Unobserved Heterogeneity and the Statistical Analysis of Highway Accident Data

Fred Mannering
University of South Florida

Highway Accidents

- Cost the lives of 1.25 million people per year
- Leading cause of death for people aged 15 to 29 years old
- These numbers have persisted in spite of advanced vehicle safety features, advances in highway design, and various safety-countermeasure policies

Analysis of Highway Accident Data

- Existing data bases include information from accident reports, local weather stations, highway-asset-management databases, etc.
- However current databases only cover a small fraction of the large number of elements that define human behavior, vehicle and roadway characteristics, traffic characteristics, and environmental conditions that determine the likelihood of an accident and its resulting injury severity (big data will not solve this issue).

Factors likely to be unobserved:

- Weather and lighting conditions change continually over time as do the driver reactions to these conditions
- Once an accident has occurred, the characteristics of energy dissipation through the vehicle structure and the resulting effect on individuals, which may vary widely based on which of the vehicle safety features deployed as well as bone mass, overall health, physical dimensions, and so on will not be known

Heterogeneity models

- Attempt to address data deficiencies with advanced statistical and econometric approaches
- Allow analysts to make more accurate inferences by explicitly accounting for observation-specific variations in the effects of influential factors
- Ignoring unobserved heterogeneity can lead to erroneous inferences

Example: The effect of gender

- Consider the indicator variable 1 if male 0 if not
- Attempts to account for general physiological and behavior differences

Problem:

- There is great variation across people of the same gender, including differences in height, weight, bone density, driving experience, response to stimuli and other factors that are generally unavailable to the analyst

Example: Type of accident

- Consider accident types such as “angle” or “head on” and their effect on injury severity

Problem:

- Vehicle-to-vehicle kinematic interactions relating to vehicle speed differences, differences in vehicle size, variations in vehicle impact locations, variations in structural integrity of the vehicles, and variations in angle of impact all comprise a significant portion of heterogeneity in collision-type effects

Example: Roadway lighting

- Consider the presence of roadway lighting at night (1 if present, 0 otherwise)
- Accounts for safety provided by lighting

Problem:

- There will be unobserved variations across roadway segments in lighting type, the ambient lighting from land uses nearby, as well as the light-output and types of lighting used

Ignoring unobserved heterogeneity:

- Restricting the variable effect to be the same across observations:
  - Model will be misspecified
  - Estimates will be bias and inefficient
  - Erroneous inferences and predictions

Addressing unobserved heterogeneity:

- Random Parameters Models
  - Effects of variables can vary across observations according to a specified distribution
- Finite Mixture (Latent Class) Models
  - Group of observations have the same parameters
- Combination of Finite Mixture/Random Parameters
- Markov-Switching Models (Temporal Heterogeneity)

Random parameters formulations:

- **Note:** Random effects models are not random parameters models
- Random effects necessitate panel data
- Random parameter models can be estimated with panel and cross-sectional data

Random parameters accident likelihood models

- Count-data models (negative binomial, zero-inflated)
- Duration models (time between accidents)
- Ordered probability models to generalize counts
- Tobit regression (for accident rates on highway segments)

Introduction of random parameters

- The effect of explanatory variables in model formulations for highway entity $\boldsymbol{i}$ (roadway segment):

$$\boldsymbol{b}'\boldsymbol{x}_i$$

where $\boldsymbol{x}_i$ is a vector of explanatory variables (including a constant) and $\boldsymbol{b}'$ is a vector of estimable parameters

Introduction of random parameters

- Each estimable parameter on explanatory variable l in the vector $\boldsymbol{x}_i$ can be written as,):

$$\beta_{il} = b_l + \varphi_{il}$$

where β_{il} is the parameter on the lth explanatory variable for observation i, b_l is the mean parameter estimate across all observations for the lth explanatory variable, and φ_{il} is a randomly distributed scalar term that captures unobserved heterogeneity across observations.

Introduction of random parameters

- The term φ_{il} can assume an analyst-specified distribution (such as the normal, lognormal, triangular, uniform and Weibull distributions).
- Estimation of random parameters models of this form is typically achieved with simulated maximum likelihood

Random parameters and injury severities

- Injury severity outcomes such as no injury, possible injury, evident injury, disabling injury and fatality are modeled with a variety of discrete-outcome models:
  - Multinomial logit
  - Ordered probability models (ordered probit)

Introduction of random parameters

- Done in much the same way with the random parameters multinomial logit written as (the probability of accident *i* resulting in injury severity level *k*),

$$P_i\left(k\right) = \int \frac{e^{\alpha_k + \boldsymbol{\beta}'\boldsymbol{x}_{ik}}}{\sum_m e^{\alpha_m + \boldsymbol{\beta}'\boldsymbol{x}_{im}}} f\left(\boldsymbol{\beta} \,|\, \varphi\right) d\boldsymbol{\beta}$$

where $f\,(\boldsymbol{\beta}\,|\,\varphi)$ is the continuous density function accounting for unobserved heterogeneity and $\boldsymbol{\varphi}$ is a vector characterizing the chosen density function.

