Observational Studies Versus RCTs

Kevin E. Thorpe

Introduction

These simulations illustrate the effect of selection bias on the estimates of treatment effect compared to randomized controlled trials. A common goal in medical research is to determine whether one treatment is better than another. Such inference is challenging with observational data because of something called (treatment) selection bias. Selection bias is a process whereby certain patient characteristics are responsible for the choice os treatment a physician makes for a patient. However, these same characteristics are also related to subsequent outcomes.

For example, it could be the case that *sicker* patients are mare likely to get one treatment however these patients are also less likely to *recover*. In a randomized trial, treatment choice is randomly assigned, thereby removing the selection bias present in the observational setting.

The simulations that follow illustrate the impact that selection bias can have on the estimation of a treatment effect and how randomization solves the issue. The selection bias for all simulations will be set up as follows: `rx` will be a binary treatment variable (1 for treatment and 0 for control) and `x` will be a variable associated with outcome and treatment selection (selection bias or confounder in epidemiology language).

$$\texttt{X}_i \sim \text{Ber}(0.5)$$

$$\texttt{RX}_i \sim \begin{cases} \text{Ber}(0.8), \texttt{x} = 0 \\ \text{Ber}(0.2), \texttt{x} = 1 \end{cases}$$

Finally, for the RCT, the treatment variable will be `rx1` and will be generated as $\texttt{RX1}_i \sim \text{Ber}(0.5)$.

For all simulations, the response will be continuous with larger values corresponding to better outcomes and the variable `x` will be assoiciated with worse response when present (i.e. equal to 1).

Simulation 1

Assume no treatment effect. In otherwords, the true model is:

$$y_i = 50 + 0 \times \texttt{rx}_i - 10\texttt{x}_i + \varepsilon_i, \varepsilon_i \sim N(0, 15^2)$$

```
set.seed(123654)

x <- rbinom(500, 1, 0.5)
rx <- sapply(x, function(x) if (x==0) rbinom(1,1,0.8) else rbinom(1,1,0.2))
eps <- rnorm(500,0,15)
y <- 50 - 10*x + eps
rx1 <- rbinom(500, 1, 0.5) # RCT
```

Notice that, despite the dependence on `x`, the treatment group is quite *balanced*.

```r
table(rx)
```

```
## rx
##   0   1
## 241 259
```

First, I present the unadjusted analysis for the treatment effect.

```r
summary(lm(y~rx))
```

```
##
## Call:
## lm(formula = y ~ rx)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.731 -10.205  -1.081  10.913  44.731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.2413     0.9721  43.453  < 2e-16 ***
## rx            7.1273     1.3507   5.277 1.97e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.09 on 498 degrees of freedom
## Multiple R-squared:  0.05295,    Adjusted R-squared:  0.05105
## F-statistic: 27.84 on 1 and 498 DF,  p-value: 1.965e-07
```

```r
confint(lm(y~rx))
```

```
##                 2.5 %    97.5 %
## (Intercept) 40.331357 44.151276
## rx           4.473581  9.781073
```

From this analysis, the inference obtained is that the treatment is strongly associated with an improvement in outcome. This is clearly the wrong inferencce.

Now, consider the analysis in which the selection feature is adjusted for.

```r
summary(lm(y~rx+x))
```

```
##
## Call:
## lm(formula = y ~ rx + x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.636  -9.122  -0.551  10.042  42.826
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.184      1.627   30.84  < 2e-16 ***
## rx             1.089      1.651    0.66     0.51
## x             -9.868      1.650   -5.98 4.27e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 14.59 on 497 degrees of freedom
## Multiple R-squared:  0.1165, Adjusted R-squared:  0.113
## F-statistic: 32.77 on 2 and 497 DF,  p-value: 4.269e-14
```

```
confint(lm(y~rx+x))
```

```
##                 2.5 %    97.5 %
## (Intercept)  46.987389 53.381563
## rx           -2.154219  4.332413
## x           -13.109681 -6.625385
```

In this analysis, we would not infer a treatment effect and we see the effect of the selection feature is close to a reduction of 10 (the true relationship). To understand what is going on, observe:

```
xtabs(~x+rx)
```

```
##    rx
## x     0   1
##   0  47 209
##   1 194  50
```

We have $\texttt{x} = 1$ is associated with worse outcome but also most of the most of the control patients have $\texttt{x} = 1$ or conversely, most of the treatment subjects correspond to $\texttt{x} = 0$ which have a better outcome.

