A bivariate approach to the Mincerian earnings equation

Danúbia R. Cunha[1,3] , **Helton Saulo**[2], **Sandro E. Monsueto**[3], **Jose A. Divino**[1]

[1]Department of Economics, Catholic University of Brasília, Brasília, Brazil

[2]Department of Statistics, University of Brasília, Brasília, Brazil

[3]FACE, Federal University of Goiás, Goiânia, Brazil

Resumo. *Este artigo apresenta uma alternativa à estimação univariada clássica da equação minceriana, propondo a estimação de regressões bivariadas para rendimento e horas trabalhadas. A modelagem inclui covariáveis tanto comuns quanto específicas para o vetor bivariado de variáveis dependentes. As estimações usam dados extraídos da Pesquisa Nacional por Amostras de Domicílio (PNAD) no período de 2013 a 2015. Dentre as distribuições utilizadas, os critérios de informação e a distância Mahalanobis indicaram que a distribuição t foi a que melhor se ajustou aos dados, tanto para modelos univariados quanto bivariados. Dentre as vantagens da regressão bivariada estão a modelagem com estrutura de correlação entre as variáveis dependentes, a identificação de efeitos antagônicos oriundos de covariáveis comuns sobre rendimento e horas trabalhadas e a flexibilidade para se assumir distintas distribuições bivariadas. Dentre os resultados obtidos, destaca-se que as covariáveis representativas de educação, tipo de contrato de trabalho e localização geográfica, na regressão bivariada, apresentaram sinais e magnitudes diferentes para rendimento e horas trabalhadas. Assim, o uso da modelagem bivariada surge como uma importante alternativa àquela tradicionalmente utilizada na estimação da equação minceriana de rendimento.*

Palavras-chave: Distribuições bivariadas; Equação minceriana; Rendimento; Horas trabalhadas; Regressão bivariada.

Abstract. *This paper proposes the estimation of bivariate regressions for earnings and hours worked as an alternative to the classical univariate estimation of the Mincerian earnings equation. The estimated models include both common and specific covariates for the bivariate vector of dependent variables and use data for the Brazilian economy extracted from the National Household Sample Survey (PNAD) from 2013 to 2015. Among the distributions used, the information criteria and Mahalanobis distance indicated that the Student t distribution produced the best fit to the data, for both univariate and bivariate cases. Advantages of the bivariate regression include estimation with a correlation structure between the dependent variables, identification of antagonistic effects from common covariates on earnings and hours worked, and flexibility to assume different bivariate distributions. The results indicate that there is a positive and statistically significant correlation between earnings and hours worked. In addition, the covariates education, type of employment contract, and geographical location presented different signals and magnitudes for the estimated coefficients by the bivariate regression for earnings and hours worked. Thus, the bivariate approach emerges as an important alternative to the univariate estimation that is traditionally applied for the Mincerian earnings equation.*

Keywords: Bivariate distributions; Mincerian equation; Earnings; Hours worked; Bivariate regression.

Classificação JEL: J20; C50.

Área 8: Microeconomia, Métodos Quantitativos e Finanças

1 Introduction

The Mincerian earnings equation introduced by Mincer (1974) is the basis of a broad empirical literature on labor economics, including contributions by Senna (1976), Willis and Rosen (1979), Cain (1986), Garen (1984), Robinson and Tomes (1984), Blackburn and Neumark (1993), Weisberg (1995), and Card (1999). These studies generally seek to estimate the returns to education and experience on the earnings-hour received by the worker. Originally, Mincer proposed that the distribution of earnings of workers among their different occupations is positively related to the amount of investment made in human capital, which has an impact on productivity and economic growth.[1]

Initially, the Mincerian earnings equation was structured by a linear regression in which the earnings-hour variable is explained by schooling and experience. Following the model proposed by Mincer (1974), other explanatory variables were incorporated into the regression, such as individual characteristics of gender and race that are used to assess the presence of discrimination in the labor market. In Cain (1986), the model was expanded to incorporate other individual characteristics, such as marital status and region of residence, to capture regional effects on wage differentials.

When deciding to join the labor market, a worker chooses the number of hours that he will make available to the market. Sedlacek and Santos (1991) used data from the Brazilian National Household Sample Survey (PNAD) and analyzed the relationship between the husband's income and the labor supply by their female spouses. They observed that the higher the husband's income, the higher the reservation wage and the less likely their wives will work. Moreover, the younger and more children the family has, the less likely they are to join the labor market or, when they do, they offer fewer hours of work.

As far as estimation methods are concerned, since Mincer (1974), the literature has used the traditional ordinary least squares (OLS) method and its variants with instrumental variables, quantile regression, sample selection, and procedures based on maximum likelihood estimation [Chatterjee and Price (1991), Stapleton (1995), Heckman (1974), Heckman (1976), Heckman (1979), Garen (1984), Buchinsky (2001)]. In Brazil, the greater availability of microdata and the improvement of the computational capacity contributed to the expansion of the empirical evidence since 2000, highlighting Maciel et al. (2001), Sachcida et al. (2004), Menezes-Filho (2002), Giuberti and Menezes-Filho (2005), Madalozzo (2010).

A common factor in the literature is the use of earnings-hour as the dependent variable in the Mincerian equation. This variable, in general, is obtained by the simple division of the income earned by the hours worked in the period. Such an approach, however, implies the agglutination, in a single variable, of two distinct components, represented by earnings and hours worked, which should be modeled separately. The determinants of income, which appear in the Mincerian equation, and of labor supply, which affect the decision to participate in the labor market, are not necessarily the same, in either quantitative (magnitudes) or qualitative (signals) terms.

Hours worked and wage income have distinct determinants, a feature that is not captured by traditional estimates of the Mincerian equation that uses hourly wage as the dependent variable. The stock of human capital, measured by formal education and experience, for instance, tends to increase the workers' remuneration, but may also reduce the willingness to offer working hours in the labor market. Those who are more qualified might receive higher remuneration by working less hours than

[1] Human capital is understood as the set of attributes acquired by a worker through education, skill, and experience that favor work and production. This term was introduced by Mincer (1958) and later explored by Becker (1993) and Heckman et al. (2000), among others.

those who are less qualified. These antagonistic effects of education on wages and hours worked are not captured by the univariate estimates of the Mincerian earnings equation.

Therefore, there is a gap in the literature that this study seeks to fill. The common practice of using the earnings-hour dependent variable might hide the effects of covariates that would be distinct if separately assessed on wage and hours worked. In contrast to the classical approach, this paper aims to estimate a bivariate regression for the Mincerian equation considering earnings and hours worked as a bivariate vector of dependent variables. The regressions include both common and specific covariates for the bivariate vector of earnings and hours worked. The bivariate Normal, Student t, and Birnbaum-Saunders (BS)[2] distributions are used in the estimation. For the sake of comparison, the univariate Mincerian earnings equation will also be estimated, considering a single dependent variable represented by earnings per hour worked. Estimates will be made for the Brazilian economy using data extracted from the PNAD from 2013 to 2015.

Advantages of the bivariate regression approach include the possibility of modeling a correlation structure among the dependent variables. If there is a correlation, the estimation of univariate regressions separately for earnings and hours worked might provide biased results [Marchant et al. (2016)]. The bivariate framework allows to identify antagonistic effects of common covariates on the different dependent variables. Finally, there is flexibility to assume different bivariate distributions for the earnings and hours-worked model. As in Heckman (1976), the parameters will be estimated by maximum likelihood, which is efficient according to Mittelhammer et al. (2000). Thus, the bivariate model emerges as an important alternative to the univariate one that is traditionally used in the literature for the estimation of the Mincerian earnings equation.

The results indicate that some common explanatory variables have different signals and magnitudes of the estimated coefficients in the bivariate regression of earnings and hours-worked. Specifically, the estimated coefficients for education, type of employment contract, and geographical location have distinct signals and different magnitudes for earnings and hours-worked regressions. Considering education, for instance, more years in school imply in higher average wage and lower supply of hours in the labor market. In the univariate regression, however, only the positive impact of an additional year of study on the earnings per hour is observed. Furthermore, the bivariate model captures the correlation between the two dependent variables, which implies in a more robust estimation than separate univariate regressions. There are, therefore, important advantages associated to the bivariate approach when compared to the univariate regression, suggesting that the former is more suitable for the estimation of the Mincerian earnings equation.

The paper is organized as follows. Section 2 describes the empirical model, presents the database, reports, and discusses the results. Finally, the third section presents the concluding remarks.

