

(Random Utility) Discrete Choice

Suppose that there are 2 ways to fly from OC to New York. Route 1 is nonstop and expensive and Route 2 stops in Chicago and is cheaper. If consumers have utility given by:

$$U_{ij} = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{Cost}_{ij} + v_{ij}, \left(v_{ij} \sim iidN(0, \sigma^2) \right)$$

The consumer will choose the nonstop flight if:

$$U_{i1} - U_{i2} = \beta_1 (\text{Time}_{i1} - \text{Time}_{i2}) + \beta_2 (\text{Cost}_{i1} - \text{Cost}_{i2}) + v_{i1} - v_{i2} > 0$$

If we observe a sample of consumers and code $Y_i=1$ if i takes the nonstop flight, then we can estimate the parameters using a probit model.

Variance Normalization

Probit model we estimate is given by:

$$\Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1\Delta time + \beta_2\Delta cost)$$

Since Φ is the cdf of the standard Normal, this implies that the variance of $(v_{i1} - v_{i2}) = 1$

The model is equivalent to: $Y_i=1$ if

$$\frac{(U_{i1} - U_{i2})}{\sqrt{2\sigma^2}} = \frac{\beta_1}{\sqrt{2\sigma^2}}(Time_{i1} - Time_{i2}) + \frac{\beta_2}{\sqrt{2\sigma^2}}(Cost_{i1} - Cost_{i2}) +$$

$$\frac{v_{i1} - v_{i2}}{\sqrt{2\sigma^2}} > 0$$

This is unchanged if we multiply standard deviation and slope parameters by the same constant!

Heteroskedastic Probit

Suppose that the variance of the error terms for business and pleasure travelers are different. Then for the business travelers we are estimating:

$$\frac{(U_{i1} - U_{i2})}{\sqrt{2\sigma_b^2}} = \frac{\beta_1}{\sqrt{2\sigma_b^2}} (\Delta Time_i) + \frac{\beta_2}{\sqrt{2\sigma_b^2}} (\Delta Cost_i) + \frac{u_{i1} - u_{i2}}{\sqrt{2\sigma_b^2}}$$

and for pleasure travelers we are estimating:

$$\frac{(U_{i1} - U_{i2})}{\sqrt{2\sigma_p^2}} = \frac{\beta_1}{\sqrt{2\sigma_p^2}} (\Delta Time_i) + \frac{\beta_2}{\sqrt{2\sigma_p^2}} (\Delta Cost_i) + \frac{u_{i1} - u_{i2}}{\sqrt{2\sigma_p^2}}$$

Heteroskedastic Probit 2

This is equivalent to estimating the following model on the entire data:

$$P(Y_i = 1) = \Phi \left[\beta_{1b} \Delta Time_{ib} + \beta_{1p} \Delta Time_{ip} + \beta_{2b} \Delta Cost_{ib} + \beta_{2p} \Delta Cost_{ip} \right]$$

Subject to the constraint that

$$\frac{\beta_{1b}}{\beta_{1p}} = \frac{\beta_{2b}}{\beta_{2p}} \quad \text{and where } \Delta Time_{ib} \text{ is the time difference for}$$

business travelers and = 0 otherwise

Marginal Effects for Probit

In the linear model $Y_i = \beta_0 + \beta_1 X_i + \nu_i$ we can interpret β_1 as the marginal impact of X on $E(Y_i | X_i) = \frac{\partial(\beta_0 + \beta_1 X_i)}{\partial X_i}$ which doesn't depend on X or β_0 . But in the Probit model:

$$\frac{\partial E(Y_i | X_i)}{\partial X_i} = \frac{\partial \Phi(\beta_0 + \beta_1 X_i)}{\partial X_i} = \phi(\beta_0 + \beta_1 X_i) \beta_1$$

Note that this depends on X and β_0 and takes its maximum value where $\beta_0 + \beta_1 X_i = 0$. Note that:

$$E\left(\phi(\beta_0 + \beta_1 X_i) \beta_1\right) \neq \phi(\beta_0 + \beta_1 E(X_i)) \beta_1$$

Marginal Effects for logit

Recall that for the Logit Model the probability of $Y=1$:

$$F(\beta_0 + \beta_1 X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

The marginal effect is given by:

$$\begin{aligned} \frac{\partial E(Y_i | X_i)}{\partial X_i} &= \frac{\partial \left(\frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \right)}{\partial X_i} = \frac{e^{-(\beta_0 + \beta_1 X)}}{\left(1 + e^{-(\beta_0 + \beta_1 X)}\right)^2} \beta_1 \\ &= F(\beta_0 + \beta_1 X)(1 - F(\beta_0 + \beta_1 X)) \beta_1 \end{aligned}$$

Note that this also depends on X and β_0 and takes its maximum value where $\beta_0 + \beta_1 X_i = 0$.

Multinomial (or Conditional) Logit

Suppose that household i chooses one out of a set of vehicles j ($j = 1, \dots, J$):

$$U_{ij} = V_{ij}(\beta, x_{ij}) + \varepsilon_{ij}, (\varepsilon_{ij} \sim iid \text{ Extreme Value})$$

The probability that household i choose alternative j is

$$P(U_{ij} - U_{ik}) \geq 0 \text{ for all } j \neq k$$

If $V_{ij} = \beta x_{ij}$ then scale of parameters is not identified.

iid assumption is strong and implies Independence from Irrelevant Alternatives.

Advantages of MNL

1. Likelihood function is globally concave so maximum likelihood is easy for even large models.
2. IIA property allows sampling of alternatives in situations with very large number of discrete alternatives (for example, multi-car household vehicle choice).
3. Prediction and welfare calculations are straightforward and do not require “predicting” random utility components.

Non IIA Multinomial Discrete Choice

If ε_{ij} follows a GEV distribution, then choice model no longer has the IIA property. Most common GEV model is Nested Logit. Partitions choice set into subgroups with common unobserved errors, but choice within subgroups is MNL. Nested Logit can be sequentially estimated for large models using MNL code.

Problems:

1. Not clear how to set subgroups
2. Parameters governing correlation (coefficients of inclusive value) are subject to frequently binding inequality constraints.
3. Likelihood function is badly behaved as function of inclusive value coefficients.

