Multiple Treatment Effects Regression

Michal Kolesár

January 15, 2024

Contents

The package `multe` implements contamination bias diagnostics for regressions with multiple treatments developed in Goldsmith-Pinkham et al. [2022]. This vignette illustrates the methods using data from Fryer and Levitt [2013].

First, we fit a regression of test scores on a race dummy (treatment of interest) and a few controls, weighting using sampling weights:

```
library("multe")
## Regression of IQ at 24 months on race indicators
## and baseline controls
r1 <- stats::lm(std_iq_24 ~ race + factor(age_24) + female +
    SES_quintile, weight = W2C0, data = fl)
## Compute alternatives estimates free of
## contamination bias
m1 <- multe(r1, "race", cluster = NULL)
print(m1, digits = 3)
#> Estimates on full sample:
#>               PL     OWN     ATE      EW      CW
#> Black    -0.2574 -0.2482 -0.2655 -0.2550 -0.2604
#> SE        0.0281  0.0291  0.0298  0.0289  0.0292
#> Hispanic -0.2931 -0.2829 -0.2992 -0.2862 -0.2944
#> SE        0.0260  0.0267  0.0299  0.0268  0.0279
#> Asian    -0.2621 -0.2609 -0.2599 -0.2611 -0.2694
#> SE        0.0343  0.0343  0.0418  0.0343  0.0475
#> Other    -0.1563 -0.1448 -0.1503 -0.1447 -0.1522
#> SE        0.0369  0.0370  0.0359  0.0368  0.0370
```

```
#>
#> P-values for null hypothesis of no propensity score variation:
#> Wald test: 0, LM test: 0
```

The package reports five different estimators:

1. PL: The uninteracted regression estimator based on the partially linear (PL) model.
2. OWN: The own-treatment effect component of the contamination bias decomposition. If OWN is close to PL, as above, this indicates negligible contamination bias.
3. ATE: the unweighted ATE estimator.
4. EW: the efficiently weighted ATE estimator that runs one-treatment-at-a-time regression
5. CW: weighted ATE estimator using efficient common weights, as in Corollary 2 in Goldsmith-Pinkham et al. [2022].

Precise definitions of these estimators are given in the Methods section below.

In this example, the propensity score varies significantly with covariates, as indicated by the p-values of the Wald and LM tests.

Including many controls may result in overlap failure, as the next example demonstrates:

```
r2 <- stats::lm(std_iq_24 ~ race + factor(age_24) + female +
    SES_quintile + factor(siblings) + family_structure,
    weight = W2C0, data = fl)
m2 <- multe(r2, treatment = "race")
#> For variable factor(siblings) the following levels fail overlap:
#> 6
#> Dropping observations with these levels
print(m2, digits = 3)
#> Estimates on full sample:
#>               PL     OWN     ATE      EW      CW
#> Black    -0.2438 -0.2043 -0.2482 -0.2180 -0.2408
#> SE        0.0308  0.0332  0.0355  0.0328  0.0389
#> Hispanic -0.2928 -0.2801 -0.2878 -0.2850 -0.2964
#> SE        0.0259  0.0267  0.0300  0.0267  0.0299
#> Asian    -0.2739 -0.2836 -0.2742 -0.2839 -0.2916
#> SE        0.0342  0.0343  0.0420  0.0343  0.0459
#> Other    -0.1520 -0.1277      NA -0.1295 -0.1459
#> SE        0.0369  0.0374      NA  0.0371  0.0385
#>
#> Estimates on overlap sample:
#>               PL     OWN     ATE      EW      CW
#> Black    -0.2444 -0.2049 -0.2505 -0.2191 -0.2426
#> SE        0.0309  0.0334  0.0357  0.0329  0.0388
#> Hispanic -0.2915 -0.2791 -0.2871 -0.2839 -0.2974
#> SE        0.0259  0.0267  0.0300  0.0267  0.0299
#> Asian    -0.2766 -0.2863 -0.2769 -0.2865 -0.2924
#> SE        0.0344  0.0345  0.0421  0.0345  0.0459
#> Other    -0.1522 -0.1280 -0.1397 -0.1298 -0.1465
#> SE        0.0369  0.0374  0.0362  0.0371  0.0385
```

```
#>
#> P-values for null hypothesis of no propensity score variation:
#> Wald test: 0, LM test: 0
```

The issue is that observations with 6 siblings never have race equal to `other`:

```
table(fl$race[fl$siblings == 6])
#>
#>    White    Black Hispanic    Asian    Other
#>       18       10       12        5        0
```

Thus, the ATE estimator comparing `other` to `white` cannot be computed. The package drops observations with 6 siblings from the sample to form an "overlap sample" (see Methods section below for precise construction of this sample), where the all estimators are identified.

