HEDG Health, Econometrics and Data Group

THE UNIVERSITY of York

HEDG Working Paper 06/07

Gender Differences in Smoking Behavior

Thomas Bauer
Silja Göhlmann
Mathias Sinning

August 2006

ISSN 1751-1976

york.ac.uk/res/herc/hedgwp

Gender Differences in Smoking Behavior

Thomas Bauer
RWI Essen, IZA Bonn and CEPR London

Silja Göhlmann
RWI Essen

Mathias Sinning
RWI Essen

August 2006

Abstract. This paper investigates gender differences in smoking behavior using data from the German Socio-economic Panel (SOEP). We develop a Blinder-Oaxaca decomposition method for count data models which allows to isolate the part of the gender differential in the number of cigarettes daily smoked that can be explained by differences in observable characteristics from the part attributable to differences in coefficients. Our results reveal that the major part of the gender smoking differential is attributable to differences in coefficients indicating substantial differences in the smoking behavior between men and women rather than differences in characteristics.

JEL classification: C25, I12.

Keywords: Blinder-Oaxaca decomposition, count data models, SOEP, smoking, gender differences.

All correspondence to Silja Göhlmann, Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI Essen), Hohenzollernstraße 1-3, D-45128 Essen, Germany. Fax: +49-201-8149-200. Email: silja.goehlmann@rwi-essen.de.

1 Introduction

There are numerous studies on the determinants of tobacco consumption focussing on issues such as, for example, the impact of the smoking behavior of parents and peers or education on smoking incidence (see, for example, Gruber and Zinman (2000). However, surprisingly little is known about gender differences in tobacco consumption. Existing empirical studies usually include only a gender dummy variable in their specification or focus on differences in the price and income elasticities of tobacco consumption between males and females (see, among others, Townsend et al. (1994), Chaloupka and Pacula (1999), Hersch (2000) and Yen (2005)). In addition, some studies try to explain gender differences by a different responsiveness to anti-smoking policies, such as clean indoor air restrictions or youth access laws (see, among others, Townsend et al. (1994) and Chaloupka and Pacula (1999)).

It is well known that more males than females smoke. According to Jhaetal (2002) about 47% of all men but only 11% of all woman smoke. In addition, there are remarkable differences in the smoking prevalence of males and females across countries. Whereas there are about 12 times as many men smokers as women smokers in India and four times as many in Japan and Pakistan, almost as many women as men smoke in the European countries and the US (WHO (2005)). In 2005, about 22% of all women and 32% of all men in Germany smoked.[1] In many developed countries the share of smokers among women recently approached the respective share among men, mainly because of a sharply decreasing smoking prevalence among the latter (WHO (2005)).

From a policy perspective it is of interest whether these gender differences in smoking prevalence could mainly be explained by differences in core economic characteristics or whether they are mainly due to behavioral differences. This knowledge would help, for example, to design anti-smoking policies, such as media campaigns, in a more efficient way by addressing specific target groups. The psychological literature concludes that gender differences in tobacco consumption are mainly due to different behavior, having its roots in traditional sex roles. Waldron (1991), for example, identifies three main reasons for gender differences in smoking behavior: (i) general characteristics of traditional sex roles lead to social pressure against female smoking, (ii) traditional sex role norms cause differences in personal characteristics leading to more or less acceptance of smoking (e.g. rebelliousness among males is more accepted than among women and causes higher smoking rates), (iii) sex roles influence the assessment of costs and utility of smoking (e.g. a thin women's beauty ideal makes weight control more important for women and therefore increases the benefits of

[1] Federal Statistical Office, www.destatis.de.

smoking).

This paper provides a detailed descriptive picture of gender differences in the number of cigarettes smoked in Germany and decomposes this difference into a part that is due to differences in socioeconomic characteristics between males and females and a part that is due to differences in coefficients. The latter will be interpreted as gender differences in smoking behavior. For this purpose, we develop a decomposition method similar to the method proposed by Blinder (1973) and Oaxaca (1973) for count data models. The Blinder-Oaxaca-decomposition and its various generalizations have almost exclusively been used in linear regression models. A decomposition method for models with binary dependent variables has been developed by Fairlie (1999, 2003). Bauer and Sinning (2005) have derived a decomposition method for Tobit-models, which allows the decomposition of differences in corner solution outcome variables between two groups.

Our empirical results indicate that gender differences in cigarette consumption are mainly due to a different smoking behavior rather than differences in observable characteristics. Almost 86% of the gender difference in the number of cigarettes smoked per day is due to differences in the estimated coefficients and only 14% due to different characteristics. This result is very robust across different regression models and different data sets.

The remainder of the paper is structured as follows. In the next section we present the data used for the empirical analysis and in Section 3 we develop a decomposition method for count data models. Section 4 presents our estimation results. Section 5 concludes.

2 Data

Our empirical analysis employs data from the German *Socio-economic Panel* (SOEP).[2] The SOEP contains smoking related questions in the years 1998, 1999, 2001, 2002 and 2004. Unfortunately, the question regarding our dependent variable, i.e. the average number of cigarettes smoked per day in the week before the interview, was not included in the questionnaire in 1999. We are further not able to differentiate between the consumption of cigarettes, pipes and cigars for 2001. Therefore we utilize only the years 1998, 2002 and 2004.[3]

[2] For more information on the SOEP see $http://www.diw.de/english/sop/uebersicht/index.html$.

