Pontificia Universidad Católica de Chile
Facultad de Economía & Administración
Instituto de Economía

Problem Set # 2

Jose Carlo Bermúdez

jcbermudez@uc.cl

Econometric Theory II

Professor: Tomás Rau

May 10th, 2024

Summary

This report includes my solution to the problem set # 2 for the Econometric Theory II graduate course. The first section of the report dives into the theoretical fundamentals behind the generalized method of moments (GMM) in the context of censored and truncated models. The second part of the document relies on different empirical settings to apply techniques such as counting models, survival analysis, and GMM. The replication package for the empirical analysis is available at my GitHub account. The Weibull distribution in exercise #2 does not converge, thus, to replicate the code interruptedly, I strongly suggest running the do-file "$0\_master.do$" at once, and not every code line by line.

Contents

1 Theoretical Problem

Estimation of the truncated regression model by GMM. The truncated regression model can also be estimated by GMM. Consider the following $K+1$ orthogonality conditions. The first K of them are:

$$\mathbf{g_1}\left(\mathbf{w}_t;\boldsymbol{\delta},\gamma\right)=\left(y_t-\frac{1}{\gamma}\mathbf{x}_t'\boldsymbol{\delta}-\frac{\lambda\left(v_t\right)}{\gamma}\right)\cdot\mathbf{x}_t \tag{1}$$

where $\mathbf{w}_t=(y_t,\mathbf{x}_t')'$ and $v_t\equiv\gamma c-\mathbf{x}_t'\boldsymbol{\delta}$. The $(K+1)$-th orthogonality condition is:

$$g_2\left(\mathbf{w}_t;\boldsymbol{\delta},\gamma\right)=y_t^2-\frac{1}{\gamma^2}\left[1+\lambda\left(v_t\right)\gamma c+\left(\mathbf{x}_t'\boldsymbol{\delta}\right)^2+\lambda\left(v_t\right)\mathbf{x}_t'\boldsymbol{\delta}\right] \tag{2}$$

1. Verify that $\mathbb{E}\left[\mathbf{g}_1\left(\mathbf{w}_t;\boldsymbol{\delta}_0,\gamma_0\right)\right]=\mathbf{0}$ and $\mathbb{E}\left[g_2\left(\mathbf{w}_t;\boldsymbol{\delta}_0,\gamma_0\right)\right]=0$.

   Solution[1]**.** My strategy for proving each orthogonality condition consists of two steps: i) First, I will find general closed-form expressions for the expectations $\mathbb{E}(y|\mathbf{x})$, $\mathbb{E}(y^2|\mathbf{x})$ assuming y is a truncated random variable; ii) I will combine these closed-form expressions found in the previous step and the law of total expectations for the expected values and prove the conditions requested by the problem.

   - Let $\mathbf{x}\sim N\left(\mu,\sigma^2\right)$, then let y be a truncated normal $Y\sim TN\left(\mu,\sigma^2,a,b\right)$ random variable. Then, we know that we can represent its distribution function as:

$$f_Y(y)=\frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)}\mathbb{1}_{[a,b]}(y)=\frac{\phi(\frac{a-\mu}{\sigma})}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)}\mathbb{1}_{[a,b]}(y) \tag{3}$$

   where $a,b\in\mathbb{R}$. The indicator function $\mathbb{1}_{[a,b]}(y)=1$ if $a\leq y\leq b$, and is zero otherwise. Let ϕ be the standard normal probability distribution function, and Φ is the standard normal cumulative distribution function. Let α be a convenient expression such that:

$$\alpha=\frac{1}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)}$$

   Then, according to the definition of expectation we have:

$$E(y)=\int_a^b yf_Y(y)dy$$

   Hence

$$\begin{aligned}\frac{\mathbb{E}(y)}{\alpha}&=\int_a^b y\cdot\left\{\frac{1}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)}\right\}^{-1}\cdot\frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)}dy\\&=\int_a^b\frac{y}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\}dy\\&=\int_a^b\left(\frac{y-\mu}{\sigma}\right)\frac{1}{\sqrt{2\pi}}\exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\}dy+\frac{\mu}{\sigma}\frac{1}{\sqrt{2\pi}}\int_a^b\exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\}dy\\&=\int_a^b\left(\frac{y-\mu}{\sigma}\right)\frac{1}{\sqrt{2\pi}}\exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\}dy+\mu\int_a^b\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\}dy\end{aligned}$$

   With the integrand of the last integral being the pdf of a $N\left(\mu,\sigma^2\right)$ distribution. Let z =

[1] For this proof, I will follow Professor David Olive notes very closely.

$(y-\mu)/\sigma$. Thus $dz = dy/\sigma$, and

$$\begin{aligned}
\frac{\mathbb{E}(y)}{\alpha} &= \frac{\mu}{\alpha} + \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \sigma \frac{z}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \\
&= \frac{\mu}{\alpha} + \frac{\sigma}{\sqrt{2\pi}} \left(-e^{\frac{-z^2}{2}} \right) \Bigg|_{\frac{b-\mu}{\sigma}}^{\frac{a-\mu}{\sigma}} \\
&= \frac{\mu}{\alpha} + \left\{ \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{a-\mu}{\sigma}} - \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{b-\mu}{\sigma}} \right\} \\
&= \frac{\mu}{\alpha} + \sigma \left\{ \phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right) \right\}
\end{aligned}$$

Solving the expression above and multiplying both sides by α we have:

$$\mathbb{E}(y) = \mu + \frac{\sigma \left\{ \phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right) \right\}}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

The latter expression constitutes the first unconditional moment for a random variable that is truncated normal. Notice that when $b = +\infty$, then $\phi(.) = 0$ and $\Phi(.) = 1$ we get:

$$\mathbb{E}(y) = \mu + \sigma \cdot \frac{\phi\left(\frac{a-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{a-\mu}{\sigma}\right)} \tag{4}$$

Next, let's find an expression for $\mathbb{E}(y^2)$. Again, following the definition for the expected value:

$$\mathbb{E}\left(y^2\right) = \int_a^b y^2 f_Y(y) dy$$

Hence

$$\begin{aligned}
\frac{\mathbb{E}\left(y^2\right)}{\alpha} &= \int_a^b y^2 \cdot \left\{ \frac{1}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} \right\}^{-1} \cdot \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} dy \\
& \int_a^b \frac{y^2}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-(y-\mu)^2}{2\sigma^2} \right\} dy \\
&= \sigma \int_a^b \left(\frac{y^2}{\sigma^2} - \frac{2\mu y}{\sigma^2} + \frac{\mu^2}{\sigma^2} \right) \frac{1}{\sqrt{2\pi}} \exp\left\{ \frac{-(y-\mu)^2}{2\sigma^2} \right\} dy ... \\
& ... + \sigma \int_a^b \frac{2y\mu - \mu^2}{\sigma^2} \frac{1}{\sqrt{2\pi}} \exp\left\{ \frac{-(y-\mu)^2}{2\sigma^2} \right\} dy \\
&= \sigma \int_a^b \left(\frac{y-\mu}{\sigma} \right)^2 \frac{1}{\sqrt{2\pi}} \exp\left\{ \frac{-(y-\mu)^2}{2\sigma^2} \right\} dy + \frac{2\mu}{\alpha} \mathbb{E}(y) - \frac{\mu^2}{\alpha}
\end{aligned}$$

Let $z = (y-\mu)/\sigma$. Then $dz = dy/\sigma, dy = \sigma dz$, and $y = \sigma z + \mu$. Thus

$$\frac{\mathbb{E}\left(y^2\right)}{\alpha} = \frac{2\mu}{\alpha} \mathbb{E}(y) - \frac{\mu^2}{\alpha} + \sigma \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \sigma \frac{z^2}{\sqrt{2\pi}} e^{-z^2/2} dz$$

Next integrate by parts with $w = z$ and $dv = ze^{-z^2/2}dz$. Then:

$$\begin{aligned}\frac{\mathbb{E}(y^2)}{\alpha} &= \frac{2\mu}{\alpha}\mathbb{E}(y) - \frac{\mu^2}{\alpha} + \frac{\sigma^2}{\sqrt{2\pi}}\left[\left(-ze^{-z^2/2}\right)\Big|_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} + \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} e^{-z^2/2}dz\right] \\ &= \frac{2\mu}{\alpha}\mathbb{E}(y) - \frac{\mu^2}{\alpha} + \sigma^2\left[\left(\frac{a-\mu}{\sigma}\right)\phi\left(\frac{a-\mu}{\sigma}\right) - \left(\frac{b-\mu}{\sigma}\right)\phi\left(\frac{b-\mu}{\sigma}\right) + \frac{1}{\alpha}\right]\end{aligned}$$

Again, multiplying by α:

$$\mathbb{E}(y^2) = 2\mu\mathbb{E}(y) - \mu^2 + \sigma^2\left[\frac{\left(\frac{a-\mu}{\sigma}\right)\phi\left(\frac{a-\mu}{\sigma}\right) - \left(\frac{b-\mu}{\sigma}\right)\phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} + 1\right]$$

Making $b = +\infty$, and replacing Equation 4 in $\mathbb{E}(y)$ we get:

$$\mathbb{E}(y^2) = \mu^2 + \sigma^2 + \mu\sigma\frac{\phi\left(\frac{a-\mu}{\sigma}\right)}{1-\Phi\left(\frac{a-\mu}{\sigma}\right)} + a\sigma\frac{\phi\left(\frac{a-\mu}{\sigma}\right)}{1-\Phi\left(\frac{a-\mu}{\sigma}\right)} \tag{5}$$

- Equation 4 and Equation 5 offer closed-form representations for expected values of random variables with truncated normal distribution. Now let us focus on our example case. Suppose we have an i.i.d. sample $\{y_t, \mathbf{x}_t\}$ satisfying

$$y_t = \mathbf{x}_t'\boldsymbol{\beta}_0 + \varepsilon_t; \quad \varepsilon_t|\mathbf{x}_t \sim N(0, \sigma_0^2)$$

Before truncation the distribution takes the form $y_t|\mathbf{x}_t \sim N(\mathbf{x}_t'\boldsymbol{\beta}_0, \sigma_0)$. And the t-th observation is in the sample if and only if $y_t > c$, with c any constant. Thus, replacing $a = c$, $\mu = \mathbf{x}_t'\boldsymbol{\beta}$, and $\sigma = \sigma_0$ in Equation 4 and Equation 5 we get the following expressions when t-th observation is in the sample:

$$\mathbb{E}(y_t|\mathbf{x}_t) = \mathbf{x}_t'\boldsymbol{\beta}_0 + \sigma_0\frac{\phi\left(\frac{c-\mathbf{x}_t'\boldsymbol{\beta}_0}{\sigma_0}\right)}{1-\Phi\left(\frac{c-\mathbf{x}_t'\boldsymbol{\beta}_0}{\sigma_0}\right)}$$

$$\mathbb{E}(y_t^2|\mathbf{x}_t) = (\mathbf{x}_t'\boldsymbol{\beta}_0)^2 + \sigma_0^2 + (\mathbf{x}_t'\boldsymbol{\beta}_0)\sigma_0\frac{\phi\left(\frac{c-\mathbf{x}_t'\boldsymbol{\beta}_0}{\sigma_0}\right)}{1-\Phi\left(\frac{c-\mathbf{x}_t'\boldsymbol{\beta}_0}{\sigma_0}\right)} + c\sigma_0\frac{\phi\left(\frac{c-\mathbf{x}_t'\boldsymbol{\beta}_0}{\sigma_0}\right)}{1-\Phi\left(\frac{c-\mathbf{x}_t'\boldsymbol{\beta}_0}{\sigma_0}\right)}$$

