USING PENALIZED LIKELIHOOD TO SELECT PARAMETERS IN A RANDOM COEFFICIENTS MULTINOMIAL LOGIT MODEL

by

Joel L. Horowitz
Department of Economics
Northwestern University
Evanston, IL 60201
U.S.A

and

Lars Nesheim
CeMMAP
University College London
Institute for Fiscal Studies
London, U.K.

September 2019

Abstract

The multinomial logit model with random coefficients is widely used in applied research. This paper is concerned with estimating a random coefficients logit model in which the distribution of each coefficient is characterized by finitely many parameters. Some of these parameters may be zero or close to zero in a sense that is defined. We call these parameters small. The paper gives conditions under which with probability approaching 1 as the sample size approaches infinity, penalized maximum likelihood estimation (PMLE) with the adaptive LASSO (AL) penalty function distinguishes correctly between large and small parameters in a random-coefficients logit model. If one or more parameters are small, then PMLE with the AL penalty function reduces the asymptotic mean-square estimation error of any continuously differentiable function of the model's parameters, such as a market share, the value of travel time, or an elasticity. The paper describes a method for computing the PMLE of a random-coefficients logit model. It also presents the results of Monte Carlo experiments that illustrate the numerical performance of the PMLE. Finally, it presents the results of PMLE estimation of a random-coefficients logit model of choice among brands of butter and margarine in the British groceries market.

Key words: Penalized estimation, adaptive LASSO, random coefficients, logit model

JEL Codes: C13, C18, C25

Research carried out in part while Joel L. Horowitz was a visitor to the Department of Economics, University College London. We gratefully acknowledge financial support from the Economic and Social Research Council (ESRC) through the ESRC Centre for Micordata Methods and Practice (CeMMAP) grant number ES/1034021/1. Data were provided by TNS UK Ltd. The use of TNS UK Ltd. data in this research does not imply endorsement by TNS UK Ltd. of the interpretation or analysis of the data. All errors and omissions are the responsibility of the authors.

USING PENALIZED LIKELIHOOD TO SELECT PARAMETERS IN A RANDOM COEFFICIENTS A MULTINOMIAL LOGIT MODEL

1. INTRODUCTION

The multinomial logit model with random coefficients is widely used in demand modeling, empirical industrial organization, marketing, and transport economics. See, for example, Train (2009); Keane and Wasi (2013); and Ackerberg, Benkard, Berry, and Pakes (2007). Random coefficients enable taste or utility function parameters to vary among individuals in ways that are not explained by the variables available in the data. Random coefficients also enable the model to approximate any discrete-choice model arbitrarily well (McFadden and Train 2000). This paper is concerned with estimating a random coefficients model in which the distribution of each coefficient is characterized by finitely many parameters, for example the mean and variance. Some of these parameters may be zero or close to zero in a sense that will be defined. The paper describes a penalized likelihood method for selecting and estimating the non-zero parameters.

In applied research, the objects of interest in a discrete-choice model, such as market shares, the value of travel time, and elasticities, are smooth functions of the parameters. Some parameters, such as the mean coefficient of a price, may also be objects of interest. The mean square estimation errors of objects of interest can be reduced by identifying and dropping from the model parameters whose values are close but not necessarily equal to zero. We call these parameters "small." Thus, for example, if the mean and variance of the coefficient of a certain variable are both small, then the mean-square estimation errors of market shares and other objects of interest can be reduced by dropping that variable from the model. Parameters that are not small are called "large." In applications, it is not known *a priori* which parameters are large and small. This paper gives conditions under which penalized maximum likelihood estimation (PMLE) with the adaptive LASSO (AL) penalty function distinguishes correctly between large and small parameters asymptotically, thereby reducing the asymptotic mean-square estimation errors of large parameters and other objects of interest in applied research. We also show that the PMLE estimates of large parameters are $n^{-1/2}$-consistent and asymptotically normally distributed, where n is the size of the estimation sample. The estimates of the large parameters have the same asymptotic normal distribution that they would have if it were known *a priori* which parameters are large and small, the small parameters were set equal to zero, and the large parameters were estimated by maximum likelihood. This property is called oracle efficiency. We illustrate the numerical performance of our PMLE method with the results of Monte Carlo experiments and an empirical application to choice among brands of butter and margarine in the British groceries market.

Penalization can also have computational advantages. Penalized estimation with a suitable penalty function can yield parameter estimates that are true zeroes, often within a few iterations of the numerical algorithm. This is especially important in high-dimensional random coefficients models. Estimation of these models requires high-dimensional numerical integration. Dropping variance parameters that are zero or close to zero and treating the associated coefficients as fixed reduces the dimension of the integral as well as the dimension of the parameter vector, thereby increasing the speed of computation and the numerical accuracy with which the non-zero parameters are estimated. Kittel and Metaxoglu (2014) explore the numerical accuracy and consequences of numerical inaccuracy in estimation of random coefficients logit models.

This paper makes the following main contributions.

1. It shows that with probability approaching 1 as $n \to \infty$, PMLE with the AL penalty function distinguishes correctly between large and small parameters in a random-coefficients logit model. The estimates of the large parameters are oracle efficient.
2. It shows that if one or more parameters are small, then PMLE with the AL penalty function reduces the asymptotic mean-square estimation error of any continuously differentiable function of the model's parameters, including predicted market shares and elasticities.
3. It describes a method for computing the PMLE of a random-coefficients logit model with the AL penalty function.
4. It presents the results of Monte Carlo experiments that illustrate the numerical performance of the PMLE of a random-coefficients logit model with the AL penalty function.
5. It presents the results of PMLE estimation of a random-coefficients logit model of choice among brands of butter and margarine in the British groceries market.

Contributions 1 and 2 above extend results of Fan and Li (2001), Zou (2006), and Horowitz and Huang (2013) as well as the very large literature on penalized estimation of high-dimensional models. Fan, Lv, and Qi (2011), Horowitz (2015), and Bülmann and van de Geer (2011) review and provide references to that literature. Contribution 3 provides a new method to carry out PMLE computation that avoids the need for maximizing a non-smooth objective function and permits the use of recent advances in algorithms for solving constrained optimization problems.

The remainder of this paper is organized as follows. Section 2 describes the random-coefficients logit model that we consider, PMLE with the AL penalty function, asymptotic properties of the parameter estimates, and asymptotic properties of smooth functions of the PMLE parameter estimates. Section 3 describes our method for computing the PMLE parameter estimates. Section 4 presents the results of the Monte Carlo experiments. Section 5 presents the application to choice among brands of butter and

margarine, and Section 6 presents conclusions. Section 7 presents the proofs of this paper's theoretical results.