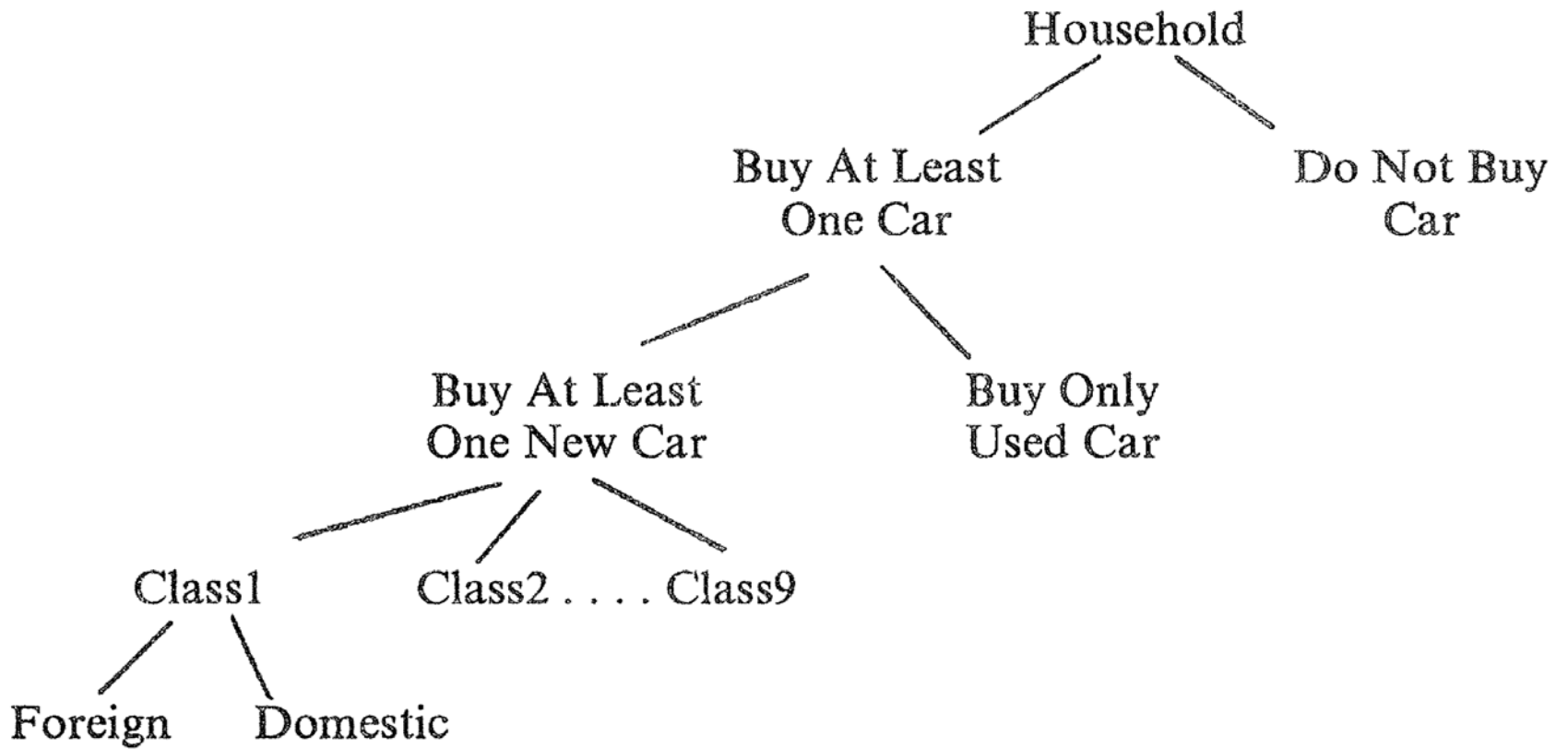


Figure 1
Automobile Choice Model

Mixed logit: General distribution for η and extreme value for ε

$$U_i = \beta'x_i + [\eta_i + \varepsilon_i]$$

$U = \beta'X + [\eta + \varepsilon]$ where $V(\varepsilon) = \alpha I$ with known (i.e., normalized) α and $V(\eta)$ is general

Density of η is $f(\eta|\Omega)$, where Ω are the fixed parameters of the distribution.

Given η , the conditional choice probability is simply logit:

$$L_i(\eta) = \exp(\beta'x_i + \eta_i) / \sum_j \exp(\beta'x_j + \eta_j)$$

the unconditional choice probability is:

$$P_i = \int L_i(\eta) f(\eta|\Omega) d\eta$$

this is approximated by:

$$SP_i = (1/R) \sum_{r=1, \dots, R} L_i(\eta^r)$$

Estimate by Max simulated log-likelihood function:

$$\sum_n \ln(SP_{ni})$$

- SP_i is an unbiased estimate of P_i for any R
- Variance decreases as R increases
- It is strictly positive for any R , such that $\ln(SP_i)$ is always defined
- It is smooth (i.e., twice differentiable)
- The simulated probabilities sum to one over alternatives, which is useful in forecasting.
- McFadden and Train have shown that any random utility model can be approximated by mixed logit

LM tests for Mixed Logit

$$z_{in} = (x_{in} - x_{Cn})^2$$

$$x_{Cn} = \sum_j x_{jn} P_{jn}$$

Hypothesis of no random coefficients on attribute x is rejected if coefficient of z significantly different from zero

Mixed Logit Problems

1. Mixed Logit likelihood can be very badly behaved so that most applications use independent error components.
2. Estimates may be sensitive to simulation methods and number of draws.
3. Identification can be very tricky (see M. Ben-Akiva and J. Walker).
4. If the model is used for forecasting, there is the problem of forecasting the random parameters for new observations and/or alternatives.

Bottom line – use Mixed Logit LM tests as a specification test for MNL – try hard to find a good MNL specification!

Bayesian Discrete Choice

- Provide a principled approach for incorporating non-sample information
- Provide finite sample inference
- Easy handling of model uncertainty
- Parameters are random variables instead of fixed constants

Prior distribution $\pi(\theta)$ Likelihood function :

$f(x | \theta)$ Observe data and get posterior distribution:

$$p(\theta | x) = \frac{f(x | \theta)\pi(\theta)}{\int f(x | \theta)\pi(\theta)d(\theta)}$$

In many cases the posterior mean is the optimal estimator. The key problem is computing high-dimensional integrals.

- Bayes confidence intervals
 - fixed regions containing θ with specified coverage probability
 - conditions on observed data
- Classical confidence intervals
 - region with random endpoints containing true θ over independent repeated replications of the data
 - depends on distribution of unobserved realizations of the data

Bayesian Model Uncertainty

Suppose there are M competing models, and let

π_m Be the prior probability that model m is correct

$$f_m(x) = \int f_m(x | \theta) p_m(\theta) d\theta \quad \text{Marginal density}$$

$$\bar{\pi}_m = \frac{\pi_m f_m(x)}{\sum_{j=1}^M \pi_j f_j(x)} \quad \text{Posterior probability that model } m \text{ is correct}$$

Unless there is a clear correct model, it is better to average over competing models :

$$\bar{p}(\theta | x) = \sum_{j=1}^M \bar{\pi}_j p_j(\theta | x)$$

The competing models do not need to be nested, and each model can be analyzed separately.

This only makes sense if you are averaging something with the same meaning (marginal effect or policy simulation!)

Bayesian computation:

Posterior distributions for discrete choice models (other than conditional logit) cannot be computed in closed form.

Use Markov Chain Monte Carlo methods to simulate draws from posterior distributions. The output files from these simulations can be saved and used to generate predictions. They can be reweighted to approximate the impact of changing prior distributions.

Bayesian Mixed Probit (Allenby and Rossi, 99)

$$p(\eta_i, \beta, \theta | x) \propto f(x | \eta_i, \beta) \pi(\eta_i | \theta) \pi(\theta) \pi(\beta)$$

f is likelihood for multinomial probit, η_i

is random effect for observation i . θ are the parameters of the distribution of the random effects over the sample.

This Bayesian formulation permits inference for the individual effects (repeated SP experiments). Revelt and Train (1999) also give classical methods.

Poorly Identified Models:

Likelihood function will be almost flat, so classical methods will have trouble converging to the optimum.

If proper prior distributions are used, Bayesian methods will have no computational problems, but posterior will look like prior for poorly identified parameters. There are no problems for inference on well-identified parameters.

Measurement Error:

Discrete choice applications in transportation are plagued by serious measurement errors. It is very difficult to directly observe key variables for unchosen alternatives, so it is common practice to impute travel times and costs from network models. Unfortunately this practice yields inconsistent parameter estimates and overstates the precision of these estimates.

Multiple Imputations:

General method for consistent inference using imputed values for missing or erroneous observations.

If the imputed values are somehow produced to match the first two moments of the correct unobserved values, then standard estimation methods that treat the imputed values as if they are correct will yield consistent parameter estimates. Unfortunately the standard errors produced by this approach will be inconsistent and downward biased because they ignore the errors introduced by the imputation process.

Rubin (1987) proposed solving this problem by independently drawing multiple imputed values. The component of variance due to the imputation error is then estimated by the variability of the estimates across the different imputed data sets.

If no data are missing, we use estimator:

$\tilde{\theta}$ and covariance estimator $\tilde{\Omega}$

Draw m independent imputations and compute corresponding parameter and covariance estimators:

$\tilde{\theta}_j$ and $\tilde{\Omega}_j$ Final estimates are given by:

$$\hat{\theta} = \sum_{j=1}^m \tilde{\theta}_j / m \quad \hat{\Sigma} = U + \left(1 + m^{-1}\right)B,$$

$$B = \sum_{j=1}^m (\tilde{\theta}_j - \hat{\theta})(\tilde{\theta}_j - \hat{\theta})' / (m - 1)$$

$$U = \sum_{j=1}^m \tilde{\Omega}_j / m.$$

These final estimates are consistent for any fixed number of imputations, and they only require estimation of the model where all data are observed without error.

Multiple imputations can be drawn once and stored so they can be used for estimating different models.

Proper multiple imputations:

Draw from the Bayesian posterior predictive distribution of the missing values under a specified model.

Any proper imputation procedure must condition on all observed data, and different sets of imputed values must be drawn independently so that they reflect all sources of uncertainty in the response process.

