RegCombin:
Partially Linear Regression under Data Combination

Xavier D'Haultfoeuille
CREST

Christophe Gaillac
University of Oxford

Arnaud Maurel
Duke University

Abstract

This vignette presents the package **RegCombin** associated to D'Haultfoeuille, Gaillac, and Maurel (2022) (DGM hereafter). **RegCombin** implements partially linear regression when the outcome of interest and some of the covariates are observed in two different datasets that cannot be linked. The package allows for common regressors observed in both datasets. The package also handles shape restrictions and lower bound on the R^2 of the long regression, as well as combinations of these. This paper illustrates the usage of **RegCombin** with several simulated and real examples. R and the package **RegCombin** are open-source software projects and can be freely downloaded from CRAN: //CRAN.R-project.org/package=RegCombin.

Keywords: Partially Linear Model; Data combination; Partial Identification; R.

1. How to get started

R is an open source software project and can be freely downloaded from the CRAN website. The R package **RegCombin** can also be downloaded from //CRAN.R-project.org/package=RegCombin, or directly installed using the command

```
install.packages("RegCombin")
```

Online help is available in two ways: either help(package="RegCombin") or ?regCombin. The first gives an overview over the available commands in the package. The second gives detailed information about a specific command. A valuable feature of R help files is that the examples used to illustrate commands are executable, so they can be pasted into an R session or run as a group with a command like example(regCombin). **Please do not hesitate to email christophe.gaillac@economics.ox.ac.uk if you have any remark, question, or suggestion to improve the package.**

2. Introduction

We consider the following linear model:

$$E(Y|X) = f(X_c) + X_{nc}'\beta_0, \quad X = (X_{nc}, X_c), \tag{1}$$

in a data combination environment where the distributions F_{Y,X_c} and F_{X_{nc},X_c} are supposed to be identified, but the joint distribution $F_{Y,X}$ is not. The variables X_c are thus common to both datasets, whereas the variables X_{nc} are only observed in one of the two datasets. In this setup, $\beta_0 = (\beta_{01}, ..., \beta_{0p})'$ is generally not point-identified, and as a result we focus on the identified set of either β_0 or β_{0k} for some $k \in \{1, ..., p\}$.

In the next section, we recall the identification results and inference method of DGM. We present the **regCombin** package in Section 4. Finally, several examples are considered in Section 5.

3. Theory

3.1. Identified set without common regressors

We first suppose that there is no X_c, so that $X = X_{nc}$ and $f(X_c) = 0$ in (1). Assuming that $E(Y^2) < \infty$, $E(\|X\|^2) < \infty$, $V(Y) > 0$ and $V(X)$ is non-singular, and $E(Y|X) = \alpha_0 + X'\beta_0$ for some $(\alpha_0, \beta_0) \in \mathbb{R} \times \mathbb{R}^p$, Theorem 1 in DGM states that the identified set $\mathcal{B}$ of β_0 can be described as

$$\mathcal{B} = \left\{\lambda q : \; q \in \mathcal{S}, \; 0 \leq \lambda \leq S(F_{Y_0}, F_{X_0'q})\right\},$$

where $\mathcal{S}$ is the unit sphere in $\mathbb{R}^p$, we let $A_0 = A - E(A)$ for any random variable A with $E[|A|] < \infty$ and

$$S(F, G) = \inf_{\alpha \in (0,1)} R(\alpha, F, G),$$
$$R(\alpha, F, G) = \frac{\int_\alpha^1 F^{-1}(t)dt}{\int_\alpha^1 G^{-1}(t)dt}.$$

Theorem 1 in DGM also shows that the identified set is a subset of the following one, based on the variances of X and Y only:

$$\begin{aligned}\mathcal{B} \subseteq \mathcal{B}^V &= \{\beta \in \mathbb{R}^p : \; \beta' V(X)\beta \leq V(Y)\} \\ &= \left\{\lambda q : q \in \mathcal{S}, \; 0 \leq \lambda \leq S^V\left(F_Y, F_{X'q}\right)\right\},\end{aligned}$$

where $S^V\left(F_Y, F_{X'q}\right) = \left(V(Y)/(q'V(X)q)\right)^{1/2}$. Hereafter, we refer to bounds based on $\mathcal{B}^V$ as the "Variance bounds".

3.2. Identified set with common regressors

With common regressors, the key is to note that we can get back to the previous setup without common regressors once we compute the following residuals, for all x in the support of X_c:

$$X^x = X_{nc} - E(X_{nc}|X_c = x),$$
$$Y^x = Y - E(Y|X_c = x).$$

Specifically, Theorem 2 in DGM shows that the identified sets $\mathcal{B}^c$ and $\mathcal{F}$ of β_0 and the function f satisfy, if we assume that $E(Y^2) < \infty$ and $E(X^x X^{x\prime}|X_c = x)$ is nonsingular for all $x \in \text{Supp}(X_c)$,

$$\mathcal{B}^c = \left\{\lambda q : q \in \mathcal{S}, \; 0 \leq \lambda \leq \overline{S}(F_{Y,X_c}, F_{X_{nc}'q,X_c})\right\},$$
$$\mathcal{F} = \left\{x \mapsto E(Y|X_c = x) - E(X_{nc}|X_c = x)'\beta : \; \beta \in \mathcal{B}^c\right\},$$

where $\overline{S}(F_{Y,X_c}, F_{X_{nc}'q,X_c}) = \inf_{x \in \text{Supp}(X_c)} S(F_{Y^x|X_c=x}, F_{X^{x\prime}q|X_c=x})$.

