Marginal Unit Interpretation of Unconditional Quantile Regression and Recentered Influence Functions using Centered Regression

Fernando Rios-Avila
Levy Economics Institute
Bard College
Annandale-on-Hudson, NY, USA

John P. de New[1]
Melbourne Institute
University of Melbourne
Melbourne, VIC, Australia

May 24, 2022

Abstract – Unconditional quantile regressions, as introduced by Firpo, Fortin, and Lemieux (2009), are a special case of Recentered Influence Functions (RIF) Regressions that can be used to relate how *small* changes in the distributions of explanatory variables affect an unconditional statistic of interest. While there is general understanding with regards to the analysis and interpretation of changes in continuous variables, difficulties remain when interpreting changes in qualitative characteristics (dummies). On the one hand, the implicit inter-relationship among binary variables is usually ignored, and on the other hand, that standard RIF regressions only capture effects at the margins, not distributional treatment effects. This paper suggests the use of restricted least squares regression analysis based on Haisken-DeNew and Schmidt (1997), combined with the use of centered continuous variables, and re-scaling, to isolate the intercept cleanly as the distributional statistic of interest and more appropriately interpret the results of RIF-regressions in the presence of dummy variables. The author-written Stata add-on command creg.ado implements this methodology.

JEL Codes: C21, C23, D63
Keywords : Recentered Influence Functions, Unconditional Quantile Regression, Restricted Least Squares

[1] Corresponding author: Prof. Dr. John P. de New, Melbourne Institute, University of Melbourne, 111 Barry St, Carlton VIC 3053, Australia, johnhd@unimelb.edu.au. Contact for Dr. Fernando Rios-Avila, Levy Economics Institute, Bard College, Annandale-on-Hudson, NY, USA, friosavi@levy.org. We thank Austin Nichols at the Stata Oceania Feb 2022 workshop for pointing out Benn Jann's (2008) devcon.ado and its implications.

1. Introduction

Recentered Influence Functions (RIF) regression is a statistical tool introduced by Firpo, Fortin, and Lemieux (2009), for analyzing unconditional partial effects on unconditional quantiles in the framework of regression analysis. The methodology is based on the use of Influence functions (IF), which have been long used for analyzing the robustness of distributional statistics, and as a simplified strategy to estimate standard errors of those statistics (Cowell and Flachaire 2007; Cowell and Flachaire 2015; Deville 1999; Jann 2020), using a linearization strategy similar to the Delta Method (Cameron and Trivedi, 2009). While Firpo, Fortin, and Lemieux (2009) originally focused on the analysis of partial effects of explanatory variables on unconditional quantiles of the dependent variable, the principles behind this methodology have been extended to other distributional statistics (see Foster, Greer, Thorbecke 1984, 2010 for their FGT(a) inequality measures, Rios-Avila 2020, Jann 2020 and the references therein for further details).

In principle, unconditional quantile regressions (UQR), as introduced by Firpo, Fortin, and Lemieux (2009), aim to identify how *small location shifts* in the distribution of explanatory variables affect a statistic of interest. This is done by using the RIF for a particular distributional statistic as dependent variable, and by estimating average marginal effects on the RIF of all explanatory variables. In this framework, the idea of a *small location shifts* is better understood as a change in the unconditional mean of the explanatory variables, that is independent from changes in other variables.

While there is general agreement in the interpretation of *small location shifts* for the case of continuous variables in the framework of RIF-regressions, some difficulties remain for understanding this thought experiment when interpreting set of dummies that describe qualitative characteristics. In particular, as discussed elsewhere in the literature (Rothe 2010, Rios-Avila 2020 and Rios-Avila and Maroto 2021), RIF-regressions in general, and UQR in particular, *cannot* be used to identify distributional treatment effects.[2] This happens for three reasons: (a) standard RIF-regressions are only linear approximations of nonlinear functionals; (b) the implicit inter-relationship within binary variables sets are usually ignored; and (c) marginal effects are estimated assuming a 1 unit change in the independent variables (going in a discrete jump from 0 to 1 in the case of dummies), which is simply too large to be considered a *small location shift*. Instead, when dummy variables are used, the equivalent to the *small location shifts* to be considered, should be a change in the *percent-point proportion* of observations in a particular group along with a decline in similar magnitude of some other groups.

In this paper, we suggest an alternative estimation of the marginal effects for RIF-OLS, that may provide a more intuitive framework for the interpretation of RIF regressions not only in the

[2] RIF regressions can be used to identify treatment effects *only* if the RIF function for a statistic does not depend on the sample used for its construction, i.e., $RIF(y, v(.)) = RIF(y, v(.|X))$. This is the case for the mean, and for FGT(a) type of poverty indicators, described in Foster, Greer, Thorbecke (1984, 2010).

presence of dummy variables, but also when continuous variables are considered. Specifically, we propose using post-estimation restricted least squares (RLS) regression analysis, as described in Haisken-DeNew and Schmidt (1997), combined with centered continuous variables in the regression analysis, to provide a regression output that should be easier to understand and interpret for applied researchers. We denote this a "centered regression".

This approach has crucial benefits for RIF coefficient interpretation: (1) it allows us to capture effects associated with dummy variables that do not depend on arbitrary omitted base-categories; instead, the coefficients of *all* dummy categories are compared to a pseudo-unit with "average" characteristics; (2) if we combine the restricted OLS with the use of centered continuous variables, which has no impact on the estimated slope coefficients, the strategy allows us to identify a "clean" intercept that now represents the unconditional mean of the dependent variable. In other words, in the framework of RIF-regressions, the intercept will capture the *unconditional* statistic of interest for the observed population, whereas the (re-scaled) slopes will provide a measure of the direction one expects the statistic to change if the distribution of that characteristic changes. This provides a more intuitive interpretation of the regression coefficient interpretation, as it follows *exactly* the definition of the RIF being a function of the functional and the Influence Function. Finally (3), by exposing the constant cleanly, elasticities can be calculated for each explanatory variable as the coefficient divided by the constant, using the correct non-linearly calculated standard errors.

The rest of the paper is as follows: Section 2 outlines the methodology; section 3 an empirical application and section 4, the conclusions. We include a set of appendices illustrating the linear transformations and the underlying technical details.

2. Methodology

2.1. Restricted Linear Regression: Interpreting the intercept

Consider the following linear regression model, in which x is a continuous variable:

$$y_i = b_0 + b_1 x_i + e_i \qquad (1)$$
$$E(y_i|x,z) = b_0 + b_1 x_i$$

If this model were to be estimated, it is well known that, under standard linear regression (LR) assumptions, the coefficients b_1 capture the change in the conditional mean of y due to a *1 unit* change in x. It is also well known that the intercept b_0 can be interpreted as the average of y_i, conditional on x_i to be equal to zero.

As discussed in many introductory econometric books (see for example Wooldridge (2016), p30), in many setups, the intercept may not be interpretable in the economic sense. This happens

because often no single real observation is represented by imposing characteristics to have a value of zero.[3]

This interpretation of the intercept changes further, and potentially becomes even more nebulous, when we introduce many sets of dummy variables, each representing mutually exclusive qualitative characteristics of individuals. Consider, for example, that we aim to introduce the single variable D, which has J categories, in a model specification:

$$y_i = b_0 + b_1 x_i + \sum_{j=1}^{J} \delta_j 1(D_i = j) + e_i \tag{2}$$

We know well that equation (2) cannot be estimated directly because of the so-called dummy variable trap. This is caused because of a perfect linear combination of the dummies colliding with the intercept. The general suggestion is to impose the condition that $\delta_j = 0$ for any one arbitrary category (henceforth the base category).[4]

$$y_i = \gamma_0 + b_1 x_i + \sum_{j=2}^{J} \delta_j 1(D_i = j) + e_i \tag{3}$$

Under this specification, the intercept could now be interpreted as the expected value of the outcome when $x_i = 0$, for the **base** category, and the δ_j of other categories represent deviations from the base category, assuming we are comparing units with otherwise same characteristics. Although this approach for the estimation of LR is standard, it has the disadvantage that a useful interpretation of the intercept is lost, and that coefficients related to other dummies within the same set of categories will vary depending on the (arbitrary) selection of the excluded categories.

One option described in some econometric textbooks (see for example Wooldridge 2016, p178) to facilitate the interpretation of marginal effects in the presence of interactions and squared parameters is the use of centered variables. If this strategy is combined with the approach proposed in Haisken-DeNew and Schmidt (1997), it is possible to obtain an intercept that has a *direct and intuitive interpretation* from an economic point of view.

Consider again equation (2), and assume all coefficients δ_i can be identified. If we obtain the unconditional expectation of all variables in this model, we could obtain the following:

$$\bar{y} = b_0 + b_1 \bar{x} + \sum_{j=1}^{J} \delta_j \bar{D}_j \tag{4}$$

$$b_0 = \bar{y} - b_1 \bar{x} - \sum_{j=1}^{J} \delta_j \bar{D}_j \tag{5}$$

[3] In a typical Mincerian wage regression specification, for example, the constant represents someone with no years of education, no experience and/or zero years of age.

[4] While this restriction is standard to avoid the variable dummy trap, other restrictions are also possible.

where $\bar{y}$ and $\bar{x}$ are the unconditional means of y and x, and $\overline{D}_j$ is the proportion of observations that belong to group $D_i = j$. If we substitute (5) in (2), we have:

$$y_i = \bar{y} - b_1\bar{x} - \sum_{j=1}^{J} \delta_j \overline{D}_j + b_1 x_i + \sum_{j=1}^{J} \delta_j 1(D_i = j) + e_i \tag{6}$$

$$y_i = \bar{y} + b_1(x_i - \bar{x}) + \sum_{j=1}^{J} \delta_j \left(1(D_i = j) - \overline{D}_j\right) + e_i \tag{7}$$

Finally, if we impose the identifying restriction $\sum_{j=1}^{J} \delta_j \overline{D}_j = 0$, as outlined in Haisken-DeNew and Schmidt (1997), and assume $\bar{y}$ is an additional parameter that needs estimation, then equation (7) becomes:[5]

$$y_i = a_0 + b_1(x_i - \bar{x}) + \sum_{j=1}^{J} \delta_j 1(D_i = j) + e_i \tag{8}$$

$$s.t.\ E(y_i) = a_0 + b_1 E(x_i - \bar{x}) + \sum_{j=1}^{J} E\left(\delta_j 1(D_i = j)\right) + E(e_i) = a_0$$

In this specification the intercept a_0 has a clear and useful interpretation. It represents the unconditional mean of the dependent variable (with an estimated standard error for hypothesis testing). From the micro perspective, it represents the expected outcome for a synthetic individual with average continuous characteristics ($x_i = \bar{x}$), but who is also average in terms of qualitative characteristics.

While nothing changes in terms of the interpretation of the slope coefficient b_1, change in the conditional mean of y given a 1 unit change in x, having a "clean" intercept that represents the unconditional mean of y helps to ascertain the true magnitude of all marginal effects, by obtaining elasticity like parameters (i.e., $\beta_1 = b_1/a_0$). In the case of dummies, the coefficients δ_j should now be interpreted as how much higher (lower) average outcome of group $D = j$ is compared to the *unconditional mean*, or that *synthetic comparison person*, regardless of arbitrary (and potentially misleading[6]) choice of reference dummy chosen, and after controlling for other characteristics.

