"A pack a day for twenty years": Smoking and cigarette pack sizes.

Lisa Farrell[a], Tim R.L. Fry[b] and Mark N. Harris[c]
[a]Department of Economics
University of Melbourne
[b]School of Economics and Finance
Royal Melbourne Institute of Technology.
[c]Department of Econometrics and Business Statistics
Monash University

July 13, 2004

Abstract

The extensive literature on smoking behaviour has focused on numerous aspects such as the factors that influence the decision to start smoking and, for smokers, what factors influence consumption and decisions to quit. This study focuses on the determinants of the typical daily volume of cigarette consumption. In particular the impact of cigarette pack sizes on the typical daily consumption of smokers is investigated. Results are presented from a statistical model which allows for 'pack-effects' in daily consumption levels.

Keywords: Smoking, count data, captivity, digit preferencing
JEL classification: C25, D12, I18.
Acknowledgements: We wish to acknowledge financial support from Monash University.

1 Introduction

It has long been acknowledged that there are significant health risks associated with cigarette consumption. Yet around 46 million adults in the US, and 12 million in the UK, smoke. Large amounts of public expenditures are directed towards health education programs aimed reducing participation rates. In the US smoking causes roughly 400,000 deaths per year at an estimated cost of more than $75 billion. The addictive nature of tobacco makes policies aimed at reducing cigarette consumption a particularly difficult process. Most governments have taken numerous measures to try and address the strain smoking puts on health services. As well as health promotion programs, many governments have imposed laws banning or highly restricting tobacco advertising.

The focus of this study is the relationship between consumption and cigarette pack sizes. Whilst much analysis has been undertaken into the factors that determine the decision to start and stop smoking, less has been done on the factors that determine the level of smoking for smokers. Market research has concentrated on the impact of elements of the marketing mix on consumption levels, but there appears to be no research on the impact of pack sizes on daily consumption patterns. This is important as, given the addictive nature of tobacco many smokers will try to regulate their consumption via the number of packets that they smoke in a given time period.

Whilst governments do closely regulate the sale of tobacco few have experimented with pack size regulation. Most governments dictate the smallest pack size that can be sold (and legislate against selling cigarettes individually) and it is the choice of the manufacturers with regard to what pack size varieties to offer (influenced, in part, by tradition and excise duty). The current legislation is designed mainly to prevent children from smoking. In England (were this study is focused) it is illegal to sell cigarettes individually and the smallest pack size allowed to be sold is 10 cigarettes. Purchasing cigarettes in packs raises the cost of smoking to a level outside of a child's income range. However, there is much debate regarding the effectiveness of this legislation. Studies have shown that the price elasticity for cigarettes by youths is more inelastic than that of older smokers - a result of brand loyalty by young smokers, Harris and Chan (1999). The tax system in England is indifferent to pack sizes. Taxes are levied as a tax per 1000 cigarettes sold, and an additional *ad valorem* tax per pack sold. These relatively high tax rates are designed to raise prices, so discouraging consumption. However, there is much debate in the literature regarding the relative merits/success of regulation versus taxation as means of reducing and preventing cigarette consumption, Wasserman, Manning, Newhouse, and Winkler (1991)

This paper contributes to this debate by hypothesizing that the variety of pack sizes available to consumers significantly impacts upon the number of cigarettes an individual smokes in a given period. Thus one policy to aid smokers reduce their consumption is to allow the sale of cigarettes in small packs (in contrast to the existing regulation). In considering these factors, this study provides a much clearer understanding of the determinants of levels of cigarette consumption.

2 Background

The literature on smoking is multidisciplinary in nature. The focus of this study is on the regulation of consumption by smokers according to pack sizes, thus evidence regarding the addictive nature of tobacco is presented. However, cigarette consumption is a natural two step process: participation and conditional consumption. The evidence on the factors influencing the decision to start smoking is considered first, then that on conditional consumption.

The literature concerned with the decision to participate in smoking has concentrated on the impact of family background and parental smoking behaviour (as well as standard demographics). A large body of literature has also looked at the decision to start smoking amongst teenage children - the time of life when smoking is most prone to start, DeCicca, Kenkel, and Mathios (2002). Indeed, the biggest growth in smokers in recent years has been amongst young females, Boreham and Shaw (2002). It has also been shown that socioeconomic status plays a key role in determining smoking participation. Issues relating to tax, economic welfare and social class patterns of smoking were investigated by Townsend (1987).

The literature on the intensity of smoking has largely concentrated on the addictive nature of tobacco. There is evidence from both the social and medical sciences, suggesting that tobacco is an addictive substance. Psychologists refer to cigarette consumption as part of a script, where a script is a set of inter-locking consumption patterns which have a re-enforcing quality. For many smokers the script involves tobacco and alcohol, but it might equally be the “slow” cigarette after a meal.

Pharmacologists, on the other hand, stress that the desire to smoke is real, as well as psychological. Nicotine is known to give smokers a ‘hit’ and is a stimulant. The lack of success for herbal cigarette replacements for those trying to quit, compared to solutions which break the smoking habit whilst slowing removing the body’s dependence on the nicotine, is a clear example of the strength of the addictive nature, Harris (1993).

Economists have generally measured addiction through the relative inelastic price elasticity of demand for cigarettes (see, for example, Young 1983, Godfrey 1986, Conniffe 1995, Harris and Chan 1999). Extensive work has been undertaken applying Becker and Murphy’s (1988) theory of rational addiction to smoking to explain addiction in terms of an individuals stock of addiction from past smoking behaviour (see Becker and Stigler 1977, Chaloupka 1991).

Given the addictive nature of cigarettes many smokers will try to regulate their consumption in order to “control” their addiction. Indeed, it is likely that smokers will do this by regulating their consumption by the number of packets of cigarettes that they smoke in a given period. The idea that pack size is an important determinant of consumption behaviour has been acknowledged in the marketing literature since the work of Ehrenberg and his colleagues in the 1960’s and 1970’s (see, for example, Ehrenberg 2000). The fact that cigarettes can not be bought individually, reinforces this. If smokers are constrained to purchase cigarettes in given quantities it is natural to expect their consumption to reflect these quantity constraints. This is consistent with the ideas in Ehrenberg (2000) who claims that, for a wide range of products, the number of packs bought on a given purchase occasion appear to be constant but the purchase interval (time between purchase occasions) tends to be the same, regardless of the size of the pack. Thus, if a consumer wishes to regulate consumption, they may do so by choice of pack size rather than by changing their frequency of purchase. This literature suggests that people have a standard shopping list and should they wish to change their consumption patterns, they adjust the quantities they purchase rather than the frequency of purchase. In essence there are menu costs. Dropping items on and off the shopping list increases the risk of accidentally not purchasing a needed item. Quantity reduction however, eradicates this risk. Thus smokers wishing to cut back their cigarette consumption will switch to purchasing smaller packs. Hence the variety of pack sizes available will impact on the level of consumption.

