Estimating a counterfactual wage distribution using survey data

Mihaela-Cătălina Anastasiade\textsuperscript{1}, Alina Matei\textsuperscript{2} and Yves Tillé\textsuperscript{2}

Swiss Federal Statistical Office\textsuperscript{1}, University of Neuchâtel\textsuperscript{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).
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 
\[
p(s) = \Pr(S = s), \quad \forall s \subseteq U.
\]

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).
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 $X_g$ is a set of covariates.

In each subset $U_g$,

$$Y_{k,g} \mid X_g = x_{k,g} \sim D(\gamma_{k,g}, \delta_g), k \in U_g.$$ 

We assume that $D(\gamma_{k,g}, \delta_g), k \in U_g$ is a continuous distribution and that $\gamma_{k,g} = h(x_{k,g}^\top \beta_g)$, 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 CDF of the counterfactual wage distribution is defined as

\[ F^C(y) = \int_{\mathcal{X}_M} F^{Y_F|X_F}(y \mid x) dF^{X_M}(x), \]

where \( \mathcal{X}_M \) is the support of \( X_M \) and \( \mathcal{X}_F \) the support of \( X_F \).

It is assumed that \( \mathcal{X}_M \subseteq \mathcal{X}_F \).

It is interpreted as the distribution function of wages that would be obtained for women if their characteristics were same as those of men.

\[ F^F(y) = \int_{\mathcal{X}_F} F^{Y_F|X_F}(y \mid x) dF^{X_F}(x), \]

\[ F^M(y) = \int_{\mathcal{X}_M} F^{Y_M|X_M}(y \mid x) dF^{X_M}(x). \]
The CDF of the counterfactual wage distribution is defined as

\[
F^C(y) = \int_{\mathcal{X}_M} F_{Y_F|X_F}(y | x) dF_{X_M}(x)
\]

\[
= \int_{\mathcal{X}_F} F_{Y_F|X_F}(y | x) \frac{dF_{X_M}(x)}{dF_{X_F}(x)} dF_{X_F}(x)
\]

\[
= \int_{\mathcal{X}_F} F_{Y_F|X_F}(y | x) \psi(x) dF_{X_F}(x),
\]

where \(\mathcal{X}_M\) is the support of \(X_M\) and \(\mathcal{X}_F\) the support of \(X_F\). It is assumed that \(\mathcal{X}_M = \mathcal{X}_F\).

\[
F^F(y) = \int_{\mathcal{X}_F} F_{Y_F|X_F}(y | x) dF_{X_F}(x),
\]
The weighted DiNardo, Fortin and Lemieux method

- DiNardo et al. (1996) write the reweighting factor \( \psi(x_k) = \frac{dF^X_M(x_k)}{dF^X_F(x_k)} \) as

\[
\psi(x_k) = \psi_k = \frac{P(\text{Gender}_k = \text{‘man’} \mid x_k)/P(\text{Gender}_k = \text{‘man’})}{P(\text{Gender}_k = \text{‘woman’} \mid x_k)/P(\text{Gender}_k = \text{‘woman’})}
\]

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

\[
\hat{X}_C = \hat{X}_M,
\]

where

\[
\hat{X}_C = \sum_{k \in S_F} \hat{\psi}_k w_k x_k / \sum_{k \in S_F} \hat{\psi}_k w_k \quad \text{and} \quad \hat{X}_M = \sum_{k \in S_M} w_k x_k / \sum_{k \in S_M} w_k.
\]

- The factor \( \psi(x_k) \) can be estimated by using a probit or a logistic regression model (DiNardo et al., 1996) or by calibration (Anastasiade and Tillé, 2017).
If Δ(α) 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 \( \alpha \) of men, counterfactual and women wage distributions, respectively.

\[ \hat{\Delta}(\alpha) = \hat{Q}^M(\alpha) - \hat{Q}^F(\alpha) = (\hat{Q}^M(\alpha) - \hat{Q}^C(\alpha)) + (\hat{Q}^C(\alpha) - \hat{Q}^F(\alpha)). \]
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

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

The quantile of order $\alpha$ of $y$ is estimated by

$$\hat{Q}_{\alpha, emp}(y) = \inf\{\hat{F}_{emp}(y) \geq \alpha\}.$$
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 $X$ in the estimation of $F(.)$ by assuming that $Y_k \mid X_k \sim D(\gamma_k = h(x_k^\top \beta), \delta_k), k \in U$.
- We write

$$F_U(y) = \sum_{k \in U} \lambda_k F_D(\gamma_k, \delta)(y \mid x_k, g) = \frac{1}{N} \sum_{k \in U} F_D(\gamma_k, \delta)(y \mid x_k, g),$$

where $\lambda_k = 1/N$, $F_D(\gamma_k, \delta)(\cdot \mid x_k)$ is the CDF of the distribution $D(\gamma_k = h(x_k^\top \beta), \delta_k), k \in U$, and $h(.)$ is a continuous function.
We propose to estimate the quantile of order $\alpha$ of $Y$ as

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

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

$$
\hat{F}_U(y) = \sum_{k \in S} w_k \hat{F}_{D(\hat{\gamma}_k, \delta)}(y_k \mid x_k) / \sum_{k \in S} w_k.
$$
Method 2 for quantile estimation

