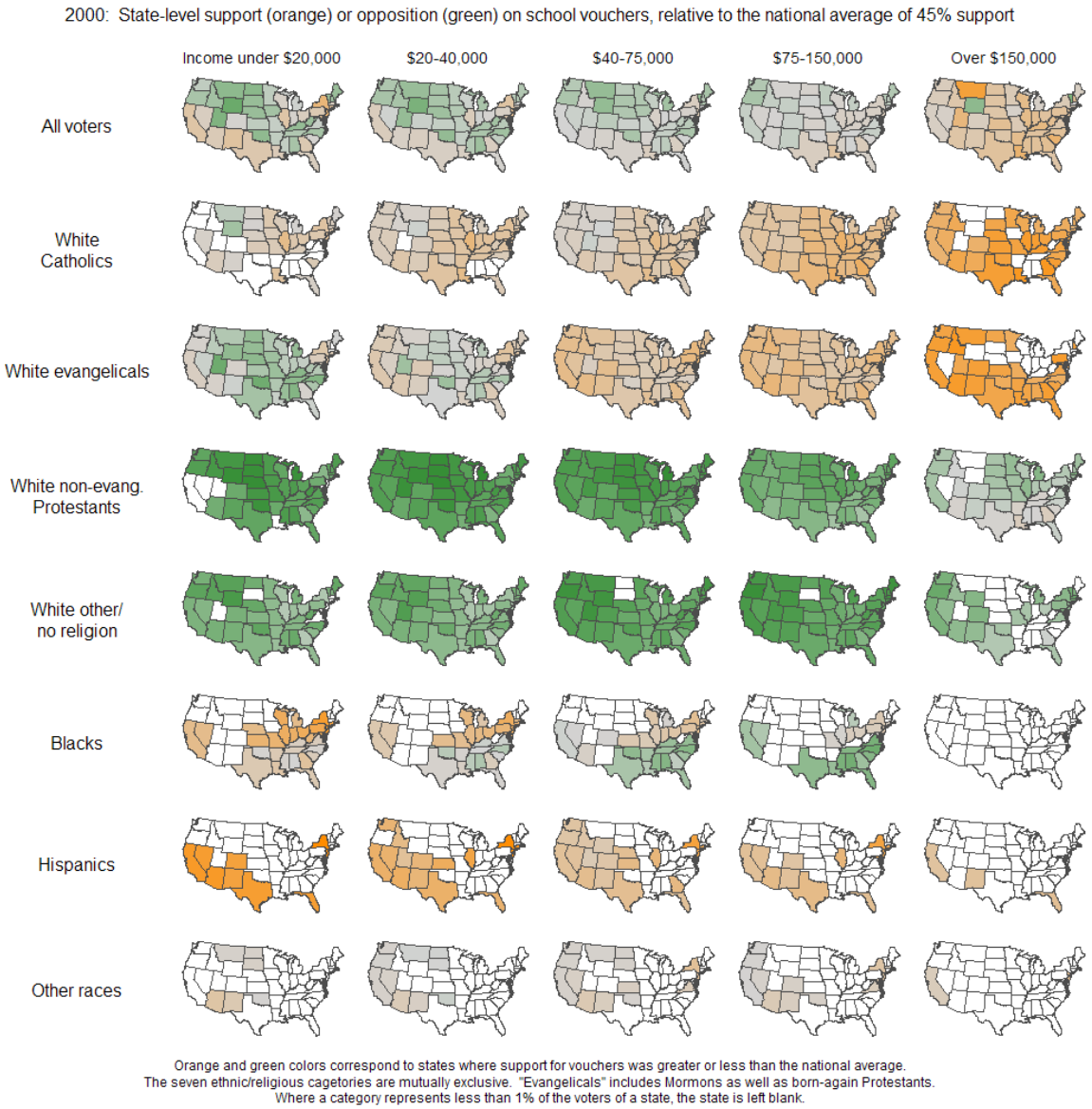
2000: State-level support (orange) or opposition (green) on school vouchers, relative to the national average of 45% support
Income under $20,000
$20-40,000
$40-75,000
$75-150,000
Over $150,000
All voters
White Catholics
White evangelicals
White non-evang. Protestants
White other/ no religion
Blacks
Hispanics
Other races
Orange and green colors correspond to states where support for vouchers was greater or less than the national average.
The seven ethnic/religious cagetories are mutually exclusive. "Evangelicals" includes Mormons as well as born-again Protestants.
Where a category represents less than 1% of the voters of a state, the state is left blank.