To check whether there is a significant difference between PL and the alternative estimators, the data frames `cb_f` and `cb_o` report differences between the estimates, along with standard errors for the full and overlap sample, respectively:

```
print(m2$cb_f, digits = 3)
#>          PL      OWN       ATE       EW       CW
#> Black    NA -0.03947  0.004397 -0.02573 -0.00292
#> pop_se   NA  0.01204  0.017297  0.01027  0.02078
#> Hispanic NA -0.01269 -0.004962 -0.00783  0.00362
#> pop_se   NA  0.00571  0.013143  0.00497  0.01274
#> Asian    NA  0.00975  0.000265  0.01001  0.01769
#> pop_se   NA  0.00487  0.024595  0.00480  0.02847
#> Other    NA -0.02430        NA -0.02246 -0.00605
#> pop_se   NA  0.00738        NA  0.00684  0.01334
```

We see statistically significant difference between the OWN and PL estimate (i.e. significant contamination bias) for all races.

Standard errors

The package also computes "oracle" standard errors, in addition to the usual standard errors reported above. These can be accessed in the `est_f` component of the output:

```
print(m1$est_f, digits = 3)
#>                PL     OWN     ATE      EW      CW
#> Black     -0.2574 -0.2482 -0.2655 -0.2550 -0.2604
#> pop_se     0.0281  0.0291  0.0298  0.0289  0.0292
#> oracle_se      NA      NA  0.0298  0.0288  0.0290
#> Hispanic  -0.2931 -0.2829 -0.2992 -0.2862 -0.2944
#> pop_se     0.0260  0.0267  0.0299  0.0268  0.0279
#> oracle_se      NA      NA  0.0299  0.0268  0.0278
#> Asian     -0.2621 -0.2609 -0.2599 -0.2611 -0.2694
#> pop_se     0.0343  0.0343  0.0418  0.0343  0.0475
#> oracle_se      NA      NA  0.0418  0.0344  0.0465
#> Other     -0.1563 -0.1448 -0.1503 -0.1447 -0.1522
```

```
#> pop_se     0.0369  0.0370  0.0359  0.0368  0.0370
#> oracle_se      NA      NA  0.0359  0.0360  0.0366
```

These oracle standard errors don't account for estimation error in the propensity score, see Methods section below. Specifying the `cluster` argument allows for computation of clustered standard errors:

```
## cluster in interviewer ID
m1alt <- multe(r1, "race", cluster = factor(factor(fl$interviewer_ID_24)))
print(m1alt, digits = 3)
#> Estimates on full sample:
#>               PL     OWN     ATE      EW      CW
#> Black    -0.2574 -0.2482 -0.2655 -0.2550 -0.2604
#> SE        0.0412  0.0425  0.0410  0.0422  0.0420
#> Hispanic -0.2931 -0.2829 -0.2992 -0.2862 -0.2944
#> SE        0.0441  0.0454  0.0495  0.0457  0.0474
#> Asian    -0.2621 -0.2609 -0.2599 -0.2611 -0.2694
#> SE        0.0521  0.0523  0.0619  0.0523  0.0675
#> Other    -0.1563 -0.1448 -0.1503 -0.1447 -0.1522
#> SE        0.0404  0.0416  0.0410  0.0414  0.0428
#>
#> P-values for null hypothesis of no propensity score variation:
#> Wald test: 0, LM test: 0
```

Methods

This section describes the implementation of the bias decomposition formula and the implementation of alternative estimators. Relative to Goldsmith-Pinkham et al. [2022], we generalize the setup to allow for sampling weights ω_i^2 (setting the sampling weights to one recovers unweighted formulas). We also explicitly deal with overlap issues.

We are interested in the effect of treatment $D_i \in \{0, 1, \ldots, K\}$ on an outcome Y_i. Let $X_i = \mathbb{1}\{D_i = 1, \ldots, D_i = K\}$ denote a vector of treatment indicators, let $X_{i0} = \mathbb{1}\{D_i = 0\}$, and let $Z_i = (1, W_i')'$ denote a vector of controls, including an intercept. We focus on the case where the controls enter linearly, so that control functions take the form $\mathcal{G} = \{z'\gamma \colon \gamma \in \mathbb{R}^{1+\dim(W_i)}\}$.