2. THE MODEL AND ADAPTIVE LASSO ESTIMATION

Section 2.1 describes the random-coefficients logit model and the penalized maximum likelihood estimation procedure that we use. Section 2.2 presents asymptotic distributional properties of the PMLE parameter estimates and functions of the estimates.

2.1 The Model and Estimation Procedure

Let each of n individuals choose among J exhaustive and mutually exclusive alternatives. Let $X \in \mathbb{R}^K$ denote the vector of the model's observed covariates, and let X_{ij} denote the value of X for individual i and alternative j $(j = 1,...,J)$. The indirect utility of alternative j to individual i $(i = 1,...,n)$ is

$$U_{ij} = (\beta' + \varepsilon_i')X_{ij} + \nu_{ij},$$

where ν_{ij} is a random variable with the Type I extreme value distribution, ν_{ij} and $\nu_{i'j'}$ are independent of one another if $i \neq i'$ or $j \neq j'$, β is a $K \times 1$ vector of constant coefficients, and ε_i is a $K \times 1$ vector of unobserved random variables that have means of 0 and are independently and identically distributed among individuals. In this paper, we assume that $\varepsilon_i \sim N(\boldsymbol{0}, \Sigma)$ for each $i = 1,...,n$, where $\boldsymbol{0}$ is a K-vector of zeroes and Σ is a positive-semidefinite $K \times K$ matrix. However, the paper's theoretical results hold with other distributions. Let $\phi(\xi; \boldsymbol{0}, \Sigma)$ denote the probability density function of the $N(\boldsymbol{0}, \Sigma)$ distribution evaluated at the point ξ. Then the probability that individual i chooses alternative j is

$$\pi_{ij}(\beta, \Sigma; X_{i1},...,X_{iJ}) = \int \left\{ \frac{\exp[(\beta' + \varepsilon')X_{ij}]}{\sum_{k=1}^{J} \exp[(\beta' + \varepsilon')X_{ik}]} \right\} \phi(\varepsilon; 0, \Sigma) d\varepsilon . \tag{2.1}$$

Let $\Sigma = CC'$ denote the Cholesky factorization of Σ, $\tilde{\varepsilon} \sim N(0_{K\times K}, I_{K\times K})$, and ϕ_K denote the $N(0_{K\times K}, I_{K\times K})$ probability density function. The standard Cholesky factorization applies to full rank matrices. However, when $rank(\Sigma) = r < K$, there is a unique Cholesky factorization with $K - r$ zeroes along the diagonal of C. Therefore (2.1) can be written as

(2.2) $$\pi_{ij}(\beta, C; X_{i1}, ..., X_{iJ}) = \int \left\{ \frac{\exp[(\beta' + \tilde{\varepsilon}' C') X_{ij}]}{\sum_{k=1}^{J} \exp[(\beta' + \tilde{\varepsilon}' C') X_{ik}]} \right\} \phi_K(\tilde{\varepsilon}) d\tilde{\varepsilon} .$$

The integral in (2.2) reduces to an r dimensional integral when $r < K$.

Define the choice indicator

$$d_{ij} = \begin{cases} 1 \text{ if individual } i \text{ chooses alternative } j \\ 0 \text{ otherwise} \end{cases}$$

Let $\{d_{ij}, X_{ij} : i = 1, ..., n; j = 1, ..., J\}$ be the observed choice indicators and covariates of an independent random sample of n individuals. Define $\theta = vec(\beta, C)$ and $L = \dim(\theta)$. The log-likelihood function for estimating θ is

$$\log L(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{J} d_{ij} \log \pi_{ij}(\theta; X_{i1}, ..., X_{iJ}) .$$

Define the maximum likelihood estimator

$$\bar{\theta} = \arg\max_{\theta} \log L(\theta) .$$

The penalized log-likelihood function that we treat here is

(2.3) $$\log L_P(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{J} d_{ij} \log \pi_{ij}(\theta; X_{i1}, ..., X_{iJ}) - \lambda_n \sum_{\ell=1}^{L} w_\ell |\theta_\ell| ,$$

where $\lambda_n > 0$ is a constant whose value may depend on n and the w_ℓ's are non-negative weights. Penalized maximum likelihood estimation with the adaptive LASSO penalty function consists of the following two steps.

Step 1: Let $\tilde{\theta}$ be a $n^{-1/2}$ -consistent estimator of θ_0 , possibly but not necessarily $\bar{\theta}$. Depending on how $\tilde{\theta}$ is obtained, some of its components may be zero. Define weights

$$\tilde{w}_\ell = \begin{cases} 1/|\tilde{\theta}_\ell| \text{ if } \tilde{\theta}_\ell \neq 0 \\ 0 \text{ if } \tilde{\theta}_\ell = 0. \end{cases}$$

Step 2: Let θ^* be a $L \times 1$ vector whose ℓ component is zero if $\tilde{\theta}_\ell = 0$ and whose remaining components are unspecified, Let $\pi_{ij}(\theta^*, X_{i1}, ..., X_{iJ})$ be the probability that individual i chooses alternative j when the parameter value is θ^* . The second-step penalized log-likelihood function is

(2.4) $$\log L_P(\theta^*) = \sum_{i=1}^{n} \sum_{j=1}^{J} d_{ij} \log \pi_{ij}(\theta^*; X_{i1}, ..., X_{iJ}) - \lambda_n \sum_{\ell=1}^{L} \tilde{w}_\ell |\theta_\ell^*| .$$

The second-step parameter estimator is

$$\hat{\theta} = \arg\max_{\theta^*} \log L_P(\theta^*),$$

where maximization is over the non-zero components of θ^*. Thus, $\hat{\theta}$ is obtained by setting any parameters estimated to be 0 in the first stage equal to 0 in the π_{ij}'s and penalty function, and maximizing the penalized log-likelihood function (2.4) over the remaining parameters. Asymptotic distributional properties of $\hat{\theta}$ and functions of $\hat{\theta}$ are described in Section 2.2.

2.2 Asymptotic Properties $\hat{\theta}$

This section describes asymptotic distributional properties of the second-step PMLE estimator $\hat{\theta}$ and smooth functions of $\hat{\theta}$. Let θ_0 denote the true but unknown value of θ. Make the high-level assumption

Assumption 1: (i) θ_0 is uniquely identified and (ii) $n^{1/2}(\overline{\theta} - \theta_0) \to^d N(0, \overline{\Omega})$ as $n \to \infty$, where $\overline{\Omega}$ is non-singular and equal to the inverse of the (non-singular) information matrix.

Amemiya (1985) among many others gives primitive conditions under which assumption 1 holds.

Let θ_{0k} denote the k'th component of θ_0. Any parameter θ_{0k} may be larger or smaller than the random sampling errors of the unpenalized MLE, which are $O_p(n^{-1/2})$. We represent this mathematically by allowing the components of θ_0 to depend on n. Call θ_{0k} small if $n^{1/2} |\theta_{0k}| \to 0$ as $n \to \infty$. Call θ_{0k} large if $|\theta_{0k}| > 0$.[1]

Assumption 2: All components of θ_0 are either large or small.

