# Analytic Methods in Accident Research: The Current Methodological Frontier and Future Directions

Fred Mannering
Charles Pankow Professor of Civil Engineering
Purdue University

## Traditional Crash Data

- Available mostly from police and possibly other reports
  - Road conditions
  - Estimates of injury severity (property damage only, possible injury, evident injury, disabling injury, fatality)
  - Occupant characteristics (age, gender)
  - Vehicle characteristics
  - Crash description, primary cause, etc.

## Emerging Data Sources

- Data from driving simulators
- Data from naturalistic driving
- Other “non-traditional” sources (detailed hospital injury data, etc.)

## Why Analyze Traditional Crash Data?

- Identify crash-prone locations
- Hoping that data analysis will suggest effective countermeasures
- Evaluate the effectiveness of implemented countermeasures

## Traditional Analysis Approaches:

- Modeling of crash frequency
- Modeling of crash-injury severity
- Some modeling approaches seek to combine the two (frequency and severity)



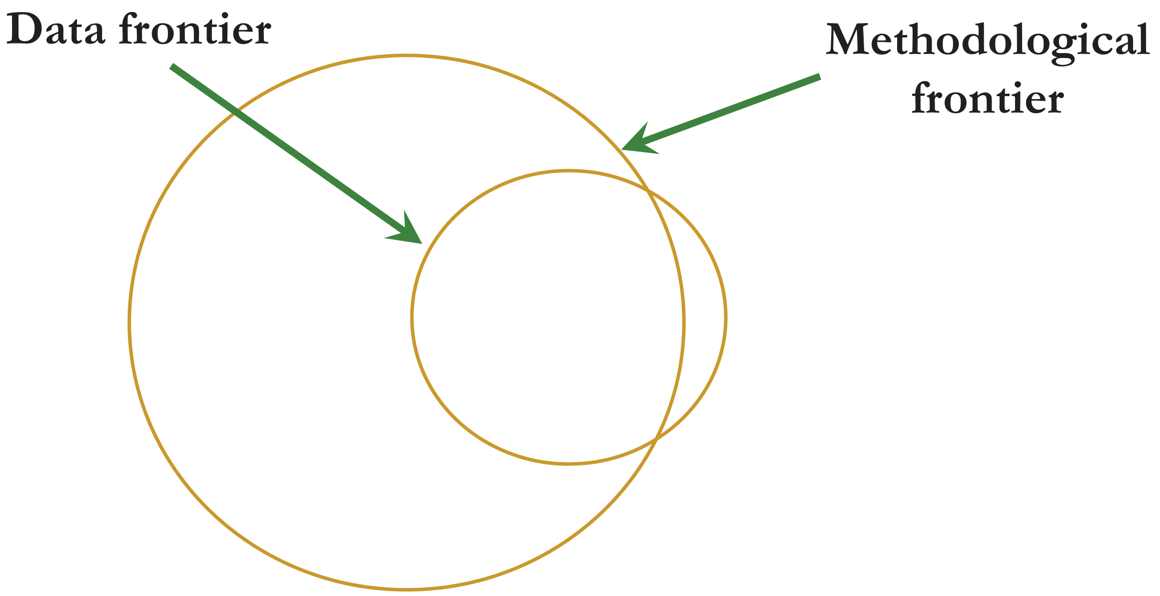
## Crash Frequency Models:

- Study the number of crashes over some specified time and space
- Various count-data and other methods have been used
- Explanatory variables:
  - Traffic conditions
  - Roadway conditions
  - Weather conditions

## Crash Injury Severity Models:

- Study injury severities of specific crashes
- Models are conditional on a crash having occurred
- Various discrete-outcome methods
- Explanatory variables:
  - Traffic Conditions, Roadway conditions, Weather conditions
  - Specific crash data: Vehicle information, Occupant information, Crash specific characteristics

## Traditional Crash Data



## What Methodological Barriers have Encountered:

- Unobserved Heterogeneity
- Endogeneity
  - Self-selectivity
- Temporal Correlation
- Spatial Correlation

Traditional crash data

## Unobserved Heterogeneity:

- Many factors influencing the frequency and severity of crashes are simply not observed
- If these are correlated with observed factors, incorrect inferences could be drawn

## Example:

### A study finds age to be an important factor in crash frequency/severity

- Problem:
  - Age is correlated with many underlying factors such as physical/mental health, attitudes, income, life-cycle factors, etc.
- Naive methodological application:
  - Effects of age are a proxy for unobserved factors – the correlation may not be stable over time and inferences will be incorrect



## Example:

### Impact of ice warning signs on frequency/severity of ice-related crashes

- Analyze the frequency/severity of crashes when ice warning signs are present vs. not present
- Problem:
  - Ice warning signs are put at locations with a high frequency and severity of ice crashes
- Naive methodological application:
  - Effectiveness of ice-warning signs understated (may find they actually increase frequency and severity)