2 Empirical approach

2.1 Empirical model

The Mincerian earnings equation is typically described by the following univariate regression:

$$\log(\boldsymbol{w}) = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{1}$$

[2]See Johnson et al. (1995), Balakrishnan and Lai (2009), Santos-Neto et al. (2012) and Saulo et al. (2018, 2019)

where $\log(\boldsymbol{w})$ is a vector with the logarithm of the wage per hour (dependent variable), $\boldsymbol{\gamma}$ is a vector of coefficients, $\boldsymbol{X}$ is a matrix of explanatory variables, such as schooling, experience, race, gender and others, and $\boldsymbol{\epsilon}$ is a random error vector. It is usually assumed that the error follows a normal distribution and the earnings per hour is calculated as:

$$\text{earnings-hour} = \text{monthly earnings}/(\text{hours worked } \times 4.33).$$

The differential of the present paper is to model the earnings equation (1) as a bivariate regression of earnings and hours worked separately in order to capture different effects of the explanatory variables on wages and labor supply. Furthermore, as earnings and hours worked are correlated, the bivariate regression is more appropriate than the univariate estimation of separate regressions.

In the bivariate environment, the model can be estimated as a vector of dependent variables $\mathbf{Y}_i = (Y_{1i}, Y_{2i})^\top$, where Y_{1i} is the wage in the main job and Y_{2i} represents the hours dedicated to the main job by each individual i. This vector might be modeled as a set $\mathbf{x}$ of explanatory variables by using one of the three bivariate distributions described in the Appendix with the respective sets of equations such that:

i) Bivariate Normal distribution:

$$\begin{aligned} \mathrm{E}[\log(Y_{1i})|\mathbf{X}_i = \mathbf{x}_i] = \mu_{1i} = \mathbf{x}_i^\top \boldsymbol{\beta}_1, \quad i = 1, \ldots, n, \\ \mathrm{E}[\log(Y_{2i})|\mathbf{X}_i = \mathbf{x}_i] = \mu_{2i} = \mathbf{x}_i^\top \boldsymbol{\beta}_2, \quad i = 1, \ldots, n; \end{aligned} \tag{2}$$

ii) Bivariate t distribution:

$$\begin{aligned} \mathrm{E}[\log(Y_{1i})|\mathbf{X}_i = \mathbf{x}_i] = \mu_{1i} = \mathbf{x}_i^\top \boldsymbol{\beta}_1, \quad i = 1, \ldots, n, \\ \mathrm{E}[\log(Y_{2i})|\mathbf{X}_i = \mathbf{x}_i] = \mu_{2i} = \mathbf{x}_i^\top \boldsymbol{\beta}_2, \quad i = 1, \ldots, n; \end{aligned} \tag{3}$$

iii) Bivariate BS distribution:

$$\begin{aligned} \mathrm{E}[Y_{1i}|\mathbf{X}_i = \mathbf{x}_i] = \mu_{1i} = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_1), i = 1, \ldots, n, \\ \mathrm{E}[Y_{1i}|\mathbf{X}_i = \mathbf{x}_i] = \mu_{2i} = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_2), i = 1, \ldots, n. \end{aligned} \tag{4}$$

Note that in the Normal and t cases, we assume that the dependent variables have bivariate log-normal and log-t distributions, which implies that the logarithm of the variables provides the Normal and t bivariate distributions, respectively [Vanegas and Paula (2016)]. For the bivariate BS distribution, it is not necessary to apply the logarithm due to the parameterization as a function of the averages of this distribution [Saulo et al. (2018, 2019)]. Based on the literature, we defined the set of covariates used in the regression estimations [Borjas (2012)]. We separated the covariates that affect both earnings and hours worked simultaneously from those that affect only one of these variables.

The common covariates, which affect both earnings and hours worked, are:

- Gender: is a dummy variable that assumes value 1 for men and 0 for women;
- Race: is a dummy variable that assumes value 1 for caucasians and 0 for non-caucasians;
- Marital status: is a dummy variable that assumes value 1 for married individuals and 0 for unmarried individuals;

- Age and Age^2: represent the age of the individual and its square, usually used as a proxy for experience;
- Education: this variable has as a proxy the formal years of study, ranging from 0 to 16 years;
- Category (high, high mean, mean, low mean, low): set of binary variables used to capture occupancy category, segmented according to socioeconomic criteria and having the low category as a reference [3];
- Employment contract (with an employment record card, with no employment record card, autonomous, civil servant): the variables "no employment record card", "autonomous" and "civil servant" are dummies that seek to capture the type of occupation of the individual in the labor market, having "individuals with an employment record card" as the base category;
- Metropolitan region (Belém-PA, Fortaleza-CE, Recife-PE, Salvador-BA, Belo Horizonte- MG, Rio de Janeiro-RJ, Curitiba-PR, Porto Alegre-RS, Brasília-DF and São Paulo-SP): set of binary variables that designate the metropolitan regions of residence of the individuals in the sample, taking São Paulo as the base category;
- Year (2013, 2014, and 2015): are time dummy variables for the years of the sample, having 2013 as the base year;
- Sector (agriculture, industry, construction, trade, food and others, education, health, and social services): dummy variables that seek to capture cluster effects by sector of activities of the individuals, having as reference individuals working in the public sector.

The covariates that affect only the earnings are:

- Trade Union: dummy variable that assumes value 1 for individuals who were associated with some trade union in the reference month and 0 for those who were not associated;
- Social Security: dummy variable that assumes value 1 for individuals who were taxpayers of some type of social security in the reference month and 0 for those who were not taxpayers;
- Experience: number of years in the main job, which can range from 0 to 56 years.

The covariates that affect only the hours worked are:

- Head: dummy variable to capture the condition of the individual in the household, assuming value 1 if the reference individual of the household is head of the family and value 0 otherwise (non-head);
- Minors: dummy variable used to capture if there are children under 10 years old in the household;
- Inactivity: dummy variable that assumes value 1 for households that had unemployed people in the reference month and 0 for those households with no unemployed.

[3]The occupational classification is based on Jannuzzi (2001).

The database came from the PNAD in the period from 2013 to 2015. This survey is annually collected and published by the Brazilian Institute of Geography and Statistics (IBGE) and provides a wide set of demographic and socioeconomic information about the Brazilian population at the individual and household levels. Here, we consider a sample of individuals aged between 18 and 65 years with complete information on earnings and hours worked, totalizing 167,271 observations. The data refer to the 10 major metropolitan regions of the country, namely Belém-PA, Fortaleza-CE, Recife-PE, Salvador- BA, Belo Horizonte-MG, Rio de Janeiro-RJ, Curitiba-PR, Porto Alegre-RS, Brasília-DF, and São Paulo-SP. The nominal values of earnings were deflated by the National Consumer Price Index (INPC). There is no inclusion of variables to control groups of individuals each year. Thus, the database is characterized as a pooled cross-section. All the results were obtained by the R statistical software [`https://www.r-project.org/`].

Table 1 provides some descriptive statistics for earnings and hours worked at level and logarithmic scales, including sample size, minimum and maximum values, average (avg), median, standard deviation (SD), coefficients of variation (CV), asymmetry (CA), and kurtosis (CK). These measures indicate that earnings in level has a high asymmetry and a significant kurtosis, suggesting that a distribution with asymmetry and heavy tails is better to fit the data. On the other hand, hours worked in level show low asymmetry and moderate kurtosis.The application of the logarithm tends to produce symmetry, especially in the case of earnings. Figure 1 shows histograms of earnings and hours worked at level and logarithmic scales.

Table 1: Descriptive statistics for earnings and hours worked (level and logarithmic scales).

Data	n	Min.	Max.	Median	Mean	SD	CV	CA	CK
Earnings	167,271	1.222	146,677.6	1,344.544	2,278.51	3,087.265	135.495%	7.056	116.514
Hours worked	167,271	1	98	40	39.909	11.554	28.951%	-0.550	3.445
log(Earnings)	167,271	0.201	11.896	7.204	7.204	0.801	10.895%	0.409	2.008
log(Hours worked)	167,271	0.000	4.585	3.689	3.605	0.526	14.589%	-4.132	21.874

Source: Authors' elaboration based on the PNAD-IBGE.

