Economia Aplicada, v. 23, n. 2, 2019, pp. 127-144

THE WAGE PENALTY OF SMOKING IN BRAZIL: EVIDENCE FROM THE SPECIAL SURVEY ON TOBACCO ADDICTION

Marcelo Justus [*]
Elder G. Sant'Anna [†]
Eloá S. Davanzo [‡]
Gustavo C. Moreira [§]

Abstract

The aim of this study is to investigate the hypothesis that smoking reduces earnings. We use data from the Special Survey on Tobacco Addiction, which was jointly carried out with the 2008 Brazilian National Household Sample Survey. Our results support the hypothesis that smoking reduces wages. Furthermore, we found that about two-thirds of wage differential between male smokers and non-smokers is due to observable characteristics.

Keywords: smoking, tobacco, discrimination, human capital.

Resumo

O principal objetivo deste artigo é analisar a hipótese de que o tabagismo reduz os rendimentos do trabalho. Utilizamos dados da Pesquisa Especial de Tabagismo (PETab) realizada durante a PNAD 2008. Nossos resultados sustentam a hipótese de que fumar cigarro afeta negativamente o salário. Além disso, encontramos que as características observáveis explicam aproximadamente dois terços do diferencial salarial observado entre homens fumantes e não-fumantes.

Palavras-chave: tabagismo, tabaco, discriminação, capital humano.
JEL classification: I12, J24.
DOI: http://dx.doi.org/10.11606/1980-5330/ea142568

[*] Professor in the Intitute of Economics at the University of Campinas, São Paulo, Brazil. mjustus@unicamp.br.
[†] PhD student at the University of São Paulo, São Paulo. eldergenerozo@gmail.com.
[‡] PhD student in the Intitute of Economics at the University of Campinas, São Paulo, Brazil. eloadavanzo@gmail.com.
[§] Professor at the University of São João del-Rei, Minas Gerais, Brazil. gustavocmoreira@ufsj.edu.br.

Recebido em 23 de janeiro de 2018 . Aceito em 09 de outubro de 2018.

1 Introduction

Cigarettes are consumed by almost 1 billion adults in the world. Without a doubt smoking is associated with a higher risk of developing serious diseases such as cancer, emphysema and cardiovascular diseases. As a consequence, it causes the death of more than six million smokers every year. In Brazil, tobacco use often takes the form of consumption of manufactured cigarettes. In 2013, there were 21.5 million smokers in Brazil, 18.5 million of whom were daily smokers. According to the Brazilian Ministry of Health, about 200,000 deaths per year are related to tobacco consumption.

Regarding the private economic costs of smoking, a clear reduction in smoker's disposable income can be observed since smokers not only buy cigarettes, but may also be forced to spend money on medical treatment due to smoking. However, the costs associated with the effects of smoking on the labor market are not so evident.

There are few empirical studies published supporting the hypothesis that smoking can severely affect labor market outcomes through multiple channels such as wage (Van Ours 2004, Lye & Hirschberg 2004, Levine et al. 1997), absences (Leigh 1995, Ault et al. 1991), accidents (Leistikow et al. 2000), and less chance of participation (Lee et al. 1991). Indirect effects can also be caused by non-observed preferences and by the behavior of persistent smokers (Grafova & Stafford 2009). With regard to wage, these studies conclude from earnings equations where smoking is controlled. Ideally, these studies would have to solve two serious problems that arise in the attempt to identify the causal effect of smoking in wages, namely: sample selection - resulting from the decision to participate in the labor market - and smoking endogeneity. Undoubtedly, the endogeneity imposes a major difficulty in identifying the effect of smoking on earnings.

On the one hand, some empirical studies have found that smoking reduces earnings (Levine et al. 1997, Auld et al. 1998, Lee et al. 1999, Braakmann 2008, Anger & Kvasnicka 2010). On the other hand, other studies did not reject that smoking does not affect earnings (Van Ours 2004, Heineck & Schwarze 2003, Braakmann 2008). However, none of these studies solved both sample selection and smoking endogeneity problems at the same time.

The relationship between cigarette demand and income already evidenced in the literature (Levine et al. 1997) is a source of endogeneity between smoking and earnings. The endogeneity occurs because someone with a high intertemporal discount rate invests less in human capital and would be more predisposed to smoking (Almeida & Araújo Júnior 2017). As a result, the negative effect of such preferences on current consumption could be attributed to the fact that the individual is a smoker.

Recently, Almeida & Araújo Júnior (2017) found, using a instrumental quantile regression approach, that brazilians workers who smoke receive 15.2% to 36.5% less than others workers. Our study is based on the same Special Survey on Tobacco Addiction, but we consider gender differences and exploit a different set of instrumental variables. Another novel contribution of our approach lies in the decomposition of earnings applying the Oaxaca-Blinder decomposition to measure the wage gap between male smokers and non-smokers, which has been poorly investigated in the literature.

The aim of this study is to test the hypothesis that smoking reduces earnings. For this purpose, an empirical strategy to jointly deal with smoking

endogeneity and sample selection was applied. This reduction in wages can occur through mechanisms such as increased absenteeism, reduced productivity, and discrimination in the labor market. An important discussion on the discrimination hypothesis is presented by (Levine et al. 1997). The authors analyzed the effects of smoking on income and raised the hypothesis that discrimination occurred over the years as public intolerance to smoking became gradually stronger. Many employers institutionalized their own policies to ban smoking from their premises and some institutions adopted employment policies to hire non-smokers only. Those authors argue that in such scenario discriminatory employment practices can be adopted and reduce the wages of smokers and their expectations of employment.

This paper is sctructured as follows. Section 2 presents the data and sample. Section 3 describes the methodological procedures. Sections 4 and 5 show the main results and concluding remarks, respectively.

