cemmap
centre for microdata methods and practice

Counterfactual analysis in R: a vignette

Mingli Chen
Victor Chernozhukov
Iván Fernández-Val
Blaise Melly

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP64/17

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

COUNTERFACTUAL ANALYSIS IN R: A VIGNETTE

MINGLI CHEN, VICTOR CHERNOZHUKOV, IVÁN FERNÁNDEZ-VAL, AND BLAISE MELLY

ABSTRACT. The R package `Counterfactual` implements the estimation and inference methods of Chernozhukov et al. (2013) for counterfactual analysis. The counterfactual distributions considered are the result of changing either the marginal distribution of covariates related to the outcome variable of interest, or the conditional distribution of the outcome given the covariates. They can be applied to estimate quantile treatment effects and wage decompositions. This vignette serves as an introduction to the package and displays basic functionality of the commands contained within.

1. INTRODUCTION

Using econometric terminology, we can often think of a counterfactual distribution as the result of a change in either the distribution of a set of covariates X that determine the outcome variable of interest Y, or the relationship of the covariates with the outcomes, that is, a change in the conditional distribution of Y given X. Counterfactual analysis consists of evaluating the effects of such changes. The R package `Counterfactual` implements the methods of Chernozhukov et al. (2013) for counterfactual analysis. It contains commands to estimate and make inference on quantile effects constructed from counterfactual distributions. The counterfactual distributions are estimated using regression methods such as classical, duration, quantile and distribution regressions. The inference on the quantile effect function can be pointwise at a specific quantile index or uniform over a range of specified quantile indexes.

We start by giving a simple example of counterfatual analysis. Suppose we would like to analyze the wage differences between men and women. Let 0 denote the population of men and let 1 denote the population of women. The variable Y_j denotes wages and X_j denotes job market-relevant characteristics that affect wages for populations $j = 0$ and $j = 1$. The conditional distribution functions $F_{Y_0|X_0}(y|x)$ and $F_{Y_1|X_1}(y|x)$ describe the stochastic assignment of wages to workers with characteristics x, for men and women, respectively. Let $F_{Y\langle 0|0\rangle}$ and $F_{Y\langle 1|1\rangle}$ represent the observed distribution function of wages for men and women, and let $F_{Y\langle 0|1\rangle}$ represent the distribution function of wages that would have prevailed for women had they faced the men's wage schedule $F_{Y_0|X_0}$:

$$F_{Y\langle 0|1\rangle}(y) := \int_{\mathcal{X}_1} F_{Y_0|X_0}(y|x) dF_{X_1}(x).$$

The latter distribution is called counterfactual, since it does not arise as a distribution from any observable population. Rather, this distribution is constructed by integrating the conditional

Version: July 16, 2017. We gratefully acknowledge research support from the NSF..

distribution of wages for men with respect to the distribution of characteristics for women. This quantity is well defined if $\mathcal{X}_0$, the support of men's characteristics, includes $\mathcal{X}_1$, the support of women's characteristics, namely $\mathcal{X}_1 \subset \mathcal{X}_0$.

Let $F^{\leftarrow}$ denote the quantile or left-inverse function of the distribution function F. The difference in the observed wage quantile function between men and women can be decomposed in the spirit of Oaxaca (1973) and Blinder (1973) as

$$F^{\leftarrow}_{Y\langle 1|1\rangle} - F^{\leftarrow}_{Y\langle 0|0\rangle} = [F^{\leftarrow}_{Y\langle 1|1\rangle} - F^{\leftarrow}_{Y\langle 0|1\rangle}] + [F^{\leftarrow}_{Y\langle 0|1\rangle} - F^{\leftarrow}_{Y\langle 0|0\rangle}], \tag{1}$$

where the first term in brackets is due to differences in the wage structure and the second term is a composition effect due to differences in characteristics. These counterfactual effects are well defined econometric parameters and are widely used in empirical analysis, for example, the first term of the decomposition is a measure of gender wage discrimination. In Section 3.2 we consider an empirical example where 0 denotes the population of nonunion workers and 1 denotes the population of union workers. In this case the the wage structure effect corresponds to the treatment effect of union or union premium. It is important to note that these effects do not necessarily have a causal interpretation without additional conditions that are spelled out in Chernozhukov et al. (2013).

2. The Counterfactual Package

2.1. **Getting Started.** To get started using the package `Counterfactual` for the first time, issue the command

```
> install.packages("Counterfactual")
```

into your R browser to install the package in your computer. Once the package has been installed, you can use the package `Counterfactual` during any R session by simply issuing the command

```
> library(Counterfactual)
```

Now you are ready to use the function `counterfactual` and data sets contained in `Counterfactual`. For general questions about the package you may type

```
> help(package = "Counterfactual")
```

to view the package help file, or for more questions about a specific function you can type `help(`*`function-name`*`)`. For example, try:

```
> help(counterfactual)
```

or simply type

```
> ?counterfactual
```

The command `counterfactual` has the general syntax:

```
> counterfactual(formula, data, weights, na.action = na.exclude,
+                group, treatment = FALSE, decomposition = FALSE,
+                counterfactual_var, transformation = FALSE,
+                quantiles = c(1:9)/10, method = "qr", discrete = FALSE,
+                trimming = 0.005, nreg = 100, scale_variable,
```

```
+                 counterfactual_scale_variable,
+                 censoring = 0, right = FALSE, nsteps = 3,
+                 firstc = 0.1, secondc = 0.05, noboot = FALSE,
+                 weightedboot = FALSE, seed = 8, robust = FALSE,
+                 reps = 100, alpha = 0.05, first = 0.1,
+                 last = 0.9, cons_test = 0, printdeco = TRUE,
+                 sepcore = FALSE, ncore=1)
```

To describe the different options of the command we need to provide some background on methods for counterfactual analysis.

2.2. **Setting for Counterfactual Analysis.** Consider a general setting with two populations labeled by $k \in \mathcal{K} = \{0, 1\}$. For each population k there is the d_x-vector X_k of covariates and the scalar outcome Y_k. The covariate vector is observable in all populations, but the outcome is only observable in populations $j \in \mathcal{J} \subseteq \mathcal{K}$. Let F_{X_k} denote the covariate distribution in population $k \in \mathcal{K}$, and $F_{Y_j|X_j}$ and $Q_{Y_j|X_j}$ denote the conditional distribution and quantile functions in population $j \in \mathcal{J}$. We denote the support of X_k by $\mathcal{X}_k \subseteq \mathbb{R}^{d_x}$, and the region of interest for Y_j by $\mathcal{Y}_j \subseteq \mathbb{R}$. The refer to j as the reference population(s) and to k as the counterfactual population(s).