3 Theoretical Framework

The market for cigarettes is unusual in that, whilst there are multiple pack sizes, there are no quantity premiums or discounts. Thus traditional price discrimination arguements do not apply. Many product line models exist where a monopolist sets pack sizes and prices in order to sort consumers in the most profitable way (see, for example Maskin and Riley 1984, Gernster and Holthausen 1986). These papers show that the variety of pack sizes offered by a monopolistic producer reflects the number of types of individuals that exist. However, such models can not explain why cigarette manufacturers would choose to supply differential pack sizes but not simultaneously price discriminate. In the market for cigarettes there is product differentiation in terms of pack sizes but no price discrimination. Price premiums on large packs do not exist as smokers would simply purchase multiple smaller packs at a cheaper per unit cost. Further large quantity discounts would lead to conflict with legistaion as it would be seen to encourage smokers to purchase large quantities. Thus cigarettes are supplied at the same unit cost regardless of pack size though superficially it is unclear why pack size variety exisits. Naturally, producers will only supply a variety of sizes if this leads to greater profits. The only rational for this observed supplier behaviour is that pack size variation must impact on consumption levels and therefore increase profits. The model presented in this sections attempts to explain how pack size variation can lead to higher consumption levels and hence higher profits.

Assume two types of consumers: *light* smokers (*type 1*) and *heavy* smokers (*type 2*) such that $\alpha_1 + \alpha_2 = 1$, where α_t is the proportion of type t consumers in the market. Given that the unit price of cigarettes is the same regardless of the pack size this suggest that the two types of consumers have different demand curves. Standard economic theory tells us that more addicted consumers will have a more inelastic demand curve (relative to less addicted consumers). Thus, if consumption levels reflect levels of addiction to cigarettes, the demand curve for type 2 consumers (*heavy*) relative to type 1 consumers (*light*) will have a steeper gradient. This is illustrated in Figure 1

Whilst this shows that there are two types of consumers in the market, the question still remains: "given the identical unit price per cigarette (regardless of pack size), why does pack size variation exist?". The ability of producers to identify different demand curves in their market usually results in price

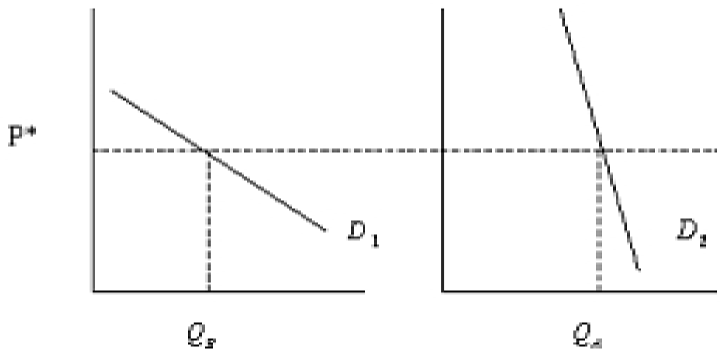

Figure 1: Demand Elasticities

discrimination, not solely product differentiation via pack sizes. Consider consumers with the following utility functions[1]

$$U = U(x_t, t) - C(Q_i, t). \tag{1}$$

Where t denotes the types of consumers in the market $t = \{1, 2\}$, i denotes the pack sizes available in the market, x_t is the consumption rate of type t consumers *i.e.* the number of cigarettes smoked per day, Q_i are the variety of pack sizes available for purchase. The associated cost function given by:

$$C(Q_i, t) = \frac{g_t x_t}{Q_i} + h_t(Q_i - x_t), \tag{2}$$

where g_t are the transactions costs of purchasing a packet of cigarettes and h_t are the storage costs of cigarettes. Transactions cost are essentially trips to the shop to purchase cigarettes and attention shall be restricted to the case where only one pack can be purchased per shopping trip. This is consistent with the work of Ehrenberg (2000) who provides evidence supporting the

[1] Assuming, separability in the budget process.

hypothesis that consumers like to maintain frequency of purchase. Storage costs are those associated with the excess quantity purchased in time period n but not consumed in this period and therefore stored, at some cost, until time period $n+1$.

Assuming that transactions costs and storage costs are independent of the type of consumer, and in the first instance, that storage costs are zero, the cost functions simplifes to,

$$C(Q_i, t) = \frac{g_t x_t}{Q_i}. \tag{3}$$

It can be seen that both types of consumers (*light* and *heavy* smokers) will maximise utility by purchasing in the largest pack available, hence miminising the transactions costs.

It is not obvious how the assumption of type-independant transaction costs ought to be relaxed. Whilst shopping costs involve the time taken to shop (and the opportuntity cost of this) which may well vary over the two types of consumers. However, there are also many other attributes that make up the total transaction costs; such as the distance from the shops of the consumer and the shopping skills of the consumer. It is not clear that these attributes are distributed nonrandomly across the types of consumers. Indeed, Kinsey (1981) suggests that those with the lowest opportunity cost of time may have the highest shopping skills. Thus the assumption of type-invariant transaction costs are retained throughout the rest of the analysis.

If transactions costs provide no explanation for pack size differentiation in the market for cigarettes, the story must be one relating to storage costs. Consider the case where storage costs are positive ($h_t > 0$), but are type-invariant *i.e.,* $h_1 = h_2 = h$. That is, there is a given unit cost per cigarette at which they can be stored until the next period. For a given pack size Q_A with $Q_A - x_2 \geq 0$ then,

$$h(Q_A - x_1) > h(Q_A - x_2). \tag{4}$$

Showing that type 1 consumers (*light* smokers) do better to purchase smaller pack sizes than type 2 consumers (*heavy* smokers). Hence consumers will self-select themselves into large and small pack sizes such that $Q_A > Q_B$. Given that x_t is the type's level of consumption per period, it must be the case that $Q_A - x_2 \geq 0$ and $Q_B - x_1 \geq 0$. Since storage costs are a function of the residual number of cigarettes at the end of the period, consumers can minimise these costs by minimising the number of unsmoked cigarettes in a given period. Thus consumers will maximise utility according to equation (1) by consuming up to the point where the marginal utlity from consuming cigarettes is equal to the storage costs.

$$\frac{dU(x_t, t)}{dx_t} = h \tag{5}$$

Allowing h_t to be type-dependent seems an obvious expansion. It is logical to think of a type's ability to store cigarettes as being a function of their level of addiction. Hence rather than being a monetary value, the cost of storing cigarettes is the acossiated disutility of having purchased them but refraining from smoking them. Given the two types of consumers considered here, it is clear that the *heavy* consumers are more addicted than *light* consumers and thus levels of consumption act as a proxy for levels of addiction.[2]

[2] This is consistent with the model of rational addiction (Becker and Murphy 1988),

Assuming $h_1 < h_2$ implies for more addicted consumers the utility from the consumption of the marginal cigarette is larger and the storage costs associated with storage are also larger than for less addicted smokers. For heavily addicted consumers it will be optimal to remove the disutility associated with the storage of cigarettes by consuming the entire pack in the current period. Hence a profit maximising supplier will do best by providing two pack sizes corresponding to the points where expression (6) is satisfied. In this way consumers will smoke the entire pack in each period and purchase one pack of their chosen size each period. Hence a profit maximising firm, in the market of an addictive good, will benefit by supplying a number of packs equal to the number of identifiable types of conumers in the market. Pack size variation can lead to an increase in consumption, and hence increased profits, without the need to price discriminate across pack sizes.

$$\frac{dU(x_1, 1)}{dx_1} = h_1 \text{ and } \frac{dU(x_2, 2)}{dx_2} = h_2 \tag{6}$$

In summary, the model presented here suggests that in the market for an addictive good, pack size variety can lead to increased consumption levels by consumers, and the number of packs offered will equal the number of types of consumers in the market. This suggests that when looking at the consumption patterns of smokers, one would expect to see clusterings around the available pack sizes. The rest of this paper investigates if such outcomes are observed in smokers consumption patterns, and if so, how to model the process.

where past consumption represent a stock of addiction that impacts on todays consumption.