After some convenient re-arrangements, we arrive at:

Conditional expectations for truncated random variable y_t

$$\mathbb{E}(y_t|\mathbf{x}_t) = \frac{\mathbf{x}_t'\boldsymbol{\delta}}{\gamma} + \sigma_0\lambda(\upsilon_t) \tag{6}$$

$$\mathbb{E}(y_t^2|\mathbf{x}_t) = \left(\frac{\mathbf{x}_t'\boldsymbol{\delta}}{\gamma}\right)^2 + \sigma_0^2 + \lambda(\upsilon_t)\frac{\mathbf{x}_t'\boldsymbol{\delta}}{\gamma^2} + \frac{c}{\gamma}\lambda(\upsilon_t) \tag{7}$$

Where $\boldsymbol{\delta} = \frac{\boldsymbol{\beta}_0}{\sigma_0}$; $\gamma = \frac{1}{\sigma_0}$, and $\upsilon_t \equiv \gamma c - \mathbf{x}_t'\boldsymbol{\delta}$. Notice that $\lambda(\upsilon_t)$ is the inverse Mill's ratio.

- Finally, we can prove orthogonality conditions in Equation 1 and Equation 2 using the law of total expectation as follows:

$$
\begin{aligned}
\mathbb{E}\left[\mathbf{g_1}\left(\mathbf{w}_t;\boldsymbol{\delta},\gamma\right)\right] &= \mathbb{E}\left\{\mathbb{E}\left[y_t\cdot\mathbf{x}_t-\frac{1}{\gamma}\mathbf{x}_t'\boldsymbol{\delta}\cdot\mathbf{x}_t-\frac{\lambda\left(\upsilon_t\right)}{\gamma}\cdot\mathbf{x}_t\middle|\mathbf{x}_t\right]\right\}\\
&= \mathbb{E}\left\{\mathbb{E}(y_t|\mathbf{x}_t)\cdot\mathbf{x}_t-\mathbb{E}\left(\frac{1}{\gamma}\mathbf{x}_t'\boldsymbol{\delta}\cdot\mathbf{x}_t+\frac{\lambda\left(\upsilon_t\right)}{\gamma}\cdot\mathbf{x}_t\middle|\mathbf{x}_t\right)\right\}\\
&= \mathbb{E}\left\{\left(\underbrace{\frac{\mathbf{x}_t'\boldsymbol{\delta}}{\gamma}+\sigma_0\lambda(\upsilon_t)}_{\text{Equation 6}}-\frac{1}{\gamma}\mathbf{x}_t'\boldsymbol{\delta}-\frac{\lambda\left(\upsilon_t\right)}{\gamma}\right)\cdot\mathbf{x}_t\right\}\quad\text{with } \sigma_0=\frac{1}{\gamma}\\
&= \mathbb{E}\left\{\left(\cancel{\frac{\mathbf{x}_t'\boldsymbol{\delta}}{\gamma}}+\cancel{\frac{\lambda\left(\upsilon_t\right)}{\gamma}}-\cancel{\frac{1}{\gamma}\mathbf{x}_t'\boldsymbol{\delta}}-\cancel{\frac{\lambda\left(\upsilon_t\right)}{\gamma}}\right)\cdot\mathbf{x}_t\right\}\\
&= \cancelto{0}{\mathbb{E}\{0\cdot\mathbf{x}_t\}}\\
\therefore\ \mathbb{E}\left[\mathbf{g_1}\left(\mathbf{w}_t;\boldsymbol{\delta},\gamma\right)\right] &= 0
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left[g_2\left(\mathbf{w}_t;\boldsymbol{\delta},\gamma\right)\right] &= \mathbb{E}\left\{\mathbb{E}\left[y_t^2-\frac{1}{\gamma^2}\left(1+\lambda\left(\upsilon_t\right)\gamma c+\left(\mathbf{x}_t'\boldsymbol{\delta}\right)^2+\lambda\left(\upsilon_t\right)\mathbf{x}_t'\boldsymbol{\delta}\right)\right]\right\}\\
&= \mathbb{E}\left\{\mathbb{E}(y_t^2|\mathbf{x}_t)-\mathbb{E}\left(\frac{1}{\gamma^2}+\frac{\lambda\left(\upsilon_t\right)c}{\gamma}+\left(\frac{\mathbf{x}_t'\boldsymbol{\delta}}{\gamma}\right)^2+\frac{\lambda\left(\upsilon_t\right)\mathbf{x}_t'\boldsymbol{\delta}}{\gamma^2}\middle|\mathbf{x}_t\right)\right\}\\
&= \mathbb{E}\Bigg\{\underbrace{\left(\frac{\mathbf{x}_t'\boldsymbol{\delta}}{\gamma}\right)^2+\sigma_0^2+\frac{\lambda(\upsilon_t)\mathbf{x}_t'\boldsymbol{\delta}}{\gamma^2}+\frac{c\lambda(\upsilon_t)}{\gamma}}_{\text{Equation 7}}-\frac{1}{\gamma^2}-\frac{\lambda\left(\upsilon_t\right)c}{\gamma}-\left(\frac{\mathbf{x}_t'\boldsymbol{\delta}}{\gamma}\right)^2-\frac{\lambda\left(\upsilon_t\right)\mathbf{x}_t'\boldsymbol{\delta}}{\gamma^2}\Bigg\}\\
&\text{with}\quad \sigma_0=\frac{1}{\gamma}\\
&= \mathbb{E}\left\{\cancel{\left(\frac{\mathbf{x}_t'\boldsymbol{\delta}}{\gamma}\right)^2}+\cancel{\frac{1}{\gamma^2}}+\cancel{\frac{\lambda(\upsilon_t)\mathbf{x}_t'\boldsymbol{\delta}}{\gamma^2}}+\cancel{\frac{c\lambda(\upsilon_t)}{\gamma}}-\cancel{\frac{1}{\gamma^2}}-\cancel{\frac{\lambda\left(\upsilon_t\right)c}{\gamma}}-\cancel{\left(\frac{\mathbf{x}_t'\boldsymbol{\delta}}{\gamma}\right)^2}-\cancel{\frac{\lambda\left(\upsilon_t\right)\mathbf{x}_t'\boldsymbol{\delta}}{\gamma^2}}\right\}\\
&= \cancelto{0}{\mathbb{E}\{0\}}\\
\therefore\ \mathbb{E}\left[g_2\left(\mathbf{w}_t;\boldsymbol{\delta},\gamma\right)\right] &= 0
\end{aligned}
$$

2. Show that $(\boldsymbol{\delta}^*, \gamma^*)$ satisfies the likelihood equations if and only if it satisfies the $K+1$ orthogonality conditions. *Hint:* Let $s_2(\mathbf{w}_t; \boldsymbol{\delta}, \gamma)$ be the $(K+1)$-th element of $\mathbf{s}(\mathbf{w}_t; \boldsymbol{\delta}, \gamma)$ in the score of the likelihood function. Then:

$$s_2(\mathbf{w}_t; \boldsymbol{\delta}, \gamma) + \gamma g_2(\mathbf{w}_t; \boldsymbol{\delta}, \gamma) = (\mathbf{x}_t' \boldsymbol{\delta}) \left(y_t - \frac{1}{\gamma} \mathbf{x}_t' \boldsymbol{\delta} - \frac{\lambda(v_t)}{\gamma} \right) \tag{8}$$

Solution. According to Equation 3, when the t-th observation is in the sample and given β_0, σ_0^2 the post-truncation density defined over $(c, +\infty)$ is:

$$f(y_t | \mathbf{x}_t) = \frac{\frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(\frac{-(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2}{2\sigma_0^2} \right)}{1 - \Phi\left(\frac{c - \mathbf{x}_t'\boldsymbol{\beta}}{\sigma_0} \right)}$$

This is the density (conditional on $\mathbf{x}_t$) of observation t in the sample. Taking logs and replacing (β_0, σ_0^2) by their hypothetical values (β, σ^2), we obtain the log-likelihood for observation t

$$\log f\left(y_t \mid \mathbf{x}_t; \boldsymbol{\beta}, \sigma^2\right) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma_0^2\right) - \frac{1}{2}\left(\frac{y_t - \mathbf{x}_t'\boldsymbol{\beta}}{\sigma_0}\right)^2 - \log\left[1 - \Phi\left(\frac{c - \mathbf{x}_t'\boldsymbol{\beta}}{\sigma_0}\right)\right]$$

The parametrized version is given by

$$\log \tilde{f}(y_t \mid \mathbf{x}_t; \lambda) = \left[-\frac{1}{2}\log(2\pi) + \log(\gamma) - \frac{1}{2}\left(\gamma y_t - \mathbf{x}_t'\boldsymbol{\delta}\right)^2\right] - \log\left[1 - \Phi\left(\gamma c - \mathbf{x}_t'\delta\right)\right]$$

Now, we get the score as the first derivative for the log-likelihood above

$$\underset{((K+1)\times 1)}{\mathbf{s}(\mathbf{w}_t; \boldsymbol{\delta}, \gamma)} \equiv \begin{bmatrix} \frac{\partial \log \tilde{f}(y_t | \mathbf{x}_t; \lambda)}{\partial \delta} \\ \frac{\partial \log \tilde{f}(y_t | \mathbf{x}_t; \lambda)}{\partial \gamma} \end{bmatrix} = \begin{bmatrix} (\gamma y_t - \mathbf{x}_t'\boldsymbol{\delta})\,\mathbf{x}_t \\ \frac{1}{\gamma} - (\gamma y_t - \mathbf{x}_t'\boldsymbol{\delta})\, y_t \end{bmatrix} + \lambda(v_t) \begin{bmatrix} -\mathbf{x}_t \\ c \end{bmatrix}$$

Let $s_2(\mathbf{w}_t; \boldsymbol{\delta}, \gamma)$ be the the $K+1$-th element of the score, thus:

$$s_2(\mathbf{w}_t; \boldsymbol{\delta}, \gamma) = \frac{1}{\gamma} - (\gamma y_t - \mathbf{x}_t'\boldsymbol{\delta})\, y_t + c\lambda(v_t)$$