The reference and counterfactual populations in the wage examples correspond to different groups such as men and women or nonunion and union workers. We can also generate counterfactual populations by artificially transforming a reference population. Formally, we can think of X_k as being created through a known transformation of X_j:

$$X_k = g_k(X_j), \quad \text{where} \quad g_k : \mathcal{X}_j \to \mathcal{X}_k. \tag{2}$$

This case covers adding one unit to the first covariate, $X_{1,k} = X_{1,j} + 1$, holding the rest of the covariates constant. The resulting quantile effect becomes the *unconditional* quantile regression, which measures the effect of a unit change in a given covariate component on the unconditional quantiles of Y. For example, this type of counterfactual is useful for estimating the treatment effect of smoking during pregnancy on infant birth weights. Another possible transformation is a mean preserving redistribution of the first covariate implemented as $X_{1,k} = (1 - \alpha)E[X_{1,j}] + \alpha X_{1,j}$. These and more general types of transformation defined in (2) are useful for estimating the effect of a change in taxation on the marginal distribution of food expenditure or the effect of cleaning up a local hazardous waste site on the marginal distribution of housing prices (Stock (1991)). We give an example of this type of transformation in Section 3.1.

The reference and counterfactual populations can be specified to `counterfactual` in two ways that accomodate the previous two cases:

(1) If the option `group` has been specified, then j is the population defined by `group=0` and k is the population defined by `group=1`. This means that both X and Y are observed in `group=0`, but only X needs to be observed in `group=1`. When both X and Y are observed in `group=1`, the option `treatment=TRUE` specifies that the structure or treatment

effect should be computed, whereas the dafault option `treatment=FALSE` specifies that the composition effect should be computed; see the definition of the structure and composition effects in the decomposition (1). If in addition to `treatment=TRUE` the option `decomposition=TRUE` is selected, then the entire decompostion (1) is reported including the composition, structure and total effects. Note that we can reverse the roles of the populations defined by an indicator variable `vargroup` by setting either `group=vargroup` or `group=1-vargroup`.

(2) Alternatively, the option `counterfactual_var` can be used to specify the covariates in the counterfactual population. In this case, the names on the right handside of `formula` contain the variables in X_j and `counterfactual_var` contains the variables in X_k. The option `transformation=TRUE` should be used when X_k is generated as a transformation of X_j, e.g., equation (2). The list passed to `counterfactual_var` must contain exactly the same number of variables as the list of independent variables in `formula` and the order of the variables in the list matters.

Counterfactual distribution and quantile functions are formed by combining the conditional distribution in the population j with the covariate distribution in the population k, namely:

$$F_{Y\langle j|k\rangle}(y) := \int_{\mathcal{X}_k} F_{Y_j|X_j}(y|x)dF_{X_k}(x), \quad y \in \mathcal{Y}_j,$$
$$Q_{Y\langle j|k\rangle}(\tau) := F^{\leftarrow}_{Y\langle j|k\rangle}(\tau), \quad \tau \in (0,1),$$

where $(j,k) \in \mathcal{JK}$, and $F^{\leftarrow}_{Y\langle j|k\rangle}(\tau) = \inf\{y \in \mathcal{Y}_j : F_{Y\langle j|k\rangle}(y) \geq \tau\}$ is the left-inverse function of $F_{Y\langle j|k\rangle}$. The main interest lies in the quantile effect (QE) function, defined as the difference of two counterfactual quantile functions over a set of quantile indexes $\mathcal{T} \subset (0,1)$:

$$\Delta_C(\tau) = Q_{Y\langle j|k\rangle}(\tau) - Q_{Y\langle j|j\rangle}(\tau), \ \tau \in \mathcal{T},$$

where $j \in \mathcal{J}$ and $k \in \mathcal{K}$. In the example of Section 1, we obtain the composition effect with $j = 0$ and $k = 1$. When Y_k is observed, then we can construct the structure effect or treatment effect on the treated

$$\Delta_S(\tau) = Q_{Y\langle k|k\rangle}(\tau) - Q_{Y\langle j|k\rangle}(\tau), \ \tau \in \mathcal{T},$$

by specifying the option `group` and setting `treatment=TRUE`. In the example of Section 1, we obtain the wage structure effect with $j = 0$ and $k = 1$, i.e. setting `group=1` and `treatment=TRUE`. If in addition we select the option `decomposition=TRUE`, then we obtain the entire decomposition (1) including the composition, structure and total effects. The total effect is

$$\Delta_T(\tau) = Q_{Y\langle k|k\rangle}(\tau) - Q_{Y\langle j|j\rangle}(\tau), \ \tau \in \mathcal{T}.$$

The set $\mathcal{T}$ is specified with the option `quantiles`, which enumerates the quantile indexes of interested and should be a vector containing numbers between 0 and 1.

To estimate the QE function we need to model and estimate the conditional distribution $F_{Y_j|X_j}$ and covariate distribution F_{X_k}. We estimate the covariate distribution using the empirical distribution, and consider several regression based methods for the conditional distribution including classical, quantile, duration, and distribution regression. Given the estimators of the

conditional and covariate distributions $\hat{F}_{Y_j|X_j}$ and $\hat{F}_{X_k}$, the estimator of each counterfactual distribution is obtained by the plug-in rule, namely

$$\hat{F}_{Y\langle j|k\rangle}(y) = \int_{\mathcal{X}_k} \hat{F}_{Y_j|X_j}(y|x) d\hat{F}_{X_k}(x), y \in \mathcal{Y}_j.$$

Then, the estimator of the QE function is also obtained by the plug-in rule as

$$\hat{\Delta}_C(\tau) = \hat{F}^{\leftarrow}_{Y\langle j|k\rangle}(\tau) - \hat{F}^{\leftarrow}_{Y\langle j|j\rangle}(\tau), \quad \tau \in \mathcal{T},$$

or

$$\hat{\Delta}_S(\tau) = \hat{F}^{\leftarrow}_{Y\langle k|k\rangle}(\tau) - \hat{F}^{\leftarrow}_{Y\langle j|k\rangle}(\tau), \quad \tau \in \mathcal{T},$$

if we define the counterfactual population with `group` and set `treatment=TRUE`. If in addition to `treatment=TRUE`, we select `decomposition=TRUE`, then the plug-in estimator of the total effect is

$$\hat{\Delta}_T(\tau) = \hat{F}^{\leftarrow}_{Y\langle k|k\rangle}(\tau) - \hat{F}^{\leftarrow}_{Y\langle j|j\rangle}(\tau), \quad \tau \in \mathcal{T}.$$

2.2.1. *Estimation of Conditional Distribution.* In this section we assume that we have samples $\{(Y_{ji}, X_{ji}) : i = 1, \ldots, n_j\}$ composed of independent and identically distributed copies of (Y_j, X_j) for all populations $j \in \mathcal{J}$. The conditional distribution $F_{Y_j|X_j}$ can be modeled and estimated directly, or throught the conditional quantile function, $Q_{Y_j|X_j}$, using the relation

$$F_{Y_j|X_j}(y|x) \equiv \int_{(0,1)} 1\{Q_{Y_j|X_j}(u|x) \leq y\} du. \tag{3}$$