[3] The data used in this paper was extracted from the SOEP Database provided by the DIW Berlin (http://www.diw.de/soep) using the Add-On package SOEP Menu v2.0 (Jul 2005) for

For our analysis we use the following set of explanatory variables: age and age squared; years of education and years of education squared; two dummy variables for the marital status (i.e. a dummy variable for being married, and a dummy variable for being separated or widowed with singles acting as reference group); two dummy variables for income[4], one taking the value one for an income between 1,000 and 2,000 Euros and one dummy variable that takes the value one for an income above 2,000 Euros with those having an income below 1,000 Euros acting as reference group; dummy variables for individuals having children younger than 2 and between 2 and 18 years old; a dummy variable for foreigners; a dummy variable for persons living in East-Germany, a dummy variable for persons living in a city with more than 99,999 residents; two dummy variables for the education of the parents (i.e. whether they have a high or medium schooling degree with those having parents with a low or no schooling degree acting as a reference group); four dummy variables indicating the labor market status (unemployed, in training, full-time job and part-time job), and interaction terms between the dummy variables for having a full-time and a part-time job, respectively, with dummy variables indicating whether the person has a white-collar and a dummy variable indicating another type of job, with blue-collar workers acting as reference group. Eliminating all observations with missing values for at least one of the used variables results in a sample of 47,066 person-year-observations of 22,748 individuals available for the empirical analysis.

To test the robustness of our results, we also utilize data from the *Population Survey on the Consumption of Psychoactive Substances* in Germany (PSCPS) collected by the Institute for Therapy Research (Institut für Therapieforschung), IFT Munich (see Kraus and Augustin (2001) for a detailed description). The surveys were collected in the years 1980, 1986, 1990, 1992, 1995, 1997, and 2000. Because we also want to analyze whether our results change when including parental smoking behavior as explaining variables when estimating smoking participation, only the first four waves might be considered, when questions about parental smoking were asked in the survey. However, the 1992 wave of the PSCPS lacks information on the number of inhabitants of the city of a respondent. In our empirical analysis, we therefore focus only on the first three waves. Differences to the SOEP data exist in particular with respect to the sampling frame of the PSCPS, which aims especially

Stata(R). SOEP Menu was written by Dr. John P. Haisken-DeNew (john@soepmenu.de). John P. Haisken-DeNew and Markus Hahn supplied the SOEP Menu Plugins used to ensure longitudinal consistency. The SOEP Menu generated DO file to retrieve the SOEP data used here and any SOEP Menu Plugins are available upon request. Any data or computational errors in this paper are our own. Haisken-DeNew (2005) describes SOEP Menu in detail.

[4] Defined as household net income/$\sqrt{\text{household size}}$. We also experimented with other definitions of equivalence income (i.e. the new and old definition of OECD and the definition of the "Bundessozialhilfegesetz". The results, however, do not vary with the chosen definition of income.

at younger respondents (aged 12 to 24 in 1980, aged 12 to 29 in 1986 and aged 12 to 39 in 1990). Moreover, in contrast to the SOEP, the PSCPS does not include foreign citizens. Furthermore, no information is available on the vocational degree of a person with the consequence that different to the SOEP years of education is measured only based on schooling degrees. The estimates based on the PSCPS include a variable indicating the number of children. It further lacks information about the type of job (white / blue collar worker).

Descriptive statistics on all variables used for both samples are shown in Table 1. The differences between the SOEP and the PSCPS can mainly be attributed to the different sampling frame of the two data sets. In both samples men smoke more cigarettes per day than females; about 1.7 times more in the SOEP and almost 1.5 times more in the PSCPS. Furthermore, in the PSCPS both men and women smoke on average more cigarettes per day than in the SOEP, indicating that younger persons smoke relatively more than the average person.

3 Empirical Strategy

To investigate gender differences in smoking behavior, we apply a Blinder-Oaxaca-type decomposition, which permits the decomposition of gender differences in the number of cigarettes smoked per day into a part that is caused by differences in observable characteristics and part that is explained by differences in estimated coefficients. In the following, we will interpret the latter part as the component that reflects gender differences in smoking behavior. Since our outcome measure is given by a count data variable, the application of the conventional Blinder-Oaxaca-decomposition for linear models is not appropriate. We therefore derive a Blinder-Oaxaca-type decomposition method for count data models.

Consider the following linear regression model, which is estimated separately for the groups $g = m, f$

$$C_{ig} = \mathbf{X}_{ig}\beta_g + \varepsilon_{ig}, \tag{1}$$

where C_{ig} represents the number of daily smoked cigarettes of individual i ($i = 1, ..., N_g$) in group g, $\mathbf{X}_{ig}$ is a vector of observable characteristics (as described in section 2), β_g denotes a vector of parameters to be estimated, and ε_{ig} is a standard error term. For these models, Blinder (1973) and Oaxaca (1973) propose the decomposition

$$\begin{aligned}\overline{C}_m - \overline{C}_f = & [E_{\beta_m}(C_{im}|\mathbf{X}_{im}) - E_{\beta_m}(C_{if}|\mathbf{X}_{if})] \\ & +[E_{\beta_m}(C_{if}|\mathbf{X}_{if}) - E_{\beta_f}(C_{if}|\mathbf{X}_{if})],\end{aligned} \tag{2}$$

where $\overline{C}_g = N_g^{-1} \sum_{i=1}^{N_g} C_{ig}$ and $\overline{\mathbf{X}}_g = N_g^{-1} \sum_{i=1}^{N_g} \mathbf{X}_{ig}$. $E_{\beta_g}(C_{ig}|\mathbf{X}_{ig})$ refers to the conditional expectation of C_{ig} evaluated at the parameter vector β_g. The first term on the right hand side of equation (2) displays the difference in the outcome variable between the two groups that is due to differences in observable characteristics, whereas the second term shows the differential that is due to differences in coefficient estimates. In a linear regression model, equation (2) reduces to the well-known formula for the Blinder-Oaxaca decomposition: $\overline{C}_m - \overline{C}_f = \Delta^{OLS} = (\overline{X}_m - \overline{X}_f)\widehat{\beta}_m + \overline{X}_f(\widehat{\beta}_m - \widehat{\beta}_f)$.

