Design and Analysis of Sample Surveys

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

Class 6b: Inference for regression coefficients

Inference for regression coefficients

- Option 1: weighted regression
- Option 2: unweighted regression, including in the model all variables that affect the probability of inclusion in the survey
- Discuss
- Population mean as a special case of a regression coefficient
- Population difference as a special case of a regression coefficient

Some operations using the “survey” package and directly in R

- Example: estimating McCain vote share in 2008 election using Pew survey
- Unweighted average (and standard error)
- Weighted average (and standard error)
- Unweighted regression
- Weighted regression
- Ratio estimate

Stratification and poststratification

- Pretend the survey is stratified by age
- Poststratify by sex
- Poststratify by sex, ethnicity, and age

Cluster sampling

- Pretend the survey is clustered by state
- Design effect

Simulation study to check the variance formula for dollar-weighted averages

- Create population data
- Simple random sampling
- Estimates and variance formula
- Check under 5000 replications

Social Indicators Survey

- Telephone survey every 2 years of NYC families
- Administered by Columbia Univ 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

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	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

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, $\hat{\theta} = \bar{y}$
  - Classical regression with just a constant term
  - Equivalent weights: $w_i = 1$

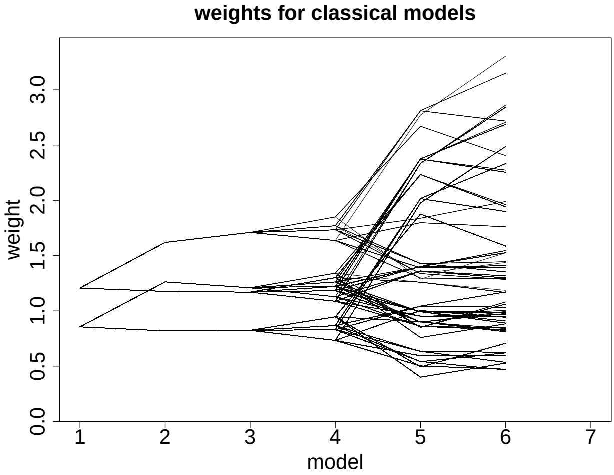
Weights corresponding to classical regressions

- Regression $y = X\beta + \epsilon$ followed by poststratification
  - $\hat{\beta}$ is a linear combination of data y
  - Vector of equivalent weights: $\frac{n}{N}(N^{\rm pop})^t X^{\rm pop}(X^t X)^{-1} X^t$
  - 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 $\times$ black/white
  - also 4 age categories
  - also 4 education categories
  - also age $\times$ education

Classical weights for CBS polls



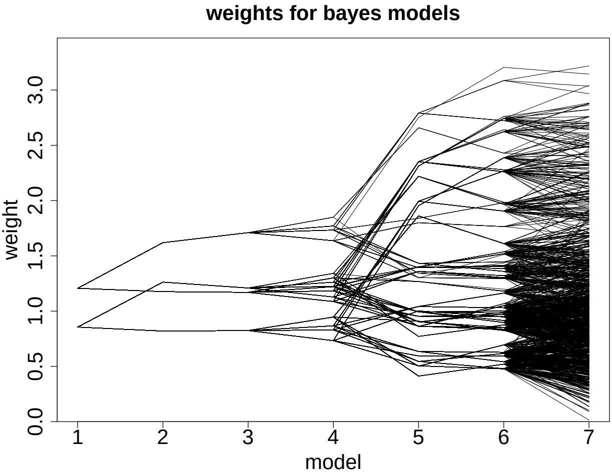
Weights corresponding to hierarchical regressions

- 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

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 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"

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!
    - 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
  - 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
  - Equivalent weights
  - When different methods give different results, we can track it back to an interaction