The option `formula` specifies the outcome Y as the left hand side variable and the covariates X as the right hand side variable(s). The option `method` allows to select the method to estimate the conditional distribution. The following methods are implemented:

(1) `method = "qr"`, which is the default, implements the quantile regresion estimator of the conditional distribution

$$\hat{F}_{Y_j|X_j}(y|x) = \varepsilon + \int_{(\varepsilon, 1-\varepsilon)} 1\{x'\hat{\beta}_j(u) \leq y\} du, \tag{4}$$

where ε is a small constant that avoids estimation of tail quantiles, and $\hat{\beta}(u)$ is the Koenker and Bassett (1978) quantile regression estimator

$$\hat{\beta}_j(u) = \arg\min_{b \in \mathbb{R}^{d_x}} \sum_{i=1}^{n_j} [u - 1\{Y_{ji} \leq X'_{ji}b\}][Y_{ji} - X'_{ji}b].$$

The quantile regression estimator calls the R package `quantreg` (Koenker, 2016). The option `trimming` specifies the value of the trimming parameter ε, with default value $\varepsilon = 0.005$. The option `nreg` sets the number of quantile regressions used to approximate the integral in (4), with a default value of 100 such that $(\varepsilon, 1-\varepsilon)$ is approximated by the grid $\{\varepsilon, \varepsilon + (1-2\varepsilon)/99, \varepsilon + 2(1-2\varepsilon)/99, \ldots, 1-\varepsilon\}$. This method should be used only with continuous dependent variables.

(2) `method = "loc"` implements the estimator of the conditional distribution

$$\hat{F}_{Y_j|X_j}(y|x) = \frac{1}{n_j}\sum_{i=1}^{n_j} 1\{Y_{ji} - X_{ji}'\hat{\beta}_j \leq y - x'\hat{\beta}_j\}, \tag{5}$$

where $\hat{\beta}_j$ is the least square estimator

$$\hat{\beta}_j = \arg\min_{b\in\mathbb{R}^{d_x}} \sum_{i=1}^{n_j}(Y_{ji} - X_{ji}'b)^2. \tag{6}$$

The estimator (5) is based on a restrictive location shift model that imposes that the covariates X only affect the location of the outcome Y.

(3) `method = "locsca"` implements the estimator of the conditional distribution

$$\hat{F}_{Y_j|X_j}(y|x) = \frac{1}{n_j}\sum_{i=1}^{n_j} 1\left\{\frac{Y_{ji} - X_{ji}'\hat{\beta}_j}{\exp(X_{2ji}'\hat{\gamma}_j/2)} \leq \frac{y - x'\hat{\beta}_j}{\exp(x_{2j}'\hat{\gamma}_j/2)}\right\}, \tag{7}$$

where $\hat{\beta}_j$ is the least square estimator (6), $X_{2j} \subseteq X_j$ with $\dim X_{2j} = d_{x_2}$, and

$$\hat{\gamma}_j = \arg\min_{g\in\mathbb{R}^{d_{x_2}}} \sum_{i=1}^{n_j}(\log(Y_{ji} - X_{ji}'\hat{\beta}_j)^2 - X_{2ji}'g)^2.$$

The option `scale_variable` specifies the covariates X_{2j} that affect the scale of the conditional distribution. The option `counterfactual_scale_variable` selects the counterfactual scale variables when the counterfactual population is specified using `counterfactual_var`. By default, `R` would use all the covariates as `scale_variable` and `counterfactual_scale_variable = counterfactual_var`. The estimator (7) is based on a restrictive location scale shift model that imposes that the covariates X only affect the location and scale of the outcome Y.

(4) `method = "cqr"` implements the censored quantile regression estimator of the conditional distribution, which is the same as (4) with $\hat{\beta}(u)$ replaced by the Chernozhukov and Hong (2002) censored quantile regression estimator. The options `trimming` and `nreg` apply to this method with the same functionality as for the `qr` method. Moreover, a variable containing a censoring indicator C_j must be specified with `censoring`. The censored quantile regression estimator has three-steps by default. The number of steps can be increased by the option `nsteps`. In the first step, the censoring probabilities are estimated by a logit regression of the censoring indicator C_j on all the covariates X_j. Then, for each quantile index u, the observations with sufficiently low censoring probabilities relative to u are selected. We allow for misspecification of the logit by excluding the observations that could theoretically be used but have censoring probabilities in the highest `firstc` quantiles, with a default of 0.1, i.e. 10% of the observations. In the second step, standard linear quantile regressions are estimated on the samples defined in step one. Using the estimated quantile regressions, we define a new sample of observations that can be used. This sample consists of all observations for which the estimated conditional quantile is above the censoring point. Again, we throw away observations in

the lowest `secondc` quantiles of the distribution of the residuals, with a default of 0.05, i.e. 5% of the observations. Step three consists in a new linear quantile regression using the sample defined in step two. Step three is repeated if `nsteps` is above 3. This method should be used only with censored dependent variables.

(5) `method = "cox"` implements the duration regression estimator of the conditional distribution function

$$\hat{F}_{Y_j|X_j}(y|x) = 1 - \exp(-\exp(\hat{t}(y) - x'\hat{\beta})), \tag{8}$$

where $\hat{\beta}$ is the Cox estimator of the regression coefficients and $\hat{t}(y)$ is the Cox estimator of the baseline integrated hazard function (Cox, 1972). The Cox estimator calls the R package `survival` (Therneau, 2015). The estimator (8) is based on a restrictive transformation location shift model that imposes that the covariates X only affect the location of a monotone transformation of the outcome $t(Y)$, i.e.

$$t(Y_j) = X_j'\beta_j + V_j,$$

where V_j has an extreme value distribution and is independent of X_j. This method should be used only with nonnegative dependent variables.

(6) `method = "logit"` implements the distribution regression estimator of the conditional distribution with logistic link function

$$\hat{F}_{Y_j|X_j}(y|x) = \Lambda(x'\hat{\beta}(y)), \tag{9}$$

where Λ is the standard logistic distribution function, and $\hat{\beta}(y)$ is the distribution regression estimator

$$\hat{\beta}(y) = \arg\max_{b\in\mathbb{R}^{d_x}} \sum_{i=1}^{n_j} \left[1\{Y_{ji} \leq y\} \log \Lambda(X_{ij}'b) + 1\{Y_{ij} > y\} \log \Lambda(-X_{ji}'b)\right]. \tag{10}$$

The estimator (9) is based on a flexible model where each covariate can have a heterogenous effect at different parts of the distribution. This method can be used with continuous dependent variables and censored dependent variables with fixed censoring point.

