# Survey weighting and hierarchical regression: some successes and struggles

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University

18 October 2004

# Survey weighting and regression modeling

- Success: state-level opinions from national polls
  - Hierarchical modeling and poststratification
- Struggle: the Social Indicators Survey of NYC families
- Unrelated topic: interactions in before-after studies
- collaborators:
  - David Park, Dept of Political Science, Washington University
  - Joe Bafumi, Dept of Political Science, Columbia University
  - Shouhao Zhao, Dept of Statistics, Columbia University
  - John Carlin, Dept of Biostatistics, University of Melbourne
  - Julien Teitler and Sandra Garcia, Social Work, Columbia Univ
  - Rod Little, Dept of Biostatistics, University of Michigan

# Survey weighting and regression modeling

- Success: state-level opinions from national polls
  - Hierarchical modeling and poststratification
- Struggle: the Social Indicators Survey of NYC families
  - Reconciling survey weighting and regression
  - Weighting from a hierarchical Bayes perspective
- Unrelated topic: interactions in before-after studies
- collaborators:
  - David Park, Dept of Political Science, Washington University
  - Joe Bafumi, Dept of Political Science, Columbia University
  - Shouhao Zhao, Dept of Statistics, Columbia University
  - John Carlin, Dept of Biostatistics, University of Melbourne
  - Julien Teitler and Sandra Garcia, Social Work, Columbia Univ
  - Rod Little, Dept of Biostatistics, University of Michigan

# Survey weighting and regression modeling

- Success: state-level opinions from national polls
  - Hierarchical modeling and poststratification
- Struggle: the Social Indicators Survey of NYC families
  - Reconciling survey weighting and regression
  - Weighting from a hierarchical Bayes perspective
- Unrelated topic: interactions in before-after studies
- collaborators:
  - David Park, Dept of Political Science, Washington University
  - Joe Bafumi, Dept of Political Science, Columbia University
  - Shouhao Zhao, Dept of Statistics, Columbia University
  - John Carlin, Dept of Biostatistics, University of Melbourne
  - Julien Teitler and Sandra Garcia, Social Work, Columbia Univ
  - Rod Little, Dept of Biostatistics, University of Michigan

# Survey weighting and regression modeling

- Success: state-level opinions from national polls
  - Hierarchical modeling and poststratification
- Struggle: the Social Indicators Survey of NYC families
  - Reconciling survey weighting and regression
  - Weighting from a hierarchical Bayes perspective
- Unrelated topic: interactions in before-after studies
- collaborators:
  - David Park, Dept of Political Science, Washington University
  - Joe Bafumi, Dept of Political Science, Columbia University
  - Shouhao Zhao, Dept of Statistics, Columbia University
  - John Carlin, Dept of Biostatistics, University of Melbourne
  - Julien Teitler and Sandra Garcia, Social Work, Columbia Univ
  - Rod Little, Dept of Biostatistics, University of Michigan

# Survey weighting and regression modeling

- Success: state-level opinions from national polls
  - Hierarchical modeling and poststratification
- Struggle: the Social Indicators Survey of NYC families
  - Reconciling survey weighting and regression
  - Weighting from a hierarchical Bayes perspective
- Unrelated topic: interactions in before-after studies

- collaborators:
  - David Park, Dept of Political Science, Washington University
  - Joe Bafumi, Dept of Political Science, Columbia University
  - Shouhao Zhao, Dept of Statistics, Columbia University
  - John Carlin, Dept of Biostatistics, University of Melbourne
  - Julien Teitler and Sandra Garcia, Social Work, Columbia Univ
  - Rod Little, Dept of Biostatistics, University of Michigan

# Survey weighting and regression modeling

- Success: state-level opinions from national polls
  - Hierarchical modeling and poststratification
- Struggle: the Social Indicators Survey of NYC families
  - Reconciling survey weighting and regression
  - Weighting from a hierarchical Bayes perspective
- Unrelated topic: interactions in before-after studies

- collaborators:
  - David Park, Dept of Political Science, Washington University
  - Joe Bafumi, Dept of Political Science, Columbia University
  - Shouhao Zhao, Dept of Statistics, Columbia University
  - John Carlin, Dept of Biostatistics, University of Melbourne
  - Julien Teitler and Sandra Garcia, Social Work, Columbia Univ
  - Rod Little, Dept of Biostatistics, University of Michigan

# Survey weighting and regression modeling

- Success: state-level opinions from national polls
  - Hierarchical modeling and poststratification
- Struggle: the Social Indicators Survey of NYC families
  - Reconciling survey weighting and regression
  - Weighting from a hierarchical Bayes perspective
- Unrelated topic: interactions in before-after studies

- collaborators:
  - David Park, Dept of Political Science, Washington University
  - Joe Bafumi, Dept of Political Science, Columbia University
  - Shouhao Zhao, Dept of Statistics, Columbia University
  - John Carlin, Dept of Biostatistics, University of Melbourne
  - Julien Teitler and Sandra Garcia, Social Work, Columbia Univ
  - Rod Little, Dept of Biostatistics, University of Michigan

# Survey weighting and regression modeling

- Success: state-level opinions from national polls
  - Hierarchical modeling and poststratification
- Struggle: the Social Indicators Survey of NYC families
  - Reconciling survey weighting and regression
  - Weighting from a hierarchical Bayes perspective
- Unrelated topic: interactions in before-after studies

- collaborators:
  - David Park, Dept of Political Science, Washington University
  - Joe Bafumi, Dept of Political Science, Columbia University
  - Shouhao Zhao, Dept of Statistics, Columbia University
  - John Carlin, Dept of Biostatistics, University of Melbourne
  - Julien Teitler and Sandra Garcia, Social Work, Columbia Univ
  - Rod Little, Dept of Biostatistics, University of Michigan

# National opinion trends

# State-level opinion trends

- Goal: estimating time series within each state
- One poll at a time: small-area estimation
- It works! Validated for pre-election polls
- Combining surveys: hierarchical model for parallel time series
- Straightforward hierarchical modeling + poststratification
- Poststratification cells: sex $\times$ ethnicity $\times$ age $\times$ education $\times$ state

# State-level opinion trends

- Goal: estimating time series within each state
- One poll at a time: small-area estimation
- It works! Validated for pre-election polls
- Combining surveys: hierarchical model for parallel time series
- Straightforward hierarchical modeling + poststratification
- Poststratification cells: sex × ethnicity × age × education × state

# State-level opinion trends

- Goal: estimating time series within each state
- One poll at a time: small-area estimation
- It works! Validated for pre-election polls
- Combining surveys: hierarchical model for parallel time series
- Straightforward hierarchical modeling + poststratification
- Poststratification cells: sex × ethnicity × age × education × state

# State-level opinion trends

- Goal: estimating time series within each state
- One poll at a time: small-area estimation
- It works! Validated for pre-election polls
- Combining surveys: hierarchical model for parallel time series
- Straightforward hierarchical modeling + poststratification
- Poststratification cells: sex × ethnicity × age × education × state

# State-level opinion trends

- Goal: estimating time series within each state
- One poll at a time: small-area estimation
- It works! Validated for pre-election polls
- Combining surveys: hierarchical model for parallel time series
- Straightforward hierarchical modeling + poststratification
- Poststratification cells: sex × ethnicity × age × education × state

# State-level opinion trends

- Goal: estimating time series within each state
- One poll at a time: small-area estimation
- It works! Validated for pre-election polls
- Combining surveys: hierarchical model for parallel time series
- Straightforward hierarchical modeling + poststratification
- Poststratification cells: sex $\times$ ethnicity $\times$ age $\times$ education $\times$ state

# Hierarchical modeling to estimate state opinions

- Hierarchical model for the data
  - $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
  - $X$ includes demographic and geographic predictors
  - Hierarchical model for the 50 state coefficients

# Hierarchical modeling to estimate state opinions

- Hierarchical model for the data
  - $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
  - $X$ includes demographic and geographic predictors
  - Hierarchical model for the 50 state coefficients
  - Bayesian inference, summarize by posterior simulations of $\beta$:

| Simulation | $\theta_1$ | $\cdots$ | $\theta_{75}$ |
|---|---|---|---|
| 1 | ** | $\cdots$ | ** |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| 1000 | ** | $\cdots$ | ** |

# Hierarchical modeling to estimate state opinions

- Hierarchical model for the data
  - $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
  - $X$ includes demographic and geographic predictors
  - Hierarchical model for the 50 state coefficients
  - Bayesian inference, summarize by posterior simulations of $\beta$:

| Simulation | $\theta_1$ | $\cdots$ | $\theta_{75}$ |
|---|---|---|---|
| 1 | ** | $\cdots$ | ** |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| 1000 | ** | $\cdots$ | ** |

# Hierarchical modeling to estimate state opinions

- Hierarchical model for the data
  - $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
  - $X$ includes demographic and geographic predictors
  - Hierarchical model for the 50 state coefficients
  - Bayesian inference, summarize by posterior simulations of $\beta$:

| Simulation | $\theta_1$ | $\cdots$ | $\theta_{75}$ |
|---|---|---|---|
| 1 | ** | $\cdots$ | ** |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| 1000 | ** | $\cdots$ | ** |

# Hierarchical modeling to estimate state opinions

- Hierarchical model for the data
  - $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
  - $X$ includes demographic and geographic predictors
  - Hierarchical model for the 50 state coefficients
  - Bayesian inference, summarize by posterior simulations of $\beta$:

| Simulation | $\theta_1$ | $\cdots$ | $\theta_{75}$ |
|---|---|---|---|
| 1 | ** | $\cdots$ | ** |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| 1000 | ** | $\cdots$ | ** |

# Hierarchical modeling to estimate state opinions

- Hierarchical model for the data
  - $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
  - $X$ includes demographic and geographic predictors
  - Hierarchical model for the 50 state coefficients
  - Bayesian inference, summarize by posterior simulations of $\beta$:

| Simulation | $\theta_1$ | $\cdots$ | $\theta_{75}$ |
|---|---|---|---|
| 1 | ** | $\cdots$ | ** |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| 1000 | ** | $\cdots$ | ** |