2 Data and Sample

We use data from the Special Survey on Tobacco Addiction (PETab, in the Brazilian acronym), which was jointly carried out with the 2008 Brazilian National Household Sample Survey (2008 PNAD, in the Brazilian acronym). The survey was conducted through a partnership between the Brazilian Institute for Geography and Statistics (IBGE in Brazilian acronym), the Ministry of Health, the National Cancer Institute, the Health Surveillance Secretariat and the National Health Surveillance Agency. It should be noted that the PETab survey is carried out in Brazil as part of an initiative of the World Health Organization and of the Centers for Disease Control and Prevention. This partnership was established with the aim of promoting part of a survey conducted in 14 countries, including Brazil, entitled Global Adults Tobacco Survey.

PNAD is a multipurpose random household survey that investigates several socioeconomic characteristics of the population, some on a permanent basis and others with variable periodicity, such as health status and smoking habit. An interesting advantage of this survey lies in its national coverage; additionally, this survey collects data on many other variables related to household structure and socioeconomic aspects of household members (labor, wage, education, housing characteristics, age, etc.). In addition to their representativeness at national level, this data covers several aspects related to tobacco addiction such as: tobacco use, attempts to quit smoking, exposure to tobacco, access to awareness-raising campaigns and perceptions about the risks of smoking, as well as aspects related to buying cigarettes and tobacco products.

The data was collected from a sub-sample of households surveyed through the 2008 PNAD, covering individuals aged 15 and above in about 51,000 Brazilian households. The individuals included in that sub-sample answered questions related to the use of tobacco products, their attempts to quit smoking, their exposure to smoke and their access to awareness-raising campaigns and to information on the risks of smoking, among other issues related to the main topic. For other people interviewed through the survey, information is also available on the habit of smoking, type of tobacco product used, and amount consumed.

To make the sample suitable for empirical modeling, we excluded all in-

dividuals under 18 years old or over 60 years old to reduce the labor market participation selection bias problem. Thus, our sample is restricted to individuals in the 18-60 age bracket; we also excluded individuals with ill-defined occupations, individuals who were still studying, individuals who worked but had no earnings, individuals who did not state their income, and individuals with wages in excess of R$ 100,000.00 (Brazilian currency).

After applying the mentioned filters to the sample and with missing values, our empirical exercises began with two subsamples: 95,626 women and 95,090 men. Sample expansion factors associated to each observation were used.

3 Methodology

We know that OLS estimates for the earnings equation may be biased due to an individual's decision to participate in the labor market. Thus, with the aim of identifying the effect of smoking, Heckman's procedure was applied to correct the sample selectivity bias (see Heckman 1979). Additionally, the smoker variable might be endogenous. This makes it more difficult to identify the hypothetical effect of smoking on earnings. Thus, an empirical strategy was applied to simultaneously address the sample selection bias resulting from the decision to participate in the labor market and the smoking endogeneity (see Wooldridge 2010, 567-570). As a final step, we are able to decompose any existent difference to measure the influence of observable and non-observable characteristics using the Oaxaca-Blinder decomposition method (Oaxaca 1973, Blinder 1973). We will now present each of the methodologies used, duly specifying the variables used in each model.

Heckit Estimator

The modelling exercise began with the standard linear regression model estimated by OLS:

$$y_{2i} = \boldsymbol{\beta}' x_i + \varepsilon_i \tag{1}$$

where y_{2i} is the logarithm of hourly earnings from the main job for individual i; x_i is a row vector containing a *dummy* variable labeled by smoker - which assumes value 1 if the individual is a smoker and 0 otherwise, other control variables (which will be described later) and a constant; $\boldsymbol{\beta}$ is a column vector of coefficients and ε_i is the random disturbance with $\varepsilon_i \sim N(0, \sigma_\varepsilon)$.

This earnings equation is separately estimated by gender. As usual in earnings equations, education was proxied by years of schooling. We also considered the existence of a threshold effect, besides the years of schooling variable based on previous evidence from Brazil found by Hoffmann & Simão (2009) and Justus et al. (2015). The returns on education are positive, suggesting that increases in earnings are substantially higher from 10 years of schooling. It should be noted that the first year of schooling yielding the highest return is that of the 11th grade, the last grade of high school. Therefore, we considered the existence of a threshold effect, besides the years of schooling variable, and included variable $S^\lambda = Z(S-\lambda)$ in the specification, where $\lambda = 10$ is the threshold, i.e. the value of schooling from which the return on education increases, and Z is a *dummy* variable that assumes value 0 for $S \neq \lambda$ and value 1 for $S > \lambda$.

Other control variables are experience in the labor market, usually measured using a typical mincer model by *proxy* defined by the difference between the actual age of the person and that at which he or she began to work, and the square of this variable; a *dummy* variable to distinguish between white (Caucasian, Asian people) and non-white (black, mulatto, indigenous people); a *dummy* variable to distinguish between residence in an urban or rural area; a *dummy* variable for labor union membership; two *dummy* variables to distinguish between three activity sectors: agriculture (base group), industry and services; three *dummy* variables to distinguish between three positions: employer, employee and self-employed (base group). Controls were included in all models for Brazil's 27 federated units (26 *dummy* variables).

However, the estimates obtained by OLS from equation (1) are biased since we only observe wages for those who are working. That is, wages are related to the decision to participate in the labor market or not, which can be denoted by:

$$\Pr[y_{1i} = 1] = \Phi(\boldsymbol{\beta}_1' x_1) \tag{2}$$

where y_{1i} is an indicator variable that assumes value one for those who work and zero otherwise, and x_1 is a vector of characteristics related to labor market participation.

