A Binary-Ordered Probit Model of Cigarette Demand

Panagiotis P. Kasteridis
Department of Economics
505A Stokely Management Center
The University of Tennessee
Knoxville, TN 37996-0550, USA
E-mail: pkasteri@utk.edu

Murat K. Munkin
Department of Economics
531 Stokely Management Center
The University of Tennessee
Knoxville, TN 37996-0550, USA
E-mail: mmunkin@utk.edu

Steven T. Yen
Department of Agricultural Economics
302 Morgan Hall
The University of Tennessee
Knoxville, TN 37996-4518, USA
E-mail: syen@utk.edu

May 2007

Abstract

This study analyzes the demand for cigarettes fitting observed zero outcomes with a trivariate model consisting of an equation for the starting smoking decision, an equation for the quitting decision, and an equation that models the level of cigarettes consumed. Five competing specifications are considered to explain level, with the ordered probit, which accommodates pile-ups of counts in the dependent variable, providing the best fit. Marginal effects of explanatory variables are calculated providing strong evidence of race and gender differences in consumption patterns. The estimated marginal effects are robust to alternative categorizations of the level of cigarettes.

Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Portland, OR, July 29-August 1, 2007.

1 Introduction

This paper proposes an approach to analyze the demand for cigarettes utilizing a rich dataset that identifies non-smokers, potential smokers, quitters and actual smokers in the sample. This information brings additional gains in efficiency than that accomplished in the double-hurdle approach (e.g., Jones, 1989a) which has been the primary econometric tool in cigarette demand modeling. Following a general approach suggested in Jones (1989a), the proposed model consists of an equation that explains the decision to be a non-smoker, an equation that explains the quitting decision for those who started smoking in the past, and an equation that models the level (number) of cigarettes consumed. Five competing specifications are considered for the level equation, some of which have been attempted in the literature.

There is a long-standing interest in the empirical analysis of cigarette smoking by individuals because the health effect of cigarette smoking is an important public-health issue. Many studies are based on micro survey data, which allow for investigation of the roles of detailed socio-demographic characteristics. In modeling cigarette demand with microdata, it has become a standard approach to assume that cigarette consumption is subject to two decisions: whether to smoke and how much to smoke (Fry and Pashardes, 1994; Garcia and Labeaga, 1996; Jones, 1989a, 1989b, 1995; Labeaga, 1999; Mullahy, 1985). These models vary in specifications of the two censoring mechanisms. The double-hurdle model, a bivariate generalization of the Tobit model (Tobin, 1958), assumes that zero observations are attributed to both nonparticipation and censoring (e.g., Atkinson et al., 1984; Jones, 1989a). Specifically, with two separate processes to govern participation and consumption in the double-hurdle model, zero observations are generated by those who are non-smokers and those who are potential smokers but choose not to consume. In many datasets potential

smokers are usually not identifiable. However, availability of such information can lead to efficiency gains and simplifications of functional forms. Jones (1989a) introduces a trivariate model which features three stochastic processes accommodating starting, quitting and the level of cigarette smoking. Not only do we know individuals in our dataset who never smoked but also we can identify those who started smoking in the past and quit. Those who identify themselves as non-smokers and quitters have zero cigarette consumption. It is interesting to note that a few individuals among those who are smokers choose not to smoke, perhaps, trying to quit smoking. All this information is utilized in our approach.

Another important issue this paper addresses relates to the distribution of the level variable. The number of cigarettes is a count variable. Many of the double-hurdle model applications have been based on the bivariate normal distribution. While the normal distribution may be appropriate for applications based on cigarette expenditure data (Atkinson et al., 1984; Jones and Labeaga, 2003) or weekly consumption data (Jones, 1989a), it is unlikely to accommodate other forms of reported consumption. It is plausible to assume the Poisson or Negative Binomial distribution for the count variable. However, a close examination of the dependent variable in our dataset shows that the underlying distribution is neither normal nor Poisson related and it is difficult to expect a good fit from such models. Since most of the observed consumption values are reported as a fraction of a pack of cigarettes it is reasonable to group observations in categories and utilize the ordered outcome approach. We consider two competing specifications along that line: the ordered probit and sequential probit models.

In this paper, we accommodate our particular form of data by following the trivariate model of Jones (1989a) and trying alternative distributions for the level equation. Specifically, the model features a starting equation, a quitting equation, and an equation that

explains the level of consumption with five alternative specifications. First we analyze a joint sample that includes both male and female individuals. We find that the ordered probit model specification is preferred to the others based on the Akaike information criterion (Akaike, 1973). To facilitate interpretations of results we propose to calculate the marginal effects of variables for the preferred ordered probit model. The calculated effect with respect to gender suggests that further analysis applied to the male and female subsamples is desirable. We also check sensitivity of the calculated marginal effects to alternative categorization of the level variable.

The rest of the paper is organized as follows. Section 2 develops the model and presents five competing specifications of the level equation. Section 3 describes the data and variables used in the study. Section 4 considers an application and presents the results. Section 5 concludes the paper. The computational appendix presents theformulae for the marginal effects of explanatory variables.

2 Econometric Specification

Assume we observe N independent observations where the dependent variable of interest, y_i $(i = 0, 1, ..., N)$, is a count variable that displays a large proportion of zeros. Each individual i belongs to one of three groups. The first group includes individuals who had not started smoking by the time the survey was conducted. The second group are those who had smoked in the past but decided to quit and consider themselves quitters. The last group are current smokers. The binary decisions to start and quit smoking are both modeled with probit models. The starting equation is characterized by a latent equation

$$s_i^* = X_i\alpha_1 + \varepsilon_{1i}, \tag{1}$$

where latent variable s_i^* measures the difference in utility derived by individual i from starting and not starting smoking, X_i is a vector of exogenous variables, α_1 is a conformable parameter vector, and the error terms ε_{1i} are independently and identically distributed as standard normal, that is, $\varepsilon_{1i} \sim N(0,1)$. The observed binary variable for starting (S_i) relates to the latent variable (s_i^*) such that

$$S_i = \begin{cases} 1 \text{ iff } s_i^* > 0 \\ 0 \text{ otherwise} \end{cases}, \tag{2}$$

taking a value of 1 if the individual ever started smoking. Then the probability of starting is

$$\Pr(S_i = 1) = \Phi\left(X_i \alpha_1\right), \tag{3}$$

where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard normal. Likewise, quitting is characterized by latent variable q_i^*, which measures the difference in utility derived from smoking and quitting states, and specified as

$$q_i^* = X_i \alpha_2 + \varepsilon_{2i}, \tag{4}$$

where $\varepsilon_{2i} \sim N(0,1)$. The observed binary variable for quitting is defined as

$$Q_i = \begin{cases} 1 \text{ iff } q_i^* > 0 \\ 0 \text{ otherwise} \end{cases}, \tag{5}$$

where $Q_i = 1$ if a person continues smoking and $Q_i = 0$ if she quits, and

$$\Pr(Q_i = 1) = \Phi\left(X_i \alpha_2\right). \tag{6}$$

Let y_i be our dependent variable, measuring the level of smoking. Variable y_i takes the value of zero for either non-smokers ($S_i = 0$) or quitters ($Q_i = 0 | S_i = 1$). Non-zero values can only be observed conditional on ($S_i = 1$ and $Q_i = 1$). The joint likelihood of observing

the sample is

$$
\begin{aligned}
L(y|\theta) &= \prod_{S_i=0} \left[1 - \Pr(S_i = 1)\right] \prod_{S_i=1, Q_i=0} \Pr(S_i = 1) \left[1 - \Pr(Q_i = 1|S_i = 1)\right] \\
&\times \prod_{S_i=1, Q_i=1} \Pr\left(S_i = 1\right) \Pr\left(Q_i = 1|S_i = 1\right) \Pr\left(y_i|S_i = 1, Q_i = 1\right)
\end{aligned} \tag{7}
$$

where θ is a vector containing all parameters in the model, and the products are taken over sample observations satisfying $(S_i = 0)$, $(S_i = 1, Q_i = 0)$ and $(S_i = 1, Q_i = 1)$ conditions respectively.