# Poststratification to estimate state opinions

- Implied inference for $\theta_j = \text{logit}^{-1}(X\beta)$ in each of 3264 cells $j$ (e.g., black female, age 18–29, college graduate, Connecticut)
- Poststratification
  - Within each state $s$, average over 64 cells:
    $\sum_{j \in s} N_j \theta_j \Big/ \sum_{j \in s} N_j$
    $N_j$ = population in cell $j$ (from Census)
    1000 simulation draws propagate to uncertainty for each $\theta_j$

# Poststratification to estimate state opinions

- Implied inference for $\theta_j = \text{logit}^{-1}(X\beta)$ in each of 3264 cells $j$ (e.g., black female, age 18–29, college graduate, Connecticut)
- Poststratification
  - Within each state $s$, average over 64 cells: $\sum_{j \in s} N_j \theta_j \Big/ \sum_{j \in s} N_j$
  - $N_j$ = population in cell $j$ (from Census)
  - 1000 simulation draws propagate to uncertainty for each $\theta_s$

# Poststratification to estimate state opinions

- Implied inference for $\theta_j = \text{logit}^{-1}(X\beta)$ in each of 3264 cells $j$ (e.g., black female, age 18–29, college graduate, Connecticut)
- Poststratification
  - Within each state $s$, average over 64 cells: $\sum_{j \in s} N_j \theta_j \Big/ \sum_{j \in s} N_j$
  - $N_j$ = population in cell $j$ (from Census)
  - 1000 simulation draws propagate to uncertainty for each $\theta_j$

# Poststratification to estimate state opinions

- Implied inference for $\theta_j = \text{logit}^{-1}(X\beta)$ in each of 3264 cells $j$ (e.g., black female, age 18–29, college graduate, Connecticut)
- Poststratification
  - Within each state $s$, average over 64 cells: $\sum_{j \in s} N_j \theta_j \Big/ \sum_{j \in s} N_j$
  - $N_j$ = population in cell $j$ (from Census)
  - 1000 simulation draws propagate to uncertainty for each $\theta_j$

# Poststratification to estimate state opinions

- Implied inference for $\theta_j = \text{logit}^{-1}(X\beta)$ in each of 3264 cells $j$ (e.g., black female, age 18–29, college graduate, Connecticut)
- Poststratification
  - Within each state $s$, average over 64 cells: $\sum_{j \in s} N_j \theta_j \Big/ \sum_{j \in s} N_j$
  - $N_j$ = population in cell $j$ (from Census)
  - 1000 simulation draws propagate to uncertainty for each $\theta_j$

# Poststratification to estimate state opinions

- Implied inference for $\theta_j = \text{logit}^{-1}(X\beta)$ in each of 3264 cells $j$ (e.g., black female, age 18–29, college graduate, Connecticut)
- Poststratification
  - Within each state $s$, average over 64 cells: $\sum_{j \in s} N_j \theta_j \Big/ \sum_{j \in s} N_j$
  - $N_j =$ population in cell $j$ (from Census)
  - 1000 simulation draws propagate to uncertainty for each $\theta_j$

# CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:

# CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:

# CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:

# CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:

## CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:
  - No pooling: 10.4%
  - Complete pooling: 5.4%
  - Hierarchical model with poststratification: 4.5%

# CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:
  - No pooling: 10.4%
  - Complete pooling: 5.4%
  - Hierarchical model with poststratification: 4.5%

# CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:
  - No pooling: 10.4%
  - Complete pooling: 5.4%
  - Hierarchical model with poststratification: 4.5%

# CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:
  - No pooling: 10.4%
  - Complete pooling: 5.4%
  - Hierarchical model with poststratification: 4.5%

# CBS/New York Times pre-election polls from 1988

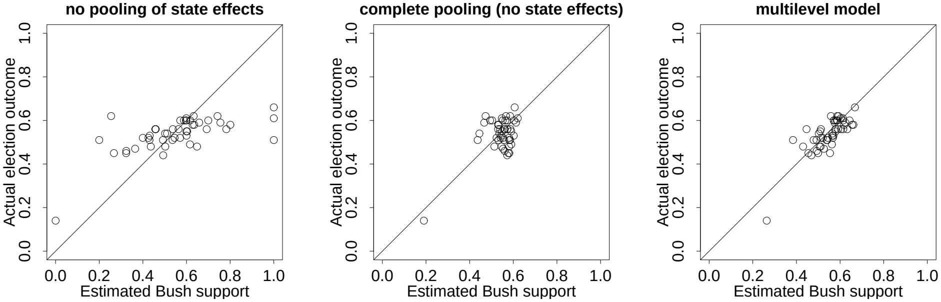
- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:
  - No pooling: 10.4%
  - Complete pooling: 5.4%
  - Hierarchical model with poststratification: 4.5%

# CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:
  - No pooling: 10.4%
  - Complete pooling: 5.4%
  - Hierarchical model with poststratification: 4.5%

# Validation study: comparison of state errors

1988 election outcome vs. poll estimate



# National opinion trends

# Death penalty opinions from General Social Survey, 1975–2000

- Goal: time series of opinions in each state
- For each state $s$ at time $t$, sum over 64 poststrat cells $j$: $\sum_{j \in s} N_{jt} \theta_{jt} \Big/ \sum_{j \in s} N_{jt}$
- Logistic regression: $\theta_{jt} = \text{logit}^{-1}((X\beta)_{jt})$
- Time series model for the state coefficients $\beta$
- Estimate the $\beta$'s from the survey data
- We just presented the model in "reverse order"

# Death penalty opinions from General Social Survey, 1975–2000

- Goal: time series of opinions in each state
- For each state $s$ at time $t$, sum over 64 poststrat cells $j$: $\sum_{j \in s} N_{jt}\theta_{jt} \Big/ \sum_{j \in s} N_{jt}$
- Logistic regression: $\theta_{jt} = \text{logit}^{-1}((X\beta)_{jt})$
- Time series model for the state coefficients $\beta$
- Estimate the $\beta$'s from the survey data
- We just presented the model in "reverse order"

# Death penalty opinions from General Social Survey, 1975–2000

- Goal: time series of opinions in each state
- For each state $s$ at time $t$, sum over 64 poststrat cells $j$: $\sum_{j \in s} N_{jt}\theta_{jt} \Big/ \sum_{j \in s} N_{jt}$
- Logistic regression: $\theta_{jt} = \text{logit}^{-1}((X\beta)_{jt})$
- Time series model for the state coefficients $\beta$
- Estimate the $\beta$'s from the survey data
- We just presented the model in "reverse order"

# Death penalty opinions from General Social Survey, 1975–2000

- Goal: time series of opinions in each state
- For each state $s$ at time $t$, sum over 64 poststrat cells $j$: $\sum_{j \in s} N_{jt}\theta_{jt} \Big/ \sum_{j \in s} N_{jt}$
- Logistic regression: $\theta_{jt} = \text{logit}^{-1}((X\beta)_{jt})$
- Time series model for the state coefficients $\beta$
- Estimate the $\beta$'s from the survey data
- We just presented the model in "reverse order"

# Death penalty opinions from General Social Survey, 1975–2000

- Goal: time series of opinions in each state
- For each state $s$ at time $t$, sum over 64 poststrat cells $j$: $\sum_{j\in s} N_{jt}\theta_{jt} \Big/ \sum_{j\in s} N_{jt}$
- Logistic regression: $\theta_{jt} = \text{logit}^{-1}((X\beta)_{jt})$
- Time series model for the state coefficients $\beta$
- Estimate the $\beta$'s from the survey data
- We just presented the model in "reverse order"

# Death penalty opinions from General Social Survey, 1975–2000

- Goal: time series of opinions in each state
- For each state $s$ at time $t$, sum over 64 poststrat cells $j$: $\sum_{j \in s} N_{jt}\theta_{jt} \Big/ \sum_{j \in s} N_{jt}$
- Logistic regression: $\theta_{jt} = \text{logit}^{-1}((X\beta)_{jt})$
- Time series model for the state coefficients $\beta$
- Estimate the $\beta$'s from the survey data
- We just presented the model in "reverse order"

# Death penalty opinions from General Social Survey, 1975–2000

- Goal: time series of opinions in each state
- For each state $s$ at time $t$, sum over 64 poststrat cells $j$: $\sum_{j \in s} N_{jt}\theta_{jt} \Big/ \sum_{j \in s} N_{jt}$
- Logistic regression: $\theta_{jt} = \text{logit}^{-1}((X\beta)_{jt})$
- Time series model for the state coefficients $\beta$
- Estimate the $\beta$'s from the survey data
- We just presented the model in "reverse order"

# Death penalty opinion trends by state

# Survey weighting is a mess

- Using weights
  - Weighted mean: $\bar{y}_w = \sum_{i=1}^n w_i y_i / \sum_{i=1}^n w_i$
  - Estimating a ratio: $r_w = \sum_{i=1}^n w_i y_i / \sum_{i=1}^n w_i x_i$
  - Estimating anything more complicated: ???
- Regression modeling as an alternative

# Survey weighting is a mess

- Using weights
  - Weighted mean: $\bar{y}_w = \sum_{i=1}^n w_i y_i \big/ \sum_{i=1}^n w_i$
  - Estimating a ratio: $r_w = \sum_{i=1}^n w_i y_i \big/ \sum_{i=1}^n w_i x_i$
  - Estimating anything more complicated: ???
- Regression modeling as an alternative

# Survey weighting is a mess

- Using weights
  - Weighted mean: $\bar{y}_w = \sum_{i=1}^n w_i y_i \big/ \sum_{i=1}^n w_i$
  - Estimating a ratio: $r_w = \sum_{i=1}^n w_i y_i \big/ \sum_{i=1}^n w_i x_i$
  - Estimating anything more complicated: ???
- Regression modeling as an alternative

# Survey weighting is a mess

- Using weights
  - Weighted mean: $\bar{y}_w = \sum_{i=1}^n w_i y_i \big/ \sum_{i=1}^n w_i$
  - Estimating a ratio: $r_w = \sum_{i=1}^n w_i y_i \big/ \sum_{i=1}^n w_i x_i$
  - Estimating anything more complicated: ???
- Regression modeling as an alternative
  - Need to control for many potential confounders
  - Hierarchical modeling as a (potential) solution