The linear regression model, however, may lead to biased estimates of the parameter vector and hence misleading results of the decomposition, if the outcome variable C_{ig} is given by a count data variable. In this case, regression models for count data are required to obtain consistent parameter estimates. Using the Poisson regression model as a benchmark for the analysis of count data (Winkelmann (2000)), we derive a general decomposition method for count data regression models. The Poisson regression model (P) assumes that the dependent variable C_{ig} conditional on the covariates $\mathbf{X}_{ig}$ is Poisson distributed with density

$$f(C_{ig}|\mathbf{X}_{ig}) = \frac{\exp(-\mu_{ig})\mu_{ig}^{C_{ig}}}{C_{ig}!}, \quad C_{ig} = 0, 1, 2, ... \tag{3}$$

and conditional expectation

$$E(C_{ig}|\mathbf{X}_{ig}) = \mu_{ig} = \exp(\mathbf{X}_{ig}\beta_g^P). \tag{4}$$

Equation (4) reveals that the conventional Blinder-Oaxaca decomposition of the outcome variable is not appropriate for count data variables. However, one can use equation (2) to derive a Blinder-Oaxaca-type decomposition for count data models.

Given a *sample counterpart* of the conditional expectation of C_{ig} evaluated at β_g,

$$E_{\beta_g}(C_{ig}|\mathbf{X}_{ig}) = S(\hat{\beta}_g, \mathbf{X}_{ig}), \tag{5}$$

the components of equation (2) can be estimated by

$$\begin{aligned}\hat{\Delta} &= \left[S(\hat{\beta}_m, \mathbf{X}_{im}) - S(\hat{\beta}_m, \mathbf{X}_{if})\right] \\ &\quad + \left[S(\hat{\beta}_m, \mathbf{X}_{if}) - S(\hat{\beta}_f, \mathbf{X}_{if})\right].\end{aligned} \tag{6}$$

For the Poisson-model, the sample counterpart of $E_{\beta_g}(C_{ig}|\mathbf{X}_{ig})$ is given by

$$S(\hat{\beta}_g^P, \mathbf{X}_{ig}) = \overline{C}_{g,\hat{\beta}_g^P} = \frac{1}{N_g} \sum_{i=1}^{N_g} \exp(\mathbf{X}_{ig}\hat{\beta}_g^P). \tag{7}$$

The Poisson-model is based on the assumption that the dependent variable has the same mean and variance $\mu_{ig} = \exp(\mathbf{X}_{ig}\beta_g^P)$. If this assumption is violated, an alternative conditional distribution of the dependent variable may be specified that permits a more flexible modeling of the variance of the dependent variable. An alternative to the Poisson regression model is given by the negative binomial (Negbin) regression model (NB), which relaxes the assumption of equality of the conditional mean and the variance of the dependent variable, while it assumes the same form of the conditional mean as the Poisson-model. Consequently, the sample counterpart of the conditional mean of the Negbin regression model is

$$S(\hat{\beta}_g^{NB}, \mathbf{X}_{ig}) = \overline{C}_{g,\hat{\beta}_g^{NB}} = \frac{1}{N_g}\sum_{i=1}^{N_g} \exp(\mathbf{X}_{ig}\hat{\beta}_g^{NB}). \tag{8}$$

However, in the Negbin model a quadratic relationship between the variance and the mean is assumed:

$$V(C_{ig}|\mathbf{X}_{ig}) = \mu_{ig} + \alpha\mu_{ig}^2.$$

where α is a scalar parameter.

In addition to the Poisson and Negbin regression models, zero-inflated models are frequently used when analyzing count data. These models take into account that real-life data may contain excess zeros, causing a higher probability of zero values than is consistent with the Poisson and negative binomial distribution. In this case it could be assumed that zeros and positive values do not come from the same data generating process. Winkelmann (2000) provides an overview of zero-inflated Poisson and Negbin models.

In order to investigate the probability of excess zeros, Lambert (1992) proposed a zero-inflated Poisson model, that allows for two different data generating regimes: the outcome of regime 1, R1, is always zero, whereas the outcome of regime 2, R2, is generated by a poisson process. In this model, the (so-called) ”unconditional” expectation of the dependent variable consists of the conditional probability of observing regime 2 and the conditional expectation of the zero-truncated density:

$$E(C_{ig}|\mathbf{X}_{ig}) = (1 - Pr(R1|\mathbf{X}_{ig}))E(C_{ig}|R2, \mathbf{X}_{ig}). \tag{9}$$

Lambert (1992) specifies the conditional probability of regime 1, that always leads to a zero outcome, as a Logit model:

$$Pr(R1|\mathbf{X}_{ig}) = \frac{\exp(\gamma_g \mathbf{Z}_{ig})}{1 + \exp(\gamma_g \mathbf{Z}_{ig})},$$

where $\mathbf{Z}_{ig}$ contains the covariates of the conditional probability of excess zeros and γ_g is the parameter vector to be estimated. Consequently, the unconditional mean of the dependent variable specified by equation (9) can be estimated by

$$S(\hat{\beta}_g^{ZIP}, \mathbf{X}_{ig}) = \frac{1}{N_g}\sum_{i=1}^{N_g}(1 - (\widehat{Pr}(R1)|\mathbf{X}_{ig}))\hat{\mu}_{ig} = \frac{1}{N_g}\sum_{i=1}^{N_g}\frac{\exp(\hat{\beta}_g^{ZIP}\mathbf{X}_{ig})}{1+\exp(\hat{\gamma}_g\mathbf{Z}_{ig})}. \tag{10}$$