2.2. Re-centered Influence functions (RIF) and RIF-Regressions

2.2.1. Re-centered Influence functions (RIF): Definition

Influence functions (IF) are analytical tools that can be used to analyze the robustness of distributional statistics to small changes in data (Cowell and Flachaire 2007), or to estimate asymptotic variances of complex statistics (Cowell and Flachaire 2015; Deville 1999). In the framework of

[5] Details on the identification of the coefficients and standard errors when this condition is imposed can be found in Appendix A. For further details in terms of the estimation approach, and estimation of the corresponding correct standard errors can be found in in Haisken-DeNew and Schmidt (1997).

[6] Selecting an unusually large positive or negative category as the reference makes all other coefficients appear arbitrarily very large and significant.

regression analysis, when IF's are used as dependent variables, they can be used to analyze how (small) location changes in the distribution of explanatory variables will affect the distribution of the dependent variable, when those changes are measured using a particular distributional statistic of interest, like the unconditional quantile (Firpo, Fortin and Lemieux 2009).

In principle, the IF measures the rate of change of a distributional statistic caused by a change in the distribution of the underlying variable of interest, when this distributional change is *infinitesimally* small. This idea is what lies behind the Gateaux derivative, which is a generalization of the directional derivative of a functional.

To formalize this concept, consider a variable y which has a cumulative distribution F_y, and a distributional statistic $v(.)$. Also consider the function $H_{y_i}(y)$, and cumulative function G_y which are defined as:

$$H_{y_i}(y) = 1(y \geq y_i) \tag{9}$$

$$G_y = (1-\varepsilon)F_y + \varepsilon H_{y_i}(y) \tag{10}$$

Function G_y is also known as a *contaminated function*, because it represents the distribution of y that would be observed after an additional observation with income y_i would be introduced to the sample. Given this, the IF is defined as:

$$IF\big(y_i, v(F_Y)\big) = lim_{\varepsilon \to 0} \frac{v\big(G_y\big) - v\big(F_y\big)}{\varepsilon} = lim_{\varepsilon \to 0} \frac{v\Big((1-\varepsilon)F_y + \varepsilon H_{y_i}(y)\Big) - v\big(F_y\big)}{\varepsilon} \tag{11}$$

This function indicates the rate of change in the distributional statistic v, when the distribution "moves" towards H_{y_i}.

Rather than using the IF function, Firpo, Fortin and Lemieux (2009) propose using the *recentered* influence function (RIF), which is equivalent to the first two terms of the von Mises (1947) *linear approximation* of the distributional statistic v:

$$RIF\big(y_i, v(F_Y)\big) = v\big(F_y\big) + IF\big(y_i, v(F_Y)\big)$$

This statistic has two crucial properties:

$$E\left(RIF\big(y_i, v(F_Y)\big)\right) = \int RIF\big(y_i, v(F_Y)\big) f_Y(y) dy = v(F_Y) \tag{12}$$

$$\sigma^2_{v_F} = N^{-1} \int IF\big(y_i, v(F_Y)\big)^2 f_Y(y) dy = N^{-1} Var\left(RIF\big(y_i, v(F_Y)\big)\right) \tag{13}$$

Equation (12) indicates that the unconditional expected value of the RIF is the *original* statistic of interest v, whereas (13) states that the asymptotic variance of v ($\sigma^2_{v_F}$) is equal the variance of the RIF's, divided by the number of observations (N).

2.2.2. RIF-Regressions and Unconditional Partial Effects (UPE)

RIF-Regression was introduced by Firpo, Fortin and Lemieux(2009), as a computationally simple strategy to analyze effects of changes in the distribution of explanatory characteristics on

unconditional quantiles. Since then, the strategy has been extended to analyze effects on other distributional statistics as well. Paraphrasing the original article, the estimation of RIF regressions can be described as follows: assume that the joint probability density function (pdf) between the dependent variable Y, and all the exogenous explanatory variables X is defined by $dF_{Y,X}(y,x) = f_{Y,X}(y,x)$. This joint distribution determines all linear and non-linear relationships between these variables. Using this, the unconditional cumulative distribution function (CDF), and the probability density function (PDF) for Y can be defined as follows:

$$F_Y(y) = \int F_{Y,X}(y,x)dx = \int F_{Y|X}(y|x)f_X(x)dx \tag{14}$$

$$f_Y(y) = \int f_{Y,X}(y,x)dx = \int f_{Y|X}(y|x)f_X(x)dx \tag{15}$$

which indicate that if the conditional distribution $f_{Y|X}(y|x)$ is fixed, changes in the unconditional distribution of the outcome Y can be observed, if there are exogenous changes in the distribution of characteristics X (i.e. f_X).

If we combine this with (12), the unconditional statistic $v(F_Y)$ can be written as:

$$v(F_y) = \int E\big(RIF\big(y_i, v(F_Y)\big)|X = x\big)f_X(x)\,dx. \tag{16}$$

This equation indicates that the statistic of interest $v(F_y)$ can be written as an average of the conditional mean $E\big(RIF\big(y_i, v(F_Y)\big)|X = x\big)$ weighted by f_X. In combination with equation (15), this suggests that if there is a small change in the distribution (say from $f_X \rightarrow f_X'$), this will generate a change in the unconditional distribution of Y (from $F_Y \rightarrow F_y'$), what can be captured by a change in the unconditional statistic v from $v(F_Y) \rightarrow v(F_y')$.

If the change Δf_X is *small*, Firpo, Fortin and Lemieux (2009) suggest that $v(F_y')$ can be *approximated* as follows:

$$v(F_y') \cong v'(F_y) = \int E\big(RIF\big(y_i, v(F_Y)\big)|X = x\big)f_X'(x)\,dx \tag{17}$$

which assumes that changes in F_Y will be *sufficiently small* that the $RIF\big(y_i, v(F_Y)\big)$ does not need to be updated. However, if Δf_X is large, there will have a significant effect on F_y and $RIF\big(y_i, v(F_Y)\big)$ *will no longer be a good approximation* for $RIF\big(y_i, v(F_Y')\big)$. This is a crucial point.

Based on equation (17), Firpo, Fortin and Lemieux (2009) indicate that the simplest approach to estimate RIF regressions is to focus on modeling the conditional mean $E\big(RIF\big(y_i, v(F_Y)\big)|X = x\big)$. Specifically, they suggest using ordinary least squares (OLS) where one uses the estimated RIF as the dependent variable:[7]

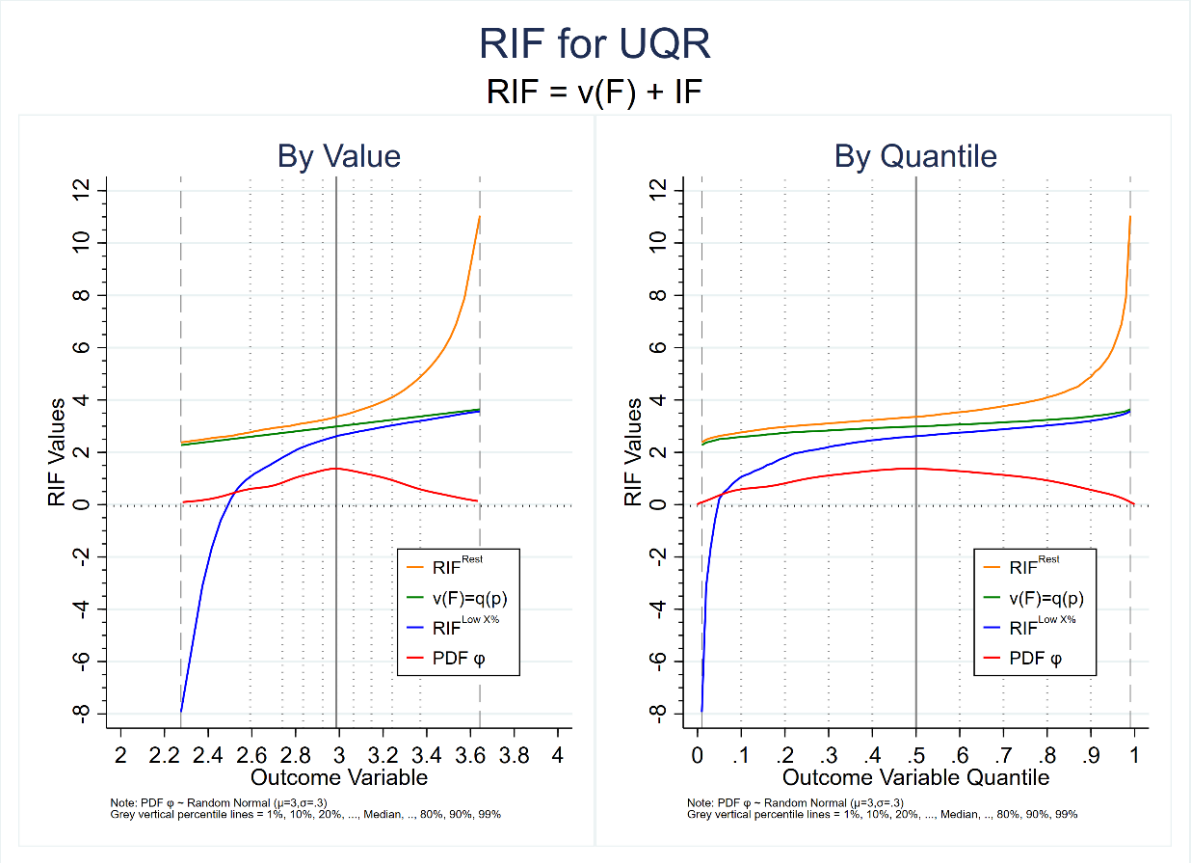
[7] It should be emphasized that one could use other methods to estimate the conditional RIF mean. For example, for the estimation of unconditional quantile regressions, Firpo, Fortin and Lemieux (2009) suggests using a binomial model, given the structure of the quantile Influence Function.

$$RIF\left(y_i, v(F_y)\right) = b_0 + \sum_{k=1}^{K} b_k x_{i,k} + u_i \rightarrow E\left(RIF\left(y_i, v(F_y)\right)\middle| X\right) = b_0 + \sum_{k=1}^{K} b_k x_{i,k} \qquad (18)$$

thus, modeling the conditional mean of the RIF as a linear function of x.

To visualize better the components and structure of a RIF for the case of unconditional quantiles, we plot in Figure 1 the RIF function for all percentiles of a variable randomly distributed variable $Y \sim N(\mu = 3, \sigma = 0.3)$ with a corresponding probability density function shown in red. The vertical grey lines bound the density at the (left-most) 1 percentile, the 10th percentile, the 20th percentile, … the median, the 60th percentile, … 90th percentile and finally the (right-most) 99th percentile. At each percentile, there are 2 (and only 2) *distinct* values for the many observations of the RIF, which finally on average produce the functional $q_y(p)$ or $v(F_Y)$ as $E(IF(y_i, v(F_Y))) = 0$. The lower of the two values is given by the blue line, which for any $q_y(p)$ percentile, takes on the value RIF_{Low} and the lowest $q_y(p)\%$ are assigned this value. The remaining (100- $q_y(p)$)% are assigned the higher of the two values, RIF_{High}, given by the orange line. Thus, the weighted sum or the two RIF values is identically equal to the $q_y(p)$ or $v(F_Y)$. From our derivation above, the green line is the constant, a_0 in (8), and the remaining coefficients from the centered continuous variables and transformed dummy variable sets, comprise the $IF(y_i, v(F_Y))$, or simultaneously the orange and blue lines. We show the graphic of the outcome variable *Y* with itself ("By Value") on the X-axis as well as the percentile ("By Quantile") in a corresponding graphic adjacent.