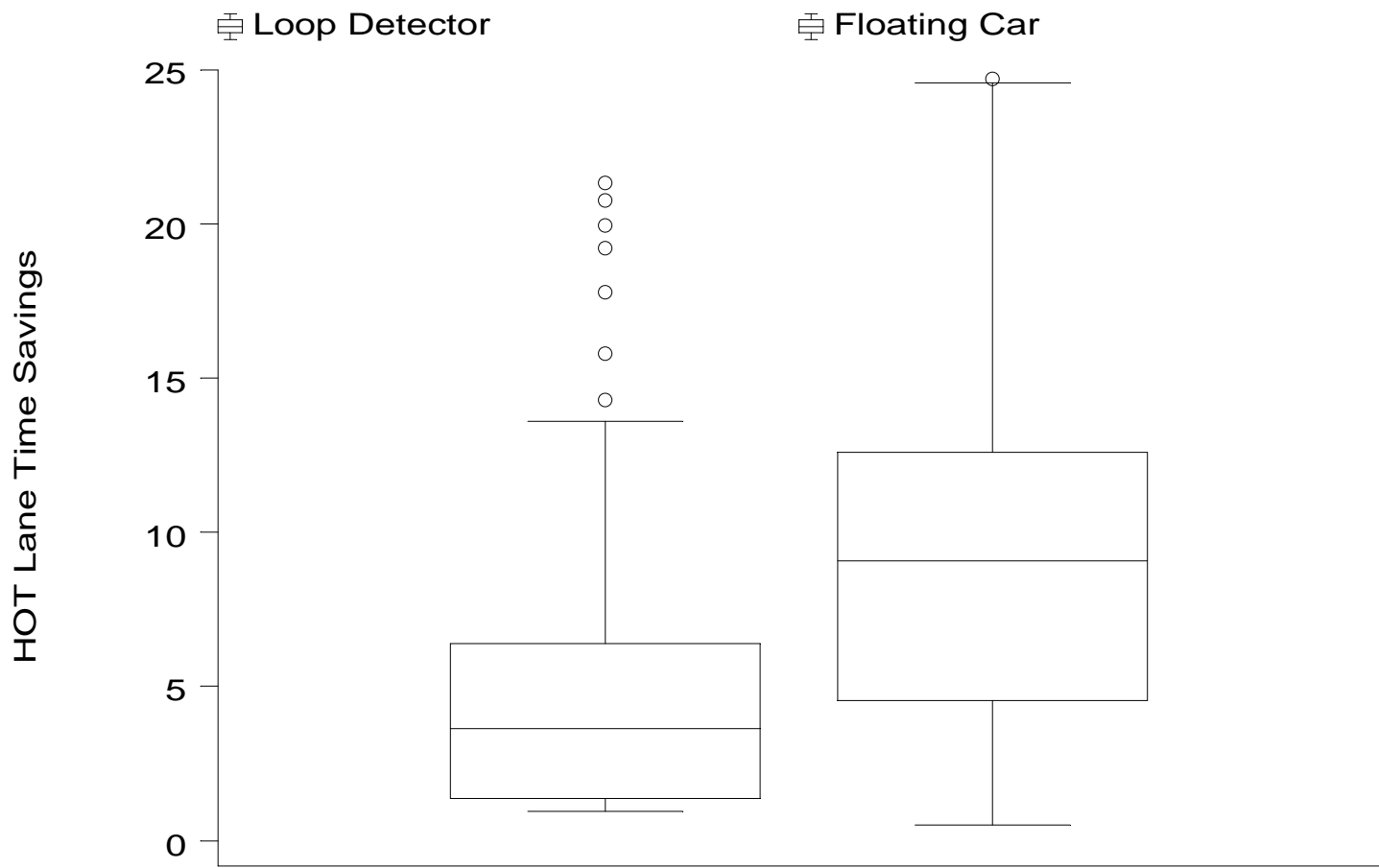
Ex. San Diego Congestion Pricing Experiment

Solo drivers can pay to use an eight-mile stretch of reversible high occupancy vehicle (HOV) lanes along Interstate Route 15 north of San Diego, California.

Per-trip fee for solo drivers is posted on changeable message signs upstream from the entrance to the lanes, and may be adjusted every six minutes to maintain free-flowing traffic conditions in the HOT lanes.

Carpoolers use HOT lanes for no charge.

Model mode choice as a function of cost and time savings



Use 5 days with both loop detector and floating car data to estimate an imputation model to predict missing floating car data. Then use multiple imputations to account for errors in imputation when estimating value of time from a conditional logit mode choice model.

First transform time savings to bound between 0 and 35 minutes:

$$\log\left(\left(\frac{t}{35}\right) / \left(1 - \frac{t}{35}\right)\right)$$

Dependent Variable: Logit of Floating Car Time Savings

$R^2 = 0.90$

Root MSE = 0.36

Independent Variables:	Coef.	Std. Err.	<i>t</i> -Stat.
Logit of Loop Detector Time Savings × Minutes Past 5:00 A.M.	0.0029	0.00031	9.3
Minutes Past 5:00 A.M.	0.222	0.0149	14.8
(Minutes Past 5:00 A.M.) ²	-0.00138	0.000121	-11.4
(Minutes Past 5:00 A.M.) ³	2.73E-06	2.91E-07	9.38
Toll	-0.229	0.188	-1.22
Toll × Minutes Past 5:00 A.M.	0.00222	0.00126	1.77
Constant	-11.4	0.52	-22.1

Implied value of time saved from mode choice model

Value of Time (\$/hour)	Corrected	Loop Data
95 th Percentile	108.70	105.60
90 th Percentile	72.12	73.63
75 th Percentile	31.30	35.27
50 th Percentile	18.71	23.37
25 th Percentile	10.30	16.55
10 th Percentile	-20.72	14.43
5 th Percentile	-83.02	14.08
Mean	25.63	32.64

The Impact of Residential Density on Vehicle Usage and Energy Consumption

David Brownstone and Tom Golob
University of California Irvine

dbrownst@uci.edu

<http://www.economics.uci.edu/~dbrownst/JUESprawlV3final.pdf>

How does residential density affect travel?

- How can we quantify the impacts of urban land use density for assessing impacts of “urban sprawl” and for evaluating densification policies
- Conventional measures:
 - household car ownership
 - and, for all household members:
 - trip generation (under-reporting of walk trips?)
 - trip distances
 - mode choices

Potentially more useful measures

- 1. Total travel distance by all household vehicles
 - captures: car ownership, trip generation, mode choices, and trip distances
- 2. Total fuel usage on all vehicles
 - captures vehicle type choice and implicit choice of fleet fuel efficiency

Must control for selectivity biases

- Different types of households choose to live in neighborhoods of varying densities
 - long list of potentially relevant demographic and socio-economic variables
- Persons choosing different lifestyles also choose to live in neighborhoods of varying densities
 - may not be fully captured by demographic and socio-economic variables
- These household effects influence travel simultaneously with density

