Biostatistics II Exam

2025-10-30

Question 1

The Missingness in Cannabis Use does not depend on the outcome or the covariates in the Figure 1A. There is no path going from $Miss_{cu}$ to *Outcome*, *Exposure* and the *covariates*, hence we can say that $Miss_{cu}$ is d-separated from them and is independent of them. So, the probability of missingness does not depend on any variable in the Figure 1A, therefore, missingness in cannabis use is missing completely at random (MCAR).

Question 2

By definition, Missing at Random (MAR) is when the missingness in Cannabis Use can be explained by associations with the observed data, specifically fully observed data of covariates or outcome or both. Which is what we observe in Figure 1B, 1F and 1D respectively. In terms of d-separation, in the Figure 1B, there are multiple open paths from $Miss_{cu}$ to *Outcome* :

1. $$Miss_{cu} \leftarrow Maternal\ Substance\ Use \rightarrow Outcome$$
2. $$Miss_{cu} \leftarrow Maternal\ Substance\ Use \rightarrow Cannabis\ Use \rightarrow Outcome$$
3. $$Miss_{cu} \leftarrow Maternal\ Substance\ Use \rightarrow Cannabis\ Use \leftarrow Sex \rightarrow Outcome$$

and can all be blocked when *Maternal Substance Use* is adjusted for, where it is possible as *Maternal Substance Use* is fully observed. Then, $Miss_{cu}$ and *Outcome* are independent conditioned on *Maternal Substance Use*. Hence, as missingness can be explained with observed data, this is MAR.

Secondly in the Figure 1D, there is an additional open path compared to Figure 1B:

$$Miss_{cu} \leftarrow Outcome$$

which means that *Outcome* is causing the missingness in Cannabis Use and not the other way around. This path can be blocked by conditioning on *Outcome*, which is fully observed as well. Hence, the Missingness in Cannabis Use and *Outcome* are d-separated conditioned on *Maternal Substance Use* and *Outcome*. Therefore, missingness is MAR.

Lastly, in the Figure 1F, the only open path existing is:

$$Miss_{cu} \leftarrow Outcome$$

Hence, the Missingness in Cannabis Use and *Outcome* are d-separated conditioned on *Outcome* and is MAR.

Question 3

By definition, Missingness Not At Random (MNAR) is when missingness cannot be explained by the observed data. For example, when Missingness in Cannabis Use is explained by the Cannabis Use itself, we can say that the missingness is not at random. In terms of d-separation, firstly in Figure 1C, there are multiple open paths from $Miss_{cu}$ to $Outcome$:

1. $$Miss_{cu} \leftarrow Cannabis\ Use \rightarrow Outcome$$
2. $$Miss_{cu} \leftarrow Cannabis\ Use \leftarrow Sex \rightarrow Outcome$$
3. $$Miss_{cu} \leftarrow Maternal\ Substance\ Use \rightarrow Outcome$$
4. $$Miss_{cu} \leftarrow Maternal\ Substance\ Use \rightarrow Cannabis\ Use \rightarrow Outcome$$
5. $$Miss_{cu} \leftarrow Maternal\ Substance\ Use \rightarrow Cannabis\ Use \leftarrow Sex \rightarrow Outcome$$

Here, the first two paths can only be blocked by conditioning on $CannabisUse$, which is not fully observed. Hence, $Miss_{cu}$ and $Outcome$ cannot be d-separated and are not independent of each other. Therefore, is MNAR.

Similarly in Figure 1E, the open path of

$$Miss_{cu} \leftarrow Outcome$$

is added to the previous list of Figure 1C, which can be blocked by adjusting for $Outcome$. Still, the $Miss_{cu}$ and $Outcome$ cannot be d-separated, even after adjusting for all observed variables and is MNAR.

