The Use of Survey Weights in Regression Analysis

The Use of Test Scores in Secondary Analysis
PIAAC Methodological Seminar
Paris, 14th June 2019

Jeff Wooldridge
Department of Economics
Michigan State University
East Lansing, Michigan

1. Reasons for Using Weights
2. The Linear Model in the Population
3. Sampling Weights and Weighted Estimation
4. Regression Adjustment for Treatment Effect Estimation
5. Summary

1. Reasons for Using Weights

1. Heteroskedasticity

- No longer needed for inference.
- Could help with efficiency, even if the weights are incorrect.
- Generally inconsistent for the linear projection parameters.

2. Treatment Effects Estimation

- Propensity score weighting.
- Combined with regression, leads to "doubly robust" estimators.

3. Missing Data, Including Attrition

- Requires a kind of "missing at random" assumption.
- Mechanically similar to treatment effect estimation.

4. Survey Sampling

- The sample is not representative of the population of interest.
  - Standard stratified sampling.
  - Variable probability sampling.
  - Combinations of stratification, VP sampling, cluster sampling.
- Sampling weights are used routinely for summary statistics.

- Should the survey weights be used in regression?
  - Still a debate.
- What about for a randomized controlled trial?
- Issues:
  - Does the sampling scheme depend on the response variable, or just covariates?
  - How much are we willing to assume about our model?
  - Tradeoff between consistency and efficiency.

2. The Linear Model in the Population

- Specifying the population is important.
  - ▸ Interested in an educational production function for a specified population of students.
  - ▸ Interested in a population average treatment effect.

- The population is represented by a random vector,

$$(y, x_1, \ldots, x_K).$$

y is the dependent or response variable

x_j are explanatory variables (covariates)

- Ideally, we can estimate the conditional mean,

$$\begin{aligned} \mathrm{E}(y|x_1, x_2, \ldots, x_K) &= \mathrm{E}(y|\mathbf{x}) \\ &\equiv \mu(\mathbf{x}) \\ &= \mu(x_1, x_2, \ldots, x_K). \end{aligned}$$

- But $\mu(\mathbf{x})$ can be virtually anything.

- Linear projection of y on $(1, x_1, \ldots, x_K)$:

$$\mathrm{L}(y|1, x_1, \ldots, x_K) = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K = \beta_0 + \mathbf{x}\boldsymbol{\beta}.$$

- The population regression coefficients are

$$\underset{K\times 1}{\boldsymbol{\beta}} = [\mathrm{Var}(\mathbf{x})]^{-1}\mathrm{Cov}(\mathbf{x}, y)$$

$$\beta_0 = \mathrm{E}(y) - \mathrm{E}(\mathbf{x})\boldsymbol{\beta}.$$

- The linear projection is *definitional*.
- It exists for any y and any vector $\mathbf{x}$ (with finite second moments).

- The LP is equivalent to writing

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + u \\ \mathrm{E}(u) &= 0 \\ \mathrm{Cov}(x_j, u) &= 0, j = 1, \ldots, K \end{aligned}$$

- Why might one settle for the LP?

1. It is consistently estimated by OLS under random sampling.

- No other assumptions are needed.
- Use heteroskedasticity-robust inference.

2. The LP has an important approximation property.

- The LP is the minimum mean squared error linear approximation to the conditional mean function.
- With $\mu(\mathbf{x}) = \mathrm{E}(y|\mathbf{x})$, β_0, β_1, $\ldots$, β_K solve

$$\min_{b_0, \mathbf{b} \in \mathbb{R}^K} \mathrm{E}\{[\mu(\mathbf{x}) - b_0 - \mathbf{x}\mathbf{b}]^2\}.$$

- If $\mu(\mathbf{x})$ is linear in $\mathbf{x}$ then the CM and LP are the same:

$$\mu(\mathbf{x}) = \beta_0 + \mathbf{x}\boldsymbol{\beta}.$$

3. The LP slope parameters are often good estimates of the average partial effects.

- With $\mu(\mathbf{x}) = \mathrm{E}(y|\mathbf{x})$, if x_j is continuous, the APE is

$$\gamma_j = APE_j = E_{\mathbf{x}}\left[\frac{\partial \mu(\mathbf{x})}{\partial x_j} \right]$$

- There are some (restrictive) theoretical results that imply

$$\beta_j = \gamma_j.$$

- For example, true model is quadratic in the x_j, distribution of $\mathbf{x}$ is symmetric.

