

# Aggregate Forecasting

# Outline

- Aggregate forecasting problem
- Aggregation bias
- Forecasting methods
  - Average individual
  - Classification
  - Sample enumeration

# What's the Problem?

- The models are disaggregate
  - Based on data for individuals or households
  - Predict the behavior of individuals or households
- Predictions must be aggregate – to provide forecasts of highway volumes, transit ridership, air pollutant quantities
- How do we obtain aggregate predictions from disaggregate models?

# Aggregate forecasting

- Disaggregate models
  - Behavior of decision-maker
- Forecasting
  - Prediction of market shares
    - Population
    - Market segments

# Definition

- Disaggregate behavior

$$Y_n = h(X_n)$$

- Aggregate quantity in population

$$Y_g = \sum_{n=1}^{N_g} Y_n = \sum_{n=1}^{N_g} h(X_n)$$

$$W_g = \frac{1}{N_g} \sum_{n=1}^{N_g} Y_n = \frac{1}{N_g} \sum_{n=1}^{N_g} h(X_n)$$

# Discrete choice model

- Expected market share

$$P(i | X_n) = h^i(X_n)$$

$$W_g(i) = \frac{1}{N_g} \sum_{n=1}^{N_g} P(i | X_n) = E[P(i | X_n)]$$

- Requires knowledge of attributes of all population members

$$W_g(i) = \int_X P(i | X) f(X) dX = E[P(i | X_n)]$$

- Requires knowledge of the distribution of attributes in the population

# The problem

- Limited information of the distributions of attributes in the population
- Computation can be very expensive
- What may be available?

– Averages

$$\bar{X}_g$$

– Sample

$$X_n, \quad n = 1, \dots, N_g$$

– Assumed distributions

$$\hat{f}_g(X)$$

# Aggregation bias

- Use of averages

- Linear model

$$\bar{Y} = \beta' \bar{X}$$

- Average demand can be estimated without bias as the demand of average attributes

