Methods: Survey Weight Diagnostic Tests

Corbin Lubianski

January 9, 2024

1 Simulation Study 1: Wang *et al.* (2023) Study #1

The first study is a reproduction of the simulation studies from Wang *et al.* (2023) whom adapted from Pfeffermann & Sverchkov (1999). [4] A population size of $N = 3000$ was generated for (Y_i, X_i) in addition to the linear regression model

$$Y_i = 1 + X_i + \varepsilon_i,\ i = 1, \ldots, N,$$

where $X_i \overset{iid}{\sim} \text{Unif}(0, 1)$ and $\varepsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ with $\sigma \in \{0.1, 0.2\}$. Sample sizes $n \in \{100, 200\}$ were drawn from the population with the probability proportional to the weight as defined by

$$W_i = aY_i + 0.3X_i + \delta U_i,$$

where $a \in \{0, 0.2, 0.4, 0.6\}$ is the significance of the Y_i on the weights, noise U_i is noise drawn from $U_i \overset{iid}{\sim} \text{Unif}(0, 1)$ and amplified by $\delta \in \{1, 1.5\}$. Weights are not informative on $Y_i \mid X_i$ when $a = 0$ and informative when $a \neq 0$. [4]

Additionally, Wang *et al.* (2023) were interested in the diagnostic test performances when the weights are generated from a quadratic function of X and Y. They proposed the following weight generation model:

$$W_i = a(Y_i - 1.5a)^2 + 0.3X_i - 0.3X_i^2 + U_i,$$

where $U_i \overset{iid}{\sim} \text{Unif}(0, 1)$ and $a \in \{0, 0.5, 1.0, 1.5\}$. The quadratic function was designed with similar characteristics to the linear weights generating function with the additional characteristic that for $a = 1$, the partial correlation between W_i and Y_i is zero. They claim that this makes it difficult for the linear regression-based diagnostic tests to determine the importance of W_i on Y_i. [4]

Simulation Set Up

For each iteration b in B total iterations, $b = 1, 2, \ldots, B$:

1. For each generated unit (Y_i, X_i), $i = 1, 2, \ldots, N$:
   (a) Sample $X_i \overset{iid}{\sim} \text{Unif}(0, 1)$, $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, and $U_i \overset{iid}{\sim} \text{Unif}(0, 1)$.
   (b) For all i, generate $Y_i = 1 + X_i + \varepsilon_i$.
   (c) For all i, generate the weights W_i:
      - Linear Case: $W_i = aY_i + 0.3X_i + \delta U_i$;
      - Quadratic Case: $W_i = a(Y_i - 1.5a)^2 + 0.3X_i - 0.3X_i^2 + U_i$.
2. Sample n sizes from the population with inclusion probabilities $\pi_i = w_i^{-1}$.
3. Perform all aforementioned tests on the generated data.
4. Record the corresponding p-values.

The simulation has $2 \times 2 \times 2 \times 4 + 2 \times 2 \times 4 = 48$ scenarios. Note that the quadratic case is not dependent on δ hence 48 scenarios instead of 64. In the linear weight generating portion of the simulation, it will vary by sample sizes n, noise amplifier δ, σ, and weight informative scalar a. In the quadratic weight generating protion of the simulation, it will vary in the same way as the linear portion excluding varying on δ.

Cases:

1. Linear Weight Generation:
   (a) Sample Size: $n \in \{100, 200\}$
   (b) Noise Amplifier: $\delta \in \{1, 1.5\}$
   (c) $\sigma \in \{0.1, 0.2\}$
   (d) Weight Informativeness: $a \in \{0, 0.2, 0.4, 0.6\}$
2. Quadratic Weight Generation:
   (a) Sample Size: $n \in \{100, 200\}$
   (b) $\sigma \in \{0.1, 0.2\}$
   (c) Weight Informativeness: $a \in \{0, 0.5, 1.0, 1.5\}$

Constants:

- Iterations: $B = 1000$
- Population per iteration: $N = 3000$

2 Simulation Study 2: Consumer Expenditure Survey

In contrast to generated data Wang *et al.* (2023), this simulation study will sample and perform tests on complex survey data from the Bureau of Labor Statistics' Consumer Expenditure Survey (CE). The 2015 dataset is accessible via the `rpms` R package by Daniell Toth that contains consumer unit characteristics, assets, and expenditure data for consumers in the United States. [2] The Consumer Expenditure Survey data is collected by the U.S. Census Bureau for the Bureau of Labor Statistics by interviews and diary surveys. Visit the CE webpage for more information regarding methods and weighting. [3]

Performing simulations on existing survey data has the advantage of testing the diagnostic tests on the complex survey designs. Replicating the complex survey designs is difficult with generating data which makes it ideal to further test the survey weight diagnostic tests beyond the results found by Wang *et al.* (2023). For the CE data, it contains 68,415 observations on 47 variables regarding sample-design,location, housing and transportation, family, earner characteristics, labor status, income, assets, and expenditures information. In the CE data, a weight per observation unit represents the inverse sampling probability.

The focus of the data is set to describing the impact of consumer expenditures (TOTEXPCQ) on the amount of taxes paid (FINCBTAX). To ensure sufficient data quality, **TO-DO**. We expect to reject the null hypothesis that the weight is noninformative.

2.1 Sampling

Acting as if the CE data is the population, utilizing various complex sampling methods will essentially mimic the CE data's sampling structure. To determine the performance of the survey weight diagnostic tests in complex survey data, the following sampling methods were employed.