Figure 1: Graphical Depiction of the UQR RIF for an Example Distribution



While the estimation of RIF-regressions using OLS does not change compared to the standard regression analysis, the interpretation of the coefficients must be adjusted. In standard OLS analysis, the interpretation of slope coefficients b_k relate to how $E(y|x)$ will change if x_k increases in 1 unit, everything else held constant. For the interpretation of RIF-regressions, however, one must relate to how a change in the distribution of x will translate into a change in the unconditional statistic v.

To emphasize this thought experiment, Rios-Avila (2020) suggests that the correct interpretation of unconditional partial effects using RIF regressions should follow a two-step approach: First, the model should be re-stated in terms of the unconditional mean of all explanatory variables; and second, calculate the partial effects on the statistic of interest v with respect to changes in all unconditional means:

$$E\Big(RIF\Big(y_i, v(F_y)\Big)\Big) = v(F_y) = b_0 + \sum_{k=1}^{K} b_k \bar{x}_k \tag{19}$$

$$\frac{\partial v(F_y)}{\partial \bar{x}_k} = b_k$$

Using the properties of RIFs, this approach provides an intuitive understanding of the relationship between the statistic v and potential changes in the distribution of x. Specifically, it suggests that if the distribution of x_k changes, such that its unconditional mean increases in one unit, we would expect the unconditional statistic v to increase in b_k units. Firpo, Fortin and Lemieux (2009)

denote this as the unconditional partial effect (UPE) caused by a *small* location shift in the distribution of x. It also shows that, if there are no interactions nor higher order polynomials, the only change we can induce in the distribution of x is through its unconditional mean $\bar{x}$.[8]

While the above interpretation of RIF-regression coefficients is standard in regression analysis, one needs to be precise when interpreting the effects of categorical variables. Consider again the specification stated in equation (18), and assume that instead of using a set of continuous variables ($X's$) as controls, we use a set of mutually exclusive categorical variables $D_k, \forall\ k = 1,2, \dots, J$, such that $D_k = 1$ if observation i belongs to group k.

$$E\left(RIF\left(y_i, v(F_y)\right)\middle|D\right) = b_0 + \sum_{k=2}^{K} \delta_k D_k \tag{20}$$

As described in section 2.1, in a standard linear model, the parameters δ_k are interpreted as differences in average outcome compared to an excluded group (here when $D_1 = 1$). In the RIF regression framework, *this experiment is no longer valid*. For example, making this type of interpretation suggest one is comparing a situation where everyone belongs to group 2, compared to a situation in which everyone belongs to group 1, e.g. imagining what would happen to some income distributional statistic if counterfactually all wage earners became male (or female) is hardly an infinitesimally small change. This is clearly a *large* change in the distribution of characteristics f_X that will have a nontrivial effect the unconditional distribution of the outcome f_Y. Because of this, using equation (17) will *no longer be a valid approximation* to estimate the change in the distributional statistic v.[9]

This problem can also be observed by using the two-step approach advocated by Rios-Avila (2020). Specifically, if we obtain unconditional means for equation (20) we have:

$$v(F_y) = b_0 + \sum_{k=2}^{K} \delta_k \bar{D}_k \tag{21}$$

which indicates that the distributional statistic v is a function of the proportion of individuals in each sub-group. Thus, when analyzing the effects δ_k, one must consider a change in the proportion of individuals belonging to a particular group, say a 1%-point increase, which is coupled with an identical decrease of individuals belonging to the excluded group $\bar{D}_1$.[10]

2.2.3. UPE: Centered variables, polynomials and interactions

[8] A more formal description of the process consists in estimating the average marginal effects. However, when using quadratic terms and simple interactions, average marginal effects and marginal effects at the means are identical.

[9] To address this problem, Firpo, Fortin and Lemieux (2018) suggests using reweighted regressions when combining RIF regressions with Kitakawa-Oaxaca-Blinder decomposition.

[10] Note that this suggestion is no different from the approach used for the interpretation of coefficients in a linear-log model.

As described in section 2.1, using centered continuous variables and using restricted linear regressions for categorical variables, allows us to estimate an intercept that represents the unconditional mean of the dependent variable. From the RIF regression point of view, this is informative, because the unconditional mean will represent the distributional statistic of interest v, whereas the rest of the coefficients would identify the direction of how the statistic would change if the distribution of a particular characteristic changes (the IF).

Consider a RIF-regression, where all control variables are centered at their unconditional means $\bar{x}_k^0$. We use the superscript 0, to denote that those are the average before any change took place.

$$RIF\left(y_i, v(F_y)\right) = a_0 + \sum_{k=1}^{K} b_k\left(x_{i,k} - \bar{x}_k^0\right) + u_i \tag{22}$$

$$v(F_y) = a_0 + \sum_{k=1}^{K} b_k\left(\bar{x}_k - \bar{x}_k^0\right) = a_0 \tag{23}$$

Here, the intercept a_0 is equal to the statistic of interest $v(F_y)$, and the partial effects of $\bar{x}_k$ can be calculated as:

$$\frac{\partial v(F_y)}{\partial \bar{x}_k} = b_k \tag{24}$$

If one is interested in analyzing further changes in the distribution, it is also possible to add quadratic[11] terms and single interactions of the centered variables, to capture effects of changes in the variance x or the covariance across variables on the distributional statistic v.[12] However, to maintain the interpretation of the intercept, these terms need to be re-centered. Define $\tilde{x}_k = x_k - \bar{x}_k^0$, and define the RIF regression, and its unconditional mean as follows:

$$RIF\left(y, v(F_y)\right) = a_0 + \sum_{j=1}^{2} b_j \tilde{x}_j + \sum_{j=1}^{2} c_j\left[\tilde{x}_i^2 - E\left(\tilde{x}_i^2\right)\right] + d_1[\tilde{x}_1\tilde{x}_2 - E(\tilde{x}_1\tilde{x}_2)] + u_i \tag{25}$$

$$v(F_y) = a_0 + \sum_{j=1}^{2} b_j\left(\bar{x}_j - \bar{x}_j^0\right) + \sum_{j=1}^{2} c_j\left(\sigma_{x_j}^2 - \sigma_{x_j}^{2,0}\right) + d_1\left(\sigma_{x_1x_2} - \sigma_{x_1x_2}^0\right)$$

where a_0 is still the statistic of interest v, and the partial effects with respect to the mean, variance and covariance are given by:

$$\frac{\partial v(F_y)}{\partial \bar{x}_j} = b_j; \frac{\partial v(F_y)}{\partial \sigma_{x_j}^2} = c_j; \frac{\partial v(F_y)}{\partial \sigma_{x_1x_2}} = d_1 \tag{26}$$

2.3. UPE: Dummies and Restricted Linear Regression

[11] Although the use of quadratic terms is not explicitly implemented in our Stata ado, it can be manually implemented by simply adding centred variables in the model specification.

[12] While one can also add higher order polynomials and interactions, the equivalence between average marginal effects and marginal effects at the mean disappears, and one has to estimate average marginal effects to identify changes of the mean.

The interpretation of dummy variables in the context of RIF regressions has traditionally been less clean. Rios-Avila (2020) suggests that coefficients should be multiplied by 0.1 or 0.01, a 10%-point or 1%-point change, to emphasize the thought experiment of a *small* change in the proportion of individuals belonging to a particular group. When doing so, all effects are measured assuming that changes in the group of interested are counterbalanced by changes in the base/omitted group. Thus unconditional changes are sensitive to the baseline/reference group.

As described in section 2.1, in the standard linear regression approach, one can impose the identifying assumption that $\sum_{j=1}^{J} \delta_j \overline{D}_j = 0$ to estimate coefficients for all dummies, avoiding the dummy variable trap. In that case, the coefficients δ_k represent deviations in the conditional mean for observations in group j with respect to the unconditional mean. In the framework of RIF regressions, the estimation of partial effects requires additional attention: the thought experiment of a change in the proportion of observations in a particular group is counterbalance by a similar change in all other groups.

Consider a model similar to the one described for equation (20), but where all coefficients δ_k are identified by using the restricted regression approach. The unconditional expectations of the model could be written as follows:

$$v(F_y) = a_0 + \delta_k \overline{D}_k + \sum_{j=1 \& j \neq k}^{J} \delta_j \overline{D}_j \tag{27}$$

To estimate the effect of a change in the proportion of group k, we need to re-express the proportion of individuals in group $j = 1, \dots, J$ ($\overline{D}_j$) in terms of the original distribution and $\overline{D}_k$. Specifically:

$$\overline{D}_j = \frac{(1 - \overline{D}_k)}{(1 - \overline{D}_k^0)} \overline{D}_j^0 \tag{28}$$

where $\overline{D}_j^0$ denotes the original share of individuals in group j. Substituting (28) in (27) we have:

$$v(F_y) = a_0 + \delta_k \overline{D}_k + \sum_{j=1 \& j \neq k}^{J} \delta_j \frac{(1 - \overline{D}_k)}{(1 - \overline{D}_k^0)} \overline{D}_j^0 \tag{29}$$

Using this, we estimate the UPE of a change in the proportion of individuals belonging to group $\overline{D}_k$ as follows:

$$\frac{\partial v(F_y)}{\partial \overline{D}_k} = \delta_k - \sum_{j=1 \& j \neq k}^{J} \delta_j \frac{\overline{D}_j^0}{(1 - \overline{D}_k^0)} = \frac{\delta_k}{1 - \overline{D}_k^0} \tag{30}$$

This expression indicates that the effect of a change in $\overline{D}_k$ will depend on the proportion of units that are not part of group k, and that the proportion of individuals belonging to group k cannot increase more than $(1 - \overline{D}_k^0)$%. Furthermore, the restriction imposed by equation (28) states that the thought experiment of an increase in the proportion of individuals in group k will be counterbalanced by a decline in the proportion of individuals belonging to all other groups. Following previous advice, the

effect defined by equation (30) must be re-scaled, considering a small change in the distribution, for example, a 1%-point increase in the proportion of individuals belonging to group k.

3. Empirical Application

Here we provide two empirical applications showing the implementation and interpretation of RIF regressions based on centered-linear regressions. The applications use publicly available data from: (1) Jann (2008) and (2) Firpo, Fortin, Lemieux (2009). It should be noted that there is similar previous literature addressing some of these issues. Jann (2008) concentrated on implementing the Oaxaca decomposition, and implemented a form of Haisken-DeNew and Schmidt (1997) to remove the arbitrariness of dummy reference for decomposition purposes and Firpo, Fortin, Lemieux (2009) introduced RIF regression to begin with.

