AN EMPIRICAL ANALYSIS OF ADULT CIGARETTE DEMAND

John A. Tauras
University of Illinois at Chicago and National Bureau of Economic Research

INTRODUCTION

Between the years 1900-1964, the use of cigarettes increased dramatically in the United States. Per capita consumption of cigarettes rose from 54 in 1900 to a peak of 4,345 in 1963, the year prior to the release of the first Surgeon General's report on the health consequences of cigarette smoking [USDHHS, 1991]. Throughout the following two decades, significant progress was made in reducing cigarette smoking in all segments of the U.S. population. Overall smoking prevalence declined from 40.4 per cent in 1965 to 30.4 per cent in 1985 [USDHHS, 1989]. However, survey data from the 1990's suggest that the decades of steady decline in smoking prevalence have not been sustained. According to the National Health Interview Surveys, the prevalence of cigarette smoking among adults has remained virtually unchanged between the years 1991 and 1997 [CDC, 1999].

The discipline of economics has made substantial contributions to understanding the determinants of cigarette demand. Throughout the past three decades, economists have examined the impact of public policy, placing special emphasis on the impact of cigarette price, on the demand for cigarettes. Using diverse data, theoretical modeling, and estimation techniques, one general conclusion has emerged from these studies - cigarette prices are inversely related to cigarette demand.[1]

A majority of the previous studies employed aggregate level data in estimating cigarette demand equations. These studies have primarily examined the impact of prices and tobacco control policies on average cigarette consumption. Price elasticity estimates obtained from these studies range from –0.14 to –1.23, with a preponderance of the estimates falling in a narrower range of –0.20 to –0.50 [USDHHS, 1994].

A growing number of recent studies have estimated cigarette demand equations using individual level data. The use of individual-level data allows for the investigation of both the propensity to smoke and the intensity with which smokers smoke. The cumulative distribution of cigarette consumption can therefore be characterized as a mixed distribution, one that is neither continuous nor discrete. There exists a mass of zero outcomes, or in other words, a large number of individuals who do not smoke. In addition, there exists a continuous distribution for those individuals who do smoke. The continuous distribution is frequently skewed to the right and exhibits non-constant variance.

Economists have traditionally modeled this mixed distribution using a two-part model. In the first part of the two-part model, a logit or probit specification is employed

John A. Tauras: Department of Economics (M/C 144), University of Illinois, 601 S. Morgan, Chicago, IL 60607-7121. E-mail: tauras@uic.edu.

Eastern Economic Journal, Vol. 31, No. 3, Summer 2005

to estimate the decision to smoke, whereas in the second part of the model, ordinary least squares (OLS) is used to estimate the amount smoked by smokers. It has become standard practice to subject the dependent variable from the second part of the two-part model to a logarithmic transformation. This transformation is conducted in an attempt to stabilize non-constant error variances and/or to more nearly satisfy normality.

Recently, more attention has been directed at improving the empirical methods in health economics. Of particular importance to this manuscript is the research that has focused on the choice of functional form and estimation techniques. In a series of seminal articles, John Mullahy and Willard Manning discuss potential problems in terms of bias and precision associated with employing a traditional two-part model and suggest alternative estimation techniques to estimate mixed distributions in the face of common data problems [Mullahy, 1998; Manning, 1998; Manning and Mullahy, 2001]. One facet of Mullahy and Manning's theoretical contributions has to deal with prediction bias associated with using log-transformed dependent variables. This paper examines the magnitude of prediction bias when employing a traditional two-part model of adult cigarette demand. In particular, this paper is the first econometric study to investigate the bias associated with naively assuming that the error retransformation is homoscedastic in price.

The next section, which discusses prediction bias, draws from the recent work by Mullahy and Manning and the reader is referred to these articles for additional reading.

PREDICTION BIAS PROBLEM

The typical regression equation estimated for the second part of the two-part model of cigarette smoking can be represented by the following log-linear equation:

$$\ln(Y) = XB + \varepsilon,\ Y > 0 \tag{1}$$

The corresponding cigarette demand equation exhibits the following exponential relationship:

$$Y = e^{XB}e^{\varepsilon}. \tag{2}$$

Virtually all previous studies of cigarette demand that have employed a logarithmic transformed dependent variable for the second part of the two part model have neglected to include the error term when transforming equation (1) back from log-scale to original scale. This naïve transformation yields:

$$E(Y \mid X) = e^{XB}. \tag{3}$$

When in fact, the correct expectation of Y should be:

$$E(Y \mid X) = e^{XB}E(e^{\varepsilon}). \tag{4}$$

The omission of $E(e^{\varepsilon})$, likely stems from researchers' intimate knowledge of prediction with linear regression models of untransformed data in which the $E(e^{\varepsilon})$ and ε enter arithmetically. In sharp contrast, with a log-linear model such as equation (1),

the error term enters multiplicatively in which $E(\varepsilon) = 0$ in logarithmic units not arithmetic units. Because the mean of the error terms will not equal zero after transforming from log-scale to original scale, the error term must be retained for predictive purposes.