Because OLS estimates for earnings equations may be biased due to an individual's decision to participate in the labor market, we apply the Heckit estimator (see Heckman 1979). This empirical strategy consists in: i) estimating the participation decision (equation 2), ii) obtaining $\phi(\hat{\boldsymbol{\beta}}_1' x_1)$ and $\Phi(\hat{\boldsymbol{\beta}}_1' x_1)$ through the estimated parameters from Equation 2, iii) calculating the estimate of the inverse Mills ratio, $\lambda(\hat{\boldsymbol{\beta}}_1' x_1) = \phi(\hat{\boldsymbol{\beta}}_1' x_1)/\Phi(\hat{\boldsymbol{\beta}}_1' x_1)$, and adding this estimate as regressor in Equation 1.

Thus, rewriting Equation 1, we have:

$$y_{2i} = \boldsymbol{\beta}' x_i + \sigma_{12}\lambda(\hat{\boldsymbol{\beta}}_1' x_{1i}) + \varepsilon_i \tag{3}$$

which can be estimated by OLS, generates a consistent estimator of beta and is identified without any restrictions in the regressors (Cameron & Trivedi 2009).

It is important to note that the participation equation - probit regression - contains the same regressors as the earnings equation, except for the *dummy* variables for labor union membership, position, and activity sectors. We also included other personal and family characteristics: a *dummy* variable for non-labor income (e.g. from conditional cash transfer programs), which is 1 if the person has such income and 0 otherwise; and a *dummy* variable for marital status, which is 1 if the man is married and 0 otherwise; a *dummy* variable for children living in the same household, which is 1 if there are children in the household and 0 otherwise; a *dummy* variable for position in the family, which is 1 if the man is the head of the family and 0 otherwise.

In Table 1, we present all the variables used in the two mentioned equations. In addition, the description of the variables contains a superscript whose purpose is to identify in which of the equations the variables are used.

Table 1: Definition, mean and standard deviation of variables

Variable	Variable Definition	Men		Women	
		Mean	Std. Dev.	Mean	Std. Dev.
Hourly Earnings	Logarithm of hourly earnings from the main job[a]	1.373	0.886	1.240	0.873
Smoker	1 if is smoker and 0 otherwise[ab]	0.247	0.431	0.152	0.359
Years of Schooling	Years of Schooling[ab]	7.565	4.312	8.285	4.303
S^{λ}	Threshold for schooling[ab]	0.382	0.486	0.460	0.498
Experience	Years of experience[a]	22.952	12.568	20.896	12.361
White	1 if is white and 0 otherwise[ab]	0.442	0.497	0.468	0.499
Urban	1 if lives in an urban area and 0 otherwise[ab]	0.852	0.355	0.904	0.295
Labor Union	1 if is a labor union membership and 0 otherwise[a]	0.192	0.394	0.156	0.363
Industry	1 if works in the industry sector and 0 otherwise[a]	0.171	0.376	0.141	0.348
Service	1 if works in the service sector and 0 otherwise[a]	0.675	0.468	0.830	0.376
Agriculture	1 if works in the agricultural sector and 0 otherwise[a]	0.154	0.361	0.029	0.167
Employer	1 if is an employer and 0 otherwise[a]	0.059	0.235	0.034	0.182
Employee	1 if is an employee and 0 otherwise[a]	0.694	0.461	0.787	0.409
Self-employement	1 if is self-employed and 0 otherwise[a]	0.247	0.432	0.178	0.383
Married	1 if is married and 0 otherwise[b]	0.652	0.476	0.608	0.488
Children	1 if has children and 0 otherwise[b]	0.522	0.500	0.641	0.480
Non Labor Income	1 if earns non-labor income and 0 otherwise[b]	0.036	0.186	0.069	0.253
Householder	1 if is householder and 0 otherwise[b]	0.632	0.482	0.311	0.463
Age	Age in years[b]	37.161	11.550	37.544	11.568
Works	1 if works and 0 otherwise[b]	0.862	0.345	0.573	0.495
Number of Smokers	Number of smokers in the household[c]	0.200	0.476	0.229	0.478
Respiratory Disease	1 if has been already diagnosed with asthma or bronchitis and 0 otherwise[c]	0.027	0.163	0.042	0.201
Observations		95,090		95,626	

Note: [a] indicates variables that were olny used in earning equations;
[b] indicates variables that were only used in selection equations;
[ab] indicates variables that were used in selection and earnings equations;
[c] indicates variables that were used instrumental variables; quadratic term were used for age in the selection equation and for experience in the earnings equation; *dummy* variables for Brazilian states were used in both equations.

Endogenous Explanatory Variable (IV-GMM)

As suggested in the literature, it is possible that the variable smoker is endogenous. This imposes more difficulty in identifying the hypothetical effect of smoking on earnings. Thus, we applied an empirical strategy to deal with the sample selection bias resulting from the decision to participate in the labor market and smoking endogeneity at the same time (see Wooldridge 2010, 567-570).

In addition to the participation equation and the earnings equation, there is another equation:

$$y_{2i} = \beta' x_i + \beta_s \text{smoke}_i + \varepsilon_i \tag{4}$$

$$\text{smoke}_i = \beta' x_i + \delta' z_i + v_i \tag{5}$$

$$\Pr[y_{1i} = 1] = \Phi(\beta_1' x_{1i} + \delta' z_i) \tag{6}$$

where z is a vector with two robust instruments for variable smoker, which is now treated as an endogenous variable: i) number of smokers living in the same household (number of smokers) and ii) a *dummy* variable to indicate whether the individual had been already diagnosed with asthma or bronchitis (respiratory disease). It is assumed that both variables are correlated with smoking but do not affect earnings.

Thus, as y_1 and smoker were always observed along with z, Equation 4 can be estimated by 2SLS controlling for inverse Mills ratio, which was obtained from Equation 6 since smoke is endogenous.