Assumption 2 precludes the possibility that some components of θ_0 are proportional to $n^{-1/2}$ asymptotically. Leeb and Pötscher (2005, 2006) explain why this is necessary. Let A_s denote the set of small parameters and A_0 denote the set of large parameters. Under assumption 2, A_0 is the complement of A_s. Let $\hat{\theta}_k$ denote the k component of $\hat{\theta}$.

Assumption 3: As $n \to \infty$, $\lambda_n \to \infty$ and $n^{-1/2}\lambda_n \to 0$.

Define $\theta_{A_0} = \{\theta_{k0} : \theta_{k0} \in A_0\}$, $\theta_{A_s} = \{\theta_{k0} : \theta_{k0} \in A_s\}$, $\hat{\theta}_{A_s} = \{\hat{\theta}_k : \theta_{k0} \in A_s\}$, and $\hat{\theta}_{A_0} = \{\hat{\theta}_k : \theta_{k0} \in A_0\}$. Let $\overline{\theta}_{A_0}$ be the unpenalized MLE of θ_{A_0} when $\theta_{A_s} = 0$ in model (2.1).

---

[1] θ_0 does not depend on n in the sampled population. Allowing some components of θ_0 to depend on n is a mathematical device that keeps these components smaller than random sampling error asymptotically as $n \to \infty$

Assumption 4: As $n \to \infty$, $n^{1/2}(\bar{\theta}_{A_0} - \theta_{A_0}) \to^d N(0, \bar{\Omega})$ for some covariance matrix $\bar{\Omega}$.

Primitive conditions for assumption 4 are the same as those for assumption 1.

For any function $g(\theta)$, let $AMSE[g(\hat{\theta})]$ and $AMSE[g(\bar{\theta})]$, respectively, denote the asymptotic mean-square errors of $g(\hat{\theta})$ and $g(\bar{\theta})$ as estimators of $g(\theta_0)$. The following theorem gives the main theoretical results of this paper.

Theorem 2.1: Let assumptions 1-4 hold. As $n \to \infty$

(i) $P(\hat{\theta}_k = 0 \; \forall \; k \text{ such that } \theta_{k0} \in A_s) \to 1$

(ii) $n^{1/2}(\hat{\theta}_{A_0} - \theta_{A_0}) \to^d N(0, \bar{\Omega})$

(iii) Let $g(\theta)$ be a continuously differentiable function of $\theta \in \mathbb{R}^K$. If A_s is non-empty, then $AMSE[g(\hat{\theta})] < AMSE[g(\bar{\theta})]$.

Parts (i) and (ii) of Theorem 2.1 state that PMLE estimation with the AL penalty function distinguishes correctly between large and small parameters with probability approaching 1 as $n \to \infty$. Part (ii) states that the PMLE estimates of the large parameters are oracle efficient. That is, they have the same asymptotic normal distribution that they would have if it were known which parameters in model (2.1) are large and small, the small parameters were set equal to zero, and the large parameters were estimated by maximum likelihood. Part(iii) states that if one or more parameters are small, then PMLE with the AL penalty function reduces the asymptotic mean-square estimation error of any continuously differentiable function of the model's parameters.

3. COMPUTATION

Maximizing $\log L_P(\theta)$ presents several computational problems. There may be more than one local maximum of $\log L_P(\theta)$, the penalty function in $\log L_P(\theta)$ is not differentiable at all values of θ, and $\log L_P(\theta)$ includes high-dimensional integrals that must be evaluated numerically. We deal with the first of these problems by maximizing $\log L_P(\theta)$ repeatedly using a different initial value of θ each time.

We deal with the second by reformulating the optimization problem to one of maximizing a differentiable objective function subject to linear constraints. To do this, write $\theta = \theta^+ - \theta^-$, where θ^+ and θ^- are $L \times 1$ vectors whose components are non-negative. Then maximizing $\log L_P(\theta)$ in (2.3) is equivalent to solving the problem

$$(3.1) \qquad \underset{\theta, \theta^+, \theta^-}{\text{maximize}}: \sum_{i=1}^{n}\sum_{j=1}^{J} d_{ij} \log \pi_{ij}(\theta; X_{i1}, ..., X_{iJ}) - \lambda_n \sum_{\ell=1}^{L} w_\ell(\theta_\ell^+ + \theta_\ell^-)$$

subject to

$$\theta = \theta^+ - \theta^-$$
$$\theta^+, \theta^- \geq 0,$$

where the last inequality holds component by component. This formulation avoids the need to maximize a non-smooth objective function and permits exploitation of advances in methods for solution of constrained optimization problems.

There is a large econometric literature on numerical methods for evaluating high-dimensional integrals. See, for example, McFadden (1989); McFadden and Ruud (1994); Geweke, Keane, and Runkle (1994); Hajivassiliou, McFadaden, and Ruud (1996); Geweke and Keane (2001), and Train (2009). Available methods include Gaussian integration procedures, pseudo Monte Carlo procedures, quasi Monte Carlo procedures, and Markov chain Monte Carlo (MCMC) methods. More recently, Heiss and Winschel (2008), Skrainka and Judd (2011), and Knittel and Metaxoglou (2014) have suggested that sparse grid integration methods produce accurate approximations at low cost. To focus on the performance of the PMLE method, we have used a simple pseudo Monte Carlo integration method based on either 500 or 1500 draws from a normal random number generator.

We computed the solution to problem (3.1) by using a sequential quadratic programming algorithm for constrained optimization from the NAG Fortran Library (The Numerical Algorithms Group, Oxford U.K., www.nag.com). The algorithm is based on NPOPT, which is part of the SNOPT package described by Gill, Murray, and Saunders (2005).

4. MONTE CARLO EXPERIMENTS

This section reports the results of a Monte Carlo investigation of the numerical performance of the PMLE method. We used two designs. One is based on a small, hypothetical model. The other is based on data from the U.K. market for butter and margarine.

4.1 Design 1: A Hypothetical Model

This design consists of a model with $J = 5$ alternatives in the choice set and $K = 20$ covariates. The random coefficients are independent of one another, so their covariance matrix is diagonal. The means and variances of the coefficients are as follows:

k	Mean (β_k)	Variance $Var(\varepsilon_k)$
$1 \leq k \leq 2$	1	1
$3 \leq k \leq 5$	1	0
$6 \leq k \leq 20$	0	0

Thus, there are two non-zero random coefficients, three non-zero coefficients that are not random, and 15 non-random coefficients whose values are zero. The covariates are independently distributed as $N(0,1)$. The sample size is $n = 1000$.