Finally, consider the unadjusted and adjusted analyses of these data using the randomly assigned treatment (`rx1`).

```
# Unadjusted
summary(lm(y~rx1))
```

```
##
## Call:
## lm(formula = y ~ rx1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.974 -10.528  -0.859  10.908  47.852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45.6117     0.9865  46.235   <2e-16 ***
## rx1           0.6354     1.3868   0.458    0.647
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.5 on 498 degrees of freedom
## Multiple R-squared:  0.0004214,  Adjusted R-squared:  -0.001586
## F-statistic: 0.2099 on 1 and 498 DF,  p-value: 0.647
```

```
confint(lm(y~rx1))
```

```
##                 2.5 %    97.5 %
## (Intercept) 43.673502 47.549976
## rx1         -2.089341  3.360222
```

```
# Adjusted
summary(lm(y~rx1+x))
```

```
##
## Call:
## lm(formula = y ~ rx1 + x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.257  -9.187  -0.363  10.203  42.860
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  50.8950     1.1370  44.764  < 2e-16 ***
## rx1           0.3438     1.3061   0.263    0.793
## x           -10.5240     1.3064  -8.056  5.9e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.6 on 497 degrees of freedom
## Multiple R-squared:  0.1159, Adjusted R-squared:  0.1123
## F-statistic: 32.57 on 2 and 497 DF,  p-value: 5.126e-14
```

```
confint(lm(y~rx1+x))
```

```
##                 2.5 %    97.5 %
## (Intercept)  48.661197 53.128860
## rx1          -2.222426  2.909943
## x           -13.090710 -7.957232
```

These two analyses are now consistent and lead to the correct inference.

Simulation 2

This time we suppose treatment improves outcome by 5, on average. The true model is:

$$y_i = 50 + 5\mathtt{rx}_i - 10\mathtt{x}_i + \varepsilon_i, \varepsilon_i \sim N(0, 15^2)$$

```
y1 <- 50 + 5*rx - 10*x + eps # Observational
y2 <- 50 + 5*rx1 - 10*x + eps # RCT
```

First, consider the observational analyses.

```
# Unadjusted
summary(lm(y1~rx))
```

```
##
## Call:
## lm(formula = y1 ~ rx)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.731 -10.205  -1.081  10.913  44.731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.2413     0.9721  43.453   <2e-16 ***
## rx           12.1273     1.3507   8.979   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.09 on 498 degrees of freedom
## Multiple R-squared:  0.1393, Adjusted R-squared:  0.1376
## F-statistic: 80.62 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
confint(lm(y1~rx))
```

```
##                 2.5 %   97.5 %
## (Intercept) 40.331357 44.15128
## rx           9.473581 14.78107
```

```
# Adjusted
summary(lm(y1~rx+x))
```

```
##
## Call:
## lm(formula = y1 ~ rx + x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.636  -9.122  -0.551  10.042  42.826
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.184      1.627  30.841  < 2e-16 ***
## rx             6.089      1.651   3.689  0.00025 ***
## x             -9.868      1.650  -5.980 4.27e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.59 on 497 degrees of freedom
## Multiple R-squared:  0.1971, Adjusted R-squared:  0.1939
## F-statistic:    61 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
confint(lm(y1~rx+x))
```

```
##                  2.5 %    97.5 %
## (Intercept)  46.987389 53.381563
## rx            2.845781  9.332413
## x           -13.109681 -6.625385
```

What we see here is that the unadjusted analysis grossly overestimates the treatment effect while the adjusted analysis gets it about right.

Now consider the RCT analyses.

```
# Unadjusted
summary(lm(y2~rx1))
```

```
##
## Call:
## lm(formula = y2 ~ rx1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.974 -10.528  -0.859  10.908  47.852
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45.6117     0.9865  46.235  < 2e-16 ***
## rx1           5.6354     1.3868   4.064 5.62e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.5 on 498 degrees of freedom
## Multiple R-squared:  0.03209,    Adjusted R-squared:  0.03015
## F-statistic: 16.51 on 1 and 498 DF,  p-value: 5.616e-05
```

```
confint(lm(y2~rx1))
```

```
##                 2.5 %    97.5 %
## (Intercept) 43.673502 47.549976
## rx1          2.910659  8.360222
```

```
# Adjusted
summary(lm(y2~rx1+x))
```

```
##
## Call:
## lm(formula = y2 ~ rx1 + x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.257  -9.187  -0.363  10.203  42.860
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.895      1.137  44.764  < 2e-16 ***
## rx1            5.344      1.306   4.091  5.0e-05 ***
## x            -10.524      1.306  -8.056  5.9e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.6 on 497 degrees of freedom
## Multiple R-squared:  0.1439, Adjusted R-squared:  0.1404
## F-statistic: 41.76 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
confint(lm(y2~rx1+x))
```

```
##                  2.5 %    97.5 %
## (Intercept)  48.661197 53.128860
## rx1           2.777574  7.909943
## x           -13.090710 -7.957232
```

Again, both analyses are consistent and correct.