For the conditional density $\Pr(y_i|S_i = 1, Q_i = 1)$, we consider five different specifications: a Gaussian model that truncates the error term to ensure non-negativity, two negative binomial models, an ordered probit and a sequential ordered probit model. The Gaussian specification which is the driving mechanism in the double-hurdle model (Jones, 1989a), and the single-hurdle model (Yen, 2005) assumes that y_i follows a truncated normal distribution with mean $X_i\beta$ and variance σ^2 such that

$$y_i = X_i\beta + u_i, \tag{8}$$

$$u_i > -X_i\beta \tag{9}$$

$$\Pr(y_i|S_i = 1, Q_i = 1) = \sigma^{-1} \frac{\phi\left((y_i - X_i\beta)/\sigma\right)}{\Phi(X_i\beta/\sigma)}, \tag{10}$$

where $\phi(\cdot)$ is the probability density function (pdf) of the standard normal.

We also consider two forms of the negative binomial model, namely NB1 and NB2, with probability mass functions of the forms

$$\Pr(y_i|S_i = 1, Q_i = 1) = \frac{\Gamma(y_i + \upsilon^{-1}\lambda_i)}{\Gamma(y_i + 1)\Gamma(\upsilon^{-1}\lambda_i)} \left(\frac{\upsilon^{-1}}{\upsilon^{-1} + 1}\right)^{\upsilon^{-1}\lambda_i} \left(\frac{1}{1 + \upsilon^{-1}}\right)^{y_i}, \tag{11}$$

$$\Pr(y_i|S_i = 1, Q_i = 1) = \frac{\Gamma(y_i + \upsilon^{-1})}{\Gamma(y_i + 1)\Gamma(\upsilon^{-1})} \left(\frac{\upsilon^{-1}}{\upsilon^{-1} + \lambda_i}\right)^{\upsilon^{-1}} \left(\frac{\lambda_i}{\upsilon^{-1} + \lambda_i}\right)^{y_i}, \tag{12}$$

where $\lambda_i = \exp(X_i\beta)$, $\Gamma(\cdot)$ is the Gamma function, and υ is the overdispersion parameter.

Finally, we specify two ordered outcome models. Construction of the ordered dependent variable is discussed in section 3. Assume that the dependent variable y_i, measuring the level of smoking conditional on $S_i = 1$ and $Q_i = 1$, takes integer values from 0 to M. Our first ordered outcome model specification assumes that y_i follows the ordered probit model. The latent variable Z_i which measures the propensity to smoke at different levels is assumed to be linear in X_i through the structural equation

$$Z_i = X_i\beta + u_i, \tag{13}$$

where the error term $u_i \sim N(0,1)$. Variable y_i relates to Z_i according to the observability condition

$$y_i = \begin{cases} 1 \text{ iff } Z_i \leq \tau_1 \\ 2 \text{ iff } \tau_1 < Z_i \leq \tau_2 \\ 3 \text{ iff } \tau_2 < Z_i \leq \tau_3 \\ . \\ M \text{ iff } \tau_{M-1} < Z_i \end{cases}, \tag{14}$$

where $\tau = (\tau_1, ..., \tau_{M-1})$ are threshold parameters and τ_1 is restricted to zero for identification. To complete the likelihood function specify

$$\Pr(y_i = m | S_i = 1, Q_i = 1) = \Phi(\tau_m - X_i\beta) - \Phi(\tau_{m-1} - X_i\beta), \tag{15}$$

so that

$$\Pr(y_i | S_i = 1, Q_i = 1) = \prod_{m=1}^{M} [\Pr(y_i = m | S_i = 1, Q_i = 1)]^{d_{im}}, \tag{16}$$

where d_{im} is a binary indicator such that $d_{im} = 1$ iff $y_i \in m$th category and 0 otherwise.

Our last model specification choice is the sequential probit model which assumes that the ordered variable y_i can take the value m only after the levels $1, ..., m-1$ have been reached. Then the conditional probability of reaching level m given the levels $1, ..., m-1$

have been reached is

$$\Pr(y_i = m \mid y_i \geq m, \beta, \tau) = \Phi\left(\tau_m - X_i\beta\right), \tag{17}$$

where X_i is a vector of covariates, β is a parameter vector, $\tau = (\tau_1, ..., \tau_{M-1})$ are threshold parameters. Then, the unconditional probabilities are

$$\Pr(y_i = m|\beta, \tau) = \Phi(\tau_m - X_i\beta) \prod_{k=1}^{m-1} [1 - \Phi(\tau_k - X_i\beta)],\ m = 1, ..., M-1. \tag{18}$$

$$\Pr(y_i = M \mid \beta, \tau) = \prod_{k=1}^{M-1} \left[1 - \Phi\left(\tau_k - X_i\beta\right)\right],\ m = M \tag{19}$$

To highlight a potential advantage of this model over the ordered probit model consider the latent representation of the sequential probit model which assumes a latent variable structure that generates the count outcome. Define latent variables

$$\xi_{mi} = X_i\beta + u_{mi}, \tag{20}$$

where $u_{mi} \sim N(0, 1)$. The latent variables ξ_{mi} represent propensities to continue to the next level. Applied to the number of cigarettes smoked, level m can be reached only after an individual makes it to level $m-1$ and ξ_{mi} defines propensities to smoke additional cigarettes to move to the next level of consumption. Counts are generated according to

$$y_i = \begin{cases} 1 & \text{iff } \xi_{1i} \leq \tau_1 \\ 2 & \text{iff } \xi_{1i} > \tau_1, \xi_{2i} \leq \tau_2 \\ 3 & \text{iff } \xi_{1i} > \tau_1, \xi_{2i} > \tau_2, \xi_{3i} \leq \tau_3 \\ . & . \\ M-1 & \text{iff } \xi_{1i} > \tau_1, \xi_{2i} > \tau_2, ..., \xi_{M-2,i} > \tau_{M-2}, \xi_{M-1,i} \leq \tau_{M-1} \\ M & \text{iff } \xi_{1i} > \tau_1, \xi_{2i} > \tau_2, ..., \xi_{M-2,i} > \tau_{M-2}, \xi_{M-1,i} > \tau_{M-1} \end{cases}. \tag{21}$$

Then

$$\Pr(y_i|S_i = 1, Q_i = 1) = \prod_{m=1}^{M-1} \left[\Pr(y_i = m \mid \beta, \tau)\right]^{d_{im}} \left[\Pr(y_i = M \mid \beta, \tau)\right]^{d_{iM}}. \tag{22}$$

As can be seen from Equation 21, the propensities to move to the next level are not ordered since the threshold parameters do not follow the order condition similar to that of the

ordered probit model. It is not clear *a priori* whether propensities to smoke more cigarettes increase with the level of smoking but allowing for a flexible propensity structure seems to be justified. It remains an empirical matter whether the sequential model would perform better than the ordered probit**.**

We restrict the error covariances across equations to be zeros. This is done because our data do not provide variables that would affect one equation but not the others. This is a limitation to our study. The assumption of independence cannot be tested and the fully-specified trivariate model cannot be estimated. For the current application with independent error terms, the likelihood function (Equation 7) suggests that the starting equation can be estimated separately with the whole sample, the quitting equation with the starter sample ($S = 1$), and the level equation with the smoker sample ($S = 1, Q = 1$).

3 Data and Variables

Data are compiled from the 1994–96 Continuing Survey of Food Intakes by Individuals (CSFII), collected by the Agricultural Research Service of the US Department of Agriculture (2000). A nationally representative survey, the CSFII 1994–96 were stratified, multistage area probability samples targeting individuals of all ages. With an overall response rate of 76.1 percent, the three-year data initially included 20607 individuals of all ages, of whom 10721 age 12 or over were asked about lifestyle and cigarette smoking. We focus on individuals age 15 or over as few of those age 12–14 reported smoking any cigarettes. After deleting observations with missing values on important variables, a total of 9587 individuals (4923 men and 4664 women) remain for analysis.