We assume that $\mu_k(W_i) := E[Y_i(k) \mid W_i] \in \mathcal{G}$, so that we may write $\mu_k(W_i) = W_i'\alpha_k$ for some vectors α_k, $k = 0, \ldots, K$. The average treatment effect (ATE) conditional on W_i is then given by $\tau(W_i) = Z_i'(\alpha_k - \alpha_0)$, and α_k correspond to the coefficients in the interacted regression

$$Y_i = \sum_{k=0}^{K} X_{ik} Z_i' \alpha_k + \dot{U}_i, \tag{1}$$

with $\dot{U}_i$ conditionally mean zero. The uninteracted estimator is given by estimating

$$Y_i = \sum_{k=1}^{K} X_{ik} \beta + Z_i' \phi + U_i, \tag{2}$$

by weighted least squares (WLS), yielding $\hat{\beta} = (\sum_i \omega_i^2 \dot{X}_i \dot{X}_i')^{-1} \sum_i \omega_i^2 \dot{X}_i Y_i$, where $\dot{X}$ is the sample residual from WLS regression of X_i onto Z_i. By Proposition 1 in Goldsmith-Pinkham et al. [2022],

the population analog of $\hat{\beta}$, β, has the decomposition

$$\beta = E[\text{diag}(\Lambda_i)\tau(W_i)] + E[\Lambda_i - \text{diag}(\Lambda_i)\tau(W_i)],$$

where $\Lambda_i = E[\tilde{X}_i\tilde{X}_i']^{-1}E[\tilde{X}_i X_i \mid W_i]$, and $\tilde{X}_i$ is the population analog of $\dot{X}_i$, the population residual from regressing X_i onto W_i. Let $\hat{\alpha}_k$ denote the WLS estimates based on (1). By construction, the sample residuals from estimating (1) and Z_i are both orthogonal to $\dot{X}_i$. As a result, we obtain the exact decomposition

$$\begin{aligned}\hat{\beta} &= E_n[\dot{X}_i\dot{X}_i']^{-1}E_n[\dot{X}_i Y_i] = \\ &= E_n[\text{diag}(\hat{\Lambda}_i)\hat{\Gamma}'Z_i] + E_n[(\hat{\Lambda}_i - \text{diag}(\hat{\Lambda}_i))\hat{\Gamma}'Z_i] =: \hat{\beta}^{\text{Own}} + \hat{\beta}^{\text{CB}},\end{aligned} \tag{3}$$

where $\hat{\Gamma}$ is a matrix with columns $\gamma_k = \alpha_k - \alpha_0$, $\hat{\Lambda}_i = E_n[\dot{X}_i\dot{X}_i']^{-1}\dot{X}_i X_i'$, and $E_n[A_i] = \sum_i \omega_i^2 A_i / \sum_i \omega_i^2$ denotes the weighted sample mean.

To compute this decomposition, we don't need to explicitly compute $\hat{\Lambda}_i$. Instead, we use the identity

$$\hat{\beta}_k^{\text{Own}} = e_k' E_n[\dot{X}_i\dot{X}_i']^{-1}E_n[\dot{X}_i X_{ik} Z_i'\hat{\gamma}_k] = \hat{\delta}_{kk}'\hat{\gamma}_k,$$

where $\hat{\delta}_{kk}$ is a WLS estimator of the system of regressions

$$Z_i X_{ik} = \delta_{kk} X_{ik} + \sum_{\ell \neq k} \delta_{k\ell} X_{i\ell} + \Delta_{Z,k} Z_i + \zeta_{ik}. \tag{4}$$

Note this decomposition and associated standard errors, in the next subsection, are purely regression-based, so they remain valid even if X_i is not a set of binary indicators. Likewise, misspecification of the interacted regression only affects the interpretation of the decomposition; if μ_k is not linear, the decomposition will not consistently estimate the contamination bias.

The own treatment weights in this decomposition sum to one, and the contamination weights sum to zero, since $E_n[\hat{\Lambda}_i] = I_k$, mimicking the population decomposition. If the propensity score doesn't satisfy $p_k \in \mathcal{G}$, the implied estimate of $\Lambda(w)$,

$$\hat{\Lambda}(w) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{W_i = w\}\hat{\Lambda}_i$$

may not be positive definite; in particular, the diagonal elements may not all be positive, in line with Proposition 1 in Goldsmith-Pinkham et al. [2022].