4 The Data

The data is drawn from the Health Education Monitoring Survey (HEMS), conducted by the Office of National Statistics in the United Kingdom. The HEMS is an individual based nationally representative annual survey, that covers the adult population aged 16-74 living in England. The main purpose of the Health Education Monitoring Survey (HEMS) is to provide information on respondents' knowledge of health-related issues. In particular, for the purposes of this study, information is collected on individual characteristics, socio-demographic characteristics, alcohol consumption, physical activity and, finally, the number of cigarettes typically consumed per day. Data from the 1998 survey is used. A major innovation in this survey year was to investigate the effect of major life-events on health-related behaviour and health.

This data is ideal for the purposes of this research as it contains information on the typical daily (weekday) consumption of the number of cigarettes. The distribution of this consumption can be seen in Figure 1. It is immediately obvious that there is a large concentration of individuals that do not smoke on a typical week day. Indeed, some 74% of our sample records zero cigarette consumption. Some of these observations will also correspond to social smokers *i.e.*, those who may not smoke during the week, but whom might occassionally smoke on a weekend. It will be important in the statistical analysis allow for the fact that there are two types of individuals in our data set who may record a zero count.

The observations are clearly not evenly distributed across all the possible

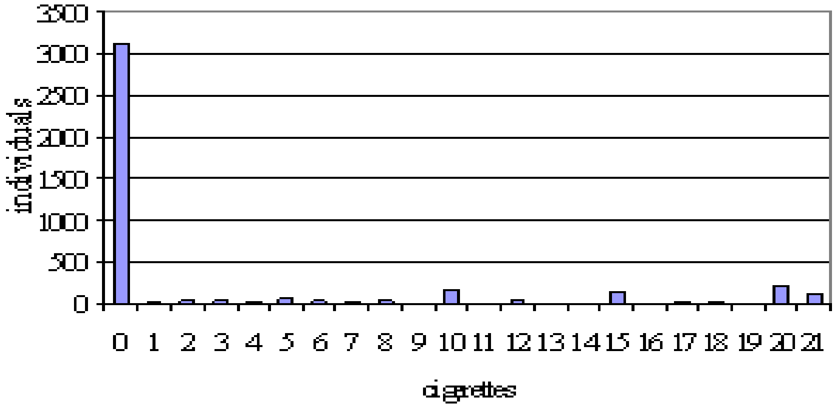

Figure 2: Unconditional Distribution of Typical Week Day Cigarette Consumption

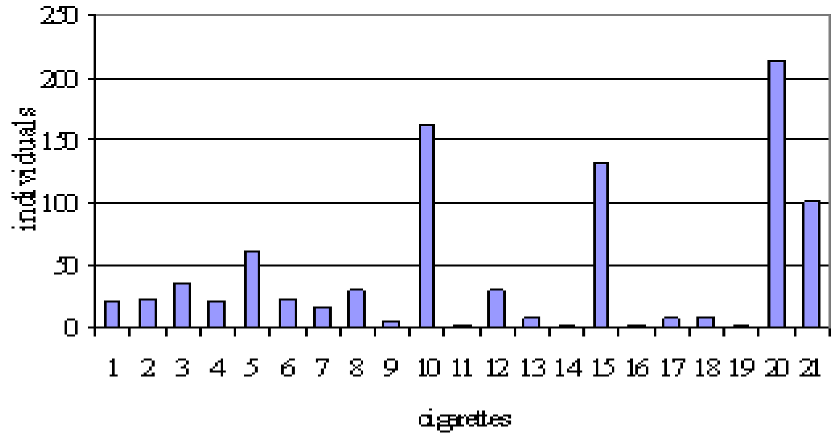

Figure 3: Conditional Distribution of Typical Week Day Cigarette Consumption

counts, there are spikes in the distribution (for those who do smoke) at 5, 10, 15, and 20 cigarettes per week day (Figure 3). Importantly the counts of 10 and 20 correspond to the packet sizes that cigarettes are sold in (in England). Moreover the remaining spikes correspond to half a pack and one and a half packs per day. This suggests that individuals are indeed regulating their cigarette consumption by counting the numbers of packets they smoke in a given time period, rather than counting the number of cigarettes directly. For example, everyone has heard of the triumphant smoker who is trying to quit, proudly announcing that "he has cut back from one pack a day to half a pack a day". Such pack counting is consistent with the observed pattern of cigarette consumption shown in the raw data.

Tables 1 and 2 provide some simple sample descriptive statistics of the raw data (the variable descriptions and motivations for their inclusion in the model will be discussed in Section 6). Table 1 presents sample means for the variables used in this analysis. The sample consists of 5,766 individuals, of whom 1,491 smoke. The average consumption is 3.48 cigarettes per weekday (13.46 cigarettes per weekday, conditional upon participation).

Table 2 shows the conditional means for smokers and nonsmokers of the variables which one might expect to be determinants for the participation decision. It is clear that there are significant differences in the demographics of smokers and nonsmokers. Importantly the conditional means are statistically significantly different between smokers and nonsmokers for all of the socioeconomic status variables. Section 7 presents the results of a more rigorous multivariate analysis.

Table 1: Sample Descriptive Statistics

Variable	Sample means	s.d.
Number of observations	5766	
Number of cigarettes	3.481	(6.72)
Standard Demographics		
age $\div 10$	4.939	(1.90)
Male $\times 1$	0.441	(0.50)
No qualifications $\times 1$	0.294	(0.46)
O-level $\times 1$	0.377	(0.48)
A-level $\times 1$	0.329	(0.47)
Single $\times 1$	0.429	(0.50)
Nonwhite $\times 1$	0.046	(0.21)
Money and stress/social capital		
Income 2,500-5,000 $\times 1$	0.194	(0.40)
Income 5,000-10,000 $\times 1$	0.203	(0.40)
Income 10,000-15,000 $\times 1$	0.128	(0.33)
Income 15,000-20,000 $\times 1$	0.115	(0.32)
Income 15,000-20,000 $\times 1$	0.161	(0.37)
Income 20,000-30,000 $\times 1$	0.198	(0.40)
Employed $\times 1$	0.580	(0.49)
Major life events		
Stress levels	2.840	(0.49)
Sad $\times 1$	0.095	(0.29)
Happy $\times 1$	0.089	(0.29)
Life style factors		
Physical activity $\times 1$	0.553	(0.50)
Heavy drinking $\times 1$	0.221	(0.41)
activity and drink interaction $\times 1$	0.144	(0.35)
Long term illness $\times 1$	0.413	(0.49)
Presence of children $\times 1$	0.293	(0.46)
Socioeconomic *status*		
Deprivation index	-0.278	(2.76)
Renting $\times 1$	0.300	(0.46)
South $\times 1$	0.495	(0.50)
North $\times 1$	0.505	(0.50)

Note: $\times 1$ indicates dummy variables.