2.2 Investigation of the best fit

Initially, the Normal, t, and BS univariate regression models, for earnings and hours worked, as well as their bivariate counterparts are estimated to investigate the distribution that best fits the data in each case. Table 2 reports the values of the Akaike (AIC) and Bayesian (BIC) information criteria, calculated as:

$$\text{AIC} = -2\ell + 2k \quad \text{and} \quad \text{BIC} = -2\ell + k\log(n),$$

where ℓ is the value of the log-likelihood function, k denotes the number of parameters, and n indicates the number of observations. These information criteria were used to select the models. According to Table 2, the univariate and bivariate models based on the t distribution yielded the best adjustments, as they resulted in the lowest values for the AIC and BIC. Thus, among the 3 distributions, the univariate and bivariate models of the t distribution shall be used according to the information criteria.

Once the best univariate and bivariate models were chosen, another evaluation of the quality of the fit was performed by the Mahalanobis distance, as proposed by Marchant et al. (2016). In the case of

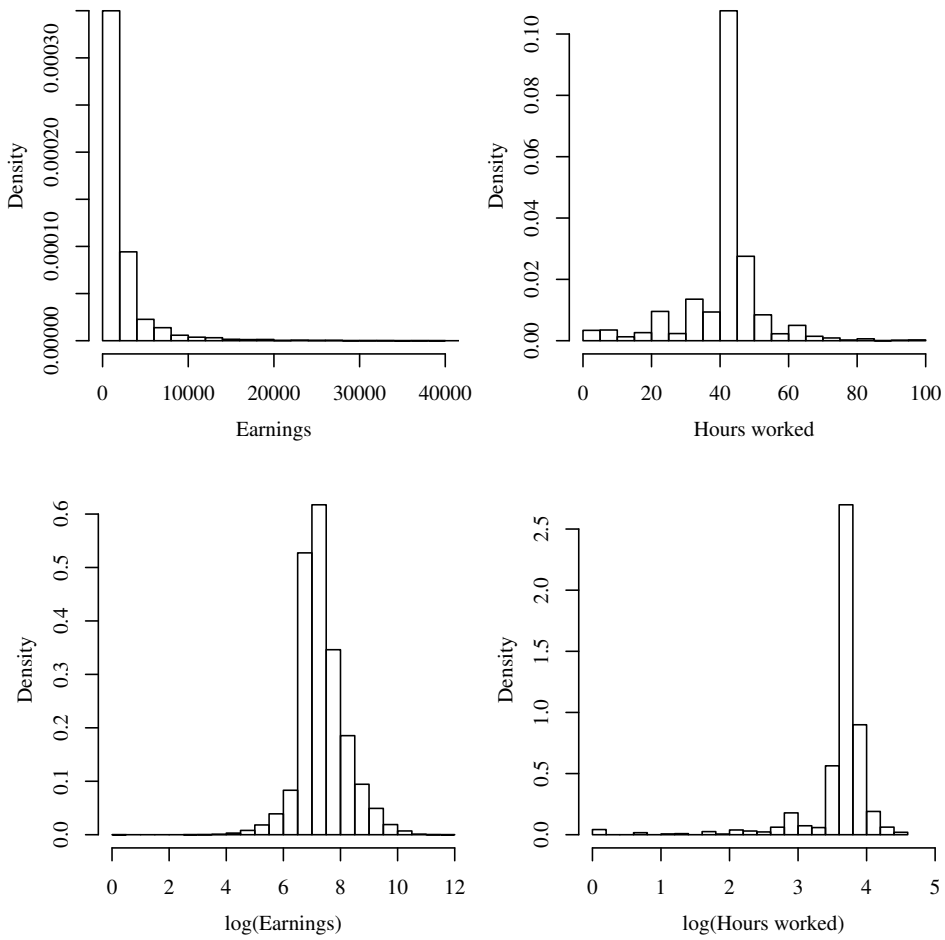

Figure 1: Histogram for earnings and hours worked (level and logarithmic scales).

Table 2: Information criteria for the univariate and bivariate models.

		Univariate models		
		Normal	t	BS
AIC	earnings	272,071.9	255,565	2,753,495
	hours worked	249,342	60,438.3	1,539,515
BIC	earnings	272,422.9	255,936	2,753,846
	hours worked	249,692.9	60,809.31	1,539,866
		Bivariate models		
		Normal	t	BS
AIC	earnings and hours worked	515,256.9	451,026	4,281,716
BIC	earnings and hours worked	515,968.8	451,737.9	4,282,428

Source: Authors' elaboration based on the PNAD-IBGE.

the bivariate t distribution, this distance is given by:

$$D = \frac{1}{2}(\boldsymbol{U} - \boldsymbol{\mu})^\top \boldsymbol{\psi}^{-1}(\boldsymbol{U} - \boldsymbol{\mu}) \sim F_{2,\nu}, \tag{5}$$

where $\mathbf{U} \sim t\text{Biv}(\mu_1, \mu_2, \sigma_1, \sigma_2, \nu, \rho)$ according to equation (14) of the Appendix, and $\boldsymbol{\psi}$ is the covariance matrix. According to (5), the Mahalanobis distance in the case of the bivariate t distribution follows a $F_{2,\nu}$ distribution. That is, F with 2 and ν degrees of freedom. In the univariate case, we have a $F_{1,\nu}$. In order to obtain the values of the Mahalanobis distance, the parameters are replaced by their maximum likelihood estimators, which results asymptotically in the same distribution in (5) [Vilca et al. (2014)]. The Wilson-Hilferty approximation can then be applied at the Mahalanobis distance to obtain a standard approximate Normal distribution in (5). Thus, the quality of the fit of the univariate and bivariate t regression models might be evaluated by the transformed distances with the Wilson-Hilferty approximation [Ibacache-Pulgar et al. (2014)]. In these cases, the distances in (5) are adapted to accommodate the regressive structure and the univariate or bivariate condition.

Figure 2 shows the probability-probability (PP) plots of the transformed Mahalanobis distance for the univariate t regression models of earnings and hours worked. The PP plot is commonly used to assess how close 2 sets of data are, which is done by plotting the 2 corresponding cumulative distribution functions. The closer the points are from the 45$\hat{0}$ line from (0.0) to (1.1), the best is the fit. Figure 2, shows the cumulative distribution function of the standard Normal versus the empirical cumulative distribution function of the transformed Mahalanobis distance. The results reveal a good fit of the univariate models.

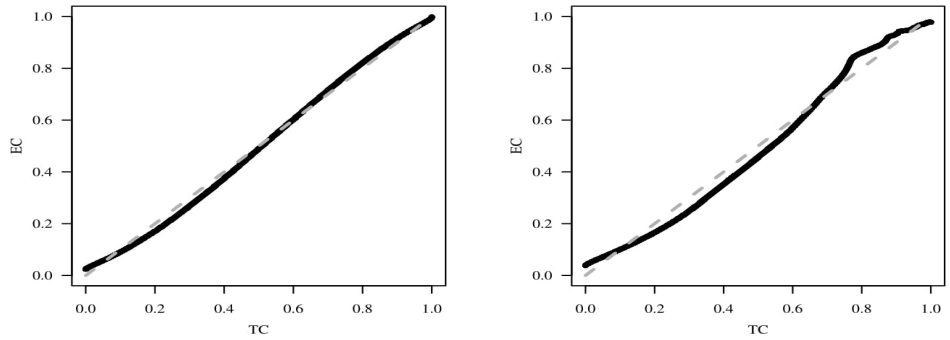

Figure 2: PP plots of the transformed distances for the univariate t regression models of earnings (left) and hours worked (right). Legend: EC = empirical probability, TC = theoretical probability.

Figure 3 shows the PP plot of the transformed Mahalanobis distance for the bivariate t regression model of earnings and hours worked. Notice that there is a good fit for the bivariate case. The results suggest that in both univariate and bivariate cases, the t models provide good adjustments and can therefore be used.

