THE UNIVERSITY of York

CENTRE FOR HEALTH ECONOMICS

THE IMPACT OF CIGARETTE SMOKING UPON THE EFFICIENT PRODUCTION OF RESPIRATORY HEALTH

Andrew Jones, Rob Manning and Matthew Sutton

CHE Technical Paper Series 5

The Impact of Cigarette Smoking upon the Efficient Production of Respiratory Health

Andrew Jones, Rob Manning and Matthew Sutton

March 1997

CENTRE FOR HEALTH ECONOMICS TECHNICAL PAPER SERIES

The Centre for Health Economics has a well established Discussion Paper series which was originally conceived as a means of circulating ideas for discussion and debate amongst a wide readership that included health economists as well as those working within the NHS and pharmaceutical industry.

The introduction of a Technical Paper Series offers a further means by which the Centre's research can be disseminated. The Technical Paper Series publishes papers that are likely to be of specific interest to a relatively specialist audience, for example papers dealing with complex issues that assume high levels of prior knowledge, or those that make extensive use of sophisticated mathematical or statistical techniques.

The series provides a maximum of 20 pages for authors. The content and its format are entirely the responsibility of the author, and papers published in the Technical Paper series are not subject to peer-review or editorial control, unlike those produced in the Discussion Paper series. Offers of further papers, and requests for information should be directed to Frances Sharp in the Publications Office, Centre for Health Economics, University of York.

THE AUTHORS

Andrew Jones is Director of the Graduate Programme in Health Economics, Department of Economics and Related Studies, University of York. Rob Manning is a Doctoral Student in the Department of Economics and Related Studies. Matthew Sutton is a Research Fellow in the Centre for Health Economics, Univesity of York.

Price £4.00

The impact of cigarette smoking upon the efficient production of respiratory health

Andrew Jones [1], Rob Manning [1] and Matthew Sutton [2]

[1] Department of Economics and Related Studies, University of York
[2] Centre for Health Economics, University of York

7 February 1997

Abstract

Using micro-level data from the Health and Lifestyle Survey (*HALS*), we employ a household production approach to examine the impact of cigarette smoking upon respiratory health. We estimate the efficiency with which individuals produce respiratory health relative to an empirical health production frontier. Instrumental variables techniques are used to explore whether estimates of the true effect of smoking are likely to be contaminated by unobserved individual heterogeneity bias. We find that smoking has a detrimental effect on respiratory health and identify the specific effects of smoking intensity, duration and recovery after quitting.

Keywords: *cigarette smoking; respiratory health; instrumental variables; health production frontiers; Health and Lifestyle Survey (HALS).*

JEL classification: ***C1, D1, I1***

Acknowledgements:
Data from the Health and Lifestyle Survey (*HALS*) were supplied by the *ESRC* Data Archive. We are grateful to Karl Claxton, Paul Contoyannis, Linda Davies, Martin Forster, Roshni Mangalore, and Nigel Rice for their comments on previous drafts.

Corresponding Author:
Dr Andrew Jones, Director of the Graduate Programme in Health Economics,
Department of Economics and Related Studies, University of York, Heslington, York YO1 5DD
tel: +44-1904-433766; fax: +44-1904-433759; e-mail: amj1@york.ac.uk

I. Introduction

This paper considers issues involved in obtaining estimates of the impact of cigarette smoking upon the production of respiratory health and of the (in)efficiency with which respiratory health is produced. The potential problem is one of endogeneity, which may arise when the selection of health inputs is influenced by unobservable characteristics that independently affect the health outcome. If this problem is present, conventional *OLS* estimates of the marginal products of these endogenous inputs will be contaminated by unobserved individual heterogeneity bias, even in large samples (Rosenzweig and Schultz, 1983). In our own analysis, this means that the impact of cigarette smoking on respiratory health may be under- or over-estimated. Following Mullahy and Portney (1990), we address this problem by using an instrumental variables approach.

However, we extend Mullahy and Portney's analysis in two ways. First, we consider the impact of both previous and current smoking behaviour, by including variables that record levels of cigarette consumption for current and ex-smokers, the age at which they started, and time since quitting. Second, we estimate the efficiency with which individuals produce respiratory health relative to a health production frontier. This is motivated by a desire to estimate a model that corresponds more closely to the theoretical production function than the empirical production functions that dominate the applied health-production literature (Kenkel, 1995; Mullahy and Portney, 1990; Rosenzweig and Schultz, 1983). The latter describe the *mean* level of health produced from given input levels, while health production frontiers indicate the *maximum* level of health attainable from given input levels.

Although frontier models have been applied widely to analyse the efficiency of hospitals and other health-care organisations (Wagstaff 1989; Newhouse 1994), they have not, to our knowledge, been used to estimate individual health production frontiers from health interview survey data. Furthermore, the estimation of stochastic production frontiers in the presence of potentially-endogenous inputs has received little attention[1]. Failure to consider this problem may result in biased estimates of (unobserved) inefficiency. While there are a number of alternative frontier specifications from which to choose (Greene, 1993), we concentrate upon the widely-used 'composite-error' specification.

We begin by summarising findings from previous econometric studies that have used micro-level health survey data to examine the relationship between smoking and health. In section III, we describe the micro-level data that we use, which was collected during the 1984/85 Health and Lifestyle Survey (HALS). In section IV, we discuss our own modifications to and extensions of previous studies, before presenting our results in section V. The conclusions that we draw from these results are presented in section VI, together with possible extensions to our work.