In addition to this decomposition, the package also computes the following alternative estimators:

1. The unweighted ATE estimator, $\hat{\beta}_k^{\text{ATE}} = E_n[W_i]'\hat{\gamma}_k$
2. The one-treatment-at-a time estimator $\hat{\beta}_k^{\text{EW}}$ that fits (2) using only observations with $D_i \in \{0, k\}$. In other words, it estimates the regression

$$Y_i = \ddot{\phi}_k + X_{ik}\ddot{\beta}_k + W_i'\ddot{\phi}_k + \ddot{U}_{ik}, \tag{5}$$

among observations with $D_i \in \{0, k\}$. This estimator weights the treatment effects using the variance-minimizing weighting scheme given in Corollary 1 in Goldsmith-Pinkham et al. [2022]. Consequently, we refer to as the efficiently weighted ATE estimator.

3. The common weights estimator $\hat{\beta}^{\text{CW}}$, given by the WLS regression of Y_i onto X_i, weighting each observation by

$$\frac{\omega_i^2 \pi_{D_i}(1-\pi_{D_i})}{\hat{p}_{D_i}(W_i)\sum_{k=0}^{K}\hat{p}_k(W_i)^{-1}},$$

where, by default, the probabilities π_k correspond to the marginal probability $E_n[X_{ik}]$ in the dataset. The propensity scores $\hat{p}_k(W_i)$ are based on fitting a multinomial logit model for the treatments. This estimator estimates a weighted ATE with weights $\lambda^{\text{CW}}(W_i) = \left(\sum_{k=0}^{K}\frac{\pi_k(1-\pi_k)}{p_k(W_i)}\right)^{-1}$. By Corollary 2 in Goldsmith-Pinkham et al. [2022], this weighting scheme minimizes the average variance, under homoskedasticity, across all treatment comparisons—comparisons of outcomes under treatment k vs treatment ℓ, if we draw the treatments k and ℓ independently from the marginal treatment distribution $(\pi_0, \ldots, \pi_K)$. Option `cw_uniform=TRUE` in the `multe` function sets these probabilities to $1/K$; setting the option to its default, `FALSE`, sets them to $(E_n[X_{i0}], \ldots, E_n[X_{iK}])$.

Standard errors

To compute cluster-robust standard errors for an asymptotically linear estimator with influence function ψ_i, we use the formula

$$\widehat{\text{se}}(\psi)^2 = \frac{G}{G-1}\sum_g\left(\sum_{G_i=g}\psi_i\right)\left(\sum_{G_i=g}\psi_i\right)'.$$

Here G_i denotes cluster membership, as specified by the `multe` argument `cluster`, and G the number of clusters. Specifying `cluster=NULL` assumes independent data, setting each observation to be in its own cluster ($G_i = i$ and $G = N$), so the reported standard errors are robust to heteroskedasticity, but not clustering.

We now describe the form of the influence function for the estimators above. For a generic WLS regression of A onto B, let $(Q_1, Q_2)\left(\begin{smallmatrix} R & S \\ 0 & 0 \end{smallmatrix}\right)\Pi'$ denote the QR decomposition of $\text{diag}(\omega_i)B$. If B has rank r, then R has dimension $r \times r$, Q has dimension $N \times r$, where r is the rank of the regressor matrix, and Π is a permutation matrix. The WLS estimator is then given by $b = \Pi\begin{pmatrix} R^{-1}Q_1'\,\text{diag}(\omega_i)A \\ \texttt{NA} \end{pmatrix}$. Denoting the regression residual by $\hat{\epsilon}_i$, the influence function is thus given by

$$\psi_i(b) = \Pi\begin{pmatrix} R^{-1}Q_{1i}\omega_i\hat{\epsilon}_i \\ \texttt{NA} \end{pmatrix}. \tag{6}$$

See the internal function `multe:::reg_if` for implementation. The influence function for the inner product of linear estimators a and b, is by the delta method given by

$$\psi_i(a'b) = a'\psi_i(b) + b'\psi_i(a),$$

while for scalars s_1, s_2, $\psi(s_1 a + s_2 b) = s_1\psi(a) + s_2\beta(b)$.

We use (6) to compute $\psi(\hat{\alpha}_k)$, as well as

$$\psi_i(\bar{Z}) = \frac{\omega_i^2(Z_i - \bar{Z})}{\sum_i \omega_i^2}$$

$$\psi_i(\hat{\delta}_{kk}) = \frac{\omega_i^2\hat{\zeta}_{ik}\ddot{X}_{ik}}{\sum_i \omega_i^2\ddot{X}_{ik}^2},$$

where $\hat{\zeta}_{ik}$ is the WLS residual based on (4), and $\ddot{X}_{ik}$ is the sample analog of $\tilde{\tilde{X}}_{ik}$, the residual from regressing X_{ik} onto $X_{i,-k}$ and Z_i. It then follows from (6) and the influence function formulas above, that