Predictions based on the naïve transformation that omit the error retransformation are biased and underestimate $E(Y|X)$ by the quantity $E(e^{\varepsilon})$. It should be noted that equation (1) consistently estimates B, which is necessary, but not sufficient for accurate prediction.

It is quite possible that the error retransformation, $E(e^{\varepsilon})$, is not a constant, but rather is heteroscedastic in X. The possible dependence of the retransformation factor on X can be represented by $E(e^{\varepsilon|X}) = \rho(x)$.

In order to correct the prediction bias, an estimate of $\rho(x)$ is required. If $\rho(x) = \phi$, where ϕ is a constant, two alternative approaches have been advanced to estimate the retransformation factor. First, if the distribution of the error term is known, the retransformation factor can be derived directly. For example if the logarithmic disturbances are normally distributed $N(0,\sigma^2)$ the error retransformation, ϕ, equals $e^{\sigma^2/2}$. σ^2 can be estimated by using the mean square of the error from the regression of equation (1). Second, if the distribution of the error term is not known, then a nonparametric smearing estimator developed by Duan [1983] can be employed. The smearing estimator "smears" the average of the exponentiated residuals across all cases: $\phi = \frac{1}{N}\sum_{i=1}^{N} e^{\hat{\varepsilon}}$.

Mullahy [1998] convincingly argues that if $\rho(x) \neq \phi$, that is, if the error retransformation is not a constant but rather depends on X, then the two approaches discussed above will in general not provide consistent estimates of the retransformation function and therefore will produce biased estimates of the derived elasticities. To see this, the true elasticity of demand for the second part of the two-part model, η_j, can be defined in terms of logarithmic derivatives:

$$\eta_j = \frac{\partial \ln E(Y \mid X)}{\partial \ln(X_j)}. \tag{5}$$

Defining $\mu(x) = e^{XB}$, equation (4) can be rewritten as:

$$E(Y \mid X) = \mu(x)\rho(X). \tag{6}$$

Plugging this into equation (5) yields the true elasticity:

$$\eta_j = \frac{\partial \ln(\mu(X))}{\partial \ln X_j} + \frac{\partial \ln(\rho(X))}{\partial \ln X_j}. \tag{7}$$

The resulting elasticity for the second part of the homoscedastic two-part model, however, yields:

$$\eta_j = \frac{\partial \ln(\mu(X))}{\partial \ln X_j} . \tag{8}$$

If heteroscedasticity exists and is ignored, the estimated elasticity will differ from the true elasticity by:

$$\frac{\partial \ln(\rho(X))}{\partial \ln X_j} \tag{9}$$

This suggests that if heteroscedasticity exists and is ignored, biased elasticity estimates will result for the second part of the two part model because the impact of X on $E(Y|X)$ which works through $\rho(X)$ is ignored.

Manning [1998] shows that if the OLS residuals on the log scale are heteroscedastic by some X, the bias for the second part of the two-part model can be corrected by using a heteroscedastic version of the Duan's smearing estimator. The heteroscedastic smearing estimator "smears" the average of the exponentiated residuals across each subgroup - where the error term is heteroscedastic by subgroup. However, if there is heteroscedasticity across numerous subgroups or if heteroscedasticity exists for a continuous covariate, the heteroscedastic retransformation can be extremely difficult – even for seasoned econometricians.

An alternative estimator that Manning and Mullahy [2001] and Blough, Madden, and Hornbrook [1999] proposed is the generalized linear model (GLM). Unlike using an OLS estimator on the $\ln(Y)$ which estimates $E(\ln(Y|X))$ and requires retransformation to $\ln(E(Y|X))$, GLM models estimate $\ln(E(Y|X))$ and consequently $E(Y|X)$ directly and preclude the need to retransform.

DATA

The empirical models that are estimated in this study employ data from the 1993 and 1994 National Health Interview Surveys (NHIS).[2] The NHIS are cross-sectional surveys covering the civilian non-institutionalized population of the United States.[3] Information on individuals is obtained continuously throughout the year and is collected through personal interviews conducted by U.S. Bureau of Census employees using a stratified multistage probability sampling technique. The primary focus of the surveys is to collect information on the quantity, distribution, and effects of disability and illness in the United States and to measure the amount of services used to treat these disabilities and illnesses.

In each year, all respondent were asked about their current and past cigarette use. These data were used to construct two alternative dependent variables: participation in cigarette smoking and average monthly smoking. The first measure is a dichotomous indicator equal to 1 for respondents who indicate that they smoked cigarettes in the 30 days prior to the survey and is equal to zero otherwise. The second dependent variable is a continuous measure of monthly cigarette consumption based on the number of days each smoker smokes each month multiplied by the average number of cigarettes smoked per day on days smoked. Table 1 includes variable definitions and basic descriptive statistics.