Random parameters with correlated parameters

- Correlation can exist between factors like weather (snow) and physical roadway conditions (pavement roughness).
- This complicates estimation and interpretation, but can be important to the specification

Correlated parameters

- assuming a multivariate normal distribution for $\boldsymbol{\beta}_i$,

$$\boldsymbol{\beta}_{\boldsymbol{i}} = \boldsymbol{b} + \mathbf{C}\boldsymbol{\varphi}_i$$

where $\boldsymbol{\beta}_i$ is a vector of random parameters corresponding to explanatory variables for observation $\boldsymbol{i}$, $\boldsymbol{b}$ is the mean parameter estimate across all observations, **C** is a lower triangular matrix for correlation among the elements $\boldsymbol{\beta}_i$, and $\boldsymbol{\varphi}_i$ is a randomly and independently distributed uncorrelated vector term.

Heterogeneity in means

- Common applications of random parameters models assume one mean for all observations
- It my be more realistic to have means vary across observations in a systematic way

Heterogeneity in means

- Allow means to be functions of explanatory variables,

$$\boldsymbol{\beta}_i = \boldsymbol{b} + \boldsymbol{\Theta}\,\boldsymbol{z}_i + \mathbf{C}\boldsymbol{\varphi}_i$$

where $\mathbf{z}_i$ is a vector of explanatory variables from $\boldsymbol{i}$ that influence the mean of the random parameter vector, $\boldsymbol{\Theta}$ is an matrix of estimable parameters

Latent class (finite mixture) models

- Idea that groups of observations (latent classes) share a parameter value as opposed to each observation having its own
- Does not require distributional assumptions like random parameters models

Latent class (finite mixture) models

- Probability of belonging to a specific latent class:

$$P_i(c) = \frac{e^{\gamma' z_{ic}}}{\sum_g e^{\gamma' z_{ig}}}$$

Where z_{ic} is a vector of explanatory variables specific to observation i and latent class c (including a constant for all latent classes except one) and γ is a vector of estimable parameters.

Latent class (finite mixture) models

- Then, the conditional models become, for example, in the injury severity case:

$$P_i(k)|c = \frac{e^{\alpha_{kc}+\boldsymbol{b}_c'\boldsymbol{x}_{ik}}}{\sum_m e^{\alpha_{mc}+\boldsymbol{b}_c'\boldsymbol{x}_{im}}}$$

- Or, if random parameters are allowed in each class:

$$P_i(k)|c = \int \frac{e^{\alpha_{kc}+\boldsymbol{\beta}_c'\boldsymbol{x}_{ik}}}{\sum_m e^{\alpha_{mc}+\boldsymbol{\beta}_c'\boldsymbol{x}_{im}}} f\left(\boldsymbol{\beta}_c|\varphi_c\right)d\boldsymbol{\beta}_c$$

Temporal heterogeneity

- Due to their low frequencies, accident data are gathered over periods of time (weeks, months, years).
- Unobserved factors may vary from one time period to the next
- Results in time-varying unobserved heterogeneity

Markov-switching approach

- Multiple hidden states exist and there is a transition between these states over time
- For a model with two states:

$$P\left(s_{t+1}=1 \mid s_t=0\right)=p_{0\rightarrow 1}\text{ , and }P\left(s_{t+1}=0 \mid s_t=1\right)=p_{1\rightarrow 0}$$

- Where state transition probabilities $p_{0\rightarrow 1}$ and $p_{1\rightarrow 0}$ can be estimated from the accident data

Unobserved heterogeneity and omitted variables bias

- Many times it is difficult to collect all relevant data, creating an omitted variables bias in parameter estimates
- These omissions will be picked up as unobserved heterogeneity
- But the model will not be able to track data as well as having omitted variables included, so it is still a problem.

Unobserved heterogeneity and transferability

- If the model is found to have significant unobserved heterogeneity, the model cannot be spatially transferred (and possibly not temporally transferred)
- If significant unobserved heterogeneity is found, that also means standard fixed parameter models most certainly cannot be spatially transferred.

Summary

- Accounting for unobserved heterogeneity has provided important new insights into accident analysis
- Advances in estimation techniques and software have allowed this
- No one heterogeneity approach is superior, and the advantages of one over another change from database to database.

Summary (cont.)

- Methodological applications are needed that address underlying data issues (unobserved heterogeneity, etc.)
- The methodological frontier needs to expand to include sophisticated new statistical and econometric methods