(7) `method = "probit"` implements the distribution regression estimator of the conditional distribution with normal link function, i.e. where Λ is the standard normal distribution function in (9) and (10).

(8) `method = "lpm"` implements the linear probability model estimator of the conditional distribution

$$\hat{F}_{Y_j|X_j}(y|x) = x'\hat{\beta}(y),$$

where $\hat{\beta}(y)$ is the least squares estimator

$$\hat{\beta}(y) = \arg\min_{b\in\mathbb{R}^{d_x}} \sum_{i=1}^{n_j} (1\{Y_{ji} \leq y\} - X_{ij}'b)^2.$$

This method might produce estimates of the conditional distribution outside the interval $[0, 1]$.

For the methods (2)–(8), the option `nreg` sets the number of values of y to evaluate the estimator of the conditional distribution function. These values are uniformly distributed among the observed values of Y_j. If `nreg` is greater than the number of observed values of Y_j, then all the observed values are used.

2.3. **Inference.** The command `counterfactual` reports pointwise and uniform confidence intervals for the QEs over a prespecified set of quantile indexes. The construction of the intervals rely on functional central limit theorems and bootstrap functional central limit theorems for the empirical QEs derived in Chernozhukov et al. (2013). In particular, the pointwise intervals are based on the normal distribution, whereas the uniform intervals are based on two resampling schemes: empirical and weighted bootstrap. Thus, the $(1-\alpha)$ confidence interval for $\Delta(\tau)$ on $\mathcal{T}$ has the form

$$\{\hat{\Delta}(\tau) \pm c_{1-\alpha}\hat{\Sigma}(\tau) : \tau \in \mathcal{T}\},$$

where $\hat{\Sigma}(\tau)$ is the standard error of $\hat{\Delta}(\tau)$ and $c_{1-\alpha}$ is a critical value. There are two options to obtain $\hat{\Sigma}(\tau)$. The default option `robust=FALSE` computes the bootstrap standard deviation of $\hat{\Delta}(\tau)$; whereas the option `robust=TRUE` computes the bootstrap interquartile range rescaled with the normal distribution, $\hat{\Sigma}(\tau) = (q_{0.75}(\tau) - q_{0.25}(\tau))/(z_{0.75} - z_{0.25})$ where $q_p(\tau)$ is the pth quantile of the bootstrap draws of $\hat{\Delta}(\tau)$ and z_p is the pth quantile of the standard normal. The pointwise critical value is $c_{1-\alpha} = z_{1-\alpha}$, and the uniform critical value is $c_{1-\alpha} = \hat{t}_{1-\alpha}$, where $\hat{t}_{1-\alpha}$ is a bootstrap estimator of the $(1-\alpha)$th quantile of the Kolmogorov-Smirnov maximal t-statistic

$$t = \sup_{\tau \in \mathcal{T}} |\hat{\Delta}(\tau) - \Delta(\tau)|/\hat{\Sigma}(\tau).$$

In addition to the intervals, `counterfactual` reports the p-values for several functional tests based on two test-statistic: Kolmogorov-Smirnov and the Cramer-von-Misses-Smirnov. The null-hypotheses considered are

(1) Correct parametric specification of the model for the conditional distribution. This test compares the empirical distribution of the outcome Y_j with the estimate of the counterfactual distribution in the reference population

$$\hat{F}_{Y\langle j|j\rangle}(y) := \int_{\mathcal{X}_j} \hat{F}_{Y_j|X_j}(y|x) d\hat{F}_{X_j}(x).$$

The power of this specification test might be low because it only uses the implications of the conditional distribution on the counterfactual distribution. For example, the test is not informative for the linear probability and logit models where the counterfactual distribution in the reference population is identical to the empirical distribution by construction. If `group` is specified and `treatment=TRUE` is selected, then the test is performed in the population defined by `group=1`. If in addition the option `decomposition=TRUE` is selected, then the test is performed in the populations defined by `group=0` and `group=1`, and in the combined population including both `group=0` and `group=1`.

(2) Zero QE at all the quantile indexes of interest: $\Delta(\tau) = 0$ for all $\tau \in \mathcal{T}$. This is stronger than a zero average effect. Other null hypotheses of constant quantile effect (but at a different level than 0) can be added with the option `cons_test`.
(3) Constant QE at all quantile indexes of interest: $\Delta(\tau) = \Delta(0.5)$ for all $\tau \in \mathcal{T}$.
(4) First-order stochastic dominance: $\Delta(\tau) \geq 0$ for all $\tau \in \mathcal{T}$.
(5) Negative first-order stochastic dominance: $\Delta(\tau) \leq 0$ for all $\tau \in \mathcal{T}$.

The options of `counterfactual` related to inference are:

(1) `noboot = TRUE` suppresses the bootstrap. The bootstrap can be very demanding in terms of computation time. Therefore, it is recommended to switch it off for the exploratory analysis of the data.
(2) `weightedboot = TRUE` selects weighted bootstrap with standard exponential weights. The default `weightedboot = FALSE` selects empirical bootstrap with multinomial weights. We recommend weighted bootstrap when the covariates include categorical variables with small cell sizes to avoid singular designs in the bootstrap draws.
(3) `reps` specifies the number of bootstrap replications. This number will matter only if the bootstrap has not be suppressed. The default is 100.
(4) `alpha` specifies the significance level of the tests and confidence intervals. Note that the confidence level of the confidence interval is 1 - `alpha`. Thus, the default value of 0.05 produces 95% confidence intervals.
(5) `first` and `last` select the subset of quantile indexes of interest for inference. The tails of the distribution should not be used because standard asymptotic does not apply to these parts. The needed amount of tail trimming depends on the sample size and on the distribution of the dependent variable. `first` sets the lowest quantile index used and `last` sets the highest quantile index used. The default values are 0.1 and 0.9 so that $\mathcal{T} = [0.1, 0.9]$.
(6) `cons_test` add tests of the null hypothesis that $\Delta(\tau) =$ `const_test` for all τ between `first` and `last`. The null hypothesis that $\Delta(\tau) = 0$ for all τ between `first` and `last` is tested by default. The null hypothesis that the quantile effects are constant is also tested by default.

2.4. **Parallel Computing.** The command `counterfactual` provides functionality for parallel computing, which is specially useful to speed up the execution of the bootstrap. There are two options related to parallel computing:

(1) `sepcore` specifies whether multiple cores should be used. The default value `sepcore = FALSE` turns off the parallel computing.
(2) `ncore` selects the number of cores to use for parallel computing. The information of this option is only used when parallel computing is switched on with `sepcore = TRUE`.

3. Empirical Examples

We consider two empirical examples to illustrate the functionality of the command `counterfactual`. The first example is an estimation of Engel curves that includes a counterfactual analysis where the counterfactual population is an artificial transformation of a reference population. The second example is wage decomposition with respect to union status where the reference and counterfactual populations correspond to two different groups.