Question 4

The linear regression model is:

$\mathbf{Outcome} \sim \mathbf{MaternalSubstanceUse} + \mathbf{Sex} + \mathbf{CannabisUse}$

$\Rightarrow \mathbf{Y} = \beta_{\mathbf{0}} + \beta_{\mathbf{1}} \times \mathbf{M} + \beta_{\mathbf{2}} \times \mathbf{S} + \beta_{\mathbf{3}} \times \mathbf{X}$

where Y stands for the $Outcome$, X denotes the exposure, which is $CannabisUse$, M is $MaternalSubstanceUse$, S is Sex and R is Missingness In Cannabis Use, $Miss_{Cu}$

Here, in both Figure 1C and 1D, we know that the exposure: $\mathbf{X} \sim \mathbf{M} + \mathbf{S}$.

Although,

in Figure 1C : $\mathbf{R} \sim \mathbf{M} + \mathbf{X}$,

in Figure 1D : $\mathbf{R} \sim \mathbf{M} + \mathbf{Y}$.

In this question, we try to calculate the Maximum Likelihood estimators of the above mentioned linear regression model of Y. Maximum likelihood here, tries to estimate the regression coefficients that makes the observable data most likely under the model we assume.

First, we simulate data based on the knowledge we know from the above mentioned causal diagrams. Since, there is missingness in exposure, we need to factor in the marginal distribution of $CannibasUse$ into the likelihood. Then, we use the `frm_em()` function to carry out the actual estimation. This function assumes that the missingness is **MAR**, that is, after we account for the observed variables, the missingness does not depend on the unobserved data.

```r
beta0 <- 0
betaX_true <- 1
betaM_true <- 1
betaS_true <- 1

generate_data <- function(n, dag) {

  pX <- function(M, S)
    plogis(M + S)

  M <- rbinom(n, 1, 0.5)
  S <- rbinom(n, 1, 0.5)
  X_continuous <- 1 + 0.8*M + 0.5*S + rnorm(n, 0, 1)
  # Convert to values 0,1,2 (categorical cannabis use), but still numeric
  X <- pmin(pmax(round(X_continuous), 0), 2)
  Y <- rnorm(n, mean = beta0 + betaM_true*M + betaS_true*S + betaX_true*X, sd=1)


  # chosen the parameters in a way to show bias evidently
  if (dag == "fig1D") {
    # scale the continuous value Y so logit doesn't explode
    R <- rbinom(n, 1, pX(0.5*M, 0.5*scale(Y)[,1]))
  } else if(dag == "fig1C") {
        R <- rbinom(n, 1, pX(0.5*M, 4*X))
  }
  Xobs <- ifelse(R==1, X, NA_integer_)
  dat <- data.frame(Y=Y, X=Xobs, M=M, S=S, R=R, X_true=X)
  return(dat)
}

# Maximum Likelihood Estimation with Missing data
fit_frm_em_once <- function(n, dag) {
  dat <- generate_data(n, dag)
  dep <- list(model="linreg", formula=Y ~ X + M + S)
  ind <- list(X = list(model="linreg", formula=X ~ M + S))
  sink(tempfile()) # to suppress the progress bar output
  fit <- frm_em(dat=dat, dep=dep, ind=ind, verbose=FALSE)
  sink()
  cf  <- coef(fit)
  beta_X  <- unname(cf["Y ON X"])
  se_X  <- unname(fit$se["Y ON X"])
  beta_M  <- unname(cf["Y ON M"])
  se_M  <- unname(fit$se["Y ON M"])
  beta_S  <- unname(cf["Y ON S"])
  se_S  <- unname(fit$se["Y ON S"])

  c(beta_X=beta_X, se_X=se_X, beta_M=beta_M, se_M=se_M, beta_S=beta_S, se_S=se_S)
}

eval_ML_grid <- function(n_vec, reps, dag){
  out <- lapply(n_vec, function(n){
    M <- replicate(reps, fit_frm_em_once(n, dag))
    beta_X <- M["beta_X",]; se_X <- M["se_X",]
```

```
    beta_M <- M["beta_M",]; se_M <- M["se_M",]
    beta_S <- M["beta_S",]; se_S <- M["se_S",]
    coverage_X <- mean((beta_X - 1.96*se_X) <= betaX_true & (beta_X + 1.96*se_X) >= betaX_true)
    coverage_M <- mean((beta_M - 1.96*se_M) <= betaM_true & (beta_M + 1.96*se_M) >= betaM_true)
    coverage_S <- mean((beta_S - 1.96*se_S) <= betaS_true & (beta_S + 1.96*se_S) >= betaS_true)
    data.frame(
      n=n,
      mean_X = mean(beta_X), bias_X = mean(beta_X)-betaX_true, coverage_X = coverage_X,
      mean_M = mean(beta_M), bias_M = mean(beta_M)-betaM_true, coverage_M = coverage_M,
      mean_S = mean(beta_S), bias_S = mean(beta_S)-betaS_true, coverage_S = coverage_S
    )
  })
  rbindlist(out)
}
```