Now, is direct to verify the equality 8:

$$\underbrace{\frac{1}{\gamma} - \gamma y_t^2 + (\mathbf{x}_t'\boldsymbol{\delta}) y_t + c\lambda(v_t)}_{s_2(\mathbf{w}_t;\boldsymbol{\delta},\gamma)} = (\mathbf{x}_t'\boldsymbol{\delta})\left(y_t - \frac{1}{\gamma}\mathbf{x}_t'\boldsymbol{\delta} - \frac{\lambda(v_t)}{\gamma}\right) - \gamma \underbrace{\left(y_t^2 - \frac{1}{\gamma^2} - \frac{\lambda(v_t)\,c}{\gamma} - \left(\frac{\mathbf{x}_t'\boldsymbol{\delta}}{\gamma}\right)^2 - \frac{\lambda(v_t)\,\mathbf{x}_t'\boldsymbol{\delta}}{\gamma^2}\right)}_{g_2(\mathbf{w}_t;\boldsymbol{\delta},\gamma)}$$

$$\iff \frac{1}{\gamma} - \gamma y_t^2 + (\mathbf{x}_t'\boldsymbol{\delta}) y_t + c\lambda(v_t) = (\mathbf{x}_t'\boldsymbol{\delta})\, y_t - \cancel{\frac{(\mathbf{x}_t'\boldsymbol{\delta})^2}{\gamma}} - \cancel{\frac{\lambda(v_t)\,\mathbf{x}_t'\boldsymbol{\delta}}{\gamma}} - \gamma y_t^2 + \frac{1}{\gamma} + c\lambda(v_t) + \cancel{\frac{(\mathbf{x}_t'\boldsymbol{\delta})^2}{\gamma}} + \cancel{\frac{\lambda(v_t)\,\mathbf{x}_t'\boldsymbol{\delta}}{\gamma}}$$

$$\iff \frac{1}{\gamma} - \gamma y_t^2 + (\mathbf{x}_t'\boldsymbol{\delta}) y_t + c\lambda(v_t) = (\mathbf{x}_t'\boldsymbol{\delta})\, y_t - \gamma y_t^2 + \frac{1}{\gamma} + c\lambda(v_t)$$

As expected, the equality 8 holds. An important thing to highlight is that we must be careful with dimensions when it comes to the score of the maximum likelihood that is usually expressed in matrix format. However, in this example case, we did not receive any information to impose restrictions on dimensionality. Hence, we can directly conclude that $(\boldsymbol{\delta}^*, \gamma^*)$ satisfies the likelihood equations.

2 Counting Models

In this exercise, we are going to dive into counting and conditional mean models. For this purpose, we have records of crimes committed in a given period for a sample of men between 20 and 45 years old.

2.1 Discussion & Summary Statistics

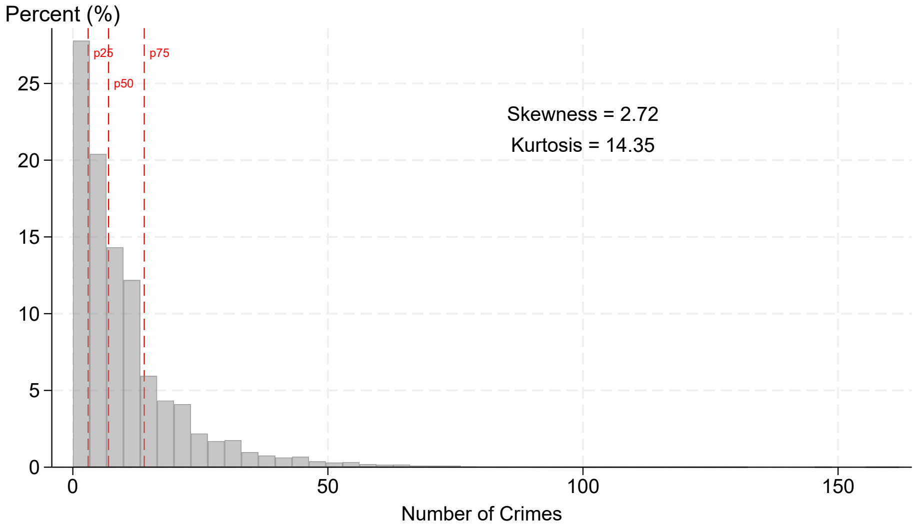
Why should we apply a count model? We have discussed non-linear discrete models where the dependent variable is categorical and usually represents one or more individual choices according to a utility-maximizing framework. Nonetheless, sometimes we might be interested in studying effects on outcomes that are not necessarily categorical but non-negative ordinary integers that represent counts of an event. These are known as models of count data. Cameron & Trivedi (2005) point out that "regression models for counts, like other limited or discrete dependent variable models such as the logit and probit, are nonlinear with many properties and special features intimately connected to discreteness and nonlinearity" (pp. 665).

Count models have some differentiating features. For instance, the dependent variable is discrete and represents the count of realized events of equal probability $\{0, 1, 2, ...\}$, which imposes restrictions on the functional form of $\mathbb{E}(y|x)$, it is non-negative and discreet. Thus, a linear model would produce predictions out of range (i.e. they could be negative). Moreover, in these types of models, the sample is concentrated on a few small discrete values, the data can be skewed to the right, and observations are intrinsically heteroskedastic with variance increasing with the mean. In some cases a high proportion of zeros in the sample may coexist with very large values of counts, creating a difficult modeling challenge. The modeling challenge is to select a functional form that can adequately capture the large mean and the high proportion of zeros. In many other examples, virtually all the data are restricted to single digits, and the mean number of events is quite low (Cameron & Trivedi, 2005, Chapter 20).

The facts discussed above motivate the implementation of count models for our example case. Figure 1 shows the raw distribution for our dependent variable of interest, the count of crimes realized by men. We can observe a highly skewed and asymmetrical distribution. This becomes even more evident by looking at the interquartile range. While the median number of offenses is 4, the percentile 75th is 14, with observations for more than 100 crimes committed by men. In sum, these characteristics suggest that the most reasonable approach to modeling the number of crimes (as a function of individual-specific covariates) is by implementing count data regressions, whether under a fully parametric approach where the restriction of non-negative integer values will be respected or by mean-variance approach, which specifies the conditional mean to be nonnegative and specifies the conditional variance to be a function of the conditional mean.

Summary Statistics. Becker (1968) set the building blocks to understand the economics behind criminal behavior. Recent literature on crime has studied the individual characteristics that predict criminal behavior and has mainly focused on race, family background (Eriksson et al., 2016), or even the interaction between criminals and victims in urban environments (Dominguez, 2022). In our empirical exercise, we explore the determinants of crime committed by males. Table 1 displays summary statistics for our set of outcomes. Stats for dependent variable "offenses" immediately confirm the graphical intuition we got from Figure 1, that is, an outcome with a variance that is substantially larger than its mean, which is a symptom of overdispersion (I am going to discuss this more extensively in a few lines down below). While there are males without any crime in their records, there is at least one person with 162 crimes committed. On the other hand, descriptives for our set of covariates also shed light on interesting traits. For instance, our sample is composed of 44% married males and 67% of individuals in our data live in an urban city. Males in our sample are, on average, 40 years old, but with a lot of dispersion in their age. We also have age squared to capture non-linearities in crime variation across the life cycle.

Figure 1: Distribution for counting of crimes



Note: This figure reports the histogram for counting the number of offenses in our dataset. Dashed red lines display the 25th, 50th, and 95th percentiles of the distribution, respectively. The y-axis is percentages, so each bar is interpreted as the proportion of observations located in each bin.

Table 1: Summary Statistics

	Mean	Var	SD	Min	Median	Max	N
Offenses	10.63	142.03	11.92	0.00	7.00	162.00	83,470
Covariates							
$\mathbb{1}(\text{Married} = 1)$	0.44	0.25	0.50	0.00	0.00	1.00	83,470
Age	40.02	132.71	11.52	20.00	40.03	60.00	83,470
Age^2	1,734.05	863,661.25	929.33	400.00	1,602.47	3,599.96	83,470
$\mathbb{1}(\text{Urban} = 1)$	0.67	0.22	0.47	0.00	1.00	1.00	83,470

Note: This table reports summary statistics for all variables included in the dataset.

Potential Sign of Coefficients. Before moving to the empirical estimations, I argue some beliefs regarding potential directions in the relationship between the number of offenses males commit and their individual-specific traits. To frame my intuition, I rely on Figure 2 which displays a matrix with crossed Pearson coefficients. The redder the box, the greater the magnitude of the correlation.

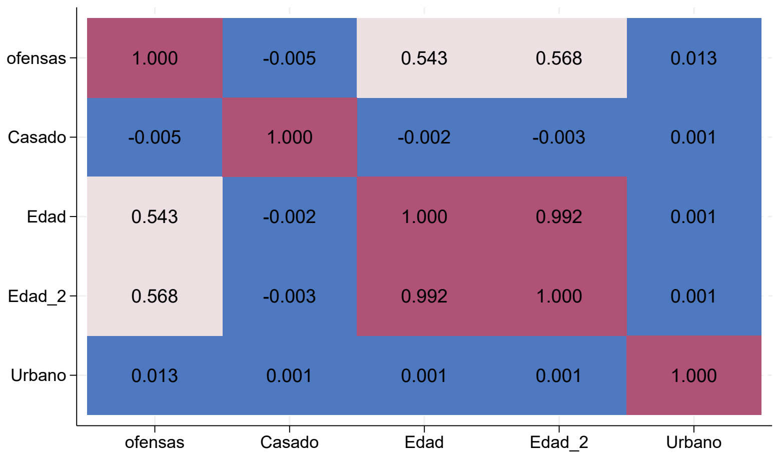
- *Being married.* It is expected that married males would be less likely to commit crimes. This intuition is somewhat supported by the sign of the Pearson coefficient in our data, with a coefficient very close to zero but positive anyway. The intuition is that married men tend to have higher levels of wealth accumulation in terms of their family which usually involves having children and thus higher levels of commitment, making crime decisions even costlier.

- *Age.* It is expected that age negatively affects the counting of crimes committed by males, but this sign should revert as age advances. That is, I expect that younger people tend to commit more crimes and at some point, these rates start to decrease. The first statement is not supported by the positive sign of the Pearson correlation, but age squared turns out to be positively correlated with the count of offenses. Nonetheless, I still expect that the sign of age would be negative in non-linear estimations. My economist intuition is that younger males tend to face higher liquidity constraints and they are also stronger and "brave" than older people. Also, the older the man the higher the probability that he has reached certain economic stability and lost the ability to get involved in challenging physical endeavors such as committing crimes, and also is less attractive for criminal organizations that usually are looking for "strong/intimidating" people.

- *Urban.* I would expect that men in urban locations would be more likely to commit a higher number of crimes. This would be not only because resources are more available in urban contexts but because exposure to social norms and lifestyle might also play a role for males who are more exposed to social distress and criminal organizations in cities than in rural zones.