2.3 Estimations and analyses

Table 3 reports the results of the maximum likelihood estimation for the bivariate t distribution regression model of earnings and hours worked, with the respective standard errors, Wald statistics, and p-values. The model based on the t distribution presented the best fit according to the AIC and BIC information criteria and the PP plot of the Mahalanobis distance reported in the previous section. The Wald statistic is used to test the following hypotheses: H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$. The Wald

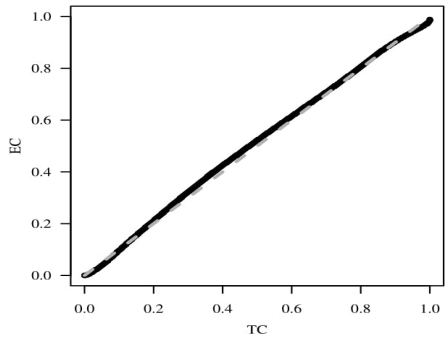

Figure 3: PP plots of the transformed distance for the bivariate t regression model of earnings and hours worked. Legend: EC = empirical probability, TC = theoretical probability.

statistic is defined by:

$$W = \frac{[\widehat{\theta} - \theta_0]}{\text{Standard Error}(\widehat{\theta})},$$

which is approximately distributed as a standard Normal under H_0, in which $\widehat{\theta}$ and θ_0 are the estimator and its proposed value under H_0, respectively. In this case, the interest lies in knowing if $\theta_0 = 0$ or H_0: $\theta = 0$ versus H_1: $\theta \neq 0$, at a significance level of $\alpha = 0.05$ (or 5%).

Regarding the interpretation of the estimated coefficients, attention should be paid to the following cases:

- When the independent variable x is quantitative (for instance, number of years in school) and the value of the coefficient estimated is: (i) out of range: $-0.05 \leq \widehat{\beta} \leq 0.05$, there is an increase (or decrease if the estimate is negative) of $(\exp(\widehat{\beta}) - 1) \times 100\%$ in the expected value (mean) of the dependent variable due to an increase of 1 unit in x; (ii) within the range $-0.05 \leq \widehat{\beta} \leq 0.05$, there is an increase (or decrease if the estimate is negative) $(\exp(\widehat{\beta}) - 1) \approx \widehat{\beta} \times 100\%$ in the expected value (mean) of the dependent variable when x increases by one unit.

- When the independent variable x is a dummy (for instance, gender) and the coefficient value is: (i) out of range $-0.05 \leq \widehat{\beta} \leq 0.05$, there is an increase (or decrease if the estimate is negative) of $(\exp(\widehat{\beta}) - 1) \times 100\%$ in the expected value (mean) of the dependent variable when x changes from 0 (women) to 1 (men); (ii) within the range $-0.05 \leq \widehat{\beta} \leq 0.05$, we have an increase (or decrease if the estimate is negative) $(\exp(\widehat{\beta}) - 1) \approx \widehat{\beta} \times 100\%$ in the expected value (mean) of the dependent variable when x changes from 0 (women) to 1 (men).

Table 3 shows that the estimated coefficient of correlation between earnings and hours worked is 0.1877 and statistically significant at the 5% significance level. This also indicates that the bivariate model is more appropriate than the univariate that estimates the equations separately, which could lead to incorrect predictions.

Table 3: Bivariate regression of earnings and hours worked with the bivariate t distribution ($\nu = 4$).

Dependent var.	Explanatory vars.	Coefficient	Standard error	Wald stat.	p-value
	σ_1	0.5456	0.0019	294.0464	<0.0001
	σ_2	0.5098	0.0018	292.0464	<0.0001
	ρ	0.1897	0.0131	14.5346	<0.0001
Earnings	(Intercept)	5.6253	0.0194	290.4145	<0.0001
	Gender	0.297	0.0031	96.2507	<0.0001
	Age	0.0514	8e-04	66.7815	<0.0001
	Age^2	-6e-04	8e-06	-57.561	<0.0001
	Race	0.0982	0.0031	31.8546	<0.0001
	Marital status	0.0065	0.0069	0.935	0.3498
	Education	0.0468	5e-04	97.6625	<0.0001
	Belém-PA	-0.3187	0.0063	-50.8093	<0.0001
	Fortaleza-CE	-0.3502	0.0059	-59.2043	<0.0001
	Recife-PE	-0.3498	0.0057	-61.2782	<0.0001
	Salvador-BA	-0.312	0.0058	-53.8771	<0.0001
	Belo Horizonte-MG	-0.0931	0.0055	-17.0357	<0.0001
	Rio de Janeiro-RJ	-0.0794	0.0052	-15.3409	<0.0001
	Curitiba-PR	0.0121	0.0065	1.85	0.0643
	Porto Alegre-RS	-0.0884	0.0051	-17.2281	<0.0001
	Brasília-DF	0.0858	0.0063	13.571	<0.0001
	2014	0.0024	0.0033	0.7242	0.469
	2015	-0.0589	0.0034	-17.5054	<0.0001
	High	0.8855	0.0084	105.5256	<0.0001
	High mean	0.3708	0.0073	50.5346	<0.0001
	Mean	0.1862	0.0069	26.8458	<0.0001
	Low mean	0.0507	0.0068	7.4931	<0.0001
	Agricultural	-0.4309	0.018	-23.969	<0.0001
	Industrial activities	-0.2036	0.0087	-23.3776	<0.0001
	Construction	-0.0781	0.0093	-8.3839	<0.0001
	Trade, food and others	-0.2107	0.0083	-25.4561	<0.0001
	Education, health and social services	-0.2543	0.0079	-32.1732	<0.0001
	Other services	-0.214	0.0083	-25.8597	<0.0001
	No employment record card	-0.1952	0.0041	-47.2089	<0.0001
	Autonomous	-0.1373	0.0042	-32.6763	<0.0001
	Civil servant	0.2383	0.0072	33.2859	<0.0001
	Trade Union	0.122	0.0038	32.4575	<0.0001
	Social Security	0.3556	0.0072	49.3444	<0.0001
	Experience	0.0127	2e-04	59.777	<0.0001
Hours worked	(Intercept)	3.2761	0.0166	196.993	<0.0001
	Gender	0.0827	0.0027	30.554	<0.0001
	Age	0.0128	7e-04	18.7405	<0.0001
	age^2	-1e-04	0	-15.1547	<0.0001
	Race	2e-04	0.0027	0.0855	0.9318
	Marital Status	-0.0092	0.0061	-1.5092	0.1312
	Education	-9e-04	4e-04	-2.2331	0.0255
	Belém-PA	0.0036	0.0055	0.6579	0.5106
	Fortaleza-CE	0.0172	0.0052	3.3292	9e-04
	Recife-PE	-0.0149	0.005	-2.9866	0.0028
	Salvador-BA	-0.044	0.0051	-8.6772	<0.0001
	Belo Horizonte-MG	0.012	0.0047	2.5458	0.0109
	Rio de Janeiro-RJ	-0.0874	0.0045	-19.5936	<0.0001
	Curitiba-PA	-0.013	0.0057	-2.2867	0.0222
	Porto Alegre-RS	0.0171	0.0044	3.8516	1e-04
	Brasília-DF	-0.027	0.0054	-5.0099	<0.0001
	2014	0.0119	0.0029	4.142	<0.0001
	2015	-0.0394	0.0029	-13.4526	<0.0001
	High	0.123	0.0073	16.8922	<0.0001
	High mean	0.0997	0.0066	15.067	<0.0001
	Mean	0.1465	0.0063	23.2992	<0.0001
	Low mean	0.155	0.0061	25.272	<0.0001
	Agricultural	0.1566	0.0152	10.3171	<0.0001
	Industrial activities	0.0204	0.0072	2.8169	0.0048
	Construction	0.0426	0.0078	5.4722	<0.0001
	Trade, food and others	0.0651	0.0069	9.4892	<0.0001
	Education, health and social services	-0.0658	0.0065	-10.0724	<0.0001
	Other services	-0.0147	0.0069	-2.1406	0.0323
	No employment record card	-0.1635	0.0036	-45.2849	<0.0001
	Autonomous	-0.1749	0.0034	-50.8516	<0.0001
	Civil servant	-0.041	0.006	-6.856	<0.0001
	Head	0.018	0.0027	6.7189	<0.0001
	Minors	-0.0028	0.0026	-1.1009	0.2709
	Inactivity	-0.0055	0.0037	-1.4954	0.1348

Source: Authors' elaboration based on the PNAD-IBGE.

Considering the estimated coefficients, the variable "Gender" indicates that men have an average income 34.58% higher than women, and that they offer 8.62% more hours worked, on average, than women do. On the other hand, the variable "Race" reveals that caucasian individuals earn, on average, 10.31% more than non-caucasians. However, when it comes to hours worked, caucasians only offer 0.02% more hours than non-caucasians. This result confirms previous studies that find discrimination in the Brazilian labor market. Cavalieri and Fernandes (1998), for example, found discrimination using data from the 1989 PNAD. They found that the income of men was higher than that of women and that the income of caucasian individuals was higher than that of non-caucasians, even after including control variables for age, number of years in school, and region of residence.