Now, we compare the ML estimates with the true value $\beta_{\mathbf{X}} = \mathbf{1}$, $\beta_{\mathbf{M}} = \mathbf{1}$ and $\beta_{\mathbf{S}} = \mathbf{1}$. I am repeating the simulation for increasing values of n reaching towards infinity to discover the asymptotic behavior accurately.

```
# should be unbiased for Fig 1D
n_vec <- c(1000, 2000, 5000, 8000, 10000)
reps  <- 100

res_1D  <- eval_ML_grid(n_vec, reps, "fig1D")
print(res_1D)
```

```
##        n   mean_X     bias_X coverage_X   mean_M     bias_M coverage_M   mean_S
##    <num>    <num>      <num>      <num>    <num>      <num>      <num>    <num>
## 1:  1000 1.029078 0.02907830       0.87 1.010887 0.01088740       0.95 1.001364
## 2:  2000 1.031134 0.03113446       0.87 1.018889 0.01888927       0.94 1.019498
## 3:  5000 1.031765 0.03176501       0.80 1.015710 0.01571023       0.95 1.016274
## 4:  8000 1.027078 0.02707833       0.75 1.014906 0.01490622       0.91 1.014726
## 5: 10000 1.024411 0.02441098       0.74 1.020694 0.02069388       0.89 1.015847
##         bias_S coverage_S
##          <num>      <num>
## 1: 0.001364479       0.92
## 2: 0.019498377       0.96
## 3: 0.016274490       0.92
## 4: 0.014726498       0.91
## 5: 0.015846913       0.91
```

As n increases, we see that the mean of ML estimates converges towards the true value 1, the biases converges towards 0, the coverage for β_M, β_S is above 95% and β_X is mostly above 75%. Hence, we can conclude that the maximum likelihood estimation of the regression coefficients is asymptotically unbiased for Figure 1D. We know that the missingness is MAR according to Figure 1D, hence, the ML estimates produced by `frm_em()` function match the true value.

Next for Figure 1C, we run ML estimation:

```
# should be biased for Fig 1C
res_1C  <- eval_ML_grid(n_vec, reps, "fig1C")
print(res_1C)
```

```
##        n   mean_X     bias_X coverage_X   mean_M     bias_M coverage_M   mean_S
##    <num>    <num>      <num>      <num>    <num>      <num>      <num>    <num>
## 1:  1000 1.052670 0.05267039       0.81 1.058232 0.05823196       0.91 1.021714
## 2:  2000 1.044082 0.04408188       0.80 1.062411 0.06241083       0.67 1.023946
```

```
## 3:  5000 1.044460 0.04446020       0.54 1.062039 0.06203924       0.52 1.030248
## 4:  8000 1.045955 0.04595547       0.32 1.064171 0.06417087       0.23 1.026879
## 5: 10000 1.047489 0.04748889       0.23 1.061752 0.06175214       0.15 1.027241
##        bias_S coverage_S
##         <num>      <num>
## 1: 0.02171447       0.95
## 2: 0.02394570       0.93
## 3: 0.03024828       0.80
## 4: 0.02687911       0.78
## 5: 0.02724081       0.73
```

Here, as n increases, the mean of the ML estimates don't converge to the true value, their biases don't converge towards 0 and the coverage for $\beta_X, \beta_M, \beta_S$ goes well below 75%. Hence, we can say that the ML estimation of these regression coefficients is biased for Figure 1C where the missingness of the exposure depends on the exposure itself. Hence, ML estimates couldn't truly represent the true values as the un-explainable missingness was causing the estimates to converge to a pseudo value instead.