Table 2: Sample Conditional Descriptive Statistics

Variable	Sample means	s.d.	Non-smokers	s.d.	Smokers	s.d.
Number of observations	5766		4275		1491	
Number of cigarettes	3.481	(6.72)	0.000	(0.00)	13.463	(6.35)*
Standard Demographics						
age/10	4.939	(1.90)	5.154	(1.92)	4.323	(1.70)*
Male $\times 1$	0.441	(0.50)	0.435	(0.50)	0.458	(0.50)
No qualifications $\times 1$	0.294	(0.46)	0.278	(0.45)	0.340	(0.47)*
O-level $\times 1$	0.377	(0.48)	0.368	(0.48)	0.402	(0.49)*
A-level $\times 1$	0.329	(0.47)	0.353	(0.48)	0.258	(0.44)*
Single $\times 1$	0.429	(0.50)	0.404	(0.49)	0.502	(0.50)*
Nonwhite $\times 1$	0.046	(0.21)	0.051	(0.22)	0.034	(0.18)*
Life style factors						
Long term illness $\times 1$	0.413	(0.49)	0.408	(0.49)	0.429	(0.50)
Presence of children $\times 1$	0.293	(0.46)	0.272	(0.45)	0.354	(0.48)*
Socioeconomic *status*						
Deprivation index	-0.278	(2.76)	-0.507	(2.68)	0.378	(2.86)*
Renting $\times 1$	0.300	(0.46)	0.247	(0.43)	0.450	(0.50)*
South $\times 1$	0.495	(0.50)	0.505	(0.50)	0.466	(0.50)*
North $\times 1$	0.505	(0.50)	0.4952	(0.50)	0.534	(0.50)*

Note: $\times 1$ indicates dummy variables, * Indicates a statistically significant difference at the 5% level between the means for smokers relative to non-smokers

5 The Econometric Model

The data to be modelled is that of an integer count. This would suggest that an appropriate statistical model might be the Poisson regression model given by

$$\Pr\left(Y = y\left|\mathbf{x}\right.\right) = \frac{\exp\left(-\lambda\right)\lambda^{y}}{y!},\ y = 0, 1, \ldots \tag{7}$$

with $\lambda = \exp\left(\mathbf{x}'\boldsymbol{\beta}\right)$.

The Poisson regression model is, in many respects, a restrictive specification. In particular, it is not able to cope with an abundance of observed zeros in the data to be modelled. The data here has 74% of the observations recorded as zero. This preponderance of zeros is inconsistent with the Poisson regression specification and tends to lead to underestimation of the expected value of the count process and potentially spurious evidence of over-dispersion.

To overcome the problems caused by the abundance of zeros often encounter in empirical counts several researchers (see, for example, Mullahey 1986, Heilbron 1989, Lambert 1992, Greene 1994, Pohlmeier and Ulrich 1995, Mullahey 1997) have suggested models that “inflate” or “augment” the Poisson process. These models account for the zeros that are additional to those expected from a simple Poisson process. The models are variously termed Zero Inflated, or Zero Augmented, Poisson (ZIP or ZAP) models. In essence what these models do is to add an additional binary statistical process to the assumed data generating process. This binary model allows for zero observations to occur separately from the Poisson count process ones.

In these models the zero observation may be a drawing from a choice set that comprises the single outcome zero, or zero from a Poisson process. The ZIP model therefore implies that the probability of the occurrence of a zero observation has two components; one from the binary process and the other from the Poisson process. This inflates the probability of a zero from that obtained from a simple Poisson process. Defining an indicator function as

$$d = \begin{cases} 0 \textit{if a zero count is observed} \\ 1 \textit{if a positive count is observed,} \end{cases}$$

the simplest form of the ZIP model is given, in two parts, by

$$\Pr\left(Y = 0 \left| \mathbf{x} \right.\right) = \Pr\left(d = 0\right) + \left[\Pr\left(d = 1\right) \times \exp\left(-\lambda\right)\right], \tag{8}$$

for the observed zero counts, and

$$\Pr\left(Y = y \left| \mathbf{x} \right.\right) = \Pr\left(d = 1\right) \times \left[\frac{\exp\left(-\lambda\right)\lambda^{y}}{y!}\right], \tag{9}$$

for $y = 1, 2, \ldots$. If the simple logistic splitting function is used, the probabilities for $d = 0$ and $d = 1$ are

$$\Pr\left(d = 0\right) = \frac{\exp\left(\theta\right)}{1 + \exp\left(\theta\right)} \tag{10}$$

and

$$\Pr\left(d = 1\right) = \frac{1}{1 + \exp\left(\theta\right)}. \tag{11}$$

There is agreement in the literature that this binary process is an oversimplification and should be specified as a binary Logit (or Probit) that allows the probability of a zero to depend upon observed characteristics. That is, the probability of observing $d = 1$ is now conditional on a set of observed characteristics $\mathbf{z}$, with unknown weights $\boldsymbol{\gamma}$, and is given by $F\left(\mathbf{z}, \boldsymbol{\gamma}\right)$

where, following the previous reasoning, F can be the logistic cumulative distribution function. This will be termed here a parameterised ZIP (ZAP) model.

This specification is particularly attractive as it allows non-participation (zero cigarettes) to depend upon individual characteristics. Further, a zero observation can be generated either from a non-participant, or from a participant who records a zero for typical week day consumption. The latter, for instance, may be an infrequent smoker who on a typical work day does not smoke, but who may smoke on an atypical workday. Alternatively, he/she may smoke at weekends only.

Whilst the parameterised ZIP model can allow for the abundance of zero observations in the data to be modelled it is unable to deal with the potential extra spikes in the observed data at values such as 5, 10, 15, 20 and 21+. It was argued earlier (Section 4) that individuals might choose to regulate their consumption through multiples of pack sizes. Additionally, a by-product of the question asked of survey respondents is that they might mentally round their responses to multiples of five (digit preference). Thus individuals are drawn to certain outcomes and there is an "attraction" to multiples of pack sizes.

To overcome this phenomena it is noted that the idea behind the ZIP formulation is the same as that in Manski (1977), where in the discrete choice context it is argued that there are two distinct probabilistic processes at work; a "choice set generation" and a "choice decision". Individuals are faced with a choice set generated from one probabilistic process and choose their preferred outcome from that set. In the ZIP models individuals choose to be participants or not, and if they participate their consumption follows a Poisson process.

