Estimating a counterfactual wage distribution using survey data

Mihaela-Cătălina Anastasiade[1], Alina Matei[2] and Yves Tillé[2]

Swiss Federal Statistical Office[1], University of Neuchâtel[2], Switzerland

Colloque francophone sur les sondages 2018
Lyon

Aims

- We present a parametric approach to estimate 'wage discrimination' at different quantiles.
- The goal is to reduce the variance of the estimates compared to the existing methods.
- We illustrate this approach using the generalized beta of the second kind distribution (hereafter, GB2).

Framework

- Consider a finite population with the labels $U = \{1, 2, \ldots, N\}$.
- From this population, we randomly select a sample S of size n, without replacement.
- The sample is selected through a sampling design $p(s) = \Pr(S = s), \forall s \subseteq U$.
- To each unit $k \in S$, a survey weight w_k is associated.
- These weights can be equal to the inverse of the inclusion probabilities or can be more complicated weights, like calibration weights.
- Let y be the variable of interest (the wage).

Superpopulation framework

- Assume that $U = U_M \cup U_F$, $U_M \cap U_F = \emptyset$ is drawn from a superpopulation.
- The superpopulation is also divided in two subsuperpopulations from where the subsets $U_g, g = \{M, F\}$ are drawn, respectively.
- The wage is a random variable Y_g and $\mathbf{X}_g$ is a set of covariates.
- In each subset U_g,

$$Y_{k,g} \mid \mathbf{X}_g = \mathbf{x}_{k,g} \sim D(\gamma_{k,g}, \boldsymbol{\delta}_g), k \in U_g.$$

- We assume that $D(\gamma_{k,g}, \boldsymbol{\delta}_g), k \in U_g$ is a continuous distribution and that $\gamma_{k,g} = h(\mathbf{x}_{k,g}^\top \boldsymbol{\beta}_g)$, where h is a known continuous function.
- The sample $S = S_M \cup S_F$, $S_M \cap S_F = \emptyset, S_g \subseteq U_g, g = \{M, F\}$.

The counterfactual wage distribution

The CDF of the counterfactual wage distribution is defined as

$$F^C(y) = \int_{\mathcal{X}_{\mathcal{M}}} F^{Y_F|\mathbf{X}_F}(y \mid \mathbf{x}) dF^{\mathbf{X}_M}(\mathbf{x}),$$

where $\mathcal{X}_{\mathcal{M}}$ is the support of $\mathbf{X}_M$ and $\mathcal{X}_{\mathcal{F}}$ the support of $\mathbf{X}_F$.
It is assumed that $\mathcal{X}_{\mathcal{M}} \subseteq \mathcal{X}_{\mathcal{F}}$.
It is interpreted as the distribution function of wages that would be obtained for women if their characteristics were same as those of men.

$$F^F(y) = \int_{\mathcal{X}_{\mathcal{F}}} F^{Y_F|\mathbf{X}_F}(y \mid \mathbf{x}) dF^{\mathbf{X}_F}(\mathbf{x}),$$

$$F^M(y) = \int_{\mathcal{X}_{\mathcal{M}}} F^{Y_M|\mathbf{X}_M}(y \mid \mathbf{x}) dF^{\mathbf{X}_M}(\mathbf{x}).$$

The counterfactual wage distribution

The CDF of the counterfactual wage distribution is defined as

$$\begin{aligned} F^C(y) &= \int_{\mathcal{X}_\mathcal{M}} F^{Y_F|\mathbf{X}_F}(y \mid \mathbf{x}) dF^{\mathbf{X}_M}(\mathbf{x}) \\ &= \int_{\mathcal{X}_\mathcal{F}} F^{Y_F|\mathbf{X}_F}(y \mid \mathbf{x}) \frac{dF^{\mathbf{X}_M}(\mathbf{x})}{dF^{\mathbf{X}_F}(\mathbf{x})} dF^{\mathbf{X}_F}(\mathbf{x}) \\ &= \int_{\mathcal{X}_\mathcal{F}} F^{Y_F|\mathbf{X}_F}(y \mid \mathbf{x}) \psi(\mathbf{x}) dF^{\mathbf{X}_F}(\mathbf{x}), \end{aligned}$$

where $\mathcal{X}_\mathcal{M}$ is the support of $\mathbf{X}_M$ and $\mathcal{X}_\mathcal{F}$ the support of $\mathbf{X}_F$. It is assumed that $\mathcal{X}_\mathcal{M} = \mathcal{X}_\mathcal{F}$.

