HARVARD UNIVERSITY

EVALUATION OF SURVEY WEIGHT DIAGNOSTIC TESTS IN REGRESSIONS WITH COMPLEX SURVEY SAMPLING

A THESIS PRESENTED TO THE DEPARTMENT OF STATISTICS IN PARTIAL FULFILLMENT OF THE HONORS REQUIREMENT FOR THE DEGREE OF BACHELOR OF ARTS

AUTHOR

CORBIN CRAIG LUBIANSKI

ADVISOR

PROFESSOR KELLY MCCONVILLE

HARVARD COLLEGE
CAMBRIDGE, MASSACHUSETTS
APRIL 2024

Abstract

When it comes to linear regressions using complex survey data, the existing research landscape remains multifaceted and nuanced. The central question — whether to incorporate survey weights and how to discern their necessity — continues to intrigue researchers. Researchers have independently developed a portfolio of diagnostic tests for informative weights, but a comprehensive comparative analysis of their performance is still absent. In addition to reviewing the collection of diagnostic tests, step-by-step instructions were developed to help researchers to systematically apply these diagnostic tests. Several simulation studies were conducted to evaluate the sizes and powers of the tests, to determine the performance of tests within complex survey data, and to evaluate the robustness of tests under violating assumptions against proposed flexible permutation test variants. The results suggest that each test may have its own favorable settings. Collectively, the simulation results indicate that diagnostic tests are sufficient for practical use by researchers who regress with complex survey data given favorable rejection rates with informative weighting, adaptiveness to complex sampling methods, and robustness to violating assumptions.

Keywords: Survey weights; complex surveys; weighted regression; hypothesis tests

Acknowledgements

Writing a senior thesis while juggling classes, extracurriculars, job searching, assistant teaching, and "socializing" is no easy feat. I extend heartfelt thanks to everyone who demonstrated patience, understanding, and unwavering support throughout this journey. A thesis not only encapsulates the writer's work, results, and discussions, but also reflects the collective efforts of those around them.

My advisor, Professor Kelly McConville, has been an incredible resource during my thesis journey. Professor McConville not only provided guidance, but fostered my deep appreciation and awe of survey statistics that I am happy to share as one of her teaching assistants. I have gained self-confidence, direction, and ambition thanks to my advisor which few have had such an impact on me.

Looking back, I can see how my four years at Harvard have led me to this thesis. Kevin Rader's support and instruction in Stat 98 enabled me to pursue a statistics thesis. Stat 110 and Stat 111 gave me the statistical skills and tools to pursue a statistics concentration. Stat 108 gave me the coding skills to construct the complex simulations and diagnostic tests. Stat 224 taught me to go beyond my comfort level to pursue graduate level statistics courses. I want to thank all of the instructors, professors, and faculty of the Statistics Department for guiding me along my academic journey.

Lastly, I want to express my heartfelt gratitude to my friends and family. Lety and Sergio, thank you for our gaming sessions during my breaks in the suite — although perhaps they occurred a bit too often! My parents, Sherri and Craig Lubianski, and my siblings, Lauren, Lindsey, and Chase, are the true reasons why I had the privilege of writing this thesis. Their unwavering support and encouragement transformed my dream of attending Harvard into a reality. They have played an integral role in shaping who I am today; this thesis belongs to all of us!

Contents

This page intentionally left blank.

1 Introduction

In the realm of survey statistics — where the question 'to weight or not to weight' is perhaps the sole occasion where statisticians pay homage to Shakespeare amidst their numerical sonnets — statisticians are unsatisfied with the state of research regarding weights in model-based analyses, especially within regression models. This question not only gets raised in statistics but is continuously raised throughout disciplines like epidemology, economics, and social sciences. Considering that the question remains unresolved, it is first worth examining why statisticians scrutinize sample design so meticulously.

Figure 1.1: *The Literary Digest September 12th, 1936 election issue predicting Kansas Governor Alf Landon would defeat President Franklin D. Roosevelt (Minnesota Libraries, 2016).*

Public polling has been a long-standing element of political life in the United States and has provided reliable public opinion data. However, not paying considerable attention to the polls' structure can easily turn a poll from reliable to misleading. *The Literary Digest* was a popular weekly news magazine founded in 1880 and served as a respectable news source to educated and well-off clientele in the United States. Like many news magazines, a portion of some editions was dedicated to editors to speculate about the presidential election. For 1916 presidential election, the *Literary Digest* asked readers to mail in ballots indicating their preferred candidate, which — after successfully predicting the winner of four out of five predicted states — began a spree of polling for six presidential polls (1916, 1920, 1924, 1928, 1932, and 1936), a 1933 New York City mayoral poll, a 1934 California gubernatorial poll, and seven policy/issue polls (Lusinchi, 2014).

With outstanding prior success, *The Literary Digest* evolved their polling into "commerical sampling" methods for the 1916 presidential election poll by sending an approximate total of 10 million ballot postcards to subscribers, people on automobile registration lists, telephone directories, and lists of registered voters. With approximately 2.3 million postcards returned, the results of the poll predicted that Republican presidential candidate Alf Landon would receive 54% of the popular vote and 41% for the Democratic presidential candidate Franklin D. Roosevelt. However, Roosevelt would end up commanding 61% of the popular vote with only 37% for Landon. The backlash from the large prediction discrepancy demolished their subscribers' trust in the new magazine with *The Literary Digest* declaring bankruptcy soon after (Lusinchi, 2014).

What went wrong? Possible sources of error were likely undercoverage and nonresponse. Households with a telephone or automobile in 1936 were more generally more affluent than other households, which was exacerbated since opinions regarding Roosevelt's economic policies were related to socioeconomic status. Furthermore, only 2.3 million postcard ballots were returned out of 10 million, with Squire (1988) reporting that people who support Landon were more likely to return the survey. Among the many lessons learned from the demise of *The Literary Digest* is that the design of the survey is much more important than its size (Lohr, 2022).