An obvious extension of the parameterised ZIP therefore, is to extend the first component to a multinomial process, where individuals might be non-participants, be drawn to multiples of pack sizes or participate and choose "freely" according to a Poisson process. The outcomes in the choice set generation process are: non-participation; "attraction" to: 5, 10, 15 and 20 cigarettes; heavy consumption and participation with "free" choice. This can be represented by a multinomial indicator

$$d = \begin{cases} 0 \textit{if non-participant} \\ 1 \textit{if "attraction" to 5} \\ \vdots \\ 6 \textit{if participant is "free"} \end{cases}$$

As in the ZIP model the probabilities of the observed outcomes are a mixture of those from the multinomial process and the Poisson process. These are given by

$$\begin{aligned}
\Pr(Y = 0 \,|\mathbf{x}) &= \Pr(d = 0) + [\Pr(d = 6) \times \exp(-\lambda)] \\
\Pr(Y = 5 \,|\mathbf{x}) &= \Pr(d = 1) + \left[\Pr(d = 6) \times \frac{\exp(-\lambda)\,\lambda^5}{5!}\right] \\
\Pr(Y = 10 \,|\mathbf{x}) &= \Pr(d = 2) + \left[\Pr(d = 6) \times \frac{\exp(-\lambda)\,\lambda^{10}}{10!}\right] \\
&\vdots \\
\Pr(Y = 21 \,|\mathbf{x}) &= \Pr(d = 5) + \left[\Pr(d = 6) \times \frac{\exp(-\lambda)\,\lambda^{21}}{21!}\right] \\
\Pr(Y = y \,|\mathbf{x}) &= \Pr(d = 6) \times \frac{\exp(-\lambda)\,\lambda^y}{y!}, \;\; y \neq 0, 5, 10, 15, 20, 21+.
\end{aligned}$$

To complete the specification it is necessary to define the multinomial (choice set) probabilities. The arguments above suggest that the non-participation outcome is likely to depend upon certain individual characteristics ($\mathbf{z}$). However, there is no clear rationale for attraction to outcomes (5, 10, 15, 20 and 21+) to depend upon such characteristics. They might be the result of either consumption regulation or digit preferences.[3] Thus the full multinomial model is parameterised as:

$$\begin{aligned}
\Pr(d=0\,|\mathbf{z}) &= \frac{\exp(\mathbf{z}'\boldsymbol{\alpha})}{1+(\exp(\mathbf{z}'\boldsymbol{\alpha})+\theta_5+\theta_{10}+\theta_{15}+\theta_{20}+\theta_{21})} \\
\Pr(d=1\,|\mathbf{z}) &= \frac{\theta_5}{1+(\exp(\mathbf{z}'\boldsymbol{\alpha})+\theta_5+\theta_{10}+\theta_{15}+\theta_{20}+\theta_{21})} \\
\Pr(d=2\,|\mathbf{z}) &= \frac{\theta_{10}}{1+(\exp(\mathbf{z}'\boldsymbol{\alpha})+\theta_5+\theta_{10}+\theta_{15}+\theta_{20}+\theta_{21})} \\
&\vdots \\
\Pr(d=6\,|\mathbf{z}) &= \frac{1}{1+(\exp(\mathbf{z}'\boldsymbol{\alpha})+\theta_5+\theta_{10}+\theta_{15}+\theta_{20}+\theta_{21})}.
\end{aligned} \tag{12}$$

Once the full set of probabilities has been specified, and given an *iid* sample from the population ($i = 1, \ldots, N$), the parameters of the model ($\boldsymbol{\phi}$) can be consistently and efficiently estimated using the maximum likelihood criteria. The log-likelihood function is

$$\ell(\boldsymbol{\phi}) = \sum_{j=1}^{J}\sum_{i=1}^{N} h_{ij} \ln\left[\Pr(Y_i = y_i\,|\mathbf{x},\mathbf{z})\right], \tag{13}$$

where the indicator function h_{ij} is

$$h_{ij} = \begin{cases} 1 \text{ if individual } i \text{ chooses count } j \\ 0 \text{ otherwise.} \end{cases} \tag{14}$$

[3] One might hypothesis, given the literature on frequencies of purchase, that "shopping habits" would be the type of variables that might be employed to characterise the higher thetas, however the dataset affords us no such variables with which to test this hypothsis.

6 Instruments used for the Participation and Levels Equations

From the discussion above, it is clear that two distinct sets of covariates are required; $\mathbf{x}$ which determine the amount (level) of cigarette consumption, and "participation" variables, $\mathbf{z}$. The latter help identify non-smokers from infrequent smokers within those observations recorded as a zero count. The HEMS is particularly useful in this respect, as it allows for identification of numerous proxies for both participation and, conditional on participation, the number of cigarettes consumed.

6.1 Participation

It is possible to conceive of three broad groups of variables that are likely to affect the probability of participation: standard demographics; socioeconomic factors; and lifestyle indicators. These are dealt with in turn below.

6.1.1 Standard Demographics

The instruments included here are: age, ethnicity (non-white) and education, all of which should decrease the extent of attractiveness to non-participation. Evidence suggests that most smokers start smoking at a relatively early age (teenage years), older non-smokers are less likely to start smoking and older smokers are more likely to (at least to attempt to) quit perhaps as a result of health related issues. The more educated, being more aware of the health risks involved, will be less likely to participate. Although there is debate in the literature whether the negative relationship between education and smoking picks up better health knowledge and understanding of the associated risks or if it proxies lower discount rates of the better educated, Sander (1995).

On the other hand, gender (male) may decrease the attraction to non-participation, as generally the incidence of smoking is higher for males, Hersch (1996). There is also evidence that minority groups generally have higher participation rates (perhaps due to cultural reasons). Hersch (1996) finds that whites choose safer products to consume, where smoking is defined as a risky consumption decision. Marital status is also included, although its effect is somewhat ambiguous *a priori* (there may be contagion effects if both partners smoke, or both do not, for example). Jones (1994) finds evidence that the presence of other smokers in the household has a significant influence on the decision to start and stop smoking. The data does not allow us to observe this interaction in such detail, but controls for the presence of a partner are included.

6.1.2 Socioeconomic Status Indicators

The data contains various instruments for socioeconomic status. Smoking participation rates tend to be much higher amongst the lower social class classifications. The proxies used for this are: income; a "deprivation index" (running from affluent to deprived); residence in rented or socially provided accommodation and region of residence (North and South).

6.1.3 Life-Style Factors

This dataset is also appropriate for this study in that it contains various lifestyle indicators. The respondent is asked whether they suffer from any long-standing illness, disability or infirmity. The general health implications of this, should ensure that it is negatively associated with smoking participation.[4] Evidence shows that those who suffer worsened health whilst smoking reduce their consumption and are more likely to quit, (Jones 2002). The presence of children is also expected to exert a negative effect on participation, for similar reasons. There have been numerous public awareness campaigns regarding the dangers of passive smoking especially on the health of young children. Blaylock and Blisard (1992) report that as the number of children in the household increases, the probability of smoking participation decreases.

6.2 The Level of Consumption

Factors relating to the level of cigarette consumption can be grouped under four main headings: standard demographics; major life events; money and stress/social capital; and lifestyle factors.