II. Previous Econometric Studies of Smoking and Health

Ostro's (1983) analysis of 1976 US health survey data is concerned primarily with estimating the impact of air pollution upon subjective measures of *general* ill health, although the effects of cigarette consumption are also considered. However, the impact of cigarette smoking upon work days lost does not have the anticipated sign and is not statistically significant. On the other hand, the partial effect of smoking on restricted activity days is positive, as predicted, although it too fails to achieve statistical significance. These results may have been affected by unobserved heterogeneity bias if cigarette

[1] Although Zuckerman *et al*'s (1994) stochastic cost frontier analysis used an instrumental-variables approach to overcome the problem of endogenous outputs and input prices.

consumption is endogenously related to health, rather than exogenously-related as assumed. Mullahy and Portney's (1990) study addresses this potential shortcoming by applying the linear generalised method-of-moments (*GMM*) estimator to the same sample as Ostro, although Mullahy and Portney use more recent data, and their health measure records the number of restricted activity days due to *respiratory* (rather than general) ill health. They find that by instrumenting their potentially-endogenous smoking variables (the number of cigarettes consumed at the time of the survey, and its square), an initial negative but nonsignificant effect changes to one that achieves statistical significance. This result implies that unobserved individual heterogeneity masks the true impact of cigarette smoking, relative to the impact of atmospheric pollution. Before describing our own modification to and extensions of Mullahy and Portney's approach (section IV), we first describe the data from which we estimate our models.

III. Data from the 1984/5 Health and Lifestyle Survey (*HALS*)

The Health and Lifestyle Survey (HALS) is a national representative sample of adults in Great Britain. The first-wave data were collected between Autumn 1984 and Summer 1985 during two home visits; a one-hour interview, followed by a nurse visit to collect physiological measurements and tests of cognitive function.

As the Health and Lifestyle Survey (HALS) is a cross-section survey, the sample should be representative of the 1984/85 adult population but not of the original birth cohorts. Because the data are used to model the impact of smoking on health, the issue of attrition due to differential mortality rates may be important. The available sample comprises 9003 individuals, aged 18 and over, living in private households. This represents a response rate of 73.5%. When information from the nurse visit and questionnaire is included, the response rate falls to 53.7% (Cox *et al*, 1987).

Cox *et al* (1987) compared the survey to the 1981 Census of Population, to gauge its representativeness. Among respondents who completed all three stages of the survey, there is a slight excess of women, particularly elderly women, and some under-representation of those individuals with low incomes and less education. However, overall, the authors conclude that 'the study appears to offer a good and representative sample of the population' (Cox *et al*, 1987).

We focus upon the effects of smoking on respiratory health, which has been found to be a better predictor of survival than other measures of health, such as blood pressure and cholesterol levels (Beaty *et al*, 1985). We use forced expiratory volume in one second (*fev1*) as our measure of respiratory health. This measure of lung function was taken during the nurse visit, using an electronic spirometer. Respondents were instructed how to use the equipment and were monitored by the attending nurse. We follow the HALS team in adopting *fev1* as our clinical measure, 'as this index is less likely to have been affected by the respondent's inappropriate use of the spirometer' (Cox *et al* 1987, p.19). Because it is an 'objective', clinical measure of respiratory health, *fev1* has the advantage over Mullahy and Portney's subjective measure of being less prone to health-state dependent measurement error. It is also continuous, in contrast to Mullahy and Portney's count measure, and therefore permits the use of more straightforward estimation techniques. *fev1* has the further advantage that the results are easier to interpret within the context of health-production frontier analysis. Although there are advantages to analysing a dimension of health which is most likely to be smoking-related, there will be broader consequences of smoking which our measure will not capture.

Because it is conventional in the clinical literature to standardise *fev1* for age, gender, height and race (Cox *et al* 1987, p.19), we include variables corresponding to these demographic characteristics in our

regressor set, together with measures of smoking behaviour and air pollution. To account for additional, *observable* individual heterogeneity, measures of education, socioeconomic group, marital and employment status also enter the structural equation for *fev1* (see section IV).

In Mullahy and Portney's (1990) model, levels of air pollution were measured directly. In our analysis, the effects of air pollution are proxied by dummy variables which indicate whether the individual lives in an urban, suburban or rural area. We also include regional dummies, to capture any regional differences in pollution levels.

The HALS provides information on various aspects of the 'life-cycle' of smoking decisions - these smoking decisions include whether to start smoking, whether to quit smoking or whether to continue to smoke - with each individual's smoking status at the time of the survey reflecting the outcome of a sequence of these decisions (see Jones, 1994). *Evrsmkd* is a binary variable that indicates whether an individual has ever been a regular smoker, whilst *Regsmkr* indicates whether they were a regular smoker at the time of the survey[2]. For all individuals who have smoked at some time, we have measures of their previous peak consumption of cigarettes (*Smklvl*), the age that they started smoking (*Agestrt*), and the number of years since they quit (*Yrsquit*). It should be noted that these are all self-reported measures and may be prone to measurement error due to strategic reporting of smoking behaviour or recall bias.

The additional instruments used to predict smoking status include measures of parental smoking and measures of the smoking behaviour of others in the respondent's household. Peer-influences on smoking have been used in other empirical studies, and the corresponding HALS variable has been found to be important in previous work (Jones, 1994; 1996). However, this measure does not capture social influences outside of the household. Furthermore, there is a potential simultaneity problem as, in principle, members of the same household may influence each other. The validity of these instruments for smoking status may appear questionable if we expect exposure to others' cigarette smoke to influence respiratory health, through the effects of passive smoking. We therefore employ an overidentification test (the so-called *J*-test) to establish whether these instruments should enter the health-production equation directly; to see if they are valid instruments.

We exclude individuals who did not respond to the nurse visit component of the survey, and hence record values for *fev1*. Also, we exclude observations with missing values for the socio-demographic and smoking variables used in the models. The resulting sample consists of 6,468 observations. This drops to 3,742 when individuals who have never smoked are excluded. Descriptions and summaries statistics for the variables are provided in Table 1.