Based on the survey data, numerous independent variables were selected to control for factors that are thought likely to influence cigarette smoking. These include: gender (female – omitted and male); marital status (single – omitted, married, and widowed or divorced); race/ethnicity (White non-Hispanic – omitted, White Hispanic, Black non-Hispanic, Black Hispanic, other Hispanic, and other non-Hispanic); indicators of education (less than high school diploma – omitted, high school graduate, and at least some college); employment status during the past two weeks (employed – omitted, not employed, and not in labor force); indicators of age (18-24, 25-44, 45-64, and greater than or equal to 65 - omitted); average real yearly family income (categories include:$0-9,999; $10,000-19,999; $20,000-34,999; and greater than $35,000; all deflated by national consumer price index, 1982-1984=100); indicators of type of community (in central city of a metropolitan statistical area (MSA), in MSA but not in central city – omitted, in a farm community not in a MSA, and not in MSA not in a farm community); poverty status (below poverty threshold – omitted and above poverty threshold); and indicator for year survey was conducted (1994 – omitted and 1993).

The family income variable poses two interesting econometric issues. First, given the sensitive nature of family income, approximately 13 per cent of the respondents did not provide information on family income. Second, individuals classify their yearly family income into one of four broadly defined income categories introducing measurement error in the family income covariate. Numerous sensitivity analyses are conducted to examine the impact missing data and explanatory variable measurement error have on the cigarette demand equation estimates.

Three alternative approaches are employed for handling missing values on income and poverty status.[4] First, a complete case method was used.[5] Second, an imputation of missing values approach was employed.[6] Finally, a modified zero order regression approach discussed in Greene [1993] was employed.[7] In addition, four alternative approaches are employed to assess the impact of measurement error in the family income variable. First, the midpoints of the categorical responses are employed as a regressor in the regressions with the highest income category coded as $50,000. Second, the lower bounds of the categorical responses are used as a regressor in the regressions. Third, the upper bounds of the categorical responses with the highest income category coded as $75,000 is used as a regressor in the regressions. Finally, dichotomous indicators for each income category are created and used as regressors in the regressions with the highest income category established as the reference category.[8]

Based on the state identifiers, cigarette prices were added to the surveys. The price data were obtained from Tobacco Institute's annual *Tax Burden on Tobacco*.[9] Each year prior to 2000, the Tobacco Institute published state-level cigarette prices as of November 1. These prices are weighted averages for a pack of 20 cigarettes based on the prices of single packs, cartons, and vending machine sales where the weights are the national proportions of each type of sale. These prices are inclusive of state level sales taxes applied to cigarettes, but are exclusive of local cigarette taxes. Since the price published is as of November 1, and the surveys are conducted throughout the year, a weighted average price for the year is computed. To account for changes in the relative price of cigarettes over time, all cigarette prices are deflated by the

national Consumer Price Index published by the Bureau of Labor Statistics (1982-1984=100).

Table 1
Variable Definitions and Basic Descriptive Statistics

Independent Variables	Definition, Sample Mean(x), Sample Standard Deviation(s)
Real Price of Cigarettes	Average yearly price of a pack of twenty cigarettes, deflated by the national Consumer Price Index, 1982-1984=100. =1.20 s=0.137
Real Average Cigarette Price Adjacent States	Average real price of cigarettes from all adjacent states. =1.16 s=0.091
Male	Dichotomous indicator equal to one if respondent is a male, and zero other wise. =0.422 s=0.494
Age 18-24	Dichotomous indicator equal to one if respondent is aged between 18 and 24, and zero otherwise. =.111 s=.314
Age 25-44	Dichotomous indicator equal to one if respondent is aged between 25 and 44, and zero otherwise. =.422 s=.494
Age 45-64	Dichotomous indicator equal to one if respondent is aged between 45 and 64, and zero otherwise. =.262 s=.440
Black	Dichotomous indicator equal to one if respondent is Black, and zero other wise. =.135 s=.342
Other Race	Dichotomous indicator equal to one if respondent is not White, Black, White Hispanic, Black Hispanic, or other Hispanic, and zero otherwise. =.040 s=.195
White Hispanic	Dichotomous indicator equal to one if respondent is White Hispanic, and zero otherwise. =.074 s=.261
Black Hispanic	Dichotomous indicator equal to one if respondent is Black Hispanic, and zero otherwise. =.004 s=.062
Other Hispanic	Dichotomous indicator equal to one if respondent is other Hispanic, and zero otherwise. =.003 s=.059
Married	Dichotomous indicator equal to one if respondent is married, and zero oth erwise. =.524 s=.499
Separated, Divorced or Widowed	Dichotomous indicator equal to one if respondent is separated, widowed, or divorced, and zero otherwise. =.270 s=.444
High School Graduate	Dichotomous indicator equal to one if respondent is a high school graduate, and zero otherwise. =.364 s=.481
Some College	Dichotomous indicator equal to one if respondent has completed some college, and zero otherwise. =.430 s=.495
Not Employed	Dichotomous indicator equal to one if respondent is unemployed, and zero otherwise. =.032 s=.177
Not in Labor Force	Dichotomous indicator equal to one if respondent is not in labor force, and zero otherwise. =.372 s=.483
Real Family Income	Average real yearly family income(in dollars) Deflated by the national Con sumer Price Index, 1982-1984=100. =157.64 s=124.17
MSA Central City	Dichotomous indicator equal to one if respondent lives in an MSA central city, and zero otherwise. =.337 s=.473
Not in MSA Non-Farm	Dichotomous indicator equal to one if respondent does not live in an MSA and does not live on a farm, and zero otherwise. =.215 s=.411
Not in MSA Farm	Dichotomous indicator equal to one if respondent lives on a farm not in an MSA, and zero otherwise. =.011 s=.104
Not under Poverty Threshold	Dichotomous indicator equal to one if respondent is not under the poverty threshold, and zero otherwise. =.789 s=.408
1993	Dichotomous indicator equal to one if the respondent was surveyed in 1993 and zero otherwise. =.513 s=.500