In the CSFII, each individual was asked (i) whether she/he had smoked 100 or more cigarettes in the entire life, and if yes, (ii) whether she/he currently smoked at the time

of the survey. Responses to these questions allow identification of starters and quitters, respectively. In addition, among the current smokers, each was asked the question: 'On average, how many cigarettes do you smoke per day?' The reported number is used as the quantity variable.

The quantities of cigarettes smoked are reported in the form of 'number per day', which feature a pile-up of counts at 10 cigarettes (0.5 pack), 20 (1 pack), 30 and 40 cigarettes, and so forth. It is unlikely that such pile-ups of counts can be adequately accommodated by the normal distribution. The histogram of the number of cigarettes smoked by the smokers only (Figure 1) has spikes at 5, 10, 15, 20, 30 cigarettes, which are self-reported consumption levels.

Since the number of cigarettes is self-reported and perhaps due to convenience of (or errors in) reporting there are disproportionately larger shares of outcomes measured in packs of 0.25, 0.5, 0.75, 1 and 1.5. Self-reported measurement errors are likely present in the sample as individuals might have rounded the number of cigarettes reported smoked to the closest integer in multiples of 5, such as 5, 10 and 20. It seems therefore reasonable to categorize the reported quantities in the following categories: 0–5 cigarettes, 6–10, 11–15, 16–20, 21–30, and over 31. The histogram of the constructed quantity variable is presented in Figure 2 for smokers in the sample. Another interesting feature of the sample is that there are a few individuals who identified themselves as smokers who never quit but whose cigarette consumption is zero.

The CSFII also includes detailed demographic information on each individual. Variables commonly used in the cigarette demand literature are included in the starting, quitting and consumption equations (Blaylock and Blisard, 1992a, 1992b; Jones, 1989a, 1989b, 1994, 1995). The explanatory variables are income, body mass, education and age, along

with dummy variables indicating urbanization (city, suburban), region (Northeast, Midwest, South), years of survey (Year95, Year96), race (Black), ethnicity (Hispanic), self-evaluated health, and whether the individual was a white-collar worker, had been diagnosed with cancer, high blood pressure or heart problems. Also included are lifestyle variables indicating whether the individuals had consumed alcohol in the past three months, had exercised regularly (no exercise and intensively), or was on any special diet (see Table 1). Price information is not available in the survey and so is included in the constant terms. However, the regional, urbanization and year dummy variables are expected to capture some price variations. In view of the literature in which socio-demographic variables are used as proxies for missing prices such as wage rate (e.g., Wales and Woodland, 1980) and the fact that, due to the insignificant role of transportation cost, cigarette prices are likely to be dominated by state level taxes, our use of regional, urbanization and time dummy variables serves as a remedial measure for the omission of the price variable[1]. Use of these variables in the literature and in the current study is elaborated in the empirical section below.

Of the final sample (N=9587), 4743 individuals (49.47%) ever started smoking, 2430 (25.35%) had smoked in the past but had quit. Among the 2313 (24.12%) current smokers, the average number of cigarettes is 18.64 per day. Detailed definitions and sample statistics for all variables by gender and for the pooled sample are presented in Table 1.

4 Application and Results

The application section is organized as follows. First we concentrate our analysis on the pooled sample including both males and females. We estimate five competing models and perform model selection based on the Akakie Information Criterion (AIC). The preferred

[1]State-level cigarette taxes would be a good proxy for price. However, for confidentiality reasons the CSFII sampling units were not identifiable by state.

model is the ordered probit for which we calculate marginal effects and partial changes. The estimated partial effect with respect to gender invites further analysis with the separate male and female subsamples, for which the ordered probit model is found to be the preferred model as well. As a robustness check we calculate marginal effects for the ordered probit model for a different segmentation of the observations, that is, with alternative definitions of the ordered dependent variable.

ML estimates of the starting and quitting equations for the pooled sample are presented in Table 2. These results are the same for all five competing models since the first two equations are independent of the conditional part and therefore parameters from the three equations are separable. The significant determinants of the starting variable at the 5% level of significance are education (negative), age, Black (negative), Hispanic (negative), healthy (negative), white-collar (negative), male, alcohol and no exercise. Quitting ($Q = 0$) is affected positively by the individual's income, body mass, education, age, the geographic variables city, Northeast and suburban, and the variables white-collar and diet. On the other hand, alcohol consumption has a negative impact on quitting. Blacks are less likely to quit smoking than others, while Hispanics and males are more successful in quitting.

The results of the starting equation should be interpreted as the effects of the covariates on the decision to be a non-smoker, which is an up-to-date decision taken throughout the entire life up to the point of the survey. Thus, all variables affecting level are included in the starting equation because they affect the current choice to be a non-smoker. The same reasoning applies to the current decision to be and remain a quitter for those who once started smoking.

We find that body mass does not affect the decision to be a non-smoker but individuals with a higher body mass are more likely to quit smoking. This result is consistent with the

findings of Jones (1994, 1995) and Blaylock and Blisard (1992b). Education improves the individual's cognitive skills regarding the risks associated with smoking, thereby discouraging starting and motivating quitting. Similar result was reported by Blaylock and Blisard (1992a). It is also in line with Hsieh (1998) who reports that the probability of quitting smoking increases with years of formal education and Ault et al. (2004) who find that those who attend high school have a higher probability of smoking and a lower probability of quitting than others. As expected, consumption of alcohol, another addictive good, decreases the probability of being a non-smoker and the probability of quitting. Individuals who exercise only rarely have a smaller probability to be a non-smoker but the lack of exercise has no effect on quitting.

Jones (1995) and Yen (2005) found that age has a negative impact on participation arguing that most smokers start smoking as a teenager or young adult. We find that age is negatively correlated with the non-smoking decision. However, older individuals are more likely to quit. Individuals who are black are more likely to be non-smokers but less successful in quitting. The empirical literature in the effects of self-perceived health status is mixed. We find that poor health is negatively related to non-smoking. As to the role of gender, males are more prone to starting smoking but are more successful in quitting.

Table 2 also presents ML results for the conditional portion of the five competing models outlined above. For the NB1, NB2, ordered probit, and sequential ordered probit models, body mass index, age, Midwest, South and sex are positive and significant while city, Black, ethnicity (Hispanic), and white-collar are negative and significant. Note that education is not significant for any of the competing models.

The ordered probit model fits the data the best according to the log-likelihood values

and the respective AIC[2]. The threshold parameters $\tau_2, ..., \tau_5$ are all significant at the 1% level, justifying the use of all six categories over combining some categories. It is known for the ordered probit model that the signs of the coefficients may not relate directly to the directions of the effects of variables on the probabilities of categories. In addition, it is useful to relate each category to the quantity level it stands for. To accomplish this goal we calculate marginal effects of each variable on the ordered probit probabilities

$$\Pr(y_i = m|X_i) = \left[\Phi\left(\tau_m - X_i\beta\right) - \Phi\left(\tau_{m-1} - X_i\beta\right)\right]\Phi(X_i\alpha_1)\Phi(X_i\alpha_2), \tag{23}$$

and the conditional mean of the dependent variable

$$E(y_i|X_i) = \sum_{m=1}^{M} \overline{y}_m \Pr(y_i = m|X_i), \tag{24}$$

where $\overline{y}_m$ is the category mean for the m^{th} category. Since the ordered model involves combining observations into categories our results are likely to depend on how the categories are defined (an issue investigated later) and how averaging takes place. To address this issue we also use category medians and modes as weights in calculating the marginal effects but since they produce similar results the corresponding marginal effects are not reported. The marginal effects on the conditional mean (Equation 24) is the sum of the marginal effects on the probabilities (Equation 23), weighted by the category means $\overline{y}_m$. The effects of each binary explanatory variable are derived by simulating a finite change (i.e., from 0 to 1) in the variable, holding all other variables constant. Analytical expressions for the marginal effects are enclosed in the appendix.