3.3. Identified set with common regressors and a lower bound constraint on the R^2

If one is ready to impose a lower bound $\underline{R}^2$ on $R^2_\ell := V(E(Y|X))/V(Y)$ (with $\underline{R}^2$ satisfying $\underline{R}^2 \geq R^2_s := V(E(Y|X_c))/V(Y)$), the identified set on β_0 becomes

$$\left\{\lambda q : q \in \mathcal{S},\ \left(\frac{(\underline{R}^2 - R^2_s)V(Y)}{q'E(V(X_{nc}|X_c))q}\right)^{1/2} \leq \lambda \leq \overline{S}(F_{Y,X_c}, F_{X'_{nc}q,X_c})\right\}.$$

3.4. Identified set with common regressors and shape constraints

We consider shape restrictions of the form $[Rf](r) \geq \underline{c}(r)$ for all $r \in \mathcal{R}$, with R a known linear operator, $\underline{c}$ a known function and $\mathcal{R}$ the domain of $[Rf]$ and $\underline{c}$. For instance, assuming that $\text{Supp}(X_c) = \{x_{c,1}, ..., x_{c,K}\} \subset \mathbb{R}$ with $K > 1$ and $x_{c,1} < ... < x_{c,K}$, if we want to impose that f is

1. non-decreasing, then $[Rf](r) = f(x_{c,r+1}) - f(x_{c,r})$ and $\underline{c}(r) = 0$ for $r \in \mathcal{R} = \{1, ..., K-1\}$;
2. convex, then $[Rf](r) = (f(x_{c,r+2}) - f(x_{c,r+1}))/(x_{c,r+2} - x_{c,r+1}) - (f(x_{c,r+1}) - f(x_{c,r}))/(x_{c,r+1} - x_{c,r})$ and $\underline{c}(r) = 0$ for $r \in \mathcal{R} = \{1, ..., K-2\}$ with $K > 2$.

Hereafter, we denote by $m_Y(\cdot) = E[Y|X_c = \cdot]$, $m_{X_{nc}}(\cdot) = E[X_{nc}|X_c = \cdot]$ and

$$\begin{aligned}
\underline{S}^c(m_Y, m_{X_{nc}}, q) &= \sup_{\substack{r \in \mathcal{R}: \\ [Rm'_{X_{nc}}q](r) \leq 0}} \lim_{u \downarrow 0} \frac{[Rm_Y - \underline{c}](r) + u}{[Rm'_{X_{nc}}q](r) - u^2}, \\
\overline{S}^c(m_Y, m_{X_{nc}}, q) &= \inf_{\substack{r \in \mathcal{R}: \\ [Rm'_{X_{nc}}q](r) \geq 0}} \lim_{u \downarrow 0} \frac{[Rm_Y - \underline{c}](r) + u}{[Rm'_{X_{nc}}q](r) + u^2},
\end{aligned}$$

where we let $\sup \emptyset = -\inf \emptyset = -\infty$. The two functions above may be infinite, and we introduce limits to deal with the cases where $[Rm'_{X_{nc}}q](r) = 0$.

Then Proposition 3 in DGM shows that the identified sets $\mathcal{B}^{\text{con}}$ and $\mathcal{F}^{\text{con}}$ of β_0 and f satisfy

$$\begin{aligned}
\mathcal{B}^{\text{con}} &= \left\{\lambda q : q \in \mathcal{S}^+,\ \underline{S}^{\text{con}}(q, F_{Y,X_c}, F_{X_{nc},X_c}) \leq \lambda \leq \overline{S}^{\text{con}}(q, F_{Y,X_c}, F_{X_{nc},X_c})\right\}, \\
\mathcal{F}^{\text{con}} &= \left\{x \mapsto E(Y|X_c = x) - E(X_{nc}|X_c = x)'\beta :\ \beta \in \mathcal{B}^{\text{con}}\right\},
\end{aligned}$$

where $\mathcal{S}^+$ denotes the upper hemisphere and

$$\begin{aligned}
\underline{S}^{\text{con}}(q, F_{Y,X_c}, F_{X_{nc},X_c}) &= \max\Big(- \overline{S}(F_{Y,X_c}, F_{-X'_{nc}q,X_c}), \underline{S}^c(m_Y, m_{X_nc}, q)\Big), \\
\overline{S}^{\text{con}}(q, F_{Y,X_c}, F_{X_{nc},X_c}) &= \min\Big(\overline{S}(F_{Y,X_c}, F_{X'_{nc}q,X_c}), \overline{S}^c(m_Y, m_{X_nc}, q)\Big).
\end{aligned}$$

3.5. Inference

Estimator and confidence regions

Without common regressors X_c. Consider now that we observe $(Y_1, ..., Y_{n_Y})$ and $(X_1, ..., X_{n_X})$, two independent samples of i.i.d. variables with the same distribution as Y and X, respectively.

An issue for estimation and inference on $\mathcal{B}$ is that when $\alpha \to 0$ or $\alpha \to 1$, $R(\alpha, F, G)$ is a ratio of two terms tending to 0. Then, its plug-in estimator may become very unstable. To regularize the problem, we consider an outer set of $\mathcal{B}$ based on the removal of extreme values of α. Specifically, we define, for any $\varepsilon \in (0, 1/2)$,

$$\begin{aligned} S_\varepsilon(F, G) &= \min_{\alpha \in [\varepsilon, 1-\varepsilon]} R(\alpha, F, G), \\ \mathcal{B}_\varepsilon &= \left\{\lambda q : q \in \mathcal{S}, 0 \le \lambda \le S_\varepsilon(F_{Y_0}, F_{X_0' q})\right\}. \end{aligned}$$