Table 1 here

IV. Methodology

We begin with a relatively simple model, which we refer to as 'model A', in which smoking behaviour is measured by the dummy variables reflecting smoking status, *Regsmkr* and *Evrsmkd*. A more sophisticated approach ('model B'), in which measures of an individual's intensity and duration of smoking and time since quitting are included in the specification, is then presented. These models estimate the impact of both current and previous smoking on respiratory health, in contrast to Mullahy and Portney's (1990) approach, which considers only current cigarette consumption.

---

[2] The *HALS* defines regular smoking as at least one cigarette per day for a period of six months or more.

A. Structural Model for fev1, with Smoking Assumed to be Exogenous

Our initial model, estimated by *OLS*, is given by

$$(1) \qquad y_i = X_i\beta + S_i\delta + \epsilon_i \quad , \quad \epsilon_i \sim iid\ (0,\sigma_\epsilon^2)$$

where y_i measures the logarithm of individual i's forced expiratory volume in one second (*Logfev1*), $\boldsymbol{X_i}$ represents the exogenous regressors, $\boldsymbol{S_i}$ records smoking behaviour and ϵ_i captures the effects on y_i of factors that are unobserved by the econometrician.

Mullahy and Portney (1986, p83) argue that, "there is little theoretical guidance as to the likely form of any functional relationship between risk exposure and illness response." However, the *translog production function* has obvious appeal on *a priori* grounds, because of its ties with the economics of production literature and its extensive use in previous frontier applications (Newhouse 1994). Therefore, we model the dependent variable *Logfev1* as a quadratic function of the logarithm of the continuous, exogenous regressors (age and height).

B. Stochastic Frontier Model for fev1, with Smoking Assumed to be Exogenous

Our aim is to estimate an individual health production function which indicates the maximum attainable level of respiratory health for a given set of observable individual characteristics. This suggests that the observed data will be clustered below this maximum and that a stochastic frontier specification will be appropriate. With cross-sectional data it is typical to assume a skewed 'composite-error' term, which comprises a symmetric component to capture 'statistical noise' and a one-sided error component which is interpreted as inefficiency (Wagstaff, 1989; Greene, 1993). To estimate a stochastic health production frontier, we modify the translog production function in equation (1), thus

$$(2) \qquad y_i = \boldsymbol{X_i}\beta + S_i\delta + (v_i + u_i) \quad , \quad u_i \leq 0\ \forall_i \quad , \quad v_i \sim N(0,\sigma_v^2)$$

in which v_i is the symmetric component of the error-term and u_i is one-sided and nonpositive. The inefficiency term u_i takes the value zero for individuals that lie on their stochastic health production frontier, given by $y_i^{max} = \boldsymbol{X_i}\beta + S_i\delta + v_i$, and a negative value for individuals that lie beneath their frontier; $y_i - y_i^{max} = u_i < 0$. For the latter group of individuals, respiratory health is said to be produced 'inefficiently', as their maximum attainable level is not reached.

We test for inefficiency based on the skewness of the estimated residuals, using the $\sqrt{b_1}$ statistic (see Wagstaff, 1989). The estimated residuals for the frontier models are decomposed into their noise (v_i) and inefficiency (u_i) parts, by Jondrow *et al*'s (1982) algorithm. This relies on the further assumption that absence of negative skewness in the compound disturbance indicates that the unexplained variation

in *Logfev1* is random and consequently that there is no inefficiency in production across the sample[3] .

The effects of misspecification of the production technology may be reflected in the estimated residuals and erroneously labelled as 'inefficiency' (Lovell, 1993). Fortunately, the overidentification test (*J*-test) used to establish the validity of our instruments also provides a test for the functional form of the model (Godfrey and Hutton, 1993). However, the distributional assumptions regarding u_i and v_i are nontestable (Newhouse, 1994). This may be problematic, as the effects of misspecifying the functional form of these terms may also be confounded with genuine (unobserved) inefficiency. As a result, we estimate maximum-likelihood stochastic frontier models by adopting the typical assumption that v_i is *iid* normal and employing half-normal, exponential and truncated-normal specifications for the inefficiency term u_i.

C. Structural Model for fev1, with Instrumented Smoking Variables

If individual characteristics that are unobserved by the analyst influence smoking behaviour ($\boldsymbol{S}_i$) and independently affect health (y_i), the true effects of smoking may be confounded with the effects of unobservable individual heterogeneity. In model (1), this would imply that the smoking variables $\boldsymbol{S}_i$ are correlated with the disturbance term ϵ_i. Consequently, *OLS* estimates of the unknown parameter δ would be biased and inconsistent, with the bias being proportional to the magnitude and direction of the correlation between $\boldsymbol{S}_i$ and ϵ_i. In principle, this problem could be overcome by randomising individuals to different smoking patterns, in practice this is not a feasible proposition.

Consequently, we are forced to rely upon nonexperimental methods to obtain consistent estimates of the structural parameters of primary interest (Heckman and Robb, 1985). These methods include instrumental variables (*IV*) techniques which require additional exogenous variables (or 'instruments'), denoted $\boldsymbol{Z}_i$, that satisfy a number of (asymptotic) properties. First, they are uncorrelated with unobservables that influence health (ϵ_i). Second, they are (strongly) correlated with the potentially-endogenous regressors $\boldsymbol{S}_i$. Thus, valid instruments are legitimately excluded from the structural health equation (1) but belong in the (reduced-form) equation that explains the problematic regressor (Heckman and Robb 1985). In our analysis, such instruments should explain smoking patterns but have no direct impact on respiratory health. We implement the instrumental variables approach using three different estimators:-

Linear-generalised methods-of-moments (*linear-GMM*) This is the estimator used by Mullahy and Portney (1990). It is based on two-stage least squares (*2SLS*) estimation of equation (1), implying linear reduced forms for $\boldsymbol{S}_i$, along with White's heteroscedasticity-corrected standard errors.