We carried out PMLE estimation with 300 simulated datasets and chose the penalty parameter λ_n to minimize the Bayes Information Criterion (BIC) using the computational procedure described in the next paragraph. Wang, Li, and Tsai (2007) and Wang, Li, and Leng (2009) have derived properties of the BIC for estimating the penalty parameter in penalized estimation of a linear model. The theoretical properties of the BIC for PMLE have not been studied. We used a pseudo Monte Carlo numerical integration procedure with antithetic variates with 500 draws from a 10-dimensional random number generator. We assumed that only 10 covariates, including the first 5, have potentially non-zero variances. Therefore, 30 parameters were estimated.

We chose λ_n by solving (2.3) for the two steps of the adaptive LASSO procedure using each point in a rectangular grid of λ_n values. There were 5 grid points for step one of the adaptive LASSO procedure, 10 points for step 2, and 50 points in total. The values of the step 1 points ranged from 10^{-4} to 10^{-3}. The values of the step 2 points ranged from 10^{-4} to 10^{-2}. The logarithms of the values in each dimension of the grid were equally spaced. We report results for the grid point of λ_n values that minimizes the BIC in step 2.

The results of the experiment are shown in Table 1. The average number of non-zero parameters in the model estimated by PMLE is 9.667, compared to 30 potentially non-zero parameters in the full model. With probability 1, unconstrained maximum likelihood estimation cannot yield estimates of zero, so unconstrained maximum likelihood estimation gives 30 non-zero parameter estimates. The mean-square errors (MSE's) of the PMLE estimates of the means of the non-zero slope coefficients (the non-zero β_k's) are less than half the MSE's of unconstrained maximum likelihood estimates. The MSE's of the PMLE estimates of the standard deviations are 90% of the MSE's of the unconstrained maximum likelihood estimates. In summary, PMLE selects a smaller model and gives estimates of important parameters with much smaller mean-square errors than does unconstrained maximum likelihood estimation.

4.2 Design 2: Butter and Margarine

This design is based on data about the UK market for butter and margarine. The data were obtained by the research company Kantar and used by Griffith, Nesheim, and O'Connell (2015). The data contain information on 10,102 households that shopped at supermarkets in the U.K. The data

include demographic characteristics of each household (e.g., household size, age, employment status, and average weekly grocery expenditure), product characteristics (e.g., brand, package size, and saturated fat content), and consumer purchase choices. On each shopping trip, each consumer chose either not to buy any product or to buy one of 47 products available in the market. Thus, the number of options in each consumer's choice set is $J = 48$.

The Kantar data contain $K = 50$ covariates, including product fixed effects. Thus, the choice model of equation (2.2) contains 99 parameters. There are 49 mean parameters (the components of β in (2.2)) and 50 variance parameters. The mean parameter for the outside option of no purchase is normalized to be zero. In the Monte Carlo experiment, we set the parameters equal to the penalized maximum likelihood estimates obtained from a random sample of 5000 observations from the Kantar data. The resulting model (the "true model") has 37 non-zero mean parameters and four non-zero random coefficient variance parameters. The remaining 58 parameters of the true model are zero. We used this model to simulate the product choices of 5000 hypothetical households. We used the simulated choice data to estimate the choice model's 99 parameters using unpenalized maximum likelihood (MLE), penalized maximum likelihood (PMLE), and the oracle MLE (maximum likelihood estimation of only the 41 non-zero parameters of the true model and the remaining parameters set equal to zero). We used 1500 antithetic variate draws from a multivariate normal random number generator to compute the numerical integral.

Table 2 summarizes results of 145 Monte Carlo replications of the foregoing simulation procedure. The number of replications was limited by the long computing time required for each replication. Columns 3-5 show the MSEs of the estimates of the non-zero parameters of the true model using each estimation method. The parameter β_1 is the mean price coefficient in the model. In all cases, the MSE of the PMLE is much smaller than that of the unpenalized MLE and close to the MSE of the oracle MLE. For example, the MSE of the PMLE of β_1 is 0.084 compared to 1.50 for the unpenalized MLE and 0.071 for the oracle MLE. The median number of non-zero parameters in the selected model is 37, and 80 percent of the replications select a model with 34-40 non-zero slope parameters. The slope of the price variable is non-zero in all replications.

We also computed the own-price elasticities of the 47 products (excluding no-purchase option) in each Monte Carlo replication. The MSE's of 37 of the 47 elasticity estimates obtained by PMLE were less than the MSE's of the corresponding elasticity estimates obtained by MLE. The median ratio of the MSEs of the MLE and PMLE elasticity estimates is 2.786. That is the median value of (MSE of MLE estimates)/(MSE of PMLE estimates) is 2.786. The median ratio of the MSEs of the PMLE and oracle MLE elasticity estimates is 1.021. Thus, the PMLE elasticity estimates, like the PMLE parameter

estimates, are more accurate than the estimates obtained from unpenalized MLE and close to the oracle estimates.

To illustrate the performance of PMLE in policy analysis, we used the PMLE, unpenalized MLE, and oracle estimates to evaluate effects of a 20% value added tax (VAT) on butter and margarine. Currently, food purchases in the UK are exempt from the VAT. The VAT increases the prices of butter and margarine, which reduces demand for these products, consumer welfare, and revenues from the sale of butter and margarine. We computed four resulting economic effects. The first is the reduction in consumer welfare as measured by the compensating income variation for the VAT. The second is the reduction in revenues to sellers of butter and margarine. The third is tax revenues resulting from the VAT. The fourth is the changes in the market shares of the products. We assumed that the pre-tax prices of butter, margarine, and any substitute products remain unchanged.

We now describe how we computed the foregoing effects. Let X_j^{notax} denote the values of the explanatory variables for product j in model (2.2) before the VAT and X_j^{tax} denote the values of the same variables after the prices of butter and margarine have been increased by 20%. Let p_j denote the before-VAT price of product j, τ denote the tax rate, and $p_j^{tax} = (1+\tau)p_j$ denote the price after the VAT has been imposed. Denote the mean and random component of the coefficient of price in (2.2) by β_1 and $\tilde{\varepsilon}_1 C_{11}$, respectively. The consumer compensating variation for the tax increase is (Small and Rosen 1981)

$$CV(\beta, C) = \sum_{i=1}^{5000} \int \frac{\log\left[\sum_{j=0}^{47} \exp(\beta + \tilde{\varepsilon}' C') X_{ij}^{notax}\right] - \log\left[\sum_{j=0}^{47} \exp(\beta + \tilde{\varepsilon}' C') X_{ij}^{tax}\right]}{\beta_1 + \tilde{\varepsilon}_1 C_{11}} \phi(\tilde{\varepsilon}) d\tilde{\varepsilon} .$$

The change in revenues is

$$\Delta R = \sum_{j=1}^{47} \sum_{i=1}^{5000} p_j [\pi_{ij}(\beta, \Sigma; X_j^{tax}) - \pi_{ij}(\beta, \Sigma; X_j^{notax})] .$$

The change in the market share of product j is

$$\Delta S_j = \sum_{i=1}^{5000} [\pi_{ij}(\beta, \Sigma; X_j^{tax}) - \pi_{ij}(\beta, \Sigma; X_j^{notax})] .$$

ΔR is the change the revenues of sellers after remitting tax revenues of τR^{tax} to the government and, therefore, does not include the factor $1+\tau$. The sums are over the 47 products and 5000 individuals in the experiment.