Survey weighting and hierarchical regression

Andrew Gelman and Rachel Schutt

2 August 2009

Survey weighting and hierarchical regression

- Discussion of today's talks
- 2 stories about survey weights
- A general approach
- Complications

Survey weighting and hierarchical regression

- Discussion of today's talks
- 2 stories about survey weights
- A general approach
- Complications

Survey weighting and hierarchical regression

- Discussion of today's talks
- 2 stories about survey weights
- A general approach
- Complications

Survey weighting and hierarchical regression

- Discussion of today's talks
- 2 stories about survey weights
- A general approach
- Complications

Survey weighting and hierarchical regression

- Discussion of today's talks
- 2 stories about survey weights
- A general approach
- Complications

Where do weights come from?

- Survey weights are not inverse probabilities of selection
- Two simple stories
- CBS/New York Times pre-election polls

Where do weights come from?

- Survey weights are not inverse probabilities of selection
- Two simple stories
- CBS/New York Times pre-election polls

Where do weights come from?

- Survey weights are not inverse probabilities of selection
- Two simple stories
- CBS/New York Times pre-election polls

Where do weights come from?

- Survey weights are not inverse probabilities of selection
- Two simple stories
- CBS/New York Times pre-election polls

Story 1: weights for men and women

- Survey of a population with 52% women, 48% men
- Simple random sampling, $n = 100$
  - 48% women, 52% men: $\hat{y}_{\text{unweighted}} = 0.48\,\bar{y}_{\text{women}} + 0.52\,\bar{y}_{\text{men}}$
  - 52% women, 48% men: $\hat{y}_{\text{weighted}} = 0.52\,\bar{y}_{\text{women}} + 0.48\,\bar{y}_{\text{men}}$
- We know the population proportions, so the selection probabilities are irrelevant
- Weights depend on the entire survey; the (y_i, w_i) paradigm is inappropriate

Story 1: weights for men and women

- Survey of a population with 52% women, 48% men
- Simple random sampling, $n = 100$
  - SRS 1: 52 women, 48 men. Weights are $w_i = 1$ for everyone
  - SRS 2: 60 women, 40 men. Weights are $w_i = \frac{52}{60}$ for women, $\frac{40}{48}$ for men
- We know the population proportions, so the selection probabilities are irrelevant
- Weights depend on the entire survey; the (y_i, w_i) paradigm is inappropriate

Story 1: weights for men and women

- Survey of a population with 52% women, 48% men
- Simple random sampling, $n = 100$
  - SRS 1: 52 women, 48 men. Weights are $w_i = 1$ for everyone
  - SRS 2: 60 women, 40 men. Weights are $w_i = \frac{52}{60}$ for women, $\frac{40}{48}$ for men
- We know the population proportions, so the selection probabilities are irrelevant
- Weights depend on the entire survey; the (y_i, w_i) paradigm is inappropriate

Story 1: weights for men and women

- Survey of a population with 52% women, 48% men
- Simple random sampling, $n = 100$
  - SRS 1: 52 women, 48 men. Weights are $w_i = 1$ for everyone
  - SRS 2: 60 women, 40 men. Weights are $w_i = \frac{52}{60}$ for women, $\frac{40}{48}$ for men
- We know the population proportions, so the selection probabilities are irrelevant
- Weights depend on the entire survey; the (y_i, w_i) paradigm is inappropriate

Story 1: weights for men and women

- Survey of a population with 52% women, 48% men
- Simple random sampling, $n = 100$
  - SRS 1: 52 women, 48 men. Weights are $w_i = 1$ for everyone
  - SRS 2: 60 women, 40 men. Weights are $w_i = \frac{52}{60}$ for women, $\frac{40}{48}$ for men
- We know the population proportions, so the selection probabilities are irrelevant
- Weights depend on the entire survey; the (y_i, w_i) paradigm is inappropriate