In short, the procedure was performed in three steps. First, the selection equation (participation equation) was estimated using all observations in the probit model and taking into account the two instruments cited, besides the previously mentioned regressors.

$$\Pr[y_{1i} = 1] = \Phi(\boldsymbol{\beta_1}' x_{1i} + \boldsymbol{\delta}' z_i) \tag{7}$$

Second, the estimated inverse Mills ratios for all observations were calculated based on this equation. Thus,

$$\lambda_2(\hat{\boldsymbol{\beta}}_{\mathbf{1}}' x_{1i} + \hat{\boldsymbol{\delta}}' z_i) = \phi(\hat{\boldsymbol{\beta}}_{\mathbf{1}}' x_{1i} + \hat{\boldsymbol{\delta}}' z_i)/\Phi(\hat{\boldsymbol{\beta}}_{\mathbf{1}}' x_{1i} + \hat{\boldsymbol{\delta}}' z_i) \tag{8}$$

Third, using the selected subsample for which wages and smoking were observed, we estimated the earnings equation also taking into account the inverse Mills ratios besides the controls variables cited previously. Thus,

$$y_{2i} = \boldsymbol{\beta}' x_i + \theta \widehat{\text{smoker}}_i + \sigma_{12} \lambda_2(\hat{\boldsymbol{\beta}}_{\mathbf{1}}' x_{1i} + \hat{\boldsymbol{\delta}}' z_i) + \varepsilon_i, \tag{9}$$

In the presence of heteroskedasticity, the GMM estimator is more efficient than the IV estimator (Baum et al. 2003). Thus, we estimated the parameters of the overidentified model using the optimal GMM.

Oaxaca-Blinder Decomposition

If smoking negatively affects earnings, then we are able to decompose this differential in order to measure the influence of observable and non-observable characteristics. Oaxaca-Blinder decomposition (Oaxaca 1973, Blinder 1973) for smoking and non-smoking individuals was applied for this purpose.

Decomposition is performed in two stages. In the first one, earning equations are estimated for each of the groups, labeled s, for smokers and ns for non-smokers. Once this is done, the difference between the logarithm of average earnings between workers in the two groups is calculated as

$$\begin{aligned} \mathrm{D} &= E(y_s) - E(y_{ns}) && (10)\\ &= E(\boldsymbol{\beta}_s' x_s + \varepsilon) - E(\boldsymbol{\beta}_{ns}' x_{ns} + \varepsilon) && (11)\\ &= E(x_s)' \boldsymbol{\beta}_s' - E(x_{ns})' \boldsymbol{\beta}_{ns}' && (12) \end{aligned}$$

where $E(\varepsilon) = 0$ was used. According to Jann (2008), this equation can be rearranged from a twofold decomposition as

$$D = [E(x_s) - E(x_{ns})]' \boldsymbol{\beta}^* + \left[E(x_s)'(\boldsymbol{\beta}_s - \boldsymbol{\beta}^*) + E(x_{ns})'(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{ns})\right] \tag{13}$$

where $\boldsymbol{\beta}^*$ represents a vector of coefficients related to non-discrimination, term $[E(x_s) - (x_{ns})]' \boldsymbol{\beta}^*$represents the earnings differential that is explained by the

mean observable characteristics of smoking and non-smoking individuals, and other component on the right side of the equation refers to the portion not explained by these characteristics.

As presented in Jann (2008), in the presence of sample selection it is necessary to deduct the effects of the sample selection from the total difference, and then apply decomposition. In practical terms, two possibilities are suggested: i) calculating the decomposition together with the Heckman procedure or ii) adjusting the decomposition with the inverse Mills ratio. In this paper, we chose to adjust the decompositions using the estimates: $\lambda(\hat{\boldsymbol{\beta}}_{\mathbf{1}}' x_1) = \phi(\hat{\boldsymbol{\beta}}_{\mathbf{1}}' x_1)/\Phi(\hat{\boldsymbol{\beta}}_{\mathbf{1}}' x_1)$. Note that since we are decomposing the wage gap between smokers and non-smokers and the procedure consists of estimating an income equation for each of the groups, it is not necessary to take the endogeneity between smoking and non-smoking and the income earned in the job market into account in the decomposition.

4 Results

Table 2 shows the earnings equations estimated by OLS, Heckman's procedure and IV-GMM with correction for sample selection bias. The selection equation estimates for Heckman's procedure and the results of the first-stage regression of the endogenous variable smoker are also presented.

In this study, we are interested in the variable smoker. However, it should be noted that for all control variables (e.g., schooling and experience) the results are the ones usually observed in studies on earnings determinants in the Brazilian labor market and international literature.

Since dependent variable is the natural logarithm of earnings, if c is the estimated value of the conditional marginal effect, the estimated percentage change in earnings due to change in a *dummy* variable is $[\exp(c)-1]\times 100$. Based on IV-GMM estimates with correction for sample selection bias, we found that smoking had a greater impact on earnings. Smoking reduces wages by 29.7 and 24.2% for men and women, respectively. This is a serious economic consequence of being a smoker.

The higher magnitude after controlling for smoking endogeneity was also verified in previous studies. In Auld et al. (1998) for example, control for simultaneity between wages and smoking suggests that smokers earn about 20% to 67% less than non-smokers. This incremental effect after using instrumental variables was also observed by Van Ours (2004) when analyzing men's earnings. It is worth remembering that none of these studies applied a correction for sample selection bias.

Our results for the reductions observed in the wages of smokers as compared to those of non-smokers are corroborated by the literature, which provides several examples of ways by which smoking influences labor income. Anger & Kvasnicka (2010) show that the wages paid to smokers can decrease due to their reduced productivity resulting from high rates of absenteeism and health problems or due to potential discrimination of smokers by employers and co-workers. Damages to one's health, however, can be irreversible. Smoking can therefore have a negative impact on both an individual's current capacity and on his or her wages in the future.