We simply estimate $R(\alpha, F_{Y_0}, F_{X_0' q})$ and $S_\varepsilon(F_{Y_0}, F_{X_0' q})$ by their empirical counterpart $R(\alpha, \widehat{F}_{Y_0}, \widehat{F}_{X_0' q})$ and $S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0' q})$. We then estimate the identified set $\mathcal{B}_\varepsilon$ by plug-in:

$$\widehat{\mathcal{B}}_\varepsilon := \left\{\lambda q : q \in \mathcal{S},\ 0 \le \lambda \le S_\varepsilon\left(\widehat{F}_{Y_0}, \widehat{F}_{X_0' q}\right)\right\}.$$

The estimator $\widehat{B}^V$ of the set obtained using the variances B^V is obtained in a similar way

$$\widehat{\mathcal{B}}^V := \left\{\lambda q : q \in \mathcal{S},\ 0 \le \lambda \le S^V\left(\widehat{F}_{Y}, \widehat{F}_{X' q}\right)\right\}.$$

To build confidence regions on β_0, we rely on subsampling. Let $n = (n_X n_Y)/(n_X + n_Y)$ and let b_n denote the size of the subsample. For any estimator $\widehat{\theta}$, let $\widehat{\theta}^*$ denotes its subsampling counterpart. For a nominal coverage of $1 - \alpha$, the confidence region on β_0 we consider is given by

$$\text{CR}_{1-\alpha}(\beta_0) = \left\{\lambda q :\ q \in \mathcal{S},\ 0 \le \lambda \le S_\varepsilon\left(\widehat{F}_{Y_0}, \widehat{F}_{X_0' q}\right) - \widehat{c}_{\alpha,\varepsilon}(q) n^{-1/2}\right\},$$

where $\widehat{c}_{\alpha,\varepsilon}(q)$ is the quantile of order α of the distribution of $b_n^{1/2}[S_\varepsilon(\widehat{F}^*_{Y_0}, \widehat{F}^*_{X_0' q}) - S_\varepsilon(\widehat{F}_{Y_0},\, \widehat{F}_{X_0' q})]$, conditional on the data.

In practice, one is often interested in conducting inference on subcomponents of β_0. The identified (outer) set $\mathcal{B}_{k,\varepsilon}$ of $\beta_{0,k}$ corresponding to $\mathcal{B}_\varepsilon$ satisfies

$$\mathcal{B}_{k,\varepsilon} = [-\sigma_\varepsilon(-e_k, F_{Y_0}, F_{X_0}),\, \sigma_\varepsilon(e_k, F_{Y_0}, F_{X_0})],$$

where $\sigma_\varepsilon(\cdot, F_{Y_0}, F_{X_0})$ denotes the support function associated to $q \mapsto S_\varepsilon(F_{Y_0}, F_{X_0' q})$ and e_k the k-th element of the canonical basis of $\mathbb{R}^p$. To construct confidence intervals on β_{0k}, we first estimate $\sigma_\varepsilon(\cdot, F_{Y_0}, F_{X_0})$ by

$$\sigma_\varepsilon(e, \widehat{F}_{Y_0}, \widehat{F}_{X_0}) = \frac{1}{\inf_{q \in \mathbb{R}^p : q'e = 1} 1/S_\varepsilon\left(\widehat{F}_{Y_0}, \widehat{F}_{X_0' q}\right)}.$$

Then, denoting by $\widetilde{c}_{\beta,\varepsilon}(e)$ the quantile of order $\beta \in (0, 1)$ of the distribution of $b_n^{1/2}(\sigma_\varepsilon(e, \widehat{F}^*_{Y_0}, \widehat{F}^*_{X_0}) - \sigma_\varepsilon(e, \widehat{F}_{Y_0}, \widehat{F}_{X_0}))$, conditional on the data, the confidence interval we consider for $\beta_{0,k}$ is

$$\text{CI}_{1-\alpha}(\beta_{0,k}) = \left[\left(-\sigma_\varepsilon(-e_k, \widehat{F}_{Y_0}, \widehat{F}_{X_0}) + \frac{\widetilde{c}_{\alpha,\varepsilon}(-e_k)}{n^{1/2}}\right)^-, \left(\sigma_\varepsilon(e_k, \widehat{F}_{Y_0}, \widehat{F}_{X_0}) - \frac{\widetilde{c}_{\alpha,\varepsilon}(e_k)}{n^{1/2}}\right)^+\right],$$

where $x^- = \min(0, x)$ and $x^+ = \max(0, x)$.

The choice of ε is made in a data-driven way. When $p = 1$, let us define, for $q \in \mathcal{S} = \{-1, 1\}$,

$$\varepsilon(q) = \underset{\varepsilon \in \mathcal{E}}{\operatorname{argmin}}\ S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0' q}) - \widehat{c}_{\alpha,\varepsilon}(q) n^{-1/2}, \qquad (2)$$

where $\mathcal{E}$ is a finite grid in $(0, 1/2]$. Hence, $\varepsilon(q)$ simply minimizes the boundary value of the confidence region in the direction $q \in \mathcal{S}$. Now consider the case $p > 1$. If one focuses on confidence intervals on β_{0k}, we need to choose the parameter ε that appears in $\sigma_\varepsilon(\pm e_k, F_{Y_0}, F_{X_0})$. To this end, we simply use $\varepsilon(q)$ as given above, with $q = \pm e_k$. If, instead, we are interested in the set $\mathcal{B}$ itself, we recommend using $\underline{\varepsilon} = \min_{q \in \mathcal{Q}} \varepsilon(q)$, where $\mathcal{Q}$ is a finite subset of $\mathcal{S}$.