The "Age" variable indicates that for each additional year of age there is an increase of 5.27% in the average income, while only 1.28% is increased in hours worked. Regarding the variable "Education", an increase of one year of study causes an increase of 4.68% in the average income. However, in terms of hours worked, this same increase in schooling leads to an average decrease of 0.09%. Thus, the higher the individual's schooling, the higher their average earnings and the lower their number of hours of work. This illustrates a fundamental advantage of the bivariate regression, since the effect of "Education" is distinct in both regressions and this is not captured by the traditional univariate estimation that considers a unique dependent variable. Lau et al. (1993) also found a positive effect of "Education" on earnings (*per capita*) due to the higher schooling. Regarding hours worked, Gonzaga et al. (2002) argue that, in Brazil, schooling is inversely related to hour of work.

Considering the metropolitan regions, Brasília-DF presents an average income 8.95% higher and offers 2.70% less of hours of work than São Paulo-SP. Here, it is also possible to identify distinct effects of an explanatory variable in the bivariate regression that cannot be captured by the traditional univariate model. In order to explain this result, the unobservable characteristics of the workers, such as skill, motivation, among others, and the specific differences of the sectors of activity and the geographical regions of the country should be observed. In this specific case, the differential is due to the location of the federal public administration in Brasília. The federal public servants receive salaries well above the average of the private sector.

Regarding the types of labor contracts and considering individuals "with employment record card" as the reference category, the estimates point out that those with "no employment record card" have an average wage 17.73% lower. In addition, they offer about 15.08% less hours worked than their peers "with employment record card". The "civil servant" category incorporates, on average, an increase of 26.90% in wage while offering an average of 4.10% less hours worked in relation to the workers "with employment record card". Meanwhile, those who work for "own account" have an average wage 12.82% lower and offer 16.04% less hours to the labor market than the reference category. It is worth mentioning that the variable "civil servant" also present antagonistic effects on earnings and hours worked, which might be captured only by bivariate estimation.

For variables that affect only earnings or hours worked, Table 3 shows that individuals who contribute to social security have an average earning 42.70% higher than those who do not contribute. The "head" variable, which affects only hours worked, confirms that heads of households offer 1.80% more hours of work on average than those who are not in this condition.

For comparison purposes, Table 4 presents the results of the traditional univariate regression in which the dependent variable is earnings per hour, as commonly used in the Mincerian regression. In principle, some results show similarity in terms of magnitude with the estimates of the bivariate regression model. However, as already pointed out, the univariate model cannot disentangle the effects of a given explanatory variable on earnings and hours worked, as were the cases of "Education",

"Brasília-DF", and "Civil Servant" in the bivariate regression. These variables reported different signals for earnings and hours worked regressions. For "Education", for instance, the higher the individual's years of schooling, the higher his average income and the lower the quantity of hours worked. However, in the univariate regression reported in Table 4, only the positive impact of an additional year of study on the average earnings per hour can be captured. In addition, the bivariate model captured a positive and significant correlation between the 2 dependent variables, which allows for a more accurate estimation than the adjustment of 2 independent regressions. Therefore, there are important advantages of the bivariate model, including the evidence that the determinants of earnings and hours worked might not be the same in both quantitative and qualitative terms. Thus, the bivariate estimation appears as an relevant alternative for the estimation of the Mincerian earnings equation.

Table 4: Univariate regression of earnings-hour.

Depend. variable	Explanatory variables	Coefficient	Standard error	Wald	p-value
Earnings-hour	(Intercept)	0.9428	0.0229	41.23	<0.0001
	Gender	0.2027	0.0037	54.42	<0.0001
	Age	0.0347	0.0009	36.99	<0.0001
	Age^2	-0.0004	0.0001	-30.31	<0.0001
	Race	0.1005	0.0037	26.91	<0.0001
	Marital status	0.0097	0.0084	1.16	0.2477
	Education	0.0479	0.0006	85.13	<0.0001
	Belém-PA	-0.3155	0.0076	-41.37	<0.0001
	Fortaleza-CE	-0.3673	0.0072	-51.15	<0.0001
	Recife-PE	-0.3321	0.0069	-47.83	<0.0001
	Salvador-BA	-0.2674	0.0070	-38.33	<0.0001
	Belo Horizonte-MG	-0.1046	0.0067	-15.73	<0.0001
	Rio de Janeiro-RJ	0.0060	0.0062	0.97	0.3332
	Curitiba-PR	0.0205	0.0080	2.57	0.0101
	Porto Alegre-RS	-0.1097	0.0062	-17.58	<0.0001
	Brasília-DF	0.1140	0.0075	15.23	<0.0001
	2014	-0.0087	0.0040	-2.15	0.0312
	2015	-0.0159	0.0041	-3.90	0.0001
	High	0.7637	0.0097	78.59	<0.0001
	High mean	0.2750	0.0088	31.24	<0.0001
	Mean	0.0399	0.0083	4.79	<0.0001
	Low mean	-0.1060	0.0081	-13.04	<0.0001
	Agricultural	-0.5832	0.0200	-29.11	<0.0001
	Industrial activities	-0.2242	0.0101	-22.24	<0.0001
	Construction	-0.1194	0.0109	-11.00	<0.0001
	Trade, food and others	-0.2780	0.0095	-29.25	<0.0001
	Education, Health and Social Services	-0.1875	0.0090	-20.87	<0.0001
	Other services	-0.2004	0.0094	-21.24	<0.0001
	No employment record card	-0.0343	0.0049	-7.00	<0.0001
	Autonomous	0.0397	0.0047	8.51	<0.0001
	Civil servant	0.2880	0.0083	34.59	<0.0001
	Trade Union	0.1199	0.0047	25.71	<0.0001
	Social Security	0.3318	0.0081	40.86	<0.0001
	Experience	0.0104	0.0002	42.06	<0.0001
	Head	0.0644	0.0036	17.92	<0.0001
	Minors	0.0122	0.0036	3.35	0.0008
	Inactivity	-0.0618	0.0052	-11.93	<0.0001

Source: Authors' elaboration based on the PNAD-IBGE.

3 Conclusion

This paper proposed an alternative approach to estimate the Mincerian earnings equation based on bivariate regression modeling. The combination of earnings and hours worked in a single dependent variable, as traditionally is done in the empirical literature, prevents capturing distinct effects of common covariates on those dependent variables. On the other hand, the univariate estimation of

independent regression for earnings and hours worked is not indicated due to the correlation between these variables, which might generate biased estimates. Thus, we estimated a regression for earnings and hours worked as a bivariate vector of dependent variables, including common and specific covariates among the explanatory variables and using the Normal, Student t and BS bivariate distributions. The estimates were made for the Brazilian economy using data at the individual level extracted from the PNAD-IBGE survey for the years from 2013 to 2015, totalizing 167,271 observations.

In the bivariate case, the Normal, t and BS distributions were used to jointly model earnings and hours worked. The AIC and BIC information criteria and the Mahalanobis distance indicated that the t distribution yielded the best fit to the estimated models. In addition, a positive and statistically significant correlation was observed between earnings and hours worked, justifying the use of bivariate regression in detriment of independent regressions, which might yield biased estimates.

The bivariate estimation indicated that a given common covariate might generate distinct effects on earnings and hours worked. The results for "Education", for instance, indicated that an additional year of study leads to an average increase of 4.68% in earnings and an average decrease of 0.9% in hours worked. This suggests that individuals with more years of schooling, on average, have higher wages and work less than those with less years of schooling. Other covariates common to the bivariate vector, such as type of employment contract and geographic location, also had antagonistic effects on earnings and hours worked. This evidence illustrates a fundamental advantage of the bivariate regression, which allows to disentangle the distinct effects of a given common covariates on earnings and hours worked. This cannot be done by the traditional estimation of the univariate regression that considers the earnings per hour as a sole dependent variable.

Thus, the bivariate regression might be considered as an alternative approach for the estimation of the Mincerian earnings equation. As further work, one might implement the Heckman two-step correction for selection bias (Heckman, 1979), since the PNAD survey refers to individuals who were actually working in the sample period. However, the individual's earnings are associated with the decision to offer work, which ultimately depends on their opportunity cost. It is advantageous to work if the wage received (or potential earnings) are greater than the opportunity cost (reservation wage). In addition, other bivariate probability distributions might be adjusted for modeling earnings and hours worked, such as Pareto and its extensions, which are commonly used for income modeling. Finally, a bivariate logistic regression model could be used to estimate the influence of individual characteristics on the probabilities of a given worker belonging to a particular income group and type of work. Some of these extensions are object of ongoing research.