Several previous studies of cigarette demand have suggested that differences in cigarette prices among states may lead smokers in high price states to purchase cigarettes in nearby low price states [Lewit and Coate, 1982; Chaloupka, 1991; Becker, Grossman, and Murphy, 1994; and Coates, 1995]. These studies point out that failing to control for these cross-border purchases will result in biased estimates of the effect of price on cigarette demand. In an attempt to control for interstate differences in cigarette prices a variable reflecting the average price of cigarettes in all neighboring states is included as a covariate in each model.[10]

Several recent studies conducted on the determinants of cigarette demand have concluded that unobserved state level sentiment toward smoking may be driving both the price of cigarettes to increase (through cigarette excise tax initiatives) and adult smoking rates to decrease [Becker, Grossman, and Murphy, 1994; Gruber and Koszegi, 2001; Gruber and Zinman, 2001; and DeCicca, Kenkel, and Mathios, 2002]. If state level sentiment toward smoking is not controlled for, there is a possibility that the estimated price elasticities of demand may be biased away from zero. In order to control for unobserved sentiment, dichotomous indicators for each state – less one are usually included as covariates in the regression specification. These state "fixed effects" are intended to capture all the factors that are unobserved at the state level that affect cigarette smoking. Given that the National Center for Health Statistics only provided two years of data with state level identifiers and detailed information on adult cigarette consumption, the inclusion of state dummies as regressors results in near perfect multicollinearity. Therefore, the results presented here do not control for unobserved state level heterogeneity.

METHODS

As was discussed in sections I and II above, it has become standard practice in the cigarette demand literature to subject the dependent variable from the second part of the two-part model to a logarithmic transformation. If the only reason that the logarithmic transformation is undertaken is to correct for heteroscedasticity, than an alternative estimator, White's [1980] heteroscedasticity consistent covariance matrix estimation technique, can be employed on the untransformed (original scale) dependent variable. The use of White's procedure on an untransformed dependent variable would circumvent any retransformation issues. It is well known that OLS estimates are unbiased when heteroscedasticity is present and ignored, however, the estimated variances and covariances of the coefficients are biased and inconsistent. White's procedure yields consistent estimates of the variances and covariances of the OLS estimates so that proper inference (hypothesis testing) can be carried out.

Unfortunately, a linear specification is rejected in favor of a log linear model. In particular a P_E test developed by Davidson and MacKinnon [1981] is employed to test the linear specification against a log-linear model. The linear model is strongly rejected whereas the log-linear model cannot be. Furthermore, a Box-Cox [1964] power transformation analysis was conducted. The transformation parameter, λ, was allowed to vary from –3 to 3 with a step size of 0.5. A transformation parameter of zero

yielded the smallest error sum of squares, and consequently the largest log-likelihood which implies that the log-linear model is the appropriate model to employ.

White's [1980] heteroscedasticity consistent covariance matrix estimation technique can be employed on log transformed dependent variables, however, when transforming back from log scale to original scale an estimate of the error retransformation must be employed. If the error retransformation depends on some X, as is the case here, then a heteroscedastic error retransformation technique, such as the heteroscedastic version of Duan's smearing estimator, can be used to yield unbiased elasticity estimates. This heteroscedastic retransformation can be intractable when heteroscedasticity exists in a continuous covariate, such as price, or exists among several subgroups. It is for this reason that almost all previous cigarette demand studies have used a naïve retransformation that ignored the impact of X on $E(Y|X)$ which works through $\rho(X)$.

This paper compares the price elasticity of conditional cigarette demand that was derived from a model that employs a naïve retransformation (from log scale to original scale) to the price elasticity of conditional demand from a model that uses a GLM framework that estimates $E(Y|X)$ directly and precludes the need to retransform. Both techniques employ White's [1980] method for obtaining consistent estimates of the variances and covariances when heteroscedasticity is present. Further, an additional cluster correction of the standard errors at the state level is employed for both techniques. That is, the standard errors are corrected for within state correlation. Clustering at the state level relaxes the assumptions of independence of observations so that observations only have to be independent across states but not within states.

Traditional Two Part Model (naive retransformation of conditional demand)

The traditional two part model that is estimated in this paper models the propensity to smoke and the intensity of smoking separately. In the first step of the traditional model, probit methods are used to estimate a smoking participation equation. In the second step, ordinary least squares methods are used to estimate the average monthly cigarette consumption by smokers, where the dependent variable is the natural logarithm of the continuous monthly consumption measure.

Modified Two-Part Model (GLM estimation of conditional demand)

A modified two-part model that allows the error retransformation to be heteroscedastic in price is estimated. The first step will be identical to that of the traditional model. However, in the second step, a generalized linear model (GLM) with log-link and Gaussian distribution will be used to estimate the average monthly cigarette consumption by smoker equations.