With common regressors X_c Inference with common regressors X_c follows the same logic as above. Consider X_c with finite support. Let $\widehat{F}_{Y^x|X_c=x}$ and $\widehat{F}_{X^{x\prime} q|X_c=x}$ denote the empirical estimators of $F_{Y^x|X_c=x}$ and $F_{X^{x\prime} q|X_c=x}$, respectively. In this case, we estimate $\overline{S}(F_{Y,X_c}, F_{X_{nc}' q, X_c})$ by

$$\widehat{\overline{S}}(q, F_{Y,X_c}, F_{X_{nc}' q, X_c}) = \min_{x \in \text{Supp}(X_c)} S_\varepsilon(\widehat{F}_{Y^x|X_c=x}, \widehat{F}_{X^{x\prime} q|X_c=x}).$$

Let $\widehat{\overline{c}}^c_{\alpha,\varepsilon}(q)$ be the quantile of order $\alpha \in (0, 1)$ of distribution of $b_n^{1/2}(\widehat{\overline{S}}^*(q, F_{Y,X_c}, F_{X_{nc},X_c})$ $-\widehat{\overline{S}}(q, F_{Y,X_c}, F_{X_{nc},X_c}))$, conditional on the data. For a nominal coverage of $1 - \alpha$, the confidence region on β_0 we consider is

$$\text{CR}^c_{1-\alpha}(\beta_0) = \left\{ \lambda q : q \in \mathcal{S},\ 0 \le \lambda \le \widehat{\overline{S}}(q, F_{Y,X_c}, F_{X_{nc},X_c}) - \widehat{\overline{c}}^c_{\alpha,\varepsilon}(q) n^{-1/2} \right\}.$$

With common regressors X_c and shape constraints. Finally, we describe how to conduct inference under constraints on the R^2 or shape restrictions. The main difference with above is that for a given direction $q \in \mathcal{S}$, both the lower and upper bounds on the identified set need to be estimated. As before, we estimate them with plug-in estimators. The only substantive difference is that in the confidence regions, we then account for the variability of both bounds. For instance, with shape restrictions, we consider the following confidence region:

$$\begin{aligned} \text{CR}^{con}_{1-\alpha}(\beta_0) = \Big\{ \lambda q : q \in \mathcal{S},\ & \widehat{\underline{S}}^{con}(q, F_{Y,X_c}, F_{X_{nc},X_c}) + \widehat{\underline{c}}^{con}_{1-\alpha/2,\varepsilon}(q) n^{-1/2} \le \lambda \\ & \le \widehat{\overline{S}}^{con}(q, F_{Y,X_c}, F_{X_{nc},X_c}) - \widehat{\overline{c}}^{con}_{\alpha/2,\varepsilon}(q) n^{-1/2} \Big\}, \end{aligned}$$

where $\widehat{\underline{c}}^{con}_{\delta,\varepsilon}(q)$ is the quantile of order δ of $b_n^{1/2}(\widehat{\underline{S}}^{con*}(q, F_{Y,X_c}, F_{X_{nc},X_c}) - \widehat{\underline{S}}^{con}(q, F_{Y,X_c}, F_{X_{nc},X_c}))$, conditional on the data and similarly for $\widehat{\overline{c}}^{con}_{\delta,\varepsilon}$.

Test of point identification

We can assess the plausibility of β_0 belonging to the boundary of the identified set $\mathcal{B}$ using a validation sample. Denoting by (Y_v, X_v) the variables corresponding to this validation sample, it becomes possible to test whether the corresponding parameter $\beta_v = V(X_v)^{-1} \text{cov}(X_v, Y_v)$ is at the boundary of the identified set one would get from the sole knowledge of F_{Y_v} and F_{X_v}. Provided that $\beta_v \neq 0$, this condition is indeed equivalent to $\|\beta_v\| = S(F_{Y_{v0}}, F_{X_{v0}' \beta_v / \|\beta_v\|})$ or, in simpler terms, $S(F_{Y_{v0}}, F_{X_{v0}' \beta_v}) = 1$. DGM develop a statistical test that can be used to check whether β_0 is at the boundary of the identified set. This boils down to testing for

$$H_0:\ S(F_{Y_{v0}}, F_{X_{v0}' \beta_v}) = 1 \quad \text{against} \quad H_1: S(F_{Y_{v0}}, F_{X_{v0}' \beta_v}) > 1.$$

The test statistic is

$$T = b_n^{1/2} \left(S_\varepsilon(\widehat{F}_{Y_{v0}}, \widehat{F}_{X_{v0}' \widehat{\beta}_v}) - 1 \right),$$

where $\widehat{\beta}_v$ is the OLS estimator of β_v. The critical value is then $q_{1-\alpha}(T^*)$, the quantile of order $1-\alpha$ (defined conditional on the data) of

$$T^* = n^{1/2} \left(S_\varepsilon(\widehat{F}^*_{Y_{v0}}, \widehat{F}^*_{X'_{v0}\widehat{\beta}^*_v}) - S_\varepsilon(\widehat{F}_{Y_{v0}}, \widehat{F}_{X'_{v0}\widehat{\beta}_v}) \right),$$

where $\widehat{F}^*_{Y_{v0}}$, $\widehat{F}^*_{X'_{v0}q}$ and $\widehat{\beta}^*_v$ are the subsampling counterpart of $\widehat{F}_{Y_{v0}}$, $\widehat{F}_{X'_{v0}q}$ and $\widehat{\beta}_v$, respectively.

4. The regCombin function in the RegCombin package

This function implements the estimators proposed in DGM. The syntax of the function `regCombin()` is as follows:

```
regCombin(Ldata,Rdata, out_var, nc_var, c_var = NULL,
constraint = NULL, nc_sign = NULL, c_sign = NULL,
weights_x = NULL, weights_y = NULL,
nbCores = 1, methods = c("DGM"), list_ex = c(),
grid = 10, alpha = 0.05, eps_default = 0.5,
R2bound = NULL, projections = FALSE,
unchanged = FALSE, ties = FALSE)
```

Ldata a dataset including Y and possibly $X_c = (X_{c,1}, ..., X_{c,q})$. X_c must be finitely supported.