References

Balakrishnan, N. and Lai, C.-D. (2009). *Continuous Bivariate Distributions*. Springer, New York.

Barros, M., Paula, G., and Leiva, V. (2008). A new class of survival regression models with heavy-tailed errors: robustness and diagnostics. *Lifetime Data Analysis*, 14:316–332.

Becker, G. (1993). *Human capital a theoretical and empirical analysis, with special reference to education*. NBER, New York.

Birnbaum, Z. and Saunders, S. (1969). A new family of life distributions. *Journal of Applied Probability*, 6:319–327.

Blackburn, M. L. and Neumark, D. (1993). Omitted-ability bias and the increase in the return to schooling. *Journal of Labor Economics*, 11:521–544.

Borjas, G. J. (2012). *Economia do trabalho*. Artmed, São Paulo, 5 edition.

Buchinsky, M. (2001). Quantile regression with sample selection: estimating women's return to education in the u.s. *Empirical Economics*, 26:87–113.

Cain, G. (1986). The economic analysis of labor market discrimination: a survey. In Ashenfelter, O. and Layard, R., editors, *Handbook of Labor Economics*. North-Holland, 1 edition.

Card, D. (1999). The causal effect of education on earnings. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 3, Part A, chapter 30, pages 1801–1863. Elsevier, 1 edition.

Cavalieri, C. H. and Fernandes, R. (1998). Diferenciais por gênero e cor: uma comparação entre as regiões metropolitanas brasileiras. *Revista de Economia Política*, 18.

Chatterjee, S. and Price, B. (1991). *Regression Analysis by Example*. John Wiley, New York.

Garen, J. (1984). The returns to scholling: A selectivity bias approach with a continuous choice variable. *Econometrica*, 52:1199–1218.

Giuberti, A. C. and Menezes-Filho, N. (2005). Discriminação de rendimentos por gênero: uma comparação entre o Brasil e os Estados Unidos. *Economia Aplicada*, 9:369–384.

Gonzaga, G., Leite, P., and Machado, D. (2002). Quem trabalha muito e quem trabalha pouco no brasil? In *Anais do XIII Encontro Nacional De Estudos Populacionais*. ABEP.

Heckman, J. (1974). Shadow prices, market wages and labor supply. *Econometrica*, 42:679–694.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5:475–492.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47:153–161.

Heckman, J., Tobias, J., and Vytlacil, E. (2000). Simple estimators for treatment parameters in a latent variable framework with an application to estimating the returns to shooling. *NBER Working Paper*, 7950.

Ibacache-Pulgar, G., Paula, G., and Galea, M. (2014). On influence diagnostics in elliptical multivariate regression models with equicorrelated random errors. *Statistical Modelling*, 16:14–21.

Jannuzzi, P. d. M. (2001). Status socioeconômico das ocupações brasileiras: medidas aproximativas para 1980, 1991 e anos 90. *Revista Brasiliera de Estatística*, 2:47–74.

Johnson, N., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, volume 2. Wiley, New York, US.

Lau, L., Jamison, D., Liu, S., and Riukin, S. (1993). Education and economic growth some cross-sectional evidence from brazil. *Journal of Development Economics*, 41:45–70.

Maciel, M., Campêlo, A., and Raposo, M. (2001). A dinâmica das mudanças na distribuição salarial e no retorno em educação para mulheres: uma aplicação de regressão quantílica. In *Anais do XXIX Encontro Nacional de Economia*, Salvador. ANPEC.

Madalozzo, R. (2010). Occupational segregation and the gender wage gap in Brazil: an empirical analysis. *Economia Aplicada*, 14:147–168.

Marchant, C., Leiva, V., and Cysneiros, F. J. A. (2016). A multivariate log-linear model for Birnbaum-Saunders distributions. *IEEE Transactions on Reliability*, 65:816–827.

Menezes-Filho, N. (2002). Equações de rendimentos: Questões metodológicas. In et al., C., editor, *Estrutura Salarial: Aspectos Conceituais e Novos Resultados para o Brasil*, pages 51–66. IPEA.

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of Political Economy*, 66:281–302.

Mincer, J. (1974). *Schooling, Experience, and Earnings*. National Bureau of Economic Research, Inc.

Mittelhammer, R. C., Judge, G. G., and Miller, D. J. (2000). *Econometric Foundations*. Cambridge University Press, New York, US.

Rieck, J. and Nedelman, J. (1991). A log-linear model for the Birnbaum-Saunders distribution. *Technometrics*, 3:51–60.

Robinson, C. and Tomes, N. (1984). Union wage differentials in the public and private sectors: a simultaneous equations. *Journal of Labor Economics*, 2:106–127.

Sachcida, A., Loureiro, P. R. A., and Mendonça, M. (2004). Um estudo sobre retorno em escolaridade no Brasil. *Revista Brasileira de Economia*, 58:249–265.

Santos-Neto, M., Cysneiros, F., Leiva, V., and Ahmed, S. (2012). On new parameterizations of the Birnbaum-Saunders distribution. *Pakistan Journal of Statistics*, 28:1–26.

Saulo, H., Leão, J., Vila, R., Leiva, V., and Tomazella, V. (2018). A bivariate Birnbaum-Saunders regression model of GLM type reparameterized by its mean applied to reliability data . *Under Review*.

Saulo, H., Leão, J., Vila, R., Leiva, V., and Tomazella, V. (2019). On mean-based bivariate birnbaum-saunders distributions: Properties, inference and application. *Communications in Statistics - Theory and Methods*, pages 1–25.

Sedlacek, G. and Santos, E. (1991). A mulher cônjuge no mercado de trabalho como estratégia de geração da renda familiar. *Pesquisa e Planejamento Econômico*, 21:449–470.

Senna, J. (1976). Escolaridade, experiência no trabalho e salários no brasil. *Revista Brasileira de Economia*, 30:163–193.

Stapleton, J. (1995). *Linear Statistical Models*. John Wiley, New York.

Vanegas, L. H. and Paula, G. A. (2016). Log-symmetric distributions: statistical properties and parameter estimation. *Brazilian Journal of Probability and Statistics*, 30:196–220.

Vilca, F., Balakrishnan, N., and Zeller, C. (2014). The bivariate sinh-elliptical distribution with applications to Birnbaum-Saunders distribution and associated regression and measurement error models. *Computational Statistics and Data Analysis*, 80:1–16.

Weisberg, J. (1995). Returns to Education in Israel: 1974 and 1983. *Economics of Education Review*, 14:145–154.

Willis, R. and Rosen, S. (1979). Education and self selection. *Journal of Political*, 87:7–36.

Appendix

Distributions and bivariate regression models

In the symmetric context, the bivariate Normal distribution has been intensely used in the literature [Johnson et al. (1995)]. However, an alternative symmetric to the bivariate Normal distribution is the Student t model, as in Johnson et al. (1995) and Balakrishnan and Lai (2009), which has heavier tails than the Normal bivariate distribution. This flexibility is important to accommodate observations with more outliers, which makes the t an alternative of interest. On the other hand, in the univariate asymmetric context, a distribution that has received considerable attention is the BS model, which was introduced by Birnbaum and Saunders (1969) whereby its genesis is motivated by material fatigue problems [Johnson et al. (1995)]. Recently, Saulo et al. (2019, 2018) proposed a bivariate BS distribution and its corresponding regression model, based on the univariate BS distribution reparameterized by the mean proposed by Santos-Neto et al. (2012). In this reparameterization, there is no need to transform the dependent variable to a logarithmic scale, which is an advantage since it can lead to difficulties in interpretation. In general terms, Normal, Student t, and BS bivariate distributions can be considered as good candidates in the context of modeling earnings and hours worked, since in the the Normal and t cases the logarithm of the data is used, i.e. the log-normal and log-t are considered for the level variables [Vanegas and Paula (2016)], and in the BS case, the data (asymmetric on the right) are used in the original scale. The Normal, Student t, and BS bivariate distributions and their respective regression models are presented in sequence.