Two-stage instrumental variables (*2SIV*) This estimates (1) with the actual values of $\boldsymbol{S}_i$ replaced with predicted values $\hat{\boldsymbol{S}}_i$ obtained from reduced-form probit models (see Blundell and Smith, 1993).

Two-stage conditional maximum likelihood (*2SCML*) Rather than including the predicted values for $\boldsymbol{S}_i$, this approach includes the actual values, along with generalised residuals from probit reduced forms. The *t*-ratios of the coefficients associated with the generalised residuals provide *H*-tests for the existence of

[3] Drawing inferences about (in)efficiency from the presence or absence of skewness in the residuals is contested by Schmidt and Sickles (1984) and Skinner (1994).

endogeneity bias (see Blundell and Smith, 1993).

If the model is overidentified, the validity of the instruments may be formally assessed by the *J*-test (see Godfrey and Hutton (1993) for a theoretical exposition, and Jones (1996) for a smoking-related application). If the instruments are found to be valid, it is then possible to establish whether the regressors are in fact endogenous, by performing the *H*-test[4]. If the regressors are found to be endogenous, we should use instrumental-variables models, otherwise we can implement the *OLS* model.

D. Stochastic Frontier Model for fev1, with Instrumented Smoking Variables

To avoid obtaining biased measures of the impact of smoking upon maximum attainable respiratory health and of the (unobserved) inefficiency with which respiratory health is produced, equation (2) can be estimated using *2SIV* or *2SCML*. The composite residuals obtained from the *IV* estimates of the non-frontier model are first tested, to establish whether they are negatively-skewed. As before, half-normal, exponential and truncated-normal specifications are used for the inefficiency term u_i.

V. Results

A. Smoking Behaviour Measured by Dummy Variables Regsmkr and Evrsmkd

In Table 2, we report the intercept and the coefficients for the smoking variables estimated in each of the models with smoking assumed to be exogenous[5].

Table 2 here

The first column contains *OLS* estimates. The full set of results show that higher educational achievement is associated with better health and that there is a clear social class gradient. The proxies for air pollution have the expected qualitative effects; those living in rural areas have better respiratory health than those in suburban areas, and both groups have better health than those living in urban areas. The smoking status variables show significant effects; smokers have poorer respiratory health, *ceteris paribus*, than those who have never smoked. The overall effect of being a current smoker on *Logfev1* is estimated by adding the coefficients of the *Regsmkr* and *Evrsmkd* variables. The estimated effect of -0.075 suggests that a smoker's maximum attainable *fev1* is 93% of a non-smoker's. There is clear evidence of skewness in the residuals, which may indicate unobserved sample-wide inefficiency in the production of respiratory health.

Table 2 also reports estimates for stochastic frontier models, using half-normal, exponential and truncated-normal specifications. The pattern of results for the demographic variable is broadly similar to *OLS*, and across the different distributional assumptions for the inefficiency term u_i. The estimated variance components indicate that measures of inefficiency are significant, with mean inefficiency of over 20% in all of the models. The coefficient on *Regsmkr* is robust to alternative specifications of the frontier model. However, the effect of *Evrsmkd* is fragile; it loses significance in all specifications and is positive

[4] The dangers of proceeding to the *H*-test, without first checking the validity of the instruments by the *J*-test, are discussed by Godfrey and Hutton (1993).

[5] Full results are available from the authors on request.

in exponential and truncated-normal specifications. This may indicate a problem of heterogeneity bias, reflected in the inefficiency component of the disturbance term. A regression of the predicted values of u_i on the regressors shows a significant correlation with *Evrsmkd* (but not *Regsmkr*). In response, we now turn to instrumental variable estimates of the production functions.

Tables 3, 4, and 5 here

Table 3 presents diagnostic tests for the reduced form models for *Regsmkr* and *Evrsmkd*. The *RESET* tests suggest that the probit reduced forms perform better than the linear probability model. The *F* and *LR* tests show that the additional instruments are highly correlated with the smoking variables. However, the relatively low goodness-of-fit statistics suggest that the *IV* estimates may have poor finite sample properties. Table 4 contains the *IV* estimates of the non-frontier models. The *J*-tests of the over-identifying restrictions suggest that the instruments are valid and that the functional form is satisfactory. The *H*-tests are not significant, indicating that the smoking variables can be treated as exogenous. The coefficient on *Evrsmkd* is positive, although not significant. This mirrors the fragility of results for *Evrsmkd* from the frontier specifications. Again there is evidence of skewness in the estimated residuals and so we turn to *IV* frontier estimates. Table 5 includes *2SCML* estimates of the frontier models with the probit generalised residuals, denoted $Gen\text{-}resid_{Regsmkr}$ and $Gen\text{-}resid_{Evrsmkd}$. For these models *Evrsmkd* is positive and significant and the overall effect is smaller than the previous specifications[6].
The main finding of this section is that the estimated effect of being a current smoker is fairly robust across all the specifications. This is confirmed when the models are re-estimated without *Evrsmkd*. Then estimates of the coefficient on *Regsmkr* are -0.069 for *OLS* (compared to a net effect of -0.075 in the model with *Evrsmkd*), -0.057 in the half-normal frontier (compared to -0.057), and -0.075 for linear-*GMM* (compared to -0.032).