Question 5

Now, for the same linear regression model, let us check whether the complete case analysis estimates are asymptotically unbiased or biased for the causal diagram in Figure 1C and 1D.

In Complete Case Analysis (CCA), we only consider the data that is observed for all variables in the main analysis. To calculate the CCA estimates of the regression coefficients, we simply fit the `lm` model. We reuse the same parameters and function to generate the data from Question 4.

```
fit_lm_once <- function(n, dag) {
  dat <- generate_data(n, dag)
  fit <- lm(formula=Y ~ X + M + S, data=dat)
  cf  <- coef(fit)
  beta_X  <- unname(cf["X"])
  beta_M  <- unname(cf["M"])
  beta_S  <- unname(cf["S"])

  c(beta_X=beta_X, beta_M=beta_M, beta_S=beta_S)
}

eval_CCA_grid <- function(n_vec, reps, dag){
  out <- lapply(n_vec, function(n){
    M <- replicate(reps, fit_lm_once(n, dag))
    beta_X <- M["beta_X",]
    beta_M <- M["beta_M",]
    beta_S <- M["beta_S",]
    data.frame(
      n=n,
      mean_X = mean(beta_X), bias_X = mean(beta_X)-betaX_true,
      mean_M = mean(beta_M), bias_M = mean(beta_M)-betaM_true,
      mean_S = mean(beta_S), bias_S = mean(beta_S)-betaS_true
    )
  })
  rbindlist(out)
}
```

Now, we compare the CCA estimates with the true value $\beta_{\mathbf{X}} = \mathbf{1}$.

```r
# should be unbiased for Fig 1C
res_1C  <- eval_CCA_grid(n_vec, reps, "fig1C")
print(res_1C)
```

```
##        n    mean_X        bias_X    mean_M        bias_M    mean_S
##    <num>     <num>         <num>     <num>         <num>     <num>
## 1:  1000 0.9856513 -0.0143486867 1.0085006  0.0085005886 1.0093546
## 2:  2000 1.0009019  0.0009019425 0.9934204 -0.0065796040 1.0030847
## 3:  5000 0.9985570 -0.0014430469 1.0002286  0.0002285886 0.9999393
## 4:  8000 1.0009847  0.0009846515 1.0003259  0.0003259403 0.9995387
## 5: 10000 0.9942979 -0.0057020952 1.0014938  0.0014937961 1.0025133
##           bias_S
##            <num>
## 1:  9.354606e-03
## 2:  3.084707e-03
## 3: -6.070272e-05
## 4: -4.612980e-04
## 5:  2.513282e-03
```

As n increases, we see that the mean of CCA estimates converges towards the true value 1 and their biases converges towards 0. Hence, we can conclude that the complete case analysis estimation of the regression coefficients is asymptotically unbiased for Figure 1C. In such cases where the missingness of the exposure depends on the exposure itself, discarding the incomplete data is more beneficial towards being unbiased.

Next for Figure 1D, we run CCA estimation:

```r
# should be biased for Fig 1D
res_1D  <- eval_CCA_grid(n_vec, reps, "fig1D")
print(res_1D)
```

```
##        n    mean_X      bias_X    mean_M      bias_M    mean_S      bias_S
##    <num>     <num>       <num>     <num>       <num>     <num>       <num>
## 1:  1000 0.9767942 -0.02320580 0.9388055 -0.06119447 0.9743542 -0.02564582
## 2:  2000 0.9696315 -0.03036849 0.9389726 -0.06102739 0.9677601 -0.03223989
## 3:  5000 0.9738943 -0.02610566 0.9313804 -0.06861958 0.9792456 -0.02075436
## 4:  8000 0.9793701 -0.02062988 0.9406836 -0.05931640 0.9741469 -0.02585314
## 5: 10000 0.9746661 -0.02533388 0.9406012 -0.05939876 0.9802368 -0.01976320
```

Here, as n increases, the mean of the CCA estimates doesn't converge to the true value and the bias doesn't converge towards 0. Hence, we can say that the CCA estimation of these regression coefficients is biased for Figure 1D where the missingness of the exposure depends on observable and explainable variables, in this case, the outcome and a covariate. Here, we are potentially discarding an entire set of outcome data as they are missing due to their outcome, leading to a bias in the CCA estimates.