Generalized linear models allow the expected value of the response variable to depend on a linear predictor through a nonlinear link function and allow the response probability distribution to be one of the distributions from the exponential family. In particular, generalized linear models take the form:

$$g[E(Y)] = X\beta\,,\; E \sim F$$

where $g[\,]$ is the link function and F is a distribution from the exponential family. The GLM employed in this paper assumes that the link function is a log-link for the expectation of the response variable (Y) given a set of covariates (X) such that $\ln(E(Y)\,|\,x)) = xB$.

Unlike taking the natural log of the dependent variable prior to estimation which yields fitted values and parameter estimates in terms of the log response, the GLM method internalizes the log transformation of the response within the model. The log link within the GLM model logs the linear predictor, XB, rather than the dependent variable to linearize the relationship between the dependent variable and covariates.

In order to choose the most appropriate exponential probability distribution to employ, one must specify a suitable variance function. Following Manning and Mullahy [2001], a modified Park test [Park, 1966] is employed to choose the most appropriate variance function. Unlike Park's original test, which was designed to test for heteroscedasticity for a specific variable, the modified Park test checks for a very specific type of heteroscedasticity, one in which the raw-scale variance is a power function of the raw-scale mean function. The modified Park test that is employed here uses ordinary least squares to estimate the following model using the residuals and predictions from a baseline GLM model with log-link and gamma distribution:

$$\ln\left(\left(y_i - \hat{y}_i\right)^2\right) = \lambda_0 + \lambda_1 \ln\left(\hat{y}_i\right) + v_i$$

The coefficient on the log of the raw-scale prediction, λ_1, will provide information as to which exponential probability distribution to employ in the second stage GLM estimation.[11] The modified Park test indicated that the raw-scale variance did not depend on the raw-scale prediction implying that the Gaussian distribution is a very good fit.

RESULTS

Estimates from the two part models of cigarette demand are presented in Table 2. Columns one and two, contain the estimates for the smoking participation equations. That is, columns one and two contain estimates for the first part of both the traditional and modified two part models. Columns three and four contain estimates when OLS is used on the natural logarithm of average monthly cigarette consumption by smokers. Columns five and six contain estimates when GLM with log-link and Gaussian distribution is used on the average monthly consumption by smokers. Given the potential for cross-border purchases, two alternative models are estimated for each econometric method. Columns one, three, and five contain models from a limited specification that includes the price of cigarettes and variables reflecting the respondent's gender, age, race, marital status, education, employment, family income, type of community, poverty status, and year surveyed. The models found in columns two, four, and six are identical to the models found in columns one, three, and five respectively, however, models two, four, and six add a variable reflecting the average price of cigarettes in all adjacent states.

The price of cigarettes has a negative and statistically significant impact on smoking participation and conditional cigarette demand using both the traditional OLS and modified GLM techniques in all model specifications. These estimates clearly show that increases in the own price of cigarettes would reduce both the number of people who smoke and the average number of cigarettes smoked by smokers. However, differences in the magnitude of the impact of price are observed between the traditional and modified two-part models as well as between models that control and do not control for potential border crossing. Table 3 contains the estimated price elasticities of demand. The estimates of the overall price elasticity of demand range from –0.541 to –1.092. The overall price elasticity estimates are smaller in absolute value for the modified two-part model than they are for the traditional two-part model. This suggests that the price elasticities of conditional demand are biased away from zero for the traditional two-part model due to the naïve assumption that the error retransformation is homoscedastic in price. In fact the magnitude of the bias is considerable, resulting in conditional demand elasticities using the traditional model to be more than twice those of the GLM model.The average price of cigarettes in adjacent states has a negative and significant impact on average consumption by smokers using both alternative methodologies.Similarly, the average price of cigarettes in adjacent states has a negative impact on smoking participation, however, it is not

TABLE 2
Cigarette Demand Equation Estimates

Independent Variables	Smoking Participation		Natural Logarithm Transformation of Conditional Use		GLM Model with Logarithm Link and Gaussian Distribution	
Real Price of Cigarettes	–0.282	–0.213	–0.531	–0.341	–0.265	–0.166
	(–3.80)	(–1.95)	(–5.39)	(–2.74)	(–3.61)	(–2.25)
Real Average Cigarette Price Adjacent States		–0.155		-0.465		–0.241
		(–0.97)		(–3.13)		(–2.47)
Male	0.203	0.202	0.174	0.174	0.210	0.210
	(7.62)	(7.58)	(5.47)	(5.44)	(13.05)	(12.93)
Age 18-24	0.887	0.889	-0.157	-0.157	-0.112	-0.113
	(15.41)	(15.41)	(-2.59)	(-2.61)	(-2.94)	(-2.98)
Age 25-44	1.027	1.027	0.126	0.126	0.119	0.119
	(23.77)	(23.74)	(2.51)	(2.53)	(3.35)	(3.33)
Age 45-64	0.813	0.811	0.277	0.277	0.210	0.210
	(28.73)	(28.77)	(6.11)	(6.15)	(7.57)	(7.54)
Black	-0.182	-0.184	-0.553	-0.559	-0.473	-0.475
	(-4.87)	(-4.90)	(-12.59)	(-12.65)	(-15.39)	(-15.49)
Other Race	-0.180	-0.193	-0.407	-0.430	-0.235	-0.235
	(-3.75)	(-3.94)	(-4.67)	(-5.04)	(-3.48)	(-3.32)
White Hispanic	-0.475	-0.478	-0.869	-0.873	-0.466	-0.469
	(-10.76)	(-10.75)	(-7.85)	(-7.93)	(-7.05)	(-7.20)
Black Hispanic	-0.513	-0.507	-0.762	-0.738	-0.529	-0.520
	(-4.46)	(-4.49)	(-4.33)	(-4.31)	(-6.66)	(-6.25)
Other Hispanic	-0.298	-0.297	-0.893	-0.898	-0.531	-0.525
	(-2.23)	(-2.20)	(-2.42)	(-2.45)	(-1.94)	(-1.94)
Married	0.057	0.057	0.142	0.142	0.065	0.064
	(2.36)	(2.35)	(4.34)	(4.31)	(2.75)	(2.72)
Separated, Divorced or Widowed	0.295	0.294	0.204	0.203	0.134	0.132
	(10.10)	(10.06)	(6.21)	(6.06)	(5.97)	(5.86)