1.1 Survey Weights

A sample is representative if it resembles the population sufficiently and provides an accurate measure of how close its estimates are to the true population values. Samples with probability sampling are generated with some specified random process in which each unit in the population has a known, nonzero probability of selection. A probability sampling procedure guarantees that each unit in the population could appear in the sample set $S = \{1, 2, \ldots, n\}$ from the population set $\mathcal{U} = \{1, 2, \ldots, N\}$. In probability sampling, the probability that each unit i in the population will appear in the selected sample is

$$\pi_i = P(\text{unit } i \text{ in sample}), \text{ with } \pi_i \in (0, 1].$$

Assume that the π_i are known in advance of sampling. The simpliest form of probability sampling is a simple random sample (SRS), where a sample of size n is taken where all population units N have the same probability of being in the sample (Lohr, 2022). Thus, each unit in an SRS has an inclusion probability $\pi_i = n/N$.

Large-scale statistical surveys seldom use SRS where more complex sampling designs are utilized to minimize sampling costs, minimize variability to increase estimator efficiency, or to improve the quality of the data sampled. To draw conclusions with samples from the population, inclusion probabilities are used to extrapolate the sample to the population. Define a survey weight as the inverse of the inclusion probability:

$$w_i = \frac{1}{\pi_i}.$$

Intuitively, a survey weight can be interpreted as the number of population units that the ith sample unit represents. For units that are very likely to be sampled where π_i is large, the unit only represents fewer population units in the population, whereas for units with low probabilities, being sampled will represent many other units. Samples where all units have the same survey weights are called self-weighting samples such as SRS where $w_i = N/n$ for $i \in U$. Survey weights are useful for reconstructing population statistics such as population size N, population total t, and population mean $\bar{y}_{\mathcal{U}}$ where

$$\hat{N} = \sum_{i \in S} w_i, \ \hat{t} = \sum_{i \in S} w_i y_i, \text{ and } \bar{y} = \frac{\sum_{i \in S} w_i y_i}{\sum_{i \in S} w_i} = \frac{\hat{t}}{\hat{N}}.$$

Survey weights are essential to avoid bias when estimating population means and proportions for descriptive analyzes. Famously, the classic Horvitz-Thompson (HT) estimator

$$\hat{t} = \sum_{i \in S} \frac{y_i}{\pi_i} = \sum_{i \in S} w_i y_i,$$

proposed in Horvitz & Thompson (1952), found that unbiased estimators can be obtained using survey weights under any sampling design with known inclusion probabilities.

1.2 Motivation

In contrast to the general consensus that survey weights are necessary for population-level estimates like means and ratios, the question of whether researchers should use survey weights to model relationships between response and explanatory variables has been widely debated in literature (Kish & Frankel (1974); Gelman (2007)). A major drawback of using weights which are not informative for the sample is they can substantially increase the variance of the model parameter estimates (Bollen *et al.*, 2016). Furthermore, the current use of survey weights in regression analysis is largely dependent on the field of study, rather than any empirical metric with the data. For instance, Bollen *et al.* (2016) claim that biostatistics and public health generally use weights, while social sciences generally do not. Metrics to give researchers an empirical justification for using weights are not widely available.

Survey weight diagnostic tests are formal model misspecification tests that determine the necessity for weighting within regression models. Ideally, these tests could help researchers determine whether to include weights in their analysis. A comprehensive review of the current literature on diagnostic tests by Bollen *et al*. (2016) highlights the lack of cross-comparisons between tests despite the considerable amount of proposed tests. Bollen *et al*. (2016) additionally classify existing tests into two groups: difference-in-coefficients (DC) and weight association (WA) tests. For the portfolio of tests, Bollen *et al*. (2016) noted that the existing Monte Carlo simulation studies on the finite sample performances of these tests are limited. Many simulation studies were designed to illustrate the new tests with limited scope to demonstrate their potential. Bollen *et al*. (2016)'s review also noted unaddressed questions regarding the tests:

1. Which tests should be used and under what conditions?
2. Based on the asymptotic properties between WA and DC tests, at what point do the tests' performances become equivalent?
3. In what situations are DC and WA tests interchangable and when is one favored over the other?
4. How flexible are the tests given heteroscedastic nested-data structure?
5. Which tests have the best finite sample properties?
6. How sensitive are the tests to various complex sampling designs?
7. How adaptable are the tests for categorical variables?

1.3 Outline

The goal of this thesis is to resolve many of the unaddressed questions in Bollen *et al*. (2016) and to provide insight for researchers to determine the necessity of weights in their regression models. Chapter 2 summarizes the broad literature of survey weight tests from Bollen *et al*. (2016) and Wang *et al*. (2023) while denoting step-by-step instructions on how to implement the tests. Chapter 3 is a simulation study to replicate the results from Wang *et al*. (2023)'s simulation studies which are adapted from Pfeffermann & Sverchkov (1999) and Wu & Fuller (2005). Chapter 4 is a simulation study determining the sensitivity of complex sampling on survey weight diagnostic tests using a 2015 Consumer Expenditure dataset from the Bureau of Labor Statistics. Chapter 5 is a simulation study that determines how robust diagnostic tests are when test assumptions are violated by comparing their rejection rates against proposed permutation tests. Chapter 6 summarizes the findings and provides recommendations for future work.

2

Diagnostic Survey Weight Tests

As often used in areas of statistics and other fields of study, regression analysis is based on a model that is presumed to describe a relationship between the explanatory variable X and a response variable Y. A simple linear regression model can be described as

$$Y_i \mid X_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where Y_i is the response variable, X_i is the explanatory variable, β_0 and β_1 are unknown coefficient parameters, and ε_i is the regression error for observation i.

While there are no assumptions needed to compute β_0 and β_1, extrapolating these calculations to infer about the true unknown linear relationship parameters β_0 and β_1 requires four main assumptions:

1. Linearity: $E(\varepsilon_i \mid X_i) = 0$, for all i;
2. Homoscedasticity: $\text{Var}(\varepsilon_i \mid X_i) = \sigma^2$, for all i;
3. Independence between observations: $\text{Cov}(\varepsilon_i, \varepsilon_j \mid X_i, X_j) = 0$, for all $i \neq j$;
4. Normality for ε_i.