6.2.1 Standard Demographics

Evidence suggests that men smoke more than women, so gender is also included here. The likelihood that the age-consumption profile will be "n-shaped" is allowed for by including a quadratic in age. Wasserman, Manning, Newhouse, and Winkler (1991) reports that the married smoke more, suggesting household contagion effects and also that, conditional on smoking, whites smoke more than non-whites.

[4] Whilst the survey does contain questions relating to the respondents current health status this variable is not included in our model as it is clearly endogenous to current smoking behaviour. However, a question relating to long term illness is included, as this will affect smoking behaviour but is also less endogenous to current cigarette consumption

6.2.2 Major Life Events

The data is unique in containing information regarding any major life events that the individuals may have experienced in the 12 months prior to the survey. These life events have been included in our analysis through the inclusion of two variables; "happy" and "sad". Happy records if the individuals state they have experienced a happy life event (such as marriage or a birth in the family). Sad records if the individual has experienced a negative life event such a the death of a close relative or friend, or personal experience of theft, a mugging or other crime. One would expect that such events are unlikely to cause an individual to start smoking. However, for those individuals who do smoke, *a priori* one would expect negative life events to be associated with higher consumption, whereas positive life events will be associated with lower consumption. There is also a self-reported measure of the respondent's overall level of stress. For smokers, increased levels of stress are likely to be associated with higher consumption levels, although this is unlikely to affect the participation decision. The assumption here is that for non-smokers a major life event is unlikely to cause them to start smoking, but for those who do smoke such life events are likely to impact on the level of consumption.

6.2.3 Money and Stress/Social Capital

Income is included for standard demand schedule reasons. Blaylock and Blisard (1992) shows income has no effect on participation but is significant in determining consumption. A dummy variable indicating whether the individual is employed is included to account for any positive effects afforded by workplace stress, peer group effects, social capital and workplace patterns.

The observed count of cigarettes smoked relates to weekday consumption and therefore it is necessary to control for whether or not a person is employed and likely to be in a workplace environment on a typical weekday. Given that this may impact on their freedom to smoke, or indeed lead to peer pressure to smoke more, or generally increase an individuals stress levels.

6.2.4 Lifestyle Factors

The data contains an "activity" variable, defined as whether the individual has undertaken any physical activity in the last month. This is a very broad measure (55% of the sample had done so). This indicator is likely to be negatively associated with the level of cigarette consumption. Given the complementarily of smoking and drinking, a "heavy drinker" indicator is also included in the Poisson process. The indicator is defined as consumption of more than the recommended daily intake of alcohol.

Given the very broad nature in which these two variables have been defined, it is unlikely that they will be able to help identify participation effects. However, they may be significant indicators of heavy cigarette consumption. It is unlikely that those who smoke heavily will participate in exercise and it is likely that given the complementarily between smoking and alcohol consumption, heavy smokers may be more prone to drink higher than the medically recommended intake. An interaction term of these two variables is also included. This variable is designed to pick up the potential bias that may occur in the coefficients for activity and alcohol consumption caused by the presence in our data of team sports players. These will engage in physical activity and also consume alcohol *i.e.* social sports players. It is an interesting line of enquiry to investigate if these social sports players also smoke.

There is significant evidence from the literature investigating health production functions that these behaviours are correlated. Individuals are seen to develope healthy or unhealthy life styles and therefore observation of their life style choices for alcohol consumption and exercise activity is likely to be a good indicator of their smoking related decisions, see Kenkel (1995).

Finally, the long term illness indicator is also included for the Poisson process for the reasons as outlined above. Note that all of the negative effects (in both processes) are likely to be smaller in absolute terms than the positive ones, due to the addictive nature of tobacco. It is easier to start smoking than to stop, and it is easier to increase consumption rather than to reduce it (all other things equal).

7 Results

The full set of results are reported in Table 3. The first-hurdle, or participation decision, appears to be extremely well-modelled given the strong significance of all of the parameters. Indeed, all of the parameters, bar the constant and the dummy variable for A-level qualifications, are significant at 5 percent size. Of the standard demographics, both males and non-whites are more likely to be participants (recalling that a negative coefficient here reduces θ_0). The more educated individuals are less likely to smoke, as are singles. As individuals age, smoking participation rates decrease, as older individuals are likely to become more health aware. With regard to the participation equation and socioeconomic status, smoking participation clearly increases with reduced social standing and poorer living environments. Finally, the lifestyle factors (of serious long term illness and the presence of any children in the household), significantly, as expected, increase the probability of non-participation.

With regard to the conditional count process, the age-consumption profile is distinctly “n-shaped”. Males and the lower educated smoke more, conditional on participation, whilst non-whites smoke less. Regarding marital status the model shows that single people smoke less. One feasible explanation is that household contagion exists if one partner smokes then the other partner will find it hard to quit. For smokers, stress increases the typical daily consumption, as does the occurrence of sad events. On the other hand, happy events decrease typical daily consumption (although as predicted smokers respond more strongly to factors that increase consumption as opposed to reducing consumption). However the statistical significance of this group of variables is surprisingly low. It seems stress related factors impact little on levels of consumption.

The consumption-income profile is somewhat erratic, but it can be seen that the highest two income categories smoke less than the lowest income group. Employment does not appear to significantly affect cigarette consumption. In terms of conditional consumption and life-style factors, the complementarity of drinking and smoking is clearly evident, maybe as a result of social scripts. Those individuals who participate in any kind of (broadly defined) physical activity, tend to have lower cigarette consumption. The measure of physical activity is too broadly defined to identify participation in smoking consumption, however it is a significant predictor of the level of consumption. In essence it works well to identify those heavy smokers with

Table 3: Estimation Results

		Coefficient	S.E.
Structural Parameters	Constant	1.285	(0.19)**
Standard Demographics	age/10	0.347	(0.08)**
	$\text{age}^2/1,000$	-0.346	(0.09)**
	Male $\times 1$	0.106	(0.05)**
	O-level $\times 1$	0.214	(0.06)**
	A-level $\times 1$	-0.001	(0.07)
	Single $\times 1$	-0.091	(0.05)*
	Non-white $\times 1$	-0.091	(0.05)**
Major Life Events	Stress levels	0.005	(0.03)
	Sad $\times 1$	0.132	(0.10)
	Happy $\times 1$	-0.098	(0.10)
Money and Employment	Income 2,500-5,000 $\times 1$	0.200	(0.06)**
	Income 5,000-10,000 $\times 1$	-0.186	(0.08)**
	Income 10,000-15,000 $\times 1$	0.022	(0.08)
	Income 15,000-20,000 $\times 1$	-0.376	(0.09)**
	Income 20,000-30,000 $\times 1$	-0.455	(0.09)**
	Employed 30,000+ $\times 1$	0.068	(0.06)
Life-Style Factors	Physical activity $\times 1$	-0.220	(0.06)**
	Heavy drinking $\times 1$	0.298	(0.09)**
	Activity $\times$ Heavy drink	-0.145	(0.12)
	Long-term illness $\times 1$	-0.065	(0.05)
First Hurdle Parameters	Constant	-0.065	(0.06)
Standard Demographics	Male $\times 1$	-0.118	(0.01)**
	Ln(age)	0.555	(0.03)**
	Single $\times 1$	0.411	(0.07)**
	O-level $\times 1$	0.779	(0.08)**
	A-level $\times 1$	-0.081	(0.07)
	Non-white $\times 1$	-0.531	(0.07)**
Socioeconomic *Status*	Deprivation index	-0.291	(0.07)**
	Renting $\times 1$	-0.505	(0.07)**
	South $\times 1$	-1.819	(0.13)**
Life-Style Factors	Long term illness $\times 1$	0.533	(0.02)**
	Children $\times 1$	1.408	(0.21)**
Theta Coefficients	θ_5	-1.674	(0.03)**
	θ_{10}	-0.773	(0.08)**
	θ_{15}	-0.867	(0.03)**
	θ_{20}	-0.952	(0.08)**
	θ_{21}	-1.075	(0.03)**
	Log likelihood	-6910	
	Number of observations	5,766	

Standard errors in parentheses. ** and * significant at 5 and 10% level, respectively (two-sided).

particularly unhealthy life-styles.