Story 1: weights for men and women

- Survey of a population with 52% women, 48% men
- Simple random sampling, $n = 100$
  - SRS 1: 52 women, 48 men. Weights are $w_i = 1$ for everyone
  - SRS 2: 60 women, 40 men. Weights are $w_i = \frac{52}{60}$ for women, $\frac{40}{48}$ for men
- We know the population proportions, so the selection probabilities are irrelevant
- Weights depend on the entire survey; the (y_i, w_i) paradigm is inappropriate

Story 1: weights for men and women

- Survey of a population with 52% women, 48% men
- Simple random sampling, $n = 100$
  - SRS 1: 52 women, 48 men. Weights are $w_i = 1$ for everyone
  - SRS 2: 60 women, 40 men. Weights are $w_i = \frac{52}{60}$ for women, $\frac{40}{48}$ for men
- We know the population proportions, so the selection probabilities are irrelevant
- Weights depend on the entire survey; the (y_i, w_i) paradigm is inappropriate

Example: CBS/New York Times pre-election polls

id	org	y	state	edu	age	adults	weight
6140	cbsnyt	NA	7	3	1	2	923
6141	cbsnyt	1	39	4	2	2	558
6142	cbsnyt	0	31	2	4	1	448
6143	cbsnyt	0	7	3	1	2	923
6144	cbsnyt	1	33	2	2	1	403

- The weight is listed as just another survey variable
- But they are actually constructed *after* the survey
- Weights $w_i = g(X_i, \theta)$

Example: CBS/New York Times pre-election polls

id	org	y	state	edu	age	adults	weight
6140	cbsnyt	NA	7	3	1	2	923
6141	cbsnyt	1	39	4	2	2	558
6142	cbsnyt	0	31	2	4	1	448
6143	cbsnyt	0	7	3	1	2	923
6144	cbsnyt	1	33	2	2	1	403

- The weight is listed as just another survey variable
- But they are actually constructed *after* the survey
- Weights $w_i = g(X_i, \theta)$

Example: CBS/New York Times pre-election polls

id	org	y	state	edu	age	adults	weight
6140	cbsnyt	NA	7	3	1	2	923
6141	cbsnyt	1	39	4	2	2	558
6142	cbsnyt	0	31	2	4	1	448
6143	cbsnyt	0	7	3	1	2	923
6144	cbsnyt	1	33	2	2	1	403

- The weight is listed as just another survey variable
- But they are actually constructed *after* the survey
- Weights $w_i = g(X_i, \theta)$

Example: CBS/New York Times pre-election polls

id	org	y	state	edu	age	adults	weight
6140	cbsnyt	NA	7	3	1	2	923
6141	cbsnyt	1	39	4	2	2	558
6142	cbsnyt	0	31	2	4	1	448
6143	cbsnyt	0	7	3	1	2	923
6144	cbsnyt	1	33	2	2	1	403

- The weight is listed as just another survey variable
- But they are actually constructed *after* the survey
- Weights $w_i = g(X_i, \theta)$

Story 2: weights for household size

- Telephone survey of households
  - Interview one adult in each sampled household
  - Pr(selection) $\propto$ 1/(# adults in household)
  - Obvious survey weight is proportional to household size
- But ... we can do better (Gelman and Little, 1998):

Story 2: weights for household size

- Telephone survey of households
  - Interview one adult in each sampled household
  - Pr(selection) $\propto$ 1/(# adults in household)
  - Obvious survey weight: # adults in household
- But ... we can do better (Gelman and Little, 1998):

Story 2: weights for household size

- Telephone survey of households
  - Interview one adult in each sampled household
  - Pr(selection) $\propto 1/(\#$ adults in household$)$
  - Obvious survey weight: # adults in household
- But ... we can do better (Gelman and Little, 1998):

Story 2: weights for household size

- Telephone survey of households
  - Interview one adult in each sampled household
  - Pr(selection) $\propto$ 1/(# adults in household)
  - Obvious survey weight: # adults in household
- But . . . we can do better (Gelman and Little, 1998):
  - Inverse-weighted estimates of household characteristics have bias and low efficiency
  - Survey weights by analogy to sampling theory: proportional to # adults in HH