In the context of sampling using complex survey sampling (i.e., departing from simple random samples), it can be hard to justify that complex survey samples meet all four main assumptions. Specifically, observations may have different inclusion probabilities π_i in complex selection designs such as stratified and cluster sampling. Complex selection designs introduce positive correlations between errors ε_i of the model which violates the assumption of independence between observations.

Furthermore, if π_i is related to Y_i — which is often the case in constructing representative survey weights W_i — failing to take into account the different probabilities of selection may lead to bias in the estimated regression parameters. See Kish & Frankel (1974) for more information on how unequal survey weights affect regression coefficients and standard errors.

2.1 Survey Weight Regressions

Consider a regression analysis from survey data of sample S with size n from a finite population $\mathcal{U}$ with size N. The observed survey data S is $\{Y_k, X_k, W_k\}_{k\in S}$ where W_k is the survey weight associated with the ith observation unit which does not necessarily have to be the inverse of the selection probability. A multivariate model for the sample S is

$$\vec{Y} = \mathbf{X}^\top \vec{\beta} + \vec{\varepsilon}$$

where $\vec{Y} = (Y_1, \ldots, Y_n)^\top$ is a vector of response variables $n \times 1$, $\mathbf{X} = (X_1^\top, \ldots, X_p^\top)^\top$ is a $n \times p$ matrix of the explanatory variables (including a column component of ones for calculating the intercept), $\vec{\beta}$ is a $p \times 1$ vector of regression coefficients, and $\vec{\varepsilon}$ is a $1 \times n$ vector of regression errors. Regarding regression weights, survey weights $\vec{W}$ are a $1 \times n$ vector.

For the observed survey data, the least squares estimators for unweighted and weighted regression coefficients $\vec{\beta}_u$ and $\vec{\beta}_w$, respectively, are

$$\hat{\vec{\beta}}_u = \frac{\mathbf{X}^\top \vec{Y}}{\mathbf{X}^\top \mathbf{X}},$$

$$\hat{\vec{\beta}}_w = \frac{\sum_{k\in S} W_k \vec{X}_k^\top Y_k}{\sum_{k\in S} W_k \vec{X}_k^\top \vec{X}_k} = \frac{\mathbf{X}^\top \mathbf{H} \vec{Y}}{\mathbf{X}^\top \mathbf{H} \mathbf{X}}, \text{ where } \mathbf{H} = \mathrm{diag}(\vec{W}).$$

Researchers are interested in testing the necessity of using survey weights in fitting their observed sample data to estimate $\vec{\beta}$ to determine whether weights are needed to obtain unbiased estimates of the population parameter $\vec{\beta}$. Specifically, unnecessarily incorporating survey weights can lead to inflated standard error estimates for the regression coefficients. Bollen *et al*. (2016) classified two large categories of survey weight diagnostic tests as difference-in-coefficients tests and weight association tests. The article concludes by establishing the asymptotic equivalence between the two test categories. In addition to the two test categories, Wang *et al*. (2023) adds to the Bollen *et al*. (2016) review by noting other diagnostic survey weight tests that do not fail under the test category umbrellas. Unless explicitly told otherwise, the following diagnostic tests assume the standard linear regression assumptions of linearity, homoskedasticity, independence conditional on weights, and Normality for $\vec{\varepsilon}$.

> Survey weight diagnostic tests are only meant to be used as a diagnostic tool to determine whether weights should be used in a regression analysis approach. Survey weight diagnostic tests should not be used to draw causal relationships between $\vec{Y}$ and $\mathbf{X}$ such that they should only be limited to testing the necessity of survey weights in regressions.

2.2 Model-Based and Design-Based Population Perspectives

Before continuing, it is important to address the different perspectives of the population taken in regression inference, as it is a common source of confusion regarding weights in the inferential analysis. Concisely, the model-based perspective is regarded as the infinite population approach, while the design-based perspective is regarded as the finite population approach. Although large-sample inferences of the two perspectives are often similar, assumptions become notable in smaller sample sizes that typically coincide with complex survey data.

2.2.1 Model-Based Inference

Consider a stochastic model that describes the relation between Y_i and $\vec{X}_i$ that holds for every observation in the population for $i \in \mathcal{U}$. One common model is to claim that Y_i is generated from $\vec{X}_i$ as

$$Y_i \mid \vec{X}_i = \vec{X}_i^T \vec{\beta} + \varepsilon_i,$$

where $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and independent of $\vec{X}_i$. If the observations in the population follow the model, then the sample design has no effect as long as the inclusion probabilities depend only on y. Some researchers may relax these assumptions such as replacing $\vec{\varepsilon} \perp\!\!\!\perp \mathbf{X}$ with $\mathrm{Cov}(\vec{\varepsilon}, \mathbf{X}) = 0$.

The model-based approach is attractive for regressions, as it easily relates to other statistical areas and is flexible in attributing aspects of the survey design such as nonresponse. In addition, model-based estimates can be used with relatively small samples and nonprobability samples. Under the aforementioned assumptions, the ordinary least squares (OLS) estimator of β is unbiased and consistent (Lohr, 2022).

2.2.2 Design-Based Inference

The design-based approach is interested in the finite population characteristics β in the model

$$\vec{Y} = \vec{X}_i^\top \vec{\beta} + \varepsilon_i,$$

where $\vec{Y}$ is an $N \times 1$ vector, $\mathbf{X}$ is an $N \times k$ matrix of explanatory variables with a column of intercepts, $\vec{\beta}$ as the $k \times 1$ vector of coefficients, and $\vec{\varepsilon}$ is an $N \times 1$ vector of model residuals. The finite population OLS estimator is

$$\vec{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \vec{Y}.$$

No assumptions are made about the fit of the model to the population or the distribution of residual errors. Inferences are based on repeated sampling from a finite population. Design-based analysts decide to fit a particular model if they believe that it is a plausible candidate to describe the population. Since all inference are based on the survey design, design-based inferences require survey weights to estimate the parameters, and the survey design to estimate the variance of the regression coefficients.