However the role of *Evrsmkd* is problematic; the *OLS* estimates are sensitive to the frontier estimation and the use of *IV*, suggesting a self-selection problem. For the frontier models, the estimated coefficient on *Evrsmkd* becomes small relative to *Regsmkr* and is not statistically significant. Furthermore, although the *H*-tests suggest that there is not a problem with endogeneity bias, the estimates for *Evrsmkd* in the *IV* models show significant positive effects. The overall impression is that *fev1* is affected by current smoking, but the role of past smoking may be confounded by unobserved heterogeneity. This leads us to look at the information on an individual's past smoking in more detail.

B. Smoking Behaviour Measured by Intensity and Duration

In this section, we concentrate on the subset of individuals that have smoked at some time in their lives, and attempt to identify the effects of different aspects of smoking on their respiratory health. We include a dummy variable for smoking status, *Exsmkr*, which equals 1 if the individual has quit smoking and 0 if they are a current smoker. In addition, the HALS data provide us with three aspects of smoking history: previous peak consumption (*Smklvl*); the age of starting smoking (*Agestrt*); and, for those who have stopped smoking, the number of years since quitting (*Yrsquit*). These three variables indicate the influence of intensity of smoking, duration of smoking, and recovery from smoking on respiratory health.

Tables 6 and 7 here

[6] *2SIV* results, which are similar to the *2SCML* results reported here, are not presented for the sake of brevity.

Table 6 contains a summary of the results for the non-*IV* estimates for the non-frontier and frontier production functions. The qualitative effects of the smoking variables imply that heavier smokers (*Smklvl*) have lower *fev*, that starting smoking at an earlier age (*Agestrt*) leads to a poorer *fev*, that the one-off effect of quitting smoking *(Exsmkr)* is to improve *fev*, and that *fev* recovers with time after quitting (*Yrsquit*). The magnitude of the effects of *Smklvl* and *Agestrt* are stable across the non-frontier and frontier specifications. For example, the estimated coefficients for *Smklvl* suggest that an additional 60 cigarettes per day on an individual's peak consumption reduces expiratory volume to between 91.2% and 93.9% of what it would otherwise be.

Although the point estimates for *Yrsquit* and *Exsmkr* show some variation across the non-frontier and frontier specifications, the two variables are highly correlated. The impact of quitting smoking on *Log(fev1)* is modelled by a one-off effect of having quit smoking, captured by *Exsmkr*, along with a period of recovery that is linear in *Yrsquit*. Changes in the coefficient on *Exsmkr* are offset by those on *Yrsquit*, and the overall effect is fairly stable across the different models. For example, an estimate of the level of expiratory volume for someone who quit 5 years ago, relative to a current smoker, suggests that their expiratory volume would be around 4% greater. As with the full sample, the frontier specifications suggest a mean level of inefficiency of around 20% relative to maximum attainable lung capacity.

Table 7 explores the possibility that these results may be prone to heterogeneity bias. J tests do not reject the validity of the instruments, and the models are estimated by 2SCML. This involves adding OLS residuals from reduced form equations for *Smklvl* and *Agestrt*, along with generalised residuals from a Tobit equation for *Yrsquit* and a probit equation for *Exsmkr*. There is no evidence from the H tests to suggest problems of endogeneity bias associated with *Smklvl* and *Agestrt,* and Table 7 presents results in which *Yrsquit* and *Exsmkr* are treated as potentially endogenous variables. The 2SCML estimates for the non-frontier model do suggest evidence of endogeneity bias in *Yrsquit*. The generalised residuals for *Exsmkr* are included in the model but are not statistically significant. However the variable does represent the same underlying decision to quit smoking at some time prior to the survey. In principle both variables should be prone to bias if an individual's respiratory health, or some unobservable factor associated with it, influences their decision to quit. However the evidence of endogeneity bias is not borne out in the frontier specifications, in which none of the generalised residuals are significant. Encouragingly, the estimates of the overall impact of smoking are broadly comparable across the frontier and non-frontier models and with the non-IV estimates in Table 6. For instance the impact of an extra 60 cigarettes on peak consumption ranges from 93.5% to 94.5%, and the improvement associated with quitting for 5 years ranges between 3% and 5%.

VI. Conclusions

Our results show that there is something to be gained by adopting a stochastic frontier approach when estimating individual health production functions. We are using a continuous measure of health (*fev1*) from a health interview survey and we would expect individual observations to be clustered below a stochastic frontier that reflects the maximum attainable lung capacity given the observed individual characteristics. Our estimates suggest a mean inefficiency of around 20% relative to maximum attainable lung capacity.

The adoption of a frontier specification has interesting implications for the investigation of endogeneity of smoking decisions in health production models. Treating smoking decisions as exogenous, we have found that, whilst the estimated impact of current smoking is robust, the estimated effect of ever having started smoking is sensitive to the role of unobservable inefficiency in health production. This may

suggest a self-selection problem, whereby individuals with a propensity to inefficient health production are more likely to start smoking.

Our alternative approach to investigate the effects of unobservable heterogeneity is to use the smoking decisions of parents and other members of the respondent's household as instruments for smoking behaviour. An overidentification test suggests that these are valid instruments. However tests of the exogeneity of the smoking status variables do not suggest rejection and, whilst the net effect of current smoking is relatively robust across different specifications, problems with the estimation of the effects of past smoking remain.

In response to these limitations, further analysis concentrates on the sub-sample who have smoked at some time in their lives. In these results we are able to distinguish between different aspects of smoking behaviour. All of these dimensions have the expected qualitative effects: increases in past peak consumption decrease lung capacity; starting smoking at a younger age reduces lung capacity; and quitting produces both one-off and recovery effects. In the sub-sample there is some evidence of endogeneity of the variable measuring years since quitting in the non-frontier models but this does not persist in the frontier specifications.