The results above also confirm the potential bias of exposure regression coefficient in CCA based on linear regression mentioned in the Table 1.

Question 6

Now, let us look at the mathematical argument as to why CCA is unbiased for Figure 1C, as seen with the help of estimates above. As we know that the missingness of the exposure in Figure 1C is dependent on a covariate and the exposure itself, hence, **MNAR**.

Hence, the missingness of the exposure in Figure 1C is independent of the outcome given exposure and the covariate $MaternalSubstanceUse$.

We can say that:

$\mathbf{R} \perp \mathbf{Y}|(\mathbf{X}, \mathbf{M}) \Leftrightarrow \mathbf{R} \perp \mathbf{Y}|(\mathbf{X}, \mathbf{M}, \mathbf{S})$

where **S** is Sex which also is not associated with the missingness in Cannabis Use.

To prove that CCA is unbiased, we can look at the true estimator and the CCA estimator of Y and check if they are identical.

The true estimator is $\mathbf{E}[\mathbf{Y} \mid \mathbf{X}, \mathbf{M}, \mathbf{S}]$.

The CCA estimator is $\mathbf{E}[\mathbf{Y} \mid \mathbf{X}, \mathbf{M}, \mathbf{S}, \mathbf{R} = \mathbf{1}]$.

But, for the figure 1C, due to the conditional independence of R and Y given (X, M, S), we can say that:

$\mathbf{P}(\mathbf{Y} \mid \mathbf{X}, \mathbf{M}, \mathbf{S}, \mathbf{R} = \mathbf{1}) = \mathbf{P}(\mathbf{Y} \mid \mathbf{X}, \mathbf{M}, \mathbf{S})$

Hence, the CCA estimator for figure 1C is deduced to:

$\mathbf{E}[\mathbf{Y} \mid \mathbf{X}, \mathbf{M}, \mathbf{S}, \mathbf{R} = \mathbf{1}] \Rightarrow \mathbf{E}[\mathbf{Y} \mid \mathbf{X}, \mathbf{M}, \mathbf{S}]$.

Therefore, we can say that the CCA is unbiased when the missingness pattern is conditionally independent of the outcome given exposure and covariates as seen in Figure 1C.

Question 7

The NPSEM corresponding to the causal diagram in the figure 1C looks like this:

$$\left.\begin{aligned} S &:= f_S(\varepsilon_S) \\ M &:= f_M(\varepsilon_M) \\ X &:= f_X(M, S, \varepsilon_X) \\ Y &:= f_Y(X, M, S, \varepsilon_Y) \\ R &:= f_R(M, X, \varepsilon_R) \end{aligned}\right\}$$

We can also rewrite this as:

$$\left.\begin{aligned} S &:= f_S(\varepsilon_S) \\ M &:= f_M(\varepsilon_M) \\ X &:= f_X(f_M(\varepsilon_M), f_S(\varepsilon_S), \varepsilon_X) \\ Y &:= f_Y(f_X(f_M(\varepsilon_M), f_S(\varepsilon_S), \varepsilon_X), f_M(\varepsilon_M), f_S(\varepsilon_S), \varepsilon_Y) \\ R &:= f_R(f_M(\varepsilon_M), f_X(f_M(\varepsilon_M), f_S(\varepsilon_S), \varepsilon_X), \varepsilon_R) \end{aligned}\right\}$$

Since, the auxiliary variables are not included in the main analysis and the absence of missing data the main analysis is assumed to give unbiased results, I have not added any additional factor U contributing to the variables. Hence, we can assume that the error terms ε_S, ε_M, ε_X, ε_R and ε_Y are independent.