Story 2: weights for household size

- Telephone survey of households
  - Interview one adult in each sampled household
  - Pr(selection) $\propto 1/(\#$ adults in household$)$
  - Obvious survey weight: $\#$ adults in household
- But . . . we can do better (Gelman and Little, 1998):
  - Instead of weights 1, 2, 3, 4, set weights to 1.0, 1.4, 1.7, 2.0
  - Lower bias *and* lower variance
  - Use weights by analogy to one-way poststratification, treating household size as a poststratification variable

Story 2: weights for household size

- Telephone survey of households
  - Interview one adult in each sampled household
  - Pr(selection) $\propto 1/(\#$ adults in household$)$
  - Obvious survey weight: $\#$ adults in household
- But ... we can do better (Gelman and Little, 1998):
  - Instead of weights 1, 2, 3, 4, set weights to 1.0, 1.4, 1.7, 2.0
  - Lower bias *and* lower variance
  - Set weights by matching to census numbers: sampling probabilities don't matter at all!

Story 2: weights for household size

- Telephone survey of households
  - Interview one adult in each sampled household
  - Pr(selection) $\propto 1/(\#$ adults in household$)$
  - Obvious survey weight: $\#$ adults in household
- But ... we can do better (Gelman and Little, 1998):
  - Instead of weights 1, 2, 3, 4, set weights to 1.0, 1.4, 1.7, 2.0
  - Lower bias *and* lower variance
  - Set weights by matching to census numbers: sampling probabilities don't matter at all!

Story 2: weights for household size

- Telephone survey of households
  - Interview one adult in each sampled household
  - Pr(selection) $\propto 1/(\#$ adults in household$)$
  - Obvious survey weight: $\#$ adults in household
- But ... we can do better (Gelman and Little, 1998):
  - Instead of weights 1, 2, 3, 4, set weights to 1.0, 1.4, 1.7, 2.0
  - Lower bias *and* lower variance
  - Set weights by matching to census numbers: sampling probabilities don't matter at all!

Story 2: weights for household size

- Telephone survey of households
  - Interview one adult in each sampled household
  - Pr(selection) $\propto 1/(\#$ adults in household$)$
  - Obvious survey weight: $\#$ adults in household
- But ... we can do better (Gelman and Little, 1998):
  - Instead of weights 1, 2, 3, 4, set weights to 1.0, 1.4, 1.7, 2.0
  - Lower bias *and* lower variance
  - Set weights by matching to census numbers: sampling probabilities don't matter at all!

The poststratification framework

- Goal is to estimate population average, θ
- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Point estimate $\hat{\theta} = \frac{\sum_j N_j \hat{\theta}_j}{\sum_j N_j}$
- Cells j might be determined by sex, age, education, ethnicity, ...
- Some estimators:

The poststratification framework

- Goal is to estimate population average, θ
- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Point estimate $\hat{\theta} = \frac{\sum_j N_j \hat{\theta}_j}{\sum_j N_j}$
- Cells j might be determined by sex, age, education, ethnicity, ...
- Some estimators:

The poststratification framework

- Goal is to estimate population average, θ
- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Point estimate $\hat{\theta} = \frac{\sum_j N_j \hat{\theta}_j}{\sum_j N_j}$
- Cells j might be determined by sex, age, education, ethnicity, ...
- Some estimators:

The poststratification framework

- Goal is to estimate population average, θ
- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Point estimate $\hat{\theta} = \frac{\sum_j N_j \hat{\theta}_j}{\sum_j N_j}$
- Cells j might be determined by sex, age, education, ethnicity, ...
- Some estimators:
  - Simple poststratification
  - Raking
  - Bayes shrinkage toward the mean
  - Bayes shrinkage toward a hierarchical model

The poststratification framework

- Goal is to estimate population average, θ
- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Point estimate $\hat{\theta} = \frac{\sum_j N_j \hat{\theta}_j}{\sum_j N_j}$
- Cells j might be determined by sex, age, education, ethnicity, ...
- Some estimators:
  - Simple poststratification: $\hat{\theta}_j = \bar{y}_j$
  - Sample mean: $\hat{\theta}_j = \bar{y}$
  - Bayes shrinkage estimate: $\hat{\theta}_j = \frac{n_j/\sigma_y^2 \bar{y}_j + 1/\tau^2 \mu}{n_j/\sigma_y^2 + 1/\tau^2}$