In case the inverse function of $\hat{F}_U(y)$ cannot be computed, we propose to use a Monte Carlo method based on parametric bootstrap.
Method 2 for quantile estimation

\[ Y_{i,k} \mid x_k \sim D(h(x_k^\top \hat{\beta}), \hat{\delta}) \]

\[
\begin{pmatrix}
  w_1 & w_2 & w_3 & \ldots & w_n \\
  y_{11} & y_{12} & y_{13} & \ldots & y_{1n} \\
  y_{21} & y_{22} & y_{23} & \ldots & y_{2n} \\
  y_{31} & y_{32} & y_{33} & \ldots & y_{3n} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  y_{m1} & y_{m2} & y_{m3} & \ldots & y_{mn}
\end{pmatrix}
\]

\[ \rightarrow \]

\[
\begin{pmatrix}
  \hat{Q}_{1}^{0.25} & \hat{Q}_{1}^{0.5} & \hat{Q}_{1}^{0.75} & \hat{Q}_{1}^{0.95} \\
  \hat{Q}_{2}^{0.25} & \hat{Q}_{2}^{0.5} & \hat{Q}_{2}^{0.75} & \hat{Q}_{2}^{0.95} \\
  \hat{Q}_{3}^{0.25} & \hat{Q}_{3}^{0.5} & \hat{Q}_{3}^{0.75} & \hat{Q}_{3}^{0.95} \\
  \vdots & \vdots & \vdots & \vdots \\
  \hat{Q}_{m}^{0.25} & \hat{Q}_{4}^{0.5} & \hat{Q}_{m}^{0.75} & \hat{Q}_{4}^{0.95} \\
  \hat{Q}_{0.25} & \hat{Q}_{0.5} & \hat{Q}_{0.75} & \hat{Q}_{0.95}
\end{pmatrix}
\]

**Remark:** Methods 1 and 2 are applied to estimate respectively the \( \alpha \)-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

$$F^C_{U_F}(y) = \frac{1}{N_C} \sum_{k \in U_F} \psi_k F(Y_F|X_F)(y_k | x_{k,F}),$$

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

- First, we estimate it by

$$\hat{F}^C_{U_F}(y) = \frac{\sum_{k \in S_F} \hat{\psi}_k w_k \hat{F}(Y_F|X_F)(y_k | x_{k,F})}{\sum_{k \in S_F} \hat{\psi}_k w_k},$$

where $\hat{F}(Y_F|X_F)(y_k | x_{k,F}) = F_{D(h(x_{k,F}^\top \hat{\beta}_F), \hat{\delta}_F)}(y_k | x_{k,F})$, and $\hat{\psi}_k$ is estimated by calibration (Anastasiade and Tillé, 2017).

- Next, $\hat{Q}^C_{(\alpha)} = \inf\{y \mid \hat{F}^C_{U_F}(y) \geq \alpha\}$. 
Method 2 for the counterfactual wage distribution

\[
Y_{i,k} | x_{k,F} \sim D(h(x_{k,F}^\top \hat{\beta}_F), \hat{\delta}_F)
\]

\[
\begin{pmatrix}
\hat{\psi}_1 w_1 & \hat{\psi}_2 w_2 & \ldots & \hat{\psi}_{n_F} w_{n_F} \\
y_{11} & y_{12} & \ldots & y_{1n_F} \\
y_{21} & y_{22} & \ldots & y_{2n_F} \\
y_{31} & y_{32} & \ldots & y_{3n_F} \\
\vdots & \vdots & \ddots & \vdots \\
y_{m1} & y_{m2} & \ldots & y_{mn_F}
\end{pmatrix} \rightarrow \begin{pmatrix}
\hat{Q}_{0.25}^C & \hat{Q}_{0.5}^C & \hat{Q}_{0.75}^C & \hat{Q}_{0.95}^C \\
\hat{Q}_1 & \hat{Q}_1 & \hat{Q}_1 & \hat{Q}_1 \\
\hat{Q}_2 & \hat{Q}_2 & \hat{Q}_2 & \hat{Q}_2 \\
\hat{Q}_3 & \hat{Q}_3 & \hat{Q}_3 & \hat{Q}_3 \\
\vdots & \vdots & \vdots & \vdots \\
\hat{Q}_m & \hat{Q}_4 & \hat{Q}_m & \hat{Q}_4 \\
\hat{C} & \hat{C} & \hat{C} & \hat{C}
\end{pmatrix}
\]
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

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

where $a$, $p$, $q$ are the shape parameters and $b$ is a scale parameter.
Example of GB2 distribution

\[ a=8 \]
\[ b=10 \]
\[ p=0.5 \]
\[ q=0.6 \]
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 $k \in U_g$, $Y_k \mid X_{k,g} = x_{k,g} \sim GB2(a_g, \exp(x_{k,g} \beta_g), p_g, q_g)$.

Thus, for each $k \in U_g$, the GB2 density becomes

$$f[y_k; a_g, \exp(x_k^\top \beta_g), p_g, q_g] = \frac{a \left[ \frac{y_k}{\exp(x_k^\top \beta_g)} \right]^{a_g p_g - 1}}{\exp(x_k^\top \beta_g) B(p_g, q_g) \left\{ 1 + \left[ \frac{y_k}{\exp(x_k^\top \beta_g)} \right]^a \right\}^{p_g + q_g}}.$$
The GB2 distribution

Change in $b$

$a=8, p=0.5, q=0.6$
We assume that
\[
\log(Y_{k,g}) = X_{k,g}^\top \beta_g + \log(\varepsilon_{k,g}),
\]
where \( Y_{k,g} \) is the wage of individual \( k \in U_g, \varepsilon_{k,g} \sim GB2(a_g, 1, p_g, q_g) \).
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 $x_k$ 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
Monte-Carlo simulation

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

![Monte-Carlo variance diagram](image)

- LN_1
- LN_2
- GB2_1
- GB2_2
- DFL
- calib.

Quantiles

<table>
<thead>
<tr>
<th>Variance Ratio</th>
<th>0.01</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>LN_1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LN_2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GB2_1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GB2_2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DFL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>calib.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Quantiles from 0.01 to 0.95 are plotted against variance ratio for different distributions.
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