Rdata a dataset including X_{nc} and the same variables X_c as in Ldata.

out_var the label of the outcome variable Y.

nc_var the labels of the regressors X_{nc}.

c_var the labels of the regressors X_c (if any).

constraint a vector of size q indicating the type of constraints (if any) on $x_{c,k} \mapsto f(x_{c,1}, ..., x_{c,q})$ for $k = 1, ..., q$: "convex", "concave", "nondecreasing", "nonincreasing", "nondecreasing_convex", "nondecreasing_concave", "nonincreasing_convex", "nonincreasing_concave", or NA for no constraint. Default is NULL, namely no constraints at all.

nc_sign a vector of size p indicating sign restrictions on each of the p coefficients of X_{nc}. For each component, -1 corresponds to a minus sign, 1 to a plus sign and 0 to no constraint. Default is NULL, namely no constraints at all.

c_sign same as nc_sign but for X_c (accordingly, it is a vector of size q).

weights_x the sampling weights for the dataset Rdata. Default is NULL.

weights_y the sampling weights for the dataset Ldata. Default is NULL.

nbCores number of cores for the parallel computation. Default is 1.

methods method(s) used for the bounds: "DGM" (Default) and/or "Variance".

grid the number of points in $\mathcal{E}$ for the grid search on ε (see Eq. (2)). If NULL, then grid search is not performed and ε is set equal to eps_default. Default is 10.

alpha one minus the nominal coverage of the confidence intervals. Default is 0.05.

eps_default a pre-specified value of ε used only if the grid search for selecting the value of ε is not performed, i.e, when grid is NULL. Default is 0.5.

R2bound the lower bound $\underline{R}^2$ on R^2_ℓ, if any. Default is NULL.

projections Boolean indicating if the identified set and confidence intervals on β_{0k} for $k = 1, ..., p$ are computed (TRUE), rather than the identified set and confidence region of β_0 (FALSE). Default is FALSE.

unchanged Boolean indicating if the categories based on X_c must be kept unchanged (TRUE). Otherwise (FALSE), a thresholding approach is taken imposing that each value appears more than 10 times in both datasets and represents more than 0.01% of the pooled dataset (of size $n_X + n_Y$). Default is FALSE.

ties Boolean indicating if there are ties in the dataset. If not (FALSE), computation is faster. Default is FALSE.

We also refer to the reference manuel or help file for additional details.

The minimial requirement for the function `regCombin()` are the two datasets Ldata and Rdata, the labels outvar and nc_var of Y and X_{nc} respectively, and the labels c_var if common regressors are observed.

If there are common regressors X_c, the `regCombin()` function starts by creating a single variable whose support is the different values taken by X_c. According to the argument unchanged, a thresholding approach is taken imposing that each value appears more than 10 times in both datasets and 0.01% is the pooled one. Then, the `regCombin()` function computes the DGM and/or variance bounds, as well as the confidence regions for the parameters associated with both X_{nc} and all dummies associated with X_c.

The grid $\mathcal{E}$ for the choice of $\varepsilon(q)$ in (2) is taken as 10 equally spaced points between $\varepsilon_{\min} = \max(9, C\ln(n))/n^{3/4}/p$ and 0.5, where $C = 5$ with common regressors X_c and $C = 1$ otherwise. Another tuning parameter we have to choose is b_n, which appears in the subsampling. Its choice seems to matter less in practice than that of ε. We fix $b_n = 0.75(0.5n - 0.3\max(n-5, 0) - 0.15\max(n - 1000, 0) - 0.5(1 - \log(3000)/\log(n))\max(n - 3000, 0))$.

The robustness of the DGM bounds to the choice of the parameter ε can be assessed using the command `regCombin_profile()`. This computes the profile of the bounds of the DGM set estimate as a function of a multiplier of the selected parameter ε.

The `regCombin()` function returns a vector containing all the different bounds, which can be represented as a knitr table using the `summary_regCombin()`. There, the table is represented in the viewer.

5. Examples

5.1. Simulations in a univariate case without common regressors

We consider the same DGP as in Section A.1 in the Online Appendix of DGM, namely:

$$Y = X_{nc}\beta_0 + U, \quad \beta_0 = 1,\ U|X_{nc} \sim \mathcal{N}(0, 1),\ X_{nc} \sim \mathcal{N}(0, 1.5).$$

The code below generates $1,000$ i.i.d. variables with the same distribution as Y and X_{nc} and stores them in the datasets Ldata and Rdata, respectively.

```
library(R.matlab)
library(MASS)
library(dplyr)
library(pracma)
library(Hmisc)
library(snowfall)
library(knitr)
library(kableExtra)
library(RegCombin)

### simulations
set.seed(3322)
n=1000
X2x =  rnorm(n,0,1.5)
X2y =  rnorm(n,0,1.5)
epsilon =  rnorm(n,0,1)
X2y=matrix(X2y,length(X2y),1)
Xnc=matrix(X2x,length(X2x),1)
Y =  X2y*1 + epsilon

## formatting
Y_all=matrix(Y,length(Y),1)
out_var = "Y"
nc_var = "X"
c_var = NULL

Ldata<- as.data.frame(Y_all)
colnames(  Ldata) <- c(out_var)
Rdata <- as.data.frame(Xnc)
colnames(Rdata) <- c(nc_var)
```

Estimation

Estimation is then performed using the main function regCombin.

```
out <- regCombin(Ldata,Rdata,out_var,nc_var)
```

