 'eq1' (equation 1)
## Model Formula: Total.ST.min ~ 1 + Total.ST.min_lag1 + weekday + semester
##
##                     Estimate Std. Error  t value Pr(>|t|)
```

```
## (Intercept)       69.717787  36.818927  1.89353 0.071514 .
## Total.ST.min_lag1 -0.132980   0.184755 -0.71976 0.479247
## weekday            0.754304  34.416560  0.02192 0.982712
## semester          69.984660  31.478735  2.22324 0.036785 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.051507 on 22 degrees of freedom
## Number of observations: 26 Degrees of Freedom: 22
## SSR: 120639.764063 MSE: 5483.625639 Root MSE: 74.051507
## Multiple R-Squared: 0.15064 Adjusted R-Squared: 0.034818
##
##
## SUR estimates for 'eq2' (equation 2)
## Model Formula: Social.ST.min ~ 1 + Social.ST.min_lag1 + weekday + semester
##
##                     Estimate Std. Error  t value Pr(>|t|)
## (Intercept)        -2.078581  18.631520 -0.11156  0.91218
## Social.ST.min_lag1  0.149428   0.158158  0.94480  0.35502
## weekday            30.240525  18.454156  1.63868  0.11550
## semester           27.655173  16.102801  1.71741  0.09995 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.122071 on 22 degrees of freedom
## Number of observations: 26 Degrees of Freedom: 22
## SSR: 31972.431358 MSE: 1453.292334 Root MSE: 38.122071
## Multiple R-Squared: 0.290509 Adjusted R-Squared: 0.19376
```

(2b)

From the estimation, it is shown that at the significance level of 0.05, only the coefficient of `semster` is significant in the model for Total Screen time.The coefficient of the 'semester' variable in equation 1 (eq1) is estimated to be 69.984660 with a p-value of 0.036785, which indicates that, all else being equal, the daily total screen time is expected to be about 70 minutes greater on days after January 10 compared to days before January 10.

The reasons that cause only the coefficient of `semester` to be significant in the model for Total Screen time might be as follows:

1. With only 26 observations and 22 degrees of freedom, the sample size may be too small to detect significant effects.
2. The model might be missing some key variables or non-linear terms that are important in explaining the variation in the dependent variables.
3. As the Social Screen time and total screen time all increase in the semster, there might be multicollinearity between the semester and lag 1 of the screen times.

(2c)

```r
library(car)
hypothesis <- c("eq1_semester = 0", "eq2_semester = 0")
wald_test <- linearHypothesis(fit_2a, hypothesis)
wald_test
```

```
## Linear hypothesis test (Theil's F test)
##
## Hypothesis:
## eq1_semester = 0
## eq2_semester = 0
##
## Model 1: restricted model
## Model 2: fit_2a
##
##   Res.Df Df      F  Pr(>F)
## 1     46
## 2     44  2 2.5414 0.09026 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The test yielded an F-statistic of 2.5414 with a corresponding p-value of 0.09026. At the significance level of 0.05, we fail to reject the null hypothesis that the coefficients of the `semester` variable in both equations are jointly equal to zero. This indicates that the data doesn‘t have sufficient evidence to confirm semester has a significant effect on the screen times.

Problem 3

(a)

By randomly assigning participants to either take A drug or take B drug, it ensures that confounders are likely be ditributed equally among the groups. In this way, any confounders have an equal chances of affecting each treatment arms. This means that unobserved confounders in the ϵ_i term are likely to be independent of the treatment assignment X_i.

Because of random assignment, the treatment variable X_i is independent of other variables that are not observed in the study. In this way, the treatment effect of drugs are independent of the outcome variable.

(b)

As X_i is coded in {1,-1}, thus, treatment effect for drug a is β_1 and treatment effect for drug b is $-\beta_1$.

(c)

$\text{Cov}(Y, X) = \text{Cov}(\beta_0 + \beta_1 X + \varepsilon, X) = \beta_1 \text{Var}(X) + \text{Cov}(\varepsilon, X) = \beta_1 \text{Var}(X)$

Therefore, without adding confounders,

$\hat{\beta}_1 = \frac{\text{Cov}(Y,X)}{\text{Var}(X)}$

After adding confounder Z, as by the randomization, Z is independent with X. Thus, we have

$\text{Cov}(Y, X) = \text{Cov}(\beta_0 + \beta_1 X + \beta_2 Z + \varepsilon, X) = \beta_1 \text{Var}(X) + \beta_2 \text{Cov}(Z, X) + \text{Cov}(\varepsilon, X) = \beta_1 \text{Var}(X) + 0 + 0$

$\hat{\beta}_1 = \frac{\text{Cov}(Y,X)}{\text{Var}(X)}$

Therefore, by adding confounder, the estimate of β_1 is still unbiased.

(d)

ATE $= \mathbb{E}[Y_A - Y_B] = \mathbb{E}[Y_A] - \mathbb{E}[Y_B]$

As X is randomized,X is independent with outcome Y. Therefore,

$\mathbb{E}[Y_A] = \mathbb{E}[Y|X=1] = \beta_0 + \beta_1, \quad \mathbb{E}[Y_B] = \mathbb{E}[Y|X=-1] = \beta_0 - \beta_1$

Thus, ATE $= \mathbb{E}[Y_A - Y_B] = \mathbb{E}[Y_A] - \mathbb{E}[Y_B] = 2\beta_1$