$$F^F(y) = \int_{\mathcal{X}_\mathcal{F}} F^{Y_F|\mathbf{X}_F}(y \mid \mathbf{x}) dF^{\mathbf{X}_F}(\mathbf{x}),$$

The weighted DiNardo, Fortin and Lemieux method

- DiNardo et al. (1996) write the reweighting factor $\psi(\mathbf{x}_k) = \frac{dF^{\mathbf{X}_M}(\mathbf{x}_k)}{dF^{\mathbf{X}_F}(\mathbf{x}_k)}$ as

$$\psi(\mathbf{x}_k) = \psi_k = \frac{P(Gender_k = \text{'}man\text{'} \mid \mathbf{x}_k)/P(Gender_k = \text{'}man\text{'})}{P(Gender_k = \text{'}woman\text{'} \mid \mathbf{x}_k)/P(Gender_k = \text{'}woman\text{'})}$$

- The idea is to reweigh the characteristics of women so that they match the characteristics of men, such that

$$\widehat{\bar{\mathbf{X}}}_C = \widehat{\bar{\mathbf{X}}}_M,$$

where $\widehat{\bar{\mathbf{X}}}_C = \sum_{k \in S_F} \widehat{\psi}_k w_k \mathbf{x}_k / \sum_{k \in S_F} \widehat{\psi}_k w_k$ and $\widehat{\bar{\mathbf{X}}}_M = \sum_{k \in S_M} w_k \mathbf{x}_k / \sum_{k \in S_M} w_k$.

- The factor $\psi(\mathbf{x}_k)$ can be estimated by using a probit or a logistic regression model (DiNardo et al., 1996) or by calibration (Anastasiade and Tillé, 2017).

Aim

If $\Delta_{(\alpha)}$ is the wage difference between men and women at a given quantile α, we can write

$$\Delta_{(\alpha)} = Q^M_{(\alpha)} - Q^F_{(\alpha)} = (Q^M_{(\alpha)} - Q^C_{(\alpha)}) + (Q^C_{(\alpha)} - Q^F_{(\alpha)}),$$

where $Q^M_{(\alpha)}$, $Q^C_{(\alpha)}$ and $Q^F_{(\alpha)}$ are the quantile of order α of men, counterfactual and women wage distributions, respectively.

$$\widehat{\Delta}_{(\alpha)} = \widehat{Q}^M_{(\alpha)} - \widehat{Q}^F_{(\alpha)} = (\widehat{Q}^M_{(\alpha)} - \widehat{Q}^C_{(\alpha)}) + (\widehat{Q}^C_{(\alpha)} - \widehat{Q}^F_{(\alpha)}).$$

Estimation of the CDF in finite populations

- The empirical CDF at the U level is defined as

$$F_{emp}(y) = \frac{\sum_{k \in U} I(y_k \leq y)}{N}.$$

- The classical design-based estimator is

$$\widehat{F}_{emp}(y) = \frac{\sum_{k \in S} w_k I(y_k \leq y)}{\sum_{k \in S} w_k}.$$

- The quantile of order α of y is estimated by

$$\widehat{Q}_{\alpha,emp}(y) = \inf\{\widehat{F}_{emp}(y) \geq \alpha\}.$$

Parametric approach

- We assume that U is selected from a superpopulation and Y is a random variable with the CDF $F(.)$.
- We include auxiliary information $\mathbf{X}$ in the estimation of $F(.)$ by assuming that $Y_k \mid \mathbf{X}_k \sim D(\gamma_k = h(\mathbf{x}_k^\top \beta), \boldsymbol{\delta}_k), k \in U$.
- We write

$$F_U(y) = \sum_{k \in U} \lambda_k F_{D(\gamma_{k,g}, \boldsymbol{\delta})}(y \mid \mathbf{x}_{k,g}) = \frac{1}{N} \sum_{k \in U} F_{D(\gamma_{k,g}, \boldsymbol{\delta})}(y \mid \mathbf{x}_{k,g}),$$

where $\lambda_k = 1/N$, $F_{D(\gamma_k, \boldsymbol{\delta})}(. \mid \mathbf{x}_k)$ is the CDF of the distribution $D(\gamma_k = h(\mathbf{x}_k^\top \beta), \boldsymbol{\delta}_k), k \in U$, and $h(.)$ is a continuous function.