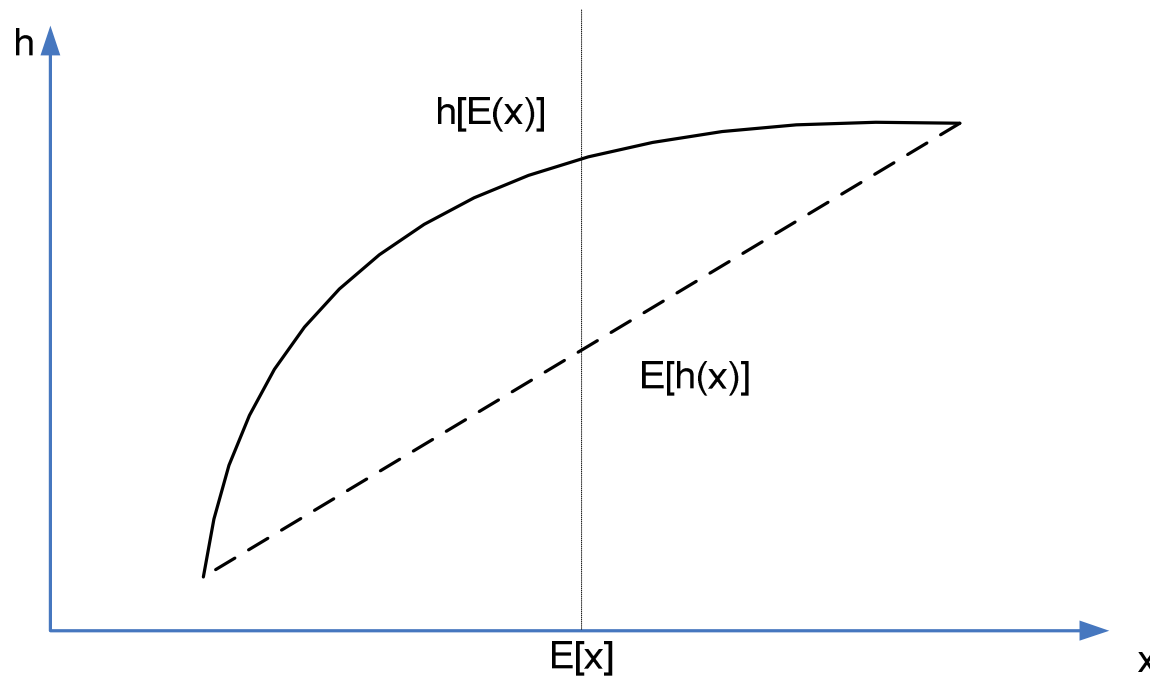
- Nonlinear model

- The curvature of the function introduces bias



# Jensen's inequality

- Concave functions  $h[E(X)] \geq E[h(X)]$ 
  - Reverse for convex functions



- Difference is aggregation bias
- Direction of bias is unknown

# Logit example

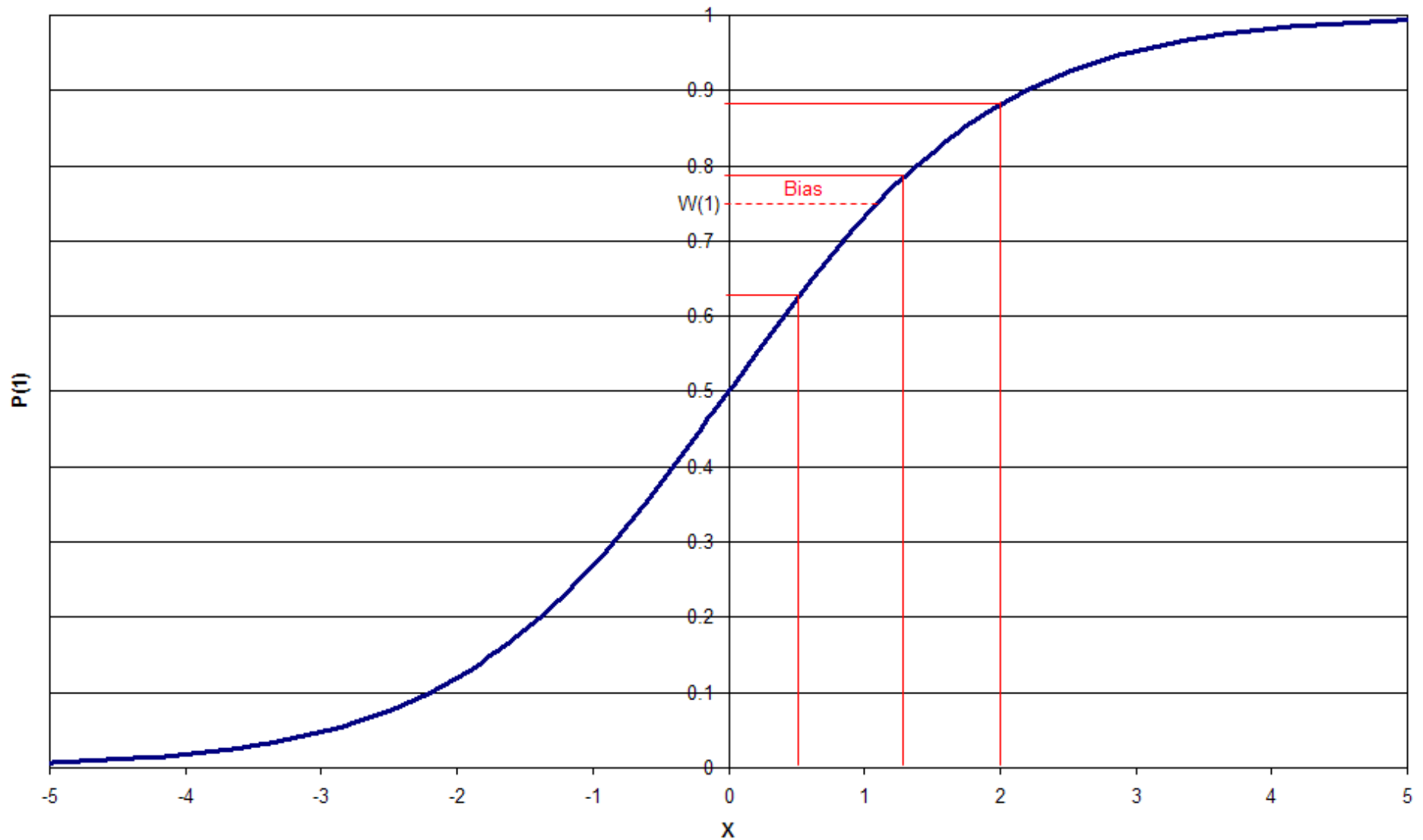
- Choice between two alternatives

$$P_n(1) = \frac{1}{1 + \exp[-(X_n)]}$$

n	X	P(1)
1	0.5	0.62
2	2.0	0.88
Average	1.25	0.78

$$W_g(1) = \frac{1}{N_g} \sum_{n=1}^{N_g} P(i | X_n) = \frac{1}{2} [0.88 + 0.62] = 0.75$$

# Graphical representation



# Forecasting methods

- Average individual
- Classification
- Sample enumeration

# Average individual

$$\hat{W}(i) = P(i | \bar{X})$$

- Appropriate
  - Homogenous population:  $\text{var}(X)$  is small
  - Model not highly nonlinear