The output results can be summarized and stored using summary_regCombin.

```
sum <- summary_regCombin(out)
```

```
Estimates of the DGM bounds in linear regression under data combination

===  ==============  ==============
Xnc  Set estimate     95% CI
===  ==============  ==============
X    [-1.212,1.237]  [-1.318,1.325]

===  ==============  ==============
Formula: Y ~ Xnc.
Outcome variable (Y): Y.
Non commun covariates (Xnc): X.
No commun covariates (Xc).
Number of observations (Y): 1000.
Number of observations (Xnc): 1000.
CI obtained using the subsampling method.
```

The robustness to the choice of ε

One can use the the function regCombin_profile to check the robustness of the results to the choice of ε, displaying the point estimates of the DGM bounds for different values of a multiplier of our data-driven selected value of ε, which is the default one.

```
profile = regCombin_profile(Ldata,Rdata,out_var,nc_var,
multipliers=seq(0.1,3,length.out=20))

### to plot the profile of the upper bounds according to the multiplier.
x11()
plot(profile$Profile_point[,1],profile$Profile_point[,3], type="l", lwd=2,col=1,
xlab="Value of the multiplier of the data-driven choice of epsilon",
ylab="Point Estimate of the upper bound")
```

Testing for point identification

In the situation where there exists a subsample where the joint distribution is observed, we can test for point identification using the function point_ident_test. In the example below, we reject the test of point identification.

```
validation <- as.data.frame(cbind(Y,X2y))
colnames(validation) <- c(out_var,nc_var)
test = point_ident_test (validation, Ldata=NULL,Rdata=NULL,out_var,nc_var)
## the p-value of the test
test$p_value
[1] 0
```

5.2. Simulations in a multivariate case without common regressor

We consider the same DGP as in Section A.2 in the Online Appendix of DGM, namely:

$$Y = \gamma_0 + X_{nc}'\beta_0 + U, \ U|X \sim \mathcal{N}(0,4),$$

where $p = 2$, the coefficients satisfy $\gamma_0 = -0.1$, $\beta_{0,1} = 1$, $\beta_{0,2} = 1$ and the variables X_{nc} follow a multivariate normal distribution with mean 0 and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & -0.2 \\ -0.2 & 1 \end{pmatrix}.$$

We first simulate the data with the following code:

```
library(MASS)
### DGM
n=800
Sigma=matrix(c(1,0.8,-0.1,0.8,1,-0.2,-0.1,-0.2,1),3,3)
norm = mvrnorm(n,mu=c(0,0,0),Sigma=Sigma)
norm_y = mvrnorm(n,mu=c(0,0,0),Sigma=Sigma)
X2=norm[,2];
X3=norm[,3];
X2_y=norm_y[,2];
X3_y=norm_y[,3];

## true value
beta0 = as.matrix(c(1,1));

X=cbind( X2, X3)
Xnc=cbind(X2_y, X3_y)
Y =  Xnc%*%beta0 + 2*rnorm(n)


##### formatting the inputs
```

```
c_var=NULL
out_var = "Y"
nc_var = c("X2","X3")
dimXnc = 2

Ldata <- as.data.frame(Y)
colnames(Ldata) <- c(out_var)
Rdata <- as.data.frame(X)
colnames(Rdata) <- c(nc_var)
```

Then estimation can be performed using regCombin.

```
out <- regCombin(Ldata,Rdata,out_var,nc_var, grid=10, projections= TRUE)
```

To save time, we can leverage parallel computing by using the snowfall library jointly with the argument nbCores=4. This indicates that 4 CPU cores are used in the computation.

```
library(snowfall)
out <- regCombin(Ldata,Rdata,out_var,nc_var,nbCores=4,projections= TRUE)
```

Results are displayed with summary_regCombin:

```
sum <- summary_regCombin(out)
Estimates of the DGM bounds in linear regression under data combination

===  ==============  ==============
Xnc  Set estimate     95% CI
===  ==============  ==============
X2   [-2.233,2.255]  [-2.471,2.449]
X3   [-2.268,2.205]  [-2.453,2.385]

===  ==============  ==============
Formula: Y ~ Xnc.
Outcome variable (Y): Y.
Non commun covariates (Xnc): X2, X3.
No commun covariates (Xc).
Number of observations (Y): 800.
Number of observations (Xnc): 800.
CI obtained using the subsampling method.
```

5.3. Simulations in a case with a common regressor

We consider the same DGP as in Section A.3 in the Online Appendix of DGM, namely:

$$Y = X_c\gamma_0 + X_{nc}\beta_0 + U,\ U|X \sim \mathcal{N}(0,4),$$

and where the coefficients satisfy $\gamma_0 = 0.3$ and $\beta_0 = 1$. The covariates satisfy $X_c = 1\{N_1 \leq 0.3\}$ and $X_{nc} = N_2$, where (N_1, N_2) follows a multivariate normal distribution with mean 0 and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1.5 \end{pmatrix}.$$