We are the first to use the restricted last squares linear combinations outlined in Haisken-DeNew and Schmidt (1997) in the application of RIF regression, implementing multiple factor variables, interactions of factor variables, and direct elasticity calculation with correct standard errors. Through the automatic centering of the explanatory variables, we maintain the constant as the estimated unconditional mean (with appropriate standard error), which we then use to calculate semi-elasticities for each explanatory variable, using non-linear combinations of the coefficients divided by the constant (and maintaining correct standard errors using Stata's nlcom command).

Application 1

We use the publicly available example Stata dataset oaxaca.dta[13] containing Swiss Household Panel (SHP) microdata based off analyses in Jann (2008). He implements a Oaxaca (1973) analysis, to illustrate the decomposition methodology. We also use the publicly available RIF estimators[14] combined with a post-estimation version of centered linear regression[15]. We implement a post-estimation command creg.ado for Stata with many options. Take for example the standard linear regression:

$$LnWage_i = \alpha + \beta_j X_i + \epsilon_i \tag{31}$$

in which $LnWage_i$ is the dependent variable, α the intercept, X_i the explanatory variables which include continuous and dummy variables, and ϵ_i the normally distributed error term. This can be mean-estimated using OLS in which case the RIF of $LnWage_i$ is simply the variable $LnWage_i$ itself. However we can estimate this for any unconditional quantile using the RIFs outline above:

[13] See http://fmwww.bc.edu/repec/bocode/o/oaxaca.dta and in Stata "net install oaxaca".
[14] See Rios-Avila (2021) for his estimator rifhdreg.ado on RIF regressions using Stata.
[15] See Haisken-DeNew and Schmidt (1997) and the Stata procedure fvhds97.ado.

$$RIF\left(y_i, q_y(p)\right) = q_y(p) + \frac{p + 1(y \leq q_y(p))}{f_y(q_y(p))} \tag{32}$$

$$RIF\left(y_i, q_y(p)\right) = \alpha + \beta_j X_i + \epsilon_i \tag{33}$$

We immediately see that the $q_y(p)$ in (32) aligns to the α in (33). The X_i the explanatory variables comprise dummy variable sets: *female*, *married*, *kids6*, *kids714*, and *isco*. The continuous variable *educ* has been made into a dummy set of discrete years of education. Continuous variables include *exper* in years. The continuous variables and dummy variables are transformed separately as outlined above.

Regression results are displayed in Table 1. Log wages are regressed using RIF regression (50th percentile) on marital status (married:0/1), 1-digit ISCO occupational status (1-9) and interactions of gender dummy (female:0/1), number of children up to age 6 (kids6:0-4) and number of children 7 through 14 (kids714:0-4) and educational status (educ: 5-18 years), exper (0-58 years) and tenure (0-44.83 years). All regressions use the Jann (2008) variable "wt" for weighting in the regression. Column (1) displays the unconditional mean of log wages using a simple OLS regression of log wages on a constant, which is 3.384. Column (2) displays the full UQR(50) regression with interactions. Of course, these parameter estimates are sensitive to the choice of the dummy reference(s). Column (3) displays the same marginal effects as calculated by the Stata command "margins". Column (4) displays the marginal effects, collapsing any interactions into their main marginal effects, as implemented in our creg.ado and produces identical estimate results to that of (3). After centering in Column (4), one can clearly see that the coefficient point-estimates for the constant in (1) and (4) are identical, albeit with different standard errors due to the presence of additional explanatory variables in (4). Column (3) and (4) are identical except for the missing information of the constant in (4), confirming correct calculation.

After all marginal effects have been calculated, column (5) performs the Haisken-DeNew and Schmidt (1997) procedure of taking deviations from a weighted average for all dummy variable sets (linear combinations or RLS), and centering any continuous variables, should they be present. Each dummy variable is interpreted as movements from the weighted average to only that dummy, i.e. from the average to only women or only men. The choice of any and all dummy reference categories in the original regressions no longer plays any role, as all coefficients are now deviations from the weighted average. Column (6) transforms the interpretation of all dummies to be interpreted as a 1%-point increase from the average, which effectively multiplies dummy coefficients in (5) by 1/(1-DummyShare)*100, for each dummy variable and respective sample dummy share respectively. Column (7) takes all non-Constant explanatory variable coefficients and divides them by the estimated constant a_0 using non-linear combinations arriving at the corresponding elasticity for each explanatory variable. This takes the unconditional partial effects (UPEs) and relativizes them by the magnitude of the dependent variable, the unconditional mean (or in this case the exposed constant 3.3842112 itself).

In Column (6) the constant corresponds to the RIF component $v(F_y)$ and the remaining coefficients correspond to the Influence Function or $IF(y_i, v(F_Y))$. We explicitly use 7 digits of precision in the tables to make the calculations clear, given the required infinitesimally small changes. See Appendix 1 for further details on the matrix algebra of RLS and further transformations.

For example, the 1.female (female=1) coefficient in (2) indicates the association of being female (however with education being 5 years, due to the interaction of gender and age) being -0.2227. In column (3) the 1.female coefficient represents the *marginal* effect of being female (taking into account all age categories) as compared to being male (over all age categories) at -0.1101. We use the Stata margins command for this. However, by definition, we cannot concurrently get an estimate of the constant with Stata's margins command. We now go to (4) and calculate the marginal effects by hand and also adjust the constant to obtain a complete set of point estimates and corresponding standard errors. We take this marginal effect and in (5), create deviations from the weighted average, and show that for females compared to the weighted average, they experience -0.0596, and males experience 0.0504. The estimates are only randomly similar in magnitude but with opposite sign, only because coincidentally the shares of males and females are similar. We now are interested in calculating a point estimate for a 1%-point increase in the share of females from the average in (6), i.e. from 51% in the population to 52%. Because now we are taking 1%-point more female-ness, we have a coefficient of -0.0011009. The population becoming 1%-point more female is associated with a median wage being 0.0011009 lower (because of negative coefficient). However, we know nothing about the relative size of this coefficient, as compared to the outcome variable (the wage). Note the complete symmetry in point estimate and standard error to that of going 1%-point more male. The male coefficient is +0.0011009 with the identical standard error.

In column (7) we calculate semi-elasticities by taking the female coefficient -0.0011009 in (6) and dividing it by the estimated constant in (1) at 3.3842112 and arrive at 0.0003253, giving us the *relative* magnitude (or elasticity) of the association of 1.female to median wages. This allows us (a) to compare much more directly coefficients between time periods for the same variable (say the Swiss Household Panel but for different years), and (b) coefficients between datasets for the same variable (say the European Community Household Panel, with standardized variables, but very different wage levels between countries).

The key to making this work in (7) is having an estimate of the constant with correct standard error in the same estimate vector and variance-covariance matrix of the other parameter estimates, which Stata cannot deliver in (3). For the continuous variables *exper* and *tenure*, the coefficients are unaffected until column (7) when they are divided by the coefficient on the constant. In general, the coefficient of a centered or non-centered continuous variable is identical; the only difference is in the point estimate of the associated constant. Only when all explanatory variables continuous and discrete

are centered does the constant provide the unconditional mean of the dependent variable, and allow interpretation of all categorical explanatory variables as being deviations from the (weighted) average. One can see that the magnitude of the unconditional partial effect from the naïve regression in (2) at -0.2227518 going *from all men to all women* is nothing like the *1%-point increase* in females of -0.0011009, appropriate to the RIF framework, and relative to the outcome variable (the elasticity), the partial effect is much smaller yet at only -0.0003253.

Application 2

We use the publicly available dataset men8385.dta from Firpo, Fortin, Lemieux (2009)[16] for men of the Merged Outgoing Rotation Group of the Current Population Survey files 1983-1985 and estimate, as an arbitrary example, unconditional quantile regressions for the 50th percentile (median). This data contains log wages, union coverage status (0/1), nonwhite (0/1), married (0/1), education in 6 dummy categories (0/5), experience in 9 dummy categories and as a continuous variable. Regression results are displayed in Table 2.

For this simple example we regress log wages on marital status and interactions between union coverage and the 6 education categories using unconditional quantile analysis for the 50th quantile. We follow the same structure as Application 1. Here the exposed constant is 1.8036749. Examining the dummy variable for union coverage (interacted with education, edc) 1.cov, in column (2) we get a coefficient of 0.4331530, but a marginal effect of 0.3327530 in in columns (3/4), but when expressed as deviations from a weighted average is only 0.2452645 in column (5). This partial effect is substantially smaller at 0.0033275 for a 1%-point increase in union coverage, and seen relative to the median wage in column (7), only 0.0018449. Again, the RIF framework deals only with infinitesimally small changes, along the lines of columns (6) or (7).

4. Conclusions

RIF regressions, as outlined in Firpo, Fortin, and Lemieux (2009), are powerful tools for distributional analysis, that allow the researcher to identify relationships between the distribution of the

[16] Citing FFL (2009) readme.txt file "The data file is: men8385.dta -- STATA10 data file containing an extract of the following variables from the Merged Outgoing Rotation Group of the Current Population Survey of 1983, 1984 and 1985. The file contains 266956 observations on males with 17 variables whose definition is given by the variable labels. More detail about the data selection and recoding (e.g. top coding, wage deflator, etc.) is found in Lemieux (2006)."

dependent and independent variables. Despite the appeal and wide use of this strategy in applied research, the estimations have until now been often misinterpreted, specifically when it comes to analyzing categorical variables. Due to the linear approximations of highly non-linear functions inherent in all RIF regressions, they fundamentally estimate the relationship between *infinitesimally small changes* of explanatory variables on the transformed outcome variable, the RIF. Standard dummy variable counterfactual interpretations, i.e. "all men" compared to "all women", are not applicable in this model and must be replaced with marginal or incremental interpretations, such as a 1%-point increase in the share of men or a 1%-point increase in the share of women.

In this paper, we suggest that using centered covariates and constrained linear regressions (Haisken-DeNew, and Schmidt, 1997) which provides a more intuitive and appropriate interpretation of unconditional partial effects on distributional statistics, as compared to how it was originally intended in Firpo, et al (2009). By centering all covariates, the constant itself becomes an interpretable estimate, *the statistic of interest itself*, while other coefficients represent the influence of the covariates on the selected statistic (unconditional quantile, FGT(a), or other inequality measure, etc.), exactly analogous to the recentered influence function composition itself. Moreover, this structure also allows us to estimate easily semi-elasticity counterparts to the unconditional partial effects, which can help assess the *true potential magnitude* of a distributional change in a straightforward manner. By identifying the constant using the method described above, we can attain the correct standard errors for the semi-elasticities, calculated directly through non-linear combinations of coefficients. This approach has the added advantage of providing more intuitive and indeed more appropriate marginal effects for the interpretation of categorical variables. Fundamentally, all previous RIF regression analyses carry these issues of subjecting the estimator to inappropriately large counterfactual changes, for which the estimator was never designed to handle.

Along with this paper, we also provide a Stata program that can be used to easily implement the suggestions we present here. This can be applied for the estimation of any linear regression model, including all of the extensions developed for the analysis of linear regression RIF models.

4. References

Colin Cameron and Parvin Trivedi (2009) *Microeconometrics: Methods and Applications*, Ch 7.2.8: "Delta Method for Confidence Intervals", Cambridge University Press.

Cowell, F. A., and E. Flachaire. (2007) "Income distribution and inequality measurement: The problem of extreme values". *Journal of Econometrics* 141: 1044–1072.