- If

$$\mu(\mathbf{x}) = \alpha + \mathbf{x}\boldsymbol{\gamma} + \sum_{j=1}^{K}\sum_{h=1}^{K}\delta_{jh}(x_j - \eta_j)(x_h - \eta_h)$$

and **x** has a symmetric distribution then

$$\beta_j = \mathrm{E}\left[\frac{\partial\mu(\mathbf{x})}{\partial x_j}\right] = \gamma_j, j = 1,\ldots,K$$

- Or $\mu(\cdot)$ is any differentiable function, **x** is multivariate normal.

4. Estimating the LP can improve efficiency in RCTs; no extra assumptions needed.

- Often called "regression adjustment" in the context of treatment effects.
- No need to have the conditional mean correctly specified.

• These four features of the LP justify the use of linear regression for discrete y.

  ▸ For example, for binary y, can justify the "linear probability model" without assume the LPM is the correct model.

• Can always linearly project onto nonlinear functions, such as squares and interactions.

  ▸ Provides a better approximation to $E(y|\mathbf{x})$.

3. Sampling Weights and Weighted Estimation

- $\mathbf{z} = (\mathbf{x}, y) \in \mathcal{Z}$.

$$y = \mathbf{x}\beta + u$$

- Partition $\mathcal{Z}$ as $\mathcal{Z}_1, \mathcal{Z}_2, ..., \mathcal{Z}_G$.
- Population (aggregate) shares:

$$\pi_g = P(\mathbf{z} \in \mathcal{Z}_g), g = 1, \ldots, G.$$

- Standard stratified sampling:

$$h_g = N_g/N \text{ (sample shares).}$$

- Weight for unit i: $w_i = \pi_{gi}/h_{g_i}$.
- "Weighted" least squares:

$$\min_{\mathbf{b}\in\mathbb{R}^K} \sum_{i=1}^{N} w_i \cdot (y_i - \mathbf{x}_i\mathbf{b})^2$$

- Consistency only requires $\mathrm{E}(\mathbf{x}'u) = \mathbf{0}$.
- Inference should account for strata as well as weights – standard in survey estimation software.

- VP sampling: p_g is the probability of retaining a draw from stratum g.
- s_i is sample selection indicator.

$$w_i = 1/p_{g_i}$$

$$\min_{\mathbf{b}\in\mathbb{R}^K} \sum_{i=1}^{N} s_i \cdot w_i \cdot (y_i - \mathbf{x}_i\mathbf{b})^2$$

- Again, consistency only requires requires $\mathrm{E}(\mathbf{x}'u) = \mathbf{0}$.

Should We Use the Sampling Weights?

A. Stratification depends (partly) on y.

- The unweighted estimator does not consistently estimate $\boldsymbol{\beta}$,

even if

$$\mathrm{E}(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$$

- Hard to justify not using weights.

B. Stratification depends on $\mathbf{x}$.

• Debate centers on two facts:

1. If $\mathrm{E}(u|\mathbf{x}) = 0$ the unweighted estimator is also consistent.

  ▸ If we add homoskedasticity in the population,

$$\mathrm{Var}(u|\mathbf{x}) = \sigma^2,$$

then the unweighted (OLS) estimator is asymptotically more efficient than the weighted estimator.

• Efficiency argues against weighting.

2. The unweighted estimator is inconsistent for $\boldsymbol{\beta}$ if we only assume

$$y = \mathbf{x}\boldsymbol{\beta} + u, \ \mathrm{E}(\mathbf{x}'u) = \mathbf{0},$$

even when stratification is based on $\mathbf{x}$.