The poststratification framework

- Goal is to estimate population average, θ
- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Point estimate $\hat{\theta} = \frac{\sum_j N_j \hat{\theta}_j}{\sum_j N_j}$
- Cells j might be determined by sex, age, education, ethnicity, ...
- Some estimators:
  - Simple poststratification: $\hat{\theta}_j = \bar{y}_j$
  - Sample mean: $\hat{\theta}_j = \bar{y}$
  - Bayesian compromises: model θ_j given covariates X_j

The poststratification framework

- Goal is to estimate population average, θ
- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Point estimate $\hat{\theta} = \frac{\sum_j N_j \hat{\theta}_j}{\sum_j N_j}$
- Cells j might be determined by sex, age, education, ethnicity, ...
- Some estimators:
  - Simple poststratification: $\hat{\theta}_j = \bar{y}_j$
  - Sample mean: $\hat{\theta}_j = \bar{y}$
  - Bayesian compromises: model θ_j given covariates X_j

The poststratification framework

- Goal is to estimate population average, θ
- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Point estimate $\hat{\theta} = \frac{\sum_j N_j \hat{\theta}_j}{\sum_j N_j}$
- Cells j might be determined by sex, age, education, ethnicity, ...
- Some estimators:
  - Simple poststratification: $\hat{\theta}_j = \bar{y}_j$
  - Sample mean: $\hat{\theta}_j = \bar{y}$
  - Bayesian compromises: model θ_j given covariates X_j

The poststratification framework

- Goal is to estimate population average, θ
- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Point estimate $\hat{\theta} = \frac{\sum_j N_j \hat{\theta}_j}{\sum_j N_j}$
- Cells j might be determined by sex, age, education, ethnicity, ...
- Some estimators:
  - Simple poststratification: $\hat{\theta}_j = \bar{y}_j$
  - Sample mean: $\hat{\theta}_j = \bar{y}$
  - Bayesian compromises: model θ_j given covariates X_j

Complications

- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Many cells j ($2 \times 4 \times 5 \times 4 \times 50$): need complicated model with many levels of interactions
- Adjusting for non-census variables (for example, religion): need to model the N_j's
- Regression of y on x

Complications

- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Many cells j ($2 \times 4 \times 5 \times 4 \times 50$): need complicated model with many levels of interactions
- Adjusting for non-census variables (for example, religion): need to model the N_j's
- Regression of y on x

Complications

- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Many cells j ($2 \times 4 \times 5 \times 4 \times 50$): need complicated model with many levels of interactions
- Adjusting for non-census variables (for example, religion): need to model the N_j's
- Regression of y on x

Complications

- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Many cells j ($2 \times 4 \times 5 \times 4 \times 50$): need complicated model with many levels of interactions
- Adjusting for non-census variables (for example, religion): need to model the N_j's
- Regression of y on x
  - Must model $y|x$ within each cell j
  - Then average over cells to estimate $E(y)$ as a function of x

Complications

- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Many cells j ($2 \times 4 \times 5 \times 4 \times 50$): need complicated model with many levels of interactions
- Adjusting for non-census variables (for example, religion): need to model the N_j's
- Regression of y on x
  - Must model $y|x$ within each cell j
  - Then average over cells to estimate $\mathrm{E}(y)$ as a function of x

Complications

- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Many cells j ($2 \times 4 \times 5 \times 4 \times 50$): need complicated model with many levels of interactions
- Adjusting for non-census variables (for example, religion): need to model the N_j's
- Regression of y on x
  - Must model $y|x$ within each cell j
  - Then average over cells to estimate $\mathrm{E}(y)$ as a function of x

Complications

- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Many cells j ($2 \times 4 \times 5 \times 4 \times 50$): need complicated model with many levels of interactions
- Adjusting for non-census variables (for example, religion): need to model the N_j's
- Regression of y on x
  - Must model $y|x$ within each cell j
  - Then average over cells to estimate $\mathrm{E}(y)$ as a function of x

2000: State-level support (orange) or opposition (green) on school vouchers, relative to the national average of 45% support

Orange and green colors correspond to states where support for vouchers was greater or less than the national average.
The seven ethnic/religious cagetories are mutually exclusive. "Evangelicals" includes Mormons as well as born-again Protestants.
Where a category represents less than 1% of the voters of a state, the state is left blank.