Table 3 shows the MSEs of the estimated effects of the VAT. The table shows the median MSEs of the estimated changes in market shares, not the MSEs of the estimated changes in the shares of

individual products. The MSEs of the unpenalized MLE and PMLE estimates of the compensating variation are similar. The MSEs of the PMLE estimates of the change in revenues to sellers (in pounds per trip per individual) and tax revenues are smaller than the MSEs of the unpenalized MLE estimates and closer to the oracle estimates. The median MSE of the PMLE estimates of the changes in market shares is smaller than the median MSE of the unpenalized MLSE estimates and close to the median MSE of the oracle estimate.

5. EMPIRICAL APPLICATION

This section summarizes the results of applying the PMLE and unpenalized MLE methods to the full Kantar data set that is described in the first paragraph of Section 4.2. We compare the own price elasticities obtained with the two methods and the results of the tax experiment described in Section 4.2. As is explained in the second paragraph of Section 4.2, the model has 99 parameters, including 49 means of the random slope coefficients and 50 standard deviations. All of the unpenalized parameter estimates are non-zero, and the empirical Hessian matrix has full rank. Only 34 of the penalized estimates are non-zero, including 30 slope coefficients and 4 standard deviation parameters.

Table 4 shows summary statistics for own price elasticities. The PMLE elasticity estimates are smaller in magnitude on average and less dispersed than the unpenalized MLE estimates. Figure 1 shows a plot of the PMLE elasticity estimates against the unpenalized MLE estimates along with the regression line obtained by using ordinary least squares (OLS) to estimate the model

(5.1) $\text{PMLE Estimate} = a + b\text{MLE Estimate} + U;\ E(U) = 0$.

The relation between the two sets of estimates appears to be scatter around a straight line. The slope of the line is $b = 0.4397$.

Table 5 shows summary statistics for changes in market shares and product revenues in (in units of pounds per shopping trip per individual) the tax experiment. The mean change in market share is zero because the sum of the shares must equal one. The PMLE estimates of the changes in market shares and revenues are less dispersed than the unpenalized MSE estimates. Figure 2 shows a plot of the PMLE estimates of changes in market shares against the unpenalized MLE estimates along with the regression line obtained by applying OLS to (5.1). The slope of the line is $b = 1.368$. Figure 3 shows a similar plot for changes in revenues. The slope of the line is 1.723. The PMLE and unpenalized MLE estimates of the compensating variation for the tax increase are 0.3405 and 0.3452 pounds per shopping trip, respectively. The PMLE and MLE estimates of tax revenue, respectively, are 0.3234 and 0.3301 pounds per shopping trip. The two methods give similar estimates of the compensating variation and tax revenues.

6. CONCLUSIONS

This paper has been concerned with estimating a random coefficients logit model in which the distribution of each coefficient is characterized by finitely many parameters. Some of these parameters may be zero or close to zero. We call such parameters "small." The paper has given conditions under which with probability approaching one as the sample size approaches infinity, penalized maximum likelihood estimation (PMLE) with the adaptive LASSO (AL) penalty function distinguishes correctly between large and small parameters in a random-coefficients logit model. The estimates of the large parameters are oracle efficient. If one or more parameters are small, then PMLE with the AL penalty function reduces the asymptotic mean-square estimation error of any continuously differentiable function of the model's parameters, such as a predicted market share. The paper has described a method for computing the PMLE of a random-coefficients logit model. It has presented the results of Monte Carlo experiments that illustrate the numerical performance of the PMLE. The paper has also presented the results of PMLE estimation of a random-coefficients logit model of choice among brands of butter and margarine in a British grocery chain.

The Monte Carlo results show that PMLE estimates have lower mean-square errors than unpenalized MLE estimates with sample sizes similar to those used in marketing and empirical industrial organization. PMLE estimation is tractable computationally, and the PMLE method can be modified easily for use in generalized method of moments estimation.

7. PROOF OF THEOREM 2.1

Parts (i) and (ii): Let I_{full} denote the information matrix of model (2.2). Let $\hat{\theta}_{A_0}$ and $\hat{\theta}_{A_s}$ denote the subvectors of $\hat{\theta}$ corresponding to θ_{A_0} and θ_{A_s}. Define the vector u_0 by

$$\theta = \theta_0 + n^{-1/2}u$$

for any θ. Let θ_{A_0} be the first L_0 components of θ_0 and θ_{A_s} be the remaining $L - L_0$ components. Order the components of u similarly. Define

$$D_n(u) = \log L_P(\theta_{A_0} + n^{-1/2}u) - \log L_P(\theta_0)$$
$$+ \lambda_n \left[\sum_{\ell=1}^{L_0} |\tilde{\theta}_\ell|^{-1} (|\theta_{0\ell} + n^{-1/2}u_\ell| - |\theta_{0\ell}|) + \sum_{\ell=L_0+1}^{L} w_\ell |n^{-1/2}u_\ell| \right]$$

$$\leq n^{-1/2} \frac{\partial \log L_P(\theta_0)}{\partial \theta'} u - (1/2) u' I_{full} u [1 + o_p(1)]$$

$$+ \lambda_n \left[\sum_{\ell=1}^{L_0} |\tilde{\theta}_\ell|^{-1} (|\theta_{0\ell} + n^{-1/2}u_\ell| - |\theta_{0\ell}|) + \sum_{\ell=L_0+1}^{L} w_\ell |n^{-1/2}u_\ell| \right].$$

Write the penalty term above as

$$n^{-1/2}\lambda_n \left[\sum_{\ell=1}^{L_0} |\tilde{\theta}_\ell|^{-1} n^{1/2} (|\theta_{0\ell} + n^{-1/2}u_\ell| - |\theta_{0\ell}|) + \sum_{\ell=L_0+1}^{L} w_\ell |u_\ell| \right]$$