Given that the zero generating process is based on a logistic distribution, a similar sample counterpart can be derived for the unconditional mean of C_{ig} in the zero-inflated Negbin model:

$$S(\hat{\beta}_g^{ZINB}, \mathbf{X}_{ig}) = \frac{1}{N_g}\sum_{i=1}^{N_g}\frac{\exp(\hat{\beta}_g^{ZINB}\mathbf{X}_{ig})}{1+\exp(\hat{\gamma}_g\mathbf{Z}_{ig})}. \tag{11}$$

Hurdle models represent another modification of count data models. The hurdle model may be interpreted as a two-part model, where the first part is a binary outcome model, and the second part a truncated count data model. The unconditional mean of the dependent variable is given by:

$$E(C_{ig}|\mathbf{X}_{ig}) = Pr(C_{ig} > 0|\mathbf{X}_{ig})E(C_{ig}|C_{ig} > 0, \mathbf{X}_{ig}). \tag{12}$$

According to Cameron and Trivedi (1998) the conditional expected values of C_{ig} of the hurdle Poisson (HP) and the hurdle Negbin (HNB) model are given by

$$E(C_{ig}|C_{ig} > 0, \mathbf{X}_{ig}) = \frac{\exp(\beta_g^{HP}\mathbf{X}_{ig})}{1-\exp(-\exp(\beta_g^{HP}\mathbf{X}_{ig}))} \tag{13}$$

and

$$E(C_{ig}|C_{ig} > 0, \mathbf{X}_{ig}) = \frac{\exp(\beta_g^{HNB}\mathbf{X}_{ig})}{1-(1+\alpha\exp(\beta_g^{HNB}\mathbf{X}_{ig}))^{-\frac{1}{\alpha}}}. \tag{14}$$

Consequently, assuming a logistic distribution for the underlying zero generating process, the unconditional expected values can be estimated by the following expressions:

$$S(\hat{\beta}_g^{HP}, \mathbf{X}_{ig}) = \frac{1}{N_g}\sum_{i=1}^{N_g}\frac{\exp(\hat{\beta}_g^{HP}\mathbf{X}_{ig})}{(1-\exp(-\exp(\hat{\beta}_g^{HP}\mathbf{X}_{ig})))(1+\exp(\hat{\gamma}_g\mathbf{Z}_{ig}))} \tag{15}$$

and

$$S(\hat{\beta}_g^{HNB}, \mathbf{X}_{ig}) = \frac{1}{N_g}\sum_{i=1}^{N_g}\frac{\exp(\hat{\beta}_g^{HNB}\mathbf{X}_{ig})}{(1-(1+\alpha\exp(\hat{\beta}_g^{HNB}\mathbf{X}_{ig}))^{-\frac{1}{\alpha}})(1+\exp(\hat{\gamma}_g\mathbf{Z}_{ig}))}. \tag{16}$$

In the following, we present the estimates and decomposition results of the different count data models described in this section.

4 Results

To investigate differences in the smoking prevalence between males and females, we estimate the count data models described in the last section separately for men and women, i.e. we estimate Poisson and Negbin models as well as Hurdle and Zero-inflated Poisson and Negbin models. Using likelihood-ratio tests and Voung tests (Vuong (1989)) for non-nested models, we test the different models against each other. The descriptive statistics reported in Table 1 already suggest that our dependent variable suffers from over-dispersion. Hence it is not surprising that all our tests reject the different Poisson models in favor of the Negbin-models for both males and females. The likelihood ratio test further rejects the Negbin-model in favor of the Hurdle Negbin model. Testing the hurdle models against the Zero-inflated model using the Voung test finally shows that the zero-inflated Negbin model describes the data best for both gender groups. This result indicates that there are two types of individuals reporting zero consumption of cigarettes: (i) non-smokers, who will never smoke; and (ii) potential smokers, for some of which zero consumption is, for example, a strictly economic decision.

Table 2 presents the estimation results of the zero-inflated Negbin model.[5] For both males and females the potential of being a non-smoker follows an U-shaped pattern with age. Compared to those not participating in the labor market, unemployed and blue-collar workers have a lower probability of being a non-smoker. Generally, white collar workers have a significantly higher probability of being non-smoker than blue collar workers. Separated, divorced or widowed individuals are significantly less likely non-smokers, as are individuals living in urban compared to those living in rural areas. There are some remarkable differences between males and females. Whereas females in East-Germany are more likely non-smokers than those in West-Germany, this difference appears not being significant for males. A similar result appears for females with a foreign citizenship. In turn, males in educational training have a significantly lower and males with a monthly income above 2000 Euros a significantly higher probability of being a non-smoker than those not participating in the labor market and those with a monthly income below 1000 Euros, respectively. For the female sample the respective coefficients are statistically insignificant.