The following code simulates the data, compute the bounds and confidence intervals and prints the results:

```
library(RegCombin)
library(MASS)
library(dplyr)
library(pracma)
library(Hmisc)
library(snowfall)
library(knitr)
library(kableExtra)


### simulation and formatting
set.seed(3322)
n=2000
norm = mvrnorm(n,mu=c(0,0),Sigma= matrix(c(1,0.8,0.8,1.5),2,2))
normy = mvrnorm(n,mu=c(0,0),Sigma= matrix(c(1,0.8,0.8,1.5),2,2))
Xc_x= matrix(1*(norm[,1]<= 0.3), n,1)
Xc_y= matrix(1*(normy[,1]<= 0.3),n,1)
Xnc=norm[,2];
Xncy=normy[,2];
epsilon = 2*rnorm(n)
beta0 = 1
X=cbind(Xc_x, Xnc)
Xnc=matrix(Xnc,n,1)
Xncy=matrix(Xncy,n,1)
Y = 0.3*Xc_y +  Xncy*beta0 + epsilon

out_var = "Y"
nc_var = "X"
c_var =  "Xc"
Ldata<- as.data.frame(cbind(Xc_y,Y))
colnames(  Ldata) <- c(c_var,out_var)
Rdata <- as.data.frame(cbind(Xc_x,Xnc))
colnames(Rdata) <- c(c_var,nc_var)

#### estimation and printing results
out <- regCombin(Ldata,Rdata,out_var,nc_var,c_var)
sum <- summary_regCombin(out)

Estimates of the DGM bounds in linear regression under data combination

===  ===  ==============  ==============  ============================  =========================
Xnc  Xc   Set estimate     95% CI         Set estimate, with constraint  95% CI,  with  constraint
===  ===  ==============  ==============  ============================  =========================
X         [-2.075,2.132]  [-2.238,2.309]  [-2.075,2.132]                [-2.263,2.33]
     Xc
     1    [-3.741,1.79]   [-4.146,2.213]  [-3.741,1.79]                 [-4.224,2.272]
===  ===  ==============  ==============  ============================  =========================
Formula: Y ~ Xnc.
Outcome variable (Y): Y.
Non commun covariates (Xnc): X.
Commun covariates (Xc): Xc.
Number of observations (Y): 2000.
Number of observations (Xnc): 2000.
CI obtained using the subsampling method.
```

We can also add constraints on γ_0 or on the R^2 of the long regression to reduce the confidence intervals on β_0. Let us start by imposing $\gamma_0 \geq 0$:

```
out_sign <- regCombin(Ldata,Rdata,out_var,nc_var,c_var,constraint = "nondecreasing")
sum_sign <- summary_regCombin(out_sign)
Estimates of the DGM bounds in linear regression under data combination

===  ===  ==============  ==============  ============================  =========================
Xnc  Xc   Set estimate     95% CI         Set estimate, with constraint  95% CI,  with  constraint
===  ===  ==============  ==============  ============================  =========================
X         [-2.075,2.132]  [-2.238,2.309]  [0.77,2.132]                  [0.505,2.33]
     Xc
     1    [-3.741,1.79]   [-4.146,2.213]  [0,1.79]                      [0,2.272]
===  ===  ==============  ==============  ============================  =========================
Formula: Y ~ Xnc.
Outcome variable (Y): Y.
Non commun covariates (Xnc): X.
Commun covariates (Xc): Xc.
Number of observations (Y): 2000.
Number of observations (Xnc): 2000.
CI obtained using the subsampling method.
```

Again, the function regCombin_profile can be used to check the robustness of the results to the choice of ε, displaying the point estimates of the DGM bounds for different values of a multiplier of our data-driven selected value of ε, which is the default one.

```
profile = regCombin_profile(Ldata,Rdata,out_var,nc_var,constraint = "nondecreasing",
 multipliers = seq(0.1,3,length.out=20))
x11()
plot(profile$Profile_point[,1],profile$Profile_point[,3], type="l", lwd=2,col=1,
xlab="value of the multiplier of the data-driven choice of epsilon",
ylab="Point Estimate of the upper bound")
```

Note that for multidimensional common regressors X_c, the contraints are specified using a vector. For instance, in a two-dimensional case, constraint =c("nondecreasing", NA) would impose that the effect of the first component of X_c is nondecreasing and no constraint on the second component.

5.4. Application to a real dataset

For illustration purposes, we consider the real dataset Schooling from the package Ecdat, used in Card (1995). This dataset is a subset of the National Longitudinal Survey of Young Men (NLSYM), in which we observe the wages of these young men in 1976 as well as their IQ score and other characteristics at the age of 14. Note that this dataset is different from that used in DGM, but has the advantage of being freely accessible. We want here to replicate on this dataset our analysis of the black-white wage gap, revisiting the seminal work of Neal and Johnson (1996) on this question. We restrict the sample to individuals between 26 and 29 in 1976, which gives 980 observations. Contrary to the dataset used in DGM, information about Hispanics is not available in this data. We then consider the following model:

$$Y = \gamma_{c,0} + X_c\gamma_c + X_{nc}\beta_{nc} + \epsilon, \quad E\left[\epsilon|X_c, X_{nc}\right] = 0,$$

where X_{nc} denotes a IQ score and X_c is a dummy variable for being black. While (Y, X_c, X_{nc}) is jointly observed in this dataset, we proceed in the following as if the IQ score were not observed jointly with wages.

We first format the data and create the two datasets:

```
### load and format the data ###########
library(Ecdat)
data(Schooling)
util <- Schooling
util$log_wage <- util$lwage76
util <- util[util$log_wage>0,]
util <- util[util$age76<=29 & util$age76>=26,]
util <- util[!is.na(util$log_wage),]
util <- util[!is.na(util$iqscore),]
table(util$age76)

dim(util)
```

```
util$X1 = 1*(util$black=="yes")
names = c("Black")

### set sign constraints Xc
c_sign =c(0)
### set sign constraints Xnc
nc_sign = c(1)

out_var= "log_wage"
nc_var = c("iqscore")
c_var = c("X1")

# artificially splitting the data in two
set.seed(02121)
lsample = sample(1:dim(util)[1],floor(dim(util)[1]/2), replace= FALSE)
lutil = util[lsample,]
rutil = util[-c(lsample),]
```