In relation to absenteeism, Halpern et al. (2001) show that the rate of absenteeism measured for workers who smoke currently was higher than that

Table 2: Earnings equations using OLS, Heckman's procedure and IV-GMM with correction for sample selection bias: Brazilian individuals aged from 18 to 60 years old, by gender

Variables	Men					Women				
	OLS	Heckit		IV-GMM		OLS	Heckit		IV-GMM	
		2^{nd} Stage	1^{st} Stage	2^{nd} Stage	1^{st} Stage		2^{nd} Stage	1^{st} Stage	2^{nd} Stage	1^{st} Stage
Smoker	−0.0633***	−0.0550***	−0.0747***	−0.2126***	−0.0688***	−0.0264***	−0.0245***	−0.0163	−0.4194***	−0.0206
	(0.0059)	(0.0061)	(0.0139)	(0.0353)	(0.0141)	(0.0091)	(0.0092)	(0.0136)	(0.0561)	(0.0137)
Years of Schooling	0.0816***	0.0793***	0.0363***	0.0773***	0.0346***	0.1049***	0.1085***	0.0710***	0.1032***	0.0706***
	(0.0011)	(0.0011)	(0.0015)	(0.0012)	(0.0016)	(0.0015)	(0.0019)	(0.0012)	(0.0020)	(0.0012)
S^{λ}	0.0408***	0.0446***		0.0357***		−0.0529***	−0.0515***		−0.0684***	
	(0.0082)	(0.0082)		(0.0084)		(0.0108)	(0.0108)		(0.0113)	
Experience	0.0373***	0.0315***		0.0323***		0.0262***	0.0281***		0.0293***	
	(0.0008)	(0.0009)		(0.0011)		(0.0009)	(0.0011)		(0.0011)	
Experience Squared	−0.0005***	−0.0004***		−0.0004***		−0.0003***	−0.0004***		−0.0004***	
	(0.0000)	(0.0000)		(0.0000)		(0.0000)	(0.0000)		(0.0000)	
White	0.1440***	0.1401***	0.0577***	0.1384***	0.0567***	0.1460***	0.1432***	−0.0377***	0.1371***	−0.0402***
	(0.0055)	(0.0055)	(0.0134)	(0.0055)	(0.0134)	(0.0067)	(0.0068)	(0.0105)	(0.0069)	(0.0104)
Urban	0.1165***	0.1466***	−0.4233***	0.1488***	−0.4105***	0.1328***	0.1396***	0.0672***	0.1525***	0.0664***
	(0.0085)	(0.0088)	(0.0204)	(0.0092)	(0.0203)	(0.0131)	(0.0132)	(0.0159)	(0.0135)	(0.0159)
Labor Union	0.1757***	0.1730***		0.1670***		0.2641***	0.2630***		0.2619***	
	(0.0068)	(0.0068)		(0.0069)		(0.0089)	(0.0089)		(0.0091)	
Industry	0.2747***	0.2714***		0.2662***		0.0284	0.0280		0.0044	
	(0.0101)	(0.0101)		(0.0103)		(0.0251)	(0.0251)		(0.0258)	
Service	0.2555***	0.2546***		0.2536***		0.1818***	0.1823***		0.1713***	
	(0.0093)	(0.0093)		(0.0094)		(0.0238)	(0.0238)		(0.0244)	
Employer	0.6447***	0.6441***		0.6375***		0.6795***	0.6786***		0.6710***	
	(0.0149)	(0.0149)		(0.0150)		(0.0245)	(0.0245)		(0.0249)	
Employee	0.0555***	0.0541***		0.0586***		0.0595***	0.0616***		0.0577***	
	(0.0070)	(0.0070)		(0.0070)		(0.0104)	(0.0104)		(0.0105)	

Robust standard errors in parentheses; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. *dummy* variables for Brazilian states were used.

Table 2: Earnings equations using OLS, Heckman's procedure and IV-GMM with correction for sample selection bias: Brazilian individuals aged from 18 to 60 years old, by gender (continuation)

Variables	Men					Women				
	OLS	Heckit		IV-GMM		OLS	Heckit		IV-GMM	
		2^{nd} Stage	1^{st} Stage	2^{nd} Stage	1^{st} Stage		2^{nd} Stage	1^{st} Stage	2^{nd} Stage	1^{st} Stage
Mills Ratio				−0.2368***					0.0335	
				(0.0307)					(0.0261)	
Married			0.4098***		0.3915***			−0.2428***		−0.2351***
			(0.0209)		(0.0212)			(0.0130)		(0.0127)
Children			0.0706***		0.0754***			−0.1338***		−0.1387***
			(0.0196)		(0.0198)			(0.0123)		(0.0121)
Non Labor Income			−0.4855***		−0.5890***			−0.3004***		−0.3028***
			(0.0311)		(0.0283)			(0.0193)		(0.0192)
Householder			0.3501***		0.3285***			0.2390***		0.2474***
			(0.0147)		(0.0150)			(0.0133)		(0.0129)
Age			0.1110***		0.1103***			0.1400***		0.1419***
			(0.0036)		(0.0036)			(0.0032)		(0.0031)
Age Squared			−0.0016***		−0.0016***			−0.0018***		−0.0018***
			(0.0000)		(0.0000)			(0.0000)		(0.0000)
Number of Smokers					−0.0405***					0.0132
					(0.0119)					(0.0102)
Respiratory Disease					−0.1359***					−0.0632***
					(0.0352)					(0.0234)
Constant	0.1028***	0.2095***	−0.9891***	0.2493***	−0.9160***	−0.1031***	−0.2189***	−2.7169***	−0.0843*	−2.7621***
	(0.0223)	(0.0235)	(0.0731)	(0.0252)	(0.0739)	(0.0325)	(0.0491)	(0.0644)	(0.0490)	(0.0610)
Number of Observations	81,974	81,974	95,090	81,974	95,090	54,772	54,772	95,626	54,772	95,626
GMM *C* (Difference-in-Sargan)				21.5960					54.7778	
Hansen's *J* Test				0.0169					0.3563	

Robust standard errors in parentheses; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. *dummy* variables for Brazilian states were used.

calculated for those who never smoked. It should be mentioned that in the group of individuals who were smokers in previous periods, absenteeism declined as they stopped smoking. With similar results, Weng et al. (2013) found that current smokers face a 33% higher risk of absenteeism than non-smokers. Those in the former group were absent from work for 2.64 more days per year on average than those in the latter.