# Survey weighting is a mess

- Using weights
  - Weighted mean: $\bar{y}_w = \sum_{i=1}^n w_i y_i \big/ \sum_{i=1}^n w_i$
  - Estimating a ratio: $r_w = \sum_{i=1}^n w_i y_i \big/ \sum_{i=1}^n w_i x_i$
  - Estimating anything more complicated: ???
- Regression modeling as an alternative
  - Need to control for many potential confounders
  - Hierarchical modeling as a (potential) solution

# Survey weighting is a mess

- Using weights
  - Weighted mean: $\bar{y}_w = \sum_{i=1}^n w_i y_i \big/ \sum_{i=1}^n w_i$
  - Estimating a ratio: $r_w = \sum_{i=1}^n w_i y_i \big/ \sum_{i=1}^n w_i x_i$
  - Estimating anything more complicated: ???
- Regression modeling as an alternative
  - Need to control for many potential confounders
  - Hierarchical modeling as a (potential) solution

# Survey weighting is a mess

- Using weights
  - Weighted mean: $\bar{y}_w = \sum_{i=1}^n w_i y_i \big/ \sum_{i=1}^n w_i$
  - Estimating a ratio: $r_w = \sum_{i=1}^n w_i y_i \big/ \sum_{i=1}^n w_i x_i$
  - Estimating anything more complicated: ???
- Regression modeling as an alternative
  - Need to control for many potential confounders
  - Hierarchical modeling as a (potential) solution

# Survey weighting is a mess

- Using weights
  - Weighted mean: $\bar{y}_w = \sum_{i=1}^n w_i y_i \Big/ \sum_{i=1}^n w_i$
  - Estimating a ratio: $r_w = \sum_{i=1}^n w_i y_i \Big/ \sum_{i=1}^n w_i x_i$
  - Estimating anything more complicated: ???
- Regression modeling as an alternative
  - Need to control for many potential confounders
  - Hierarchical modeling as a (potential) solution

# Where do weights come from?

- Survey weights are not inverse probabilities of selection
- Simple theoretical example
- CBS/New York Times pre-election polls
- NYC Social Indicators Survey

# Where do weights come from?

- Survey weights are not inverse probabilities of selection
- Simple theoretical example
- CBS/New York Times pre-election polls
- NYC Social Indicators Survey

# Where do weights come from?

- Survey weights are not inverse probabilities of selection
- Simple theoretical example
- CBS/New York Times pre-election polls
- NYC Social Indicators Survey

# Where do weights come from?

- Survey weights are not inverse probabilities of selection
- Simple theoretical example
- CBS/New York Times pre-election polls
- NYC Social Indicators Survey

# Where do weights come from?

- Survey weights are not inverse probabilities of selection
- Simple theoretical example
- CBS/New York Times pre-election polls
- NYC Social Indicators Survey

# Simple theoretical example

- 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, $\frac{48}{40}$ 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

# Simple theoretical example

- 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

# Simple theoretical example

- 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

# Simple theoretical example

- 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

# Simple theoretical example

- 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

# Simple theoretical example

- 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

# Simple theoretical example

- 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

# 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 $w_i = g(X_i, \theta)$:

- Goal is to estimate national and statewide averages

# 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 $w_i = g(X_i, \theta)$:
  - $X_i$ are sex, age, education, ...
  - $\theta$ are parameters depending on the entire survey and on Census population info
- Goal is to estimate national and statewide averages

# 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 $w_i = g(X_i, \theta)$:
  - $X_i$ are sex, age, education, ...
  - $\theta$ are parameters depending on the entire survey and on Census population info
- Goal is to estimate national and statewide averages

# 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 $w_i = g(X_i, \theta)$:
  - $X_i$ are sex, age, education, ...
  - $\theta$ are parameters depending on the entire survey and on Census population info
- Goal is to estimate national and statewide averages

# 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 $w_i = g(X_i, \theta)$:
  - $X_i$ are sex, age, education, ...
  - $\theta$ are parameters depending on the entire survey and on Census population info
- Goal is to estimate national and statewide averages

# 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 $w_i = g(X_i, \theta)$:
  - $X_i$ are sex, age, education, ...
  - $\theta$ are parameters depending on the entire survey and on Census population info
- Goal is to estimate national and statewide averages

# 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 $w_i = g(X_i, \theta)$:
  - $X_i$ are sex, age, education, ...
  - $\theta$ are parameters depending on the entire survey and on Census population info
- Goal is to estimate national and statewide averages

# Social Indicators Survey

- Telephone survey every 2 years of NYC families
- Administered by Columbia University School of Social Work
- Questions such as, "Do you rate the schools as poor, fair, good, or very good?"
- Weighting to match Current Population Survey: #adults and children in family, marital status, ethnicity, age, education
- Goal is to estimate changes over time
- Bias-variance tradeoff in constructing weights:

# Social Indicators Survey

- Telephone survey every 2 years of NYC families
- Administered by Columbia University School of Social Work
- Questions such as, "Do you rate the schools as poor, fair, good, or very good?"
- Weighting to match Current Population Survey: #adults and children in family, marital status, ethnicity, age, education
- Goal is to estimate changes over time
- Bias-variance tradeoff in constructing weights:

# Social Indicators Survey

- Telephone survey every 2 years of NYC families
- Administered by Columbia University School of Social Work
- Questions such as, "Do you rate the schools as poor, fair, good, or very good?"
- Weighting to match Current Population Survey: #adults and children in family, marital status, ethnicity, age, education
- Goal is to estimate changes over time
- Bias-variance tradeoff in constructing weights:

# Social Indicators Survey

- Telephone survey every 2 years of NYC families
- Administered by Columbia University School of Social Work
- Questions such as, "Do you rate the schools as poor, fair, good, or very good?"
- Weighting to match Current Population Survey: #adults and children in family, marital status, ethnicity, age, education
- Goal is to estimate changes over time
- Bias-variance tradeoff in constructing weights:

# Social Indicators Survey

- Telephone survey every 2 years of NYC families
- Administered by Columbia University School of Social Work
- Questions such as, "Do you rate the schools as poor, fair, good, or very good?"
- Weighting to match Current Population Survey: #adults and children in family, marital status, ethnicity, age, education
- Goal is to estimate changes over time
- Bias-variance tradeoff in constructing weights:
  - Weights adjust for potential confounders
  - But we want weighted estimates to be stable

# Social Indicators Survey

- Telephone survey every 2 years of NYC families
- Administered by Columbia University School of Social Work
- Questions such as, "Do you rate the schools as poor, fair, good, or very good?"
- Weighting to match Current Population Survey: #adults and children in family, marital status, ethnicity, age, education
- Goal is to estimate changes over time
- Bias-variance tradeoff in constructing weights:
  - Weights adjust for potential confounders
  - But we want weighted estimates to be stable

# Social Indicators Survey

- Telephone survey every 2 years of NYC families
- Administered by Columbia University School of Social Work
- Questions such as, "Do you rate the schools as poor, fair, good, or very good?"
- Weighting to match Current Population Survey: #adults and children in family, marital status, ethnicity, age, education
- Goal is to estimate changes over time
- Bias-variance tradeoff in constructing weights:
  - Weights adjust for potential confounders
  - But we want weighted estimates to be stable

# Social Indicators Survey

- Telephone survey every 2 years of NYC families
- Administered by Columbia University School of Social Work
- Questions such as, "Do you rate the schools as poor, fair, good, or very good?"
- Weighting to match Current Population Survey: #adults and children in family, marital status, ethnicity, age, education
- Goal is to estimate changes over time
- Bias-variance tradeoff in constructing weights:
  - Weights adjust for potential confounders
  - But we want weighted estimates to be stable

# Social Indicators Survey

- Telephone survey every 2 years of NYC families
- Administered by Columbia University School of Social Work
- Questions such as, "Do you rate the schools as poor, fair, good, or very good?"
- Weighting to match Current Population Survey: #adults and children in family, marital status, ethnicity, age, education
- Goal is to estimate changes over time
- Bias-variance tradeoff in constructing weights:
  - Weights adjust for potential confounders
  - But we want weighted estimates to be stable

# Poststratification for the CBS polls

- We don't actually use the "weights"
- We model $y$ conditional on the variables used in the weighting
- These define poststratification cells $j = 1, \ldots, J = 3264$
- $2 \times 2 \times 4 \times 4 \times 51$: sex $\times$ ethnicity $\times$ age $\times$ education $\times$ state
- Poststratified average, $\theta = \frac{\sum_{j=1}^{J} N_j \theta_j}{\sum_{j=1}^{J} N_j}$
- $N_j$ = population in cell $j$ (from Census)
- Same Census that was used to create the survey weights

# Poststratification for the CBS polls

- We don't actually use the "weights"
- We model $y$ conditional on the variables used in the weighting
- These define poststratification cells $j = 1, \ldots, J = 3264$
- $2 \times 2 \times 4 \times 4 \times 51$: sex $\times$ ethnicity $\times$ age $\times$ education $\times$ state
- Poststratified average, $\theta = \frac{\sum_{j=1}^{J} N_j \theta_j}{\sum_{j=1}^{J} N_j}$
- $N_j$ = population in cell $j$ (from Census)
- Same Census that was used to create the survey weights

# Poststratification for the CBS polls

- We don't actually use the "weights"
- We model $y$ conditional on the variables used in the weighting
- These define poststratification cells $j = 1, \ldots, J = 3264$
- $2 \times 2 \times 4 \times 4 \times 51$: sex $\times$ ethnicity $\times$ age $\times$ education $\times$ state
- Poststratified average, $\theta = \frac{\sum_{j=1}^{J} N_j \theta_j}{\sum_{j=1}^{J} N_j}$
- $N_j$ = population in cell $j$ (from Census)
- Same Census that was used to create the survey weights

# Poststratification for the CBS polls

- We don't actually use the "weights"
- We model $y$ conditional on the variables used in the weighting
- These define poststratification cells $j = 1, \ldots, J = 3264$
- $2 \times 2 \times 4 \times 4 \times 51$: sex $\times$ ethnicity $\times$ age $\times$ education $\times$ state
- Poststratified average, $\theta = \frac{\sum_{j=1}^{J} N_j \theta_j}{\sum_{j=1}^{J} N_j}$
- $N_j$ = population in cell $j$ (from Census)
- Same Census that was used to create the survey weights