Conditional on being a potential smoker, the number of cigarettes smoked per day follows an inverted U-shaped age profile for both males and females. East Germans smoke significantly less cigarettes than individuals living in the West, and those living in urban areas smoke significantly more than persons living in rural

[5] The results of the other models as well of the Likelihood Ratio- and Vuong-test are available from the authors upon request.

areas. With respect to the rest of the coefficients there appears a different pattern by gender. Potential female smokers in educational training smoke significantly less than those not participating in the labor force as are females with a foreign nationality if compared to German females. Potential male smokers with a blue-collar full-time or part-time job as well as unemployed males consume significantly more cigarettes than those not participating in the labor market and male white-collar workers smoke less cigarettes per day than male blue-collar workers. The results, however, do indicate a significant difference between male white-collar workers and men not-participating in the labor market. Being separated, divorced or widowed increases cigarette consumption of potential male smokers if compared to single males. Males with a monthly income above 2000 Euros have a lower probability of being a potential smoker, but conditional on being a potential smoker they smoke significantly more cigarettes than those with lower income.

The estimation results for the PSCPS data are presented in Table 3. Although the estimated coefficients are insignificant in many cases, the findings suggest that the estimates in Tables 2 and 3 do not differ substantially from each other. Again, the number of cigarettes smoked per day follows an inverted U-shaped pattern with increasing age. Persons in urban areas smoke significantly more than comparable persons residing in rural areas. Moreover, separated, divorced or widowed men smoke significantly more than single men while females smoke significantly less if they are married.

The results of the decomposition analysis for count data models described in the last section are reported in Table 4 for the SOEP, and in Table 5 for the PSCPS. For all but the hurdle Negbin model the results of the decomposition analysis are rather stable across the different models and across the two data sets. Note that only the hurdle Negbin model does a poor job in predicting the gender difference in cigarette consumption.

Overall it appears that most of the differences in the daily consumption of cigarettes between males and females is due to differences in the estimated coefficients and hence differences in smoking behavior rather than differences in observable characteristics. Referring to the Zero-inflated Negbin model - the model which appears to describe the underlying data generating process best - 86% of the difference could be explained by differences in coefficients and only 14% by different observable characteristics. A similar picture emerges when using the PSCPS. Here more than 96% of the gender difference in cigarette consumption is due to differences in coefficients and only 4% due to differences in observable characteristics.[6] The differences in

[6] The results do not change when controlling for parental smoking behavior in the smoking

the results of the PSCPS if compared to the respective results based on the SOEP may be explained by the fact that we can not account for the type of job (white vs. blue-collar worker) in the PSCPS.

5 Conclusion

In almost all countries less females smoke than males. From a policy perspective it is of great interest whether the gender differences in smoking prevalence could mainly be explained by differences in core characteristics or by a different smoking behavior. Having evidence on the sources of the differences in tobacco consumption between males and females may help, for example, to make anti-smoking policies more effective by enabling the policy makers to address specific target groups. The results of Chaloupka and Pacula (1999) indicate for example that clean indoor air laws were correlated with a decreased smoking participation only for (white) males.

In this paper we provide a detailed analysis of the determinants of cigarette consumption of males and females in Germany. In order to decompose the gender difference in cigarette consumption into a part that can be explained by different characteristics and a part that can be explained by a different smoking behavior, we develop a decomposition method for count data models that follows the well-known Blinder-Oaxaca decomposition for linear regression models. The results from our empirical analysis show that more than 86% of the gender difference in the number of cigarettes smoked per day can be explained by a different smoking behavior, indicating that anti-smoking policies can be more effective if they take these behavioral differences into account.

participation equation.

References

Bauer T K, Sinning M. Blinder-Oaxaca Decomposition for Tobit Models, *RWI Discussion Paper* 2005; **32**: 1-10.

Blinder A S. Wage Discrimination: Reduced Form and Structural Estimates, *J Hum Resour* 1973; **8**: 436-455.

Cameron A C, Trivedi P K. *Regression Analysis of Count Data.* Cambridge University Press: Cambridge, 1998.

Chaloupka F J, Pacula R L. Sex and Race Differences in Young People's Responsiveness to Price and Tobacco Control Policies, *Tob Control* 1999; **8**: 373-377.

Fairlie R W. The Absence of the African-American Owned Business: An Analysis of the Dynamics of Self-Employment, *J Labor Econ* 1999; **17**: 80-108.

Gruber J, Zinman J. Youth Smoking in the U.S.: Evidence and Implications, *NBER Work Pap Ser* 2000; **7780**: 1-50.

Haisken-DeNew J P. SOEP Menu: A Menu-Driven Stata/SE Interface for Accessing the German Socio-Economic Panel, *mimeo, http://www.soepmenu.de* 2005.

Hersch J. Gender, Income Levels, and the Demand for Cigarettes, *J Risk Uncertain* 2000; **21**: 263-282.

Jha P, Ranson M K, Nguyen S N, Yach D. Estimates of Global and Regional Smoking Prevalence in 1995, by Age and Sex, *American Public Health Association* 2002; **92**(6).

Lambert D. Zero-inflated Poisson Regression with an Application to Defects in Manufacturing, *Technometrics* 1992; **34**: 1-14.

Oaxaca R L. Male-Female Wage Differentials in Urban Labor Markets, *International Economic Review* 1973; **14**: 693-709.

Townsend J, Roderick P, Cooper J. Cigarette smoking by socioeconomic group, sex, and age: effects of price, income, and health publicity, *Br Med J* 1994; **309**: 923-927.

Voung Q. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses, *Econometrica* 1989; **57**: 307-334.

Waldron I. Patterns and Causes of Gender Differences in Smoking, *Social Science and Medicine* 1991; **32**(9): 989-1005.

WHO. Gender in Lung Cancer and Smoking Research, 2005; 1-43.

Winkelmann R. *Econometric Analysis of Count Data*. Springer Verlag: Berlin, Heidelberg, 2000.