Simulation 3

Finally, suppose rx actually worsens response by 5 on average. The true model is:

$$y_i = 50 - 5\texttt{rx}_i - 10\texttt{x}_i + \varepsilon_i, \varepsilon_i \sim N(0, 15^2)$$

```r
y3 <- 50 - 5*rx - 10*x + eps # Observational
y4 <- 50 - 5*rx1 - 10*x + eps # RCT
```

Again, start with the observational analysis.

```r
# Unadjusted
summary(lm(y3~rx))
```

```
##
## Call:
## lm(formula = y3 ~ rx)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.731 -10.205  -1.081  10.913  44.731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.2413     0.9721  43.453   <2e-16 ***
## rx            2.1273     1.3507   1.575    0.116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.09 on 498 degrees of freedom
## Multiple R-squared:  0.004956,   Adjusted R-squared:  0.002958
## F-statistic: 2.481 on 1 and 498 DF,  p-value: 0.1159
```

```r
confint(lm(y3~rx))
```

```
##                   2.5 %    97.5 %
## (Intercept) 40.3313566 44.151276
## rx          -0.5264192  4.781073
```

```r
# Adjusted
summary(lm(y3~rx+x))
```

```
##
## Call:
## lm(formula = y3 ~ rx + x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.636  -9.122  -0.551  10.042  42.826
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.184      1.627  30.841  < 2e-16 ***
## rx            -3.911      1.651  -2.369   0.0182 *
## x             -9.868      1.650  -5.980 4.27e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.59 on 497 degrees of freedom
## Multiple R-squared:  0.07174,    Adjusted R-squared:  0.06801
## F-statistic: 19.21 on 2 and 497 DF,  p-value: 9.242e-09
```

```r
confint(lm(y3~rx+x))
```

```
##                 2.5 %     97.5 %
## (Intercept)  46.987389 53.3815627
## rx           -7.154219 -0.6675871
## x           -13.109681 -6.6253852
```

Again we see the unadjusted analysis getis it completely wrong while the adjusted analysis gets it right.

Finally, we have the RCT analysis.

```r
# Unadjusted
summary(lm(y4~rx1))
```

```
##
## Call:
## lm(formula = y4 ~ rx1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.974 -10.528  -0.859  10.908  47.852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45.6117     0.9865  46.235  < 2e-16 ***
## rx1          -4.3646     1.3868  -3.147  0.00175 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.5 on 498 degrees of freedom
## Multiple R-squared:  0.0195, Adjusted R-squared:  0.01753
## F-statistic: 9.904 on 1 and 498 DF,  p-value: 0.001748
```

```r
confint(lm(y4~rx1))
```

```
##                 2.5 %    97.5 %
## (Intercept) 43.673502 47.549976
## rx1         -7.089341 -1.639778
```

```r
# Adjusted
summary(lm(y4~rx1+x))
```

```
##
## Call:
## lm(formula = y4 ~ rx1 + x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.257  -9.187  -0.363  10.203  42.860
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.895      1.137  44.764  < 2e-16 ***
## rx1           -4.656      1.306  -3.565 0.000399 ***
## x            -10.524      1.306  -8.056  5.9e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 14.6 on 497 degrees of freedom
## Multiple R-squared:  0.1327, Adjusted R-squared:  0.1293
## F-statistic: 38.03 on 2 and 497 DF,  p-value: 4.265e-16
```

```
confint(lm(y4~rx1+x))
```

```
##                 2.5 %    97.5 %
## (Intercept)  48.661197 53.128860
## rx1          -7.222426 -2.090057
## x           -13.090710 -7.957232
```

As with the other simulations, both the unadjusted and adjusted analyses get the *right* answer.

Implications

You may still be wondering by bother with RCTs when all you need to do is adjust for the variable responsible for treatment selection bias. If reality were that simple, we could. The problem is that there are always multiple factors at work that inluence the choice of treatment physicians make for their patients. Moreover, we cannot actually measure all of them. That means that in practice we cannot adjust for the treatment selection bias.

On the other hand, what we have seen is that the RCT removes selection bias and that we do not need to adjust for it in order to obtain correct results, subject to the usual statistical errors.