$$\begin{aligned}
\psi_i(\hat{\beta}_k^{\text{Own}}) &= \hat{\delta}_{kk}' \psi_i(\hat{\gamma}_k) + \hat{\gamma}_k' \psi_i(\hat{\delta}_{kk}) \\
\psi_i(\hat{\beta}) &= \left(\sum_i \omega_i^2 \dot{X}_i \dot{X}_i' \right)^{-1} \omega_i^2 \dot{X}_i \hat{U}_i \\
\psi_i(\hat{\beta}_k^{\text{EW}}) &= \frac{\mathbb{1}\{D_i \in \{0,k\}\} \omega_i^2 \hat{X}_{ik} \hat{U}_{ik}}{\sum_i \mathbb{1}\{D_i \in \{0,k\}\} \omega_i^2 \hat{X}_{ik}^2} \\
\psi_i(\hat{\beta}_k^{\text{ATE}}) &= \bar{Z}' \psi_i(\hat{\alpha}_k - \hat{\alpha}_0) + \psi_i(\bar{Z})(\hat{\alpha}_k - \hat{\alpha}_0).
\end{aligned}$$

where $\hat{X}_{ik}$ is the residual from regressing X_{ik} onto Z in the subset with $D_i \in \{0,k\}$, and $\hat{U}_{ik}$ the residual from regressing Y_i onto X_{ik} and Z_i in this subsample.

When the treatment is binary and overlap holds, the formula for $\psi_i(\hat{\beta}_k^{\text{ATE}})$ is similar to that discussed on page 29 in Imbens and Wooldridge [2009], except we don't assume that the regression error V_i in (1) is conditionally mean zero, so that the standard error is robust to misspecification.

Derivations in the last section show that the influence function for the common weights estimator is given by $\hat{\psi}_i(\hat{\beta}_k^{\text{CW}}) = \hat{\psi}_i(\hat{\alpha}_k^{\text{CW}} - \hat{\alpha}_k^{\text{CW}})$, where

$$\hat{\psi}_i(\hat{\alpha}_k^{\text{CW}}) = \frac{1}{\sum_i \lambda^{\text{CW}}(W_i)} \left(\frac{\lambda^{\text{CW}}(W_i) X_{ik}}{\pi_k(W_i;\hat{\theta})} (Y_i - \hat{\alpha}_k^{\text{CW}}) + a_i(\hat{\theta}) \right). \tag{7}$$

with the formula for a_i given in (12) below, θ corresponds to the parameters in the multinomial logit model, $\pi_k(W_i;\hat{\theta})$ to the fitted probabilities in this model, and $\hat{\alpha}_k^{\text{CW}}$ is the estimate based on (11) below.

Oracle standard errors

The package also reports "oracle" standard errors, which interprets the alternative estimators as estimates of the contrasts

$$\beta_{\lambda,k} = \frac{\sum_{i=1}^N \lambda(W_i)(\mu_k(W_i) - \mu_0(W_i))}{\sum_{i=1}^N \lambda(W_i)},$$

with $\lambda(W_i) = 1$ for the unweighted ATE, $\lambda(W_i) = \lambda^{\text{CW}}(W_i)$ for the common weights estimator and $\lambda(W_i) = \frac{p_k(W_i)p_0(W_i)}{p_k(W_i)+p_0(W_i)}$ for the efficiently weighted ATE estimator. In contrast, the standard errors in the previous subsection set the estimands to be the population counterparts to these weighted averages, replacing the sums in the above display with population expectations. In addition, the oracle standard errors don't account for estimation error in the propensity score $p(W_i)$.

For the unweighted ATE, the oracle standard error is based on the influence function $\tilde{\psi}_i(\hat{\beta}_k^{ATE}) = \bar{Z}' \psi_i(\hat{\alpha}_k - \hat{\alpha}_0)$. From the derivation in the last section, it follows that the oracle standard error for $\hat{\beta}_k^{\text{EW}}$ is given by

$$\psi_i(\hat{\beta}_k^{\text{EW}}) = \frac{\mathbb{1}\{D_i \in \{0,k\}\} \omega_i^2 \hat{X}_{ik} \hat{\hat{U}}_i}{\sum_i \mathbb{1}\{D_i \in \{0,k\}\} \omega_i^2 \hat{X}_{ik}^2}, \tag{8}$$

where $\hat{\hat{U}}_i$ is the interacted regression residual based on (1).