Again, it is important to reaffirm that intuition from pure correlations has to be taken with precaution because this is probably an imprecise measure of non-linear variations arising in our data. Hence, the discussion above is merely informative and will be subject to empirical scrutiny after the estimation of the Poisson model.

Figure 2: Matrix of Correlations



Note: This figure reports the matrix of raw correlations in heatmap format.

2.2 Poisson Model: Estimation & Interpretation

Poisson regression is often used for modeling count data. Poisson regression has several extensions useful for count models. In our case, we run regressions where the dependent variable is the number of crimes committed by males as a function of individual-specific covariates such as age, age squared, being married, or living in an urban city. As suggested by Cameron & Trivedi (2009), I include robust standard errors for the parameter estimates to control for mild violation of underlying assumptions. In this case, the maximized likelihood is actually a pseudolikelihood. The model to be estimated is as follows:

$$\mathbb{E}(\text{crime}_i|x_i) \equiv \lambda_i = e^{x_i'\beta} = exp[\beta_0 + \beta_1 \cdot \mathbb{1}(\text{Married} = 1) + \beta_2 \cdot \text{Age} + \beta_3 \cdot \text{Age}^2 + \beta_4 \cdot \mathbb{1}(\text{Urban} = 1)] \quad (9)$$

Results for our Poisson model are displayed in Table 2. Column (1) displays point estimates for our covariates. For this specific case, the interpretation is "for one unit change in the independent variable, the difference in the expected count change by $\hat{\beta}$ times, while holding constant the rest of the variables of the model". Column (2) displays the marginal effects (at means) of the estimated coefficients, and their interpretation would be that a unit change in the covariate increases/decreases the probability of counting more crimes, but marginal effects can also be interpreted as the predicted number of events when it comes to categorical variables with respect to a baseline category. Hence, it is important to pay attention to the scale of the covariate. Coefficients for continuous variables are usually interpreted as semi-elasticities and coefficients for categorical variables tend to be adjusted as $e^{\hat{\beta}_k} - 1$. This is why I include column (3) with Incidence Rate Ratios (IRR), which is the exponentiated version of the coefficients in our baseline model. While the coefficients $\hat{\beta}$ have an additive effect in the log(y) scale, the IRRs have a multiplicative effect in the y scale. The interpretation for each covariate is as follows:

- **Married.** The $\hat{\beta}$ coefficient is the expected difference in log count between married and non-married males. Compared to non-married males, the expected log count of crimes decreases by ≈ 0.005, while holding all other variables constant. The marginal effect preserves its sign and suggests that the predicted number of events for being married is lower at ≈ 0.04. In addition, the IRR indicates that the incidence rate of crimes for married males is ≈ 0.99 times the incidence rate for the reference group (non-married males). However, all these results did not turn out to be statistically significant at any standard level.

  Even when my results are not statistically different from zero, I still want to address the discussion surrounding the effects of being married on crime. For instance, Andersen et al. (2015) analyzed whether the effect of marriage on recidivism varied by spousal criminality in Denmark, and found that marriage reduced recidivism compared to non-marriage only when the spouse had no criminal record. Airaksinen et al. (2023) uses longitudinal register data and between- and within-individual analyses to examine how cohabitation and marriage were associated with suspected crime in Finland. In that setting, cohabitation reduces the risk of being suspected of a crime. Finally, in a report for the Department of Justice of the United States, Thompson & Tapp (2023) shows that from 2021 to 2022, the rate of violent victimization increased for persons who were never married, married, divorced, and separated. However, married people still face lower rates of crime in comparison to other marital statuses.

- **Age.** The coefficient for age means that the expected decrease in log count for a one-unit increase in age is ≈ 0.02, holding all other variables constant. The marginal effect suggests that the average decrease in the predicted count from age is about 0.17. The IRR indicates the percent change in the incident rate of male crimes is a decrease of $\approx 2\%$ $(0.980 - 1)$ for every unit increase in age. This intuition is reinforced by the sign of the coefficient for age^2, which is positive as the men become older. For reference, Thompson & Tapp (2023) show that the highest rate of victimization for violent crime in 2022 in the US was for people between 18-24 years old, and the lowest rate was for people older than 50 years old. Hence, it seems to be reasonable to assume non-linearities in the variation of age and its relation with offenses committed by males.

- **Urban.** The estimated coefficient for urban suggests that, compared to non-urban resident males, the expected log count of crimes increases by ≈ 0.03, while holding all other variables constant. The marginal effect preserves its sign and suggests that the predicted number of events for living in an urban zone is higher at ≈ 0.3. In addition, the IRR indicates that the incidence rate of crimes for urban-resident males is ≈ 1.03 times the incidence rate for the reference group (non-urban-resident males), holding all other variables constant. Again, the intuition of these results seems to be quite reasonable in comparison to data observed in developed contexts, like the US. For instance, Thompson & Tapp (2023) in their report for the Department of Justice, state that the highest rate of crime is in urban regions compared to suburban or regional and rural zones, respectively.

Table 2: Poisson Model

Variables	(1) Estimates for $\hat{\beta}$	(2) Marginal effects	(3) Incidence rate ratios
$\mathbb{1}(\text{Married} = 1)$	-0.00451	-0.0394	0.996
	(0.00627)	(0.0548)	(0.00625)
Age	-0.0199***	-0.174***	0.980***
	(0.00208)	(0.0182)	(0.00204)
Age^2	0.000899***	0.00786***	1.001***
	(2.55e-05)	(0.000224)	(2.56e-05)
$\mathbb{1}(\text{Urban} = 1)$	0.0316***	0.276***	1.032***
	(0.00662)	(0.0579)	(0.00683)
Constant	1.387***		4.003***
	(0.0401)		(0.161)
Observations	83,470	83,470	83,470

Note: This table reports estimates from Poisson regression models. Robust standard errors in parenthesis.

2.3 Overdispersion: Implications & Tests

So far, we have conveniently assumed that Poisson regression is the appropriate model for our data-generating process, but this is not necessarily the case. At this point, we might need to talk about overdispersion. An extensive discussion can be found in chapter 20.2.4 in the Cameron & Trivedi (2005) textbook. About the problem in counting regression models, they discuss that "the fundamental problem is that the distribution is parameterized in terms of a single scalar parameter so that all moments of the count of crimes are a function of that scalar parameter. In our case, this parameter is the mean and the variance of the Poisson distribution" (p. 670), which are actually the same $\mathbb{E}(y) = V(y) = \lambda$. One way this restrictiveness manifests itself is that in many applications a Poisson density predicts the probability of a zero count to be considerably less than is actually observed in the sample. This is what they call "the excess zeros problem", as there are more zeros in the data than the Poisson predicts. A second deficiency of the Poisson model they describe is that for count data the variance usually exceeds the mean. This is what they call "overdispersion" and has qualitatively similar consequences to the failure of the assumption of homoskedasticity in the linear regression model.

Our data arguably suggests that we are facing a problem of overdispersion. As shown in Figure 1, while there is only $\approx 5\%$ of offenses equal to zero in our sample (so the "excess of zeros problem" is not a threat here), the distribution for the number of crimes committed by males is highly skewed, with a variance that is fourteen times its mean. Cameron & Trivedi (2005) discusses some of the consequences of the overdispersion:

- First, in more complicated settings such as with truncation and censoring, overdispersion leads to the more fundamental problem of inconsistency.
- Second, even in the simplest settings, large overdispersion leads to grossly deflated standard errors and grossly inflated statistics in the usual Maximum Likelihood output, and hence it is important to use the previously given robust variance estimator.
- Third, if one wants to estimate probabilities of the number of events, rather than merely the conditional mean, these depend on additional parameters of the data generating process.

Next, I discuss some flexible ways of testing for overdispersion after running a Poisson model:

- **Pearson goodness-of-fit test.** Following Stata user's guide for a reference of this test, let M be the total number of covariate patterns among the N observations. View the data as collapsed on covariate patterns $j = 1, 2, ..., M$, and define m_j as the total number of observations having covariate pattern j and y_j as the total number of positive responses among observations with covariate pattern j. Define p_j as the predicted probability of a positive outcome in covariate pattern j. The Pearson χ^2 goodness-of-fit statistic is

$$\chi^2 = \sum_{j=1}^{X} \frac{(y_j - m_j p_j)^2}{m_j p_j (1 - p_j)} \tag{10}$$

  This χ^2 statistic has approximately $M - k$ degrees of freedom for the estimation sample, where k is the number of independent variables, including the constant. For a sample outside the estimation sample, the statistic has M degrees of freedom.

- **Hosmer–Lemeshow goodness-of-fit test.** Following Stata user's guide for a reference of this test, we have that is calculated similarly as the Pearson test, except that rather than using the M covariate patterns as the group definition, the quantiles of the predicted probabilities are used to form groups. Let G = # be the number of quantiles requested. The smallest index $1 \leq q(i) \leq M$, such that:

$$W_q(i) = \sum_{j=1}^{m_j \geq N} G \tag{11}$$

  gives $p_q(i)$ as the upper boundary of the i^{th} quantile for $i = 1, 2, ..., G$. Let $q(0) = 1$ denote the first index. The resulting F_2 statistic has approximately $G-2$ degrees of freedom for the estimation sample. For a sample outside the estimation sample, the statistic has G degrees of freedom.

- **Cameron & Trivedi test.** Following (Cameron & Trivedi, 2005, pp.670-671) for a reference of this test, most count models with overdispersion specify overdispersion to be of the form

$$V[y_i|x_i] = \mu_i + \alpha g(\mu_i)$$

  where α is an unknown parameter and $g(\cdot)$ is a known function, most commonly $g(\mu) = \mu^2$ or $g(\mu) = \mu$. It is assumed that under both null and alternative hypotheses the mean is correctly specified as, for example, $\exp(x_i\beta)$, whereas under the null hypothesis $\alpha = 0$ so that $V[y_i|x_i] = \mu_i$. A simple overdispersion test statistic for $H_0 : \alpha = 0$ versus $H_1 : \alpha \neq 0$ or $H_1 : \alpha > 0$ can be computed by estimating the Poisson model, constructing fitted values $\hat{\mu}_i = \exp(x_i\hat{\beta})$, and running the auxiliary OLS regression (without constant).

$$\frac{(y_i - \hat{\mu}_i)^2 - y_i}{\hat{\mu}_i} = \alpha \frac{g(\hat{\mu}_i)}{\hat{\mu}_i} + u_i \tag{12}$$

  where u_i is an error term. The reported t-statistic for α is asymptotically normal under the null hypothesis of no overdispersion.

In the next sub-section I will formally implement these three tests to explore the presence of overdispersion in our data.