Method 1 for quantile estimation

We propose to estimate the quantile of order α of Y as

$$\widehat{Q}_{(\alpha)} = \inf\{y \mid \widehat{F}_U(y) \geq \alpha\},$$

where $\widehat{F}_U(y)$ is the estimator of $F_U(y)$ in the point y given by

$$\widehat{F}_U(y) = \sum_{k \in S} w_k \widehat{F}_{D(\widehat{\gamma}_k, \widehat{\boldsymbol{\delta}})}(y_k \mid \mathbf{x}_k) / \sum_{k \in S} w_k.$$

Method 2 for quantile estimation

In case the inverse function of $\widehat{F}_U(y)$ cannot be computed, we propose to use a Monte Carlo method based on parametric bootstrap.

Method 2 for quantile estimation

$Y_{i,k} \mid \mathbf{x}_k \sim D(h(\mathbf{x}_k^\top \widehat{\beta}), \widehat{\delta})$

$$\begin{pmatrix} w_1 & w_2 & w_3 & \dots & w_n \\ \hline y_{11} & y_{12} & y_{13} & \dots & y_{1n} \\ y_{21} & y_{22} & y_{23} & \dots & y_{2n} \\ y_{31} & y_{32} & y_{33} & \dots & y_{3n} \\ \dots & \dots & \dots & \dots & \\ y_{m1} & y_{m2} & y_{m3} & \dots & y_{mn} \end{pmatrix} \longrightarrow \begin{pmatrix} \widehat{Q}_1^{0.25} & \widehat{Q}_1^{0.5} & \widehat{Q}_1^{0.75} & \widehat{Q}_1^{0.95} \\ \widehat{Q}_2^{0.25} & \widehat{Q}_2^{0.5} & \widehat{Q}_2^{0.75} & \widehat{Q}_2^{0.95} \\ \widehat{Q}_3^{0.25} & \widehat{Q}_3^{0.5} & \widehat{Q}_3^{0.75} & \widehat{Q}_3^{0.95} \\ \dots & \dots & \dots & \dots \\ \widehat{Q}_m^{0.25} & \widehat{Q}_4^{0.5} & \widehat{Q}_m^{0.75} & \widehat{Q}_4^{0.95} \\ \hline \widehat{\overline{Q}}_{0.25} & \widehat{\overline{Q}}_{0.5} & \widehat{\overline{Q}}_{0.75} & \widehat{\overline{Q}}_{0.95} \end{pmatrix}$$

Remark: Methods 1 and 2 are applied to estimate respectively the α-quantiles in each group $g \in \{M, F\}$.

Method 1 for the counterfactual wage distribution

We redefine the counterfactual CDF at the U_F level as

$$F^C_{U_F}(y) = \frac{1}{N_C} \sum_{k \in U_F} \psi_k F^{(Y_F|\mathbf{X}_F)}(y_k \mid \mathbf{x}_{k,F}),$$

where $N_C = \sum_{k \in U_F} \psi_k$.

- First, we estimate it by

$$\widehat{F}^C_{U_F}(y) = \frac{\sum_{k \in S_F} \widehat{\psi}_k w_k \widehat{F}^{(Y_F|\mathbf{X}_F)}(y_k \mid \mathbf{x}_{k,F})}{\sum_{k \in S_F} \widehat{\psi}_k w_k},$$

  where $\widehat{F}^{(Y_F|\mathbf{X}_F)}(y_k \mid \mathbf{x}_{k,F}) = F_{D(h(\mathbf{x}_{k,F}^\top \widehat{\boldsymbol{\beta}}_F), \widehat{\boldsymbol{\delta}}_F)}(y_k \mid \mathbf{x}_{k,F})$, and $\widehat{\psi}_k$ is estimated by calibration (Anastasiade and Tillé, 2017).
- Next, $\widehat{Q}^C_{(\alpha)} = \inf\{y \mid \widehat{F}^C_{U_F}(y) \geq \alpha\}$.