## Example:
### Effectiveness of Side-Impact Airbags

- Analyze the severity of crashes involving vehicles with and without side-impact airbags
- Problem:
  - People owning side-impact airbags are not a random sample of the population (likely safer drivers)
- Naive methodological application:
  - Side-impact airbag effectiveness is overstated

## Good Morning America

- http://abcnews.go.com/Video/playerIndex?id=2530346

## Ignoring self-selectivity will almost always overstate the effectiveness of new safety features due to self-selectivity

- May mask important factors relating to possible risk compensation, etc.
- Statistical corrections must be used



## The issue of risk compensation

- In the presence of advanced safety features drivers can drive faster, engage in distracted driving, etc. and maintain an acceptable level of safety
- Simulator study of the effectiveness of an in-vehicle hazard warning system:
  - Slow down during the hazard
  - Speed up afterward to make up for lost time

## Example:
### Use of Airbags or other safety features as an Explanatory Variable

- Use airbag deployment as an explanatory variable finding it reduces injury severity
- Problem:
  - People owning airbag cars are not a random sample of the population (likely safer drivers)
- Naive methodological application:
  - Airbag effectiveness is overstated



## Example:
### Effectiveness of Motorcycle Safety Courses

- Analyze the frequency and severity of crashes involving riders with and without course experience
- Problem:
  - People taking the course are not a random sample of the population (likely less skilled)
- Naive methodological application:
  - Effectiveness of the course understated (course participants may have higher crash rates)

## Underlying issue:

- There is unobserved heterogeneity about drivers that can manifest itself as a self-selectivity problem
- This can mask causality and lead to erroneous inferences and policies



## Temporal and Spatial Correlation

- Crashes in close spatial proximity will share correlation due to unobserved factors associated with space (unobserved visual distractions, sight obstructions, etc.)
- Crashes in occurring near the same or similar times will share correlation due to unobserved factors associated with time (precise weather conditions, similar sun angle, etc.)
- Much to be learned from spatial econometrics

## Omitted Variables

- Many crash frequency models use few explanatory variables (some only use traffic)
- This creates a massive bias in parameter estimates that most certainly will lead to incorrect and temporally unstable inferences



## Building on Bad Research

- Highway Safety Manual (HSM) in the U.S. is an important practice-oriented document
- However, it is several methodological generations behind the cutting-edge econometrics in the field
- Problem: Some researchers view the HSM as the cutting edge and they base their work on terribly outdated methods and thinking

# Current Methodological Frontier

- **Accident Frequency Models**
  - Random parameters count models (track unobserved heterogeneity) – Random parameters negative binomial
  - Markov switching/finite mixture count-data models (track unobserved heterogeneity by allowing observations to switch between two or more states)



- **Accident Severity Models**
  - Random parameters Logit model
  - Random parameters ordered probit models with thresholds that move as a function of explanatory variables
  - Markov switching/finite mixture variants

## Accident Severity Models – Ordered vs. Unodered

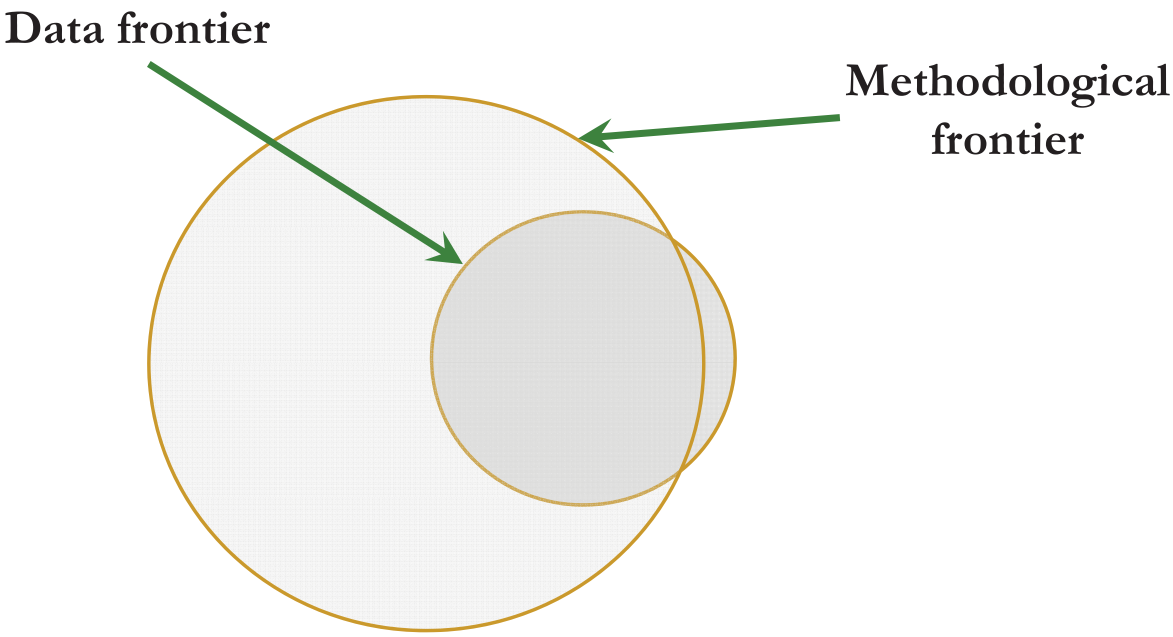
- Unordered models have far more flexible functional form, may perform better in under-reporting of minor crashed
- Ordered models account for the ordering of injury severity levels (no injury, possible injury, evident injury, fatality)