2.2.3 Model-Assisted Inference

As a combination of design-based and model-based approaches, Särndal *et al.* (1992) proposed a model-assisted approach that is assumed to be generated from a finite population, yet all inference is based on the survey design. Thus, survey weights are used to estimate regression coefficients β and standard errors are calculated by survey design. This produces a consistent estimator

$$\vec{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \vec{Y}$$

if $\mathrm{Cov}(W_i, Y_i \mid \vec{X}_i) = 0$ for all i (Särndal *et al.*, 1992). Acknowledging this other approach is imperative to understanding most of the weighting tests as much of the surrounding literature utilize this perspective implicitly. Bollen *et al.* (2016) argues this a major source of researchers' misunderstanding about survey weights.

2.3 Difference-in-Coefficient Tests

Difference-in-coefficients (DC) tests compare the coefficients of the weighted and unweighted regressions to determine whether the coefficient differences are statistically significantly different from zero. Starting with

$$\vec{Y} = \mathbf{X}\beta + \vec{\varepsilon}, \text{ assuming } E(\vec{\varepsilon} \mid \mathbf{X}) = 0 \text{ and } \mathrm{Var}(\vec{\varepsilon} \mid \mathbf{X}) = \sigma^2 \mathbf{I}.$$

Hausman (1978) created the basis of the DC test as a test for general misspecifications. Consider two linear regressions which output two equally sized estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ of the β estimators. In a correctly specified model, the asymptotic value of $(\hat{\beta}_1 - \hat{\beta}_2)$ should be zero. Otherwise, if there is misspecification, then $(\hat{\beta}_1 - \hat{\beta}_2)$ should be nonzero. Hausman's proposed test statistic T_H is

$$T_H = (\hat{\beta}_1 - \hat{\beta}_2)' \hat{V}_H^{-1} (\hat{\beta}_1 - \hat{\beta}_2)$$

where $\hat{V}_H = \hat{V}(\hat{\beta}_1) - \hat{V}(\hat{\beta}_2)$ as the estimator of the asymptotic covariance matrix. Lastly, $T_H \sim \chi^2_k$ with degrees of freedom equal to the dimension of $\hat{\beta}$ (Hausman, 1978).

2.3.1 Hausman-Pfeffermann DC Test

Pfeffermann (1993) proposed using the Hausman test for misspecification as a test to compare the coefficients of weighted and unweighted regressions as $\hat{\beta}_1 = \hat{\beta}_w$ referring to the coefficients of the weighted regression and $\hat{\beta}_2 = \hat{\beta}_u$ as the coefficients of the unweighted regression. This also corresponds with the covariance matrix estimator $\hat{\mathbf{V}} = \hat{\mathbf{V}}(\hat{\beta}_w) - \hat{\mathbf{V}}(\hat{\beta}_u)$.

A notable issue with this test statistic is the event in which the covariance estimator is negative, which could correspond to a negative chi-squared test statistic. As probability theory defines the variance of random variables as non-negative, Hausman (1978) proposed this covariance estimator under the null hypothesis, $\mathrm{Cov}(\hat{\beta}_u, \hat{\beta}_w - \hat{\beta}_u) = 0$. Unfortunately, this estimator is not necessarily positive definite, especially for small and moderate sample sizes when $\hat{\beta}_w$ can inflate.

Asparouhov-Muthen Variance Estimator Adjustment

Asparouhov & Muthen (2007) extended the Hausman-Pfeffermann test by proposing a different estimator for V that is always positive definite. Specifically, they proposed

$$\hat{\mathbf{V}}_{AM} = \hat{\mathbf{V}}(\hat{\beta}_w) + \hat{\mathbf{V}}(\hat{\beta}_u) - 2\hat{C}$$

where $\mathbf{C}$ is an estimator of the covariance matrix of the two estimators as

$$\mathbf{C} = \left(\frac{\partial^2 L_1(\hat{\beta}_{w_1})}{(\partial\beta)^2}\right)^{-1} \mathbf{M} \left(\frac{\partial^2 L_1(\hat{\beta}_{w_1})}{(\partial\beta)^2}\right)^{-1'}$$

$$M = \sum_{k\in S} w_{1,k} w_{2,k} \frac{\partial l_k(\hat{\beta}_{w_1})}{\partial\beta} \left(\frac{\partial l_k(\hat{\beta}_{w_2})}{\partial\beta}\right)'.$$

Let $w_{1,k} = 1$ and $w_{2,k} = W_k$ and l_k be the log-likelihood for the kth observation. The proposed estimator of $\mathbf{V}$ is positive definite, even for small sample sizes (Asparouhov & Muthen, 2007). However, $\mathbf{C}$ can be difficult to compute if the standard linear regression assumptions do not hold for a sample S. Asparouhov & Muthen (2007) conducted a limited simulation study comparing the Hausman-Pfeffermann test with its variance estimator $\hat{V}$ and found their modifications to reduce the large Type I error rates associated with the Hausman-Pfeffermann test under sufficiently large sample sizes (Bollen *et al.*, 2016).

Kott Variance Estimator Adjustment

Kott (2018) proposed an explicit variance estimator using a "model-based design-sensitive" regression approach. The estimation procedure is to assign copies of each observation unit to identical sampling PSUs, then assign one of the copies with equal inclusion probability weights as the unweighted regression coefficients $\vec{\beta}_u$ and the other with unequal inclusion probability weights as the weighted regression coefficients $\vec{\beta}_w$. Then, the unweighted copy covariates $\mathbf{x}_k^\top$ are replaced by $\mathbf{x}_k^\top \mathbf{x}_k^\top$ and the weighted copy is $\mathbf{x}_k^\top \mathbf{0}^\top$. Finally, running a linear regression to obtain the regression coefficients $\mathbf{d} = (\vec{\beta}_u, \vec{\beta}_w - \vec{\beta}_u)^\top$ is simple with design-based statistical software (Kott, 2018).