# Poststratification for the CBS polls

- We don't actually use the "weights"
- We model $y$ conditional on the variables used in the weighting
- These define poststratification cells $j = 1, \ldots, J = 3264$
- $2 \times 2 \times 4 \times 4 \times 51$: sex $\times$ ethnicity $\times$ age $\times$ education $\times$ state
- Poststratified average, $\theta = \frac{\sum_{j=1}^{J} N_j \theta_j}{\sum_{j=1}^{J} N_j}$
- $N_j$ = population in cell $j$ (from Census)
- Same Census that was used to create the survey weights

# Poststratification for the CBS polls

- We don't actually use the "weights"
- We model $y$ conditional on the variables used in the weighting
- These define poststratification cells $j = 1, \ldots, J = 3264$
- $2 \times 2 \times 4 \times 4 \times 51$: sex $\times$ ethnicity $\times$ age $\times$ education $\times$ state
- Poststratified average, $\theta = \frac{\sum_{j=1}^{J} N_j \theta_j}{\sum_{j=1}^{J} N_j}$
- $N_j$ = population in cell $j$ (from Census)
- Same Census that was used to create the survey weights

# Poststratification for the CBS polls

- We don't actually use the "weights"
- We model $y$ conditional on the variables used in the weighting
- These define poststratification cells $j = 1, \ldots, J = 3264$
- $2 \times 2 \times 4 \times 4 \times 51$: sex $\times$ ethnicity $\times$ age $\times$ education $\times$ state
- Poststratified average, $\theta = \frac{\sum_{j=1}^{J} N_j \theta_j}{\sum_{j=1}^{J} N_j}$
- $N_j$ = population in cell $j$ (from Census)
- Same Census that was used to create the survey weights

# Poststratification for the CBS polls

- We don't actually use the "weights"
- We model $y$ conditional on the variables used in the weighting
- These define poststratification cells $j = 1, \dots, J = 3264$
- $2 \times 2 \times 4 \times 4 \times 51$: sex $\times$ ethnicity $\times$ age $\times$ education $\times$ state
- Poststratified average, $\theta = \frac{\sum_{j=1}^{J} N_j \theta_j}{\sum_{j=1}^{J} N_j}$
- $N_j$ = population in cell $j$ (from Census)
- Same Census that was used to create the survey weights

# Estimating time trends in NYC

- Compare 1999 and 2001 Social Indicators Surveys
- Goal is to estimate $\bar{Y}^{2001} - \bar{Y}^{1999}$, for various survey responses $y$
- Estimate from weighted average, $\bar{y}_w^{2001} - \bar{y}_w^{1999}$
- Or, estimate using regression:

# Estimating time trends in NYC

- Compare 1999 and 2001 Social Indicators Surveys
- Goal is to estimate $\bar{Y}^{2001} - \bar{Y}^{1999}$, for various survey responses $y$
- Estimate from weighted average, $\bar{y}_w^{2001} - \bar{y}_w^{1999}$
- Or, estimate using regression:
  - Combine two surveys into a single data matrix
  - Add an indicator that is 1 for 2001 and 0 for 1999
  - Fit regression, look at coefficient for the "2001" indicator

# Estimating time trends in NYC

- Compare 1999 and 2001 Social Indicators Surveys
- Goal is to estimate $\bar{Y}^{2001} - \bar{Y}^{1999}$, for various survey responses $y$
- Estimate from weighted average, $\bar{y}_w^{2001} - \bar{y}_w^{1999}$
- Or, estimate using regression:
  - Combine two surveys into a single data matrix
  - Add an indicator that is 1 for 2001 and 0 for 1999
  - Fit regression, look at coefficient for the "2001" indicator

# Estimating time trends in NYC

- Compare 1999 and 2001 Social Indicators Surveys
- Goal is to estimate $\bar{Y}^{2001} - \bar{Y}^{1999}$, for various survey responses $y$
- Estimate from weighted average, $\bar{y}_w^{2001} - \bar{y}_w^{1999}$
- Or, estimate using regression:
  - Combine two surveys into a single data matrix
  - Add an indicator that is 1 for 2001 and 0 for 1999
  - Fit regression, look at coefficient for the "2001" indicator

# Estimating time trends in NYC

- Compare 1999 and 2001 Social Indicators Surveys
- Goal is to estimate $\bar{Y}^{2001} - \bar{Y}^{1999}$, for various survey responses $y$
- Estimate from weighted average, $\bar{y}_w^{2001} - \bar{y}_w^{1999}$
- Or, estimate using regression:
    - Combine two surveys into a single data matrix
    - Add an indicator that is 1 for 2001 and 0 for 1999
    - Fit regression, look at coefficient for the "2001" indicator

# Estimating time trends in NYC

- Compare 1999 and 2001 Social Indicators Surveys
- Goal is to estimate $\bar{Y}^{2001} - \bar{Y}^{1999}$, for various survey responses $y$
- Estimate from weighted average, $\bar{y}_w^{2001} - \bar{y}_w^{1999}$
- Or, estimate using regression:
  - Combine two surveys into a single data matrix
  - Add an indicator that is 1 for 2001 and 0 for 1999
  - Fit regression, look at coefficient for the "2001" indicator

# Estimating time trends in NYC

- Compare 1999 and 2001 Social Indicators Surveys
- Goal is to estimate $\bar{Y}^{2001} - \bar{Y}^{1999}$, for various survey responses $y$
- Estimate from weighted average, $\bar{y}_w^{2001} - \bar{y}_w^{1999}$
- Or, estimate using regression:
  - Combine two surveys into a single data matrix
  - Add an indicator that is 1 for 2001 and 0 for 1999
  - Fit regression, look at coefficient for the “2001” indicator

# Comparing estimates from weighting and regression

| Question | weighted averages 1999 | weighted averages 2001 | (a) time change in percent | (b) linear regression coefficient of time |
|---|---|---|---|---|
| Adult in good/excellent health | 75% | 78% | 3.4% (2.4%) | 6.6% (1.4%) |
| Child in good/excellent health | 82% | 84% | 1.7% (1.5%) | 1.2% (1.3%) |
| Neighborhood is safe/very safe | 77% | 81% | 4.5% (2.3%) | 4.1% (1.5%) |

- The estimates can be very different!
- Which to believe?
- Same pattern with logistic regression

# Comparing estimates from weighting and regression

| Question | weighted averages 1999 | weighted averages 2001 | (a) time change in percent | (b) linear regression coefficient of time |
|---|---|---|---|---|
| Adult in good/excellent health | 75% | 78% | 3.4% (2.4%) | 6.6% (1.4%) |
| Child in good/excellent health | 82% | 84% | 1.7% (1.5%) | 1.2% (1.3%) |
| Neighborhood is safe/very safe | 77% | 81% | 4.5% (2.3%) | 4.1% (1.5%) |

- The estimates can be very different!
- Which to believe?
- Same pattern with logistic regression

# Comparing estimates from weighting and regression

| Question | weighted averages 1999 | weighted averages 2001 | (a) time change in percent | (b) linear regression coefficient of time |
|---|---|---|---|---|
| Adult in good/excellent health | 75% | 78% | 3.4% (2.4%) | 6.6% (1.4%) |
| Child in good/excellent health | 82% | 84% | 1.7% (1.5%) | 1.2% (1.3%) |
| Neighborhood is safe/very safe | 77% | 81% | 4.5% (2.3%) | 4.1% (1.5%) |

- The estimates can be very different!
- Which to believe?
- Same pattern with logistic regression

# Summary so far

- Hierarchical modeling + poststratification works well for estimating state-level opinions from national polls
- We're not sure what to do with the Social Indicators Survey
  - Tangle of regression coefficients
  - No simple structure (as in the hierarchical model for 50 states)
- Larger goal:

# Summary so far

- Hierarchical modeling + poststratification works well for estimating state-level opinions from national polls
- We're not sure what to do with the Social Indicators Survey
  - Tangle of regression coefficients
  - No simple structure (as in the hierarchical model for 50 states)
- Larger goal:

# Summary so far

- Hierarchical modeling + poststratification works well for estimating state-level opinions from national polls
- We're not sure what to do with the Social Indicators Survey
  - Tangle of regression coefficients
  - No simple structure (as in the hierarchical model for 50 states)
- Larger goal:
  - Believable estimates using regression
  - "Backward compatibility" to simple weighted averages

# Summary so far

- Hierarchical modeling + poststratification works well for estimating state-level opinions from national polls
- We're not sure what to do with the Social Indicators Survey
  - Tangle of regression coefficients
  - No simple structure (as in the hierarchical model for 50 states)
- Larger goal:
  - Believable estimates using regression
  - "Backward compatibility" to simple weighted averages

# Summary so far

- Hierarchical modeling + poststratification works well for estimating state-level opinions from national polls
- We're not sure what to do with the Social Indicators Survey
  - Tangle of regression coefficients
  - No simple structure (as in the hierarchical model for 50 states)
- Larger goal:
  - Believable estimates using regression
  - "Backward compatibility" to simple weighted averages

# Summary so far

- Hierarchical modeling + poststratification works well for estimating state-level opinions from national polls
- We're not sure what to do with the Social Indicators Survey
  - Tangle of regression coefficients
  - No simple structure (as in the hierarchical model for 50 states)
- Larger goal:
  - Believable estimates using regression
  - "Backward compatibility" to simple weighted averages

# Summary so far

- Hierarchical modeling + poststratification works well for estimating state-level opinions from national polls
- We're not sure what to do with the Social Indicators Survey
  - Tangle of regression coefficients
  - No simple structure (as in the hierarchical model for 50 states)
- Larger goal:
  - Believable estimates using regression
  - "Backward compatibility" to simple weighted averages

# Regression models and implied weights

- Fit a regression and poststratify:
  - $\hat{\theta} = \sum_{j=1}^{J} N_j \hat{\theta}_j \Big/ \sum_{j=1}^{J} N_j$
  - From regression, $\hat{\theta}_j$'s are linear combinations of the data $y$
  - We can write $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i$
  - $w_i$'s are *implied weights*
- Classical regression
- Hierarchical regression