Finally, the oracle standard errors for $\hat{\beta}_k^{\text{CW}}$ are based on the influence function $\tilde{\psi}_i(\hat{\beta}_k^{\text{CW}}) = \tilde{\psi}_i(\hat{\alpha}_k^{\text{CW}}) - \tilde{\psi}_i(\hat{\alpha}_0^{\text{CW}})$, where

$$\tilde{\psi}_i(\hat{\alpha}_k^{\text{CW}}) = \frac{\omega_i^2 \lambda^{\text{CW}}(W_i;\hat{\theta})}{\sum_i \omega_i^2 \lambda^{\text{CW}}(W_i;\hat{\theta})} \frac{X_{ik}}{\pi_k(W_i;\hat{\theta})} \hat{U}_i. \tag{9}$$

Overlap sample

The package applies the above formulas to the full sample. In cases with poor overlap, this may not yield well-defined estimates or bias decomposition for all treatments. For components of the decomposition and alternative estimators that are not identified, the package returns `NA`. In such cases, the package also returns results for a trimmed "overlap sample", where the decomposition and alternative estimators are all identified. The overlap sample is constructed as follows:

1. Find a factor variable among the controls with the greatest number of levels. If there are no factor variables, skip this step. If for some levels of this variable, we don't see observations that take on one or more of the $K+1$ possible treatments, drop observations with these levels.
2. For the remaining controls, if a control doesn't display any variation in the subset of the data with treatment $k = 0, \ldots, K$, drop the control.

Wald and LM tests

We now give the form of the Wald and LM tests for variation in the propensity score. First, we give a general derivation of these tests in a likelihood context when the Hessian may be reduced rank. We then specialize the formulas to the case where the likelihood corresponds to the that for the multinomial logit model.

Consider a log-likelihood $\ell_n(\theta)$ for a p-dimensional parameter θ, with score function S that's approximately normal with covariance matrix Ω, and Hessian H. We're interested in testing the hypothesis that last r elements of θ are zero, $H_0 \colon \theta_2 = 0$. We assume that the submatrix H_{11} of the Hessian corresponding to the restricted model is full rank, but the full matrices Ω or H may not be invertible.

The score evaluated at the unrestricted estimator $\hat{\theta}$ satisfies

$$0 = \begin{pmatrix} S_1(\hat{\theta}_1, \hat{\theta}_2) \\ S_2(\hat{\theta}_1, \hat{\theta}_2) \end{pmatrix} = \begin{pmatrix} S_1(\theta_1, 0) \\ S_2(\theta_1, 0) \end{pmatrix} + \begin{pmatrix} H_{11}(\hat{\theta}_1 - \theta_1) + H_{12}\hat{\theta}_2 \\ H_{21}(\hat{\theta}_1 - \theta_1) + H_{22}\hat{\theta}_2, \end{pmatrix}$$

ignoring in the notation that the Hessian evaluated needs to be evaluated at intermediate values. Rearranging,

$$\begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ (H_{22} - H_{21}H_{11}^{-1}H_{12})\hat{\theta}_2 \end{pmatrix} = \begin{pmatrix} -H_{11}^{-1}S_1(\theta_1, 0) - H_{11}^{-1}H_{12}\hat{\theta}_2 \\ H_{21}H_{11}^{-1}S_1(\theta) - S_2(\theta) \end{pmatrix}$$

This yields the Wald statistic

$$W = \hat{\theta}_2'(H_{22} - H_{21}H_{11}^{-1}H_{12})' \operatorname{var}(S_2(\theta_1, 0) - H_{21}H_{11}^{-1}S_1(\theta_1, 0))^+ (H_{22} - H_{21}H_{11}^{-1}H_{12})\hat{\theta}_2,$$

where A^+ denotes a generalized inverse. By Lemma 9.7 in Newey and McFadden [1994], the statistic has an asymptotic χ^2 distribution with degrees of freedom equal to the rank of the variance.

The score evaluated at the restricted estimator $\bar{\theta}_1$ satisfies

$$\begin{pmatrix} 0 \\ S_2(\bar{\theta}_1, 0) \end{pmatrix} = \begin{pmatrix} S_1(\theta_1, 0) \\ S_2(\theta_1, 0) \end{pmatrix} + \begin{pmatrix} H_{11}(\bar{\theta}_1 - \theta_1) \\ H_{21}(\bar{\theta}_1 - \theta_1) \end{pmatrix},$$

which implies $\bar{\theta}_1 - \theta_1 = -H_{11}^{-1} S_1(\theta_1, 0)$, and hence

$$S_2(\bar{\theta}_1, 0) = S_2(\theta_1, 0) - H_{21} H_{11}^{-1} S_1(\theta_1, 0).$$

Thus the statistic

$$LM = S_2(\bar{\theta}_1, 0)' \operatorname{var}(S_2(\theta_1, 0) - H_{21} H_{11}^{-1} S_1(\theta_1, 0))^+ S_2(\bar{\theta}_1, 0)$$

will again have a χ^2 distribution.