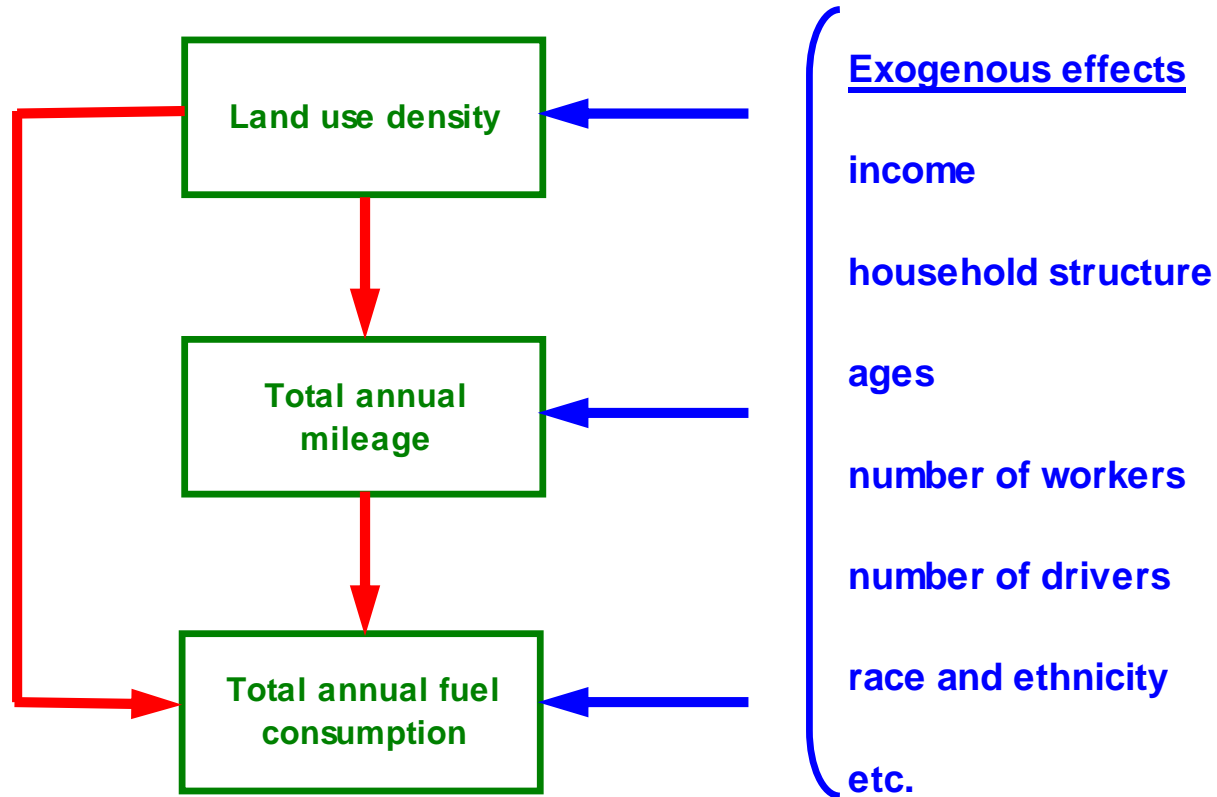
Previous Literature

- Newman and Kenworthy, 1999
- Bento et al. (2005)
- Boarnet and Sarmiento (1998)
- Bhat and Guo (2007)
- Ewing and Cervero, 2001, and Badoe and Miller, 2000, give literature reviews.

Our approach to the problem

- Make choice of residential density endogenous
- Simultaneous equations with two sets of endogenous variables
 - residential density
 - annual mileage and fuel consumption
- Both sets explained by demographic and socio-economic variables
- The residential density variable affects the two travel variables
- Estimate as a simultaneous system

Simultaneous system: 3 endogenous variables



2001 U.S. National Household Transportation Survey (NHTS)

- annual mileage for all household vehicles
 - fuel usage for all household vehicles
 - census data on land use density
 - 24-hour travel diaries for all members
 - 28-day record of long-distance travel (50 mi.+)
 - demographics and socio-economics
-
- <http://nhts.ornl.gov/2001/index.shtml>

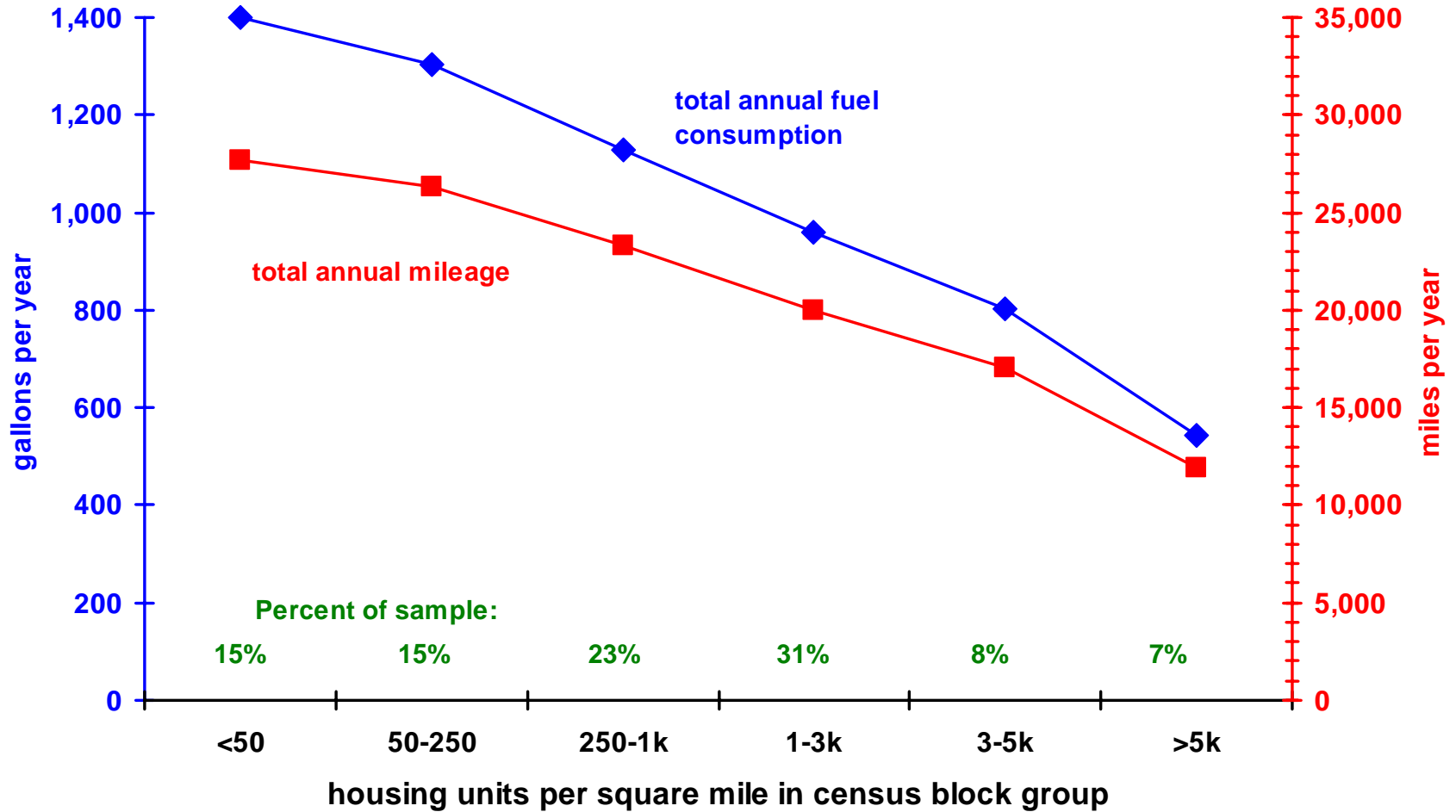
2001 NHTS data

- Annual mileage for all household vehicles
 - derived from two odometer readings or imputed
- fuel usage for all vehicles
 - according to vehicle make, model and vintage
 - see: Schipper and Pinckney (2003), *Supplementing the 2001 NHTS with Energy-related Data* (online)
- Census data on land use density
 - housing units per sq. mi. (block and tract levels)
 - population per sq. mi. (block and tract)
 - jobs per sq. mi. (tract level)