Wang-Wang-Yan Estimator Adjustment

In the Wang *et al.* (2023) review of diagnostic tests and simulation study, they proposed a more direct estimator of $\hat{\mathbf{V}} = \hat{\sigma}^2 \mathbf{A}\mathbf{A}^\top$, where

$$\mathbf{A} = (\mathbf{X}^\top \mathbf{H}\mathbf{X})^{-1}\mathbf{H} - (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$$

with $\mathbf{H} = \text{diag}(\vec{W})$ and $\hat{\sigma}^2$ is the estimator of the least squares σ^2 under the null hypothesis of noninformative weights.

Steps for performing the Hausman-Pfeffermann DC Test with Wang-Wang-Yan variance estimator, given $\{Y_k, \vec{X}_k, W_k\}_{k \in S}$:

1. Estimate $\vec{\beta}_u = (\mathbf{X}^\top \mathbf{X})^{-1}(\mathbf{X}^\top \vec{Y})$.
2. With $\mathbf{H} = \text{diag}(\vec{W})$, estimate $\vec{\beta}_w = (\mathbf{X}^\top \mathbf{H}\mathbf{X})^{-1}(\mathbf{X}^\top \mathbf{H}\vec{Y})$.
3. Compute $\hat{\sigma}^2 = (n - p + 1)^{-1} \sum_{k=1}^{n} \varepsilon_k$ where $\varepsilon_k = Y_k - \vec{X}_k^\top \hat{\beta}_u$.
4. Estimate $\hat{V} = \hat{\sigma}^2 \mathbf{A}\mathbf{A}^\top$ where $\mathbf{A} = (\mathbf{X}^\top \mathbf{H}\mathbf{X})^{-1}\mathbf{H} - (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$.
5. Calculate the chi-square test statistic $T_H = (\hat{\beta}_w - \hat{\beta}_u)^\top \hat{V}^{-1}(\hat{\beta}_w - \hat{\beta}_u)$.
6. Determine p-value with $T_H \sim \chi^2_p$.

2.4 Weight Association Tests

The basis for many weight association (WA) tests stems from Hausman (1978) misspecification tests with the intention of assessing the statistical significance of $\vec{\beta}_M$ in equation

$$Y = X\vec{\beta} + X_M\vec{\beta}_M + \varepsilon$$

where X_M is the transformed version of X. The null hypothesis is $H_0 : \beta_M = 0$ such that the regression coefficients of the weighted explanatory variables are non-information of Y after already regressing on the untransformed covariates. Many WA tests require the normality assumption for ε_i to perform F tests, which is not assumed as in DC tests. More generally, weight association tests evaluate whether

$$H_0 : E(\vec{Y} \mid \mathbf{X}, \vec{W}) = E(\vec{Y} \mid \mathbf{X})$$

$$H_A : E(\vec{Y} \mid \mathbf{X}, \vec{W}) \neq E(\vec{Y} \mid \mathbf{X}).$$

2.4.1 DuMouchel-Duncan WA Test

Although Hausman (1978) only specified the regression as a misspecification test, DuMouchel & Duncan (1983) extended the test to determine the necessity of weighting in regressions. As a simple WA test, consider the regression

$$\vec{Y} = \mathbf{X}\vec{\beta}_u + \mathbf{X}_w\vec{\beta}_w + \vec{\varepsilon}.$$

DuMouchel & Duncan (1983) recommend estimating the regression model with ordinary least squares (OLS) and then testing the null hypothesis of $H_0 : \vec{\beta}_w = 0$ using an F-test to determine whether weights are needed in the analysis.

Steps for performing the DuMouchel-Duncan WA Test, given $\{Y_k, \vec{X}_k, W_k\}_{k \in S}$:

1. Create $\tilde{\mathbf{X}} = \mathbf{HX}$, then augment the matrices $\mathbf{X}$ and $\tilde{\mathbf{X}}$ to form the covariate matrix for the full model $\mathbf{X}_{\text{full}}$ such that $\mathbf{X}_{\text{full}} = [\mathbf{X}, \tilde{\mathbf{X}}]$. For the reduced model, let $\mathbf{X}_{\text{reduced}} = \mathbf{X}$. Both covariate matrices should include a column of ones for the intercept.
2. For full and reduced models, compute β estimates.
3. For full and reduced models, calculate the sum of squared errors (SSE) by summing the squared differences between $\hat{\vec{Y}}$ and $\vec{Y}$.
4. Compute test statistic T_H with p being the number of regression parameters under the models as

$$T_H = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/(p_{\text{full}} - p_{\text{reduced}})}{SSE_{\text{full}}/(n - p_{\text{full}} - 1)}.$$

5. Calculate p-value with

$$T_H \sim F_{df_{\text{reduced}} - df_{\text{full}},\, df_{\text{full}}}.$$

2.4.2 Pfeffermann-Sverchkov (1999) WA Test

Pfeffermann and Sverchkov proposed multiple WA tests in a sequence of works. Pfeffermann & Sverchkov (1999) derived several tests in which they investigate the relationships between the unweighted residuals of the sample and the weights in a regression. They argue that if the sample distribution of the residuals is the same as the population distribution, then you can ignore the weights to then use an unweighted regression (Bollen *et al.*, 2016). Let $\hat{\varepsilon}_u = \vec{Y} - \mathbf{X}\hat{\beta}_u$, be the unweighted residuals. Firstly, Pfeffermann & Sverchkov (1999) considered the null hypotheses

$$H_{0,j} : \text{Corr}(\hat{\varepsilon}_u^j, \vec{W}) = 0, j = 1, 2, \ldots$$

For a given j, the sample correlation after Fisher transformation asymptotically follows a Normal distribution. Although the range of j is not specified, the j choices of two or three are generally sufficient to test the null hypothesis.