Zou (2006, Theorem 2) shows that if $\theta_{\ell 0} \neq 0$, then

$$|\tilde{\theta}_\ell|^{-1} n^{1/2} (|\theta_{0\ell} + n^{-1/2}u_\ell| - |\theta_{0\ell}|) \rightarrow^p u_\ell \operatorname{sgn}(\theta_{0\ell}),$$

where $\operatorname{sgn}(v)$ for any scalar v equals 1, -1, or 0 according to whether v is positive, negative, or zero. Therefore, the terms of the penalty function corresponding to components of θ_{A_0} converge in probability to 0. Zou (2006) also shows that the terms in the penalty function corresponding to θ_{A_s} diverge to ∞. If the components of u corresponding to θ_{A_s} are non-zero, D_n is dominated by the penalty term, which increases without bound as $n \rightarrow \infty$. Arguments identical to those of Zou (2006, proof of Theorem 2) except with the least-squares objective function replaced by $-\log L(\theta)$, show that if $\theta_j \in A_s$, then $P(\hat{\theta}_j \in A_0) \rightarrow 0$. Therefore, D_n is dominated asymptotically by $\log L_P(\theta_{A_0} + n^{-1/2}u_0, 0) - \log L(\theta_{A_0}, 0)$, where u_0 denotes the components of u corresponding to components of θ_{A_0} and argument 0 corresponds to θ_{A_s}. Therefore, standard results for maximum likelihood estimates yield parts (i) and (ii). Q.E.D.

Part (iii): Let Ω_0 and $\bar{\Omega}$, respectively, be the covariance matrices of the asymptotic normal distributions of $n^{1/2}(\hat{\theta} - \theta_{A_0})$ and $n^{1/2}(\bar{\theta} - \theta_0)$. It follows from $\theta_{A_s} = o(n^{-1/2})$ and an application of the delta method that

$$AMSE[g(\hat{\theta})] = \frac{\partial g(\theta_0)}{\partial \theta'} \Omega_0 \frac{\partial g(\theta_0)}{\partial \theta}$$

and

$$AMSE[g(\hat{\theta})] = \frac{\partial g(\theta_0)}{\partial \theta'} \bar{\Omega} \frac{\partial g(\theta_0)}{\partial \theta}.$$

Therefore, it suffices to show that $\bar{\Omega} - \Omega_0$ is positive definite. Partition I_{full} as

$$I_{full} = \begin{pmatrix} I_{11} & I_{12} \\ I_{12}' & I_{22} \end{pmatrix},$$

where I_{11} is the submatrix of I_{full} corresponding to θ_{A_0}, I_{22} is the submatrix of components of I_{full} corresponding to components of θ_{A_s}, and I_{12} is the submatrix corresponding to the covariance of the estimators of θ_{A_0} and θ_{A_s}. Then

$$\bar{\Omega} = I_{full}^{-1} = I_{11}^{-1} + I_{11}^{-1} I_{12} [(I_{22} - I_{12}' I_{11}^{-1} I_{12})^{-1}] I_{12}' I_{11}^{-1} > I_{11}^{-1} = \Omega_0.$$

Q.E.D.

REFERENCES

Ackerberg, D.C., L. Benkard, S. Berry, and A. Pakes (2007). Econometric tools for analyzing market outcomes. *Handbook of Econometrics*, Vol. 6, ed. by E.E. Leamer and J.J. Heckman. Amsterdam: North-Holland, pp. 4171-4276.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data*. Heidelberg: Springer-Verlag.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.

Fan, J., J. Lv, and L. Qi (2011). Sparse high-dimensional models in economics. *Annual Review of Economics*, 3, 291-317.

Genz, A. and K.-S. Kwong (2000). Numerical evaluation of singular multivariate normal distributions. *Journal of Statistical Computation and Simulation*, 68, 1-21.

Geweke, J. and M. Keane (2001). Computationally intensive methods for integration in econometrics. *Handbook of Econometrics*, Vol. 5, ed. by E.E. Leamer and J.J. Heckman. Amsterdam: North-Holland, pp. 3463-3568.

Geweke, J., M. Keane, and D. Runkle (1994). Alternative computational approaches to inference in the multinomial probit model. *Review of Economics and Statistics*, 76, 609-632.

Gill, P.E., W. Murray, and M.A. Saunders (2005). Users' guide for SNOPT 7.1: a Fortran package for large-scale linear and nonlinear programming. Report NA 05-2, Department of Mathematics, University of California, San Diego. http://www.ccom.uscd.edu/~peg/papers/sndoc7.pdf.

Griffith, R. L. Nesheim, and M. O'Connell (2015). Income effects and the welfare consequences of tax in differentiated product oligopoly. Working paper, Institute for Fiscal Studies, London, U.K.

Hajivassiliou, V., McFadden, D. and P. Ruud (1996). Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results. *Journal of Econometrics*, 72, 85-134.

Heiss, F. and V. Winschel (2008). Likelihood approximation by numerical integration on sparse grids. *Journal of Econometrics*, 144, 62-80.

Horowitz, J.L. (2015). Variable Selection and Estimation in High-Dimensional Models. *Canadian Journal of Economics*, 48, 389-407.

Horowitz, J.L. and J. Huang (2013). Penalized Estimation of High-Dimensional Models under a Generalized Sparsity Condition. *Statistica Sinica*, 23, 725-748.

Keane, M. and N. Wasi (2013). Comparing alternative models of heterogeneity in consumer choice behavior. *Journal of Applied Econometrics*, 28, 1018-1045.

Knittel, C.R. and K. Metaxoglou (2014). Estimation of random-coefficent demand models: Two empiricists' perspective. *Review of Economics and Statistics*, 96, 34-59.

Leeb, H. and B.M. Pötscher (2005). Model selection and inference: facts and fiction. *Econometric Theory*, 21, 21-59.

Leeb, H. and B. Pötscher (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics*, 34, 2554-2591.

McFadden, D. 1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, 57, 1027-1057.

McFadden, D. and P.A. Ruud (1994). Estimation by simulation. *Review of Economics and Statistics*, 76, 591-608.

McFadden, D. and K. Train (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15, 447-470.

Skrainka, B.S. and K.L. Judd (2011). High performance quadrature rules: How numerical integration affects a popular model of product differentiation. Cemmap working paper CWP03/11, Institute for Fiscal Studies, London, U.K.

Train, K.E. (2009). *Discrete Choice Methods with Simulation*. Cambridge, U.K.: Cambridge University Press.

Wang, H., R. Li, and C.-L. Tsai (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94, 553-568.

Wang, H., B. Li, and C. Leng (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society, Series B*, 71, 671-683.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418-1429.