To apply these formulas in the context of a multinomial logit model, we use the score and the Hessian

$$S(\theta) = \sum_i \omega_i^2 (X_i - \pi(Z_i; \theta)) \otimes Z_i, \quad H(\theta) = -\sum_i \omega_i^2 (\operatorname{diag}(\pi(Z_i; \theta)) - \pi(Z_i;)\pi(Z_i)') \otimes Z_i Z_i'$$

Derivations

We first derive (7). Observe first that the common weights estimator is identical to the two-step GMM estimator that in the first step, fits a multinomial logit model

$$P(D_i = k \mid W_i) = \frac{e^{Z_i' \theta_k}}{\sum_{k'=0}^K e^{Z_i' \theta_{k'}}} =: \pi_k(W_i, \theta), \tag{10}$$

with the normalization $\theta_0 = 0$. In the second step, we use the moment condition

$$E\left[\frac{\lambda^{\mathrm{CW}}(W_i; \theta) X_{ik}}{\pi_k(W_i; \theta)} (Y_i - \alpha_k^{\mathrm{CW}}) \right] = 0. \tag{11}$$

to obtain estimates $\hat{\alpha}_k^{\mathrm{CW}}$, and set $\hat{\beta}_k^{\mathrm{CW}} = \hat{\alpha}_k^{\mathrm{CW}} - \hat{\alpha}_0^{\mathrm{CW}}$.

Let

$$\zeta_k(W_i; \hat{\theta}) = \frac{\lambda^{\mathrm{CW}}(W_i; \hat{\theta})}{\pi_k(W_i, \hat{\theta})} = \frac{e^{-Z_i' \hat{\theta}_k}}{\sum_{j=0}^K \pi_j (1 - \pi_j) e^{-Z_i' \hat{\theta}_j}}$$

By equation (6.6) in Newey and McFadden [1994], the influence function of this two-step estimator is given by

$$\psi_i(\hat{\alpha}_k^{\mathrm{CW}}) = \frac{1}{E[\lambda^{\mathrm{CW}}(W_i)]} \left(\frac{\lambda^{\mathrm{CW}}(W_i) X_{ik}}{\pi_k(W_i)} (Y_i - \alpha_k^{\mathrm{CW}}) + M_k(\theta) \psi_i(\theta) \right),$$

where $\psi_i(\theta)$ is the influence function of the multinomial logit estimator $\hat{\theta}$, and $M_k(\theta)$ is the derivative of (11) wrt θ.

Since $\partial \zeta_k(W_i; \theta) / \partial \theta_j = Z_i \zeta_k(W_i; \theta)[\pi_j(1 - \pi_j)\zeta_k(W_i; \theta) - \mathbb{1}\{k = j > 0\}]$, it follows that

$$M_k(\theta) = E[(\eta - e_k) \otimes Z_i \cdot \zeta_k X_{ik} (Y_i - \alpha_{\lambda^C, k})], \qquad \eta = (\pi_j(1 - \pi_j)\zeta_j, \ldots, \pi_K(1 - \pi_K)\zeta_K)'.$$

Since the multinomial logit log-likelihood is given by $\ell_i = \sum_{k=0}^K X_k \log(\pi_k) = \sum_{k=0}^K X_k Z_i' \theta_k - \log(\sum_{k=0}^K e^{Z_i' \theta_k})$, the score and the Hessian are

$$S_i(\theta) = (X_i - \pi(W_i; \theta)) \otimes Z_i, \qquad H(\theta) = -E_n[(\operatorname{diag}(\pi(W_i)) - \pi(W_i)\pi(W_i)') \otimes Z_i Z_i'],$$

Since $\psi_i(\theta) = -H(\theta)^{-1} S_i(\theta)$, this yields

$$a_i(\theta) = \hat{M}_k(\hat{\theta})\hat{H}(\hat{\theta})^{-1} S_i(\theta), \qquad S_i(\theta) = (X_i - \pi(W_i;\theta)) \otimes Z_i, \tag{12}$$

with $\hat{M}_k(\theta) = \left(\sum_i (\zeta - e_k) \otimes Z_i \cdot \zeta_k X_{ik}(Y_i - \alpha_{\lambda^C,k})\right)$, and $\hat{H}(\theta) = \sum_i (\text{diag}(\pi(W_i)) - \pi(W_i)\pi(W_i)') \otimes Z_i Z_i'$. When $\pi_k = 1/(K+1)$ this formula reduces to that in Theorem 1 in Li and Li [2019].