Finally, the pack-size effects are all strongly significant.[5] The largest effect is afforded by 10, then 15, 20, 21 and 5. As noted above, the count of 10 corresponds to one of the pack-sizes available in England. Many smokers are clearly regulating their consumption to a "small pack a day". Alternatively, this corresponds to a standard pack (of 20), every other day.

8 Model Evaluation and Predictions

In Figure 4 observed sample proportions are plotted alongside probabilities (evaluated at covariate sample means) of: our model (the Parameterised Dogit Poisson); ZAP and Poisson models. In essence this is the discrete choice analogue of the actual versus predicted plot of more traditional linear models. Ignoring the concentration of observations at zero, clearly leads to misspecified models and erroneous inference as is evidenced by the poor results of the standard count data model (Poisson).

Augmenting the standard count data model to account for the preponderance of zeros improves performance somewhat. However, in effectively ignoring the zeros in the count process the conditional mean of the Poisson component of the ZAP model is biased towards the larger counts. Thus this model is unduly influenced by the spikes in the distribution at larger counts (those corresponding to pack sizes and digit preferenced outcomes).

Finally, the Parameterised Dogit Poison model clearly out performs the other statistical models. The participation/non-participation choice is mod-

[5] Experimentation with a specification that included a theta for all the possible counts was conducted but this specification was found to effectively "over fit" the data.

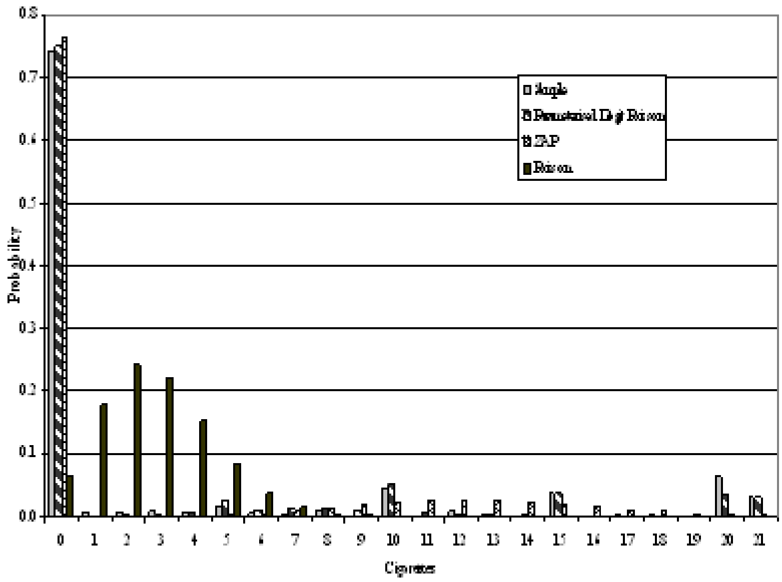

Figure 4: Sample Proportions and Predicted Counts for Different Models

elled well and almost exactly replicates the observed sample proportion. Moreover, the model clearly picks up the increased probabilities at the observed spikes in the distribution. Specifically, the model proposed here excels in its ability to replicate the probability mass of observations at pack sizes and provides a much clearer understanding of the drivers of typical cigarette consumption.

Using the estimated parameters it is interesting to ask what level of cigarette consumption the model predicts for a number of individuals with differing characteristics. In particular, a few stylized sets of characteristics will be analyzed. In the first instance a typical nonsmoker is considered, secondly an individual with the characteristics of a heavy smoker is profiled. These are illustrated in Figure 5.

The model predicts the probability of smoking zero cigarettes a day for a

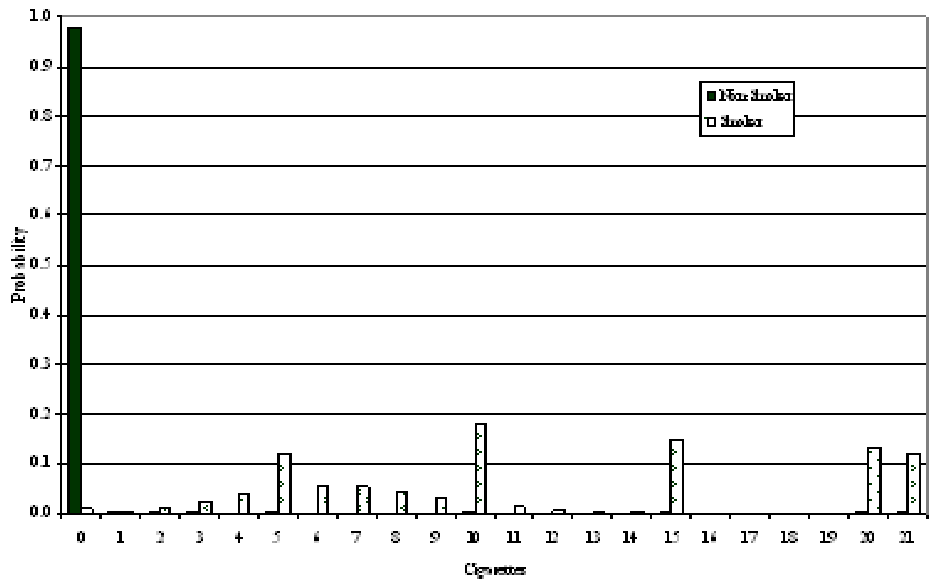

Figure 5: Predicted Probabilities for Sylised Individuals

nonsmoker is, not surprisingly, very high. For a typical smoker there is effectively zero probability of non-participation. Maximum probability occurs at 15 cigarettes closely followed by counts of 10, 20 and 21 plus cigarettes per week day. There is also significant mass at the larger levels of consumption.

In addition to estimating the predicted probabilities of the various count outcomes the model can also be used to estimate expected values of counts for stylized individuals. For an average person in the data set (*i.e.,* setting all variables to sample means) the expected number of cigarettes smoker per weekday is 3. This corresponds to the number of meals in the day and is consistent with the literature on scripts. For an individual with the characteristics of a nonsmoker this expectation falls to zero cigarettes per day. For a typical smoker the expected number of cigarettes smoked per day is 10. This is the quantity contained in the smallest pack size available

Table 4: Expected Values for Differing Characteristics

	Dummy Variable Value	
	0	1
Expected Number of Cigarettes Smoked		
Presence of children	4	1
Long-term illness	4	2
Rented accommodation	3	4
Non-white	3	4
South	1	6

to smokers in England. The model predicts an expected value for smokers consistent with the hypothesis of pack size consumption regulation.