The marginal effects of explanatory variables are presented in Table 3. It is interesting to note that significance and signs of the marginal effects vary with categories and the

[2]Note that the ordered probit specification for conditional level can be extended to the multinomial probit (or logit). The lack of variations among some of the explanatory variables however prevented this pursuit without consolidating the ordered dependent variable.

weighted average. Income has a negative and significant effect on the cigarette consumption level, with higher income individuals smoking fewer cigarettes than others. The roles of income in cigarette smoking are largely inconclusive in the literature. Blaylock and Blisard (1992a) find that cigarettes are an inferior good. Tansel (1993) reports a low but positive income elasticity of demand, which is consistent with the addictive nature of cigarettes, whilst Goel and Nelson (2005) find that income does not affect smoking prevalence among adults.

Body mass, education and age all have significant and overall negative effects. Body mass and age both have positive effects on the probability of consuming in the lowest (0–5) category and negative effects on the probabilities of the higher categories, whereas education has negative effects on the probabilities of all but the highest (> 30) categories. The persistently negative effects of body mass, education and age translate into the significant and negative effects on the level of cigarette consumption.

Also presented in Table 3 are the discrete effects of binary explanatory variables. Positive effects on level are seen in Midwest, South and male, which are due to the positive (negative) effects of these variables on the probabilities of consuming in the higher (lower) categories. Thus, men consume more cigarettes than women, and individuals residing in the Midwest and the South consume more cigarettes than those residing in the West. Consumption of alcohol and lack of exercise are also positive and significant. Opposite effects are seen in the other variables, which include residing in the city or a suburban area (relative to rural area), residing in the Northeast, being on a special diet and being Black, Hispanic, healthy and white-collar worker. Individuals with these characteristics consume fewer cigarettes than others. Overall, the effects of most variables on the probabilities and conditional level, though statistically significant, are fairly small. The more notable effects of variables are

seen in ethnicity, gender, race, job status, use of alcohol and diet, and lack of exercise, with individuals of the Hispanic origin smoking 3.43 fewer cigarettes (per day) than non-Hispanics, men smoking 1.58 more cigarettes than women, alcohol consumers smoking 2.21 more cigarettes than non-consumers, and Blacks smoking 1.17 fewer cigarettes than non-Blacks. The effects of other variables are all very small, with individuals 10 years older smoking only 0.77 fewer cigarette than their younger counterparts, and with all other binary variables having the effects of less than 1 cigarette per day on average.

The estimated marginal effects with respect to gender in the pooled men-women sample invite further analysis for men and women separately. Such separate analysis can provide further insight into gender differences in cigarette consumption. We perform a formal test to determine whether male and female samples can be pooled or should be used separately in modelling cigarette demand. The results, reported in Table 4, suggest the hypothesis of equal parameters between genders is rejected for all competing models, favoring estimation of each model for men and women separately.

Table 5 presents estimation results for men and women, using the preferred ordered probit specification for the conditional level. White-collar men are more likely to be non-smokers and more successful in quitting. In contrast, having a white-collar job has no effect on the starting or quitting variables among women. This result for women contradicts the usual argument that antismoking messages have the greatest effect on women in better jobs. Men on a special diet are more likely to be non-smokers and they smoke fewer cigarettes on average; these effects of special diet are not seen in women. Other differences between genders are observed in the effects of city, Hispanic, healthy and alcohol on quitting, and the effects of age and Northeast on the level of smoking. Table 6 presents the mean marginal effects and average partial changes. The signs and significance of the effects are similar to

those for the pooled sample. However, the marginal effects of body mass and age on the level, as well as the discrete effects of city, suburban, Midwest, South, no exercise, Hispanic, and white-collar are smaller for women than for men. Yen (2005) calculated elasticities with respect to the same variables using hurdle models and found smaller effects for women as well. On the other hand our marginal effects with respect to Black and healthy are larger for women, whilst Yen documents the opposite. Unconditional on smoking, a black woman smokes 1.41 fewer cigarettes than other women while a black man smokes 0.97 fewer cigarettes than other men. In general, our marginal effects are smaller than the those reported by Yen for both men and women.

Finally, we investigate sensitivity of the marginal effects to the choice of categories. Our original categorization of quantity, labeled "Categorization 1" in Table 6, was based on the assumption that self-reported numbers of cigarettes smoked are rounded upward to the closest fraction of a cigarette pack during reporting. Self-reported consumption levels may be subject to non-systematic measurement errors. While our paper does not address the measurement errors per se, as a robustness check of the marginal effects we categorized quantity with an alternative set of cut-offs (Categorization 2): 0–4 cigarettes, 5–9, 10–14, 15–19, 20–29, and over 30. The marginal effects calculated at the conditional means based on this alternative categorization are presented in Table 6. The results show that the calculated mean effects are similar between the two categorization schemes.

5 Concluding Remarks

We address the issues of zero observations in modeling cigarette smoking. Our dataset contains information which allows investigation of three key elements of cigarette smoking: starting, quitting and the level of consumption. Despite a lack of exclusion restrictions,

a shortcoming, which prevents estimation of the fully specified model with dependent covariances, the special feature of the data allows us to construct a statistical model that accommodate skewness of distribution and pile-ups of counts in the number of cigarettes smoked. We model starting and quitting as binary variables, and use alternative specifications to accommodate the level of smoking. The ordered probit model provides the best fit to the data. Unlike other ordered probit specifications based entirely on the category information, the approach we follow allows a way to relate each category to the level associated with the category, thereby allowing the calculation of marginal effects of variables on the level of cigarettes smoked.

The empirical analysis was carried out separately for men and women, and we find strong evidence of gender differences in cigarette consumption in terms of parameter estimates and marginal effects of explanatory variables. Estimation of our preferred model, with the ordered probit specification for the conditional level of smoking, requires categorization of the quantity variable and we find our results are robust to alternative categorizations of the variable. Although our data do not allow estimation of the dependent trivariate model due to a lack of exclusion restrictions, we present the likelihood function which can be used in future applications when the data become available. Whilst we apply the statistical model to cigarette smoking, the model can be useful in other applications, such as consumption of soft drink or vegetables which are likely to be reported in cans or servings, in which dependent variable may be censored and may feature skewness in distribution and pile-ups at certain values.

Appendix

Marginal Effects for the Trivariate Ordered Probit with Independence

Differentiating the conditional mean of the dependent variable (Equation 24) and averaging the derivatives over the sample, the mean (average) marginal effect of the conditional mean with respect to the explanatory (continuous) variable X_{ik} is

$$\begin{aligned} MME &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial E\left(y_i \mid X_i\right)}{\partial X_{ik}} \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{\sum_{m=0}^{M}\frac{\partial\left[\overline{y}_m \Pr\left(y_i = m \mid X_i\right)\right]}{\partial X_{ik}}\right\} \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{\sum_{m=1}^{M}\overline{y}_m\frac{\partial\left[\Pr\left(y_i = m \mid S_i = 1, Q_i = 1\right)\Pr(S_i = 1)\Pr(Q_i = 1)\right]}{\partial X_{ik}}\right\}. \end{aligned}$$

Note that for $m = 0$, y_i is zero so that $\overline{y}_m = 0$.This simplifies our calculations since the term $\overline{y}_m \Pr\left(y_i = m \mid X_i\right)$ vanishes for $m = 0$. One can calculate the partial derivative with respect to X_{ik} as

$$\frac{\partial\left[\Pr\left(y_i = m \mid S_i = 1, Q_i = 1\right)\Pr(S_i = 1)\Pr(Q_i = 1)\right]}{\partial X_{ik}} = \Pr(y_i = m|X_i)$$

$$\times\left[\frac{\beta_k[\phi\left(\tau_{m-1} - X_i\beta\right) - \phi\left(\tau_m - X_i\beta\right)]}{\Phi\left(\tau_m - X_i\beta\right) - \Phi\left(\tau_{m-1} - X_i\beta\right)} + \frac{\alpha_{1k}\phi(X_i\alpha_1)}{\Phi(X_i\alpha_1)} + \frac{\alpha_{2k}\phi(X_i\alpha_2)}{\Phi(X_i\alpha_2)}\right],$$

where $\Pr(y_i = m|X_i)$ is defined in the text (Equation 23).