# Classification

- Steps
  - Divide population to (homogenous) segments
    - Mutually exclusive collectively exhaustive
  - Choose a representative set of variables for each group

$$\bar{X}_g \quad g = 1, \dots, G$$

- Apply average individual method for each segment and approximate  $W(i)$

$$\hat{W}(i) = \sum_{g=1}^G \frac{N_g}{N_T} P(i | \bar{X}_g)$$

# Determining segments

- Ideally, different ranges of utility,  $V_i$
- Cannot realistically use all  $X$ 's
  - Important  $X$ 's in the model
  - Segment averages can be reasonably estimated
    - Zones, geographic classification
  - Ease of estimating segment population size
  - Avoid very small segments
- Open segments
  - Minimize the size of open ended segments, i.e., income over \$100,000
- Account for availability of alternatives
  - i.e., households with no vehicles
- Zones or zone pairs provide a “natural” classification

# Sample enumeration

- Use a sample to represent population
- Random

$$\hat{W}(i) = \frac{1}{N} \sum_{n=1}^N P(i | X_n)$$

- Non-random

$$\hat{W}(i) = \frac{\sum_{n=1}^N W_n P(i | X_n)}{\sum_{n=1}^N W_n}$$

– Weight inversely related to sampling rates

- Useful for segment-based policies and forecasts<sub>16</sub>



# Choice simulation

- Sample enumeration requires tracking all probabilities
- Large choice sets
  - e.g., destination and route choice
- Multidimensional choices
  - Model systems
  - Dynamic choice processes over time
- Computationally demanding
  - Many “paths” through decision tree

# Monte Carlo simulation

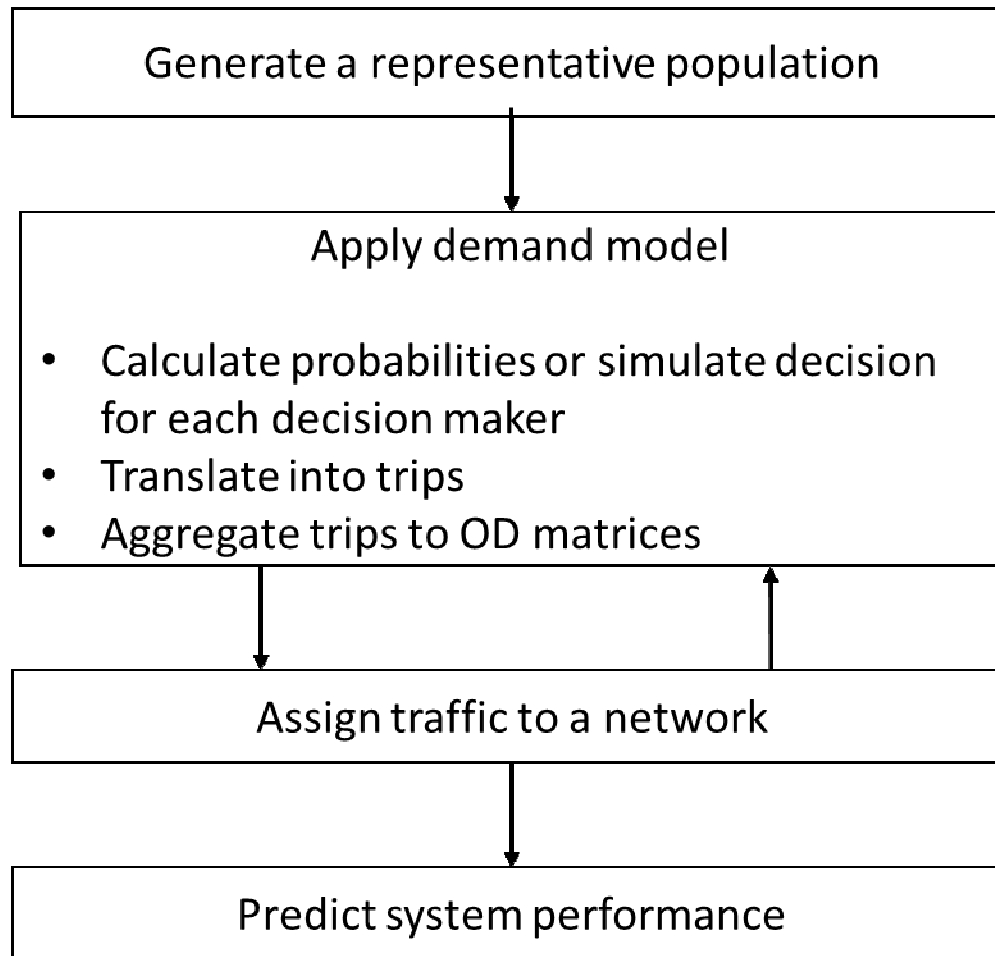
- Generate realizations for each choice
  - Repeat to estimate shares
- Draw random variables from a uniform [0,1] distribution
- Make simulated choices
- Use simulated choice in subsequent models
- Sum up simulated choices