3.1. **Engel Curves.** We use the classical Engel 1857 dataset to estimate the relationship between food expenditure (`foodexp`) and annual household income (`income`), and then report the estimates of the QE of a change in the distribution of the annual household income that might be induced for example by a variation in income taxation.[1] We estimate the conditional distribution with the quantile regression method, i.e., `method ="qr"`.

First, we generate the variable `counterfactual_income` with the counterfactual values of income and plot the reference and counterfactual income distributions. The counterfactual distribution corresponds to a mean preserving spread of the distribution in the reference population that reduces standard deviation by 25%.

```
> library(quantreg)
> data(engel)
> attach(engel)
> counter_income <- mean(income)+0.75*(income-mean(income))
> cdfx <- c(1:length(income))/length(income)
> plot(c(0,4000),range(c(0,1)), xlim =c(0, 4000), type="n", xlab = "Income",
+      ylab="Probability")
> lines(sort(income), cdfx)
> lines(sort(counter_income), cdfx, lwd = 2, col = 'grey70')
> legend(1500, .2, c("Original", "Counterfactual"), lwd=c(1,2),bty="n",
+        col=c(1,'grey70'))
```

To estimate the QEs of this counterfactual change we turn on the option `transformation` of `counterfactual` by setting `transformation = TRUE`, and specify that the counterfactual values of the covariate `income` are in the generated variable `counter_income` by setting `counterfactual_var = counter_income`. This yields:

```
> qrres <- counterfactual(foodexp~income, counterfactual_var
+         = counter_income, transformation = TRUE, sepcore = TRUE, ncore = 2)
cores used= 2

Conditional Model:                       linear quantile regression
Number of regressions estimated:          100
```

[1] This is the same data set as in the quantile regression package `quantreg`, see Koenker (2016).

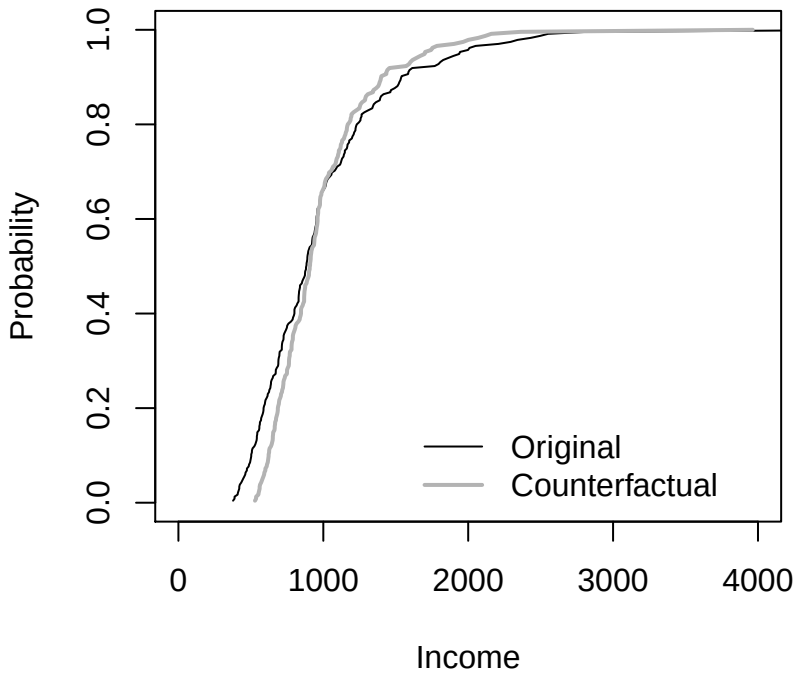

FIGURE 1. Observed and counterfactual distributions of income

```
The variance has been estimated by bootstraping the results 100 times.

No. of obs. in the reference group:      235
No. of obs. in the counterfactual group: 235


                    Quantile Effects -- Composition
-----------------------------------------------------------------------
                     Pointwise      Pointwise           Functional
 Quantile     Est.    Std.Err   95% Conf.Interval   95% Conf.Interval
     0.1     55.2      4.48      46.4        64      42.9      67.5
     0.2     48.1      4.06      40.1        56      36.9      59.3
     0.3     38.9      4.42      30.3      47.6      26.8      51.1
     0.4     27.2      4.36      18.6      35.7      15.2      39.2
     0.5     16.6      4.18      8.44      24.8      5.11      28.1
     0.6     5.86      4.56     -3.07      14.8      -6.7      18.4
     0.7    -5.84      5.37     -16.4      4.69     -20.6      8.97
```

```
     0.8     -30.6      8.68     -47.6     -13.6     -54.5     -6.66
     0.9     -78.3      12.4      -103       -54      -113     -44.1


      Bootstrap inference on the counterfactual quantile process
------------------------------------------------------------------------
                                                    P-values
                                                ======================
NULL-Hypthoesis                                 KS-statistic CMS-statistic
========================================================================
Correct specification of the parametric model       0.38       0.23
No effect: QE(tau)=0 for all taus                   0.00       0.00
Constant effect: QE(tau)=QE(0.5) for all taus       0.00       0.00
Stochastic dominance: QE(tau)>0 for all taus        0.00       0.00
Stochastic dominance: QE(tau)<0 for all taus        0.00       0.00
```

We reject the simultaneous hypotheses of zero, constant, positive and negative effect of the income redistribution at all the deciles. The QR model for the conditional distribution cannot be rejected at conventional significance levels.

Finally, we reestimate the QE function on the larger set of quantiles $\{0.01, 0.02, \ldots, 0.99\}$, and plot a uniform confidence band over the subset $\{0.10, 0.11, \ldots, 0.90\}$ constructed by empirical bootstrap with 100 replications. In Figure 2 we can visually reject the functional hypotheses of zero, constant, positive and negative effect at the percentiles considered. We use the option `printdeco = FALSE` to supress the display of the table of results.

```
> taus  <- c(1:99)/100
> first <- sum(as.double(taus <= .10))
> last  <- sum(as.double(taus <= .90))
> rang  <- c(first:last)
> rqres <- counterfactual(foodexp~income, counterfactual_var=counter_income,
+                         nreg=100, quantiles=taus, transformation = TRUE,
+                         printdeco = FALSE, sepcore = TRUE,ncore=2)

cores used= 2

> duqf   <- (rqres$resCE)[,1]
> l.duqf <- (rqres$resCE)[,5]
> u.duqf <- (rqres$resCE)[,6]

> plot(c(0,1), range(c(min(l.duqf[rang]), max(u.duqf[rang]))), xlim = c(0,1),
+      type = "n", xlab = expression(tau), ylab = "Difference in Food Expenditure",
+      cex.lab=0.75)
> polygon(c(taus[rang], rev(taus[rang])), c(u.duqf[rang], rev(l.duqf[rang])),
+         density = -100, border = F, col = "grey70", lty = 1, lwd = 1)
```