References

Akaike H. (1973) Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F. (eds), *2nd International Symposium on Information Theory*, Budapest: Akademiai Kiado, 267–81.

Atkinson, A.B., Gomulka, J. and Stern, N. (1984) Household expenditure on tobacco 1970–1980: evidence from the Family Expenditure Survey, London: London School of Economics.

Ault, R.W., Ekelund, R.B., Jackson, J.D. and Saba, R.P. (2004) Smokeless tobacco, smoking cessation and harm reduction: an economic analysis, *Applied Economics*, **36**, 17–29.

Blaylock, J.R. and Blisard, W.N. (1992a) US cigarette consumption: the case of low income women, *American Journal of Agricultural Economics*, **74**, 698–705.

Blaylock, J.R. and Blisard, W.N. (1992b) Self-evaluated health status and smoking behavior, *Applied Economics*, **24**, 429–35.

Fry, V. and Pashardes, P. (1994) Abstention and aggregation in consumer demand: zero tobacco expenditures, *Oxford Economic Papers*, **46**, 502–18.

Garcia, J. and Labeaga, J.M. (1996), Alternative approaches to modelling zero expenditure: an application to Spanish demand for tobacco, *Oxford Bulletin of Economics and Statistics*, **58**, 489–506.

Goel, R.K. and Nelson, M.A. (2005) Tobacco policy and tobacco use: differences across tobacco types, gender and age, *Applied Economics*, **37**, 765–71.

Hsieh,C.R. (1998) Health risk and the decision to quit smoking, *Applied Economics*, **30**, 795–804.

Jones, A.M. (1989a) A double-hurdle model of cigarette consumption, *Journal of Applied Econometrics*, **4**, 23–39.

Jones, A.M. (1989b) The UK demand for cigarettes 1954–86, a double-hurdle approach, *Journal*

of Health Economics, **8**, 133–41.

Jones, A.M. (1994) Health, addiction, social interaction and the decision to quit smoking, *Journal of Health Economics*, **13**, 93–110.

Jones, A.M. (1995) A microeconometric analysis of smoking in the UK Health and Lifestyle Survey, Discussion Paper 139, The University of York, September.

Jones, A.M. and Labeaga, J.M. (2003). Individual heterogeneity and censoring in panel data estimates of tobacco expenditure, *Journal of Applied Econometrics*, **18**, 157–77.

Labeaga, J.M. (1999) A double-hurdle rational addiction model with heterogeneity: estimating the demand for tobacco, *Journal of Econometrics*, **93**, 49–72.

Mullahy, J. (1985) Cigarette smoking, habits, health concerns, and heterogeneous unobservables in a microeconometric analysis of consumer demand, Ph.D. dissertation, University of Virginia, 1985.

Tansel, A. (1993) Cigarette demand, health scares and education in Turkey, *Applied Economics*, **25**, 521–29.

Tobin, J. (1958) Estimation of relationships for limited dependent variables, *Econometrica*, **26**, 24–36.

US Department of Agriculture (2000) Continuing Survey of Food Intakes by Individuals 1994–96 and 1998, CD-ROM, Agricultural Research Service, Washington, DC.

Wales, T.J., and Woodland A.D. (1980) Sample selectivity and the estimation of labor supply functions, *International Economic Review*, **21**, 437–68.

Yen, S.T. (2005) Zero observations and gender differences in cigarette consumption, *Applied Economics*, **37**, 1839–49.

Table 1. Summary statistics

Variable	Definition	Mean		
		Pooled (N=9587)	Male (N=4923)	Female (N=4664)
Cigarettes	Number per day (full sample)	4.50 (9.98)	5.19 (11.04)	3.76 (8.68)
	Number per day (consuming sample)	18.64 (12.23)	20.29 (13.03)	16.66 (10.89)
	Proportion of smokers	24.12%	25.59%	22.58%
	Proportion ever started smoking	49.47%	56.65%	41.90%
	Proportion quitters	25.35%	31.06%	19.32%
Income	Per capita income in thousand USD	15.29 (12.54)	15.84 (13.02)	14.70 (11.98)
Body mass	Quetelet's body mass index[a]	26.06 (5.25)	26.26 (4.55)	25.84 (5.90)
Education	Years of formal education	12.50 (3.09)	12.56 (3.18)	12.43 (2.99)
Age	Age in years	47.19 (18.95)	47.42 (18.97)	46.95 (18.93)
Dummy variables (yes = 1, no = 0)				
Male	Gender is male	0.51		
City	Resides in central city	0.29	0.28	0.31
Suburban	Resides in suburban area	0.45	0.46	0.43
Rural	Resides in rural area (reference)	0.26	0.26	0.26
Northeast	Resides in the North or Northeast	0.18	0.18	0.18
Midwest	Resides in the Midwest	0.24	0.24	0.25
South	Resides in the South	0.37	0.36	0.37
West	Resides in the West (reference)	0.21	0.22	0.20
Black	Race is Black	0.12	0.10	0.13
Hispanic	of Hispanic origin	0.04	0.04	0.04
Healthy	Self-evaluated health is fair or better	0.83	0.84	0.82
White-collar	A white-collar worker	0.23	0.25	0.21
Cancer	Has been diagnosed of cancer	0.06	0.06	0.06
BP-heart	Has had blood pressure/heart problems	0.27	0.28	0.27
Alcohol	Consumed alcohol in past 3 months	0.62	0.67	0.56
Intensive exercise	Exercises 2–4 times per week or more	0.50	0.57	0.42
Moderate exercise	Exercises 1–4 times per month (reference)	0.12	0.12	0.14
No exercise	Exercises rarely or never	0.38	0.31	0.44
Diet	On a special diet	0.17	0.14	0.20
Year94	Survey conducted in 1994 (reference)	0.33	0.33	0.34
Year95	Survey conducted in 1995	0.34	0.34	0.34
Year96	Survey conducted in 1996	0.33	0.33	0.32

Note: Standard deviations in parentheses.
[a]Quetelet's body mass index=(weight in kg)/(height in metres).