$$\hat{W}_t(i) = \frac{N^t(i)}{N_s}$$

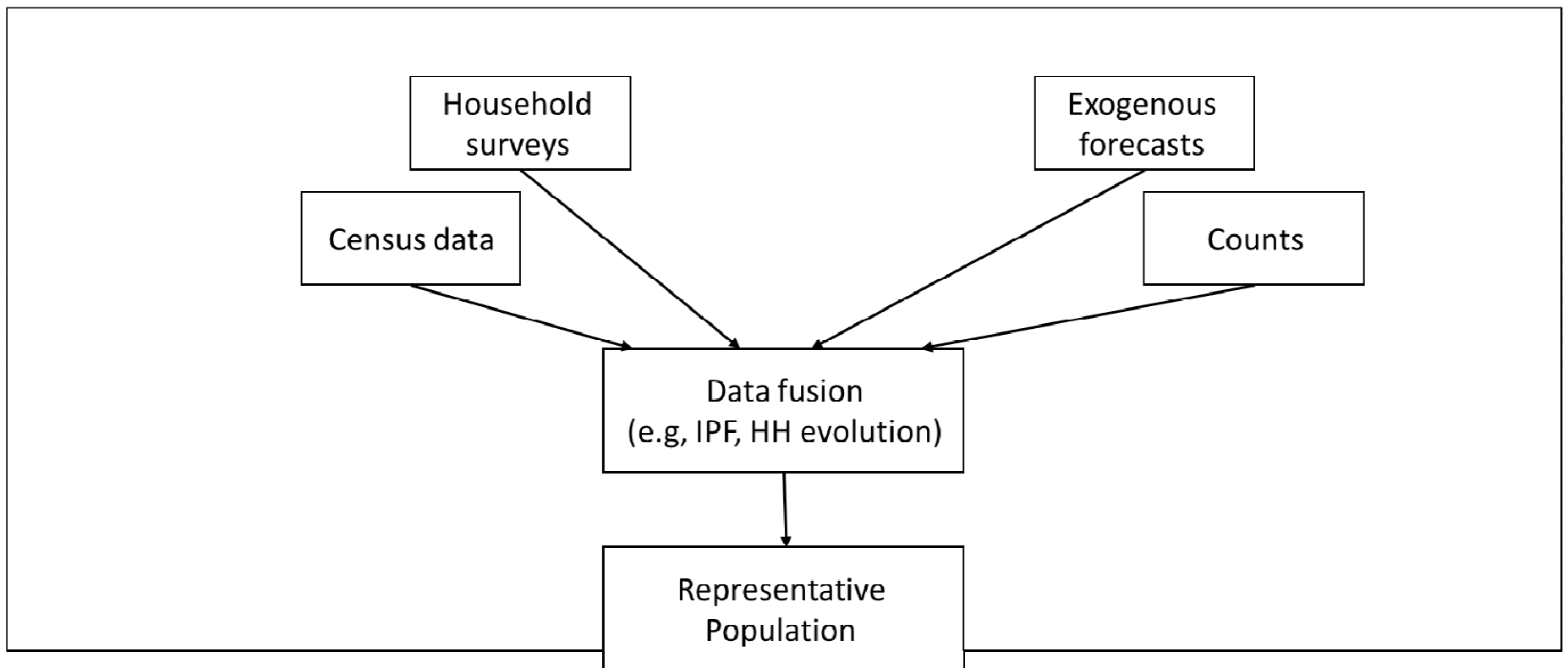
# Sample Enumeration Characteristics

- Predictions are subject to sampling error; they will have large errors when the choice probabilities or sample size are small
- The predictions are a consistent estimate when the parameters are consistent
- Predictions for socioeconomic groups are easy to obtain
- Simulation predictions have higher variances
- Convenient for testing policy changes which affect entire groups

# Disaggregate Prediction



# Generate Disaggregate Population



# Conclusions

- Average individual prediction should be avoided
- Classification typically used when spatial results are needed
- Classification should be based on groups having differences in alternative availability
- Sample enumeration typically used to provide area wide forecasts
- Errors due to aggregation across individuals can be made small without great difficulty