Method 2 for the counterfactual wage distribution

$$Y_{i,k} \mid \mathbf{x}_{k,F} \sim D(h(\mathbf{x}_{k,F}^\top \widehat{\beta}_F), \widehat{\delta}_F)$$

$$\left(\begin{array}{cccc}
\widehat{\psi}_1 w_1 & \widehat{\psi}_2 w_2 & \dots & \widehat{\psi}_{n_F} w_{n_F} \\
\hline
y_{11} & y_{12} & \dots & y_{1n_F} \\
y_{21} & y_{22} & \dots & y_{2n_F} \\
y_{31} & y_{32} & \dots & y_{3n_F} \\
\dots & \dots & \dots & \dots \\
y_{m1} & y_{m2} & \dots & y_{mn_F}
\end{array}\right)
\longrightarrow
\left(\begin{array}{cccc}
\widehat{Q}_1^{0.25} & \widehat{Q}_1^{0.5} & \widehat{Q}_1^{0.75} & \widehat{Q}_1^{0.95} \\
\widehat{Q}_2^{0.25} & \widehat{Q}_2^{0.5} & \widehat{Q}_2^{0.75} & \widehat{Q}_2^{0.95} \\
\widehat{Q}_3^{0.25} & \widehat{Q}_3^{0.5} & \widehat{Q}_3^{0.75} & \widehat{Q}_3^{0.95} \\
\dots & \dots & \dots & \dots \\
\widehat{Q}_m^{0.25} & \widehat{Q}_4^{0.5} & \widehat{Q}_m^{0.75} & \widehat{Q}_4^{0.95} \\
\hline
\widehat{\overline{Q}}^{C}_{0.25} & \widehat{\overline{Q}}^{C}_{0.5} & \widehat{\overline{Q}}^{C}_{0.75} & \widehat{\overline{Q}}^{C}_{0.95}
\end{array}\right)$$

The two methods

- The proposed methods aim to reduce the variance of the estimated quantiles compared to the estimation given by the empirical CDF.
- The methods are correct if the underlined conditional distribution is correct.
- Departures from this assumption can be managed by using a GB2 distribution.

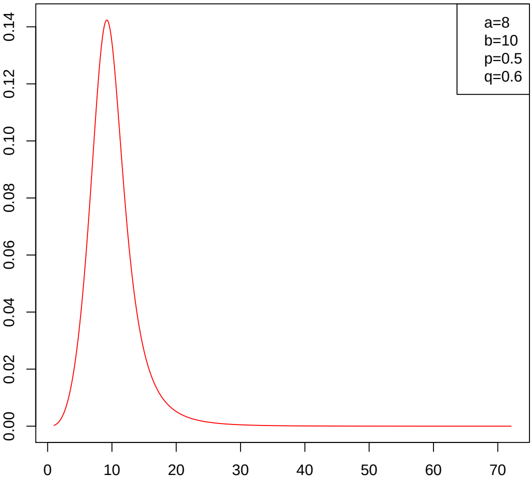
The GB2 distribution

The GB2 distribution is characterized by four parameters, namely a, b, p and q. The probability density function of a $GB2(a, b, p, q)$ distribution is given by

$$f(y; a, b, p, q) = \frac{a\left(\frac{y}{b}\right)^{ap-1}}{b\mathrm{B}(p, q)[1 + \left(\frac{y}{b}\right)^{a}]^{p+q}},$$

where a, p, q are the shape parameters and b is a scale parameter.

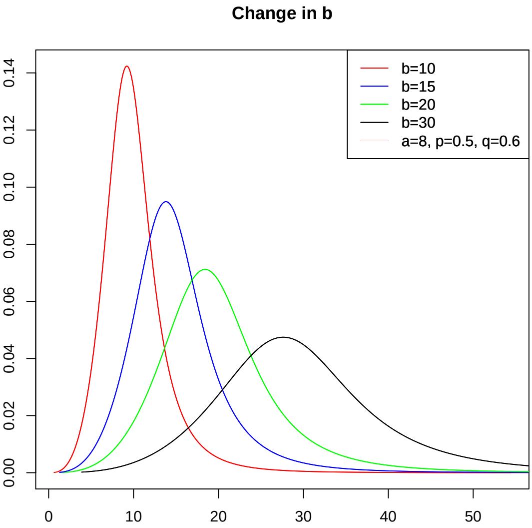
Example of GB2 distribution



- We borrow from McDonald and Butler (1990) the idea of changing the scale parameter, by expressing it as a function of the observed characteristics.
- In each group, $g \in \{M, F\}$, we assume that the conditional wage of $k \in U_g$, $Y_k \mid \mathbf{X}_{k,g} = \mathbf{x}_{k,g} \sim GB2(a_g, \exp(\mathbf{x}_k \boldsymbol{\beta}_g), p_g, q_g)$.
- Thus, for each $k \in U_g$, the GB2 density becomes