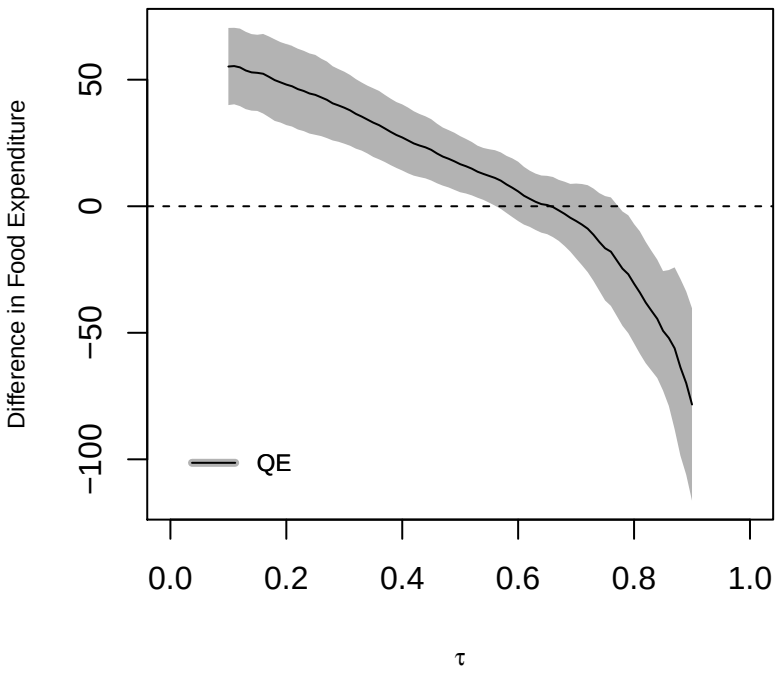

FIGURE 2. Quantile effects of income redistribution on food consumption

```
> lines(taus[rang], duqf[rang])
> abline(h = 0, lty = 2)
> legend(0, -90, "QE", cex = 0.75, lwd = 4, bty = "n", col = "grey70")
> legend(0, -90, "QE", cex = 0.75, lty = 1, bty = "n", lwd = 1)
```

3.2. **Union Premium.** We use an extract of the U.S. National Longitudinal Survey of Young Women (NLSW) for employed women in 1988 to estimate a wage decomposition with respect to union status.[2] The outcome variable Y is the log hourly wage (`lwage`), the covariates X include job tenure in years (`tenure`), years of schooling (`grade`), and total experience (`ttl_exp`), and the union indicator `union` defines the reference and counterfactual populations. We estimate the conditional distributions by distribution regression with logistic link and duration regression, i.e., `method ="logit"` and `method ="cox"`. We use weighted bootstrap for the construction of uniform confidence intervals and hypothesis tests and run parallel computing with 2 nodes.

[2]This dataset is available from the Stata's sample datasets at `http://www.stata-press.com/data/r9/nlsw88.dta`.

We start by estimating the wage decomposition by logistic distribution regression, where the counterfactual population is specified with *group=union* with the options *treatment=TRUE* and *decomposition=TRUE* to estimate the composition, structure and total effects. The estimates of the total effect show that the union wage gap is positive throughout the distribution and decreasing in the quantile index. Indeed, we reject the hypothesis that the differences between the deciles of wages between union and nonunion workers are all equal or negative, but cannot reject that they are all positive at conventional significance levels.

In this example the composition effect captures differences in tenure, education and experience between union and nonunion workers, and the structure effect corresponds to the treatment effect of union on the treated or union premium. The estimates of the composition and structure effects show that differences in worker characteristics account for about half of the gap in the upper tail of the distribution, whereas in the rest of the distribution the union premium explains most of the gap. The deciles of the structure effect are all positive and heterogeneous. The deciles of the composition effect are also positive, but we cannot reject that they are constant. All the specification tests for the parametric model report *NA* because these tests are uninformative about the logistic distribution regression model.

```
> data(nlsw88)
> attach(nlsw88)
> lwage    <- log(wage)
> logitres <- counterfactual(lwage~tenure+ttl_exp+grade,
+                            group = union, treatment=TRUE,
+                            decomposition=TRUE, method = "logit",
+                            weightedboot = TRUE, sepcore = TRUE, ncore=2)
cores used= 2

Conditional Model:                       logit
Number of regressions estimated:          100

The variance has been estimated by bootstraping the results 100 times.

No. of obs. in the reference group:      1407
No. of obs. in the counterfactual group: 459


                    Quantile Effects -- Structure
------------------------------------------------------------------------
                     Pointwise      Pointwise           Functional
 Quantile     Est.    Std.Err   95% Conf.Interval   95% Conf.Interval
     0.1     0.239     0.0439     0.154     0.325     0.111     0.368
     0.2     0.208      0.038     0.133     0.282    0.0968     0.319
```

0.3	0.218	0.0277	0.163	0.272	0.137	0.299
0.4	0.19	0.0326	0.126	0.254	0.0947	0.285
0.5	0.157	0.0327	0.0925	0.221	0.0609	0.252
0.6	0.15	0.0301	0.0909	0.209	0.0618	0.238
0.7	0.0714	0.0319	0.00886	0.134	-0.022	0.165
0.8	0.0173	0.0264	-0.0345	0.069	-0.06	0.0946
0.9	-0.00794	0.0516	-0.109	0.0932	-0.159	0.143

Bootstrap inference on the counterfactual quantile process

	P-values	
NULL-Hypthoesis	KS-statistic	CMS-statistic
Correct specification of the parametric model	NA	NA
No effect: QE(tau)=0 for all taus	0	0
Constant effect: QE(tau)=QE(0.5) for all taus	0	0.01
Stochastic dominance: QE(tau)>0 for all taus	0.89	0.9
Stochastic dominance: QE(tau)<0 for all taus	0	0

Quantile Effects -- Composition

		Pointwise	Pointwise		Functional	
Quantile	Est.	Std.Err	95% Conf.Interval		95% Conf.Interval	
0.1	0.0606	0.0269	0.00797	0.113	-0.0223	0.144
0.2	0.0545	0.0239	0.0076	0.101	-0.0194	0.128
0.3	0.0691	0.0254	0.0193	0.119	-0.00942	0.148
0.4	0.0821	0.0272	0.0288	0.135	-0.00192	0.166
0.5	0.0982	0.0278	0.0438	0.153	0.0124	0.184
0.6	0.112	0.0247	0.0637	0.161	0.0358	0.189
0.7	0.115	0.0331	0.0505	0.18	0.0133	0.217
0.8	0.0975	0.0221	0.0543	0.141	0.0294	0.166
0.9	0.0613	0.039	-0.0151	0.138	-0.0591	0.182