Bivariate Normal distribution

Let $\mathbf{Y} = (Y_1, Y_2)^\top$ be a bivariate random vector following a bivariate normal distribution with means μ_1 e μ_2, and standard deviations σ_1 e σ_2. In addition to these 4 parameters there is a correlation coefficient ρ between Y_1 and Y_2 defined by $-1 < \rho < 1$. Therefore, denoting $\mathbf{Y} \sim$ NBiv$(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. The joint probability density function (PDF)of Y_1 and Y_2 can be written as:

$$f(y_1, y_2; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \tag{6}$$
$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(y_1-\mu_1)^2}{\sigma_1^2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2}\right]\right\}.$$

The joint PDF of the random vector $\mathbf{Z} = (Z_1, Z_2)^\top$ following a bivariate standard Normal distribution (means zero and variances one) is given by:

$$\phi_2(z_1, z_2; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(y_1^2 + y_2^2 - 2\rho y_1 y_2)\right\}. \tag{7}$$

The corresponding joint cumulative distribution function (CDF) associated with (6) is given by:

$$\Phi_2(z_1, z_2; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{z_1}\int_{-\infty}^{z_2} \exp\left\{-\frac{1}{2(1-\rho^2)}(u^2 + v^2 - 2\rho uv)\right\}. \tag{8}$$

When $\rho = 0$, i.e., when the Normal variables are uncorrelated, (6) can be expressed as the product of 2 Normal CDFs.

Normal bivariate regression model

Consider $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ such that $\mathbf{Y}_i = (Y_{1i}, Y_{2i})^\top$ follows a Normal bivariate model, i.e., $\mathbf{Y}_i \sim$ NBiv$(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Consider that there are r and s covariates, lets say $\mathbf{x}_i^{(1)} = (x_{i1}^{(1)}, x_{i2}^{(1)}, \ldots, x_{ir}^{(1)})^\top$ and $\mathbf{x}_i^{(2)} = (x_{i1}^{(2)}, x_{i2}^{(2)}, \ldots, x_{is}^{(2)})^\top$, associated with Y_{1i} and Y_{2i}, respectively, such that

$$\mathrm{E}[Y_{1i}|\mathbf{X}_i^{(1)} = \mathbf{x}_i^{(1)}] = \mu_{1i} = \mathbf{x}_i^{(1)\top}\boldsymbol{\beta}_1 = \beta_{11}x_{i1}^{(1)} + \beta_{12}x_{i2}^{(1)} + \cdots + \beta_{1l}x_{il}^{(1)}, \quad i = 1, \ldots, n, \tag{9}$$

$$\mathrm{E}[Y_{2i}|\mathbf{X}_i^{(2)} = \mathbf{x}_i^{(2)}] = \mu_{2i} = \mathbf{x}_i^{(2)\top}\boldsymbol{\beta}_2 = \beta_{21}x_{i1}^{(2)} + \beta_{22}x_{i2}^{(2)} + \cdots + \beta_{2l}x_{il}^{(2)}, \quad i = 1, \ldots, n, \tag{10}$$

where $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kl})$ is a vector of l unknown parameters, and $\mathbf{x}_i^{(k)}$ is the i-th line of matrix $\mathbf{X}^{(k)}$, whose dimension is $n \times l$, for $k = 1, 2$ and $l = r, s$. Thus, we have the following Normal bivariate model:

$$\begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_i^{(1)\top}\boldsymbol{\beta}_1 \\ \mathbf{x}_i^{(2)\top}\boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{pmatrix} \quad i = 1, \ldots, n, \tag{11}$$

where $(\epsilon_{1i}, \epsilon_{2i}) \sim$ NBiv$(0, 0, \sigma_1, \sigma_2, \rho)$, and they are independently distributed.

To estimate the unknown parameters σ_1, σ_2, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and ρ based on a random sample of size n, i.e., $\{(y_{1i}, y_{2i}, \mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}); i = 1, \ldots, n\}$, the maximum likelihood method is used. The likelihood and log-likelihood functions of the observed sample can be written respectively as

$$L = \prod_{i=1}^{n} f(y_{1i}, y_{2i}; \mu_{1i}, \mu_{2i}, \sigma_1, \sigma_2, \rho), \tag{12}$$

$$\ell = \sum_{i=1}^{n} \log(f(y_{1i}, y_{2i}; \mu_{1i}, \mu_{2i}, \sigma_1, \sigma_2, \rho)), \tag{13}$$

where f is the joint PDF of the bivariate normal distribution. The model parameter estimates must be obtained by maximizing the log-likelihood function (13). This is done by solving a nonlinear iterative optimization process, particularly the quasi-Newton Broyden-Fletcher-Goldfarb-Shanno (BFGS) method can be used. The BFGS method is implemented in R software (`http://cran.r-project.org`), using the `optim` and `optimx` functions.

Bivariate t distribution

Let $\mathbf{U} = (U_1, U_2)^\top$ be a bivariate random vector following a bivariate t distribution with location parameters μ_1 and μ_2, scale parameters σ_1 e σ_2, degrees of freedom ν, and correlation coefficient $-1 < \rho < 1$ between U_1 and U_2, denoted by $\mathbf{U} \sim t\text{Biv}(\mu_1, \mu_2, \sigma_1, \sigma_2, \nu, \rho)$. Therefore, the joint PDF of U_1 and U_2 is given by:

$$f(u_1, u_2; \mu_1, \mu_2, \sigma_1, \sigma_2, \nu, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \tag{14}$$
$$\times \left[1 + \frac{1}{\nu(1-\rho^2)}\left(\frac{(u_1-\mu_1)^2}{\sigma_1^2} + \frac{(u_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(u_1-\mu_1)(u_2-\mu_2)}{\sigma_1\sigma_2}\right)\right]^{-(\nu+2)/2}.$$

The joint PDF of the random vector $\mathbf{U} = (U_1, U_2)^\top$ following a standard bivariate t distribution is given by:

$$f(u_1, u_2, \nu, \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}}\left[1 + \frac{1}{\nu(1-\rho^2)}\left(u_1^2 + u_2^2 - 2\rho u_1 u_2\right)\right]^{-(\nu+2)/2}.$$

The corresponding joint CDF associated to (15) is:

$$F(u_1, u_2; \nu, \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{u_1}\int_{-\infty}^{u_2}\left[1 + \frac{1}{\nu(1-\rho^2)}\left(u^2 + v^2 - 2\rho uv\right)\right]^{-(\nu+2)/2}. \tag{15}$$

3.1 Bivariate t regression model

Bivariate t regression model

Consider $\mathbf{U}_1, \ldots, \mathbf{U}_n$ such that $\mathbf{U}_i = (U_{1i}, U_{2i})^\top$ follows a bivariate t distribution, i.e., $\mathbf{U}_i \sim t\text{Biv}(\mu_1, \mu_2, \sigma_1, \sigma_2, \nu, \rho)$ with PDF (14). Assume that there are r and s covariates, $\mathbf{x}_i^{(1)} = (x_{i1}^{(1)}, x_{i2}^{(1)}, \ldots, x_{ir}^{(1)})^\top$ and $\mathbf{x}_i^{(2)} = (x_{i1}^{(2)}, x_{i2}^{(2)}, \ldots, x_{is}^{(2)})^\top$ say, associated with U_{1i} and U_{2i}, respectively, such that

$$\text{E}[U_{1i}|\mathbf{X}_i^{(1)} = \mathbf{x}_i^{(1)}] = \mu_{1i} = \mathbf{x}_i^{(1)\top}\boldsymbol{\beta}_1 = \beta_{11}x_{i1}^{(1)} + \beta_{12}x_{i2}^{(1)} + \cdots + \beta_{1l}x_{il}^{(1)}, i = 1, \ldots, n, \tag{16}$$
$$\text{E}[U_{2i}|\mathbf{X}_i^{(2)} = \mathbf{x}_i^{(2)}] = \mu_{2i} = \mathbf{x}_i^{(2)\top}\boldsymbol{\beta}_2 = \beta_{21}x_{i1}^{(2)} + \beta_{22}x_{i2}^{(2)} + \cdots + \beta_{2l}x_{il}^{(2)}, i = 1, \ldots, n, \tag{17}$$

where $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kl})$ is a vector of l unknown parameters, and $\mathbf{x}_i^{(k)}$ is the i-th line of matrix $\mathbf{X}^{(k)}$, whose dimension is $n \times l$, for $k = 1, 2$ and $l = r, s$. Therefore, we have the following bivariate t regression model

$$\begin{pmatrix} U_{1i} \\ U_{2i} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_i^{(1)\top}\boldsymbol{\beta}_1 \\ \mathbf{x}_i^{(2)\top}\boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{pmatrix} \quad i = 1, \ldots, n, \tag{18}$$

where $(\epsilon_{1i}, \epsilon_{2i}) \sim t\text{Biv}(0, 0, \sigma_1, \sigma_2, \nu, \rho)$ are independently distributed.