A selection of expected values are presented in Table 4. The presence of children in the household and having a long term illness, both have large effects in terms of reducing the expected number of cigarettes to be consumed. Alternatively, living in rented accommodation, being in the South of England and being non-white all increase the expected number of cigarettes consumed, holding all other variables at sample means.

9 Conclusions

The paper presents a new statistical model which combines the traditional Poison model for count data, with the Dogit model for discrete data. This allows us to model a count data process where there are a number of spikes in the distribution at particular counts. In the context of the application presented here the focus is on the importance of pack sizes on cigarette consumption. Given the addictive nature of tobacco, in an attempt to control their addiction, many smokers may monitor, or regulate, their consumption. Evidence is presented here that smokers regulate their consumption in accordance with the size of packets that are available. That is, daily consumption patterns reflect the available pack sizes. Such discreteness in the quantities in which cigarettes can be purchased correspond to spikes in daily consumption of cigarettes. It is perhaps not surprising to find that smokers count their consumption in terms of the number of packs smoked in a given period, since they can only purchase cigarettes in packs. This phenomenon is confirmed by the estimation results, which suggest that the (expected) number of cigarettes smoked by a typical smoker is 10 per day - equivalent to the amount contained in the smallest packet that consumers can purchase in England.

The results presented here suggest that governments wishing to reduce cigarette consumption can do so by allowing cigarettes to be sold in smaller packets, and by increasing the variety of pack sizes available to consumers. Alternatively, cigarette manufacturers could potentially increase sales by restricting the sale of cigarettes to large pack sizes. However, these are only tentative policy recommendations to test this hypothesis further we would need to observe individuals consumptions patterns over a period when the pack sizes available to consumers changed. Unfortunately, we know of no data that corresponds to such a natural experiment.

In terms of the remaining results, they confirm those of the existing literature in terms of the impact of the covariates included. The participation decision is dominated by: gender; age; marital status; education; ethnicity; the presence of children in the household; and health issues. There are also strong socioeconomic status effects that influence the decision to smoke. Conditional upon being a smoker the model shows significant impacts for: gender; age; marital status; education; and ethnic origin. Importantly, life style factors, such as participation in exercise and heavy alcohol intake, impact on the level of consumption (supporting the UK government's intensive health promotion campaigns). The paper also investigated the impact of major life events, such as marriage, divorce and stress levels, on the consumption of smokers, but found that the response coefficients were imprecisely estimated.

References

Becker, G., and K. Murphy (1988): "A Theory of Rational Addiction," *Journal of Political Economy*, 96(4), 675–700.

Becker, G., and G. Stigler (1977): "De Gustibus Non Est Disputandum," *American Economic Review*, 68(1), 76–90.

Blaylock, J., and W. Blisard (1992): "U.S. Cigarette Consumption: The Case of Low Income Women," *American Journal of Agricultural Economics*, 74, 698–705.

Boreham, R., and A. Shaw (2002): *Drugs Use, Smoking and Drinking Among Young People in England in 2001.* The Stationary Office.

Chaloupka, F. (1991): "Rational Addictive Behaviour and Cigarette Smoking," *Journal of Political Economy*, 99(4), 722–742.

Conniffe, F. (1995): "Models of Irish Tobacco Consumption," *Economic and Social Review*, 26(4), 331–347.

DeCicca, P., D. Kenkel, and A. Mathios (2002): "Putting Out the Fires: Will Higher Taxes Reduce the Onset of Youth Smoking?," *Journal of Political Economy*, 110(1), 144–169.

Ehrenberg, A. (2000): "Repeat Buying," *Journal of Empirical Generalisations in Marketing Science*, 5, 392–770.

Gernster, E., and D. Holthausen (1986): "Profitable Pricing When Market Segments over-Lap," *Marketing Science*, 5, 55–69.

Godfrey, C. (1986): "Price and Advertising Elasticities of the Demand for Tobacco," Discussion paper, University of York, ESRC Addiction Research Centre.

Greene, W. (1994): "Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models," Working Paper EC-94-10, Stern School of Business, New York University, Stern School of Business, New York University.

Harris, J. (1993): *Deadly Choices: Coping with Health Risks in Everyday Life*. Basic Books, New York.

Harris, J., and S. Chan (1999): "The Continuum-of-Addiction: Cigarette Smoking in Relation to Price Among Americans Aged 15-29," *Health Economics Letters*, 8, 81–86.

Heilbron, D. (1989): "Generalized Linear Models for Altered Zero Probabilities and Overdispersion in Count Data," Discussion paper, University of California, University of California, San Francisco.

Hersch, J. (1996): "Smoking, Seat Belts, and Other Risky Consumer Decisions: Differences by Gender and Race," *Managerial and Decision Economics*, 17(5), 471–481.

Jones, A. (1994): "Health, Addiction,Social Interaction and the Decsion to Quit Smoking," *Journal of Health Economics*, 13, 93–110.

——— (2002): "Do Health Changes Affect Smoking? Evidence from British Panel Data," *Journal of health economics*, 21(4), 533–562.

Kenkel, D. (1995): "Should You Eat Breakfast? Estimates from Health Production Functions," *Health Economics*, 4, 15–29.

Kinsey, J. (1981): "Determinants of Credit Card Accounts: An Application of Tobit Analysis," *Journal of Consumer Reseach*, 8, 172–182.

Lambert, D. (1992): "Zero Inflated Poisson Regression with an Application to Defects in Manufacturing," *Technometrics*, 34, 1–14.

Manski, C. (1977): "The Structure of Random Utility Models," *Theory and Decision*, 8, 229–254.

Maskin, E., and J. Riley (1984): "Monopoly with Incomplete Information.," *Journal of Economics.*, 15, 171–96.

Mullahey, J. (1986): "Specification and Testing of Some Modified Count Data Models," *Journal of Econometrics*, 33, 341–365.

——— (1997): "Heterogeneity, Excess Zeros and the Structure of Count Data Models," *Journal of Applied Econometrics*, 12, 337–350.

Pohlmeier, W., and V. Ulrich (1995): "An Econometric Model of the Two-Part Decision-Making Process in the Demand for Health Care," *Journal of Human Resources*, 30, 339–361.

Sander, W. (1995): "Schooling and Quitting Smoking," *The review of economics and statistics*, 1, 191–199.

Townsend, J., L. (1987): "Cigarette Tax, Economic Welfare and Social Class Patterns of Smoking.," *Applied Economics*, 19(3), 355–365.

Wasserman, J., W. Manning, J. Newhouse, and L. Winkler (1991): "The Effects of Excise Taxes and Regualtions on Cigarette Smoking," *Journal of Health Economics*, 10, 43–64.

Young, T. (1983): "The Demand for Cigarettes: Alternative Specifications of Fujii's Model," *Applied Economics*, 15, 203–211.