Our analysis of stochastic production frontiers has relied on a comparison of parametric models. While our results appear to be robust, they may be sensitive to the distributional assumptions required by these models. In future work, we intend to extend Kopp and Mullahy's (1990, 1993) semi-parametric frontier estimators. Also, we plan to use the seven-year follow-up to the Health and Lifestyle Survey (HALS2), and experiment with panel data estimators, which treat inefficiency as an individual fixed effect (Schmidt and Sickles, 1984).

References

Beaty, T.H., C.A. Newill, B.H. Cohen, H.S. Tockman, S.H. Bryant, and H.A. Spurgeon. 1985. "Effects of pulmonary function on mortality." *Journal of Chronic Diseases* 38: 703-10.

Blundell, R.W., and R.J. Smith. 1993. "Simultaneous microeconometric models with censored or qualitative dependent variables." In *Handbook of Statistics No 11: Econometrics*, ed. G.S. Maddala, C.S. Rao, and H.D. Vinod, 117-143. Amsterdam: North Holland.

Cox, B.D. *et al.*1987. *The Health and Lifestyle Survey*. London: London Health Promotion Research Trust.

Greene, W.H. 1993. "The econometric approach to efficiency analysis." In *The Measurement of Productive Efficiency*: *Techniques and Applications*, ed H.O. Fried, C.A.K. Lovell, and S.S. Schmidt, 68-119. Oxford: Oxford University Press.

Godfrey, L.G., and J.P. Hutton. 1993. "Discriminating between errors-in-variables / simultaneity and misspecification in linear regression models." *Economics Letters* 44: 359-64.

Heckman, J.J., and R. Robb.1985. "Alternative methods for evaluating the impact of interventions." In *Longitudinal Analysis of Labour Market Data*, ed Heckman, J.J., and B. Singer, 156-245. Cambridge: Cambridge University Press.

Jondrow, J.C., C. Lovell, I. Materov, and P. Schmidt. 1982. "On the estimation of technical efficiency in the stochastic frontier production model." *Journal of Econometrics* 19: 233-8.

Jones, A. 1994. "Health, addiction, social interaction and the decision to quit smoking." *Journal of Health Economics* 13: 93-110.

Jones, A. 1996. "Smoking cessation and health: a response." *Journal of Health Economics* 15: 755-759.

Kenkel, D.S. 1995. "Should you eat breakfast? Estimates from health production functions." *Health Economics* 4: 15-29.

Kopp, R.J., and J. Mullahy. 1990. "Moment-based estimation and testing of stochastic frontier models." *Journal of Econometrics* 46: 165-183.

Kopp, R.J., and J. Mullahy. 1993. "Least squares estimation of econometric frontier models: consistent estimation and inference." *Scandinavian Journal of Economics* 95(1): 125-132.

Lovell, C.A.K. 1993. "Production frontiers and productive efficiency." In The Measurement of Productive Efficiency: Techniques and Applications, ed H.O. Fried, C.A.K. Lovell, and S.S. Schmidt, 3-67. Oxford: Oxford University Press.

Mullahy J and Portney PR. 1986. "Cigarette smoking, air pollution, and respiratory illness: an analysis of relative risks". Advances in Applied Micro-Economics; 4: 83-103. JAI Press Inc.

Mullahy, J., and P.R. Portney. 1990. "Air pollution, cigarette smoking, and the production of respiratory health." *Journal of Health Economics* 9: 193-205.

Newhouse, J.P. 1994. "Frontier estimation: how useful a tool for health economics?" *Journal of Health Economics* 13: 317-22.

Ostro, B.D. 1983. "The effects of air pollution on work loss and morbidity." *Journal of Environmental Economics and Management* 10: 371-82.

Rosenzweig, M.R. and T.P. Schultz. 1983. "Estimating a household production function: heterogeneity, the demand for health inputs and their effects on birth weight." *Journal of Political Economy* 91 (51): 723-46.

Schmidt, P. and R.C. Sickles. 1984. "Production frontiers and panel data." *Journal of Business and Economic Statistics* 2(4): 367-74.

Skinner, J. 1994. "What do stochastic frontier cost functions tell us about inefficiency?" *Journal of Health Economics* 13: 323-328.

Wagstaff, A. 1989. "Estimating the efficiency of the hospital sector: a comparison of three statistical cost frontier models." *Applied Economics* 21:659-72.

Zuckerman, S., J. Hadley, and L. Iezzoni. 1994. "Measuring hospital efficiency with frontier cost functions." *Journal of Health Economics* 13: 255-280.

Table 1

Definition of Variables and Sample Means

Variable	Variable definition	Sample mean (n = 6468)	Sample mean (n = 3742)
Dependent Variable			
Logfev1	logarithm of forced expiratory volume in one second	0.926	0.906
Independent Variables			
1. Demographic characteristics			
Male	Male (Yes = 1, No = 0)	0.454	0.518
Prgnt	Pregnant (Yes = 1, No = 0)	0.014	0.012
Lnage	logarithm of age in years	3.722	3.756
Lnagesq	logarithm of (age in years squared / 100)	14.011	14.251
Lnhgt	logarithm of height in inches	4.191	4.195
Lnhgtsq	logarithm of (height in inches squared / 100)	17.566	17.602
Lnaghgt	interaction for *Lnage*Lnhgt*	15.596	15.756
2. Ethnic Group Variables ('Reference group' = *White*)			
Black	Black (Yes = 1, No = 0)	0.008	0.006
Asian	Asian (Yes = 1, No = 0)	0.013	0.005
Othethnc	Other ethnic group (Yes = 1, No = 0)	0.007	0.006
3. Employment Status Variables ('Reference group' = *Full Time*)			
Partime	Part time (Yes = 1, No = 0)	0.129	0.120
Unemp	Unemployed (Yes = 1, No = 0)	0.072	0.089
Retd	Retired (Yes = 1, No = 0)	0.191	0.197
Stdnt	Student (Yes = 1, No = 0)	0.011	0.005
Keephse	House keeper (Yes = 1, No = 0)	0.143	0.138