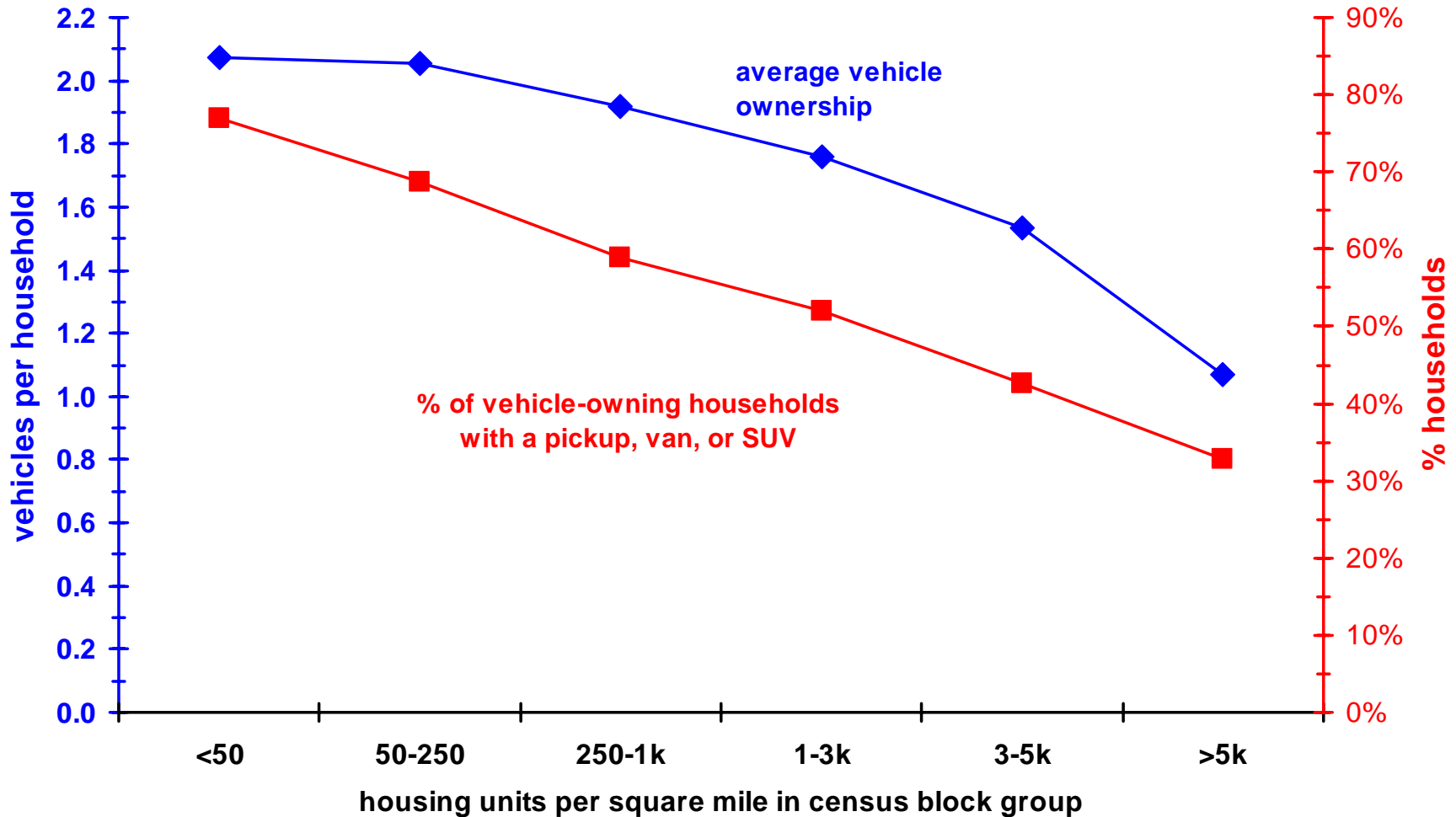
2001 NHTS sample

- National sample
 - about 26,000 households
 - 82% have complete data on fuel usage
- California subsample
 - 2583 households
 - 2079 (80%) have complete data on fuel usage
- Plus 9 add-ons for other areas
 - about 44,000 additional households
 - generally, no data on mileage and fuel usage

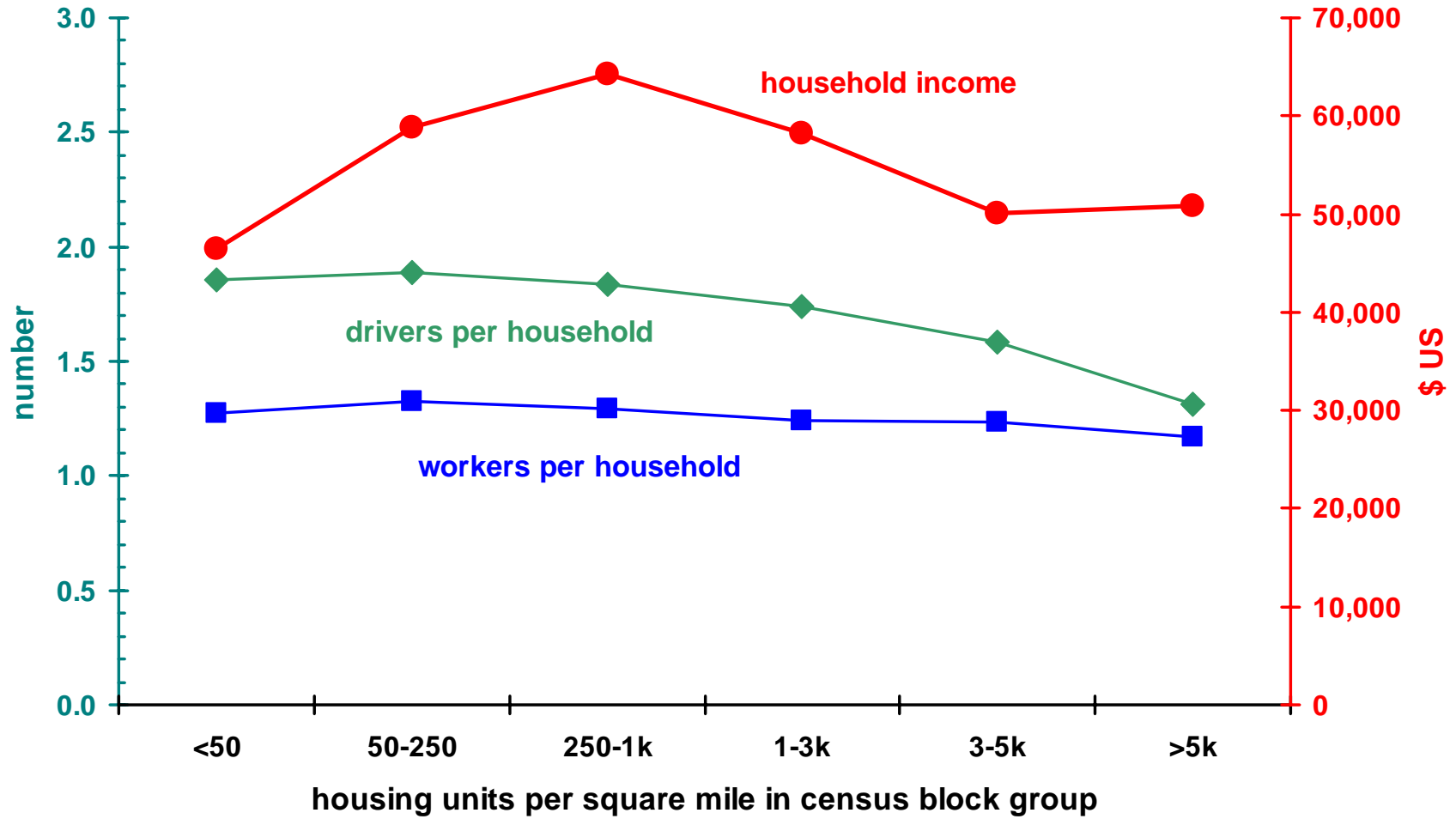
Mileage, fuel usage by residential density