2.4 Negative Binomial Model & Test for Overdispersion

We study a special case of Poisson regression models. Now, we reconvert the expected value of our baseline model in Equation 9 and make:

$$\mathbb{E}(\text{crime}_i|x_i) \equiv \lambda_i^* = e^{x_i'\beta}e^{\nu_i} \tag{13}$$

Where $\nu_i \sim \Gamma(\frac{1}{\alpha}, \alpha)$, and α is the parameter for overdispersion. Table 3 report the results from a negative binomial model. As argued by Cameron & Trivedi (2005), Poisson regression models are a specific case of the family of negative binomial models so the interpretation of coefficients is the same in both models but with a little twist. In a Poisson model, the coefficients represent the change in the log mean of the variable of interest (usually a rate or count) per unit change in the corresponding independent variable, holding all other variables constant. Therefore, they are interpreted as the effect of the predictor on the rate of occurrence of the event of interest. In a negative binomial model, the coefficients also represent the change in the log mean of the variable of interest, but due to the nature of the negative binomial distribution, this change is not linear in the rate of occurrence. Instead, the interpretation of the coefficients in a negative binomial model is in terms of the probability of success on each trial –as opposed to a rate of occurrence–, assuming that other variables are held constant. Therefore, the coefficients in a negative binomial model can be interpreted as the change in the probability of success for an additional trial, given the values of the predictors.

Table 3: Negative Binomial Model

	(1)	(2)	(3)
Variables	Estimates for $\hat{\beta}$	Marginal effects	Incidence rate ratios
$\mathbb{1}(\text{Married} = 1)$	0.00162	0.0142	1.002
	(0.00552)	(0.0483)	(0.00553)
Age	-0.0200***	-0.175***	0.980***
	(0.00188)	(0.0165)	(0.00185)
Age^2	0.000900***	0.00787***	1.001***
	(2.31e-05)	(0.000203)	(2.31e-05)
$\mathbb{1}(\text{Urban} = 1)$	0.0290***	0.253***	1.029***
	(0.00584)	(0.0511)	(0.00601)
α	-0.696***		0.499***
	(0.00618)		(0.00309)
Constant	1.387***		4.002***
	(0.0365)		(0.146)
Observations	83,470	83,470	83,470

Note: This table reports estimates from negative binomial regression models.

The interpretation for each covariate is as follows:

- **Married.** We can notice that any of the coefficients for being married turn out to be statistically insignificant at any standard level and also estimates for the coefficient and marginal effect did change their sign in comparison to those we obtained before with the Poisson model. From these results we can conclude that being married is not a relevant predictor of variations in the expected counts of crimes committed by men.

- **Age.** The coefficient for age means that the expected decrease in the probability of success for an additional trial in log count for a one-unit increase in age is ≈ 0.02, holding all other variables constant. The marginal effect suggests that the average decrease in the predicted count from age is about 0.18. The IRR indicates the percent change in the incident rate of male crimes is a decrease of $\approx 1\%$ times for every year of increase in age. This intuition is reinforced by the sign of the coefficient for age^2, which is positive as the men become older. These results are qualitatively and quantitatively similar to those obtained using the Poisson model.

- **Urban.** The estimated coefficient for urban suggests that, compared to non-urban resident males, the expected probability of success for an additional trial in the log count of crimes increases by ≈ 0.03, while holding all other variables constant. The marginal effect preserves its sign and suggests that the predicted number of events for living in an urban zone is higher at ≈ 0.25. In addition, the IRR indicates that the incidence rate of crimes for urban-resident males is ≈ 1.03 times the incidence rate for the reference group (non-urban-resident males), holding all other variables constant. Again, these results mirror those we obtained with the Poisson specification.

Next, I carry out the three statistical tests discussed in the previous subsection to look for overdispersion. Results for Pearson and Hosmer–Lemeshow goodness-of-fit tests are displayed in Table 4. In contrast, the α coefficient for residuals from the Cameron y Trivedi test are displayed in Table 5. Recall that in all of the tests, the contrast of hypotheses is H_0 : no overdispersion versus H_1 : presence of overdispersion. By looking at the p-values and standard errors, and as expected, the three tests report strong statistical evidence rejecting the null hypothesis of no overdispersion. Moreover, the coefficient I get for $\hat{\alpha}$ with the Cameron & Trivedi test is the same as reported by default from Stata in column (3) of the Table 3. Thus, we can conclude that the negative binomial model is more appropriate to study the count of the number of crimes males commit.

Table 4: Goodnes of Fit

Statistic	Pearson goodness-of-fit	Deviance goodness-of-fit
χ^2	524,739.12	492,947.92
p-value	0.000	0.000

Note: This table reports results from the goodness of fit test as in Equation 10 and Equation 11, respectively.

Table 5: Overdispersion test à la Cameron & Trivedi (2005)

Regressor	Offenses
$\hat{\alpha}$ errors	0.495***
	(0.00424)
Observations	83,470
R-squared	0.140

Note: This table reports results for an overdispersion test as in Equation 12.

3 Duration Models

In this section, we are going to study duration models. For that purpose, we will try to understand and replicate the paper *¿Cuánto Dura el Desempleo de la Población más Pobre en Chile?* (Montero, 2007).

3.1 Summary of the Paper

This paper by Montero (2007) provides an empirical evaluation of unemployment duration for Chile's poorest population. Its main objectives are to dive into the determinants of the unemployment spell, understand how much it lasts/changes across individual characteristics, and study the probability of leaving the unemployment as time goes by. To achieve these goals, the paper relies on rich survey data. Main estimates are drawn from *Encuesta Panel Chile Solidario 2004* which was applied to a sub-sample of individuals from *Encuesta de Caracterización Socioeconómica Nacional* (CASEN) 2003.

The most substantial results of the paper are mainly concentrated in Section 5 ("Estimaciones y Resultados"). The empirical strategy followed by the author is split into two stages. In the first place, the author focuses on unconditional analysis using non-parametric tools to estimate the probability of being unemployed as long as time goes by. He found that the most significant decrease in the probability of being unemployed happens around the sixth-seventh month. In the second place, the author relies on a parametric framework to run a conditional analysis on the determinants of unemployment according to several individual-specific characteristics. Results from this second stage of the analysis suggest that being a woman, being from an indigenous ethnicity, being right down below the retirement age (45-54 years old), and being highly educated are the most relevant predictors for the duration of unemployment.

In my opinion, some relevant takeaways arise from this paper. Firstly, although the results are not directly comparable or representative of the whole population, they increase our comprehension of an often understudied phenomenon. Moreover, the paper overcomes a non-negligible challenge in the empirical analysis which is having access to household representative data that recovers the duration of unemployment not only for unemployed people per se –which is the common practice in household surveys– but for employed people as well. Finally, the paper addresses a super relevant policy question, so it provides helpful insights into the better design of labor market policies, especially for underrepresented sectors.

3.2 Replication: Log-Normal Distribution

In this segment, I rely on a parametric approach to study the determinants of unemployment duration, as in table 4 of Montero (2007). An important consideration is that the model estimated by the author is an Accelerated Failure Time (AFT) model. Whereas in a Proportional Hazards (PH) model, the covariates act multiplicatively on the hazard, in an AFT model the covariates act multiplicatively on time. An AFT model proposes the following relationship between covariates and time:

$$\ln(t) = x_i\beta + W_i \quad \text{with } W_i \sim_{iid} f \tag{14}$$

The above framework describes a general class of models: depending on the distribution we specify for W. Let us suppose that we can write our AFT model as:

$$\ln(t) = x_i\beta + \sigma W_i \tag{15}$$

Note, for example, that the Weibull and lognormal models can be written this way. Then, the likelihood function is given by:

$$\ell(\beta, \sigma; x, t) = \prod_{i=1}^{n} \left[\frac{f(\upsilon_i)}{\sigma}\right]^{d_i} S(\upsilon_i)^{1-d_i} = \prod_{i=1}^{n} \left[\frac{\lambda(\upsilon_i)}{\sigma}\right]^{d_i} S(\upsilon_i) \tag{16}$$

Where f, λ and S represent the density, hazard, and survival functions for the error distribution, and $\upsilon_i = \frac{\ln(t)-x_i\beta}{\sigma}$. Assuming f to be a log-normal distribution, the survival function is:

$$S(t) = 1 - \Phi\left[\frac{\ln(t) - x_i\beta}{\sigma}\right]$$

Results are reported Table 6. Notice that the number of observations is twenty times larger than the original sample size because of the way I am implementing the expansion factor of the survey Panel 2004[2]. The interpretation of our coefficients from an AFT model is quite similar to the interpretation of any log-linear regression because $\hat{\beta}$ is the semi-elasticity of the covariate x_i on the time people are unemployed. If the coefficient is positive, then the $e^{\hat{\beta}_k}$ will be > 1, which will decelerate the event time (increase the mean survival time). Similarly, a negative coefficient will reduce the mean survival time (accelerate the event time). The particular interpretation for each covariate is as follows:

- **Gender.** Being a male reduces the expected duration of unemployment by $1 - e^{-0.333} \approx 0.28$ p.p. This result is qualitatively consistent with the one reported by Montero, as my results preserve the negative sign; however, they are quantitatively different, because Montero gets a coefficient of -0.19. But in any case, my results also preserve the statistical significance.

- **Years of schooling.** An additional year in schooling increases the expected duration of unemployment by $1 - e^{0.02} \approx 2\%$. This coefficient mirrors the results of Montero, qualitatively and quantitatively.

- **Ethnicity.** Being indigenous decreases the expected duration of unemployment by $1 - e^{-0.113} \approx 0.11$ p.p. This result is completely different from the one reported by Montero in his paper, as his finding is that being indigenous increases the expected duration of unemployment.

- **Age.** A one-unit increase in age increases the expected duration of unemployment. Moreover, this is a general result, and econometric analysis using age as a covariate often considers the presence of non-linearities. As Montero did, I also ran heterogeneity analysis for different brackets of age and found a similar intuition compared to the paper. Being older than 65 years old increases the expected duration of unemployment by $1 - e^{0.032} \approx 0.03$ p.p. Is worth noting that my magnitudes are different from those reported by Montero.

- **Being fired.** People who reported being hired from their previous employment face a lower expected duration of unemployment. This result differs from the one reported by Montero.

- **First job.** I get that looking for a job for the very first time substantially decreases the expected duration of unemployment by $1 - e^{-8.138} \approx 0.99$ p.p. This result is qualitatively similar to the one reported by Montero in his paper with the consideration that his coefficient is lower than mine.

[2] I use the **fweights** command in Stata.