In terms of productivity, smoking can reduce the net productivity of workers due to its effects on their ability to perform manual tasks (Levine et al. 1997) and to the high absenteeism rates recorded for smoking workers and/or their lower physical and mental resistance (Grafova & Stafford 2009). Considering subjective productivity (productivity as assessed by others and personal life satisfaction), Halpern et al. (2001) showed significant trends with higher figures for those who never smoked in their life, lower figures for current smokers, and intermediate figures for individuals who were smokers in previous periods.

In view of the evidence of effects of smoking on earnings in all the estimated models, we decomposed this differential in order to shows the influence of observable and non-observable characteristics. Oaxaca-Blinder decomposition (Oaxaca 1973, Blinder 1973) for smoking and non-smoking individuals was applied for this purpose. Table 3 shows the results.

Table 3: Oaxaca-Blinder decomposition for logarithm of hourly earnings

	Men		**Women**	
	OLS	**Heckit**	**OLS**	**Heckit**
Smokers	1.2159***	1.2670***	1.1302***	1.0346***
	(0.0067)	(0.0127)	(0.0106)	(0.0401)
Non-smokers	1.4555***	1.5330***	1.2826***	1.2345***
	(0.0039)	(0.0074)	(0.0044)	(0.0167)
Difference	−0.2396***	−0.2660***	−0.1524***	−0.1999***
	(0.0078)	(0.0146)	(0.0115)	(0.0434)
Explained	−0.1763***	−0.1735***	−0.1260***	−0.1312***
	(0.0054)	(0.0054)	(0.0077)	(0.0080)
Unexplained	−0.0633***	−0.0925***	−0.0264***	−0.0688
	(0.0059)	(0.0138)	(0.0091)	(0.0427)
Number of smokers	19,799	19,799	8,165	8,165
Number of non-smokers	62,175	62,175	46,607	46,607

Robust standard errors in parentheses; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Considering the results from decomposing the model with sample selection correction, around two-thirds of the wage differential between smoking and non-smoking men are due to their observable characteristics. The same proportion was found for women, but unobserved factors were not statistically significant for women. As far as we know, the only study that sought to analyze the decomposition of wage differentials between smokers and non-smokers was one conducted by Hotchkiss & Pitts (2013). The authors found that the differential between the groups was of about 24%, two-thirds of which were explained by differences in observable characteristics.

Our results corroborate by Becker & Murphy (1988) the theory of rational addiction, which suggests a higher intertemporal preference for the present

for individuals with an addiction of some kind. Considering their higher preference for the present, smokers have lower incentives to invest in human capital since they will not be able to enjoy its returns for the same period of time as non-smokers. There was virtually no change in the magnitude of wage differentials between smokers and non-smokers that can be explained by observable characteristics for both males and females and regardless of the model used for the decomposition.

Still regarding the decomposition, we could only observe a significant effect of non-observable characteristics on the earnings of smokers compared to non-smokers for men. This effect may result, for example, from a certain productive heterogeneity not controlled for by the model's exogenous variables or even from discrimination toward smokers in the labor market. This significance can be justified by studies that confirm that smokers generate higher costs for companies. Smoking workers can be more expensive for their employers due to their increased absenteeism, higher health insurance premium, higher maintenance costs, and negative effects on the company's image. Due to health problems associated with smoking, smokers themselves may prefer jobs that provide partial or full health insurance to higher-wage jobs (Levine et al. 1997).

Robustness Checks

Regarding our instrument variables, the coefficient for respiratory disease is highly significant statistically ($p < 0.001$) for both men ($\beta = -0.1359$) and women ($\beta = -0.0632$), as expected. The number of smokers variable is statistically significant at 1% level only in the first-stage equation estimated for men.

After the GMM estimation, we performed a robust test of endogeneity (orthogonality conditions). The GMM C statistic is χ^2 distributed with one degree of freedom, under the null hypothesis that the regressor is exogenous. We apply the test to our model with one potentially endogenous regressor, smoker, intrumented by number of smokers and respiratory disease. The stastistic is $\chi^2 = 21.60$ ($p = 0.0000$) and $\chi^2 = 54.78$ ($p = 0.0000$) for men and women, respectively. For both genders, the statistical test leads to the strong rejection of the null hypothesis that smoker is an exogenous variable in the earnings equations. We conclude that it is endogenous.

We also applied Hansen's J test to test the validity of the overidentified restrictions. The statistic is $\chi^2 = 0.02$ ($p = 0.8965$) and $\chi^2 = 0.36$ ($p = 0.5505$) for men and women, respectively. Because $p > 0.05$, we do not reject the null hypothesis. The failure to reject H_0 is interpreted as indicating that at least one of the instruments is valid. We conclude that overidentifying is valid. For details about this endogeneity and overidentification tests see (Cameron & Trivedi 2009).

Finally, we estimated the earnings equations once again disregarding variables related to activity sector and position in the occupation, as these are potentially endogenous characteristics, to check the robustness of the estimates. For all variables, the estimates are virtually the same as compared to those shown in Table 2. These additional results are shown in Table 4.