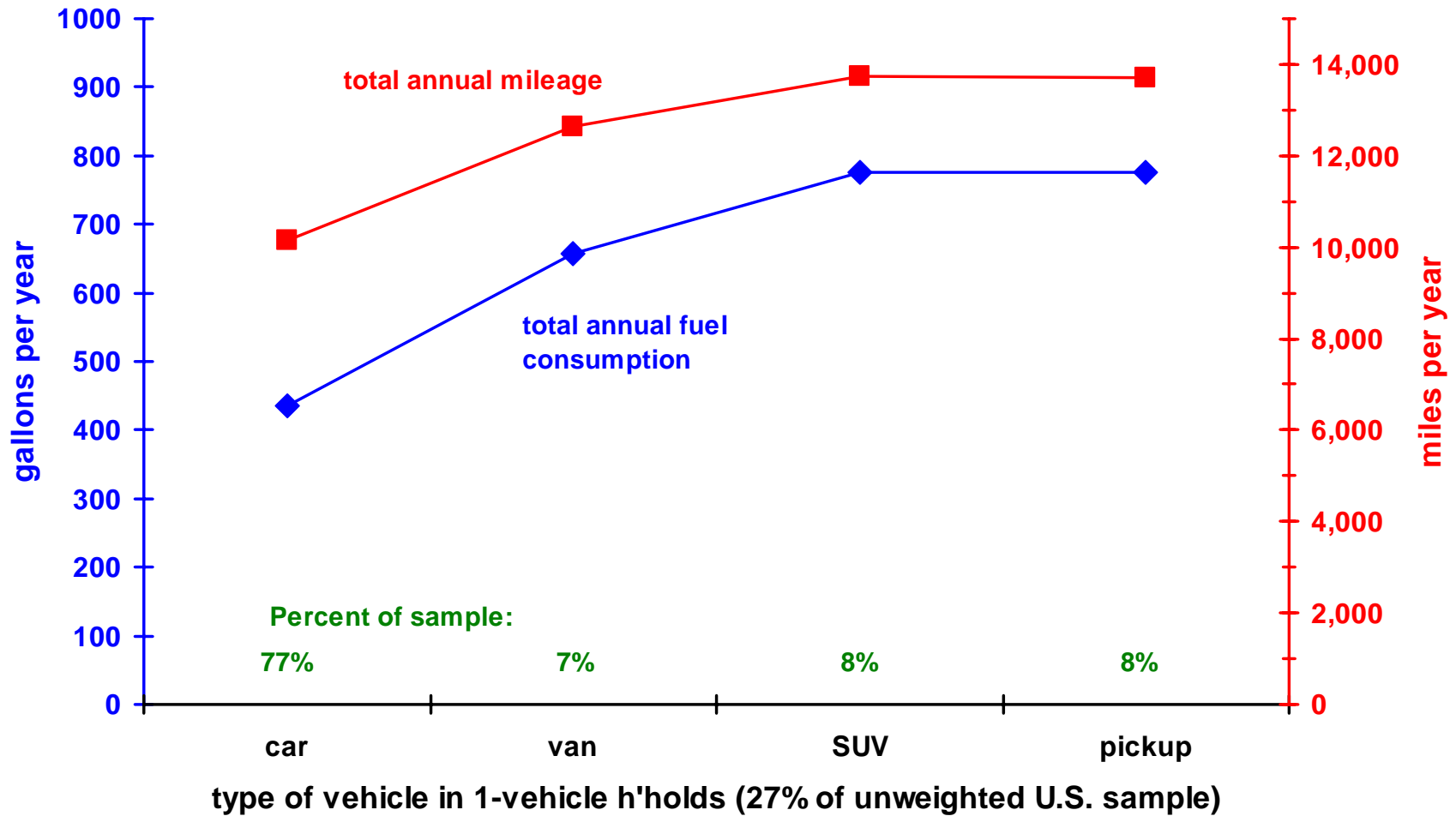
Vehicle ownership by residential density