Table 6: Determinants for the Duration of Unemployment
Log-Normal Distribution

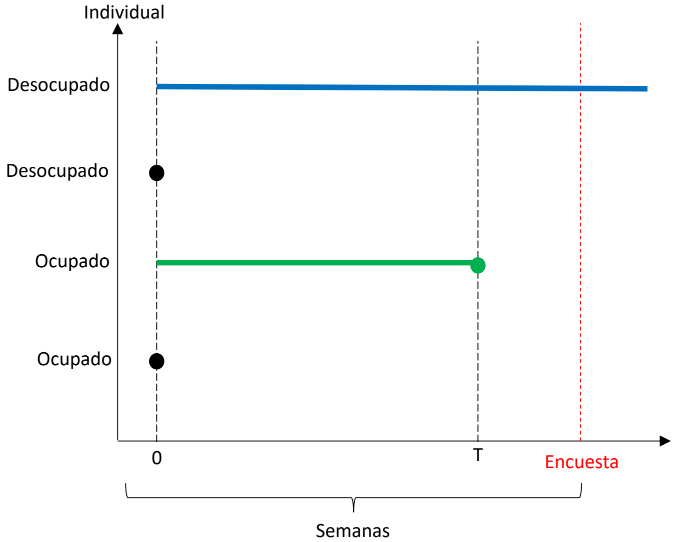
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Sigma (σ)	1.29	1.28	1.26	1.26	1.26	1.25	1.25
Constante	2.372***	2.587***	2.432***	2.445***	2.292***	2.699***	10.818
Hombre		-0.333***	-0.318***	-0.319***	-0.322***	-0.325***	-0.317***
Años de Escolaridad			0.020***	0.019***	0.022***	0.018***	0.019***
Indígena				-0.113***	-0.109***	-0.098***	-0.092***
Edad					0.003***		
25-34 años de edad						-0.154***	-0.162***
35-44 años de edad						-0.399***	-0.410***
55-64 años de edad						-0.333***	-0.333***
Mayor de 65 años						0.032	0.026
Despedido						-0.029**	-0.020
Busca trabajo por primera vez							-8.138
N observaciones	51,516	51,516	50,455	50,455	50,455	50,455	50,455
Prob>Chi2		0.00	0.00	0.00	0.00	0.00	0.00

Note: This table reports coefficients from estimating an AFT model using a log-normal distribution.

Why do my results differ to some extent from those reported by Montero? In the first place, I think it is important to flag several considerations surrounding our data. We are dealing with an interval censoring problem, so maybe there are differences in harmonizing the definition of failures to make it as comparable to the paper as possible. Figure 3 illustrates this issue. In the data that is available to us, there are three cases: i) people who reported having at least one week being unemployed before but at the time of the survey were employed (green line); ii) people who reported having at least one week being unemployed before but at the time of the survey still were unemployed (blue line); and iii) people who reported have zero weeks being unemployed but at the time of the survey were employed or unemployed (black circles). Case i) corresponds to full stories, as the failure occurred sometime before the survey, while case ii) is the censored case. Case iii) will be considered by Stata as observations that end on or before entering the state. The paper is not that clear about how the author defines failures. In my case, I defined failures as individuals who were employed at the time of the survey. This is maybe one of the reasons.

A second consideration has to do with the data structure itself. We are not very sure that we are working with the same sample as in the paper. In an attempt to have any clue about comparability, I replicate Table 3 of page 221 in the paper for the average duration of unemployment in weeks conditional on individual traits and labor occupational status at the time of the survey. Results are reported in Table 7. After a quick check, we can immediately realize that our data systematically differs from the data used in the paper. For example, the paper indicates that unemployed people in urban zones last, on average, 15.6 weeks unemployed, but in our case, this number is 12.6. We can continue to do more comparisons and will always face a different number. Thus, it is not expected that my results would mirror those reported by Montero (2007).

Figure 3: Diagram for Interval Censoring



Note: This figure reports a diagram with the survival duration of unemployment in the survey data.

Table 7: Duration of Unemployment in Weeks

	Ocupados	Desocupados	Todos
Zona			
Rural	9.1	9.0	9.1
Urbano	15.0	12.6	14.6
Género			
Hombre	11.5	12.2	11.6
Mujer	15.7	10.7	14.7
Etnia			
No indigena	13.5	11.5	13.2
Indigena	9.7	11.4	10.0
Edad			
25-34 años	14.9	10.6	14.1
35-44 años	12.0	11.5	11.9
45-54 años	12.6	13.9	12.8
55-64 años	12.8	10.9	12.6
65 o más	6.4	11.2	7.0
Escolaridad			
Sin escolaridad	6.2	7.8	6.4
1-8 años	12.4	10.4	12.1
9-12 años	15.8	15.5	15.7
13-17 años	17.8	20.0	18.0

Note: This table reports the average duration of unemployment (in weeks) by current employment situation. This is a replication of table 3 as in Montero (2007) but using our sample. This table aims to explore how similar our data is in comparison to the data used in the paper.

3.3 Replication: Weibull Distribution

I replicate Equation 16 but now considering that the distribution is Weibull. It is important to be aware that the Weibull distribution can be reparametrized as an AFT or PH model[3]. The survival function is given by $S(t) = exp(-\gamma t^{\alpha})$. Results are displayed in Table 8. A consideration is that the log-likelihood for specification (7) did not achieve convergence, but I still report the results. Interpretation is as follows:

- **Gender.** Being a male reduces the expected duration of unemployment by $1 - e^{-0.354} \approx 0.30$ p.p. This result is higher in magnitude than the one I got before using the log-normal distribution.
- **Years of schooling.** An additional year in schooling increases the expected duration of unemployment by $1 - e^{0.02} \approx 2\%$. This coefficient mirrors the ones I got with the log-normal distribution.
- **Ethnicity.** Being indigenous decreases the expected duration of unemployment by $1 - e^{-0.087} \approx 0.08$ p.p. This result differs in magnitude from the one I got using the log-normal distribution.
- **Age.** A one-unit increase in age increases the expected duration of unemployment. These results are similar to those obtained using the log-normal distribution. However, there is a notable difference as in specification (6) the coefficient for people older than 65 years old turns out negative.
- **Being fired.** People who reported being hired from their previous employment face a lower expected duration of unemployment. This result is slightly different from the one I got with the log-normal distribution.
- **First job.** I get that looking for a job for the very first time substantially decreases the expected duration of unemployment by $1 - e^{-18.112} \approx 0.99$ p.p. This result is the same as I got before using a log-normal distribution.

Table 8: Determinants for the Duration of Unemployment
Weibull Distribution

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constante	3.009***	3.233***	3.029***	3.039***	2.827***	3.422***	21.512***
Hombre		-0.354***	-0.332***	-0.329***	-0.336***	-0.320***	-0.309***
Años de Escolaridad			0.024***	0.024***	0.026***	0.021***	0.023***
Indígena				-0.087***	-0.076***	-0.048***	-0.045**
Edad					0.005***		
25-34 años de edad						-0.277***	-0.283***
35-44 años de edad						-0.546***	-0.556***
55-64 años de edad						-0.342***	-0.340***
Mayor de 65 años						-0.415***	-0.422***
Despedido						-0.097***	-0.089***
Busca trabajo por primera vez							-18.112***
N observaciones	51,516	51,516	50,455	50,455	50,455	50,455	50,455
Prob>Chi2	.	0.00	0.00	0.00	0.00	0.00	0.00

Note: This table reports coefficients from estimating an AFT model using a Weibull distribution.

[3]This is important because, by default, Stata will estimate the PH version for the Weibull distribution and report hazard ratios. Thus, we need to include the commands **nohr time** when using *streg* to get the AFT version of the Weibull distribution.

3.4 Which Distribution Fits the Model Best?

In most of the cases, results are qualitatively similar to estimations of the AFT model using a log-normal distribution but they are not the same quantitatively speaking. Thus, a natural question is why these models yield different results. The answer to this question is closely related to the assumption we made for modeling the duration of unemployment in Chile. So far, we have assumed that log-normal or Weibull distributions reasonably adapt to our data-generating process, yet this is an ad-hoc assumption. We should run a formal test to make sure that we are using an adequate density function. This is what we do in this section. I employ two different tools: the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Each criterion is estimated as follows:

$$\text{AIC} = 2(k + p + 1) - 2\log(\hat{L}) \tag{17}$$

$$\text{BIC} = k\log(n) - 2\log(\hat{L}) \tag{18}$$

Where k is the number of covariates in the model; p is the number of auxiliary coefficients; n is the number of observations and $\hat{L}$ is the log-likelihood estimated for the model. These are the criteria for selecting models among a finite set of models. It is based, in part, on the probability function. When fitting models, it is possible to increase the likelihood by adding parameters, but this can result in overfitting. Both the BIC and AIC solve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in the BIC than in the AIC. We would expect that the lower the AIC or BIC, the better the model fits the data.

Results are displayed in Table 9. I also include information criterion from the estimation of the models assuming a generalized gamma distribution. Each column corresponds to a different specification as I did in Tables 6 and 8, respectively. Notice I do not report the seventh specification (the one adding a dummy of people looking for a job for the first time) because the model does not achieve convergence. Anyways, the information criterion suggests that the distribution fitting the data best is the generalized gamma since the numbers are the lowest in most cases.

Table 9: Akaike Information Criterion

Distribution	(1)	(2)	(3)	(4)	(5)	(6)
Akaike Information Criterion						
Log-normal	148,882.53	148,168.74	143,674.95	143,642.43	143,610.87	142,928.71
Weibull	155,176.02	154,345.10	149,135.52	149,116.33	149,041.24	147,869.66
Generalized gamma	148,786.50	148,112.20	143,665.78	143,634.30	143,603.97	142,927.91
Bayesian Information Criterion						
Log-normal	148,900.23	148,195.29	143,710.27	143,686.58	143,663.85	143,017.00
Weibull	155,193.72	154,371.64	149,170.84	149,160.48	149,094.21	147,957.95
Generalized gamma	148,813.05	148,147.59	143,709.92	143,687.28	143,665.77	143,025.03

Note: This table reports results.

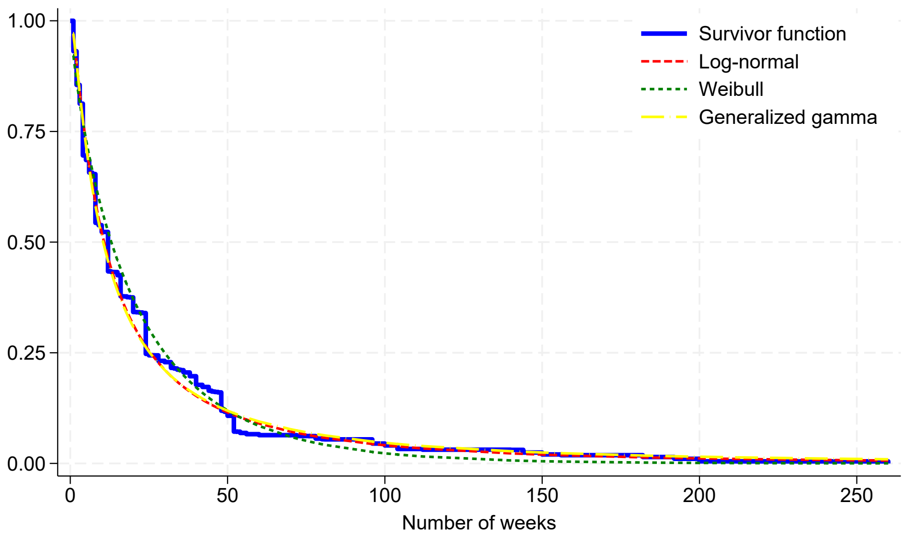
3.5 Survival & Hazard Functions

Next, I will report unconditional estimates of the survival function. For this purpose, I compute the Kaplan-Meier estimator given by:

$$\hat{S}(t) = \prod_{t_i<t} \left(\frac{n_i - d_i}{n_i} \right)$$

With d_i the number of failures right before t_i, and n_i the number of unemployed people right before time t_i. The interpretation is that for any given t_i (measured in weeks) the survival function represents the probability of remaining unemployed until week t_i. Figure 4 displays the results for non-parametric and parametric estimations of $S(t)$. Parametric estimates are obtained assuming three different distributions: log-normal, Weibull, and generalized gamma. We can see that the survival function decreases with time and the Kaplan-Meier estimate mirrors the one reported by Montero in figure 1 of the paper. We also observe that around 30 weeks there is a turning point, so around the seventh month of searching the probability of leaving unemployment increases. Also, is worth noting that the Weibull distribution does not seem to be fitting that well the empirical survival function. In Figure A1 of the Appendix, I run some additional exercises for the Kaplan-Meier estimator conditional on a selected number of covariates.