Similarly, the NPSEM of figure 1D is:

$$\left.\begin{aligned} S &:= f_S(\varepsilon_S) \\ M &:= f_M(\varepsilon_M) \\ X &:= f_X(M, S, \varepsilon_X) \\ Y &:= f_Y(X, M, S, \varepsilon_Y) \\ R &:= f_R(M, Y, \varepsilon_R) \end{aligned}\right\}$$

also, represented as:

$$\left.\begin{aligned} S &:= f_S(\varepsilon_S) \\ M &:= f_M(\varepsilon_M) \\ X &:= f_X(f_M(\varepsilon_M), f_S(\varepsilon_S), \varepsilon_X) \\ Y &:= f_Y(f_X(f_M(\varepsilon_M), f_S(\varepsilon_S), \varepsilon_X), f_M(\varepsilon_M), f_S(\varepsilon_S), \varepsilon_Y) \\ R &:= f_R(f_M(\varepsilon_M), f_Y(f_X(f_M(\varepsilon_M), f_S(\varepsilon_S), \varepsilon_X), f_M(\varepsilon_M), f_S(\varepsilon_S), \varepsilon_Y), \varepsilon_R) \end{aligned}\right\}$$

The error terms are assumed to be independent here as well.

Question 8

Let us check if we see a statistical association between $CannabisUse$ and $Outcome$ when adjusted for $MaternalSubstanceUse$ and Sex, prove a causal effect from $CannabisUse$ on the $Outcome$ in the case of Figure 1C and 1D. There are several open paths from $CannabisUse$ to $Outcome$ in both the figures. Once we block all these open paths passing through various covariates and still see a statistical association (after adjusting for these covariates), then it means that there is a causal effect from $CannabisUse$ to $Outcome$. Hence, assuming that there is no direct arrow from $CannabisUse$ to $Outcome$, if we can show that they are d-separated conditional on covariates, we can prove the causal effect.

First, for Figure 1C:

The open paths from $CannabisUse$ to $Outcome$, assuming that there is no direct arrow from $CannabisUse$ to $Outcome$:

1. $$CannabisUse \leftarrow Sex \rightarrow Outcome$$

where Sex is a fork and it needs to be conditioned for to block the path by the first rule.

2. $$CannabisUse \leftarrow MaternalSubstanceUse \rightarrow Outcome$$

where $MaternalSubstanceUse$ is a fork as well and blocks the open path when conditioned for.

3. $$CannabisUse \rightarrow Miss_{cu} \leftarrow MaternalSubstanceUse \rightarrow Outcome$$

where $Miss_{cu}$ is an inverted fork and need not be conditioned for to block the path by the second rule. Conditioning the fork $MaternalSubstanceUse$ is enough to block this path.

Hence,

$$CannabisUse \perp_d Outcome \mid MaternalSubstanceUse, Sex$$

meaning there are $CannabisUse$ and $Outcome$ are statistically independent when adjusted for $MaternalSubstanceUse$ and Sex. And if we see a statistical association in the complete case analysis, which is asymptotically unbiased for Figure 1C, we can say that CCA is a valid test to prove that there is a direct causal effect from $CannabisUse$ to $Outcome$.

Now, for Figure 1D:

The open paths from $CannabisUse$ to $Outcome$, assuming that there is no direct arrow from $CannabisUse$ to $Outcome$:

1. $$CannabisUse \leftarrow Sex \rightarrow Outcome$$

2. $$CannabisUse \leftarrow MaternalSubstanceUse \rightarrow Outcome$$

3. $$CannabisUse \leftarrow MaternalSubstanceUse \rightarrow Miss_{cu} \leftarrow Outcome$$

and they can all be blocked by conditioning on $MaternalSubstanceUse$ and Sex and not conditioning on $Miss_{cu}$. Hence, again

$$CannabisUse \perp_d Outcome \mid MaternalSubstanceUse, Sex$$

meaning they are statistically independent when adjusted for $MaternalSubstanceUse$ and Sex. And if we see a statistical association in the complete case analysis we can say that there is a direct causal effect from $CannabisUse$ to $Outcome$. Even though, CCA is asymptotically biased for Figure 1D, I would say that it can be used to prove the causal effect by statistical association, but maybe not for estimating the causal effect.

Question 9

The assignment took 21 hours to complete.