Table 4: Earnings equations without potential endogenous regressors using OLS, Heckman's procedure and IV-GMM with correction for sample selection bias: Brazilian individuals aged from 18 to 60 years old, by gender

Variables	Men					Women				
	OLS	Heckit		IV-GMM		OLS	Heckit		IV-GMM	
		2^{nd} Stage	1^{st} Stage	2^{nd} Stage	1^{st} Stage		2^{nd} Stage	1^{st} Stage	2^{nd} Stage	1^{st} Stage
Smoker	−0.0739***	−0.0654***	−0.0747***	−0.2286***	−0.0688***	−0.0257***	−0.0235**	−0.0160	−0.4447***	−0.0206
	(0.0061)	(0.0062)	(0.0139)	(0.0361)	(0.0141)	(0.0093)	(0.0094)	(0.0136)	(0.0570)	(0.0137)
Years of Schooling	0.0912***	0.0888***	0.0360***	0.0866***	0.0346***	0.1105***	0.1144***	0.0710***	0.1086***	0.0706***
	(0.0011)	(0.0012)	(0.0015)	(0.0012)	(0.0016)	(0.0015)	(0.0020)	(0.0012)	(0.0020)	(0.0012)
S^{λ}	0.0321***	0.0357***		0.0270***		−0.0513***	−0.0496***		−0.0673***	
	(0.0085)	(0.0085)		(0.0087)		(0.0110)	(0.0111)		(0.0116)	
Experience	0.0403***	0.0343***		0.0347***		0.0279***	0.0300***		0.0313***	
	(0.0008)	(0.0009)		(0.0011)		(0.0009)	(0.0011)		(0.0011)	
Experience Squared	−0.0005***	−0.0004***		−0.0004***		−0.0003***	−0.0004***		−0.0004***	
	(0.0000)	(0.0000)		(0.0000)		(0.0000)	(0.0000)		(0.0000)	
White	0.1616***	0.1577***	0.0580***	0.1552***	0.0567***	0.1562***	0.1530***	−0.0373***	0.1463***	−0.0402***
	(0.0056)	(0.0057)	(0.0134)	(0.0056)	(0.0134)	(0.0068)	(0.0069)	(0.0105)	(0.0071)	(0.0104)
Urban	0.2570***	0.2870***	−0.4234***	0.2910***	−0.4105***	0.1778***	0.1854***	0.0671***	0.1960***	0.0664***
	(0.0078)	(0.0081)	(0.0203)	(0.0085)	(0.0203)	(0.0129)	(0.0130)	(0.0159)	(0.0133)	(0.0159)
Labor Union	0.1563***	0.1532***		0.1472***		0.2526***	0.2515***		0.2502***	
	(0.0069)	(0.0069)		(0.0070)		(0.0089)	(0.0089)		(0.0091)	
Mills Ratio				−0.2600***					0.0362	
				(0.0317)					(0.0267)	

Robust standard errors in parentheses; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4: Earnings equations without potential endogenous regressors using OLS, Heckman's procedure and IV-GMM with correction for sample selection bias: Brazilian individuals aged from 18 to 60 years old, by gender (continuation)

Variables	Men					Women				
	OLS	Heckit		IV-GMM		OLS	Heckit		IV-GMM	
		2^{nd} Stage	1^{st} Stage	2^{nd} Stage	1^{st} Stage		2^{nd} Stage	1^{st} Stage	2^{nd} Stage	1^{st} Stage
Married			0.4115***		0.3915***			−0.2434***		−0.2351***
			(0.0208)		(0.0212)			(0.0130)		(0.0127)
Children			0.0763***		0.0754***			−0.1337***		−0.1387***
			(0.0196)		(0.0198)			(0.0123)		(0.0121)
Non Labor Income			−0.4744***		−0.5890***			−0.3009***		−0.3028***
			(0.0314)		(0.0283)			(0.0193)		(0.0192)
Householder			0.3485***		0.3285***			0.2384***		0.2474***
			(0.0146)		(0.0150)			(0.0134)		(0.0129)
Age			0.1106***		0.1103***			0.1398***		0.1419***
			(0.0036)		(0.0036)			(0.0032)		(0.0031)
Age Squared			−0.0016***		−0.0016***			−0.0018***		−0.0018***
			(0.0000)		(0.0000)			(0.0000)		(0.0000)
Number of Smokers					−0.0405***					0.0132
					(0.0119)					(0.0102)
Respiratory Disease					−0.1359***					−0.0632***
					(0.0352)					(0.0234)
Constant	0.1732***	0.2808***	−0.9840***	0.3343***	−0.9160***	0.0210	−0.1026**	−2.7130***	0.0295	−2.7621***
	(0.0211)	(0.0224)	(0.0730)	(0.0243)	(0.0739)	(0.0254)	(0.0464)	(0.0649)	(0.0440)	(0.0610)
Number of Observations	81,974	81,974	95,090	81,974	95,090	54,772	54,772	95,626	54,772	95,626
GMM *C* (Difference-in-Sargan)				22.3074					60.5239	
Hansen's *J* Test				0.0940					0.2609	

Robust standard errors in parentheses; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

5 Concluding Remarks

We do not reject the hypothesis that smoking reduces earnings, i.e., smoking really harms wages. Our results are in line with those recently presented by Almeida & Araújo Júnior (2017), but our estimates are higher, i.e., men (women) who smoke earn 29.7% (24.2%) less than other non-smokers workers. Furthermore, when the wage differential between smokers and non-smokers were decomposed, we saw that a significant part of this difference is due to observable characteristics for both men and women. This final exercise provides evidence that women smokers suffer less wage discrimination than men smokers.

It is well known how harmful smoking is. We provide evidence that the private costs of smoking are not limited to health-related aspects, but that they also affect wages. Productivity on the decline, high absenteeism rates, and the higher costs borne by employers are possible reasons referred to in the literature. This paper contributes to the literature by showing that the decrease in earnings is partly explained by non-observable characteristics in the labor market only for men. We provide more evidence of the hazards of smoking in an unprecedented way by addressing the topic in the context of a developing country with approximately 20 million smokers and 200,000 deaths caused by tobacco use every year.