The parameters of the model are estimated as the bivariate normal, that is, given as likelihood and

log-likelihood function,

$$L = \prod_{i=1}^{n} f(u_{1i}, u_{2i}; \mu_{1i}, \mu_{2i}, \sigma_1, \sigma_2, \nu, \rho), \tag{19}$$

$$\ell = \sum_{i=1}^{n} \log(f(u_{1i}, u_{2i}; \mu_{1i}, \mu_{2i}, \sigma_1, \sigma_2, \nu, \rho)), \tag{20}$$

respectively, where f is the joint PDF of the bivariate t distribution, the model parameter estimates, σ_1, σ_2, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ e ρ, are obtained by maximizing the log-likelihood function by solving an iterative nonlinear optimization process, particularly the quasi-Newton BFGS method. The parameter ν is estimated according to Barros et al. (2008). The profiled log-likelihood and the following steps are used:

1) Let $\nu_k = k$ be for each k, $k = 1, \ldots, 20$, compute the parameter estimates of $(\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \sigma_1, \sigma_2, \nu, \rho)^\top$ using the maximum likelihood method. Moreover, compute the log-likelihood function;

2) The final estimate of ν is the one which maximizes the log-likelihood function and the associated estimates of $(\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \sigma_1, \sigma_2, \nu, \rho)^\top$ are the final ones.

Bivariate Birnbaum-Saunders (BS) distribution

Let $\mathbf{T} = (T_1, T_2)^\top$ be a bivariate random vector following a bivariate BS distribution parameterized by means with parameters μ_1, μ_2, δ_1, δ_2 e ρ. Therefore, the joint CDF of T_1 and T_2 can be written as (Saulo et al., 2019)

$$F(t_1, t_2; \mu_1, \mu_2, \delta_1, \delta_2, \rho) = \Phi_2\left(\sqrt{\tfrac{\delta_1}{2}}\left[\sqrt{\tfrac{(\delta_1+1)t_1}{\delta_1\mu_1}} - \sqrt{\tfrac{\delta_1\mu_1}{(\delta_1+1)t_1}}\right], \sqrt{\tfrac{\delta_2}{2}}\left[\sqrt{\tfrac{(\delta_2+1)t_2}{\delta_2\mu_2}} - \sqrt{\tfrac{\delta_2\mu_2}{(\delta_2+1)t_2}}\right]; \rho\right), t_1, t_2 > 0, \tag{21}$$

where $\mu_1 > 0$, $\delta_1 > 0$, $\mu_2 > 0$, $\delta_2 > 0$, $-1 < \rho < 1$, Φ_2 is the CDF of the standard bivariate distribution given in (8). Therefore, the joint PDF associated with (21) is given by

$$f(t_1, t_2; \mu_1, \mu_2, \delta_1, \delta_2, \rho) = \phi_2\left(\sqrt{\frac{\delta_1}{2}}\left[\sqrt{\frac{(\delta_1+1)\,t_1}{\delta_1\mu_1}} - \sqrt{\frac{\delta_1\mu_1}{(\delta_1+1)\,t_1}}\right], \sqrt{\frac{\delta_2}{2}}\left[\sqrt{\frac{(\delta_2+1)\,t_2}{\delta_2\mu_2}} - \sqrt{\frac{\delta_2\mu_2}{(\delta_2+1)\,t_2}}\right]; \rho\right) \frac{\sqrt{\delta_1}}{2\sqrt{2}t_1}\left[\sqrt{\frac{(\delta_1+1)\,t_1}{\delta_1\mu_1}} + \sqrt{\frac{\delta_1\mu_1}{(\delta_1+1)\,t_1}}\right] \frac{\sqrt{\delta_2}}{2\sqrt{2}t_2}\left[\sqrt{\frac{(\delta_2+1)\,t_2}{\delta_2\mu_2}} + \sqrt{\frac{\delta_2\mu_2}{(\delta_2+1)\,t_2}}\right], \tag{22}$$

where ϕ_2 is the PDF of a normal bivariate distribution given by (7). The bivariate BS distribution with PDF (22) is denoted by BSBiv$(\mu_1, \mu_2, \delta_1, \delta_2, \rho)$.

Bivariate Birnbaum-Saunders (BS) regression model

Consider $\mathbf{T}_1, \ldots, \mathbf{T}_n$ such that $\mathbf{T}_i = (T_{1i}, T_{2i})^\top$ follows a bivariate BS distribution, that is, $\mathbf{T}_i \sim$ BSBiv$(\mu_1, \mu_2, \delta_1, \delta_2, \rho)$. Assume that there is r and s covariates, $\mathbf{x}_i^{(1)} = (x_{i1}^{(1)}, x_{i2}^{(1)}, \ldots, x_{ir}^{(1)})^\top$ and

$\mathbf{x}_i^{(2)} = (x_{i1}^{(2)}, x_{i2}^{(2)}, \ldots, x_{is}^{(2)})^\top$, associated with T_{1i} and T_{2i}, respectively. Therefore, from (21), we have the joint CDF (Saulo et al., 2019)

$$F(t_{1i}, t_{2i}; \mu_{1i}, \mu_{2i}, \delta_1, \delta_2, \rho) = \Phi_2 \left(\sqrt{\frac{\delta_1}{2}} \left[\sqrt{\frac{(\delta_1+1)t_{1i}}{\delta_1 \mu_1^i}} - \sqrt{\frac{\delta_1 \mu_1^i}{(\delta_1+1)t_{1i}}} \right], \right. \tag{23}$$
$$\left. \sqrt{\frac{\delta_2}{2}} \left[\sqrt{\frac{(\delta_2+1)t_{2i}}{\delta_2 \mu_2^i}} - \sqrt{\frac{\delta_2 \mu_2^i}{(\delta_2+1)t_{2i}}} \right]; \rho \right), \quad t_{1i}, t_{2i} > 0 \quad i = 1, \ldots, n,$$

where

$$\mathrm{E}[T_{1i}|\mathbf{X}_i^{(1)} = \mathbf{x}_i^{(1)}] = \mu_{1i} = \exp(\mathbf{x}_i^{(1)\top}\boldsymbol{\beta}_1) = \exp(\beta_{11}x_{i1}^{(1)} + \beta_{12}x_{i2}^{(1)} + \cdots + \beta_{1l}x_{il}^{(1)}), i = 1, \ldots, n, \tag{24}$$
$$\mathrm{E}[T_{2i}|\mathbf{X}_i^{(2)} = \mathbf{x}_i^{(2)}] = \mu_{2i} = \exp(\mathbf{x}_i^{(2)\top}\boldsymbol{\beta}_2) = \exp(\beta_{21}x_{i1}^{(2)} + \beta_{22}x_{i2}^{(2)} + \cdots + \beta_{2l}x_{il}^{(2)}), i = 1, \ldots, n, \tag{25}$$

where $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kl})$ is a vector of unknown l parameters, and $\mathbf{x}_i^{(k)}$ is the i-th line of matrix $\mathbf{X}^{(k)}$, whose dimension is $n \times l$, for $k = 1, 2$ and $l = r, s$. Here, differently from the BS regression model based on the classical parameterization Rieck and Nedelman (1991), there is no need for logarithmic transformation, that is, the data for the response are worked on in their original scale.

In order to estimate the parameters, as in the normal bivariate case, the maximum likelihood method is used. Consider a random sample of size n, $\{(t_{1i}, t_{2i}, \mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}); i = 1, \ldots, n\}$ say, therefore the likelihood and log-likelihood functions of the observed sample are given respectively by

$$L = \prod_{i=1}^{n} f(t_{1i}, t_{2i}; \mu_{1i}, \mu_{2i}, \delta_1, \delta_2, \rho), \tag{26}$$

$$\ell = \sum_{i=1}^{n} \log(f(t_{1i}, t_{2i}; \mu_{1i}, \mu_{2i}, \delta_1, \delta_2, \rho)), \tag{27}$$

where f is a joint PDF of the bivariate BS distribution. The parameter estimates $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, δ_1, δ_2 and ρ are obtained by maximizing the log-likelihood function (27) using an iterative non-linear optimization process, in this case, the BFGS quasi-Newton method.