Table 2. ML estimates: pooled sample

	Starting (N=9587)	Not quitting[c] (N=4743)	Level (N=2313)				
Variable			Gaussian	NB1	NB2	OP	SOP
Constant	-0.501[a]	3.163[a]	0.322	2.476[a]	2.384[a]	0.648[a]	0.683[a]
	(0.116)	(0.189)	(3.247)	(0.123)	(0.125)	(0.208)	(0.161)
Income	-0.0002	-0.005[a]	0.058	0.001	0.001	0.002	0.002
	(0.0012)	(0.002)	(0.040)	(0.001)	(0.001)	(0.002)	(0.002)
Body mass	0.001	-0.027[a]	0.343[a]	0.011[a]	0.013[a]	0.018[a]	0.014[a]
	(0.003)	(0.004)	(0.082)	(0.003)	(0.003)	(0.004)	(0.003)
Education	-0.031[a]	-0.065[a]	-0.214	-0.005	-0.003	-0.014	-0.011
	(0.005)	(0.008)	(0.170)	(0.006)	(0.006)	(0.010)	(0.007)
Age	0.012[a]	-0.028[a]	0.088[a]	0.002[a]	0.004[a]	0.004[a]	0.004[a]
	(0.001)	(0.001)	(0.029)	(0.001)	(0.001)	(0.002)	(0.001)
City	-0.022	-0.157[a]	-4.419[a]	-0.125[a]	-0.134[a]	-0.233[a]	-0.182[a]
	(0.037)	(0.055)	(1.095)	(0.036)	(0.035)	(0.060)	(0.045)
Suburban	-0.044	-0.181[a]	-1.377	-0.024	-0.027	-0.042	-0.030
	(0.037)	(0.049)	(0.934)	(0.032)	(0.031)	(0.055)	(0.041)
Northeast	-0.004	-0.208[a]	-1.602	-0.017	-0.066	-0.083	-0.098
	(0.044)	(0.064)	(1.485)	(0.050)	(0.048)	(0.078)	(0.060)
Midwest	-0.014	-0.011	3.788[a]	0.174[a]	0.129[a]	0.261[a]	0.152[a]
	(0.041)	(0.059)	(1.272)	(0.044)	(0.041)	(0.068)	(0.050)
South	-0.014	0.016	4.586[a]	0.208[a]	0.152[a]	0.331[a]	0.218[a]
	(0.038)	(0.056)	(1.162)	(0.041)	(0.038)	(0.064)	(0.048)
Black	-0.125[a]	0.289[a]	-12.677[a]	-0.366[a]	-0.415[a]	-0.652[a]	-0.526[a]
	(0.044)	(0.068)	(1.445)	(0.041)	(0.043)	(0.068)	(0.057)
Hispanic	-0.471[a]	-0.287[a]	-20.782[a]	-0.635[a]	-0.707[a]	-1.140[a]	-0.909[a]
	(0.076)	(0.133)	(4.247)	(0.100)	(0.122)	(0.170)	(0.163)
Healthy	-0.223[a]	-0.086[b]	-1.290	0.001	-0.040	-0.059	-0.070
	(0.039)	(0.052)	(1.048)	(0.037)	(0.035)	(0.063)	(0.047)
White-collar	-0.103[a]	-0.109[a]	-3.051[a]	-0.098[a]	-0.101[a]	-0.165[a]	-0.117[a]
	(0.036)	(0.054)	(1.137)	(0.038)	(0.038)	(0.063)	(0.047)
Cancer	0.107[b]	-0.108	0.316	0.021	0.004	-0.025	-0.027
	(0.060)	(0.075)	(1.780)	(0.054)	(0.060)	(0.095)	(0.072)
BP-heart	-0.026	-0.047	-0.703	-0.038	-0.019	-0.055	-0.029
	(0.035)	(0.048)	(1.012)	(0.036)	(0.034)	(0.060)	(0.045)
Male	0.360[a]	-0.158[a]	6.337[a]	0.182[a]	0.219[a]	0.345[a]	0.278[a]
	(0.027)	(0.041)	(0.830)	(0.027)	(0.027)	(0.046)	(0.035)
Alcohol	0.568[a]	0.137[a]	-0.337	-0.027	-0.020	-0.044	-0.021
	(0.030)	(0.046)	(0.886)	(0.030)	(0.031)	(0.052)	(0.041)
Intensive exercise	-0.028	-0.074	-2.096	-0.039	-0.049	-0.095	-0.022
	(0.042)	(0.066)	(1.317)	(0.047)	(0.044)	(0.074)	(0.055)
No exercise	0.144[a]	0.106	1.594	0.094[a]	0.058	0.108	0.110[a]
	(0.044)	(0.068)	(1.329)	(0.047)	(0.044)	(0.075)	(0.056)
Diet	-0.025	-0.260[a]	-1.760	-0.070	-0.068	-0.101	-0.059
	(0.038)	(0.054)	(1.303)	(0.046)	(0.043)	(0.073)	(0.055)
Year95	-0.044	0.112[a]	0.211	0.017	0.016	-0.030	-0.014
	(0.033)	(0.048)	(0.887)	(0.032)	(0.031)	(0.054)	(0.041)
Year96	-0.005	0.069	-0.681	-0.017	-0.009	-0.055	-0.044
	(0.032)	(0.048)	(0.892)	(0.033)	(0.032)	(0.053)	(0.041)

Table 2 (Continued).

	Starting (N=9587)	Not quitting[c] (N=4743)	Level (N=2313)				
Variable			Gaussian	NB1	NB2	OP[d]	SOP
σ			14.562[a]				
			(0.581)				
α				6.492[a]	0.361[a]		
				(0.263)	(0.015)		
τ_2						0.701[a]	0.365[a]
						(0.030)	(0.046)
τ_3						0.995[a]	0.134[a]
						(0.034)	(0.052)
τ_4						1.944[a]	1.421[a]
						(0.043)	(0.050)
τ_5						2.446[a]	1.267[a]
						(0.048)	(0.066)
Log-likelihood[e]			-17600.97	-17621.43	-17632.10	-12712.14	-39458.79
AIC			3.686	3.691	3.693	2.667	8.247

Note: Standard errors in parentheses.
[a,b] Denote significance at the 5% and 10% levels respectively.
[c]We label the second column as "not quitting" to be consistent with definition of the quitting equation in the text (Equation 5).
[d]OP and SOP stand for Ordered Probit and Sequential Ordered Probit respectively.
[e]The log-likelihood values and AIC reported at the bottom of each model correspond to the trivariate models.