Yen, S T. Zero observations and gender differences in cigarette consumption, *Appl Econ* 2005; **37**: 1839-1849.

Table 1: **Descriptive Statistics**

	SOEP		**PSCPS**	
	Women	Men	Women	Men
Number of cigarettes	3.626	6.186	5.483	8.095
	(7.458)	(10.193)	(8.518)	(10.191)
Age	47.788	46.547	22.483	22.160
	(17.517)	(16.640)	(4.612)	(4.434)
$\text{Age}^2/100$	25.906	24.435	5.268	5.107
	(17.886)	(16.466)	(2.342)	(2.225)
East-Germany	0.258	0.258	0.145	0.117
	(0.438)	(0.438)	(0.352)	(0.322)
Years of education	11.579	12.000	10.244	10.071
	(2.440)	(2.606)	(1.725)	(1.790)
Years of $\text{education}^2/100$	1.400	1.508	1.079	1.046
	(0.633)	(0.699)	(0.368)	(0.377)
Father high school degree	0.096	0.092	0.120	0.107
	(0.294)	(0.289)	(0.325)	(0.309)
Father medium school degree	0.124	0.133	0.130	0.123
	(0.330)	(0.339)	(0.337)	(0.328)
Father low school degree	0.780	0.775	0.666	0.689
	(0.414)	(0.418)	(0.472)	(0.463)
Mother high school degree	0.043	0.045	0.052	0.055
	(0.203)	(0.208)	(0.223)	(0.227)
Mother medium school degree	0.148	0.151	0.160	0.148
	(0.355)	(0.358)	(0.366)	(0.355)
Mother low school degree	0.809	0.804	0.762	0.773
	(0.393)	(0.397)	(0.426)	(0.419)
Full time job	0.266	0.591	0.441	0.545
	(0.442)	(0.492)	(0.497)	(0.498)
Full time job - white collar	0.197	0.275	–	–
	(0.398)	(0.446)	–	–
Full time job - not white / blue collar	0.021	0.074	–	–
	(0.144)	(0.262)	–	–
Part time job	0.213	0.036	0.069	0.020
	(0.410)	(0.185)	(0.253)	(0.138)
Part time job - white collar	0.144	0.015	–	–
	(0.351)	(0.122)	–	–
Part time job - not white / blue collar	0.012	0.007	–	–
	(0.110)	(0.082)	–	–
In educational training	0.028	0.035	0.348	0.385
	(0.164)	(0.184)	(0.477)	(0.487)
Unemployed	0.062	0.072	0.043	0.037
	(0.240)	(0.259)	(0.203)	(0.189)
Not participating	0.432	0.266	0.098	0.014
	(0.495)	(0.442)	(0.298)	(0.117)
Married	0.601	0.644	0.324	0.176
	(0.490)	(0.479)	(0.468)	(0.381)
Separated, divorced or widowed	0.203	0.104	0.031	0.014
	(0.402)	(0.306)	(0.172)	(0.117)
Single	0.197	0.251	0.646	0.810
	(0.398)	(0.434)	(0.478)	(0.392)
Children younger 2	0.025	0.030	–	–
	(0.155)	(0.171)	–	–
Children aged 2 - 18	0.325	0.321	–	–
	(0.468)	(0.467)	–	–
Number of children	–	–	0.391	0.229
	–	–	(0.747)	(0.599)
Monthly income more than 2000 Euro	0.170	0.195	0.340	0.424
	(0.376)	(0.396)	(0.474)	(0.494)
Monthly income 1000 - 1999 Euro	0.591	0.603	0.568	0.512
	(0.492)	(0.489)	(0.495)	(0.500)
Monthly income less than 1000 Euro	0.239	0.202	0.092	0.064
	(0.426)	(0.401)	(0.289)	(0.245)
Non-German nationality	0.084	0.098	–	–
	(0.278)	(0.298)	–	–
Urban	0.303	0.291	0.267	0.250
	(0.459)	(0.454)	(0.443)	(0.433)
Number of observations	22264	20761	3863	3905

Note: Standard deviations in parentheses.

Table 2: **Determinants of the Number of Cigarettes smoked per Day, Zero inflated NB Estimates (SOEP)**