# Regression models and implied weights

- Fit a regression and poststratify:
  - $\hat{\theta} = \sum_{j=1}^{J} N_j \hat{\theta}_j \Big/ \sum_{j=1}^{J} N_j$
  - From regression, $\hat{\theta}_j$'s are linear combinations of the data $y$
  - We can write $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i$
  - $w_i$'s are *implied weights*
- Classical regression
- Hierarchical regression

# Regression models and implied weights

- Fit a regression and poststratify:
  - $\hat{\theta} = \sum_{j=1}^{J} N_j \hat{\theta}_j \Big/ \sum_{j=1}^{J} N_j$
  - From regression, $\hat{\theta}_j$'s are linear combinations of the data $y$
  - We can write $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i$
  - $w_i$'s are *implied weights*
- Classical regression
- Hierarchical regression

# Regression models and implied weights

- Fit a regression and poststratify:
  - $\hat{\theta} = \sum_{j=1}^{J} N_j \hat{\theta}_j \Big/ \sum_{j=1}^{J} N_j$
  - From regression, $\hat{\theta}_j$'s are linear combinations of the data $y$
  - We can write $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i$
  - $w_i$'s are *implied weights*
- Classical regression
- Hierarchical regression

# Regression models and implied weights

- Fit a regression and poststratify:
  - $\hat{\theta} = \sum_{j=1}^{J} N_j \hat{\theta}_j \Big/ \sum_{j=1}^{J} N_j$
  - From regression, $\hat{\theta}_j$'s are linear combinations of the data $y$
  - We can write $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i$
  - $w_i$'s are *implied weights*
- Classical regression
- Hierarchical regression

# Regression models and implied weights

- Fit a regression and poststratify:
  - $\hat{\theta} = \sum_{j=1}^{J} N_j \hat{\theta}_j \Big/ \sum_{j=1}^{J} N_j$
  - From regression, $\hat{\theta}_j$'s are linear combinations of the data $y$
  - We can write $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i$
  - $w_i$'s are *implied weights*
- Classical regression
- Hierarchical regression

# Regression models and implied weights

- Fit a regression and poststratify:
  - $\hat{\theta} = \sum_{j=1}^{J} N_j \hat{\theta}_j \Big/ \sum_{j=1}^{J} N_j$
  - From regression, $\hat{\theta}_j$'s are linear combinations of the data $y$
  - We can write $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i$
  - $w_i$'s are *implied weights*
- Classical regression
- Hierarchical regression

# Weights corresponding to trivial classical regressions

- Full poststratification, $\hat{\theta} = \sum_{j=1}^{J} N_j \bar{y}_j \Big/ \sum_{j=1}^{J} N_j$
  - Classical regression on indicators for all $J$ cells
  - Equivalent weights: $w_i \propto N_j / n_j$
- No weighting, $\hat{\theta} = \bar{y}$

# Weights corresponding to trivial classical regressions

- Full poststratification, $\hat{\theta} = \sum_{j=1}^{J} N_j \bar{y}_j \Big/ \sum_{j=1}^{J} N_j$
  - Classical regression on indicators for all $J$ cells
  - Equivalent weights: $w_i \propto N_j / n_j$
- No weighting, $\hat{\theta} = \bar{y}$
  - Classical regression with just a constant term
  - Equivalent weights: $w_i = 1$

# Weights corresponding to trivial classical regressions

- Full poststratification, $\hat{\theta} = \sum_{j=1}^{J} N_j \bar{y}_j \Big/ \sum_{j=1}^{J} N_j$
  - Classical regression on indicators for all $J$ cells
  - Equivalent weights: $w_i \propto N_j / n_j$
- No weighting, $\hat{\theta} = \bar{y}$
  - Classical regression with just a constant term
  - Equivalent weights: $w_i = 1$

# Weights corresponding to trivial classical regressions

- Full poststratification, $\hat{\theta} = \sum_{j=1}^{J} N_j \bar{y}_j \Big/ \sum_{j=1}^{J} N_j$
  - Classical regression on indicators for all $J$ cells
  - Equivalent weights: $w_i \propto N_j / n_j$
- No weighting, $\hat{\theta} = \bar{y}$
  - Classical regression with just a constant term
  - Equivalent weights: $w_i = 1$

# Weights corresponding to trivial classical regressions

- Full poststratification, $\hat{\theta} = \sum_{j=1}^{J} N_j \bar{y}_j \Big/ \sum_{j=1}^{J} N_j$
  - Classical regression on indicators for all $J$ cells
  - Equivalent weights: $w_i \propto N_j / n_j$
- No weighting, $\hat{\theta} = \bar{y}$
  - Classical regression with just a constant term
  - Equivalent weights: $w_i = 1$

# Weights corresponding to trivial classical regressions

- Full poststratification, $\hat{\theta} = \sum_{j=1}^{J} N_j \bar{y}_j \Big/ \sum_{j=1}^{J} N_j$
  - Classical regression on indicators for all $J$ cells
  - Equivalent weights: $w_i \propto N_j / n_j$
- No weighting, $\hat{\theta} = \bar{y}$
  - Classical regression with just a constant term
  - Equivalent weights: $w_i = 1$

# Weights corresponding to classical regressions

- Regression $y = X\beta + \epsilon$ followed by poststratification
  - $\hat{\beta}$ is a linear combination of data $y$
  - Vector of equivalent weights: $\frac{n}{N}(N^{\text{pop}})^t X^{\text{pop}}(X^t X)^{-1} X^t$
  - These depend on population $N$'s and sample $X$'s but *not* on sample $y$'s
- Equivalent weights sum to $n$

# Weights corresponding to classical regressions

- Regression $y = X\beta + \epsilon$ followed by poststratification
  - $\hat{\beta}$ is a linear combination of data $y$
  - Vector of equivalent weights: $\frac{n}{N}(N^{\rm pop})^t X^{\rm pop}(X^t X)^{-1}X^t$
  - These depend on population $N$'s and sample $X$'s but *not* on sample $y$'s
- Equivalent weights sum to $n$

# Weights corresponding to classical regressions

- Regression $y = X\beta + \epsilon$ followed by poststratification
  - $\hat{\beta}$ is a linear combination of data $y$
  - Vector of equivalent weights: $\frac{n}{N}(N^{\rm pop})^t X^{\rm pop}(X^t X)^{-1} X^t$
  - These depend on population $N$'s and sample $X$'s but *not* on sample $y$'s
- Equivalent weights sum to $n$
  - Proof uses translation-invariance of linear regression
  - $\hat{\theta}$ is thus a *weighted average*, not just a *linear combination*

# Weights corresponding to classical regressions

- Regression $y = X\beta + \epsilon$ followed by poststratification
  - $\hat{\beta}$ is a linear combination of data $y$
  - Vector of equivalent weights: $\frac{n}{N}(N^{\mathrm{pop}})^t X^{\mathrm{pop}} (X^t X)^{-1} X^t$
  - These depend on population $N$'s and sample $X$'s but *not* on sample $y$'s
- Equivalent weights sum to $n$
  - Proof uses translation-invariance of linear regression
  - $\hat{\theta}$ is thus a *weighted average*, not just a *linear combination*

# Weights corresponding to classical regressions

- Regression $y = X\beta + \epsilon$ followed by poststratification
  - $\hat{\beta}$ is a linear combination of data $y$
  - Vector of equivalent weights: $\frac{n}{N}(N^{\text{pop}})^t X^{\text{pop}}(X^tX)^{-1}X^t$
  - These depend on population $N$'s and sample $X$'s but *not* on sample $y$'s
- Equivalent weights sum to $n$
  - Proof uses translation-invariance of linear regression
  - $\hat{\theta}$ is thus a *weighted average*, not just a *linear combination*

# Weights corresponding to classical regressions

- Regression $y = X\beta + \epsilon$ followed by poststratification
  - $\hat{\beta}$ is a linear combination of data $y$
  - Vector of equivalent weights: $\frac{n}{N}(N^{\rm pop})^t X^{\rm pop}(X^t X)^{-1} X^t$
  - These depend on population $N$'s and sample $X$'s but *not* on sample $y$'s
- Equivalent weights sum to $n$
  - Proof uses translation-invariance of linear regression
  - $\hat{\theta}$ is thus a *weighted average*, not just a *linear combination*

# Weights corresponding to classical regressions

- Regression $y = X\beta + \epsilon$ followed by poststratification
  - $\hat{\beta}$ is a linear combination of data $y$
  - Vector of equivalent weights: $\frac{n}{N}(N^{\rm pop})^t X^{\rm pop}(X^t X)^{-1} X^t$
  - These depend on population $N$'s and sample $X$'s but *not* on sample $y$'s
- Equivalent weights sum to $n$
  - Proof uses translation-invariance of linear regression
  - $\hat{\theta}$ is thus a *weighted average*, not just a *linear combination*

# Classical regression for CBS polls

- Illustration with a sequence of regressions:
  - male/female
  - also black/white
  - also male/female × black/white
  - also 4 age categories
  - also 4 education categories
  - also age × education

# Classical regression for CBS polls

- Illustration with a sequence of regressions:
  - male/female
  - also black/white
  - also male/female × black/white
  - also 4 age categories
  - also 4 education categories
  - also age × education

# Classical regression for CBS polls

- Illustration with a sequence of regressions:
  - male/female
  - also black/white
  - also male/female × black/white
  - also 4 age categories
  - also 4 education categories
  - also age × education

# Classical regression for CBS polls

- Illustration with a sequence of regressions:
  - male/female
  - also black/white
  - also male/female $\times$ black/white
  - also 4 age categories
  - also 4 education categories
  - also age $\times$ education

# Classical regression for CBS polls

- Illustration with a sequence of regressions:
  - male/female
  - also black/white
  - also male/female × black/white
  - also 4 age categories
  - also 4 education categories
  - also age × education