TABLE 2—*Continued*

Cigarette Demand Equation Estimates

Independent Variables	Smoking Participation		Natural Logarithm Transformation of Conditional Use		GLM Model with Logarithm Link and Gaussian Distribution	
High School Graduate	-0.146	-0.145	-0.105	-0.104	-0.059	-0.059
	(-4.97)	(-4.94)	(-2.01)	(-1.99)	(-2.33)	(-2.28)
Some College	-0.486	-0.486	-0.349	-0.348	-0.181	-0.181
	(-18.59)	(-18.65)	(-9.80)	(-9.84)	(-6.85)	(-6.83)
Not Employed	0.243	0.235	0.147	0.140	0.080	0.078
	(6.89)	(6.66)	(2.48)	(2.34)	(2.39)	(2.29)
Not in Labor Force	0.025	0.023	0.031	0.030	0.040	0.038
	(0.97)	(0.91)	(0.89)	(0.83)	(2.02)	(1.94)
Real Family Income	-0.001	-0.001	-0.000	-0.000	-0.000	-0.000
	(-13.34)	(-13.32)	(-1.18)	(-1.20)	(-0.38)	(-0.39)
MSA Central City	-0.065	-0.064	0.022	0.021	0.017	0.017
	(-2.59)	(-2.56)	(0.60)	(0.57)	(0.81)	(0.80)
Not in MSA Non-Farm	-0.050	-0.049	0.015	0.017	0.021	0.022
	(-1.86)	(-1.83)	(0.42)	(0.48)	(0.99)	(1.02)
Not in MSA Farm	-0.431	-0.430	-0.016	-0.009	0.039	0.041
	(-5.96)	(-5.91)	(-0.12)	(-0.06)	(0.52)	(0.54)
Not under Poverty Threshold	0.040	0.038	0.078	0.076	0.048	0.047
	(1.40)	(1.33)	(1.80)	(1.75)	(1.89)	(1.80)
1993	0.003	0.007	0.080	0.096	0.056	0.063
	(0.15)	(0.45)	(3.24)	(3.56)	(4.33)	(4.50)
Constant	-0.749	-0.651	6.348	6.653	6.333	6.494
	(-7.66)	(-4.53)	(49.25)	(43.87)	(68.23)	(55.56)

All equations also include a missing value indicator for unknown real family income. Asymptotic t-ratios are in parentheses. The critical values for the t-ratios are 2.58 (2.33), 1.96 (1.64), 1.64 (1.28) at the 1, 5, and 10% significance levels, respectively, based on a two-tailed (one-tailed) test. Sample sizes for the smoking participation and conditional use equations are 40,943 and 10,164, respectively, implying that nonsmokers (smokers) are approximately 75.18% (24.82) of the total sample.

TABLE 3

Average Price Adjacent State Not Included

	Traditional Two-Part Model (OLS)	Modified Two-Part Model (GLM)
Participation Elasticity	-0.453	-0.453
	(0.120)	(0.120)
Conditional Demand Elasticity	-0.639	-0.318
	(0.119)	(0.088)
Total Elasticity	-1.092	-0.771

Average Price Adjacent State Included

	Traditional Two-Part Model (OLS)	Modified Two-Part Model (GLM)
Participation Elasticity	-0.341	-0.341
	(0.176)	(0.176)
Conditional Demand Elasticity	-0.410	-0.200
	(0.150)	(0.089)
Total Elasticity	-0.751	-0.541

Standard errors of elasticities are in parenthesis.

significant at conventional levels. These estimates indicate that smokers would reduce (increase) their average consumption of cigarettes if the average price of cigarettes in adjacent states increased (decreased). This implies that border crossing is an important determinant of the number of cigarettes smoked by smokers. In fact, if border crossing is not controlled for, the price elasticity of participation, OLS conditional use, and GLM conditional use are inflated by approximately 33 per cent, 55 per cent, and 59 per cent, respectively.