$$f[y_k; a_g, \exp(\mathbf{x}_k^\top \boldsymbol{\beta}_g), p_g, q_g] = \frac{a \left[\frac{y_k}{\exp(\mathbf{x}_k^\top \boldsymbol{\beta}_g)} \right]^{a_g p_g - 1}}{\exp(\mathbf{x}_k^\top \boldsymbol{\beta}_g) \mathrm{B}(p_g, q_g) \left\{ 1 + \left[\frac{y_k}{\exp(\mathbf{x}_k^\top \boldsymbol{\beta}_g)} \right]^a \right\}^{p_g + q_g}}.$$

The GB2 distribution



We assume that

$$\log(Y_{k,g}) = \mathbf{X}_{k,g}^{\top}\boldsymbol{\beta}_g + \log(\varepsilon_{k,g}),$$

where $Y_{k,g}$ is the wage of individual $k \in U_g$, $\varepsilon_{k,g} \sim GB2(a_g, 1, p_g, q_g)$.

The GB2 distribution

- We estimate the parameters of the GB2 distribution using pseudo-maximum likelihood.
- We developed an algorithm to estimate the parameters of the GB2 distribution when $\mathbf{x}_k$ is introduced in the scale parameter.
- We estimate the standard errors of the estimated parameters using the sandwich estimator and a parametric bootstrap approach.

Example - Swiss Survey on Earnings Structure, 2012

- sample of 5643 employees (1880 women, 3763 men) working in the economic activity 'Manufacture of computer, electronic and optical products');
- models with the covariates: age, education level (9 categories), professional position (5 categories).

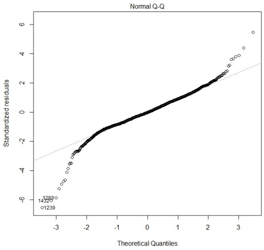


Figure: QQplot for a log-normal model

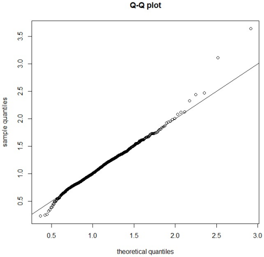
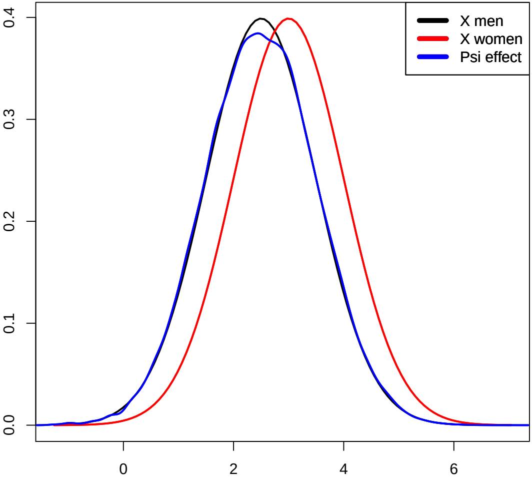


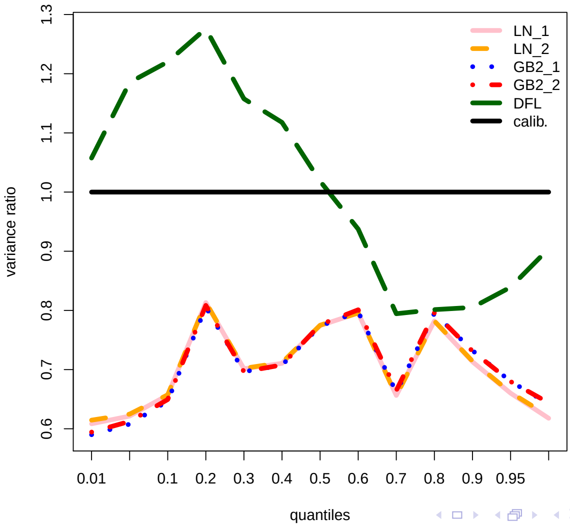
Figure: QQplot for a GB2 model

- $N_F = N_M = 50,000, n_F = n_M = 10,000,$
- $X_{k,F} \sim N(3,1), X_{k,M} \sim N(2.5,1)$, independent,
- $Y_{k,F} \sim LN(1.15 + 2.5X_{k,F}, 1), k \in U_F$
- 1000 independent srswor samples of size n_F from U_F,
- At the population level: $\psi_k = f_{N(2.5,1)}(X_{k,F})/f_{N(3,1)}(X_{k,F}), k \in U_F$.