Cowell, F. A., and E. Flachaire (2015). Statistical methods for distributional analysis. In Handbook of Income Distribution, vol. 2, ed. A. B. Atkinson and F. Bourguignon, 359–465. The Netherlands: Elsevier.

Deville, J.-C. (1999) "Variance estimation for complex statistics and estimators: Linearization and residual techniques", *Survey Methodology* 25: 193–203.

Firpo, Sergio, Nicole Fortin, Thomas Lemieux (2009) "Unconditional Quantile Regressions", *Econometrica*, 77(3), 953-973.

Firpo, Sergio, Nicole Fortin, Thomas Lemieux (2018) "Decomposing Wage Distributions Using Recentered Influence Function Regressions", *Econometrics*, 6(28), 1-40.

Foster, James, Joel Greer, Erik Thorbecke (1984) "A class of decomposable poverty measures", *Econometrica* 52, 761–776.

Foster, James, Joel Greer, Erik Thorbecke (2010) "The Foster-Greer-Thorbecke (FGT) Poverty Measures: Twenty-Five Years Later", Institute for International Economic Policy, Washington DC

Haisken-DeNew, John P. and Christoph M. Schmidt (1997) "Inter-Industry and Inter-Region Differentials: Mechanics and Interpretation", *Review of Economics and Statistics*, 79(3), 516-521.

Jann, Ben (2008) "The Blinder-Oaxaca decomposition for linear regression models", *The Stata Journal*, 8(4), 453-479.

Jann, Ben (2020) "Influence functions continued. A framework for estimating standard errors in reweighting, matching, and regression adjustment" University of Bern Social Sciences Working Paper No. 35

Lemieux, Thomas (2006) "Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill?" *American Economic Review*, 96 (3): 461-498.

Oaxaca, Ronald (1973) "Male-female wage differentials in urban labor markets", *International Economic Review* 14: 69-709.

Rios-Avila, Fernando (2020) "Recentered influence functions (RIFs) in Stata: RIF regression and RIF decomposition", *The Stata Journal*, 2(1), 51-94

Rios-Avila, F., & Maroto, M. L. (2022) "Moving Beyond Linear Regression: Implementing and Interpreting Quantile Regression Models With Fixed Effects", *Sociological Methods & Research*, https://doi.org/10.1177/00491241211036165

Rothe, C. (2010) "Nonparametric estimation of distributional policy effects", *Journal of Econometrics* 155: 56–70.

von Mises, R. (1947) "On the asymptotic distribution of differentiable statistical functions" *Annals of Mathematical Statistics* 18: 309–348.

Appendix A: Centered Regression

Table 1: Log wages: Unconditional Quantile Regression (50th percentile)

Explanatory Variables	(1) UCMean	(2) Ref1	(3) ST-Marg	(4) CRMarg	(5) HDS97	(6) RADN	(7) RADNdbc
0.female	--	0.0000000	0.0000000	0.0000000	0.0504556***	0.0011009***	0.0003253***
		(.)	(.)	(.)	(0.012935)	(0.000282)	(0.000083)
1.female	--	-0.2227518**	-0.1100923***	-0.1100923***	-0.0596368***	-0.0011009***	-0.0003253***
		(0.078248)	(0.028223)	(0.028223)	(0.015288)	(0.000282)	(0.000083)
5.educ	--	0.0000000	0.0000000	0.0000000	-0.2696779***	-0.0027469***	-0.0008117***
		(.)	(.)	(.)	(0.042569)	(0.000434)	(0.000128)
9.educ	--	0.0534898	0.0881486	0.0881486	-0.1815292***	-0.0019837***	-0.0005862***
		(0.078546)	(0.052751)	(0.052751)	(0.036396)	(0.000398)	(0.000117)
10.educ	--	0.1248709	0.1813769*	0.1813769*	-0.0883009	-0.0009254	-0.0002734
		(0.120881)	(0.077404)	(0.077404)	(0.063599)	(0.000667)	(0.000197)
11.educ	--	0.1604245*	0.2082279***	0.2082279***	-0.0614499***	-0.0011917***	-0.0003521***
		(0.065836)	(0.044372)	(0.044372)	(0.014486)	(0.000281)	(0.000083)
12.educ	--	0.3122144***	0.3385040***	0.3385040***	0.0688262*	0.0007951*	0.0002350*
		(0.077688)	(0.056429)	(0.056429)	(0.034024)	(0.000393)	(0.000116)
13.educ	--	0.3951552***	0.4383448***	0.4383448***	0.1686669***	0.0018575***	0.0005489***
		(0.087400)	(0.061060)	(0.061060)	(0.040458)	(0.000446)	(0.000132)
15.educ	--	0.4068822***	0.5050878***	0.5050878***	0.2354100***	0.0024738***	0.0007310***
		(0.085049)	(0.079044)	(0.079044)	(0.062245)	(0.000654)	(0.000193)
18.educ	--	0.3489731***	0.4652243***	0.4652243***	0.1955464***	0.0021535***	0.0006363***
		(0.084845)	(0.063955)	(0.063955)	(0.043675)	(0.000481)	(0.000142)
0.female#5.educ	--	0.0000000	--	--	--	--	--
		(.)					
0.female#9.educ	--	0.0000000	--	--	--	--	--
		(.)					
0.female#10.educ	--	0.0000000	--	--	--	--	--
		(.)					
0.female#11.educ	--	0.0000000	--	--	--	--	--
		(.)					
0.female#12.educ	--	0.0000000	--	--	--	--	--
		(.)					
0.female#13.educ	--	0.0000000	--	--	--	--	--
		(.)					
0.female#15.educ	--	0.0000000	--	--	--	--	--
		(.)					
0.female#18.educ	--	0.0000000	--	--	--	--	--
		(.)					
1.female#5.educ	--	0.0000000	--	--	--	--	--
		(.)					
1.female#9.educ	--	0.0756244	--	--	--	--	--
		(0.104656)					
1.female#10.educ	--	0.1232943	--	--	--	--	--
		(0.146501)					
1.female#11.educ	--	0.1043054	--	--	--	--	--
		(0.085915)					
1.female#12.educ	--	0.0573632	--	--	--	--	--
		(0.109203)					
1.female#13.educ	--	0.0942382	--	--	--	--	--
		(0.115565)					
1.female#15.educ	--	0.2142814	--	--	--	--	--
		(0.147038)					
1.female#18.educ	--	0.2536561*	--	--	--	--	--
		(0.114336)					
0.married	--	0.0000000	0.0000000	0.0000000	-0.0315069	-0.0005785	-0.0001709
		(.)	(.)	(.)	(0.017087)	(0.000314)	(0.000093)
1.married	--	0.0578452	0.0578452	0.0578452	0.0263383	0.0005785	0.0001709
		(0.031371)	(0.031371)	(0.031371)	(0.014284)	(0.000314)	(0.000093)
0.kids6	--	0.0000000	0.0000000	0.0000000	-0.0263238***	-0.0016559***	-0.0004893***
		(.)	(.)	(.)	(0.006198)	(0.000390)	(0.000115)
1.kids6	--	0.1069467*	0.1069467*	0.1069467*	0.0806229	0.0008952	0.0002645
		(0.047711)	(0.047711)	(0.047711)	(0.042538)	(0.000472)	(0.000140)
2.kids6	--	0.2961427***	0.2961427***	0.2961427***	0.2698189***	0.0028354***	0.0008378***
		(0.061167)	(0.061167)	(0.061167)	(0.057585)	(0.000605)	(0.000179)
3.kids6	--	0.0923144	0.0923144	0.0923144	0.0659906	0.0006669	0.0001970
		(0.104038)	(0.104038)	(0.104038)	(0.102347)	(0.001034)	(0.000306)
4.kids6	--	0.5338065***	0.5338065***	0.5338065***	0.5074827***	0.0050786***	0.0015007***
		(0.075675)	(0.075675)	(0.075675)	(0.074925)	(0.000750)	(0.000223)
0.kids714	--	0.0000000	0.0000000	0.0000000	-0.0175173*	-0.0009136*	-0.0002700*
		(.)	(.)	(.)	(0.006946)	(0.000362)	(0.000107)
1.kids714	--	0.0773452	0.0773452	0.0773452	0.0598279	0.0006696	0.0001979

		(0.044516)	(0.044516)	(0.044516)	(0.039163)	(0.000438)	(0.000130)
2.kids714	--	0.1184693*	0.1184693*	0.1184693*	0.1009520*	0.0010881*	0.0003215*
		(0.052167)	(0.052167)	(0.052167)	(0.047626)	(0.000513)	(0.000152)
3.kids714	--	0.1263489	0.1263489	0.1263489	0.1088316	0.0011002	0.0003251
		(0.130366)	(0.130366)	(0.130366)	(0.128422)	(0.001298)	(0.000384)
4.kids714	--	-0.2860688	-0.2860688	-0.2860688	-0.3035861	-0.0030427	-0.0008991
		(0.336799)	(0.336799)	(0.336799)	(0.335876)	(0.003366)	(0.000995)
1.isco	--	0.0000000	0.0000000	0.0000000	0.2254289***	0.0024104***	0.0007123***
		(.)	(.)	(.)	(0.046693)	(0.000499)	(0.000148)
2.isco	--	-0.0753253	-0.0753253	-0.0753253	0.1501037***	0.0017856***	0.0005276***
		(0.056418)	(0.056418)	(0.056418)	(0.034517)	(0.000411)	(0.000121)
3.isco	--	-0.1627713**	-0.1627713**	-0.1627713**	0.0626577**	0.0008816**	0.0002605**
		(0.053778)	(0.053778)	(0.053778)	(0.020112)	(0.000283)	(0.000084)
4.isco	--	-0.2205742***	-0.2205742***	-0.2205742***	0.0048547	0.0000574	0.0000170
		(0.062606)	(0.062606)	(0.062606)	(0.034153)	(0.000404)	(0.000119)
5.isco	--	-0.3824937***	-0.3824937***	-0.3824937***	-0.1570648***	-0.0017659***	-0.0005218***
		(0.062101)	(0.062101)	(0.062101)	(0.035080)	(0.000394)	(0.000116)
6.isco	--	-0.5227205***	-0.5227205***	-0.5227205***	-0.2972916**	-0.0030213**	-0.0008928**
		(0.119658)	(0.119658)	(0.119658)	(0.108028)	(0.001098)	(0.000324)
7.isco	--	-0.3841254***	-0.3841254***	-0.3841254***	-0.1586965***	-0.0018279***	-0.0005401***
		(0.060690)	(0.060690)	(0.060690)	(0.033860)	(0.000390)	(0.000115)
8.isco	--	-0.2452113**	-0.2452113**	-0.2452113**	-0.0197823	-0.0002056	-0.0000608
		(0.087289)	(0.087289)	(0.087289)	(0.070595)	(0.000734)	(0.000217)
9.isco	--	-0.6020859***	-0.6020859***	-0.6020859***	-0.3766570***	-0.0039077***	-0.0011547***
		(0.077343)	(0.077343)	(0.077343)	(0.058372)	(0.000606)	(0.000179)
exper	--	0.0099687***	0.0099687***	0.0099687***	0.0099687***	0.0099687***	0.0029456***
		(0.001624)	(0.001624)	(0.001624)	(0.001624)	(0.001624)	(0.000480)
tenure	--	0.0092932***	0.0092932***	0.0092932***	0.0092932***	0.0092932***	0.0027461***
		(0.001998)	(0.001998)	(0.001998)	(0.001998)	(0.001998)	(0.000591)
_cons	3.3842112***	3.1703680***	--	3.3842112***	3.3842112***	3.3842112***	3.3842112***
	(0.015693)	(0.079126)		(0.012414)	(0.012414)	(0.012414)	(0.012414)
N	1434	1434	1434	1434	1434	1434	1434