	Women		**Men**	
	Logit Model	Truncated Negbin	Logit Model	Truncated Negbin
Age	-0.089***	0.031***	-0.096***	0.049***
	(0.017)	(0.009)	(0.015)	(0.005)
Age^2/100	0.132***	-0.034***	0.136***	-0.051***
	(0.018)	(0.010)	(0.016)	(0.006)
East-Germany	0.278***	-0.234***	0.024	-0.146***
	(0.081)	(0.039)	(0.076)	(0.022)
Years of education	0.055	-0.083	0.211*	-0.026
	(0.107)	(0.054)	(0.111)	(0.038)
Years of education^2/100	0.244	0.266	-0.349	-0.026
	(0.412)	(0.230)	(0.418)	(0.152)
Father high school degree	-0.022	-0.133**	-0.222*	-0.030
	(0.123)	(0.063)	(0.134)	(0.052)
Father medium school degree	-0.037	0.047	-0.243**	-0.032
	(0.112)	(0.048)	(0.108)	(0.040)
Mother high school degree	-0.000	0.119*	0.398**	-0.073
	(0.161)	(0.067)	(0.164)	(0.060)
Mother medium school degree	-0.230**	-0.059	0.031	-0.035
	(0.107)	(0.045)	(0.114)	(0.039)
Full time job	-0.665***	0.029	-0.430***	0.077**
	(0.150)	(0.052)	(0.105)	(0.038)
Full time job - white collar	0.432***	-0.056	0.448***	-0.057**
	(0.151)	(0.052)	(0.087)	(0.029)
Full time job - other job	0.049	-0.112	0.197*	0.097**
	(0.238)	(0.094)	(0.117)	(0.038)
Part time job	-0.564***	0.080	-0.576**	0.184***
	(0.118)	(0.080)	(0.286)	(0.063)
Part time job - white collar	0.266**	-0.134	0.701**	-0.259**
	(0.128)	(0.086)	(0.344)	(0.116)
Part time job - other job	0.483*	0.090	0.356	-0.426***
	(0.254)	(0.114)	(0.425)	(0.155)
In educational training	-0.218	-0.169**	-0.352**	0.027
	(0.164)	(0.067)	(0.138)	(0.051)
Unemployed	-0.677***	0.047	-0.838***	0.096**
	(0.126)	(0.054)	(0.127)	(0.040)
Married	0.394***	-0.057	0.156	-0.022
	(0.109)	(0.046)	(0.105)	(0.031)
Separated, divorced or widowed	-0.349***	-0.016	-0.521***	0.121***
	(0.126)	(0.057)	(0.137)	(0.037)
Children younger 2	0.315	-0.002	0.182	0.039
	(0.192)	(0.075)	(0.143)	(0.043)
Children aged 2 - 18	-0.027	-0.031	0.123*	-0.012
	(0.073)	(0.038)	(0.068)	(0.023)
Monthly income more than 2000 Euro	-0.021	-0.071	0.257**	0.097***
	(0.104)	(0.058)	(0.104)	(0.035)
Monthly income 1000 - 1999 Euro	0.126*	-0.043	0.110	-0.004
	(0.070)	(0.040)	(0.067)	(0.023)
Non-German nationality	0.302**	-0.111**	-0.111	-0.004
	(0.128)	(0.053)	(0.117)	(0.032)
Urban	-0.404***	0.088***	-0.289***	0.040*
	(0.068)	(0.029)	(0.067)	(0.022)
Constant	1.226*	2.742***	-0.017	2.156***
	(0.731)	(0.347)	(0.730)	(0.253)
α	0.2621***		0.2031***	
	(0.0661)		(0.0471)	
Vuong: ZINB vs. Standard NEGBIN	51.01		61.74	
Wald-Statistic (χ^2)	175.838		445.3644	
Log Pseudolikelihood	-30560.78		-37153.02	

Notes: *** significant at 1%; ** significant at 5%; * significant at 10%. Number of observations: 22,264 women, 20,761 men. Weighted estimation using weights provided by the SOEP. Standard errors, which are reported in parentheses, are adjusted to take repeated observations into account. Reference group is a single individual, not participating at labor market with a monthly income less than 1000 Euro and with parents both having a low school degree. The regression further includes year dummies.

Table 3: **Determinants of the Number of Cigarettes smoked per Day, Zero inflated NB Estimates (PSCPS)**

	Women		**Men**	
	Logit Model	Truncated Negbin	Logit Model	Truncated Negbin
Age	-0.316***	0.163***	-0.283***	0.128***
	(0.071)	(0.029)	(0.066)	(0.020)
Age^2/100	0.586***	-0.259***	0.498***	-0.193***
	(0.141)	(0.057)	(0.132)	(0.040)
Years of education	0.183	0.071	0.438**	0.071
	(0.235)	(0.104)	(0.214)	(0.063)
Years of education^2/100	0.001	-0.004	-0.010	-0.005*
	(0.011)	(0.004)	(0.010)	(0.003)
Father high school degree	0.174	-0.161**	-0.106	-0.007
	(0.140)	(0.064)	(0.139)	(0.043)
Father medium school degree	0.129	-0.053	0.040	-0.031
	(0.120)	(0.047)	(0.115)	(0.039)
Mother high school degree	-0.260	0.113	0.533***	-0.038
	(0.189)	(0.079)	(0.192)	(0.073)
Mother medium school degree	-0.124	-0.018	-0.062	0.016
	(0.112)	(0.046)	(0.110)	(0.035)
Full time job	-0.156	-0.045	-0.115	-0.003
	(0.147)	(0.056)	(0.295)	(0.102)
Part time job	-0.215	-0.042	-0.288	-0.144
	(0.181)	(0.068)	(0.388)	(0.119)
In educational training	-0.153	-0.136**	-0.061	-0.052
	(0.159)	(0.059)	(0.297)	(0.104)
Unemployed	-0.408*	0.038	-0.495	0.013
	(0.215)	(0.075)	(0.347)	(0.107)
Married	0.302***	-0.170***	0.053	-0.013
	(0.111)	(0.042)	(0.121)	(0.032)
Separated, divorced or widowed	-0.459**	0.076	-0.380	0.195**
	(0.222)	(0.060)	(0.305)	(0.080)
Number of children	-0.125*	0.005	0.003	0.006
	(0.065)	(0.024)	(0.076)	(0.021)
Monthly income more than 2000 Euro	-0.007	-0.054	-0.066	-0.045
	(0.129)	(0.056)	(0.150)	(0.043)
Monthly income 1000 - 1999 Euro	-0.021	-0.097	-0.159	-0.009
	(0.146)	(0.062)	(0.159)	(0.046)
Urban	-0.245***	0.092***	-0.343***	0.079***
	(0.085)	(0.031)	(0.086)	(0.024)
Constant	2.488*	0.272	0.676	0.834**
	(1.514)	(0.618)	(1.398)	(0.417)
α	0.1979***		0.1376**	
	(0.0625)		(0.0585)	
Wald-Statistic (χ^2)	215.82		205.58	
Log Pseudolikelihood	-7754.044		-9299.994	

Notes: See Notes to Table 2. Number of observations: 3,905 women, 3,863 men. Weighted estimation using weights provided by the PSCPS.