2.1.1 Grouping

Grouping is a sampling technique that groups a continuous variable into groups based on whether their numeric value is within the range of the group such that X_i is in group H if $X_i \in (a, b]$ where a, b are numeric scalars and $a < b$. This tries to mimic surveys that group potential observations given continuous X that over sample certain groups when X has a strong relationship to the variable of interest Y.

With regards to calculating the inclusion probabilities, let n be the sample size and p_H be the probability of selecting an observation unit from group H. After determining the groups based on the numeric values X, the inclusion probabilities are that of stratum in a stratified sampling method such that the inclusion probabilities is

$$\pi_{H,i} = \frac{n \cdot p_H}{N} = \frac{n_H}{N},$$

where weights for the ith observation unit in group H are $w_{H,i} = \pi_H^{-1}$.

2.1.2 Probability Proportional to Size

Probability proportional to size (PPS) is a sampling design where each unit of the population has an independent probability of being selected p_i when performing one sample. PPS sets some numeric quantity x_i proportional to the probability that the ith unit will be selected in a sample is

$$p_i = \frac{x_i}{\sum_{i=1}^{N} x_i}, \text{ with } \pi_i = \frac{n \cdot p_i}{N}.$$

Yet, survey administrators rarely have complete certainty about the numeric quantity for the observation units. Thus, an element of randomness is needed to account for variability during the survey design process. Since PPS requires x_i to be positive-definite, it is problematic to suggest an additive random noise process like the model $Z_i = Y_i + \varepsilon_i, \forall i$ where Z_i is the observed response variable, Y_i is the signal derived from the dataset, and ε is the noise term. Without imposing arbitrary distributional characteristics on ε to ensure $Z_i > 0$ for all i, consider the multiplicative regression

$$Z_i = Y_i * (1 + \varepsilon_i), \text{ where } E(\varepsilon_i) = 0 \text{ and } \varepsilon_i \overset{iid}{\sim}.$$

Without specifying the distribution of ε_i, $E(Z_i) = Y_i$. Let $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, then $\text{Var}(Z_i) = \text{Var}(Y_i) + E(Y_i^2)\text{Var}(\varepsilon_i)$.

2.1.3 Stratifying

This sampling method calculates the inclusion probabilities by stratifying on some category. This aims to produce a simplified sampling design that the Bureau of Labor Statistics employs to select survey participants. The probability of including unit i of stratum h in the sample is $\pi_{h,i} = n_h/N_h$ where N_h is the number of sampling units in stratum h with sampling weights for unit i in stratum h as $w_{h,i} = \pi_{h,i}^{-1} = N_h/n_h$.

2.2 Simulation Design

In contrast to Wang *et al.* (2023) of simulating generated data and varying model parameters, the simulation on the CE data is largely centered on varying the sampling methods and sample sizes for testing the performance of the survey weight diagnostic tests in different sampling conditions.

Simulation Set Up

For each iteration b in B total iterations, $b = 1, 2, \ldots, B$:

1. Select sampling method to select n observations from N population.
2. Calculate inclusion probabilities and corresponding weights from sample method.
3. Sample n observations.
4. Perform all aforementioned tests on sampled observations.
5. Record the corresponding p-values.

The simulation has a 4 factorial design with 20 scenarios. Varying based on sampling methods will test how each survey weight diagnostic test performs in complex sampling. Additionally, the robustness of the tests in different sample sizes is of great interest given many of the tests are asymptotically correct. [1]

Cases:

1. Sampling Method:
   (a) Grouping: $\pi_{H,i} = \frac{n \cdot p_H}{N} = \frac{n_H}{N}$ with $w_{H,i} = \pi_H^{-1}$.
   (b) Probability Proportional to Size (PPS): $\pi_i = \frac{n \cdot p_i}{N}$ with $w_i = \pi_i^{-1}$.
   (c) Stratifying: $\pi_{h,i} = \frac{n_h}{N_h}$ with $w_{h,i} = \pi_{h,i}^{-1}$.
   (d) Simple Random Sampling (Control): $\pi_i = \frac{n}{N}$ with $w_i = \frac{N}{n}$.
2. Sample Size: $n \in \{25, 50, 100, 500, 1000\}$

Constants:

- Iterations: $B = 1000$
- Population per iteration: Rows of CE dataset

3 Simulation Study 3 TO-DO

Same as Simulation Study 2 but include generalized linear models. Need to construct test code before considering the parameters of the simulation study.

4 Case Studies TO-DO

Introduce diagnostic test package with applications in notable research studies and other complex surveys like NHANES, Survey of Industrial and Service Firms (SISF), and Survey of Household Income and Wealth (SHIW). Lohr's textbook as countless datasets, though not confident that many will have a large amount of observations.

References

[1] Bollen, K. A., Biemer, P. P., Karr, F. A., Tueller, S., and Berzofsky, M. E. Are survey weights needed? a review of diagnostic tests in regression analysis. *Annual Review of Statistics and Its Applications*, 3 (2016), 375–392.

[2] Toth, D. *rpms: Recursive Partitioning for Modeling Survey Data*, 2021. R package version 0.5.1.

[3] U.S. Bureau of Labor Statistics. Consumer expenditure surveys, 2023.

[4] Wang, F., Wang, H., and Jun, Y. Diagnostic tests for the necessity of weight in regression with survey data. *International Statistical Review 91*, 1 (2023), 55–71.