Briefly reviewing the estimates for the other independent variables controlling for all other covariates: Males are significantly more likely to smoke and smoke more on average that are females. With respect to race and ethnicity, Blacks, White Hispanics, Black Hispanics, other Hispanics, and other races are significantly less likely to smoke and, when they do, they smoke less on average than Caucasians. Individuals who are aged 18-24, 25-44, and 45-64 are significantly more likely to smoke than are individuals aged 65 years or older. In addition, individuals aged 25-44 and 45-64 smoke more on average than do individuals aged 65 years or older, however, individuals aged 18-24 smoke less on average than do individuals aged 65 years or older. Although age is used primarily as a control variable, caution should be used in interpreting the age specific results. In particular, given that smoking related disease is a function of both the intensity and duration of use, individuals aged 65 + (the reference category) are more likely to have died from smoking causes than are individuals of younger ages. Given the accumulation of addictive capital, it is quite possible that older cohorts of individuals would be smoking more than younger cohorts had they not died from smoking related disease.

With respect to marital status, individuals who are married or who are either widowed, separated, or divorced are significantly more likely to smoke and smoke more on average than those who are single. Individuals with at least a high school diploma are less likely to smoke and smoke fewer cigarettes than those who never completed high school. Individuals who are not employed are more likely to smoke cigarettes and consume more on average than those who are employed. No significant differences are observed in the probability of being a smoker between individuals who are not in the labor force and individuals who are working. However, when GLM is employed, smokers who are not in the labor force are found to consume significantly more cigarettes per month than are individuals who work.

Individuals with high family incomes are significantly less likely to smoke than are individuals with lower family incomes, however, no significant differences are observed between family income and average smoking. Individuals who live in the center of a metropolitan statistical area (MSA), on a farm not in an MSA, and not on a farm not in an MSA are found to be significantly less likely to smoke than are those individuals who live in an MSA, but not in the center of the city. No statistically significant differences are found between the type of community individuals reside and the amount smoked on average. Individuals who are not under the poverty threshold are statistically more likely to smoke and smoke more on average than those individuals who are under the poverty threshold, based on a one-tail test at at least the 10 per cent level. Finally, no statistically significant differences are observed

between smoking participation and the year surveyed, however, individuals surveyed in 1993 smoke more on average than those surveyed in 1994.

DISCUSSION

The results from this research imply that the price of cigarettes is a very important determinant of adult cigarette demand. The estimates imply that policies to increase the price of cigarettes will reduce not only the prevalence of smoking, but will also decrease the number of cigarettes smoked by smokers. However, the price coefficients obtained from the conditional demand equations are biased away from zero when using the traditional log-transformed dependent variable. The price elasticities are biased because the impact of price on $E(Y|X)$ which works through $\rho(x)$ is ignored. Empirically, the bias is quite substantial. The GLM model yielded conditional price elasticities less than one-half those obtained using the traditional method.

The precision of the estimates of price on conditional demand may have important policy implications. The results from this research implied that the traditional two-part model significantly overstated the negative impact of price on the conditional cigarette demand of adults.

NOTES

The author thanks three anonymous referees for very helpful comments. In addition, the author is grateful for advice offered by Willard Manning and research assistance provided by Victoria Meiners.

1. For a comprehensive review of economic studies see "The Economics of Smoking" in *The Handbook of Health Economics* (North-Holland, Elsevier Science, 2000) and the Surgeon General's reports (USDHHS, 1989, 1994, and 2000).
2. By special agreement, the National Center for Health Statistics (NCHS) has provided a restricted data set containing detailed information on cigarette use and identifiers for each respondent's state of residence only for individuals sampled in 1993 and 1994.
3. Persons in long-term care facilities, on active duty in the U.S. military, and U.S. nationals living in foreign countries are excluded from the surveys.
4. Approximately 7 per cent of observations have missing values on poverty status.
5. The complete case method deletes all observations with missing values for both income and poverty and estimates the cigarette demand equations using only observations in which there were no missing values.
6. In this approach, the missing values for income and poverty were imputed using their predicted values from regressions of income and poverty on all the other covariates in the model for which no missing values were present.
7. This approach creates separate dichotomous indicators to reflect whether or not observations have missing data on family income or poverty. These dichotomous indicators are used as regressors in the cigarette demand equations. When an observation has a missing value for income (poverty status), the dichotomous indicator for missing income (poverty status) takes on a value of 1 and the value of income is set to zero. When an observation does not have a missing value for income, the missing value indicator is set to zero and the income variable (poverty status) takes on its actual value.
8. To conserve space, the estimates from all the missing data and measurement error in covariate sensitivity analyses are not presented in this manuscript, but are available upon request. This manuscript presents the estimates from the models that employ the modified zero order regression missing value technique and an income variable that takes on the midpoint values from the possible income catego-

ries with the highest midpoint coded as $50,000. No statistically significant differences are observed with respect to the price elasticity estimates and any of the sensitivity analyses that were conducted. In addition, no statistically significant differences are observed with respect to the estimated income elasticities of demand and any of the sensitivity analyses that were conducted.