Table 1: Results of Monte Carlo Experiments with Design 1[a]

Parameter	MSE of PMLE Estimate	MSE of Unpenalized MLE Estimate
β_1	0.016	0.038
β_2	0.016	0.038
β_3	0.011	0.036
β_4	0.011	0.035
β_4	0.010	0.035
σ_1	1.722	1.926
σ_2	1.733	1.926
Average number of non-zero parameters in the model selected by PMLE	9.667	
Average value of λ in step 2	0.002	

a. Based on 300 Monte Carlo replications. σ_1 and σ_2, respectively, are the standard deviations of ε_1 and ε_2. The correct model is the model specified in design 1 with the parameter values specified in that design. The model selected by PMLE contains the correct model if the PMLE estimates of the non-zero parameters of the correct model are not zero.

Table 2: Results of Monte Carlo Experiments with Design 2[b]

Parameter	Definition of variable	MSE of PMLE Estimate	MSE of Unpenalized MLE Estimate	MSE of Oracle MLE
β_1	Price	0.08405	1.499	0.07094
β_2	Index of monthly advertising expenditure	0.002260	0.05067	0.003586
β_3	Square of index of monthly advertising expenditure	0.0006978	0.1094	0.00224
β_4	Dummy variable equal to 1 for 500 gram pack and 0 otherwise	0.7336	28.75	0.4579
β_5	Dummy variable equal to 1 for 1000 gram pack and 0 otherwise	4.314	48.92	3.111
β_6	Grams of saturated fat per pack	0.001699	0.02799	0.001329
β_7	Dummy variable equal to 1 if household size is 2 and makes no purchase and 0 otherwise	0.008787	1.576	0.06006
β_8	Dummy variable equal to 1 if household size is 3 and makes no purchase and 0 otherwise	0.1536	2.234	0.03704
β_9	Dummy variable equal to 1 if household size is 4 and makes no purchase and 0 otherwise	0.1535	0.9506	0.04824
β_{10}	Brand-specific constant	0.1350	1.241	0.07290
β_{11}	Brand-specific constant	0.3942	30.92	0.4635
β_{12}	Brand-specific constant	0.8753	79.88	0.6541
β_{13}	Brand-specific constant	2.527	130.4	1.554
β_{14}	Brand-specific constant	2.140	34.59	1.517
β_{15}	Brand-specific constant	0.7431	68.58	0.5094
β_{16}	Brand-specific constant	0.7388	13.89	0.3363
β_{17}	Brand-specific constant	0.2193	39.52	0.2519
β_{18}	Brand-specific constant	1.006	64.28	0.7423
β_{19}	Brand-specific constant	2.480	42.31	0.3746
β_{20}	Brand-specific constant	3.004	73.37	0.7231
β_{21}	Brand-specific constant	3.652	188.3	2.263
β_{22}	Brand-specific constant	2.966	79.01	1.533
β_{23}	Brand-specific constant	7.3899	122.5	3.946

Table 2, continued

β_{24}	Brand-specific constant	2.275	79.96	1.202
β_{25}	Brand-specific constant	0.8110	37.31	0.4154
β_{26}	Brand-specific constant	1.452	143.9	0.8864
β_{27}	Brand-specific constant	0.1767	53.87	0.1712
β_{28}	Brand-specific constant	0.1901	34.61	0.2385
β_{29}	Brand-specific constant	0.5416	20.20	0.5505
β_{30}	Brand-specific constant	0.2073	48.99	0.2453
β_{31}	Brand-specific constant	0.3410	58.10	0.3325
β_{32}	Brand-specific constant	1.184	70.28	0.6190
β_{33}	Brand-specific constant	0.1670	112.8	0.1521
β_{34}	Brand-specific constant	1.176	100.1	0.7542
β_{35}	Brand-specific constant	0.9511	151.8	0.6271
β_{36}	Brand-specific constant	1.331	132.0	0.8442
β_{37}	Brand-specific constant	0.2824	120.0	0.1681
σ_1	Standard deviation of coefficient of price	0.1537	1.258	0.1311
σ_6	Standard deviation of coefficient of saturated fat per pack	0.06436	1.968	0.005400
σ_{23}	Standard deviation of coefficient of a brand-specific constant	8.997	48.18	4.494
σ_{38}	Standard deviation of utility of no-purchase option for households of at least 5 persons	17.23	33.13	12.34
Average number of non-zero parameters in the model selected by PMLE		36.90		
Average value of λ in step 2		0.002611		

Based on 145 Monte Carlo replications.

Table 3: Mean Square Errors of Estimated Effects of the VAT in Monte Carlo Design 2

Effect	MSE Using MLE	MSE Using PMLE	MSE Using Oracle Model
Compensating Variation	0.0147	0.0151	0.00980
Change in Revenues to Sellers	0.0281	0.00793	0.00595
Tax Revenues	0.0186	0.0171	0.0110
Median MSE of Changes Market Share	4.71×10^{-7}	1.88×10^{-7}	1.68×10^{-7}

TABLE 4: SUMMARY STATISTICS FOR OWN PRICE ELASTICITIES

Method	Mean Elasticity	Standard Deviation of Elasticity	Maximum	Minimum
MLE	-2.811	0.9816	-1.320	-4.611
PMLE	-2.450	0.6412	-0.851	-4.091

TABLE 5: SUMMARY STATISTICS FOR CHANGES IN MARKET SHARES AND PRODUCT REVENUE

Method	Standard Deviation of Change in Share $(\times10^{-3})$	Mean Change in Revenue $\times10^{-3}$	Standard Deviation of Change in Revenue $\times10^{-3}$
MLE	4.896	-3.543	4.529
PMLE	3.541	-4.100	5.847

