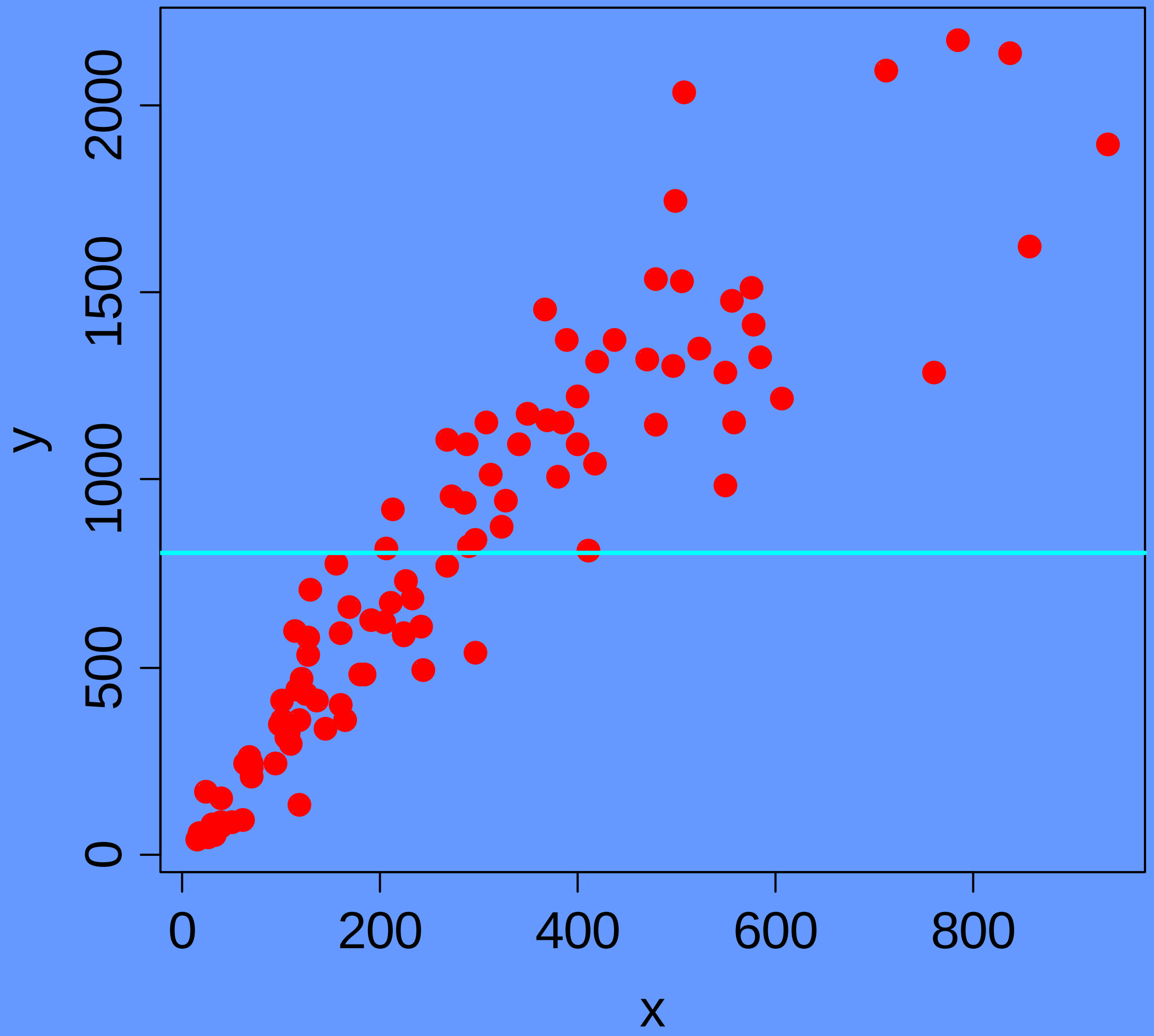
Discussion

R. Valliant

Models or Model-Free?

- Design-based inference is model-free
- An estimator can be unbiased in repeated probability sampling but biased under a model
- Easy example
  - Select simple random sample
  - Estimate population average by sample mean



0
500
1000
1500
2000
y
0
200
400
600
800
x

Design vs. Model-bias

- Design bias of sample mean is 0
- Model-bias (if straight-line thru origin) is

$$E_M\left(\bar{Y}_s - \bar{Y}_U\right) \propto \left(\bar{x}_s - \bar{x}_U\right)$$

- Model-bias has order $1/\sqrt{n}$ and so does

$$SE\left(\bar{Y}_s\right)$$

$\Rightarrow$ Confidence intervals will not have correct coverage in off-balance SRS’s

Use of Models

- Good way to develop estimators (non-Bayes or Bayes)
- Every estimator can be analyzed under a model
- If “implied” model for estimator is unrealistic, then estimator is bad
- Calibration in repeated applications needed

Long-run Calibration

- Critical to maintain acceptance
- Must be able to say we are unbiased and CI’s cover at advertised rates (regardless of methods used—design-based, model-based, Bayes, non-parametric)
- With *NR*, non-coverage (*NC*) assurance of calibration uncertain
  - Extent of and reasons for *NR*, *NC* out of our control

Coverage Problems

- HH surveys: some groups not covered by frame
  - CPS: 70% of Black males age 25-34
  - BRFSS 44 border counties: 15% of Hispanic males, 18-24
- GREG (e.g., poststrata) can correct for NR, NC
  - Useful when little known about NR's individually
- PS collapsing procedures based on cell similarity (e.g., adjacent age groups) can be biased
- Collapsing should be based on *Y*'s or coverage rates to avoid bias (Kim, Li, Valliant 2006)

How many distributions do we need?

1. Superpopulation model
2. Random selection model
3. Response model
4. Coverage model
5. Imputation model

---

6. Prior
7. Hyper-prior
8. Posterior

Logistics

- *NR*, *NC* adjustments need to consider outcomes (*Y*'s), design variables (*Z*'s), sample covariates (*X*'s), *R* (response/nonresponse)
- Weighters often have access to (*Z*,*X*,*R*) or (*Z*,*R*) only
- Editing of *Y*'s and *X*'s on parallel track
- Some *Y*'s will never be available in timely way
  - Biomarker processing—blood, urine, etc

Multiple Outcome Variables

- Surveys collect many *Y*'s
- What works for one may not work for others
  - NR adjustments, important covariates for models
- How many *Y*'s to consider?
  - How to develop compromise procedures
  - Never be able to cover all *Y*'s

Response Models

- Info needed for *R*'s and *NR*'s
- Establishment surveys may have many *Z'*s on both
- Almost nothing may be known about *NR'*s in some surveys—telephone
- Response models will be wrong
  - Omitted, unknown regressors
  - Response rates are declining
  - More uncontrolled reasons for being in nonsample ⇒
    - more problems in fitting response models
    - more problems predicting values for nonsample units

Some Issues

- Prediction models for categorical variables
  - Some surveys collect no quantitative variables
  - Ordered and unordered categorical
  - Normality assumptions unreal
- Aggregation consistency
  - Low level estimates of totals need to add to higher level estimates
- Users expect weights