Note: UQR (50th percentile) regression using Swiss Household Panel from Jann (2008). Log wages are regressed using RIFreg on marital status (mar:0/1), 1-digit ISCO occupational status (1-9) and interactions of gender dummy (female:0/1) and educational status (educ:5-18), exper (0-58) and tenure (0-44.83). All regressions use the Jann (2008) variable "wt" for weighting in the regression. Column (1) displays the unconditional mean of log wages using a simple OLS regression of log wages on a constant. Column (2) displays the full UQR(50) regression with interactions. Of course these parameter estimates are sensitive to the choice of the dummy reference(s). Column (3) displays the same marginal effects as calculated by the Stata command "margins". Column (4) displays the marginal effects, collapsing any interactions into their main effects, as implemented in creg.ado. It produces identical estimate results to that of (3). After centering in Column (4), one can clearly see that the coefficient point-estimates for the constant in (1) and (4) are identical, albeit with different standard errors due to the presence of additional explanatory variables in (4). If column (3) and (4) are identical except for the missing information of the constant in (4), then we have calculated correctly. Column (5) performs the Haisken-DeNew and Schmidt (1997) procedure of taking deviations from a weighted average for all dummy variable sets (RLS), and centering any continuous variables, should they be present. Each dummy variable is interpreted as movements from the weighted average to only that dummy. The choice of any and all dummy reference categories in the original regressions no longer plays any role, as all coefficients are now deviations from the weighted average. Column (6) transforms the interpretation of all dummies to be interpreted as a 1 percentage point increase from the average, which effectively multiplies dummy coefficients in (5) by 1/(1-Dummy_Share)*100. Column (7) takes all non-Constant explanatory variable coefficients and divides them by the estimated constant using non-linear combinations arriving at the elasticity of each explanatory variable. This takes the unconditional partial effects and relativizes them by the magnitude of the dependent variable, the unconditional mean (or in this case the exposed constant 3.3842112 itself). In Column (6) the constant corresponds to the RIF component $q_y(p)$ or $v(F_Y)$) and the remaining coefficients correspond to the Influence Function or $IF\left(y_i, v(F_Y)\right)$. We explicitly use 7 digits of precision to make the calculations clear. See Appendix 1 for further details on the matrix algebra of RLS and further transformations.

Table 2: Log wages: Unconditional Quantile Regression (50th percentile)

Explanatory Variables	(1) UCMean	(2) Ref1	(3) ST-Marg	(4) CRMarg	(5) HDS97	(6) RADN	(7) RADNdbc
0.cov	--	0.0000000	0.0000000	0.0000000	-0.0874885***	-0.0033275***	-0.0018449***
		(.)	(.)	(.)	(0.000945)	(0.000036)	(0.000020)
1.cov	--	0.4331530***	0.3327530***	0.3327530***	0.2452645***	0.0033275***	0.0018449***
		(0.013605)	(0.003594)	(0.003594)	(0.002649)	(0.000036)	(0.000020)
0.edc	--	0.0000000	0.0000000	0.0000000	-0.6231018***	-0.0066782***	-0.0037026***
		(.)	(.)	(.)	(0.005524)	(0.000059)	(0.000033)
1.edc	--	0.2976684***	0.2984676***	0.2984676***	-0.3246342***	-0.0037191***	-0.0020619***
		(0.007471)	(0.006868)	(0.006868)	(0.003580)	(0.000041)	(0.000023)
2.edc	--	0.5346775***	0.5363870***	0.5363870***	-0.0867148***	-0.0014074***	-0.0007803***
		(0.006873)	(0.006226)	(0.006226)	(0.001908)	(0.000031)	(0.000017)
3.edc	--	0.7510585***	0.7310174***	0.7310174***	0.1079156***	0.0013398***	0.0007428***
		(0.007420)	(0.006732)	(0.006732)	(0.003116)	(0.000039)	(0.000021)
4.edc	--	1.1050926***	1.0041336***	1.0041336***	0.3810318***	0.0043817***	0.0024293***
		(0.007558)	(0.007110)	(0.007110)	(0.003976)	(0.000046)	(0.000025)
5.edc	--	1.1964289***	1.0925377***	1.0925377***	0.4694359***	0.0051994***	0.0028827***
		(0.007919)	(0.007154)	(0.007154)	(0.004199)	(0.000047)	(0.000026)
0.cov#0.edc	--	0.0000000	--	--	--	--	--
		(.)					
0.cov#1.edc	--	0.0000000	--	--	--	--	--
		(.)					
0.cov#2.edc	--	0.0000000	--	--	--	--	--
		(.)					
0.cov#3.edc	--	0.0000000	--	--	--	--	--
		(.)					
0.cov#4.edc	--	0.0000000	--	--	--	--	--
		(.)					
0.cov#5.edc	--	0.0000000	--	--	--	--	--
		(.)					
1.cov#0.edc	--	0.0000000	--	--	--	--	--
		(.)					
1.cov#1.edc	--	0.0030396	--	--	--	--	--
		(0.016753)					
1.cov#2.edc	--	0.0065020	--	--	--	--	--
		(0.014565)					
1.cov#3.edc	--	-0.0762244***	--	--	--	--	--
		(0.015712)					
1.cov#4.edc	--	-0.3839862***	--	--	--	--	--
		(0.017719)					
1.cov#5.edc	--	-0.3951388***	--	--	--	--	--
		(0.016868)					
0.marr	--	0.0000000	0.0000000	0.0000000	-0.1541278***	-0.0024117***	-0.0013371***
		(.)	(.)	(.)	(0.002277)	(0.000036)	(0.000020)
1.marr	--	0.2411678***	0.2411678***	0.2411678***	0.0870400***	0.0024117***	0.0013371***
		(0.003563)	(0.003563)	(0.003563)	(0.001286)	(0.000036)	(0.000020)
exper	--	0.0149103***	0.0149103***	0.0149103***	0.0149103***	0.0149103***	0.0082666***
		(0.000141)	(0.000141)	(0.000141)	(0.000141)	(0.000141)	(0.000078)
_cons	1.8036749***	0.6485083***	--	1.8036749***	1.8036749***	1.8036749***	1.8036749***
	(0.001727)	(0.006776)		(0.001475)	(0.001475)	(0.001475)	(0.001475)
N	266956	266956	266956	266956	266956	266956	266956

Note: UQR (50th percentile) regression using CPS Men data set from 1983-1985 from Firpo, Fortin, Lemieux (2009). Log wages are regressed using RIFreg on marital status (mar:0/1) and interactions of union coverage (cov:0/1) and educational status (edc:0/5) and exper (0-58). All regressions use the FFL(2009) variable "eweight" for weighting in the regression. Column (1) displays the unconditional mean of log wages using a simple OLS regression of log wages on a constant. Column (2) displays the full UQR(50) regression with interactions. Of course these parameter estimates are sensitive to the choice of the dummy reference(s). Column (3) displays the same marginal effects as calculated by the Stata command "margins". Column (4) displays the marginal effects, collapsing any interactions into their main effects, as implemented in creg.ado. It produces identical estimate results to that of (3). After centering in Column (4), one can clearly see that the coefficient point-estimates for the constant in (1) and (4) are identical, albeit with different standard errors due to the presence of additional explanatory variables in (4). If column (3) and (4) are identical except for the missing information of the constant in (4), then we have calculated correctly. Column (5) performs the Haisken-DeNew and Schmidt (1997) procedure of taking deviations from a weighted average for all dummy variable sets (RLS), and centering any continuous variables, should they be present. Each dummy variable is interpreted as movements from the weighted average to only that dummy. The choice of any and all dummy reference categories in the original regressions no longer plays any role, as all coefficients are now deviations from the weighted average. Column (6) transforms the interpretation of all dummies to be interpreted as a 1 percentage point increase from the average, which effectively multiplies dummy coefficients in (5) by 1/(1-Dummy_Share)*100. Column (7) takes all non-Constant explanatory variable coefficients and divides them by the estimated constant using non-linear combinations arriving at the elasticity of each explanatory variable. This takes the unconditional partial effects and relativizes them by the magnitude of the dependent variable, the unconditional mean (or in this case the exposed constant 1.8036749 itself). In Column (6) the constant corresponds to the RIF component $q_y(p)$ or $v(F_Y)$) and the remaining coefficients correspond to the Influence Function or $IF\left(y_i, v(F_Y)\right)$. We explicitly use 7 digits of precision to make the calculations clear. See Appendix 1 for further details on the matrix algebra of RLS and further transformations.

Appendix 1

We require an effective method to introduce *infinitesimally small* changes into the IF. Haisken-DeNew and Schmidt (1997) introduced Restricted Least Squares as a post-estimation linear combination of all dummy variable sets *and* the intercept. The restriction implemented is that the weighted sum of the coefficients of each dummy variable is set identically equal 0. This is implemented after running a standard regression of *J-1* dummies for each dummy variable set, dropping the *J*th dummy as the reference.

Assuming there are no interactions in the regression, we can continue straightforwardly. If not, the marginal effects must first be calculated[17] and then the remaining steps can be carried out. Here for example, a linear regression, using the 4 directions on the map, is run using *North*, *South* and *East*, whilst *West* is dropped due to perfect collinearity. We do the same for two gender dummies, *Male* and *Female*[18], dropping *Female*. By augmenting the $\tilde{\beta}$ with two additional columns, we regain the *West* and *Female* coefficients after the transformation with the weighting matrix *W*, indicating a matrix of the sample shares of the respective dummy variables, as depicted below. Additionally we include notation for continuous variables, which we simplify using only one indicative variable, but there can be arbitrarily many. The first 7 rows of the matrix *W* address adjustments made to the original coefficient vector $\tilde{\beta}$ to arrive at $\tilde{\tilde{\beta}}$ through the linear combinations, or weighting matrix. Row 8 address the adjustments made to the constant in the centering process. Intuitively, any means subtracted from the coefficients must be re-applied/added back to the constant for the equation still to hold.

$$W = \begin{bmatrix} 1_{Continuous} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1-\overline{North} & -\overline{South} & -\overline{East} & -\overline{West_{Ref}} & 0 & 0 & 0 \\ 0 & -\overline{North} & 1-\overline{South} & -\overline{East} & -\overline{West_{Ref}} & 0 & 0 & 0 \\ 0 & -\overline{North} & -\overline{South} & 1-\overline{East} & -\overline{West_{Ref}} & 0 & 0 & 0 \\ 0 & -\overline{North} & -\overline{South} & -\overline{East} & 1-\overline{West_{Ref}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1-\overline{Male} & -\overline{Female} & 0 \\ 0 & 0 & 0 & 0 & 0 & -\overline{Male} & 1-\overline{Female_{Ref}} & 0 \\ \overline{Continuous} & \overline{North} & \overline{South} & \overline{East} & \overline{West_{Ref}} & \overline{Male} & \overline{Female} & 1_{Const} \end{bmatrix}$$

$$\tilde{\tilde{\beta}} = W \cdot \tilde{\beta}$$

$$V(\tilde{\tilde{\beta}}) = W \cdot V(\tilde{\beta}) \cdot W'$$

The coefficient vector $\tilde{\tilde{\beta}}$ and variance-covariance matrix $V(\tilde{\tilde{\beta}})$ now include the intercept, adjusted for the deviations from the weighted average of the example two dummy sets (N, S, E, W) and (M, F). With any additional centered continuous variables, $\tilde{\tilde{\beta}}$ exposes the intercept as being exactly the $q_y(p)$

[17] Intuitively one would carry this out in Stata with the "margins" command. However, we cannot do this here and must complete this by hand, as Stata can either calculate marginal effects of the coefficients -or- provide the constant (as unconditional mean of the dependent variable), but not both at the same time.