Bootstrap inference on the counterfactual quantile process

	P-values

	P-values	
NULL-Hypthoesis	KS-statistic	CMS-statistic
Correct specification of the parametric model	NA	NA
No effect: QE(tau)=0 for all taus	0	0
Constant effect: QE(tau)=QE(0.5) for all taus	0.46	0.31
Stochastic dominance: QE(tau)>0 for all taus	0.91	0.91
Stochastic dominance: QE(tau)<0 for all taus	0	0

Quantile Effects -- Total

		Pointwise	Pointwise		Functional	
Quantile	Est.	Std.Err	95% Conf.Interval		95% Conf.Interval	
0.1	0.3	0.0437	0.214	0.386	0.186	0.414
0.2	0.262	0.0417	0.181	0.344	0.153	0.372
0.3	0.287	0.0374	0.214	0.36	0.189	0.385
0.4	0.272	0.0353	0.203	0.341	0.18	0.364
0.5	0.255	0.0373	0.182	0.328	0.157	0.353
0.6	0.262	0.0348	0.194	0.33	0.171	0.353
0.7	0.187	0.0294	0.129	0.244	0.11	0.264
0.8	0.115	0.0289	0.0581	0.171	0.0392	0.19
0.9	0.0534	0.0451	-0.0351	0.142	-0.0647	0.172

Bootstrap inference on the counterfactual quantile process

	P-values	
NULL-Hypthoesis	KS-statistic	CMS-statistic
Correct specification of the parametric model	NA	NA
No effect: QE(tau)=0 for all taus	0	0
Constant effect: QE(tau)=QE(0.5) for all taus	0.03	0.02
Stochastic dominance: QE(tau)>0 for all taus	0.9	0.9
Stochastic dominance: QE(tau)<0 for all taus	0	0

Next, we reestimate the QE function on the larger set of quantiles $\{0.01, 0.02, ..., 0.99\}$, and plot a uniform confidence band over the subset $\{0.10, 0.11, ..., 0.90\}$) constructed by weighted bootstrap with 100 replications. Figure 3 shows that the findings for the deciles carry over to the percentiles considered. Thus, we can visually test that the structure effect is heterogeneous

and explains most of the union wage gap below the third quartile. The composition effect is homogeneous and explains most of the wage gap above the ninth decile. When we consider the finer grid of percentiles, however, we can no longer reject that there is no composition effect at the 5% significance since the dashed line at zero is fully covered by the bands.

```
> taus  <- c(1:99)/100
> first <- sum(as.double(taus <= .10))
> last  <- sum(as.double(taus <= .90))
> rang  <- c(first:last)
> logitres <- counterfactual(lwage~tenure+ttl_exp+grade,
+           group = union, treatment=TRUE, quantiles=taus,
+           method="logit", nreg=100, weightedboot = TRUE,
+           printdeco=FALSE, decomposition = TRUE,
+           sepcore = TRUE,ncore=2)
cores used= 2
> duqf_SE   <- (logitres$resSE)[,1]
> l.duqf_SE <- (logitres$resSE)[,5]
> u.duqf_SE <- (logitres$resSE)[,6]
> duqf_CE   <- (logitres$resCE)[,1]
> l.duqf_CE <- (logitres$resCE)[,5]
> u.duqf_CE <- (logitres$resCE)[,6]
> duqf_TE   <- (logitres$resTE)[,1]
> l.duqf_TE <- (logitres$resTE)[,5]
> u.duqf_TE <- (logitres$resTE)[,6]
> range_x <- min(c(min(l.duqf_SE[rang]), min(l.duqf_CE[rang]),
+                 min(l.duqf_TE[rang])))
> range_y <- max(c(max(u.duqf_SE[rang]), max(u.duqf_CE[rang]),
+                 max(u.duqf_TE[rang])))
> par(mfrow=c(1,3))
> plot(c(0,1), range(c(range_x, range_y)), xlim = c(0,1), type = "n",
+      xlab = expression(tau), ylab = "Difference in Wages", cex.lab=0.75,
+      main = "Total")
> polygon(c(taus[rang],rev(taus[rang])),
+         c(u.duqf_TE[rang], rev(l.duqf_TE[rang])), density = -100, border = F,
+         col = "grey70", lty = 1, lwd = 1)
> lines(taus[rang], duqf_TE[rang])
> abline(h = 0, lty = 2)
> plot(c(0,1), range(c(range_x, range_y)), xlim = c(0,1), type = "n",
+      xlab = expression(tau), ylab = "", cex.lab=0.75, main = "Structure")
> polygon(c(taus[rang],rev(taus[rang])),
+         c(u.duqf_SE[rang], rev(l.duqf_SE[rang])), density = -100, border = F,
```

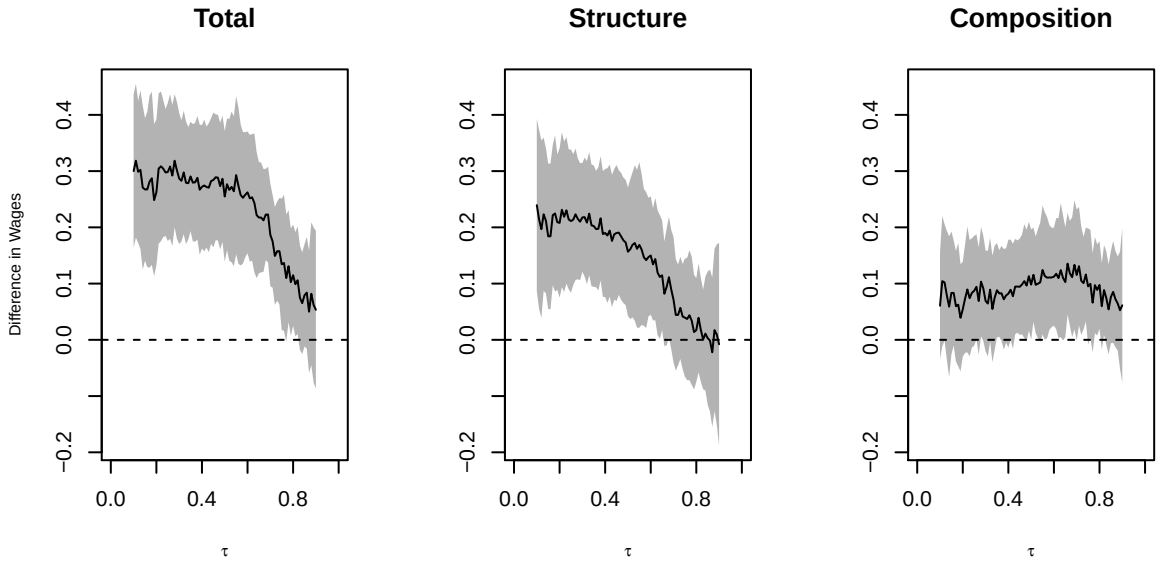

FIGURE 3. Wage decomposition with respect to union: logit regression estimates

```
+         col = "grey70", lty = 1, lwd = 1)
> lines(taus[rang], duqf_SE[rang])
> abline(h = 0, lty = 2)
> plot(c(0,1), range(c(range_x, range_y)), xlim = c(0,1), type = "n",
+      xlab = expression(tau), ylab = "", cex.lab=0.75, main = "Composition")
> polygon(c(taus[rang],rev(taus[rang])),
+         c(u.duqf_CE[rang], rev(l.duqf_CE[rang])), density = -100, border = F,
+         col = "grey70", lty = 1, lwd = 1)
> lines(taus[rang], duqf_CE[rang])
> abline(h = 0, lty = 2)
```