- ▸ The weighted estimator consistently estimates the coefficients $\boldsymbol{\beta}$ in the linear projection.
- ▸ $\mathrm{E}(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$ is a special case.

• Consistency argues for weighting.

- Practical Issue:

(i) Weighted and unweighted estimates differ.

(ii) Weighted estimates imprecise.

- Can we justify the usual OLS estimates?
- Suggests a Hausman test.
- But it should be made robust.

$$\sum_{i=1}^{N} w_i \mathbf{x}_i' \left(y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_{WLS} \right) = \mathbf{0}$$

$$\sum_{i=1}^{N} w_i \mathbf{x}_i' \left(y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_{OLS} \right) \approx \mathbf{0}?$$

- Use OLS on

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + w_i \mathbf{x}_i \boldsymbol{\gamma} + u_i$$

$$\mathrm{H}_0 : \boldsymbol{\gamma} = \mathbf{0}$$

- Heteroskedasticity-robust Wald test.

4. Regression Adjustment for Treatment Effect Estimation

- d is a binary treatment indicator.
- For each unit in the population, two potential outcomes, $y(0)$ and $y(1)$.
- Population average treatment effect:

$$\tau_{ate} = \mathrm{E}[y(1) - y(0)]$$

- Along with d we observe

$$y = (1 - d) \cdot y(0) + d \cdot y(1)$$

• If d is randomized, can use simple difference in means:

$$\bar{y}_1 - \bar{y}_0$$

• With nonrandom sampling, generally have to use the sampling weights for consistency.

$$\min_{\alpha,\tau} \sum_{i=1}^{N} w_i \cdot (y_i - \alpha - \tau d_i)^2$$

• Might try to improve efficiency by using regression adjustment with covariates **x**.

- Negi and Wooldridge (2018) under random sampling.
- Separate regressions for the control and treatment groups.

$$\hat{\tau}_{ate} = N^{-1}\sum_{i=1}^{N}\left[\left(\hat{\alpha}_1 + \mathbf{x}_i\hat{\boldsymbol{\beta}}_1\right) - \left(\hat{\alpha}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}_0\right)\right]$$
$$= (\hat{\alpha}_1 - \hat{\alpha}_0) + \bar{\mathbf{x}}\left(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0\right)$$

- The asymptotic variance of $\hat{\tau}_{ate}$ is no greater than that of $\bar{y}_1 - \bar{y}_0$, and usually smaller.
- Key: No extra assumptions are made.
- The separate regressions consistently estimate the linear projections $\mathrm{L}[y(0)|1, \mathbf{x}]$, $\mathrm{L}[y(1)|1, \mathbf{x}]$.

• With survey sampling, have to use the weights for two reasons.

1. To consistently estimate the parameters in the linear projections.

  ▸ The separate regressions use the survey weights.

2. To consistently estimate $\boldsymbol{\mu}_{\mathbf{x}} = \mathrm{E}(\mathbf{x})$:

$$\tilde{\tau}_{ate} = (\tilde{\alpha}_1 - \tilde{\alpha}_0) + \left(N^{-1} \sum_{i=1}^{N} w_i \cdot \mathbf{x}_i \right) (\tilde{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_0)$$

• If weights are not used in both stages, $\tilde{\tau}_{ate}$ not guaranteed to be consistent.

5. Summary

- To ensure consistency use the sampling weights – if they can be trusted.
- Desirable to estimate the best linear approximation in the population.
  - ▸ Might get good estimates of average partial effects, including average treatment effects.

• For RCTs, need to use weights whether or not one uses simple difference in means or regression adjustment.

  ▸ The linear RA case extends to particular nonlinear methods, such as logit and Poisson.

- Can use sampling weights along with propensity score weights.

1. Treatment effects estimation.
   - ▸ Use sampling weights when estimating propensity score.
   - ▸ Combine weights with regression adjustment.
2. Missing data.
   - ▸ Again, use sampling weights when estimating propensity score.