9. The Tobacco Institute stopped publishing the *Tax Burden on Tobacco* in 1999. Orzechowski and Walker have taken on the responsibility of publishing the *Tax Burden on Tobacco* post 1999.
10. Models that included the lowest price of cigarettes in an adjacent state were also estimated. No significant differences were found in the own-price elasticities of cigarette demand between these models and those presented in Table 2.
11. If the raw-scale variance does not depend on the raw-scale prediction, $\lambda_1=0$, then use a Gaussian distribution; if the raw-scale variance is proportional to the raw-scale prediction, $\lambda_1=1$, then use the poisson distribution; if the raw-scale variance is quadratic in the raw-scale prediction, $\lambda_1=2$, then use the gamma distribution; if the raw-scale variance is cubic in the raw-scale prediction, $\lambda_1=3$, then use the inverse Gaussian distribution.

REFERENCES

Becker, G. S., Grossman, M., and Murphy, K. M. An Empirical Analysis of Cigarette Addiction. *American Economic Review*, June 1994.

Belsley, D., Kuh, E., and Welsch, R. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley, 1980.

Blough, D. K., Madden, C. W., and Hornbrook, M. C. Modeling Risk Using Generalized Linear Models. *Journal of Health Economics*, April 1999.

Box, G. and Cox, D. An Analysis of Transformation. *Journal of the Royal Statistical Society, Series B,* 1964, 212-264.

Centers for Disease Control and Prevention's Tobacco Information and Prevention Sourcepage, Adult Prevalence Data. Centers for Disease Control and Prevention, 1999.

Chaloupka, F. J. Rational Addictive Behavior and Cigarette Smoking. *Journal of Political Economy*, August 1991.

Chaloupka, F. J., and Warner, K. E. *The Economics of Smoking*, in The Handbook of Health Economics, edited by Joseph P. Newhouse and Anthony J. Culyer. New York: North-Holland, Elsevier Science B.V., 2000, 1539-1627.

Coats, R. M. A Note on Estimating Cross-Border Effects of State Cigarette Taxes. *National Tax Journal*, December, 1995.

Cragg, J. G. Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica*, September 1971.

Davidson, R. and MacKinnon, J. Several Tests for Model Specification in the Presence of Alternative Hypotheses. *Econometrica*, May 1981.

DeCicca, P., Kenkel, D. and Mathios, A. Putting Out the Fires: Will Higher Taxes Reduce the Onset of Youth Smoking. *Journal of the Political Economy*, February 2002.

Duan, N. Smearing estimate: a nonparametric retransformation method. *Journal of the Ameridan Statistical Association,* September 1983.

Greene, W. H. *Econometric Analysis.* Englewood Cliffs, NJ: Prentice Hall, 1993.

Gruber, J. and Zinman, J. Youth Smoking in the United States: Evidence and Implications, *in Risky Behavior Among Youth*, edited by Jonathan Gruber. Chicago: The University of Chicago Press, 2001, 69-120.

Gruber, J. and Koszegi, B. Is Addiction "Rational"? Theory and Evidence. *The Quarterly Journal of Economics*, November 2001.

Johnston, L. D., O'Malley, P. M., and Bachman, J. G. Cigarette Smoking Among American Teens Continues Gradual Decline. University of Michigan News and Information Services, Ann Arbor, MI., December 1999.

_____. Cigarette Smoking Among American Teens Declines Sharply in 2001. University of Michigan News and Information Services. Ann Arbor, MI., December 2001.

Jones, A. M. *Health Econometrics*, in The Handbook of Health Economics, edited by Joseph P. Newhouse and Anthony J. Culyer. New York: North-Holland, Elsevier Science B.V., 2000, 265-344.

Lewit, E. M., and Coate, D. The Potential for Using Excise Taxes to Reduce Smoking. *Journal of Health Economics*. August 1982.

Manning, W. G. The Logged Dependent Variable, Heteroscedasticity, and the Retransformation Problem. *Journal of Health Economics*, June 1998.

Manning, W. G., and Mullahy, J. Estimating Log Models: To Transform or Not to Transform. *Journal of Health Economics*, July 2001.

Mullahy, J. Much Ado About Two: Reconsidering Retransformation and the Two-Part Model in Health Econometrics. *Journal of Health Economics*, June 1998.

Orzechowski and Walker. The Tax Burden on Tobacco, *Historic Compilation*. 35, 2000.

Park, R. Estimation with Heteroscedastic Error Terms. *Econometrica*, October, 1966.

Tobacco Institute. *The Tax Burden on Tobacco*. Washington, D.C.: Tobacco Institute, Historic Compilation, 34, 1999.

U.S. Department of Health and Human Services. *Reducing the Health Consequences of Smoking: 25 Years of Progress. A Report of the Surgeon General,* Rockville, MD: U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 1989.

_____. *Strategies to Control Tobacco Use in the United States: a Blueprint for Public Health Action in the 1990's.* Smoking and Tobacco Control Monographs. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute, 1991.

_____. *Preventing Tobacco Use Among Young People: A Report of the Surgeon General*, Atlanta GA: Public Health Service, Center for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 1994.

_____. *Reducing Tobacco Use: A Report of the Surgeon General*, Atlanta GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2000.

White, H. A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity. *Econometrica*, May 1980.