Finally, we repeat the point and interval estimation using the duration regression method with the option `method = "cox"`. Despite of relying on a more restrictive model for the conditional distribution, the duration regression estimates in Figure 4 are similar to the logit regression estimates in Figure 3.

```
> coxres <- counterfactual(lwage~tenure+ttl_exp+grade,
+           group = union, treatment=TRUE, quantiles=taus,
+           method="cox", nreg=100, weightedboot = TRUE,
+           printdeco = FALSE, decomposition = TRUE, sepcore = TRUE,ncore=2)
cores used= 2
> duqf_SE   <- (coxres$resSE)[,1]
> l.duqf_SE <- (coxres$resSE)[,5]
```

```
> u.duqf_SE <- (coxres$resSE)[,6]
> duqf_CE   <- (coxres$resCE)[,1]
> l.duqf_CE <- (coxres$resCE)[,5]
> u.duqf_CE <- (coxres$resCE)[,6]
> duqf_TE   <- (coxres$resTE)[,1]
> l.duqf_TE <- (coxres$resTE)[,5]
> u.duqf_TE <- (coxres$resTE)[,6]
> range_x = min(c(min(l.duqf_SE[rang]), min(l.duqf_CE[rang]),
+                 min(l.duqf_TE[rang])))
> range_y = max(c(max(u.duqf_SE[rang]), max(u.duqf_CE[rang]),
+                 max(u.duqf_TE[rang])))

> par(mfrow=c(1,3))
> plot(c(0,1), range(c(range_x, range_y)), xlim = c(0,1), type = "n",
+      xlab = expression(tau), ylab = "Difference in Wages", cex.lab=0.75,
+      main = "Total")
> polygon(c(taus[rang],rev(taus[rang])),
+         c(u.duqf_TE[rang], rev(l.duqf_TE[rang])), density = -100, border = F,
+         col = "grey70", lty = 1, lwd = 1)
> lines(taus[rang], duqf_TE[rang])
> abline(h = 0, lty = 2)
> plot(c(0,1), range(c(range_x, range_y)), xlim = c(0,1), type = "n",
+      xlab = expression(tau), ylab = " ", cex.lab=0.75, main = "Structure")
> polygon(c(taus[rang],rev(taus[rang])),
+         c(u.duqf_SE[rang], rev(l.duqf_SE[rang])), density = -100, border = F,
+         col = "grey70", lty = 1, lwd = 1)
> lines(taus[rang], duqf_SE[rang])
> abline(h = 0, lty = 2)
> plot(c(0,1), range(c(range_x, range_y)), xlim = c(0,1), type = "n",
+      xlab = expression(tau), ylab = "", cex.lab=0.75, main = "Composition")
> polygon(c(taus[rang],rev(taus[rang])),
+         c(u.duqf_CE[rang], rev(l.duqf_CE[rang])), density = -100, border = F,
+         col = "grey70", lty = 1, lwd = 1)
> lines(taus[rang], duqf_CE[rang])
> abline(h = 0, lty = 2)
```

4. Acknowledgments

We wish to thank Lorenz Thomschke and an anonymous referee for insightful comments. We gratefully acknowledge research support from the NSF.

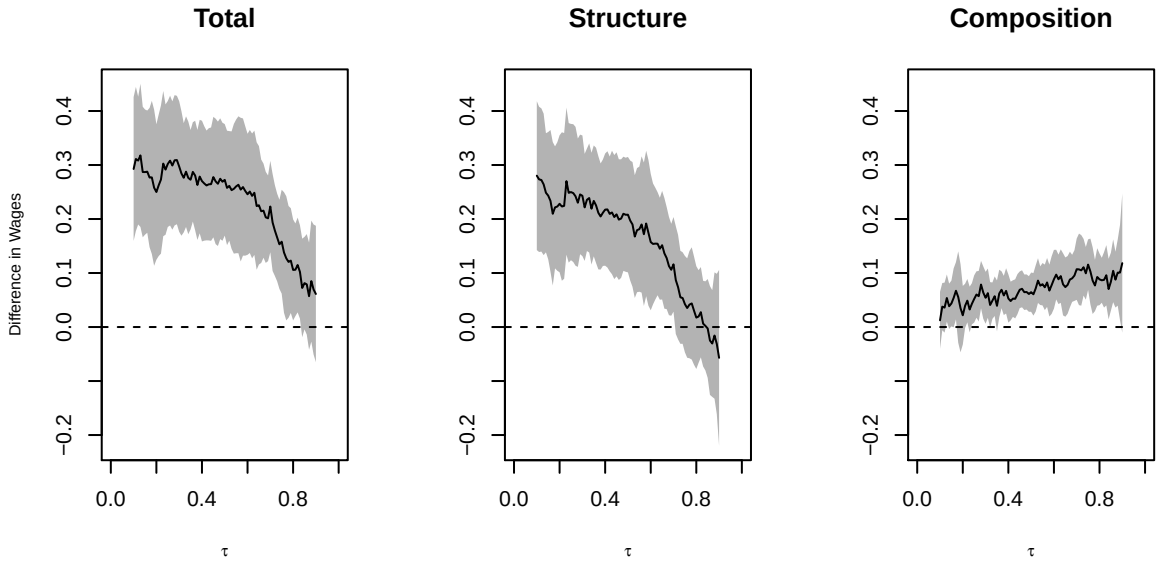

FIGURE 4. Wage decomposition with respect to union: duration regression estimates

REFERENCES

Blinder, A. (1973). Wage discrimination: reduced form and structural estimates. *Journal of Human resources*, 436–455.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013). Inference on counterfactual distributions. *Econometrica 81*(6), 2205–2268.

Chernozhukov, V. and H. Hong (2002). Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association 97*(459), 872–882.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 187–220.

Koenker, R. (2016). *quantreg: Quantile Regression*. R package version 5.21.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International economic review*, 693–709.

Stock, J. (1991). Nonparametric policy analysis: an application to estimating hazardous waste cleanup benefits. *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, 77–98.

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. version 2.38.