Table 1 (*contd*)

Definition of Variables and Sample Means

Variable	Variable definition	Sample mean (n = 6468)	Sample mean (n = 3742)
Independent Variables (*contd*)			
4. Qualifications Variables ('Reference group' = *No* or *Other Work-related Qualifications*)			
Qcse25	Highest qualification CSE grade 2-5 (Yes = 1, No = 0)	0.036	0.031
Qolevel	Highest qualification O' Level (Yes = 1, No = 0)	0.152	0.142
Qalevel	Highest qualification A' Level (Yes = 1, No = 0)	0.047	0.039
Qdegree	Highest qualification Degree (Yes = 1, No = 0)	0.052	0.043
Qteach	Highest qualification Teacher Training Qualification (Yes = 1, No = 0)	0.022	0.015
Qnurse	Highest qualification Teacher Training Qualification (Yes = 1, No = 0)	0.024	0.023
Qprof	Highest qualification Professional Qualification (Yes = 1, No = 0)	0.037	0.034
5. Socioeconomic Group Variables ('Reference Group' = *SEG4*)			
SEG1	Socio-economic group 1 (Yes = 1, No = 0)	0.056	0.047
SEG2	Socio-economic group 2 (Yes = 1, No = 0)	0.183	0.176
SEG3	Socio-economic group 3 (Yes = 1, No = 0)	0.204	0.177
SEG5	Socio-economic group 5 (Yes = 1, No = 0)	0.159	0.173
SEG6	Socio-economic group 6 (Yes = 1, No = 0)	0.056	0.060
SEGARMY	Armed forces (Yes = 1, No = 0)	0.007	0.008
6. *Marital Status Variables* ('Reference Group' = *Married*)			
Widow	Widow (Yes = 1, No = 0)	0.072	0.066
Single	Single (Yes = 1, No = 0)	0.165	0.132
Seprtd	Separated (Yes = 1, No = 0)	0.021	0.026
Divorced	Divorced (Yes = 1, No = 0)	0.039	0.044

Table 1 (*contd*)

Definition of Variables and Sample Means

Variable	Variable definition	Sample mean (n = 6468)	Sample mean (n = 3742)
Independent Variables (*contd*)			
7. Type of Area Variables ('Reference Group' = *Urban*)			
Suburb	House / flat situated in a suburban area (Yes = 1, No = 0)	0.462	0.469
Rural	House / flat situated in a rural area (Yes = 1, No = 0)	0.204	0.192
8. *Regional Variables* ('Reference Group' = *London*)			
Scot	Standard Region Scotland (Yes = 1, No = 0)	0.043	0.044
Wales	Standard Region Wales (Yes = 1, No = 0)	0.064	0.073
North	Standard Region North (Yes = 1, No = 0)	0.115	0.126
Nwest	Standard Region North West (Yes = 1, No = 0)	0.086	0.085
Yorks	Standard Region Yorkshire (Yes = 1, No = 0)	0.089	0.084
Wmids	Standard Region West Midlands (Yes = 1, No = 0)	0.080	0.075
Emids	Standard Region East Midlands (Yes = 1, No = 0)	0.041	0.045
Anglia	Standard Region Anglia (Yes = 1, No = 0)	0.082	0.080
Swest	Standard Region South West (Yes = 1, No = 0)	0.195	0.184
Seast	Standard Region South East (Yes = 1, No = 0)	0.102	0.104
9. Smoking Variables			
Regsmkr	Regular Smoker at Time of Survey (Yes = 1, No = 0)	0.331	
Evrsmkd	Regular Smoker at any Time (Yes = 1, No = 0)	0.588	
Exsmkr	Used to be a Regular Smoker (Yes = 1, No = 0)		0.43
Smklvl	Peak cigarette consumption		22.58
Agestrt	Age started smoking in years		16.98
Yrsquit	Years since quitting smoking		5.43

Table 1 (*contd*)

Definition of Variables and Sample Means

Variable	Variable definition	Sample mean (n = 6468)	Sample mean (n = 3742)
Additional instruments			
Smkothr	Other smokers in the household (Yes =1, No = 0)	0.250	0.440
Parsmk1	Only mother smoked (Yes =1, No = 0)	0.121	0.072
Parsmk2	Only father smoked (Yes =1, No = 0)	0.55	0.470
Parsmk3	Both parents smoked (Yes =1, No = 0)	0.102	0.364

Table 2

Structural equation and stochastic frontier models, with smoking assumed to be exogenous

Variables	Structural model (*OLS*)*	Frontier model (half-normal)	Frontier model (exponential)	Frontier model (truncated-normal)
Intercept	3.059 (0.199)	3.847 (0.355)	3.464 (0.341)	3.251 (0.316)
Regsmkr	-0.057 (-6.096)	-0.054 (-8.312)	-0.058 (-9.774)	-0.058 (-9.602)
Evrsmkd	-0.018 (-2.172)	-0.003 (-0.541)	0.003 (0.461)	0.002 (0.346)
σ_v^2		0.007	0.010	0.010
σ_u^2		0.136	0.047	1.022
$E(u)$		0.279	0.218	0.224
skewness: $\sqrt{b_1}$ +	-1.800			
R^2 (*adj*)	0.605			