Additionally, Pfeffermann & Sverchkov (1999) proposed regressing $\vec{W}$ on $\hat{\varepsilon}_u^j$ such that

$$E(\vec{W} \mid \hat{\varepsilon}_u^k) = \alpha + \beta^{(j)}\hat{\varepsilon}_u^j, \text{ with } j \in \{1, 2, 3\},$$

with intercept α and slope coefficient $\beta^{(j)}$. For a given j, perform a t-test with $H_{0,j} : \beta^{(j)} = 0$. For any of jth t-tests, a statistically significant p-value is sufficient to reject the null hypothesis of noninformative weights for the model. Pfeffermann & Sverchkov (1999) report that the two variations of the WA test have similar performance, but an adjustment by Wang *et al.* (2023) will likely improve the performance more.

Wang-Wang-Yan Adjustment

Wang *et al.* (2023) sought to address two limitations of the test: multiple testing issues for $j \in \{1, 2, 3\}$ and the regression model for $\vec{W}$ does not condition on $\mathbf{X}$ which may harbor high correlation between $\vec{W}$ and ε_u due to $\mathbf{X}$. They propose a simple modification by regressing $\vec{W}$ on the first two moments and an interaction with $\mathbf{X}$:

$$E(\vec{W} \mid \varepsilon_u) = f(\mathbf{X}; \eta) + \sum_{j=1}^{2} \beta^{(j)} \varepsilon_u^j + \text{diag}(\varepsilon_u)\mathbf{X}\gamma,$$

where $f(\mathbf{X}; \eta)$ is a function of $\mathbf{X}$ with scalar parameter η, scalar coefficients $\beta^{(1)}$ and $\beta^{(2)}$, and γ is a $p \times 1$ coefficient vector for the interaction between $\mathbf{X}$ and ε. Finally, test the null hypothesis $H_0 : \beta^{(1)} = \beta^{(2)} = \gamma = 0$ by an F-test (Wang *et al.*, 2023).

Steps for performing the Pfeffermann-Sverchov (1999) WA Test with Wang-Wang-Yan adjustment, given $\{Y_i, \vec{X}_i, W_i\}_{k \in S}$:

1. Compute the unweighted regression $E\left(\vec{Y} \mid \mathbf{X}\right)$ and calculate the residuals $\varepsilon_u = \vec{Y} - \mathbf{X}\hat{\beta}_u$.
2. Construct the full model matrix $\mathbf{X}_{full} = [\mathbf{X}, \varepsilon, \varepsilon^2, \tilde{\mathbf{X}}]$ with $\tilde{\mathbf{X}} = \text{diag}(\varepsilon)\mathbf{X}$. For the reduced model, let $\mathbf{X}_{\text{reduced}} = \mathbf{X}$. Both covariate models should include a column of ones for the intercept. Given the specified function $f(\mathbf{X}; \eta)$, the full and reduced covariate matrices can change. Simple forms of $f(\mathbf{X}; \eta)$ are linear and quadratic.
3. For full and reduced models, compute β estimates.
4. For full and reduced models, calculate the sum of squared errors (SSE) by $SSE = \sum_{k \in S}(\hat{W}_k - W_k)^2$.
5. Compute test statistic T_H as
$$T_H = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/(p_{\text{full}} - p_{\text{reduced}})}{SSE_{\text{full}}/(n - p_{\text{full}} - 1)}.$$
6. Calculate p-value with
$$T_H \sim F_{df_{\text{reduced}} - df_{\text{full}},\, df_{\text{full}}}.$$

2.4.3 Pfeffermann-Sverchkov (2007) WA Test

Pfeffermann & Sverchkov propose another WA test based on regressing $\vec{W}$ on both $\mathbf{X}$ and $\vec{Y}$ such that

$$E\left(\vec{W} \mid \mathbf{X}, \vec{Y}\right) = \eta\mathbf{X} + \gamma\vec{Y}.$$

Conducting a t-test for the null hypothesis $H_0 : \gamma = 0$ determines informative weights for $\vec{Y}$ if the null hypothesis is rejected (Pfeffermann & Sverchkov, 2007). Note that the test was created in the context of small area estimation while Bollen *et al.* (2016) presented it as a more general test for weights.

Wang-Wang-Yan Adjustment

Wang *et al.* (2023) critiques the regression model $E(\vec{W} \mid \mathbf{X}, \vec{Y})$ since it would only captures a linear relationship between $\vec{W}$ and $(\mathbf{X}, \vec{Y})$. Thus, Wang *et al.* (2023) suggest capturing possible non-linear relationships by considering

$$E\left(\vec{W} \mid \mathbf{X}, \vec{Y}\right) = f(\mathbf{X}; \eta) + \sum_{j=1}^{2} \gamma_j \vec{Y}^j,$$

where $f(\mathbf{X}; \eta)$ is a function of $\mathbf{X}$ with parameter η, coefficient γ_j of $\vec{Y}^j$. Finally, test the null hypothesis $H_0 : \gamma_1 = \gamma_2 = 0$ with an F-test to determine whether $\vec{W}$ and $\vec{Y}$ are associated conditional on $\mathbf{X}$ (Wang *et al.*, 2023).

Steps for performing the Pfeffermann-Sverchov (2007) WA Test with Wang-Wang-Yan adjustment, given $\{Y_k, \vec{X}_k, W_k\}_{k \in S}$:

1. Construct the full model matrix $\mathbf{X}_{full} = [\mathbf{X}, \vec{Y}, \vec{Y}^2]$. For the reduced model, let $\mathbf{X}_{\text{reduced}} = \mathbf{X}$. Both covariate models should include a column of ones for the intercept.
2. For full and reduced models, compute β estimates.
3. For full and reduced models, calculate the sum of squared errors (SSE) by summing the squared differences between $\hat{\vec{Y}}$ and $\vec{Y}$.
4. Compute test statistic T_H as
$$T_H = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/(p_{\text{full}} - p_{\text{reduced}})}{SSE_{\text{full}}/(n - p_{\text{full}} - 1)}.$$
5. Calculate p-value with
$$T_H \sim F_{df_{\text{reduced}} - df_{\text{full}},\, df_{\text{full}}}.$$

2.4.4 Wu-Fuller WA Test

As another special case of the Hausman (1978) misspecification regression test, Wu & Fuller (2005) extended the model in DuMouchel & Duncan (1983) by changing the way $\mathbf{X}$ is transformed in the regression. Consider the regression

$$\vec{Y} = \mathbf{X}^\top \beta + \widetilde{\mathbf{X}} \widetilde{\beta} + \vec{\varepsilon},$$

where $\widetilde{\mathbf{X}} = \mathbf{Q}\mathbf{X}$, $\mathbf{Q} = \text{diag}(q_1, q_2, \ldots, q_n)$, and $q_k = W_k \hat{W}_k^{-1}(\vec{X}_k)$ where $\hat{W}_k$ is estimated by regressing W_k on $f(\vec{X}_k; \eta)$.