Next, we show (8). Note that it follows from Proposition 2 in Goldsmith-Pinkham et al. [2022] that the efficient influence function is given by

$$\begin{aligned}\psi_i(\alpha_{\lambda_k,k0}) = \frac{\lambda(W_i)}{E[\lambda(W_i)]} \left(\frac{X_{ik}}{p_k(W_i)}(Y_i - \mu_k(W_i)) - \frac{X_{i0}}{p_0(W_i)}(Y_i - \mu_0(W_i)) \right) \\ = \frac{X_{ik} + X_{i0}}{E[\lambda(W_i)]} (X_{ik} - r_k) V_i = \frac{(X_{ik} + X_{i0})(X_{ik} - r_k)V_i}{E[(X_{ik} + X_{i0})(X_{ik} - r_k)^2]},\end{aligned}$$

where $r_k = r_k(W_i) = E[X_{ik} \mid W_i, X_{ik} + X_{i0} = 1]$. The result then follows since $\hat{X}_{ik}$ is an estimator of $X_{ik} - r_k(W_i)$.

Finally, we show (9). The derivative of the moment condition (11) with respect to $\pi_k = p_k$ (assuming correct specification of the propensity score) is given by

$$-E[\lambda \frac{X_{ik}}{p_k^2(W_i)}(\mu_k - \alpha_k^{\text{CW}})\dot{p}_k(W_i)],$$

where we write λ for $\lambda^{\text{CW}}(W_i)$. Since p_k is a projection, by Proposition 4 in Newey [1994], the influence function for $\hat{\alpha}_k^{\text{CW}}$ is given by

$$\begin{aligned}\frac{1}{E[\lambda]} \left(\lambda \frac{X_{ik}}{p_k(W_i)}(Y_i - \alpha_k^{\text{CW}}) - \frac{\lambda}{p_k(W_i)}(\mu_k(W_i) - \alpha_k^{\text{CW}})(X_{ik} - p_k(W_i)) \right) \\ = \frac{1}{E[\lambda]} \left(\lambda \frac{X_{ik}}{p_k(W_i)}(Y_i - \mu_k(W_i)) + \lambda(\mu_k(W_i) - \alpha_k^{\text{CW}}) \right).\end{aligned}$$

Next, as noted in Abadie et al. [2014], we can view $\tilde{\alpha}_k^{\text{CW}} = \sum_i \lambda \mu_k(W_i) / \sum_i \lambda$ as an estimator of α_k^{CW} based on the moment condition $E[\lambda(\mu_k(W_i) - \alpha_k^{\text{CW}})] = 0$, which by standard arguments has influence function given by $\frac{\lambda}{E[\lambda]}(\mu_k(W_i) - \alpha_k^{\text{CW}})$. Since $\hat{\alpha}_k^{\text{CW}} - \alpha_k^{\text{CW}} = (\hat{\alpha}_k^{\text{CW}} - \alpha_k^{\text{CW}}) - (\alpha_k^{\text{CW}} - \tilde{\alpha}_k^{\text{CW}})$, we subtract this influence function from the preceding display to obtain (9).

References

Alberto Abadie, Guido W. Imbens, and Fanyin Zheng. Inference for misspecified models with fixed regressors. *Journal of the American Statistical Association*, 109(508):1601–1614, December 2014. doi: 10.1080/01621459.2014.928218.

Roland G Fryer and Steven D Levitt. Testing for racial differences in the mental ability of young children. *American Economic Review*, 103(2):981–1005, April 2013. doi: 10.1257/aer.103.2.981.

Paul Goldsmith-Pinkham, Peter Hull, and Michal Kolesár. Contamination bias in linear regressions. ArXiv:2106.05024, August 2022. URL https://arxiv.org/abs/2106.05024.

Guido W. Imbens and Jeffrey Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, March 2009. doi: 10.1257/jel.47.1.5.

Fan Li and Fan Li. Propensity score weighting for causal inference with multiple treatments. 13(4): 2389–2415, December 2019. doi: 10.1214/19-AOAS1282.

Whitney K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6): 1349–1382, November 1994. doi: 10.2307/2951752.

Whitney K. Newey and Daniel L. McFadden. Large sample estimation and hypothesis testing. In Robert F. Engle and Daniel L. McFadden, editors, *Handbook of Econometrics*, volume 4, chapter 36, pages 2111–2245. Elsevier, New York, NY, 1994. doi: 10.1016/S1573-4412(05)80005-4.