Design and Analysis of Sample Surveys

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

Class 4b: Ratio and regression estimation

Using auxiliary information to improve a survey

- Example: QMSS enrollment
  - Number accepted and planning to attend, 16 June 1999: 7
  - Number of entering students, Fall 1999: 11
  - Number accepted and planning to attend, 20 June 2000: 14
  - Number of entering students, Fall 2000: ?
- Ratio estimate:
  - $11 \cdot \frac{14}{7}$
- Regression estimate:
  - Get more data, fit model $y = a + bx + \text{error}$
  - Estimate is $11 + b(14 - 7)$

Using auxiliary information to improve a survey

- Example: what percentage of people volunteer for political campaigns?
  - Direct approach: estimate the percentage of people who volunteer
  - Ratio: estimate the percentage who volunteer, divided by the percentage who vote
  - Regression: get lots of data, estimate regression of percentage who attend estimate the percentage who volunteer, divided by the percentage who vote
- Formulas:
  - Direct approach: $\bar{y}$
  - Ratio: $\bar{y}\frac{\bar{X}}{\bar{x}}$
  - Regression: $\bar{y} + b(\bar{X} - \bar{x})$ (estimating b using data from many surveys)
  - $\bar{x}$ = proportion of people in survey who vote
  - $\bar{X}$ = proportion of population who vote

Ratio estimation in R

- Direct calculations:

```
is_old_white_male <- ifelse (pew$age==4 & pew$eth==1 &
  pew$male==1, 1, 0)
r.w <- mean(pew$pop.weight*pew$rvote*is_old_white_male)/
  mean(pew$pop.weight*is_old_white_male)
s2.w <- (1/(n-1))*sum((pew$pop.weight*pew$rvote*
  is_old_white_male -
  r.w*pew$pop.weight*is_old_white_male)^2)
se.w <- (1/sqrt(n))*sqrt(s2.w)/
  mean(pew$pop.weight*is_old_white_male)
```

- Using the "survey" package:

```
weighted_design <- update (weighted_design,
  is_old_white_male = (age==4 & eth==1 & male==1))
svyratio (~I(rvote*is_old_white_male),
  ~is_old_white_male, weighted_design)
```

Ratio estimation of a population ratio

- Example: volunteering and voter turnout
  - $y_i = 1$ if person i volunteered for a campaign
  - $x_i = 1$ if person i voted
  - Assume that all volunteers are voters
  - Suppose $\overline{X} = 0.6$, $\overline{x} = 0.75$, $\overline{y} = 0.15$
  - In this example, sample is more politically active than population
- Estimating $\overline{Y}/\overline{X}$ in *population*
  - Estimate is $\overline{y}/\overline{x} = 0.2$ in *sample*
  - 20% of voters in sample are volunteers

Ratio estimation of a population average or total

- Example
  - $y_i = 1$ if person i volunteered for a campaign
  - $x_i = 1$ if person i voted
  - Suppose $\overline{X} = 0.6$, $\overline{x} = 0.75$, $\overline{y} = 0.15$
- Simple estimate of $\overline{Y}$ (proportion of voters who volunteered)
  - Estimate is $\overline{y} = 0.15$
- Ratio estimate of $\overline{Y}$
  - Estimate is $\overline{y}\frac{\overline{X}}{\overline{x}} = 0.15\frac{0.6}{0.75} = 0.12$
- Estimating $N\overline{Y}$ (total number of people of volunteered)
  - Estimate is $N\overline{y}\frac{\overline{X}}{\overline{x}} = 0.12N$

Example of ratio estimation

- You have a population of insurance claims
- What is the average liability?
- Audit a random sample of claims, $i = 1, \ldots, n$
- Direct approach:
  - $y_i =$ liability of claim i
  - Estimated avg liability is $\overline{y}$, std err is $\sqrt{1-f}\,\frac{1}{\sqrt{n}} s_y$
- Ratio estimation ...?