[18] It is recognised that there are more gender identities than simply Male and Female following LGBTIQ+ definitions. This is merely a simplified illustrative example.

or $v(F_Y)$. Collectively, the linear combination of all centered continuous variables and all dummy variable sets comprise the $IF(y_i, v(F_Y))$, identically equal 0 in sum.

We transform the coefficients one step further to allow for a 1 percentage point increase in the share of a dummy variable from its mean, as opposed to going from the mean to full 100% of the dummy variable, multiplying the coefficient vector $\tilde{\tilde{\beta}}$ and variance-covariance matrix $V(\tilde{\tilde{\beta}})$ by 1/(1-$\overline{Dummy}$), which $\overline{Dummy}$ is the respective dummy sample share, to arrive at the coefficient vector $\ddot{\beta}$ and variance-covariance matrix $V(\ddot{\beta})$, both of which include the intercept and a continuous variable.

$$\ddot{\beta} = TT \cdot \tilde{\tilde{\beta}}$$

$$V(\ddot{\beta}) = TT' \cdot V(\tilde{\tilde{\beta}}) \cdot TT$$

where the new weighting matrix TT for dummy variables is given by an identity matrix with 1/(1-$\overline{Dummy}$)*100 on the diagonal element for any *Dummy*. The constant and any continuous variables remain unchanged, as indicated by the diagonal element of "1".

The advantage of doing this is that we now have symmetrical marginal effects for two-outcome dummy variables. For a continuous variable this is nothing new: symmetry is guaranteed with the single coefficient. Yet, for a simple dummy variable with two outcome categories, a 1 percent point increase in the first category is now *the exact same magnitude but opposite in sign* to that of the second category. It remains that standard errors of each of the coefficients in the two-outcome dummy case are identical.

Finally, in order to put the relative magnitude of the estimated coefficients into perspective, we relativize the estimated coefficients of the explanatory variables to the magnitude of the now cleanly estimated a_0 or statistic of interest $v(F_y)$, providing us with elasticities. Thus, we take each estimated explanatory coefficient (except for the constant a_0 itself) and divide through by the constant a_0 in a *non*-linear combination (taking into account correct standard error transformations) using the Delta Method[19].

$$\check{\beta} = \left[\frac{\ddot{\beta}}{a_0}\right]$$

[19] This is implemented straightforwardly in Stata using the command <nlcom> which performs the non-linear transformation, taking into account the standard error of the constant. In our Stata ado, this is run in the background for every explanatory variable, other than the constant itself.

Appendix C: Stata Command Files

```
//                  EXAMPLE 1

//                  Use Oaxaca data: Jann (2008)
use                 oaxaca, clear
compress

drop                if lnwage==. | married==. | female==. | isco==. ///
                       | exper==. | educ==. | lfp==. | tenure==.

replace             educ=round(educ)

compress


//                  Define model specification
global              rhs1      i.female##i.educ i.marr i.kids6 i.kids714 i.isco exper tenure
global              wgt       [aweight=wt]


//                  Get unconditional mean of statistic q(50)
//                  (1) Unconditional Mean: just constant
Rifhdreg            lnwage  $wgt ,    ///
                    rif(q(50)  kernel(gaussian) bw(0.06))   ///
                    robust
eststo              umean

//                  (2) Coefficients as is, including interactions

rifhdreg            lnwage $rhs1 $wgt ,      ///
                    rif(q(50) kernel(gaussian) bw(0.06))    ///
                    robust
eststo              set1_orig

//                  (3) Marginal Calculation: margins

margins,            dydx(*) post
eststo              set1_marg

//                  (4) Marginal Calculation: creg
//                  (5) HDS97 based on marginals
//                  (6) RADN based on 1ppt

rifhdreg            lnwage $rhs1 $wgt ,      ///
                    rif(q(50)  kernel(gaussian) bw(0.06))   ///
                    robust
creg,               eval      radn   eststub(set1)

//                  (7) RADN based on 1ppt & divide by constant

rifhdreg            lnwage $rhs1 $wgt ,      ///
                    rif(q(50)  kernel(gaussian) bw(0.06))   ///
                    robust
creg,               eval      radn   eststub(set2) divbycons


//                  (1) Unconditional Mean: just constant
//                  (2) Coefficients as is, including interactions
//                  (3) Marginal Calculation: margins
//                  (4) Marginal Calculation: creg
//                  (5) HDS97 based on marginals
//                  (6) RADN based on 1ppt
//                  (7) RADN based on 1ppt & divide by constant


esttab              umean      set1_orig   set1_marg    set1_crmarg             ///
                    set1_hds97 set1_radn   set2_radn    using tab2.rtf, replace  ///
                    cells(b(star fmt(7) vacant(--)) se(par(( )) fmt(6)))         ///
                    mlabels("UCMean" "Ref1" "ST-Marg" "CRMarg"  "HDS97" "RADN"   ///
                                    "RADNdbc" ) collabels(none)

//                  ============================================================================
```

```
//                  EXAMPLE 2

//                  Use FFL (2009) data
Use                 men8385, clear


//                  Use FFL's variable definitions
gen                 edc=0 if educ<9
Replace             edc=1 if educ<12 & educ>=9
Replace             edc=2 if educ>=12 & educ<13
Replace             edc=3 if educ>=13 & educ<=15
Replace             edc=4 if educ==16
Replace             edc=5 if educ>16
Gen                 cov=covered
compress

//                  Define model specification
global              rhs1      i.cov##i.edc i.marr exper
global              wgt       [aweight=eweight]


//                  Get unconditional mean of statistic q(50)
//                  (1) Unconditional Mean: just constant
Rifhdreg            lwage  $wgt ,     ///
                    rif(q(50)  kernel(gaussian) bw(0.06))   ///
                    robust
eststo              umean

//                  (2) Coefficients as is, including interactions

rifhdreg            lwage $rhs1 $wgt ,       ///
                    rif(q(50) kernel(gaussian) bw(0.06))    ///
                    robust
eststo              set1_orig

//                  (3) Marginal Calculation: margins

margins,            dydx(*) post
eststo              set1_marg

//                  (4) Marginal Calculation: creg
//                  (5) HDS97 based on marginals
//                  (6) RADN based on 1ppt

rifhdreg            lwage $rhs1 $wgt ,       ///
                    rif(q(50)  kernel(gaussian) bw(0.06))   ///
                    robust
creg,               eval      radn   eststub(set1)

//                  (7) RADN based on 1ppt & divide by constant

rifhdreg            lwage $rhs1 $wgt ,       ///
                    rif(q(50)  kernel(gaussian) bw(0.06))   ///
                    robust
creg,               eval      radn   eststub(set2) divbycons


//                  (1) Unconditional Mean: just constant
//                  (2) Coefficients as is, including interactions
//                  (3) Marginal Calculation: margins
//                  (4) Marginal Calculation: creg
//                  (5) HDS97 based on marginals
//                  (6) RADN based on 1ppt
//                  (7) RADN based on 1ppt & divide by constant


esttab              umean      set1_orig   set1_marg    set1_crmarg              ///
                    set1_hds97 set1_radn   set2_radn    using tab1.rtf, replace  ///
                    cells(b(star fmt(7) vacant(--)) se(par(( )) fmt(6)))         ///
                    mlabels("UCMean" "Ref1" "ST-Marg" "CRMarg"  "HDS97" "RADN"   ///
                                "RADNdbc" ) collabels(none)

//                  ============================================================================
```

Appendix D: “creg” help file

help creg

---

creg -- Centered linear regression (all RHS variable coefficients are treated as resulting from "de-meaned" explanatory variables)

Syntax

creg [, options]

options	Description
Handling simulations appropriate for use after rifhdreg or any reg-based linear model	
eval	eval activate transformations a la Haisken-DeNew and Schmidt (1997).
radn	radn activate transformations a la Rios-Avila and de New (2022).
divbycons	divbycons divide all coefficients by the constant a la Rios-Avila and de New (2022).
pp(#)	pp is set by default to 1 percentage point; this is used when one is simulating a 1 percentage-point increase in a dummy variable using the option radn.
eststub(string)	eststub is a string indicating a name for the set of estimation results for use in estimates table.

Postestimation command for linear regression models

creg (Post-Estimation command with options)

creg is a post-estimation command run after any linear regression command like regress, areg, rifhdreg or xtreg. It adjusts coefficients and the variance-covariance matrix.

creg requires all factor variables in the main command (regress, areg, rifhdreg or xtreg) to be explicitly declared using the standard "c." or "i." variable prefixes. All other variables are assumed to be continuous.

All dummy variable coefficient sets are adjusted to be deviations from a weighted average.

All continuous variable coefficients are adjusted as if the variables had been deviations from their means.

The result is a "centered regression" such that the overall constant of the regression is equal to the unconditional mean of the dependent variable.

Description

creg implements the restricted least squares (RLS) procedure for dummy variable sets as described by Haisken-DeNew and Schmidt (1997). For example, log wages are regressed on a group of k-1 industry/region/job/etc dummies using Stata's factor variable notation (e.g. i.gender). The k-th dummy is the omitted reference dummy.

Using the factor variable notation, one can select the desired reference dummy (e.g. b2.gender or b1.gender). It does not matter for RLS. Using RLS, all k dummy coefficients and standard errors are reported. The coefficients are interpreted as deviations from the dummy group weighted average.

If the preceeding regression command has a constant, it is adjusted to include the dummy set averages. It does not matter how many sets of dummies are included in the previous regression. All sets will be handled if using factor variable notation.

This ado corrects problems with the Krueger and Summers (1988) Econometrica methodology of overstated differential standard errors, and understated overall dispersion. The command creg is run after regress, areg, rifhdreg or xtreg.

The coefficients of continuous variables are also affected by creg. All continuous variable coefficients reflect a "demeaned variable", such that this variable mean (times the coefficient) is added to the constant.

Also, all results calculated in creg are independent of the choice of the reference category. By the way, for all dummy variable sets having only two outcomes, i.e. male/female, the t-values of the creg adjusted coefficients are always equal in magnitude, but opposite in sign.

creg currently can deal with some interactions. You may only use the "i" and "b" notation when using factor variables, e.g. i.race or b2.race but not i.race#i.industry or i.race##i.industry or i.race#c.grade.