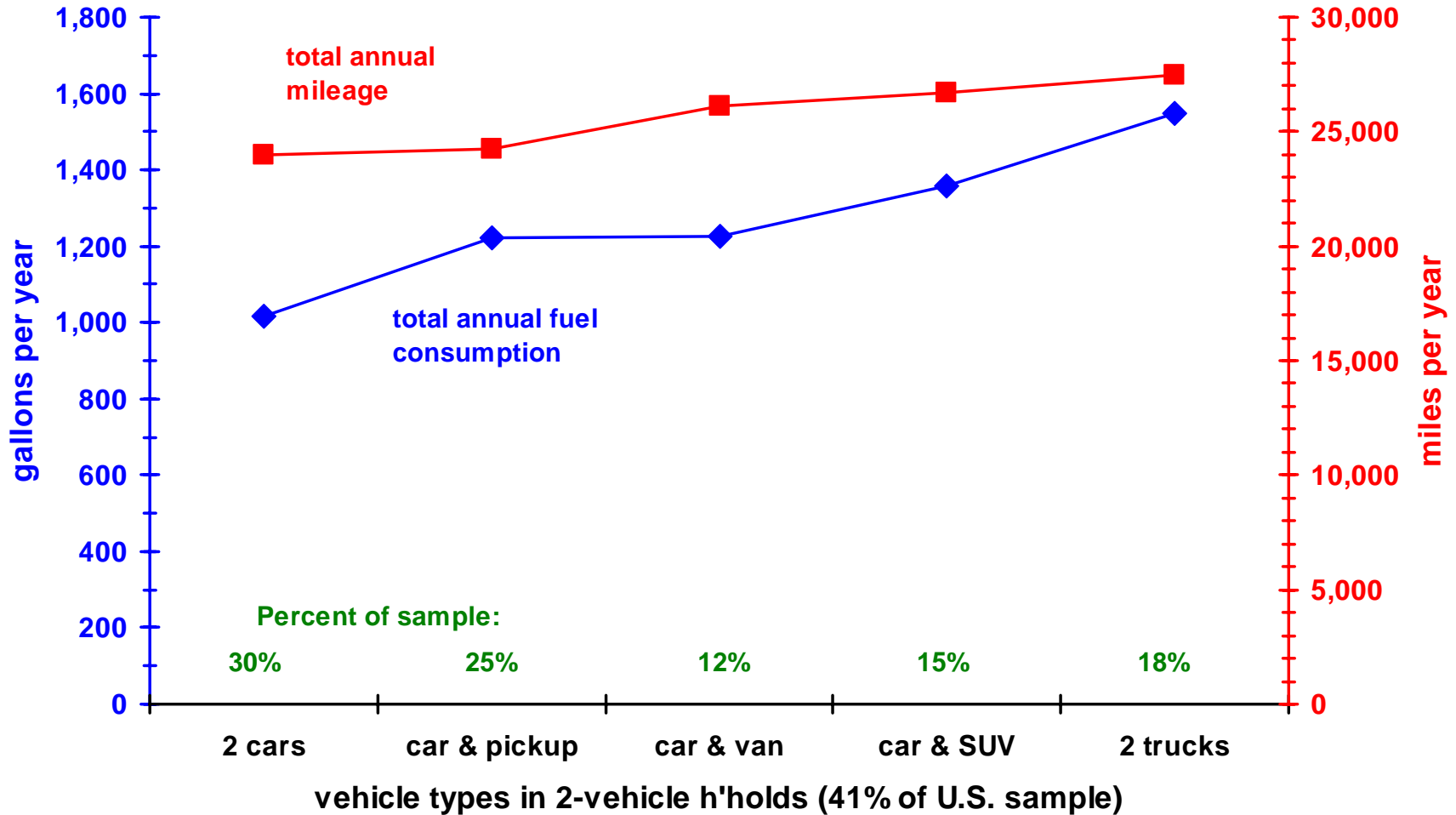
Demographics by residential density



Mileage & fuel by vehicle type in single-vehicle households



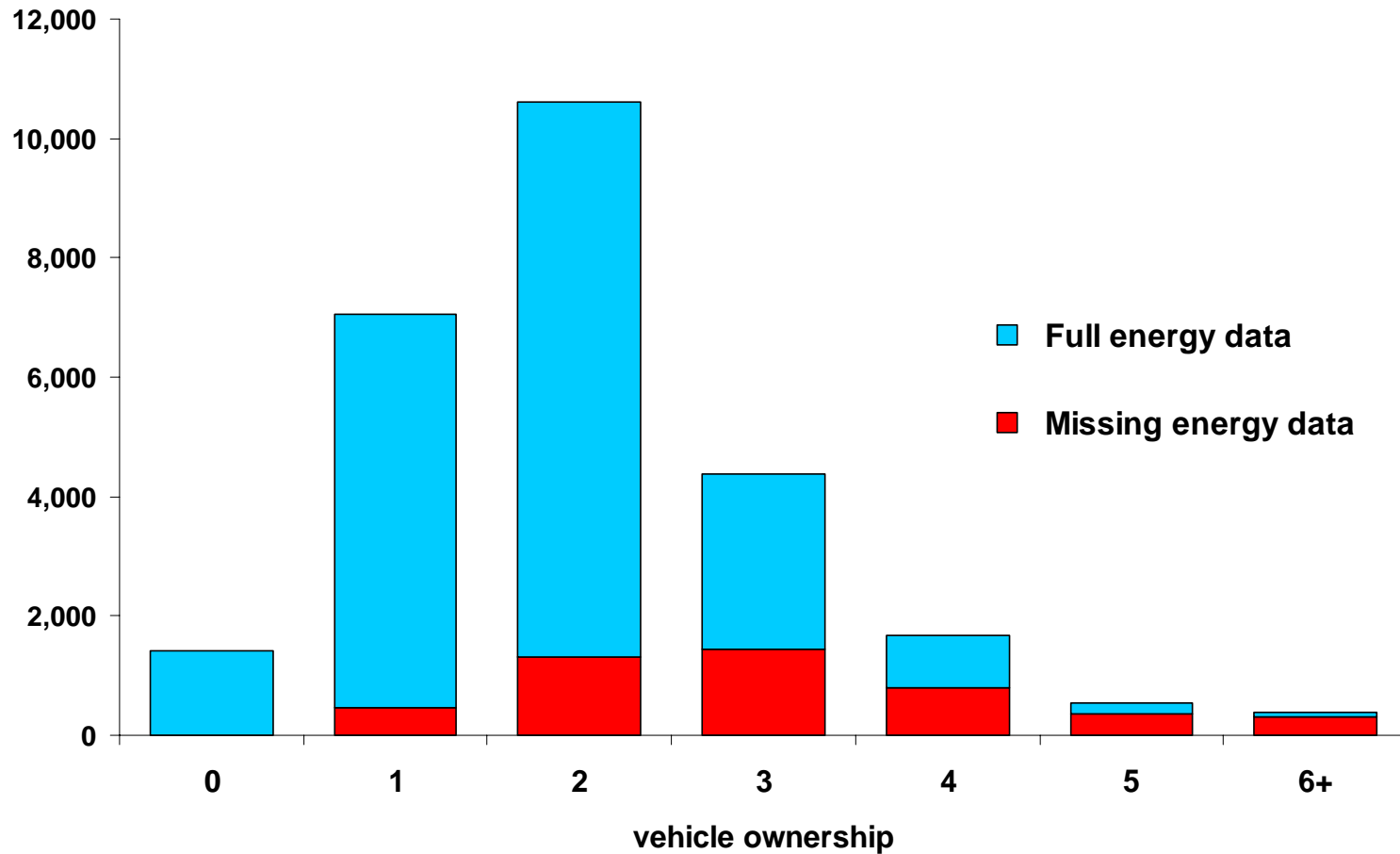
Mileage and fuel by vehicle type pairs in two-vehicle households



Important exogenous variables

- Income
- Number of drivers (continuous + dummy vars.)
- Number of workers (continuous + dummy vars.)
- Number of children (plus dummy for child > 15)
- Education (2 dummy vars.)
- Single-person household dummy
- Retired household dummy
- Race/ethnicity (4 dummy vars.)

Weighting to account for missing data



Biases due to missing data

- Missing data strongly related to number of vehicles, and this is closely related to endogenous mileage and fuel usage
- Sample selection problem
 - Structural approach (Heckman, 1979 *Econometrica*)
 - Weighting (Manski and Lerman, 1977, *Econometrica*) – use WESMLE

Endogenous Missing Data

- Structural (Heckman) approach results are very sensitive to model specification.
 - But cannot reject hypothesis that missing data are exogenous for preferred specification.
- WESMLE is not efficient, but is less sensitive to model specification. Also allows easy implementation of error heteroskedasticity.

WESMLE

- Model is
$$y_i = Ay_i + Bx_i + \varepsilon_i$$
$$\text{Cov}(\varepsilon_i) = \Omega$$

- WESMLE is

$$\min \sum w_i \left((I - A)y_i - Bx_i \right)' \Omega^{-1} \left((I - A)y_i - Bx_i \right)$$

weight is inverse of probability of selection

- WESMLE easy to compute, but variance=

$$V = \Psi^{-1} \Lambda \Psi^{-1}$$

$$\Psi = -E \left(\frac{\partial^2 w_i L_i(\theta, \mathbf{x}_i)}{\partial \theta \partial \theta'} \right)$$

$$\Lambda = E \left(\left(\frac{\partial w_i L_i(\theta, \mathbf{x}_i)}{\partial \theta} \right) \left(\frac{\partial w_i L_i(\theta, \mathbf{x}_n)}{\partial \theta'} \right) \right)$$

Alternative: Bootstrap – we used Wild Bootstrap to get covariance estimator that is consistent under arbitrary error heteroskedasticity

Wild Bootstrap

- Estimate model and get residuals, μ_i .
- Multiply vector μ_i by a draw from

$$\left(1 - \sqrt{5}\right)/2 \text{ with PR} = \left(1 + \sqrt{5}\right)/(2\sqrt{5})$$
$$\left(1 + \sqrt{5}\right)/2 \text{ with PR} = 1 - \left(1 + \sqrt{5}\right)/(2\sqrt{5})$$

Add resulting bootstrap residual to y_i to get
bootstrap samples

Specification testing

- Reduced form: $y_i = Cx_i + \mu_i$
- Structural restrictions:

$$C = (I - A)^{-1} B$$

$$\text{Cov}(\mu_i) = (I - A)^{-1'} \Omega (I - A)^{-1}$$

Bootstrap variance of difference between restricted and unrestricted estimates of C

Is this algebra important?

- Wrong standard errors for WESMLE are between 10 – 1000% downward biased.
- WESMLE estimates are statistically and operationally significantly different from unweighted estimates in many specifications.
- Standard overidentification tests reject many structural models that are accepted by bootstrap test – but bootstrap test does reject many specifications.

Model fit on California NHTS data

- Preferred Model structure is recursive with uncorrelated structural errors Ω
 - 3 endogenous variables
 - 19 exogenous variables
 - 44 free parameters (vs. 63 for unrestricted model)
- Model fits well
 - overall goodness-of-fit statistics all excellent. R^2 are .94 for fuel, .37 for mileage, and .11 for density.

Variable	Mean	Std. Dev.
Annual household fuel consumption in gallons	1173	1201
Total mileage per year for all household vehicles	25018	28486
Thousand dwelling units per sq. mile - Census block group	2.61	1.91
Annual household income in units of \$10,000	7.08	5.66
Number of children in household	0.69	1.07
Number of workers in household	1.43	1.08
Number of drivers in household	1.86	1.03

Structural Parameter Estimates

Influenced endogenous variable	Causal endogenous variable	
	Dwelling units per sq. mile in units of 1,000 – census block group	Total mileage per year on all household vehicles
Total mileage per year on all household vehicles	-1171 (-4.97)	
Household fuel usage per year in gallons	-20 (Total -65) (-5.12)	0.0382 (17.3)

Restricted reduced form

Endogenous variable

Exogenous variable	fuel usage	mileage	Density
Income in units of \$10,000	24.2	276	-0.017
Number of children in household	55.0	271	-0.232
Number of workers in household	-129	-211	0.180
1-worker household	422	8493	
2-worker household	761	13316	
3-or-more-worker household	1274	23327	

Exogenous variable	fuel usage	mileage	Density
Number of drivers in household	596	13815	-0.139
1-driver household	-128	-3716	-0.701
2-driver household	-315	-8792	-1.013
3-or-more-driver household	-265	-7515	-1.078
respondent has only college degree	-45.9		
respondent has postgraduate degree	-74.9		

Exogenous variable	fuel usage	mileage	Density
respondent is retired	129	4208	-0.409
youngest child at least 16-21 and at least 2 adults not retired	-400	-10850	-0.700
single-person household not retired	-14.1	-256	0.218
Asian	-199	-3989	0.601
Hispanic	-172	-3456	0.684
Black	-58.7	-1063	0.908
mixed White & Hispanic	-46.1	-835	0.713

Exogenous variable	Endogenous variable		
	fuel usage	mileage	Density
Income in units of \$10,000	24.2	276	-0.017
Number of children in household	55.0	271	-0.232
Number of workers in household	-129	-211	0.180
1-worker household	422	8493	
2-worker household	761	13316	
3-or-more-worker household	1274	23327	

