# Contamination Bias in Linear Regressions

Sai Zhang

August 09, 2023

## Contents



## 1 Introduction

The package `multeR` implements contamination bias decomposition and bias-aware estimation for linear regressions considered in Goldsmith-Pinkham et al. [2022]. In this vignette, we demonstrate the implementation using Project STAR data as in Krueger [1999], which are included in the package as a data frame `star`.

## 2 Model

```
library(multeR)
```

## 3 Bias-aware treatment effect estimation

```
multe(score ~ treatment | school, data = star, base_val = "regular")
#> 5868 observations without missing values in the treatment variable or control variables kept
#> Treatment Effect Estimates (with robust standard errors):
#> ATE
#>            Coef. Robust S.E.    z P>|z| 95% Confidence Interval
#> aide  0.0698381   0.7082876 0.10 0.921 (-1.3183802, 1.4580563)
```

```
#> small 5.5614049   0.7626419 7.29 0.000  (4.0666542, 7.0561555)
#> One-at-a-time
#>           Coef. Robust S.E.    z P>|z| 95% Confidence Interval
#> aide  0.2630480   0.7145257 0.37 0.713 (-1.1373966, 1.6634927)
#> small 5.2951470   0.7751078 6.83 0.000  (3.7759636, 6.8143305)
#> Common weights
#>            Coef. Robust S.E.     z P>|z| 95% Confidence Interval
#> aide  -0.0028816   0.7123410 -0.00 0.997 (-1.3990443, 1.3932810)
#> small  5.5630023   0.7640788  7.28 0.000  (4.0654355, 7.0605691)
multe(score ~ treatment | school, data = star, base_val = "regular",
    vce = "oracle")
#> 5868 observations without missing values in the treatment variable or control variables kept
#> Treatment Effect Estimates (with oracle standard errors):
#> ATE
#>           Coef. Oracle S.E.    z P>|z| 95% Confidence Interval
#> aide  0.0698381   0.6938996 0.10 0.920 (-1.2901802, 1.4298564)
#> small 5.5614049   0.7444408 7.47 0.000  (4.1023278, 7.0204819)
#> One-at-a-time
#>           Coef. Oracle S.E.    z P>|z| 95% Confidence Interval
#> aide  0.2630480   0.6912518 0.38 0.704 (-1.0917805, 1.6178766)
#> small 5.2951470   0.7429330 7.13 0.000  (3.8390251, 6.7512689)
#> Common weights
#>            Coef. Oracle S.E.     z P>|z| 95% Confidence Interval
#> aide  -0.0028816   0.6950073 -0.00 0.997 (-1.3650708, 1.3593075)
#> small  5.5630023   0.7420179  7.50 0.000  (4.1086740, 7.0173306)
```

## 4 Contamination bias decomposition

```
multe(score ~ treatment | school, data = star, base_val = "regular",
    decomp = TRUE)
#> 5868 observations without missing values in the treatment variable or control variables kept
#> Treatment Effect Estimates (with robust standard errors):
#> ATE
#>           Coef. Robust S.E.    z P>|z| 95% Confidence Interval
#> aide  0.0698381   0.7082876 0.10 0.921 (-1.3183802, 1.4580563)
#> small 5.5614049   0.7626419 7.29 0.000  (4.0666542, 7.0561555)
#> One-at-a-time
#>           Coef. Robust S.E.    z P>|z| 95% Confidence Interval
#> aide  0.2630480   0.7145257 0.37 0.713 (-1.1373966, 1.6634927)
#> small 5.2951470   0.7751078 6.83 0.000  (3.7759636, 6.8143305)
#> Common weights
#>            Coef. Robust S.E.     z P>|z| 95% Confidence Interval
```

```
#> aide  -0.0028816   0.7123410 -0.00 0.997 (-1.3990443, 1.3932810)
#> small  5.5630023   0.7640788  7.28 0.000  (4.0654355, 7.0605691)
#>
#> Contamination Bias Decomposition:
#>             Coef.  Own Effect        Bias
#> aide    0.1769329   0.3595971  -0.1826642
#>       (0.7198171) (0.7138820) (0.1493920)
#> small   5.3571292   5.2020420   0.1550872
#>       (0.7782534) (0.7775252) (0.1603352)
multe(score ~ treatment | school, data = star, base_val = "regular",
    decomp = TRUE, minmax = TRUE)
#> 5868 observations without missing values in the treatment variable or control variables kept
#> Treatment Effect Estimates (with robust standard errors):
#> ATE
#>           Coef. Robust S.E.    z P>|z| 95% Confidence Interval
#> aide  0.0698381   0.7082876 0.10 0.921 (-1.3183802, 1.4580563)
#> small 5.5614049   0.7626419 7.29 0.000  (4.0666542, 7.0561555)
#> One-at-a-time
#>           Coef. Robust S.E.    z P>|z| 95% Confidence Interval
#> aide  0.2630480   0.7145257 0.37 0.713 (-1.1373966, 1.6634927)
#> small 5.2951470   0.7751078 6.83 0.000  (3.7759636, 6.8143305)
#> Common weights
#>            Coef. Robust S.E.     z P>|z| 95% Confidence Interval
#> aide  -0.0028816   0.7123410 -0.00 0.997 (-1.3990443, 1.3932810)
#> small  5.5630023   0.7640788  7.28 0.000  (4.0654355, 7.0605691)
#>
#> Contamination Bias Decomposition:
#>             Coef.  Own Effect        Bias    Min Bias    Max Bias
#> aide    0.1769329   0.3595971  -0.1826642  -1.5292799   1.5302748
#>       (0.7198171) (0.7138820) (0.1493920) (0.1757464) (0.1771296)
#> small   5.3571292   5.2020420   0.1550872  -1.6538811   1.6696356
#>       (0.7782534) (0.7775252) (0.1603352) (0.1850315) (0.1874256)
```