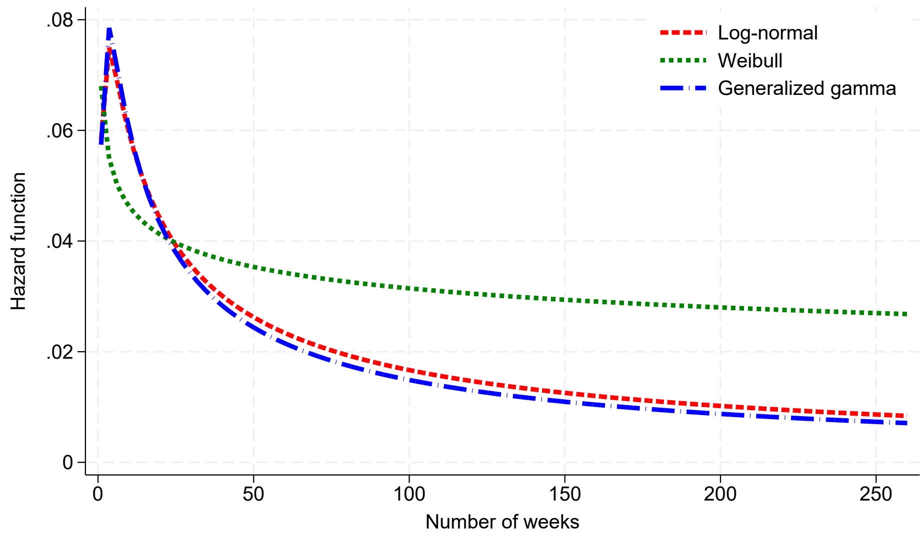
Figure 4: Parametric & Non-Parametric Survival Functions



Note: This figure reports survival estimates for unemployment. The blue thick line is the non-parametric Kaplan-Meier estimate for the survival function, as in figure 1 in the paper of Montero (2007).

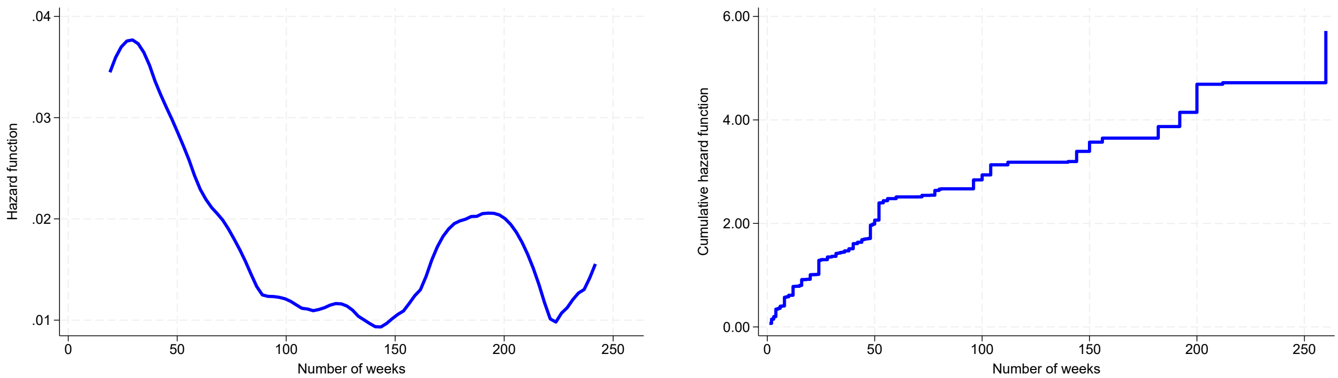
Another interesting estimator for survival analysis is the hazard rate. Hazard is defined as the slope of the survival curve. Thus, we could say that the hazard function is the probability that if you survive to time t_i, you will experience the event in the next instant, in other words, the hazard function gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t_i. The hazard rate, as the survival function, can be estimated parametrically and non-parametrically. Results for the parametric version are reported in Figure 5. We can notice how divergent estimates with the Weibull distribution are. My results for the log-normal distribution suggest that, at the beginning of the process, the first few weeks as unemployed, the probability of exiting unemployment increases, reaching a maximum point around 30 weeks, which is similar to the pattern reported by Montero in his paper. Finally, for illustration purposes, I include estimations for non-parametric versions of the Hazard function. Figure 6a shows the smoothed hazard estimate, as in figure 2 in the paper of Montero, and Figure 6b shows the Nelson-Aalen cumulative hazard rate.

Figure 5: Parametric Hazard Functions



Note: This figure reports hazard functions for unemployment, as in figure 6 in the paper of Montero (2007).

Figure 6: Non-Parametric Hazard Functions



(a) Smoothed hazard estimate

(b) Nelson-Aalen cumulative

Note: This panel presents non-parametric estimates for hazard functions.

4 Generalized Method of Moments

In this section, we are going to extend our knowledge of the Generalized Method of Moments (GMM). For that purpose, we are going to conduct an empirical application following "*Imposing Moment Restrictions from Auxiliary Data by Weighting*" by Hellerstein & Imbens (1999).

4.1 Summary of the Paper

In their paper, the authors analyze the estimation of coefficients in models under moment restrictions that are drawn from auxiliary data, and how these orthogonality conditions offer enough traceability to get weights for each observation in the auxiliary data that can be implemented in regression analysis afterward. Hence, in my interpretation, the paper addresses an underlying fundamental question ¿in contexts of moment restrictions, how important weighting strategies are to get consistent estimators that are drawn from auxiliary samples?

To tackle this question, the authors follow two stages of analysis: a theoretical and an empirical. In the first part of the study, they propose a weighted estimation methodology retrieved from moment restrictions using auxiliary data under the assumption that the moment restrictions are correctly specified, and then, they extend this estimator to the context where the target and auxiliary samples differ. The methodology above relies on two assumptions. To make results interpretable they assume that the population of interest is the same as the sampled population –in terms that the individuals appear in both data sources– and also that these populations differ –in the sense that non-response and attrition might be present in the auxiliary data–. In the second stage of the analysis, they carry out an empirical application of their methodology to estimate wage regressions using data from the 1980 wave of the National Longitudinal Survey Young Men's Cohort (NLS) –auxiliary data– with weights derived from readily available 1980 Census records –target population–.

The paper contributed to several strands of the literature. Firstly, they propose a weighted estimator that is flexible enough to be computed using GMM, but known parameters are necessary for defining the overidentifying moments. Secondly, they propose an estimator that accounts for sampling errors in the moment restrictions, thus serving as an extension to contingency analysis with known marginals. Finally, they propose a weighted estimator that does not rely on tight parametric assumptions, so they can easily interpret results when the population of interest and auxiliary sample differ from each other. In this sense, their empirical results are striking as they propose an estimator that alleviates concerns around how representative econometric results are with respect to the true population, or to a general extent, they offer an estimator that tackles internal and external validation. The underlying mechanism for this is that they impose moment restrictions that yield significant improvements in weighting estimations compared to those estimations that would be obtained in the absence of omitted variables bias.

4.2 Estimations without Overidentification

Now we are going to analyze the case when the moment restrictions are identified (the same number of parameters as the number of equations). We have the following condition constraint:

$$\mathbb{E}[\psi(\log(earn), educ, expr, expr^2, iq; \beta)]$$

Where:

$$\psi = \begin{bmatrix} \log(earn)_i - \beta_0 - \beta_1 \cdot \text{educ}_i - \beta_2 \cdot \text{expr}_i - \beta_3 \text{expr}_i^2 - \beta_4 \cdot \text{iq}_i \\ educ \times (\log(earn)_i - \beta_0 - \beta_1 \cdot \text{educ}_i - \beta_2 \cdot \text{expr}_i - \beta_3 \text{expr}_i^2 - \beta_4 \cdot \text{iq}_i) \\ expr \times (\log(earn)_i - \beta_0 - \beta_1 \cdot \text{educ}_i - \beta_2 \cdot \text{expr}_i - \beta_3 \text{expr}_i^2 - \beta_4 \cdot \text{iq}_i) \\ expr^2 \times (\log(earn)_i - \beta_0 - \beta_1 \cdot \text{educ}_i - \beta_2 \cdot \text{expr}_i - \beta_3 \text{expr}_i^2 - \beta_4 \cdot \text{iq}_i) \\ iq \times (\log(earn)_i - \beta_0 - \beta_1 \cdot \text{educ}_i - \beta_2 \cdot \text{expr}_i - \beta_3 \text{expr}_i^2 - \beta_4 \cdot \text{iq}_i) \end{bmatrix}$$

The identification assumption is given by: $\mathbb{E}(\mathbf{x}'\upsilon) = 0$, with $\mathbf{x}_i \equiv [1, educ_i, expr_i, expr_i^2, iq_i]$. We are asked to estimate the moment restriction described above using Ordinary Least Squares (OLS) with the Eicker- Huber-White correction, that is, with robust standard errors, and compare these results with the two-step GMM estimator and robust weighting matrix. I estimate the following equation:

$$\log(earn)_i = \beta_0 + \beta_1 \cdot \text{educ}_i + \beta_2 \cdot \text{expr}_i + \beta_3 \text{expr}_i^2 + \beta_4 \cdot \text{iq}_i + \upsilon_i \tag{19}$$

$educ_i$ is the years of education; $expr_i$ is defined as "potential experience"$= age - 6 - years\ of\ education$; iq_i are the scores from an IQ test. Results for equation 19 are displayed in Table 10. We can immediately realize that both methods yield similar results, quantitatively and qualitatively. Moreover, standard errors reported by OLS are larger than those reported by GMM by a factor of sqrt(74/71). This is because the Stata command for OLS makes a small-sample adjustment to the estimated variance matrix but does not for the GMM estimator[4]. Recall we are estimating a log-linear model where the dependent variable is scaled in logs, but the right-hand side of the model is levels. Interpretation for each coefficient is as follows:

- **Education.** An increase of one year in the years of education increases earnings by $\approx 6.3\%$, holding all other variables constant.
- **Experience.** An increase of one year in experience yields a change of $\approx (0.068 - 2 * 0.0016 expr) \times 100$, holding all other variables constant.
- **IQ.** A unit increase in the IQ test score increases earnings by $\approx 0.6\%$, holding all other variables constant.