* *adjusted for heteroscedasticity by White's method.*

\+ $\sqrt{b_1}^{(0.05)}_{n=5000} = -0.057$

Table 3

Diagnostics: reduced-form smoking equations

	Regsmkr (*LPM*)	*Regsmkr* (*probit*)	*Evrsmkd* (*LPM*)	*Evrsmkd* (*probit*)
R^2 (*adj*)	0.135		0.112	
McFadden R^2		0.117		0.093
Count R^2		0.707		0.655
RESET test (*t*-ratio)	2.794	-0.399	-0.938	0.173
LR test (χ^2_4)		458.412		217.716
F test (4 , 6416)	126.126		55.825	

Table 4

Structural equation models, with instrumented smoking variables (OLS model included for comparison)

Variables	Structural model (*OLS*)	Structural model (*linear-GMM*)	Structural model (*2SIV*)	Structural model (*2SCML*)
Intercept	3.059 (0.199)	4.393 (0.273)	2.634 (0.169)	2.777 (0.180)
Regsmkr	-0.057 (-6.096)	-0.116 (-2.151)	-0.117 (-2.109)	-0.096 (-1.993)
Evrsmkd	-0.018 (-2.172)	0.076 (1.011)	0.081 (1.055)	0.057 (0.857)
$Gen\text{-}resid_{Regsmkr}$				0.023 (0.788)
$Gen\text{-}resid_{Evrsmkd}$				-0.047 (-1.133)
J-test +		2.005	2.005	2.587
R^2 *(adj)*	0.605			

* *adjusted for heteroscedasticity by White's method.*

\+ $\chi_2^{2\ (0.05)} = 5.99$

Table 5

Structural equation and stochastic frontier models, estimated by 2SCML

Variables	Structural model (*2SCML*)*	Frontier model (half-normal)	Frontier model (exponential)	Frontier model (truncated-normal)
Intercept	2.777 (0.180)	3.590 (0.312)	2.909 (0.286)	2.546 (0.247)
Regsmkr	-0.096 (-1.993)	-0.082 (-2.484)	-0.096 (-3.214)	-0.090 (-2.966)
Evrsmkd	0.057 (0.857)	0.054 (1.141)	0.082 (1.943)	0.071 (1.643)
$Gen\text{-}resid_{Regsmkr}$	0.023 (0.788)	0.016 (0.837)	0.023 (1.261)	0.019 (1.023)
$Gen\text{-}resid_{Evrsmkd}$	-0.047 (-1.133)	-0.035 (-1.234)	-0.049 (-1.906)	-0.043 (-1.647)
σ_v^2		0.007	0.010	0.010
σ_u^2		0.136	0.047	1.022
$E(u)$		0.279	0.218	0.223
J-test +	2.587			
R^2 *(adj)*	0.605			

* *adjusted for heteroscedasticity by White's method.*

\+ $\chi_2^{2\ (0.05)} = 5.99$

Table 6

Structural equation and stochastic frontier models, with smoking assumed exogenous: smokers only

Variables	Structural model (*OLS*)*	Frontier model (half-normal)	Frontier model (exponential)	Frontier model (truncated-normal)
Intercept	-16.025 (-0.706)	7.232 (0.459)	6.010 (0.437)	-5.816 (-0.442)
Smklvl	-0.154E-2 (-4.471)	-0.105E-2 (-4.354)	-0.104E-2 (-4.951)	-0.111E-2 (-5.391)
Agestrt	0.276E-2 (2.757)	0.134E-2 (2.101)	0.102E-2 (1.779)	0.108E-2 (1.890)
Yrsquit	0.412E-2 (5.261)	0.292E-2 (6.531)	0.257E-2 (6.166)	0.256E-2 (6.190)
Exsmkr	0.170E-1 (1.369)	0.236E-1 (2.494)	0.303E-1 (3.679)	0.296E-1 (3.649)
σ_v^2		0.007	0.010	0.010
σ_u^2		0.156	0.055	1.148
$E(u)$		0.298	0.235	0.242
skewness: $\sqrt{b_1}$ +	-1.700			
R^2 (*adj*)	0.578			

* *adjusted for heteroscedasticity by White's method.*

\+ $\sqrt{b_1}^{(0.05)}_{n=5000} = -0.057$

\+ $\chi_2^{2\,(0.05)} = 5.99$

Table 7

Structural equation and stochastic frontier models, 2SCML estimates: smokers only

Variables	Structural model (*2SCML*)*	Frontier model (half-normal)	Frontier model (exponential)	Frontier model (truncated-normal)
Intercept	-21.738 (-0.959)	5.857 (0.370)	5.131 (0.372)	-10.142 (-0.740)
Smklvl	-0.113E-2 (-3.072)	-0.941E-3 (-3.424)	-0.979E-3 (-4.097)	-0.968E-3 (-3.956)
Agestrt	0.316E-2 (3.110)	0.142E-2 (2.138)	0.107E-2 (1.793)	0.108E-2 (1.767)
Yrsquit	0.922E-2 (3.662)	0.367E-2 (2.025)	0.297E-2 (1.871)	0.324E-2 (2.003)
Exsmkr	-0.148E-2 (-0.035)	0.314E-1 (0.908)	0.371E-1 (1.260)	0.362E-1 (1.195)
$Gen\text{-}resid_{Yrsquit}$	-0.566E-2 (-2.079)	-0.848E-3 (-0.445)	-0.444E-3 (-0.264)	-0.692E-3 (-0.403)
$Gen\text{-}resid_{Exsmkr}$	0.554E-1 (1.229)	0.136E-2 (0.043)	-0.995E-3 (-0.036)	0.116E-2 (0.041)
σ_v^2		0.007	0.010	0.010
σ_u^2		0.156	0.055	1.074
$E(u)$		0.298	0.235	0.242
J-test	2.657			
R^2 *(adj)*	0.579			

* *adjusted for heteroscedasticity by White's method.*

\+ $\chi_2^{2\,(0.05)} = 5.99$