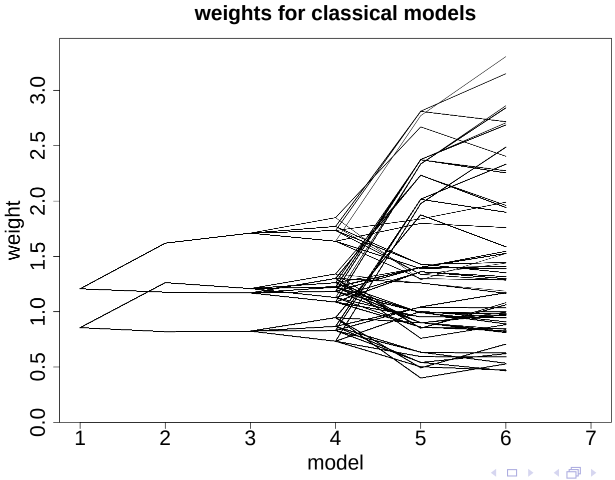
# Classical regression for CBS polls

- Illustration with a sequence of regressions:
  - male/female
  - also black/white
  - also male/female $\times$ black/white
  - also 4 age categories
  - also 4 education categories
  - also age $\times$ education

# Classical regression for CBS polls

- Illustration with a sequence of regressions:
  - male/female
  - also black/white
  - also male/female × black/white
  - also 4 age categories
  - also 4 education categories
  - also age × education

# Classical weights for CBS polls



# Weights corresponding to hierarchical regressions

- Same algebra as in classical regression
- Augment with "prior distribution"
- Vector of equivalent weights now depends on the hierarchical variance parameters (and thus indirectly on the data)
- Different vector of weights for different choices of $y$
- With noninformative prior distribution, the equivalent weights still sum to $n$
- Illustration with CBS polls
- Shrinkage of weights

# Weights corresponding to hierarchical regressions

- Same algebra as in classical regression
- Augment with "prior distribution"
- Vector of equivalent weights now depends on the hierarchical variance parameters (and thus indirectly on the data)
- Different vector of weights for different choices of $y$
- With noninformative prior distribution, the equivalent weights still sum to $n$
- Illustration with CBS polls
- Shrinkage of weights

# Weights corresponding to hierarchical regressions

- Same algebra as in classical regression
- Augment with "prior distribution"
- Vector of equivalent weights now depends on the hierarchical variance parameters (and thus indirectly on the data)
- Different vector of weights for different choices of $y$
- With noninformative prior distribution, the equivalent weights still sum to $n$
- Illustration with CBS polls
- Shrinkage of weights

# Weights corresponding to hierarchical regressions

- Same algebra as in classical regression
- Augment with "prior distribution"
- Vector of equivalent weights now depends on the hierarchical variance parameters (and thus indirectly on the data)
- Different vector of weights for different choices of $y$
- With noninformative prior distribution, the equivalent weights still sum to $n$
- Illustration with CBS polls
- Shrinkage of weights

# Weights corresponding to hierarchical regressions

- Same algebra as in classical regression
- Augment with "prior distribution"
- Vector of equivalent weights now depends on the hierarchical variance parameters (and thus indirectly on the data)
- Different vector of weights for different choices of $y$
- With noninformative prior distribution, the equivalent weights still sum to $n$
- Illustration with CBS polls
- Shrinkage of weights

# Weights corresponding to hierarchical regressions

- Same algebra as in classical regression
- Augment with "prior distribution"
- Vector of equivalent weights now depends on the hierarchical variance parameters (and thus indirectly on the data)
- Different vector of weights for different choices of $y$
- With noninformative prior distribution, the equivalent weights still sum to $n$
- Illustration with CBS polls
- Shrinkage of weights

# Weights corresponding to hierarchical regressions

- Same algebra as in classical regression
- Augment with "prior distribution"
- Vector of equivalent weights now depends on the hierarchical variance parameters (and thus indirectly on the data)
- Different vector of weights for different choices of $y$
- With noninformative prior distribution, the equivalent weights still sum to $n$
- Illustration with CBS polls
- Shrinkage of weights

# Hierarchical regression for CBS polls

- Illustration with a sequence of regressions:
  - male/female
  - also black/white
  - also male/female × black/white
  - also 4 age categories (hierarchical)
  - also 4 education categories (hierarchical)
  - also age × education (hierarchical)
  - also 50 states (hierarchical)

# Hierarchical regression for CBS polls

- Illustration with a sequence of regressions:
  - male/female
  - also black/white
  - also male/female × black/white
  - also 4 age categories (hierarchical)
  - also 4 education categories (hierarchical)
  - also age × education (hierarchical)
  - also 50 states (hierarchical)

# Hierarchical regression for CBS polls

- Illustration with a sequence of regressions:
  - male/female
  - also black/white
  - also male/female × black/white
  - also 4 age categories (hierarchical)
  - also 4 education categories (hierarchical)
  - also age × education (hierarchical)
  - also 50 states (hierarchical)

# Hierarchical regression for CBS polls

- Illustration with a sequence of regressions:
  - male/female
  - also black/white
  - also male/female × black/white
  - also 4 age categories (hierarchical)
  - also 4 education categories (hierarchical)
  - also age × education (hierarchical)
  - also 50 states (hierarchical)

# Hierarchical regression for CBS polls

- Illustration with a sequence of regressions:
  - male/female
  - also black/white
  - also male/female $\times$ black/white
  - also 4 age categories (hierarchical)
  - also 4 education categories (hierarchical)
  - also age $\times$ education (hierarchical)
  - also 50 states (hierarchical)

# Hierarchical regression for CBS polls

- Illustration with a sequence of regressions:
  - male/female
  - also black/white
  - also male/female $\times$ black/white
  - also 4 age categories (hierarchical)
  - also 4 education categories (hierarchical)
  - also age $\times$ education (hierarchical)
  - also 50 states (hierarchical)

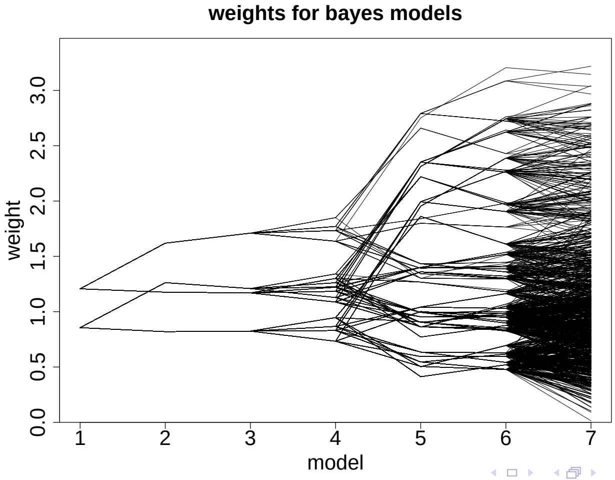
# Hierarchical regression for CBS polls

- Illustration with a sequence of regressions:
  - male/female
  - also black/white
  - also male/female $\times$ black/white
  - also 4 age categories (hierarchical)
  - also 4 education categories (hierarchical)
  - also age $\times$ education (hierarchical)
  - also 50 states (hierarchical)

# Hierarchical regression for CBS polls

- Illustration with a sequence of regressions:
  - male/female
  - also black/white
  - also male/female × black/white
  - also 4 age categories (hierarchical)
  - also 4 education categories (hierarchical)
  - also age × education (hierarchical)
  - also 50 states (hierarchical)

# Hierarchical weights for CBS polls



# Hierarchical models and smoothing of weights

- Exchangeable normal model on $J$ categories
  - Raw weights $w_i \propto N_j/n_j$ in cell $j$
  - Pooled weights $w_i = 1$
  - Equivalent weights are *approximately* partially pooled by the "shrinkage factor" $\tau^2 \Big/ \left(\frac{\sigma^2}{n_j} + \tau^2\right)$
- Hierarchical regression models:
  Shrinkage toward marginal "raking" weights
- Important for "backward compatibility"

# Hierarchical models and smoothing of weights

- Exchangeable normal model on $J$ categories
  - Raw weights $w_i \propto N_j/n_j$ in cell $j$
  - Pooled weights $w_i = 1$
  - Equivalent weights are *approximately* partially pooled by the "shrinkage factor" $\tau^2 \Big/ \left(\frac{\sigma^2}{n_j} + \tau^2\right)$
- Hierarchical regression models: Shrinkage toward marginal "raking" weights
- Important for "backward compatibility"

# Hierarchical models and smoothing of weights

- Exchangeable normal model on $J$ categories
  - Raw weights $w_i \propto N_j/n_j$ in cell $j$
  - Pooled weights $w_i = 1$
  - Equivalent weights are *approximately* partially pooled by the "shrinkage factor" $\tau^2 \Big/ \left(\frac{\sigma^2}{n_j} + \tau^2\right)$
- Hierarchical regression models:
  Shrinkage toward marginal "raking" weights
- Important for "backward compatibility"

# Hierarchical models and smoothing of weights

- Exchangeable normal model on $J$ categories
  - Raw weights $w_i \propto N_j/n_j$ in cell $j$
  - Pooled weights $w_i = 1$
  - Equivalent weights are *approximately* partially pooled by the "shrinkage factor" $\tau^2 \Big/ \left(\frac{\sigma^2}{n_j} + \tau^2\right)$
- Hierarchical regression models: Shrinkage toward marginal "raking" weights
- Important for "backward compatibility"

# Hierarchical models and smoothing of weights

- Exchangeable normal model on $J$ categories
  - Raw weights $w_i \propto N_j/n_j$ in cell $j$
  - Pooled weights $w_i = 1$
  - Equivalent weights are *approximately* partially pooled by the "shrinkage factor" $\tau^2 \Big/ \left(\frac{\sigma^2}{n_j} + \tau^2\right)$
- Hierarchical regression models: Shrinkage toward marginal "raking" weights
- Important for "backward compatibility"

# Hierarchical models and smoothing of weights

- Exchangeable normal model on $J$ categories
  - Raw weights $w_i \propto N_j/n_j$ in cell $j$
  - Pooled weights $w_i = 1$
  - Equivalent weights are *approximately* partially pooled by the "shrinkage factor" $\tau^2 \Big/ \left(\frac{\sigma^2}{n_j} + \tau^2\right)$
- Hierarchical regression models:
  Shrinkage toward marginal "raking" weights
- Important for "backward compatibility"

# Where do we stand?

- Practical limitations of weighting
- Practical limitations of modeling
- Putting it all together using hierarchical models and poststratification

# Where do we stand?

- Practical limitations of weighting
- Practical limitations of modeling
- Putting it all together using hierarchical models and poststratification

# Where do we stand?

- Practical limitations of weighting
- Practical limitations of modeling
- Putting it all together using hierarchical models and poststratification

# Practical limitations of weighting

Simple estimates for population averages and ratios, but . . .