There are some interactions which creg cannot handle directly. However, see fvint for single # dummy set interactions.

If you have specfied any weights using the previous reg, areg, rifhdreg or xtreg command, creg will automatically use these same weights to weight the means of the dummies in the dummy set to arrive at the weighted . If no weights were used in the previous command, then creg assumes no weights. Also using the if e(sample) condition, creg uses by definition the same observations as in the previous regression command.

Stored results

creg stores the following in e( ):

Matrices
e(b) creg replaces the e(b) of the previous regression.
e(V) creg replaces the e(V) of the previous regression.

Macros
e(allfactors) List of all factor variable base names, e.g. race state education.
e(all_sd) List of all factor variable standard deviation of values from a dummy variable set, e.g. for the regressor i.race, the contents of e(all_sd) would be "race_sd" and "race_sd" is the scalar register name of e(race_sd).

Scalars (Example given for the factor variable: i.race used in previous command)
e(race_sd) Assuming the factor variable i.race, e(race_sd) is the standard deviation of of the associated coefficients weighted by their sample means and taking into account their respective standard errors.
e(race_f) Assuming the factor variable i.race, e(race_f) is the F statistic associated with the joint test of all associated coefficients being equal to zero.
e(race_df) Assuming the factor variable i.race, e(race_df) is degrees of freedom of the F statistic associated with the joint test of all associated coefficients being equal to zero.
e(race_dfr) Assuming the factor variable i.race, e(race_dfr) is restricted degrees of freedom of the F statistic associated with the joint test of all associated coefficients being equal to zero.
e(race_p) Assuming the factor variable i.race, e(race_p) is p-value of the F statistic associated with the joint test of all associated coefficients being equal to zero.

References

Haisken-DeNew, John P. and Christoph M. Schmidt (1997): "Inter-Industry and Inter-Region Wage Differentials: Mechanics and Interpretation," Review of Economics and Statistics, 79(3), 516-21. Download REStat Reprint

Krueger, Alan and Lawrence Summers (1988): "Efficiency wages and the Inter-Industry Wage Structure", Econometrica, 56, 259-193. Download Econometrica Reprint

Numerical example after regress command

```
. sysuse nlsw88, clear
. numlabel, add mask("[#] ")
. tab race
```

```
. tab race

       Race |      Freq.     Percent        Cum.
------------+-----------------------------------
  [1] White |      1,637       72.89       72.89
  [2] Black |        583       25.96       98.84
  [3] Other |         26        1.16      100.00
------------+-----------------------------------
      Total |      2,246      100.00
```

```
    . regress wage b1.race
```

```
. regress   wage b1.race

      Source |       SS           df       MS      Number of obs   =     2,246
-------------+----------------------------------   F(2, 2243)      =     10.28
       Model |  675.510282         2  337.755141   Prob > F        =    0.0000
    Residual |  73692.4571     2,243  32.8544169   R-squared       =    0.0091
-------------+----------------------------------   Adj R-squared   =    0.0082
       Total |  74367.9674     2,245  33.1260434   Root MSE        =    5.7319

------------------------------------------------------------------------------
        wage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        race |
  [2] Black  |  -1.238442   .2764488    -4.48   0.000    -1.780564   -.6963193
  [3] Other  |   .4677818   1.133005     0.41   0.680    -1.754067    2.689631
             |
       _cons |   8.082999   .1416683    57.06   0.000     7.805185    8.360814
------------------------------------------------------------------------------
```

```
    . creg, eval

 Restricted Least Squares for Dummy Variable Sets (Stata Factor Variables)

 Authors     : Prof Dr John P. de New and Prof Dr Christoph M. Schmidt
               Version: 22 Dec 2021

 Citation    : Haisken-DeNew, J.P. and Schmidt C.M. (1997):
               "Interindustry and Interregion Wage Differentials:
               Mechanics and Interpretation," Review of Economics
               and Statistics, 79(3), 516-521. REStat Reprint

------------------------------------------------------------------------------
        wage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        race |
  [1] White  |   .3160504   .0737694     4.28   0.000     .1713869    .4607138
  [2] Black  |  -.9223912   .2042697    -4.52   0.000    -1.322969   -.5218139
  [3] Other  |   .7838322   1.117588     0.70   0.483    -1.407783    2.975448
             |
       _cons |   7.766949   .1209461    64.22   0.000     7.529771    8.004127
------------------------------------------------------------------------------

 Sampling-Error-Corrected Standard Deviation of Differentials
 Joint test of all coefficients in dummy variable set = 0, Prob > F = p
-------------------------------------------------------------------------------
               race |  0.521062      F(2,2243) = 10.28         p=0.0000
-------------------------------------------------------------------------------
```

Numerical example after regress command, comparing results

```
    . sysuse nlsw88, clear
    . numlabel, add mask("[#] ")

    . regress wage
    . estimates store b0

    . regress wage b1.race
    . estimates store b1
```

```
. regress wage b2.race
. estimates store b2

. regress wage b3.race
. estimates store b3

. regress wage b1.race
. creg, eval
. estimates store hds1

. regress wage b2.race
. creg, eval
. estimates store hds2

. regress wage b3.race
. creg, eval
. estimates store hds3

// Install estout if you have not already
. ssc install estout

// Now display the table of results
. estout with lots of options ...
```

The estimation results b1, b2, b3 are all different, as in each case, there is a different base or reference category. However, the estimation results hds1, hds2, hds3 are all identical, regardless of base category used.

The constant has been adjusted as well, to reflect the weighted average that had been removed from the deviations.

The constant reported will always be the unconditional mean of the dependent variable, after constant adjustments for dummy and continuous explanatory variables are made through centering.

Numerical example after xtreg command (compare to reg with i.factor)

```
// Given that i.company is already the ID in xtset,
// we should expect identical results for estimations (1) and (2):
// (1) xtreg invest mvalue kstock i.time, fe
// (2) reg invest mvalue kstock i.time i.company

. webuse grunfeld, clear
. compress
. xtset


// Examples using xtreg
. xtreg invest mvalue kstock i.time, fe
. estimates store B1
. creg, eval
. estimates store HDS1

. xtreg invest mvalue kstock b2.time, fe
. estimates store B2
. creg, eval
. estimates store HDS2

// Examples using reg
. reg invest mvalue kstock b2.company i.time
. estimates store B3
. creg, eval
. estimates store HDS3

. reg invest mvalue kstock b2.company b2.time
. estimates store B4
. creg, eval
. estimates store HDS4

// Example using areg
. areg invest mvalue kstock b3.time, absorb(company)
. estimates store B5
. creg, eval
. estimates store HDS5
```

```
// Install estout if you have not already
.  ssc install estout

// Now display the table of results
. estout with lots of options ...
```

Again, the estimation results B1, B2, B3, B4, B5 are all different, as in each case, there is a different base/reference category. The estimation results HDS1, HDS2, HDS3, HDS4, HDS5 are all identical, regardless of base category used.

The constant has been adjusted as well, to reflect the weighted average that had been removed from the deviations and continuous variables.

The weighted average of every dummy variable set has been put back into the constant. The mean X_bar of every continuous variable X is multiplied by the estimated coefficient _b[X] and also re-added to the constant.

Numerical example after reg command (using factor variable interactions)

```
. sysuse nlsw88, clear
. numlabel, add mask("[#] ")

// Make sure you have installed Benn Jann's cool center ado
. ssc install center

// Must pre-center variables used in polynomial
. center age

// Run regession using explanatory vars and interactions
. reg wage i.occupation##i.race i.industry c.c_age##c.c_age hours

. gen touse=e(sample)
// i.occupation##i.race   : the complex dummy-on-dummy interaction
//                          (i.occupation#i.race)
// i.industry             : a standard dummy set
// c.c_age##c.c_age           : a quandratic term in age (must be 2#'s)
// c.grade#i.south        : a continuous var interacted with a dummy set
// hours                  : a standard continuous var

// Now run post-estimation command to adjust e(b) and e(V) to "center" results.
// Automatically get marginal effects of interactions and polynomials.


// Because there is a quadratic in c_age and an interaction
// in i.occupation##i.race, we first need marginals. See what Stata says:
. stata margins, dydx(*)

// Re-Run regession using explanatory vars and interactions
. reg wage i.occupation##i.race i.industry c.c_age##c.c_age hours

. creg, eval

// Run naked regession without any explanatory variables
// The constant is the unconditional mean of the dependent variable wage
. reg wage if touse==1
```

Numerical example after reg command (using factor variable interactions)

```
. sysuse oaxaca, clear

. drop if lnwage==. | married==. | female==. | isco==.  | age==.

// Make sure you have installed Benn Jann's cool center ado
. ssc install center

// You must pre-center a variable to be used in a quadratic or polynomial
. center age

. label define isco 1 "Managers", modify
```

```
. label define isco 2 "Professional", modify
. label define isco 3 "Technicians and associate professionals", modify
. label define isco 4 "Clerical support workers", modify
. label define isco 5 "Service and sales workers", modify
. label define isco 6 "Skilled agricultural, forestry and fishery workers", modify
. label define isco 7 "Craft and related trades workers", modify
. label define isco 8 "Plant and machine operators, and assemblers", modify
. label define isco 9 "Elementary occupations", modify
. label values isco isco

. label define married 0 "not married", modify
. label define married 1 "married", modify
. label values married married

. label define single 0 "not single", modify
. label define single 1 "single", modify
. label values single single

. label define female 0 "not female", modify
. label define female 1 "female", modify
. label values female female

. numlabel, add mask("[#] ")

. desc

// get unconditional mean
. reg lnwage

// run simple regression
// interaction of i.married#i.isco
// interaction of i.single#i.female
. reg lnwage educ exper i.kids6 i.kids714 i.married##i.isco c.c_age##c.c_age

// get HDS97 and center any continuous vars; transform a la Rios-Avila & de New (2021)
// Continuous vars: 1 unit increase from average
// Dummy vars: 1 percentage point (PP) increase from average
// Constant: unconditional mean of LHS var or functional
. creg, eval radn pp(1)


// Make sure you have installed Fernando Rios-Avila's cool rifhdreg ado
. ssc install rifhdreg

// Do Unconditional Quantile Regression at the 25th percentile of LHS var "lnwage"
. rifhdreg lnwage educ exper i.kids6 i.kids714 i.married##i.isco c.c_age##c.c_age,
    rif(q(25))

// get HDS97 and center any continuous vars; transform a la Rios-Avila & de New (2021)
// Continuous vars: 1 unit increase from average
// Dummy vars: 1 percentage point (PP) increase from average
// Constant: unconditional mean of LHS var or functional
. creg, eval radn pp(1)

// Now do it again with Stata "margins" to check marginals
. rifhdreg lnwage educ exper i.kids6 i.kids714 i.married##i.isco c.c_age##c.c_age,
    rif(q(25))
. margins, dydx(*)

// Compare with creg marginals
. creg, eval
```

Authors

Dealing with Restricted Least Squares
Email: Prof Dr John P. de New and Prof Dr Christoph M. Schmidt

Dealing with Centered Regression with RIFs/UQR
Email: Prof Dr John P. de New and Dr Fernando Rios-Avila

We would be delighted if you cited us if you use this ado and research. Please drop us a line if you do.