Table 10: Returns to Schooling

	(1) OLS	(2) GMM
Education	0.06337***	0.06337***
	(0.00822)	(0.00820)
Experience	0.06786***	0.06786***
	(0.02277)	(0.02271)
Experience^2	-0.00157**	-0.00157**
	(0.00079)	(0.00079)
IQ	0.00611***	0.00611***
	(0.00099)	(0.00098)
Constant	0.09192	0.09192
	(0.21036)	(0.20980)
N	935	935

Note: This table reports coefficients from estimating Equation 19 using OLS and GMM. Robust standard errors are included.

[4]Retrieved from the Stata helper for the GMM command. See here on page 14 for a reference.

4.3 Estimations with Overidentification

Now, let us include four additional moment restrictions retrieved from Census data. Hence:

$$\psi = \begin{bmatrix} \log(earn)_i - \beta_0 - \beta_1 \cdot \text{educ}_i - \beta_2 \cdot \text{expr}_i - \beta_3 \text{expr}_i^2 - \beta_4 \cdot \text{iq}_i \\ educ \times (\log(earn)_i - \beta_0 - \beta_1 \cdot \text{educ}_i - \beta_2 \cdot \text{expr}_i - \beta_3 \text{expr}_i^2 - \beta_4 \cdot \text{iq}_i) \\ expr \times (\log(earn)_i - \beta_0 - \beta_1 \cdot \text{educ}_i - \beta_2 \cdot \text{expr}_i - \beta_3 \text{expr}_i^2 - \beta_4 \cdot \text{iq}_i) \\ expr^2 \times (\log(earn)_i - \beta_0 - \beta_1 \cdot \text{educ}_i - \beta_2 \cdot \text{expr}_i - \beta_3 \text{expr}_i^2 - \beta_4 \cdot \text{iq}_i) \\ iq \times (\log(earn)_i - \beta_0 - \beta_1 \cdot \text{educ}_i - \beta_2 \cdot \text{expr}_i - \beta_3 \text{expr}_i^2 - \beta_4 \cdot \text{iq}_i) \\ \log(earn) - 2,0647 \\ \log(earn)^2 - 4,5258 \\ educ - 13,97 \\ \log(earn) \times educ - 29,099 \end{bmatrix}$$

Results are reported in Table 11. The first two columns correspond to the identified case, as in Equation 19, while the third column estimates the new ψ including moment restrictions from the Census data. My results mirror estimates reported by Hellerstein & Imbens (1999). Then, we are asked to compare results obtained using the identified versus the over-identified model. In terms of coefficients, they are all the same in terms of magnitude, sign, and significance. This is why we are asked to focus on efficiency instead. In particular, standard errors from the over-identified model turn out to be substantially lower than those from the identified model. As argued by Hellerstein & Imbens (1999), this is not necessarily because of a numerical reason but because we are "forcing" the moments to be the same as in the Census, thus reducing uncertainty, as we are replicating the true data-generating process. In the paper, the authors show that imposing restrictions implied by the Census moments considerably changes the wage regression results implying that estimates based solely on NLS data may not be very robust and need not generalize to the population at large. But in the paper, they are running weighted regressions. Instead, in our example case, we have moment restrictions. The intuition behind this exercise is that we can improve point estimates either by using weighting estimators –as in the paper, but at the cost of losing efficiency–, by imposing additional moment restrictions from the population sample –as in our example case–, or a combination of both –then systematically reducing bias and improving efficiency.

Table 11: Returns to Schooling

	Identified		Over-Identified
	(1) OLS	(2) GMM	(3) GMM
Education	0.06337***	0.06337***	0.06337***
	(0.00822)	(0.00820)	(0.00185)
Experience	0.06786***	0.06786***	0.06786***
	(0.02277)	(0.02271)	(0.00597)
Experience^2	-0.00157**	-0.00157**	-0.00157***
	(0.00079)	(0.00079)	(0.00021)
IQ	0.00611***	0.00611***	0.00611***
	(0.00099)	(0.00098)	(0.00039)
Constant	0.09192	0.09192	0.09192
	(0.21036)	(0.20980)	(0.06407)
N	935	935	935

Note: This table reports coefficients from estimating Equation 19 excluding (identified) and including (overidentified) the fourth additional moment restrictions. Robust standard errors are included.

4.4 Testing for Overidentification

In this case, I implement Hansen's J statistic, which is used to determine the validity of the over-identifying restrictions in a GMM model. If the model is correctly specified in the sense that $\mathbb{E}\left[z_i u_i(\beta)\right] = 0$, then the sample analog to that condition should hold at the estimated value of β. Hansen's J statistic is valid only if the weight matrix is optimal, meaning that it equals the inverse of the covariance matrix of the moment conditions. Results are reported in Table 12. The J statistic is significant even at the 1% significance level, so we conclude that there is over-identification, thus our model is misspecified.

Table 12: Hansen Test

Statistic	Value
Hansen's J-statistic	935.00
Degrees of freedom	19
P-value	0.0000

Note: This figure reports Hansen's test for over-identification.

References

Airaksinen, J., Aaltonen, W., Tarkiainen, L., Martikainen, P., & Latvala, P. (2023). Associations between cohabitation, marriage, and suspected crime: a longitudinal within-individual study. *Journal of Developmental and Life-Course Criminology*, (9), 54–70. 10

Andersen, S., LH, A., & PE., S. (2015). Effect of marriage and spousal criminality on recidivism. *Journal of Marriage and Family*, *2*(77), 496–509. 10

Becker, G. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, *76*(2). 7

Cameron, A. C., & Trivedi, P. K. (2009). Microeconometrics using stata. Tech. rep., College Station, TX: Stata Press. 9

Cameron, C. A., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press. 7, 11, 12, 13, 14

Dominguez, P. (2022). Victim incentives and criminal activity: Evidence from bus driver robberies in Chile. *The Review of Economics and Statistics*, *104*(5), 946–961. 7

Eriksson, K. H., Hjalmarsson, R., Lindquist, M. J., & Sandberg, A. (2016). The importance of family background and neighborhood effects as determinants of crime. *Journal of Population Economics*, *29*, 219–262. 7

Hellerstein, J. K., & Imbens, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. *The Review of Economics and Statistics*, *81*(1), 1–14. 23, 25

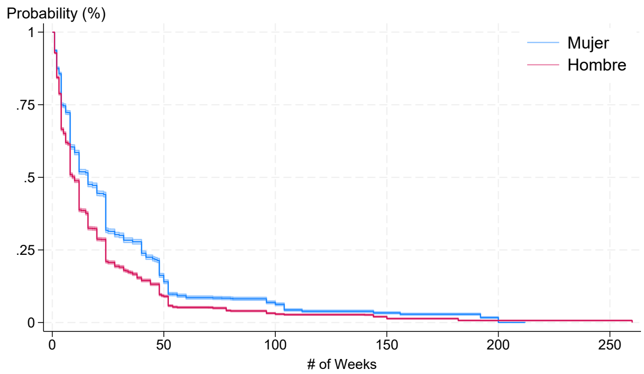
Montero, R. (2007). ¿Cuánto dura el desempleo de la población más pobre en Chile? *Cuadernos de Economía*, *44*(130), 211–231. 15, 17, 18, 21, 22

Thompson, A., & Tapp, S. N. (2023). Criminal victimization, 2022. Tech. rep., U.S. Department of Justice. 10

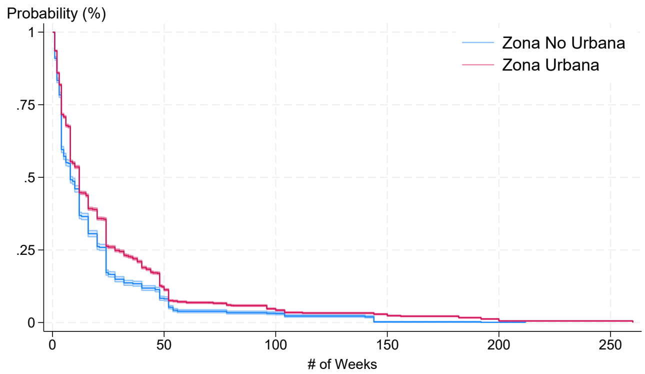
A Appendix

A.1 Additional Figures

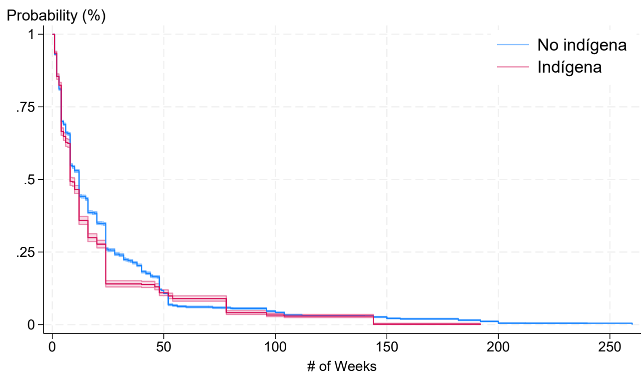
Figure A1: Survival Functions



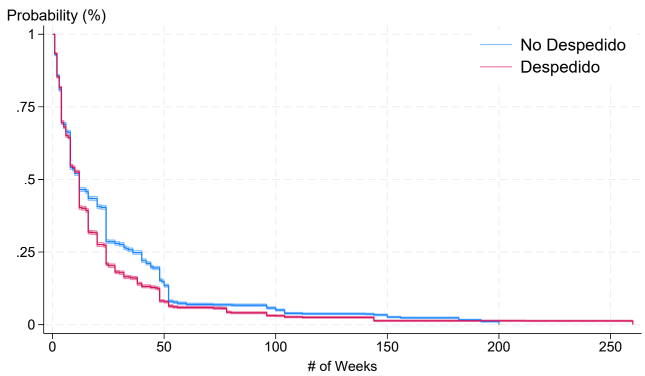
(a) Gender



(b) Urban Zone



(c) Ethnicity



(d) Last Job

Note: This panel presents non-parametric estimates for Kaplan-Meier survivor functions according to selected categorical covariates. Continuous red and blue lines correspond to point estimates while blurred areas are 95% confidence intervals. These figures are also helpful for understanding, to some extent, the proportionality assumption behind survival analysis.