Let us start by imposing a negative sign constraint on the coefficient γ_c associated with the black dummy as well as a positive sign constraint β_0 associated with the IQ score. These two constraints can be imposed byincluding constraint= c("nonincreasing") and nc_sign =c(1) in `regCombin()`.

```
########
ln= dim(lutil)[1]
rn= dim(rutil)[1]

dimXnc = length(nc_var)
dimXc = length(c_var)
####### prepare the data
if(dimXc==1){
        X1_x = as.matrix(rutil[,c_var],rn,dimXc)
        X1_y = as.matrix(lutil[,c_var],ln,dimXc)
}else{
        X1_x= rutil[,c_var]
        X1_y= lutil[,c_var]
}
# X1=  as.matrix(util[,c_var])
Y_all =  lutil$log_wage
X_all=cbind(X1_x, rutil[,nc_var])

Y_all = cbind(X1_y,Y_all)
X0 <- X_all
Y_all0 <-  Y_all
Y_all <- as.data.frame(Y_all)
colnames(Y_all) <- c(c_var,out_var)
X_all <- as.data.frame(X_all)
colnames(X_all) <- c(c_var,nc_var)

weights_x = NULL
weights_y = NULL
Ldata = Y_all
Rdata= X_all

###### Bounds using sign constraints
output1 <- regCombin(Ldata,Rdata,out_var,nc_var,c_var,
constraint= c("nonincreasing"),
nc_sign =c(1), c_sign = NULL,
weights_x = NULL,weights_y = NULL,
nbCores=1,
methods=c("DGM"),
list_ex=c(),
grid = 10,
alpha=0.05,
eps_default = 0.5,
R2bound=NULL ,
projections= FALSE,
unchanged=FALSE,
ties = TRUE)

mat1 <- summary_regCombin(output1)
```

Next, we seek to impose the constraint that $\underline{R}^2 \geq 1.3R_s^2$. Note that in the full dataset, the R^2 of the long regression is 0.0552 and the one of the short regression is 0.0328, so our constraint is satisfied. Imposing the constraint can be done specifying R2bound=1.3 in the arguments of `regCombin()`.

```
output2 <- regCombin(Ldata,Rdata,out_var,nc_var,c_var,
constraint= c("nonincreasing"),
nc_sign =c(1), c_sign = NULL,
weights_x = NULL,weights_y = NULL,
nbCores=1,
methods=c("DGM"),
list_ex=c(),
grid = 10,
alpha=0.05,
eps_default = 0.5,
R2bound=1.3,
projections= FALSE,
unchanged=FALSE,
ties = TRUE)

mat2 <- summary_regCombin(output2)
```

Gathering these bounds with the OLS estimates on the joint dataset, we obtain Table 1 below. If we focus on the main coefficient of interest γ_c, these results indicate that imposing the lower bound on the R^2 results in an identified set and confidence interval that are quite informative. Notably, the lower bound of the confidence interval is equal to -0.219 in the last case, as for the OLS estimator. Taken together, these results show that our method is able to deliver confidence intervals that are quite informative in realistic data environments.

References

- D'Haultfoeuille, X., C. Gaillac, and A. Maurel (2023). Partially Linear Models under Data Combination. arXiv prepint arXiv:2204.05175.
- Card, D. (1995) Using geographical variation in college proximity to estimate the return to schooling in Christofides, L.N., E.K. Grant and R. Swidinsky (1995) Aspects of labour market behaviour : essays in honour of John Vanderkamp, University of Toronto Press, Toronto.
- Neal, D. A. and W. R. Johnson (1996). The role of premarket factors in black-white wage differences. Journal of Political Economy 104 (5), 869–895.

	OLS	DGM		
Constraints		Without	With signs constraints	
			Only	And $\underline{R}^2 \geq 1.3R_s^2$
	(1)	(2)	(3)	(4)
Omitted variable X_{nc}				
IQ score, pt.	0.004	[-0.021,0.021]	[0,0.009]	[0.002,0.009]
CI	[0.002,0.006]	[-0.026,0.026]	[0,0.012]	[0.001,0.012]
Common variables X_c				
Black, pt.	-0.115	[-0.557,0.238]	[-0.16,0]	[-0.117,0]
CI	[-0.219,-0.012]	[-0.708,0.392]	[-0.271,0]	[-0.219,0]

Notes: Y is log wage in 1976, X_{nc} is the IQ score, X_c is a dummy for being Black. The sample size is $n = 980$, which is randomly split in two to artificially create a dataset where we observe (Y, X_c) and another one with (X_{nc}, X_c). The first column presents the OLS estimates on the full dataset, where the 95% CI have been multiplied by $\sqrt{2}$ to make it comparable with the DGM procedure using only half of it. The second column (2) presents the DGM estimates without constraints. Column (3) is the DGM estimates with a negative sign constraint on the coefficient of Black and a positive one on the coefficient of IQ score. Column (4) gathers the DGM estimates with the latter constraints plus a lower bound constraint R^2 of the long regression: $\underline{R}^2 \geq rR_s^2$, with $r = 1.3$. The R^2 on the short and long regressions are respectively 0.0328 and 0.0552.

Table 1: Bounds on the wage gap under different constraints for this NLSY sample

Affiliation:

Xavier D'Haultfœuille
CREST-ENSAE
5, avenue Henry Le Chatelier
91 120 Palaiseau, France
E-mail: `xavier.dhaultfoeuille@ensae.fr`

Christophe Gaillac
Nuffield College and the University of Oxford
10 Manor Rd
Oxford OX1 3UQ, UK
E-mail: `christophe.gaillac@economics.ox.ac.uk`

Arnaud Maurel
Duke University, NBER and IZA
Department of Economics
213 Social Sciences
Durham, NC 27708-0097, US
E-mail: `arnaud.maurel@duke.edu`