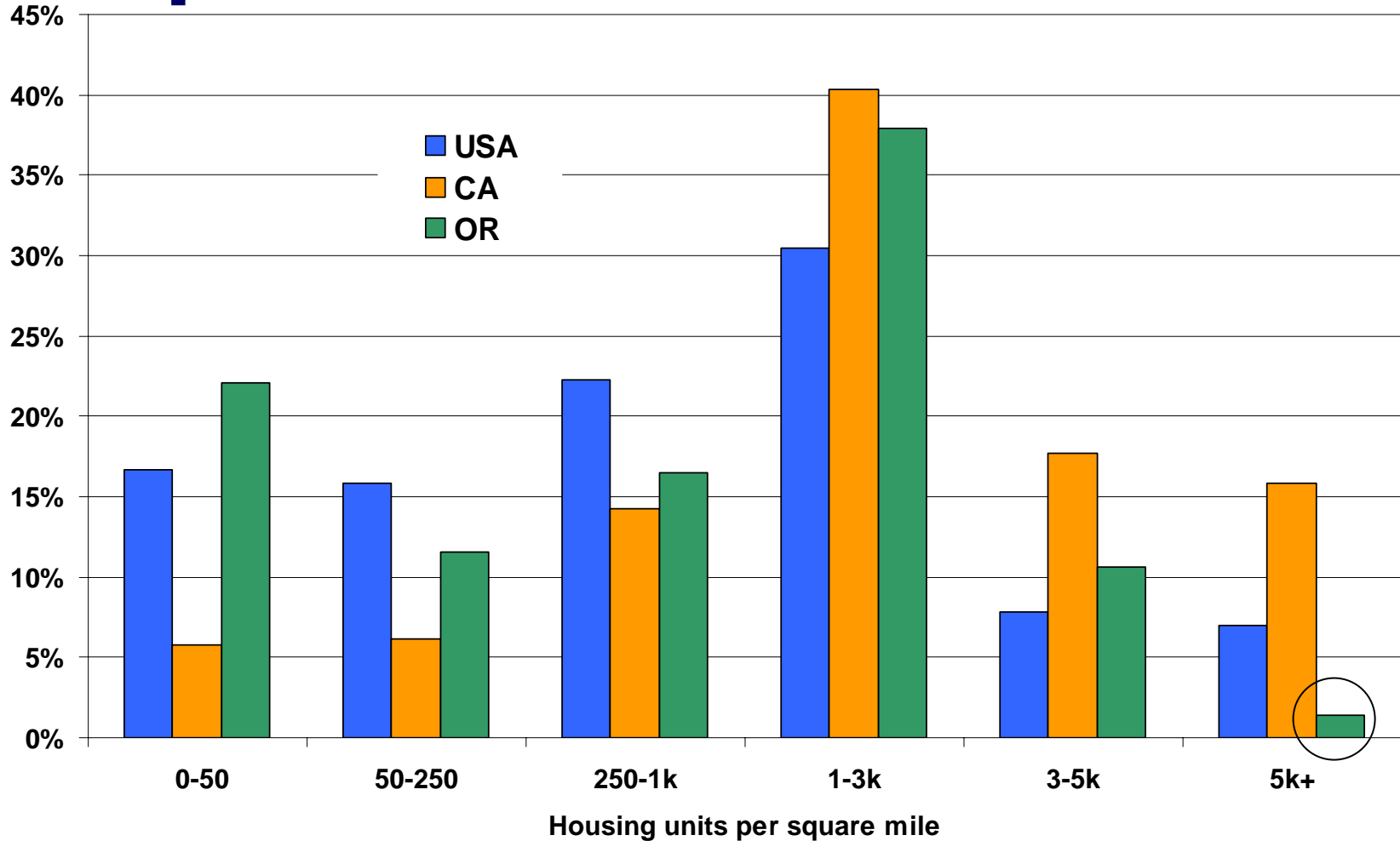
Exogenous variable	fuel usage	mileage		Density	
Number of drivers in household	596	13815		-0.139	
1-driver household	-128	-3716		-0.701	
2-driver household	-315	-8792		-1.013	
3-or-more-driver household	-265	-7515		-1.078	
respondent has only college degree	-45.9				
respondent has postgraduate degree	-74.9				

Exogenous variable	fuel usage	mileage	Density
respondent is retired	129	4208	-0.409
youngest child at least 16-21 and at least 2 adults not retired	-400	-10850	-0.700
single-person household not retired	-14.1	-256	0.218
Asian	-199	-3989	0.601
Hispanic	-172	-3456	0.684
Black	-58.7	-1063	0.908
mixed White & Hispanic	-46.1	-835	0.713

Contrasting results for 3 NHTS samples

- National
 - N = 21,347
- Portland Metro Area and rest of Oregon
 - (includes 2 counties in WA)
 - N = 325
- California
 - N = 2,079

Residential densities for 3 NHTS samples



Results by area

Increase in density of 1,000 households / sq. mi. (40%)				
Change in annual total mileage on all household vehicles		Change in annual fuel consumption due to:		
		mileage	fuel economy	Total
U.S.	- 1,630	- 74	- 16	- 90
OR	- 1,340	- 57	- 20	- 77
CA	- 1,200 (4.8%)	- 45	- 20	- 65 (5.5%)

Increasing Density by 1,000 households / sq. mi. ?

- Downs (2004) shows this requires extreme densities for new and infill development.
- Bryan, Minton, and Sarte (2007)
 - 30 out of 456 cities increased population density more than 40% between 1950 and 1990.
 - Median city decreased population density by 36%.
 - Cities with large population density increase are like Santa Ana – many poor immigrants

Conclusions

- We can measure the effects of residential density controlling for residential choice (self-selection)
- Self-selection effects are fully captured by rich demographics
- Impact of increased density statistically significant but too small for useful policy.

Vehicle Type Choice

- Previous results suggest that density may be related to the choice of vehicle fuel efficiency.
- Except for Bhat and Guo (2007) previous literature has treated density as exogenous
- Standard discrete choice has curse of dimensionality in number of vehicles.

BMOPT model with endogenous density

- Extends Fang's (2008) Bayesian Multivariate Ordered Probit and Tobit model to include an equation for density. Can be easily estimated with standard MCMC methods.
- Reduced form alternative to Bhat (2005) MCDEV model. This requires that the total household miles driven is fixed.

BMOPT Model

$$y_i^* = D_i \alpha + x_i \beta + \varepsilon_i \quad (1)$$

$$D_i = z_i \gamma + \eta_i \quad (2)$$

where y_i^* is a 4 by 1 vector of latent dependent variables for number of cars, number of trucks, mileage on cars, and mileage on trucks; D_i is a measure of density for households i at the census tract level, and is endogenous. The relation between the latent dependent variables and their observed values are:

$$y_j = 0, \text{ if } y_j^* \leq \alpha_0, j = 1, 2$$

$$y_j = 1, \text{ if } \alpha_0 < y_j^* \leq \alpha_1, j = 1, 2$$

$$y_j \geq 2, \text{ otherwise, } j = 1, 2$$

$$y_j = y_j^*, \text{ if } y_j^* > 0, j = 3, 4$$

$$y_j = 0, \text{ otherwise, } j = 3, 4$$

Data and Estimation

- Use a random subsample of 5863 observations from 2001 NHTS (25,027 total observations).
- Use MSA density as instrument for census block level density (appears strong 1st stage F-statistic about 400).
- Do out of sample forecasting tests from subsample to the remainder of the NHTS data.

Change in Vehicle Ownership for 50% increase in density

Probability changes for truck choice

$\Delta P(\text{tnum}=0)$	$\Delta P(\text{tnum}=1)$	$\Delta P(\text{tnum} \geq 2)$
.0267	-.0107	-.0159
(.0058)	(.0023)	(.0035)

Probability changes for car choice

$\Delta P(\text{cnum}=0)$	$\Delta P(\text{cnum}=1)$	$\Delta P(\text{cnum} \geq 2)$
-.0047	.0005	.0042
(.0054)	(.0007)	(.0048)

Change in Vehicle miles for 50% increase in density

Δ car miles	% Δ car miles	Δ truck miles	% Δ truck miles
14.02	.16	-610.5	-8.27
(196.79)	(2.23)	(117.66)	(1.59)

Out of Sample Predictions

	c=0	c=1	c ≥ 2	t=0	t=1	t ≥ 2
Predicted number of households (standard deviation)	1301 (28.8)	2677 (33.8)	1013 (25.5)	2413 (29.7)	1774.6 (34.9)	804 (25.9)
True number of households	1060	2601	1330	2165	1884	942

	average miles by cars	average miles by trucks
Forecast (standard deviation)	9113.6 (178.9)	7649.3 (210.6)
True	9135	7204.4

Conclusions

- No evidence for self-selection bias after controlling for rich sociodemographics
- BMOPT model works well and does tolerably well in out of sample forecasting
- Impact of increased density statistically significant but too small for useful policy.