- Not clear how to apply to regression coefs, other complicated estimands
- Standard errors are tricky
- A "quick and dirty" method? Not necessarily so quick!

# Practical limitations of weighting

Simple estimates for population averages and ratios, but ...

- Not clear how to apply to regression coefs, other complicated estimands
- Standard errors are tricky
- A "quick and dirty" method? Not necessarily so quick!
  - Arbitrary choices about which variables and interactions to include
  - Pooling of weighting cells and truncation of weights
  - X's, y's, and "canary variables"

# Practical limitations of weighting

Simple estimates for population averages and ratios, but . . .

- Not clear how to apply to regression coefs, other complicated estimands
- Standard errors are tricky
- A "quick and dirty" method? Not necessarily so quick!
  - Arbitrary choices about which variables and interactions to include
  - Pooling of weighting cells and truncation of weights
  - $X$'s, $y$'s, and "canary variables"

# Practical limitations of weighting

Simple estimates for population averages and ratios, but ...

- Not clear how to apply to regression coefs, other complicated estimands
- Standard errors are tricky
- A "quick and dirty" method? Not necessarily so quick!
  - Arbitrary choices about which variables and interactions to include
  - Pooling of weighting cells and truncation of weights
  - $X$'s, $y$'s, and "canary variables"

# Practical limitations of weighting

Simple estimates for population averages and ratios, but ...

- Not clear how to apply to regression coefs, other complicated estimands
- Standard errors are tricky
- A "quick and dirty" method? Not necessarily so quick!
  - Arbitrary choices about which variables and interactions to include
  - Pooling of weighting cells and truncation of weights
  - $X$'s, $y$'s, and "canary variables"

# Practical limitations of weighting

Simple estimates for population averages and ratios, but . . .

- Not clear how to apply to regression coefs, other complicated estimands
- Standard errors are tricky
- A "quick and dirty" method? Not necessarily so quick!
  - Arbitrary choices about which variables and interactions to include
  - Pooling of weighting cells and truncation of weights
  - $X$'s, $y$'s, and "canary variables"

# Practical limitations of weighting

Simple estimates for population averages and ratios, but . . .

- Not clear how to apply to regression coefs, other complicated estimands
- Standard errors are tricky
- A "quick and dirty" method? Not necessarily so quick!
  - Arbitrary choices about which variables and interactions to include
  - Pooling of weighting cells and truncation of weights
  - $X$'s, $y$'s, and "canary variables"

# Practical limitations of modeling

Easy to do (even hierarchical models), but . . .

- Theoretically must condition on all poststratification cells
- Models with potentially thousands of coefficients
- Lack of trust in results
- But sometimes we do trust highly-parameterized models
- ??

# Practical limitations of modeling

Easy to do (even hierarchical models), but . . .

- Theoretically must condition on all poststratification cells
- Models with potentially thousands of coefficients
- Lack of trust in results
- But sometimes we do trust highly-parameterized models
- ??

# Practical limitations of modeling

Easy to do (even hierarchical models), but . . .

- Theoretically must condition on all poststratification cells
- Models with potentially thousands of coefficients
- Lack of trust in results
- But sometimes we do trust highly-parameterized models
  - State-level estimates from national polls
  - Small-area estimation + poststratification
- ??

# Practical limitations of modeling

Easy to do (even hierarchical models), but ...

- Theoretically must condition on all poststratification cells
- Models with potentially thousands of coefficients
- Lack of trust in results
- But sometimes we do trust highly-parameterized models
  - State-level estimates from national polls
  - Small-area estimation + poststratification
- ??

# Practical limitations of modeling

Easy to do (even hierarchical models), but ...

- Theoretically must condition on all poststratification cells
- Models with potentially thousands of coefficients
- Lack of trust in results
- But sometimes we do trust highly-parameterized models
  - State-level estimates from national polls
  - Small-area estimation + poststratification
- ??

# Practical limitations of modeling

Easy to do (even hierarchical models), but ...

- Theoretically must condition on all poststratification cells
- Models with potentially thousands of coefficients
- Lack of trust in results
- But sometimes we do trust highly-parameterized models
  - State-level estimates from national polls
  - Small-area estimation + poststratification
- ??

# Practical limitations of modeling

Easy to do (even hierarchical models), but ...

- Theoretically must condition on all poststratification cells
- Models with potentially thousands of coefficients
- Lack of trust in results
- But sometimes we do trust highly-parameterized models
  - State-level estimates from national polls
  - Small-area estimation + poststratification
- ??

# Practical limitations of modeling

Easy to do (even hierarchical models), but ...

- Theoretically must condition on all poststratification cells
- Models with potentially thousands of coefficients
- Lack of trust in results
- But sometimes we do trust highly-parameterized models
  - State-level estimates from national polls
  - Small-area estimation + poststratification
- ??

# Putting it all together

- Our ideal procedure:
  - As easy to use as hierarchical regression
  - Population info included using poststratification
- Smooth transition from classical weighting

# Putting it all together

- Our ideal procedure:
  - As easy to use as hierarchical regression
  - Population info included using poststratification
- Smooth transition from classical weighting
  - Equivalent weights
  - When different methods give different results, we can trace it back to an interaction

# Putting it all together

- Our ideal procedure:
  - As easy to use as hierarchical regression
  - Population info included using poststratification
- Smooth transition from classical weighting
  - Equivalent weights
  - When different methods give different results, we can track it back to an interaction

# Putting it all together

- Our ideal procedure:
  - As easy to use as hierarchical regression
  - Population info included using poststratification
- Smooth transition from classical weighting
  - Equivalent weights
  - When different methods give different results, we can track it back to an interaction

# Putting it all together

- Our ideal procedure:
  - As easy to use as hierarchical regression
  - Population info included using poststratification
- Smooth transition from classical weighting
  - Equivalent weights
  - When different methods give different results, we can track it back to an interaction

# Putting it all together

- Our ideal procedure:
  - As easy to use as hierarchical regression
  - Population info included using poststratification
- Smooth transition from classical weighting
  - Equivalent weights
  - When different methods give different results, we can track it back to an interaction

# Our research plan

- Figuring out where the 2 estimates diverge for the Social Indicators Survey
  - Goal: believable estimates for time trends
  - Goal: a good set of weights for simple estimands
- Related problems in statistical modeling
- No "conclusions"

# Our research plan

- Figuring out where the 2 estimates diverge for the Social Indicators Survey
  - Goal: believable estimates for time trends
  - Goal: a good set of weights for simple estimands
- Related problems in statistical modeling
  - Hierarchical regressions with complex interactions
    Iterative proportional fitting, etc., using population margins
- No "conclusions"

# Our research plan

- Figuring out where the 2 estimates diverge for the Social Indicators Survey
  - Goal: believable estimates for time trends
  - Goal: a good set of weights for simple estimands
- Related problems in statistical modeling
  - Hierarchical regressions with complex interactions
  - Iterative proportional fitting, etc., using population margins
- No "conclusions"

# Our research plan

- Figuring out where the 2 estimates diverge for the Social Indicators Survey
  - Goal: believable estimates for time trends
  - Goal: a good set of weights for simple estimands
- Related problems in statistical modeling
  - Hierarchical regressions with complex interactions
  - Iterative proportional fitting, etc., using population margins
- No "conclusions"

# Our research plan

- Figuring out where the 2 estimates diverge for the Social Indicators Survey
  - Goal: believable estimates for time trends
  - Goal: a good set of weights for simple estimands
- Related problems in statistical modeling
  - Hierarchical regressions with complex interactions
  - Iterative proportional fitting, etc., using population margins
- No "conclusions"

# Our research plan

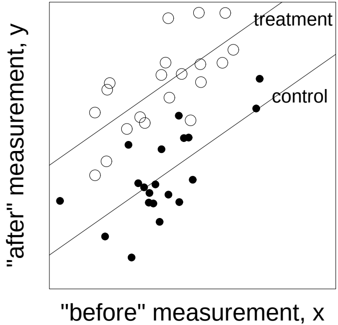
- Figuring out where the 2 estimates diverge for the Social Indicators Survey
  - Goal: believable estimates for time trends
  - Goal: a good set of weights for simple estimands
- Related problems in statistical modeling
  - Hierarchical regressions with complex interactions
  - Iterative proportional fitting, etc., using population margins
- No "conclusions"

# Our research plan

- Figuring out where the 2 estimates diverge for the Social Indicators Survey
  - Goal: believable estimates for time trends
  - Goal: a good set of weights for simple estimands
- Related problems in statistical modeling
  - Hierarchical regressions with complex interactions
  - Iterative proportional fitting, etc., using population margins
- No "conclusions"

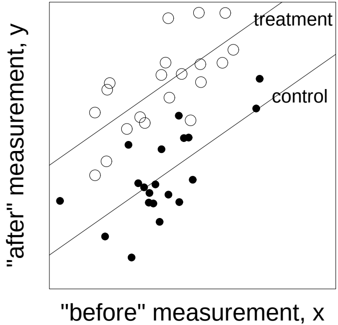
# No-interaction model

- Default analysis for experiments and observational studies: constant treatment effects
  - Fisher's classical null hyp: effect is zero for all cases
  - Regression model: $y_i = T_i\theta + X_i\beta + \epsilon_i$



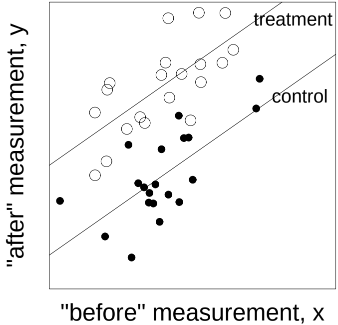
# No-interaction model

- Default analysis for experiments and observational studies: constant treatment effects
  - Fisher's classical null hyp: effect is zero for all cases
  - Regression model: $y_i = T_i\theta + X_i\beta + \epsilon_i$