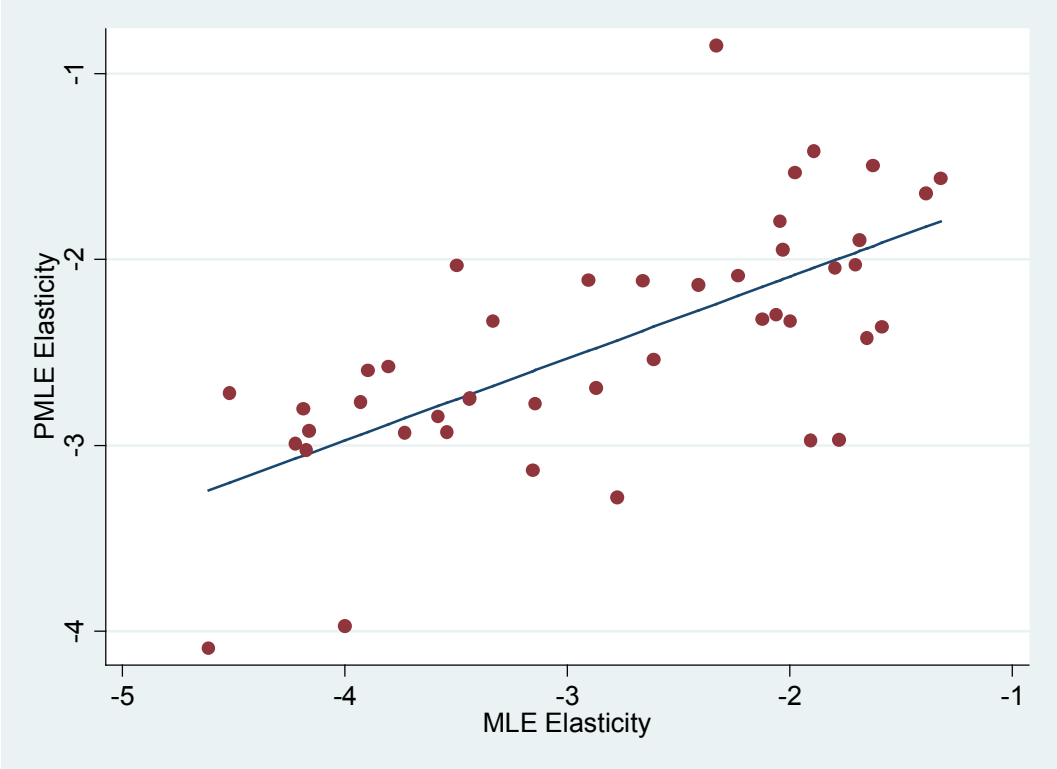

Figure 1: PMLE and unpenalized MLE estimates of own price elasticities

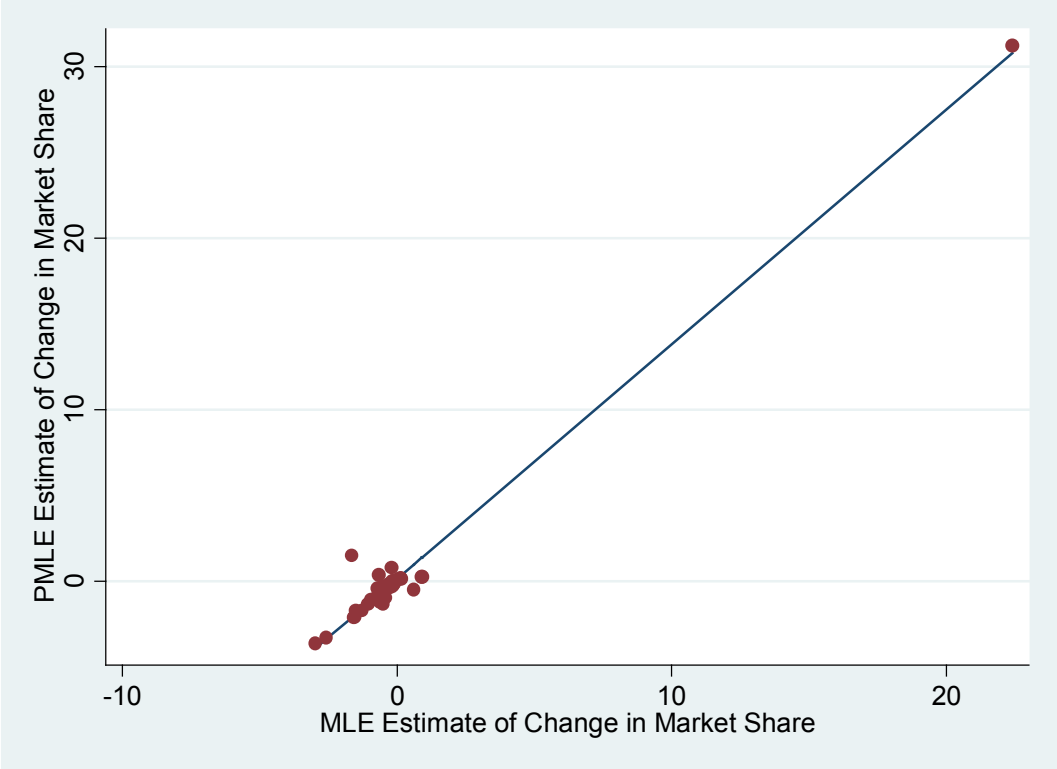

Figure 2: PMLE and unpenalized MLE estimates of changes in market shares

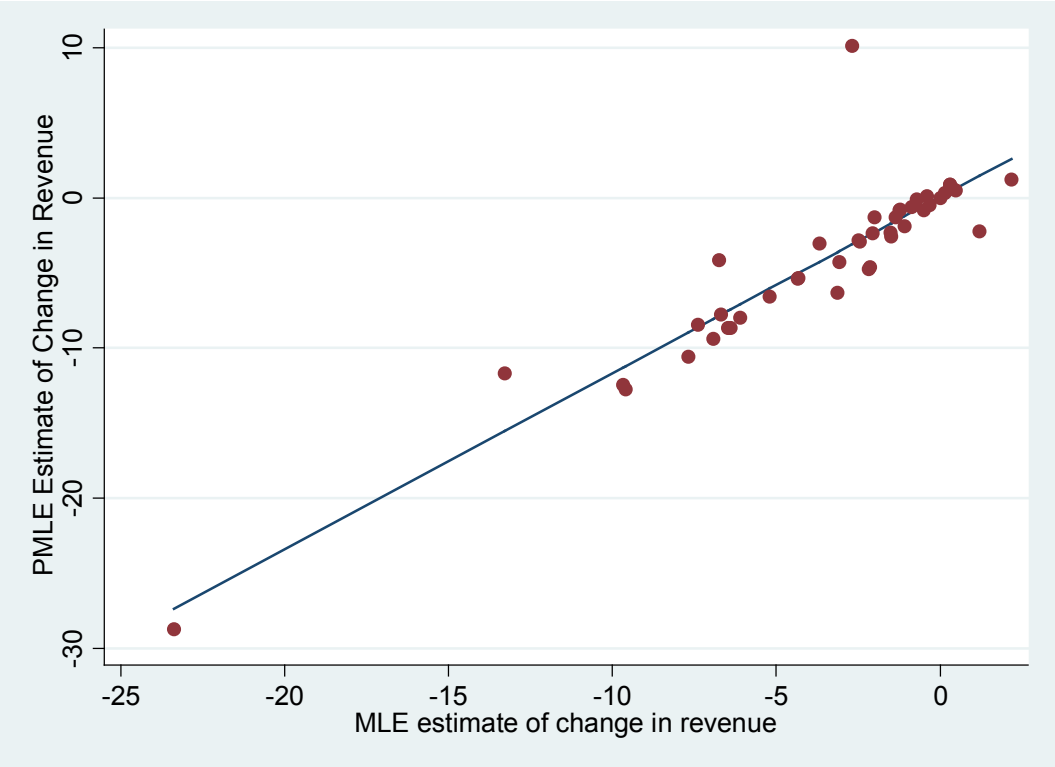

Figure 3: PMLE and unpenalized MLE estimates of changes in revenue per trip per individual