Table 4: **Decomposition Results (SOEP)**

	Poisson		**NB**	
	Observed Coef.	s.e.	Observed Coef.	s.e.
$\hat{\Delta}$	2.618***	0.134	2.619***	0.161
Explained Part	0.356**	0.177	0.383**	0.177
in % of $\hat{\Delta}$	13.595**	6.657	14.622**	6.709
Unexplained Part	2.262***	0.196	2.236***	0.220
in % of $\hat{\Delta}$	86.405***	6.657	85.378***	6.709

	Hurdle Poisson		**Hurdle NB**	
	Observed Coef.	s.e.	Observed Coef.	s.e.
$\hat{\Delta}$	0.474***	0.149	-0.192*	0.111
Explained Part	0.035**	0.016	0.003	0.006
in % of $\hat{\Delta}$	7.33	7.328	-1.333	211.240
Unexplained Part	0.439***	0.150	-0.195*	0.111
in % of $\hat{\Delta}$	92.67***	7.328	101.333	211.240

	Zero-inflated Poisson		**Zero-inflated NB**	
	Observed Coef.	s.e.	Observed Coef.	s.e.
$\hat{\Delta}$	2.630***	0.135	2.637***	0.135
Explained Part	0.351**	0.174	0.371**	0.172
in % of $\hat{\Delta}$	13.364**	6.502	14.070**	6.372
Unexplained Part	2.278***	0.192	2.266***	0.189
in % of $\hat{\Delta}$	86.637***	6.502	85.930***	6.372

Notes: Bootstrapped (100 replications) standard errors. *** significant at 1%; ** significant at 5%; * significant at 10%.

Table 5: **Decomposition Results (PSCPS)**

	Poisson		**NB**	
	Observed Coef.	s.e.	Observed Coef.	s.e.
$\hat{\Delta}$	2.673***	0.214	2.663***	0.219
Explained Part	0.065	0.182	0.064	0.213
in % of $\hat{\Delta}$	2.411	6.919	2.543	8.078
Unexplained Part	2.608***	0.256	2.595***	0.280
in % of $\hat{\Delta}$	97.588***	6.919	97.457***	8.078
	Hurdle Poisson		**Hurdle NB**	
	Observed Coef.	s.e.	Observed Coef.	s.e.
$\hat{\Delta}$	0.101**	0.040	0.026	0.093
Explained Part	0.007	0.011	-0.000	0.003
in % of $\hat{\Delta}$	7.272	15.555	-0.791	14.459
Unexplained Part	0.093**	0.041	0.026	0.094
in % of $\hat{\Delta}$	92.728***	15.555	100.791***	14.459
	Zero-inflated Poisson		**Zero-inflated NB**	
	Observed Coef.	s.e.	Observed Coef.	s.e.
$\hat{\Delta}$	2.671***	0.214	2.670***	0.214
Explained Part	0.114	0.178	0.098	0.179
in % of $\hat{\Delta}$	4.266	6.732	3.665	6.785
Unexplained Part	2.557***	0.254	2.572***	0.255
in % of $\hat{\Delta}$	95.735***	6.732	96.335***	6.785

Notes: See Notes to Table 4.

Appendix

Table A.1: **Description of Variables**

Variable	Description
Number cigarettes	Number of daily smoked cigarettes
Age	Age of individual in years
Age^2	Age squared
East-Germany	1 if individuals residents in East-Germany; 0 otherwise
Years of education	Years of individual's education
Years of education^2/100	Years of individual's education squared
Father high school degree	1 if father has a high school degree: 0 otherwise
Father medium school degree	1 if father has a medium school degree: 0 otherwise
Father low school degree	1 if father has a low school degree: 0 otherwise
Mother high school degree	1 if mother has a high school degree: 0 otherwise
Mother medium school degree	1 if mother has a medium school degree: 0 otherwise
Mother low school degree	1 if mother has a low school degree: 0 otherwise
Full time job	1 if individual has a full time job including civil-/military service; 0 otherwise
Part time job	1 if individual has a part time job; 0 otherwise
White collar job	1 if individual has a white collar job; 0 otherwise
Other job	1 if individual has a job that is self-employed, in apprenticeship or with armed forces; 0 otherwise
In educational training	1 if individual is in vocational training; 0 otherwise
Unemployed	1 if individual is unemployed and looking for a job; 0 otherwise
Not participating	1 if individual does not participate at the labor market; 0 otherwise
Actual hours worked	Number of actual hours worked
Married	1 if individual is married; 0 otherwise
Separated, divorced or widowed	1 if individual is separated, divorced or widowed; 0 otherwise
Single	1 if individual is single; 0 otherwise
Children younger 2	1 if individual has at least one children younger than two years old; 0 otherwise
Children aged 2 - 18	1 if individual has at least one children aged between 2 and 18; 0 otherwise
Number Children	Number of children
Monthly income more than 2000 Euro	1 if individual's (equivalence-) income is more than 2000 Euro; 0 otherwise
Monthly income 1000 - 1999 Euro	1 if individual's (equivalence-) income is between 1000 and 1999 Euro; 0 otherwise
Monthly income less than 1000 Euro	1 if individual's (equivalence-) income is less than 1000 Euro; 0 otherwise
Non-German nationality	1 if individuals is foreigner without German citizenship
Urban	1 if individual residents in a city with more than 100,000 inhabitants; 0 otherwise