Adapted from the regression by Pfeffermann & Sverchkov (1999) for modeling survey data, Wu & Fuller (2005) uses it to check the impact of $\vec{W}$ on $\vec{Y}$ after removing any information from $\mathbf{X}$. Testing the model with the null hypothesis $H_0 : \tilde{\beta} = 0$ determines the impact of $\vec{W}$ on $\vec{Y}$ after removing the information contained in $\mathbf{X}$ as q_k are the predictable factors of weight W_k by $\vec{X}_i$ (Wu & Fuller, 2005).

Special care should be taken to determine $f(\mathbf{X}; \eta)$ since Pfeffermann & Sverchkov (2003) warns about how mischaracterizing the relationship between $\vec{W}$ and $\mathbf{X}$ can result in incorrect size and poor power of the misspecification test. Properly determining the relationship, like through a model building process, may help improve beneficial for the test's performance.

Steps for performing the Wu-Fuller WA Test, given $\{Y_k, \vec{X}_k, W_k\}_{k \in S}$:

1. Compute the regression of $E(\vec{W} \mid \mathbf{X}) = f(\mathbf{X}; \eta)$ and estimate $\hat{W}_k$ for $k \in S$.
   - Reasonable choices for $f(\mathbf{X}; \eta)$ may include linear and quadratic relationships.
2. With $\mathbf{Q} = \mathrm{diag}(\vec{q})$, create $\widetilde{\mathbf{X}} = \mathbf{Q}\mathbf{X}$.
3. Augment the matrices $\mathbf{X}$ and $\tilde{\mathbf{X}}$ to form the covariate matrix for the full model $\mathbf{X}_{\text{full}}$ such that $\mathbf{X}_{\text{full}} = [\mathbf{X}, \tilde{\mathbf{X}}]$. For the reduced model, let $\mathbf{X}_{\text{reduced}} = \mathbf{X}$. Note that both covariate matrices should include a column of ones for the intercept.
4. For full and reduced models, compute β estimates.
5. For full and reduced models, calculate the sum of squared errors (SSE) by summing the squared differences between $\hat{\vec{Y}}$ and $\vec{Y}$.
6. Compute test statistic T_H as
$$T_H = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/(p_{\text{full}} - p_{\text{reduced}})}{SSE_{\text{full}}/(n - p_{\text{full}} - 1)}.$$
7. Calculate p-value with
$$T_H \sim F_{df_{\text{reduced}} - df_{\text{full}},\, df_{\text{full}}}.$$

2.5 Other Tests

Beyond the parametric WA and DC tests reviewed by Bollen *et al.* (2016), there are additional diagnostic tools that may help researchers determine whether weights are necessary in their regression analysis. Some consist of formal parametric tests or informal judgement calls.

1. **Bayesian statistics** provides another perspective on weighting, yet there are no proposed tests for weights from a Bayesian perspective. It is an opportunity to depart from frequentist statistics, as most survey weight diagnostic tests utilize. Bayesian inference using linear regressions is an active part of survey data inference literature and available for researchers via the `rstanarm` R-package. See Si *et al.* (2020) for more information.
2. **Standard Errors** are influenced by the survey design and consider how weighted regressions generally increase standard error estimates. Gelman (2007) provides discussion on how to navigate this issue, though does not offer a diagnostic test. Gelman (2007) recommends to use the same procedure used to create the weights to compute the standard errors.
3. **Confidence Intervals** was considered as an informal DC test by Bollen *et al.* (2016). Fitting models with and without weights and assessing whether the associated confidence intervals of the regression coefficient estimates overlap is a crude

diagnostic test. Schenker & Gentleman (2001) recommend to use confidence intervals only when more formal DC tests are not available. It is important to take into account possible heteroskedasticity when weighting. Preliminary test designs had a difficult time determining the weight informativeness since weighted and unweighted coefficient estimates $\vec{\hat{\beta}}_w$ and $\vec{\hat{\beta}}_u$, respectively, often had overlapping 95% confidence intervals due to similar coefficient magnitudes and standard errors. See the Conclusion for discussion about how informal methods compare to diagnostic tests.

2.5.1 Pfeffermann-Sverchkov Estimation Test

As a separate test from the WA and DC tests, Pfeffermann & Sverchkov (2003) propose a test that uses estimating equations to estimate β by an auxiliary regression model for $\vec{W}$ on some function of $\mathbf{X}$ with parameter η. The unweighted estimating function is

$$\delta_k(\beta) = \vec{X}_k(Y_k - \vec{X}_k^\top \beta), \;\; k \in S.$$

Define $\hat{W}_k$ as the fitted value of the regression, $q_k = W_k/\hat{W}_k$, and $R(\vec{X}_k;\beta) = \delta_k(\beta) - q_k\delta_k(\beta)$. Ignorable weights $\vec{W}$ cause $\vec{q}$ to be approximately 1. Thus, the null hypothesis is $H_0 : E(R(\mathbf{X};\beta)) = 0$. The sampling weight means $E(R(\mathbf{X};\beta))$ can be tested by a Hotelling statistic

$$\frac{n-p}{p}\bar{R}_n^{-\top}\hat{\Sigma}_{R,n}^{-1}\bar{R}_n,$$

where $\bar{R}_n$ is the sample mean and $\hat{\Sigma}_{R,n}$ is the sample variance matrix of $R(\mathbf{X};\hat{\beta}_u)$. The statistic approximately follows an F distribution with $(p, n-p)$ degrees of freedom under the null hypothesis (Pfeffermann & Sverchkov, 2003).