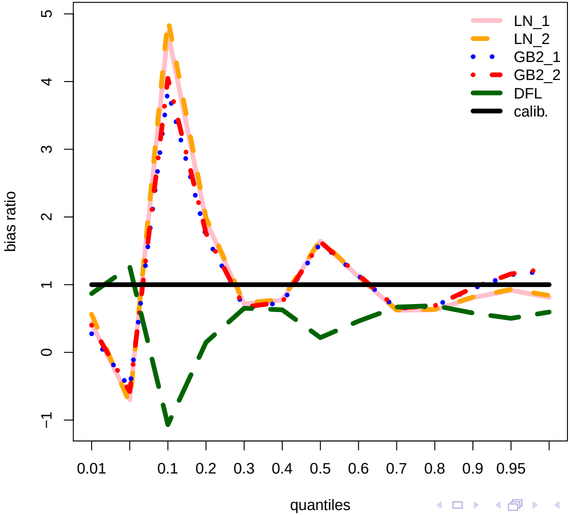
ψ effect



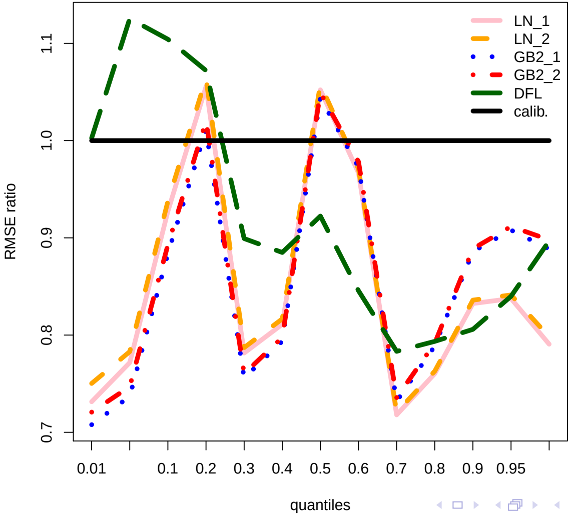
Monte-Carlo variance



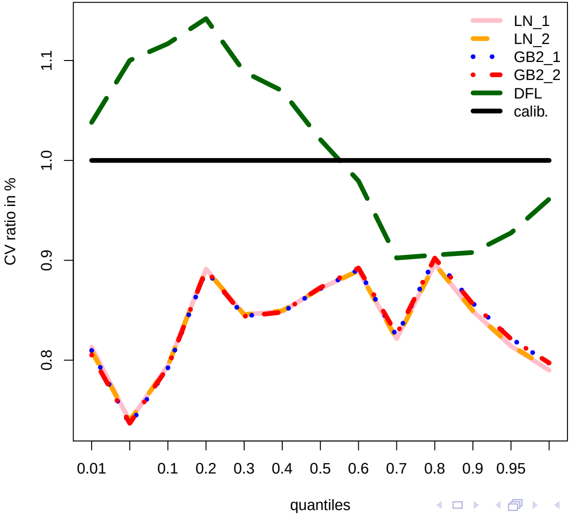
Monte-Carlo bias



Monte-Carlo RMSE



Monte-Carlo CV



Conclusions

- Wages usually have heavy-tailed distributions, which makes it difficult to fit a distribution for them.
- We propose two parametric methods to estimate the quantiles (and differences between the quantiles).
- The introduction of the covariates aims to reduce the variance of the estimates.

References

Anastasiade, M.-C. and Tillé, Y. (2017). Decomposition of gender wage inequalities through calibration: Application to the swiss structure of earnings survey. *Survey Methodology*, 43(2):211–234.

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996). Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. *Econometrica*, 64(5):1001–44.

McDonald, J. B. and Butler, R. J. (1990). Regression models for positive random variables. *Journal of Econometrics*, 43(1-2):227–251.