Table 3. Mean marginal effects and average discrete changes: pooled sample

Variable	Mean	Probabilities					
		$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$
		Marginal effects of continuous variables					
Income	-0.033[a]	0.00030	-0.00023[a]	-0.00010[a]	-0.00025[a]	-0.00005	-0.00054
	(0.018)	(0.00044)	(0.00010)	(0.00004)	(0.00012)	(0.00007)	(0.00046)
Body mass	-0.209[a]	0.00278[a]	-0.00131[a]	-0.00055[a]	-0.00110[a]	-0.00011	-0.00399[a]
	(0.036)	(0.00085)	(0.00021)	(0.00009)	(0.00027)	(0.00014)	(0.00088)
Education	-0.208[a]	-0.00507[a]	-0.00269[a]	-0.00159[a]	-0.00564[a]	-0.00230[a]	0.00077
	(0.080)	(0.00189)	(0.00046)	(0.00021)	(0.00058)	(0.00031)	(0.00196)
Age	-0.077[a]	0.00057[b]	-0.00051[a]	-0.00025[a]	-0.00070[a]	-0.00021[a]	-0.00117[a]
	(0.014)	(0.00032)	(0.00007)	(0.00004)	(0.00009)	(0.00005)	(0.00035)
		Average effects of discrete variables					
City	-1.094[a]	0.00759[a]	0.00114	-0.00171	-0.01448[a]	-0.00900[a]	-0.01286[a]
	(0.250)	(0.00356)	(0.00271)	(0.00120)	(0.00387)	(0.00200)	(0.00278)
Suburban	-0.826[a]	-0.00312	-0.00577[a]	-0.00355[a]	-0.01319[a]	-0.00563[a]	-0.00669[a]
	(0.240)	(0.00315)	(0.00258)	(0.00114)	(0.00352)	(0.00186)	(0.00288)
Northeast	-0.841[a]	-0.00086	-0.00426	0.00310[a]	-0.01303[a]	-0.00604[a]	-0.00746[a]
	(0.294)	(0.00402)	(0.00322)	(0.00143)	(0.00457)	(0.00225)	(0.00317)
Midwest	0.552[b]	-0.01242[a]	-0.00871[a]	-0.00242[b]	0.00224	0.00551[a]	0.01157[a]
	(0.311)	(0.00318)	(0.00294)	(0.00137)	(0.00439)	(0.00232)	(0.00380)
South	0.822[a]	-0.01562[a]	-0.01002[a]	-0.00239[b]	0.00535	0.00792[a]	0.01528[a]
	(0.285)	(0.00327)	(0.00278)	(0.00126)	(0.00406)	(0.00222)	(0.00349)
Black	-1.166[a]	0.04418[a]	0.02072[a]	0.00342[a]	-0.01394[a]	-0.01358[a]	-0.02010[a]
	(0.246)	(0.00649)	(0.00370)	(0.00151)	(0.00468)	(0.00201)	(0.00206)
Hispanic	-3.425[a]	0.02796[a]	-0.01457[a]	-0.01430[a]	-0.06060[a]	-0.02591[a]	-0.02823[a]
	(0.249)	(0.01106)	(0.00480)	(0.00226)	(0.00555)	(0.00199)	(0.00179)
Healthy	-1.227[a]	-0.00600	-0.00947[a]	-0.00555[a]	-0.01983[a]	-0.00821[a]	-0.00954[a]
	(0.317)	(0.00387)	(0.00320)	(0.00133)	(0.00439)	(0.00256)	(0.00399)
White-collar	-1.031[a]	0.00224	-0.00270	-0.00290[a]	-0.01493[a]	-0.00782[a]	-0.01040[a]
	(0.236)	(0.00348)	(0.00262)	(0.00113)	(0.00363)	(0.00189)	(0.00265)
Cancer	-0.052	0.00192	0.00101	0.00020	-0.00052	-0.00061	-0.00096
	(0.429)	(0.00572)	(0.00462)	(0.00202)	(0.00637)	(0.00326)	(0.00497)
BP-heart	-0.365	0.00091	-0.00081	-0.00095	-0.00514	-0.00277	-0.00381
	(0.248)	(0.00353)	(0.00268)	(0.00112)	(0.00359)	(0.00196)	(0.00300)
Male	1.579[a]	-0.00990[a]	-0.00176	0.00221[a]	0.02001[a]	0.01288[a]	0.01908[a]
	(0.197)	(0.00257)	(0.00217)	(0.00095)	(0.00304)	(0.00164)	(0.00240)
Alcohol	2.216[a]	0.02148[a]	0.02586[a]	0.01332[a]	0.04003[a]	0.01349[a]	0.01251[a]
	(0.200)	(0.00250)	(0.00228)	(0.00127)	(0.00319)	(0.00165)	(0.00241)
Intensive exercise	-0.533[b]	0.00207	-0.00045	-0.00110	-0.00710	-0.00416[b]	-0.00599
	(0.322)	(0.00425)	(0.00351)	(0.00155)	(0.00485)	(0.00248)	(0.00373)
No exercise	1.126[a]	0.00111	0.00517	0.00385[a]	0.01669[a]	0.00809[a]	0.01054[a]
	(0.371)	(0.00442)	(0.00378)	(0.00176)	(0.00547)	(0.00284)	(0.00429)
Diet	-1.114[a]	-0.00194	-0.00631[a]	-0.00434[a]	-0.01747[a]	-0.00788[a]	-0.00959[a]
	(0.262)	(0.00376)	(0.00277)	(0.00118)	(0.00387)	(0.00211)	(0.00305)

Note: Standard errors in parentheses.
[a,b] Denote significance at the 5% and 10% levels respectively.

Table 4. Likelihood-ratio tests for gender differences and AIC for male and female subsamples

Model	Pooled	Male		Female		LR	df	p-value
	Log-likelihood	Log-likelihood	AIC	Log-likelihood	AIC			
Gaussian	-17600.97 (70)	-9463.85 (67)	3.872	-8016.70 (67)	3.466	240.84	64	< 0.001
NB1	-17621.43 (70)	-9487.46 (67)	3.882	-8011.67 (67)	3.464	244.60	64	< 0.001
NB2	-17632.10 (70)	-9494.60 (67)	3.884	-8021.60 (67)	3.469	231.80	64	< 0.001
OP[a]	-12712.14 (73)	-6715.02 (70)	2.756	-5870.96 (70)	2.548	252.32	67	< 0.001
SOP	-39458.79 (73)	-21139.79 (70)	10.039	-18198.90 (70)	7.834	240.20	67	< 0.001

[a]OP and SOP stand for Ordered Probit and Sequential Ordered Probit respectively.

Table 5. ML estimation of the ordered probit parts for male and female subsamples

	Male			Female		
Variable	Starting (N=4923)	Not quitting (N=2789)	Level (N=1260)	Starting (N=4664)	Not quitting (N=1954)	Level (N=1053)
Constant	-0.334^b	3.238^a	0.518^b	-0.529^a	3.054^a	0.560^b
	(0.173)	(0.269)	(0.302)	(0.158)	(0.280)	(0.304)
Income	-0.001	-0.004^a	0.002	(0.002)	-0.007^a	0.003
	(0.002)	(0.002)	(0.003)	(0.002)	(0.003)	(0.003)
Body mass	0.005	-0.032^a	0.019^a	0.003	-0.026^a	0.023^a
	(0.004)	(0.006)	(0.007)	(0.003)	(0.005)	(0.007)
Education	-0.042^a	-0.055^a	-0.013	-0.017^a	-0.083^a	-0.002
	(0.008)	(0.010)	(0.013)	(0.008)	(0.014)	(0.015)
Age	0.019^a	-0.032^a	0.007^a	0.005^a	-0.021^a	0.003
	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)
City	-0.061	-0.101	-0.261^a	0.049	-0.252^a	-0.155^b
	(0.053)	(0.073)	(0.081)	(0.053)	(0.085)	(0.089)
Suburban	-0.074	-0.132^a	-0.042	-0.012	-0.263^a	0.020
	(0.047)	(0.064)	(0.076)	(0.049)	(0.077)	(0.082)
Northeast	-0.010	-0.149^b	0.043	0.0004	-0.285^a	-0.247^a
	(0.062)	(0.084)	(0.105)	(0.063)	(0.097)	(0.116)
Midwest	-0.003	-0.019	0.298^a	0.001	0.025	0.253^a
	(0.057)	(0.079)	(0.094)	(0.059)	(0.091)	(0.099)
South	0.017	-0.020	0.380^a	-0.045	0.081	0.328^a
	(0.052)	(0.075)	(0.087)	(0.055)	(0.087)	(0.094)
Black	-0.117^b	0.316^a	-0.594^a	-0.146^a	0.236^a	-0.773^a
	(0.065)	(0.090)	(0.089)	(0.062)	(0.103)	(0.111)
Hispanic	-0.298^a	-0.394^a	-1.154^a	-0.713^a	-0.113	-1.101^a
	(0.100)	(0.162)	(0.205)	(0.125)	(0.248)	(0.291)
Healthy	-0.183^a	-0.150^a	0.076	-0.271^a	-0.025	-0.122
	(0.058)	(0.070)	(0.090)	(0.055)	(0.078)	(0.088)
White-collar	-0.131^a	-0.207^a	-0.139^b	-0.082	0.065	-0.269^a
	(0.050)	(0.072)	(0.084)	(0.053)	(0.083)	(0.095)
Cancer	-0.104	-0.152	-0.055	0.212^a	-0.055	-0.017
	(0.088)	(0.108)	(0.155)	(0.082)	(0.109)	(0.123)
BP-heart	0.033	-0.064	-0.039	-0.079	-0.009	-0.064
	(0.050)	(0.063)	(0.083)	(0.051)	(0.075)	(0.089)
Alcohol	0.493^a	0.202^a	-0.074	0.610^a	0.076	0.006
	(0.043)	(0.062)	(0.075)	(0.043)	(0.071)	(0.075)
Intensive exercise	-0.014	-0.042	-0.055	-0.00001	-0.140	-0.146
	(0.059)	(0.091)	(0.105)	(0.059)	(0.098)	(0.107)
No exercise	0.098	0.134	0.130	0.223^a	0.050	0.117
	(0.066)	(0.095)	(0.112)	(0.061)	(0.099)	(0.103)
Diet	-0.129^a	-0.372^a	-0.239^a	0.033	-0.155^a	-0.024
	(0.057)	(0.078)	(0.120)	(0.050)	(0.077)	(0.094)
Year95	-0.115^a	0.107^b	-0.006	0.055	0.121^b	-0.022
	(0.047)	(0.063)	(0.073)	(0.046)	(0.073)	(0.082)
Year96	-0.059	0.035	-0.069	0.057	0.120	0.032
	(0.046)	(0.064)	(0.072)	(0.047)	(0.075)	(0.081)

Table 5 (Continued).