# No-interaction model

- Default analysis for experiments and observational studies: constant treatment effects
  - Fisher's classical null hyp: effect is zero for all cases
  - Regression model: $y_i = T_i\theta + X_i\beta + \epsilon_i$



# Actual data

- Treatment interacts with "before" measurement
- Before-after correlation is higher for *controls* than for *treated* units
- Examples
  - An observational study of legislative redistricting
  - An experiment with pre-test, post-test data
  - Congressional elections with incumbents and open seats

# Actual data

- Treatment interacts with "before" measurement
- Before-after correlation is higher for *controls* than for *treated* units
- Examples
  - An observational study of legislative redistricting
  - An experiment with pre-test, post-test data
  - Congressional elections with incumbents and open seats

# Actual data

- Treatment interacts with "before" measurement
- Before-after correlation is higher for *controls* than for *treated* units
- Examples
  - An observational study of legislative redistricting
  - An experiment with pre-test, post-test data
  - Congressional elections with incumbents and open seats

# Actual data

- Treatment interacts with "before" measurement
- Before-after correlation is higher for *controls* than for *treated* units
- Examples
  - An observational study of legislative redistricting
  - An experiment with pre-test, post-test data
  - Congressional elections with incumbents and open seats

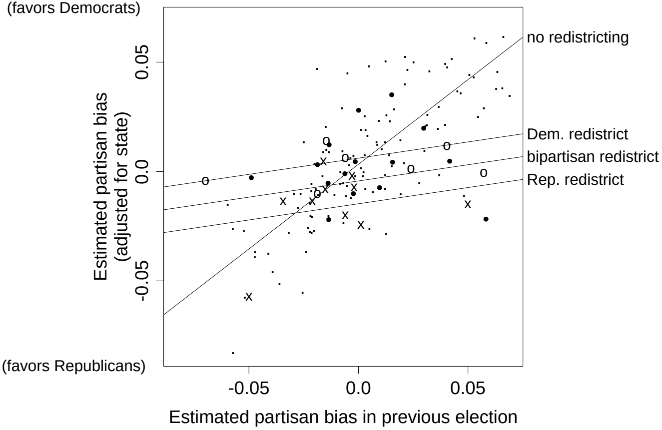
# Actual data

- Treatment interacts with "before" measurement
- Before-after correlation is higher for *controls* than for *treated* units
- Examples
  - An observational study of legislative redistricting
  - An experiment with pre-test, post-test data
  - Congressional elections with incumbents and open seats

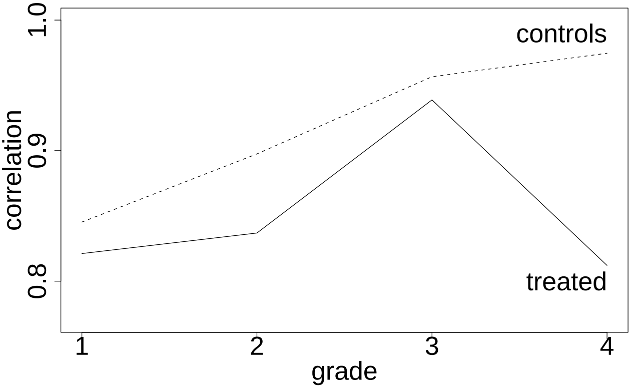
# Actual data

- Treatment interacts with "before" measurement
- Before-after correlation is higher for *controls* than for *treated* units
- Examples
  - An observational study of legislative redistricting
  - An experiment with pre-test, post-test data
  - Congressional elections with incumbents and open seats

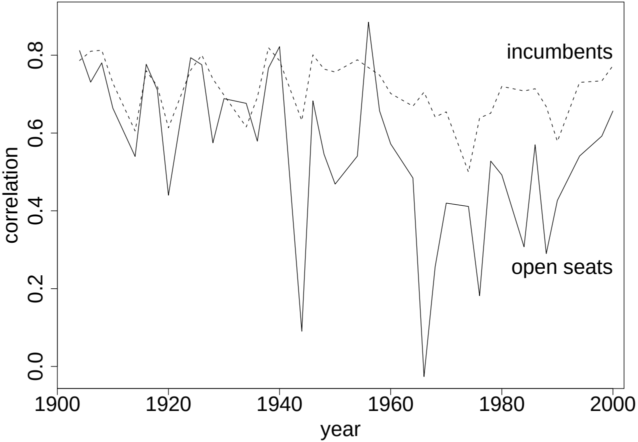
# Observational study of legislative redistricting before-after data



# Experiment: correlation between pre-test and post-test data for controls and for treated units



# Correlation between two successive Congressional elections for incumbents running (controls) and open seats (treated)



# Underlying variance components

Unit-level "error term" $\eta_i$

- For control units, $\eta_i$ persists from time 1 to time 2
- For treatment units, $\eta_i$ changes:
  - Subtractive treatment error ($\eta_i$ only at time 1)
  - Additive treatment error ($\eta_i$ only at time 2)
  - Replacement treatment error
- Under all these models, the before-after correlation is higher for controls than treated units

# Underlying variance components

Unit-level "error term" $\eta_i$

- For control units, $\eta_i$ persists from time 1 to time 2
- For treatment units, $\eta_i$ changes:
  - Subtractive treatment error ($\eta_i$ only at time 1)
  - Additive treatment error ($\eta_i$ only at time 2)
  - Replacement treatment error
- Under all these models, the before-after correlation is higher for controls than treated units

# Underlying variance components

Unit-level "error term" $\eta_i$

- For control units, $\eta_i$ persists from time 1 to time 2
- For treatment units, $\eta_i$ changes:
  - Subtractive treatment error ($\eta_i$ only at time 1)
  - Additive treatment error ($\eta_i$ only at time 2)
  - Replacement treatment error
- Under all these models, the before-after correlation is higher for controls than treated units

# Underlying variance components

Unit-level "error term" $\eta_i$

- For control units, $\eta_i$ persists from time 1 to time 2
- For treatment units, $\eta_i$ changes:
  - Subtractive treatment error ($\eta_i$ only at time 1)
  - Additive treatment error ($\eta_i$ only at time 2)
  - Replacement treatment error
- Under all these models, the before-after correlation is higher for controls than treated units

# Underlying variance components

Unit-level "error term" $\eta_i$

- For control units, $\eta_i$ persists from time 1 to time 2
- For treatment units, $\eta_i$ changes:
  - Subtractive treatment error ($\eta_i$ only at time 1)
  - Additive treatment error ($\eta_i$ only at time 2)
  - Replacement treatment error
- Under all these models, the before-after correlation is higher for controls than treated units

# Underlying variance components

Unit-level "error term" $\eta_i$

- For control units, $\eta_i$ persists from time 1 to time 2
- For treatment units, $\eta_i$ changes:
  - Subtractive treatment error ($\eta_i$ only at time 1)
  - Additive treatment error ($\eta_i$ only at time 2)
  - Replacement treatment error
- Under all these models, the before-after correlation is higher for controls than treated units

# Underlying variance components

Unit-level "error term" $\eta_i$

- For control units, $\eta_i$ persists from time 1 to time 2
- For treatment units, $\eta_i$ changes:
  - Subtractive treatment error ($\eta_i$ only at time 1)
  - Additive treatment error ($\eta_i$ only at time 2)
  - Replacement treatment error
- Under all these models, the before-after correlation is higher for controls than treated units

# Papers

- State-level estimates from national polls
  - 1997 Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*. (Andrew Gelman and Thomas C. Little)
  - 2004 Bayesian multilevel estimation with poststratification: state-level estimates from national polls. *Political Analysis*. (David K. Park, Andrew Gelman, and Joseph Bafumi)
- Struggles with survey weighting and poststratification
  - 2001 Poststratification and weighting adjustments. In *Survey Nonresponse*. (Andrew Gelman and John B. Carlin)
  - 2004 Struggles with survey weighting and regression modeling. Unpublished. (Andrew Gelman)

# Papers

- State-level estimates from national polls

  1997 Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*. (Andrew Gelman and Thomas C. Little)

  2004 Bayesian multilevel estimation with poststratification: state-level estimates from national polls. *Political Analysis*. (David K. Park, Andrew Gelman, and Joseph Bafumi)

- Struggles with survey weighting and poststratification

  2001 Poststratification and weighting adjustments. In *Survey Nonresponse*. (Andrew Gelman and John B. Carlin)

  2004 Struggles with survey weighting and regression modeling. Unpublished. (Andrew Gelman)

# Papers

- State-level estimates from national polls
  - 1997 Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*. (Andrew Gelman and Thomas C. Little)
  - 2004 Bayesian multilevel estimation with poststratification: state-level estimates from national polls. *Political Analysis*. (David K. Park, Andrew Gelman, and Joseph Bafumi)
- Struggles with survey weighting and poststratification
  - 2001 Poststratification and weighting adjustments. In *Survey Nonresponse*. (Andrew Gelman and John B. Carlin)
  - 2004 Struggles with survey weighting and regression modeling. Unpublished. (Andrew Gelman)

# Papers

- Interactions in before-after data

1994 Enhancing democracy through legislative redistricting. *American Political Science Review*. (Andrew Gelman and Gary King)

2004 Treatment effects in before-after data. In *Applied Bayesian Modeling and Causal Inference*. (Andrew Gelman)

2005 Estimating incumbency advantage and its variation, as an example of a before/after study. *Journal of the American Statistical Association*. (Andrew Gelman and Zaiying Huang)

# Papers

- Interactions in before-after data

1994 Enhancing democracy through legislative redistricting. *American Political Science Review*. (Andrew Gelman and Gary King)

2004 Treatment effects in before-after data. In *Applied Bayesian Modeling and Causal Inference*. (Andrew Gelman)

2005 Estimating incumbency advantage and its variation, as an example of a before/after study. *Journal of the American Statistical Association*. (Andrew Gelman and Zaiying Huang)