Acknowledgments

Marcelo Justus thanks the National Council for Technological and Scientific Development (CNPq) for financial support to conduct this research (process number 442483/2014-7), and is also grateful to CNPq for his Productivity in Research Grant. A preliminary version of this paper was presented at the *14th International Conference on Urban Health* (ISUH International Society for Urban Health, 2017).

References

Almeida, A. T. C. & Araújo Júnior, I. T. (2017), 'Tabagismo e penalização salarial no mercado de trabalho brasileiro', *Economia Aplicada* **21**(2), 249. doi: 10.11606/1413-8050/ea146024.

Anger, S. & Kvasnicka, M. (2010), 'Does smoking really harm your earnings so much? Biases in current estimates of the smoking wage penalty', *Applied Economics Letters* **17**(6), 561–564. doi: 10.1080/13504850802260846.

Auld, M. C. et al. (1998), *Wages, alcohol use, and smoking: simultaneous estimates*, Institute of Pharmaco-Economics.

Ault, R. W., Ekelund Jr, R. B., Jackson, J. D., Saba, R. S. & Saurman, D. S. (1991), 'Smoking and absenteeism', *Applied Economics* **23**(4), 743–754. doi: 10.1080/00036849108841031.

Baum, C. F., Schaffer, M. E., Stillman, S. et al. (2003), 'Instrumental variables and GMM: Estimation and testing', *Stata Journal* **3**(1), 1–31. doi: 10.1177/1536867X0300300101.

Becker, G. S. & Murphy, K. M. (1988), 'The theory of rational addiction', *Journal of Political Economy* **96**(4), 675–700. doi: 10.1086/261558.

Blinder, A. S. (1973), 'Wage discrimination: reduced form and structural estimates', *Journal of Human Resources* pp. 436–455. doi: 10.2307/144855.

Braakmann, N. (2008), 'The smoking wage penalty in the United Kingdom: regression and matching evidence from the British Household Panel Survey', *Working Paper Series in Economics* (96).

Cameron, A. C. & Trivedi, P. K. (2009), *Microeconometrics using Stata*, 2 edn, Stata Press, College Station.

Grafova, I. B. & Stafford, F. P. (2009), 'The wage effects of personal smoking history', *ILR Review* **62**(3), 381–393. doi: 10.1177/001979390906200307.

Halpern, M. T., Shikiar, R., Rentz, A. M. & Khan, Z. M. (2001), 'Impact of smoking status on workplace absenteeism and productivity', *Tobacco Control* **10**(3), 233–238. doi: 10.1136/tc.10.3.233.

Heckman, J. J. (1979), 'Sample selection bias as a specification error', *Econometrica* **47**(1), 153–61. doi:10.2307/1912352.

Heineck, G. & Schwarze, J. (2003), 'Substance use and earnings: the case of smokers in Germany', *IZA Discussion Paper* (173).

Hoffmann, R. & Simão, R. C. S. (2009), 'Determinantes do rendimento das pessoas ocupadas em Minas Gerais em 2000: o limiar no efeito da escolaridade e as diferenças entre mesorregiões', *Nova Economia* **15**(2), 35–62.

Hotchkiss, J. L. & Pitts, M. M. (2013), 'Even one is too much: the economic consequences of being a smoker', *FRB Atlanta Working Paper Series 2013-3* . doi: 10.2139/ssrn.2359224.

Jann, B. (2008), 'The Blinder-Oaxaca decomposition for linear regression models', *The Stata Journal* **8**(4), 453–479. doi: 10.1177/1536867X0800800401.

Justus, M., Kawamura, H. & Kassouf, A. L. (2015), 'What is the best age to enter the labor market in Brazil today?', *EconomiA* **16**(2), 235–249. doi: 10.1016/j.econ.2015.03.007.

Lee, A. J., Crombie, I. K., Smith, W. C. & Tunstall-Pedoe, H. D. (1991), 'Cigarette smoking and employment status', *Social Science and Medicine* **33**(11), 1309–1312. doi: 10.1016/0277-9536(91)90080-V.

Lee, Y. L. et al. (1999), 'Wage effects of drinking and smoking: an analysis using Australian twins data', *Working Paper* (22), 1–29.

Leigh, J. P. (1995), 'Smoking, self-selection and absenteeism', *The Quarterly Review of Economics and Finance* **35**(4), 365–386. doi: 10.1016/1062-9769(95)90046-2.

Leistikow, B. N., Martin, D. C. & Milano, C. E. (2000), 'Fire injuries, disasters, and costs from cigarettes and cigarette lights: a global overview', *Preventive Medicine* **31**(2), 91–99. doi: 10.1006/pmed.2000.0680.

Levine, P. B., Gustafson, T. A. & Velenchik, A. D. (1997), 'More bad news for smokers? The effects of cigarette smoking on wages', *Indutrial and Labour Relations Review* **50**(3), 493–509.

Lye, J. N. & Hirschberg, J. (2004), 'Alcohol consumption, smoking and wages', *Applied Economics* **36**(16), 1807–1817. doi: 10.1080/00036840410001710645.

Oaxaca, R. (1973), 'Male-female wage differentials in urban labor markets', *International Economic Review* pp. 693–709. doi: 10.2307/2525981.

Van Ours, J. C. (2004), 'A pint a day raises a man's pay; but smoking blows that gain away', *Journal of Health Economics* **23**(5), 863–886. doi: 10.1016/j.jhealeco.2003.12.005.

Weng, S. F., Ali, S. & Leonardi-Bee, J. (2013), 'Smoking and absence from work: systematic review and meta-analysis of occupational studies', *Addiction* **108**(2), 307–319. doi: 10.1111/add.12015.

Wooldridge, J. M. (2010), *Econometric analysis of cross section and panel data*, MIT Press.