	Male			Female		
Variable	Starting (N=4923)	Not quitting (N=2789)	Level (N=1260)	Starting (N=4664)	Not quitting (N=1954)	Level (N=1053)
τ_2			0.572[a]			0.841[a]
			(0.039)			(0.047)
τ_3			0.854[a]			1.140[a]
			(0.044)			(0.051)
τ_4			1.778[a]			2.148[a]
			(0.055)			(0.065)
τ_5			2.259[a]			2.699[a]
			(0.061)			(0.078)
Log-likelihood[c]			-6715.02			-5870.96
AIC			2.756			2.548

Note: Standard errors in parentheses.
[a,b] Denote significance at the 5% and 10% levels respectively.
[c]We report log-likelihood values and AIC for the trivariate models.

Table 6. Mean marginal effects and average discrete changes

	Categorization 1		Categorization 2		
	Male	Female	Pooled	Male	Female
Mean Marginal Effects					
Income	-0.031	-0.033	-0.029[b]	-0.029	-0.035[b]
	(0.024)	(0.024)	(0.016)	(0.024)	(0.021)
Body mass	-0.247[a]	-0.211[a]	-0.197[a]	-0.228[a]	-0.179[a]
	(0.056)	(0.049)	(0.034)	(0.054)	(0.046)
Education	-0.254[a]	-0.244[b]	-0.251[a]	-0.319[a]	-0.199[a]
	(0.107)	(0.130)	(0.068)	(0.095)	(0.093)
Age	-0.110[a]	-0.052[a]	-0.076[a]	-0.101[a]	-0.047[a]
	(0.021)	(0.019)	(0.013)	(0.019)	(0.015)
Average discrete changes					
City	-1.267[a]	-0.744[a]	-0.946[a]	-1.113[a]	-0.690[a]
	(0.383)	(0.295)	(0.242)	(0.374)	(0.293)
Suburban	-0.862[a]	-0.620[a]	-0.811[a]	-0.841[a]	-0.634[a]
	(0.359)	(0.280)	(0.241)	(0.335)	(0.290)
Northeast	-0.524	-1.108[a]	-0.714[a]	-0.324	-0.954[a]
	(0.450)	(0.338)	(0.290)	(0.461)	(0.346)
Midwest	0.752[b]	0.596	0.659[a]	0.859[b]	0.675[b]
	(0.450)	(0.383)	(0.301)	(0.477)	(0.380)
South	1.007[a]	0.750[a]	0.913[a]	1.121[a]	0.777[a]
	(0.434)	(0.368)	(0.279)	(0.421)	(0.356)
Black	-0.971[a]	-1.407[a]	-1.215[a]	-1.060[a]	-1.485[a]
	(0.386)	(0.277)	(0.242)	(0.415)	(0.281)
Hispanic	-3.839[a]	-3.006[a]	-3.462[a]	-3.805[a]	-3.097[a]
	(0.401)	(0.312)	(0.242)	(0.386)	(0.287)
Healthy	-0.974[a]	-1.264[a]	-1.147[a]	-0.975[a]	-1.281[a]
	(0.453)	(0.376)	(0.296)	(0.430)	(0.371)
White-collar	-1.557[a]	-0.630[a]	-1.035[a]	-1.494[a]	-0.651[a]
	(0.359)	(0.297)	(0.227)	(0.355)	(0.293)
Cancer	-1.012[b]	0.492	-0.060	-0.950	0.474
	(0.543)	(0.490)	(0.406)	(0.598)	(0.496)
BP-heart	-0.239	-0.417	-0.330	-0.281	-0.298
	(0.370)	(0.300)	(0.239)	(0.360)	(0.302)
Alcohol	2.232[a]	2.177[a]	2.237[a]	2.248[a]	2.212[a]
	(0.307)	(0.258)	(0.191)	(0.283)	(0.259)
Intensive exercise	-0.377	-0.650[b]	-0.523[b]	-0.315	-0.653[b]
	(0.471)	(0.360)	(0.317)	(0.491)	(0.372)
No exercise	1.196[a]	1.091[a]	1.141[a]	1.261[a]	1.085[a]
	(0.552)	(0.370)	(0.335)	(0.516)	(0.376)
Diet	-2.229[a]	-0.307	-1.095[a]	-2.228[a]	-0.290
	(0.379)	(0.306)	(0.241)	(0.363)	(0.304)
Male			1.527[a]		
			(0.189)		

Note: Standard errors in parentheses.
[a,b] Denotes significance at the 5% and 10% levels respectively.

Figure 1. Histogram of cigarettes smoked (smokers only).

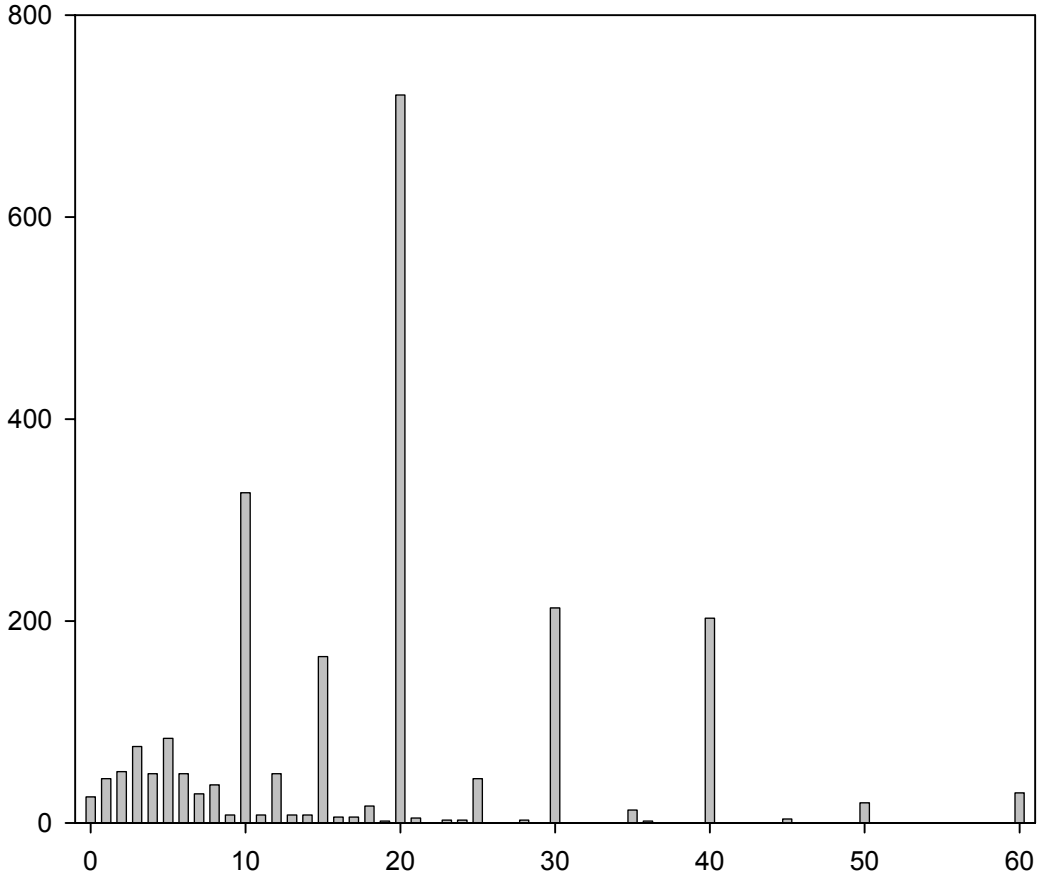

Figure 2. Histogram of the six cigarette categories (smokers only).