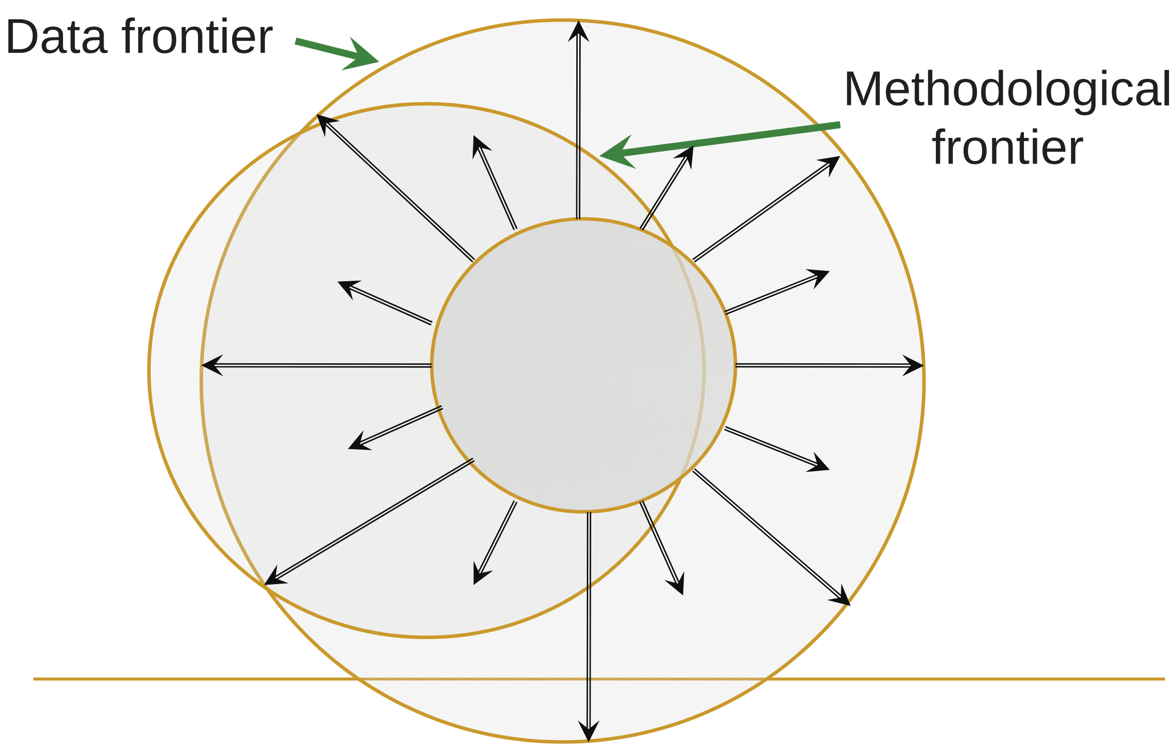
## Ordered vs. Unordered Which is better?

- Likely to depend on the crash data
- Example: Some crash data show a simple multinomial logit is best, others nested logit, etc.
- There is no generalization that can be made

# Traditional Crash Data




# Emerging Data Sources



## Naturalistic Driving and Simulator Data

- Unobserved heterogeneity
- Endogeneity
  - Self-selectivity (route choices, etc.)
- Temporal correlations
- Spatial correlations
- Subject correlations
- Realism (how does experiment affect behvior)



## Massive correlation problems

- Spatial and temporal correlation in the traditional sense are compounded subject correlation

## Realism problems

- The effect of subjects knowing they are in a simulator or naturalistic driving experiment is a constant concern

## Summary

- In the past, comparatively "static" data quality and quantity has enabled sophisticated methodological applications to extract much of the available information
- A new data-rich era is beginning
- With few exceptions, sophisticated methodologies have not been widely used on these data

## 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

## Without expanding the methodological frontier:

- **At best:** We will be effectively ignoring important information on crash and injury mitigation
- **At worst:** Incorrect analyses of data will lead to erroneous and ineffective safety policies