## 5 Save implicit regression weights and group-specific treatment effects

One can save the implicit ATE regression weights and the saturated group-specific treatment effects, then calculate the correlations to get a sense of how much contamination bias might affect estimates from such a model.

```
d <- multe(score ~ treatment | school, data = star, base_val = "regular",
    decomp = TRUE, save_lambda_as = "lambda", save_tau_as = "tau",
    print = FALSE)
#> 5868 observations without missing values in the treatment variable or control variables kept
```

```r
lambda <- d$decomposition$lambda_saved
tau <- d$decomposition$tauhat_saved
cor(lambda, tau)
#>                tau1        tau2
#> lambda11  0.2446820  0.24042214
#> lambda12 -0.1474606 -0.12698690
#> lambda21  0.1000662  0.08396965
#> lambda22 -0.2557241 -0.19388617
```

## 6 Weak overlapping

Weak overlapping strata will be dropped, as shown in this pseudo example.

```r
df <- star
df$moc_ctrl <- rep(1:5, length.out = length(df$score))
multe(score ~ treatment | school + moc_ctrl, data = df,
    base_val = "regular", decomp = T)
#> Drop 6 control strata without sufficient overlapping (79 observations dropped)
#> 5789 observations without missing values in the treatment variable or control variables kept
#> Treatment Effect Estimates (with robust standard errors):
#> ATE
#>            Coef. Robust S.E.     z P>|z| 95% Confidence Interval
#> aide  -0.3073748   0.6863340 -0.45 0.654 (-1.6525646, 1.0378151)
#> small  4.9522648   0.7195697  6.88 0.000  (3.5419341, 6.3625956)
#> One-at-a-time
#>            Coef. Robust S.E.     z P>|z| 95% Confidence Interval
#> aide  -0.0566102   0.7309134 -0.08 0.938 (-1.4891741, 1.3759536)
#> small  5.0733256   0.7814407  6.49 0.000  (3.5417299, 6.6049213)
#> Common weights
#>            Coef. Robust S.E.     z P>|z| 95% Confidence Interval
#> aide  -0.4466127   0.7293089 -0.61 0.540 (-1.8760319, 0.9828065)
#> small  5.2943652   0.7632317  6.94 0.000  (3.7984586, 6.7902718)
#>
#> Contamination Bias Decomposition:
#>             Coef.  Own Effect        Bias
#> aide   -0.0993316   0.0060547  -0.1053863
#>       (0.7328765) (0.7190665) (0.2351416)
#> small   5.2448935   4.9251195   0.3197740
#>       (0.7820081) (0.7731088) (0.2610353)
```

## Reference

Paul Goldsmith-Pinkham, Peter Hull, and Michal Kolesár. Contamination bias in linear regressions. Technical report, National Bureau of Economic Research, 2022.

Alan B Krueger. Experimental estimates of education production functions. *The quarterly journal of economics*, 114(2):497–532, 1999.