Example of ratio estimation

- Direct approach:
  - $y_i =$ liability of claim i
  - Estimated avg liability is $\overline{y}$, std err is $\sqrt{1-f}\frac{1}{\sqrt{n}}s_y$
- Ratio estimation:
  - $x_i =$ something that's fully observed in the population
  - Estimate of $\overline{Y}$ is $\overline{y}\frac{\overline{X}}{\overline{x}}$
  - What's a good "x" here?

Standard error of ratio estimation

- Estimate $\overline{Y}$ from data $(x_i, y_i), i = 1, \ldots, n$
- Simple estimate, $\overline{y}$
  - Std err is $\sqrt{1-f}\frac{1}{\sqrt{n}}s_y$
- Ratio estimate, $\overline{y}\frac{\overline{X}}{\overline{x}}$
  - Define $r = \overline{y}/\overline{x}$
  - Define $z_i = y_i - rx_i$
  - Std err of ratio estimate is $\sqrt{1-f}\frac{1}{\sqrt{n}}s_z$
- Design effect of ratio estimation
  - $\frac{\text{variance of ratio estimate}}{\text{variance of simple estimate}} = \frac{s_z^2}{s_y^2}$
  - This should be less than 1

Standard error of other ratio estimates

- Simple estimate, $\overline{y}$
  - Std err is $\sqrt{1-f}\,\frac{1}{\sqrt{n}}s_y$
- Ratio estimate, $\overline{y}\frac{\overline{X}}{\overline{x}}$
  - Define $r = \overline{y}/\overline{x}$
  - Define $z_i = y_i - rx_i$
  - Std err of ratio estimate is $\sqrt{1-f}\,\frac{1}{\sqrt{n}}s_z$
- Ratio estimate of total, $N\overline{y}\frac{\overline{X}}{\overline{x}}$
  - Std err is $N\sqrt{1-f}\,\frac{1}{\sqrt{n}}s_z$
- Ratio estimate of ratio, $\overline{y}/\overline{x}$
  - Std err is $\frac{1}{|\overline{X}|}\sqrt{1-f}\,\frac{1}{\sqrt{n}}s_z$

Regression estimation

- Example
  - y_i = liability of insurance claim i
  - x_i = (cheap) estimate of claim i
- Estimating $\overline{Y}$ (avg liability of claims)
- Fit a regression, $y_i = a + bx_i + \text{error}$
- Regression estimate of $\overline{Y}$ is $\overline{y} + b\,(\overline{X} - \overline{x})$
- Corrects for known difference between sample and population

Regression estimation: std err

- Fit a regression, $y_i = a + bx_i + \text{error}$
- Regression estimate of $\overline{Y}$ is $\overline{y} + b(\overline{X} - \overline{x})$
- Standard error
  - Compute residuals $z_i = y_i - a - bx_i$
  - Std err is $\sqrt{1-f}\,\frac{1}{\sqrt{n}}s_z$
- Design effect
  - $\frac{\text{variance of ratio estimate}}{\text{variance of simple estimate}} = \frac{s_z^2}{s_y^2}$
  - This is $1 - R^2$
  - Optimal (that is, lowest) for least-squares regression

Regression estimation as a general framework

- Fit a regression, $y_i = a + bx_i + \text{error}$
- Regression estimate of $\overline{Y}$ is $\overline{y} + b\,(\overline{X} - \overline{x})$
- Special cases:
  - $b = 0$: unadjusted sample average
  - $b = \frac{\overline{y}}{\overline{x}}$: ratio estimation
  - $b = 1$: simple adjustment
- Regression estimation is valid for any b
  - Optimal for $b =$ least-squares estimate

More on regression estimation

- Poststratification as a special case of regression estimation (indicator for each stratum)
- Use of auxiliary information

"Double robustness" of regression estimation

- Two possible assumptions
  - Data are a simple random sample
  - Linear regression model is true
- Regression estimate $\overline{y} + b\,(\overline{X} - \overline{x})$ is unbiased if *either* assumption is true
- Can include multiple x's:
  - Fit a regression, $y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots +$ error
  - Regression estimate of $\overline{Y}$ is $\overline{y} + b_1\,(\overline{X}_1 - \overline{x}_1) + b_2\,(\overline{X}_2 - \overline{x}_2) + \ldots$
- Connection to poststratification and missing-data imputation