Care should be taken for determining the auxiliary estimating function $f(\mathbf{X};\eta)$ to increase the power of the test. With the simplest form being linear regression, more flexible forms can accommodate non-linearity to possibly improve the power if some model building is made. Pfeffermann & Sverchkov (2003) suggest using the score equations if the likelihood is specified.

Steps for performing the Pfeffermann-Sverchkov Estimation Test, given $\{Y_k, \vec{X}_k, W_k\}_{k \in S}$:

1. For the auxiliary regression model of $E\left(\vec{W} \mid \mathbf{X}\right)$, use the design matrix $\mathbf{X}_{\text{design}} = \mathbf{X}$ with a column of ones for the intercept to compute the regression coefficient estimates $\hat{\eta}$. The design matrix may change depending on the auxiliary regression model.
2. Determine $\hat{W}_k$ from the estimates fitted with the auxiliary regression and calculate $q_k = W_k/\hat{W}_k$.
3. Estimate β from regressing $\vec{Y}$ on $\mathbf{X}$ and estimate the fitted $\hat{\vec{Y}}$.
4. Use the unweighted estimation function $\delta_k(\hat{\beta})$ for $k \in S$ to compute $R(\vec{X}_k; \hat{\beta}) = \delta_k(\hat{\beta}) - q_k\delta_k(\hat{\beta})$.
5. Compute test statistic T_H as

$$T_H = \frac{n-p}{p}\bar{R}_n^{-\top}\hat{\Sigma}_{R,n}^{-1}\bar{R}_n.$$

6. Calculate p-value with $T_H \sim F_{p,\,n-p}$.

2.5.2 Pfeffermann-Nathan Predictive Power Test

Pfeffermann & Nathan (1985) propose a test based on predicting the out-of-sample predictive power of weighted and unweighted estimation by a cross-validation approach of splitting the sample set S into an estimation set E and validation set V where $S = E + V$ and $E \cap V = \emptyset$. Weighted and unweighted regressions are fitted with the estimation set E to predict the observations in the validation set V.

Let $v_{u,k}$ and $v_{w,k}$ be the prediction errors of the unweighted and weighted regression fits for the kth observation in the validation set V. Under the null hypothesis of noninformative weighting,

$$H_0 : E(v_{u,k}^2 - v_{w,k}^2) = 0, \;\; k \in V$$

which can be tested by a Z-test of test statistic $Z = \bar{D}/S_D$ where $\bar{D}$ is the sample mean and S_D is the sample standard deviation of $D_k = v_{u,k}^2 - v_{w,k}^2$.

The implementation of the test requires splitting the sample into two smaller sets. Although Pfeffermann & Nathan (1985) do not recommend a split ratio, the conventional split between an estimation set E and validation set V is 80-20. Wang *et al.* (2023) utilize a 50-50 split for their sample split. The prediction errors are conditionally independent of the estimation set E, but not independent since they are calculated based on the same $\hat{\beta}_u$ and $\hat{\beta}_w$ (Wang *et al.*, 2023). Reducing the sample set into smaller sets may significantly reduce the power of the tests.

Steps for performing the Pfeffermann-Nathan Predictive Power Test, given $\{Y_k, \vec{X}_k, W_k\}_{k \in S}$:

1. With the split ratio for the sample S, create the estimation set E and validation set V accordingly.
2. Compute the unweighted linear regression of $E(Y_k \mid \vec{X}_k),\ k \in E$ to obtain $\hat{\beta}_u$. With the regression coefficient estimates, fit the unweighted regression onto the validation set V and compute the prediction errors $v_{u,k} = Y_k - \hat{Y}_k,\ k \in V$.
3. Compute the weighted linear regression of $E(Y_k \mid \vec{X}_k, W_k),\ k \in E$ to obtain $\hat{\beta}_w$. With the estimates of the regression coefficients, fit the weighted regression onto the validation set V and compute the prediction errors $v_{w,k} = Y_k - \hat{Y}_k,\ k \in V$.
4. With $D_k = v_{u,k}^2 - v_{w,k}^2$, compute $\bar{D}$ and S_D. Calculate the test statistic $Z = \bar{D}/S_D$.
5. Compute the two-sided p-value where $Z \sim \mathcal{N}(0,1)$ under the null hypothesis of $E(D) = 0$.

2.5.3 Breidt Likelihood-Ratio Test

Breidt *et al.* (2013) formally propose a likelihood-ratio test from Herndon (2014)'s dissertation that is distinct from other formal diagnostic tests. Assuming a superpopulation model with a finite population U, Breidt *et al.* (2013) proposes a weighted log-likelihood with a general weight vector $\vec{\omega}$ with true parameters $\vec{\theta} = [\vec{\beta}_\omega, \sigma^2]$ as

$$l(\theta; \vec{\omega}) = \sum_{k \in S} \omega_k \log(f(Y_k \mid \vec{X}_k; \theta)).$$

For a weighted log-likelihood estimation, $\vec{\omega}_w = \vec{W}$. For unweighted log-likelihood, $\vec{\omega}_u = N/n$ where N is the size of the finite population $\mathcal{U}$ and n is the size of sample S (Herndon, 2014).

Let $\hat{\theta}_u = \text{argmin}_\theta l(\theta; \vec{\omega}_u)$ and $\hat{\theta}_w = \text{argmin}_\theta l(\theta; \vec{\omega}_w)$. Two LR statistics are considered as

$$T_U = 2(l(\hat{\theta}_u; \vec{\omega}_u) - l(\hat{\theta}_w; \vec{\omega}_u)) \text{ and } T_W = 2(l(\hat{\theta}_w; \vec{\omega}_u) - l(\hat{\theta}_w; \vec{\omega}_w)).$$

Implementing the LR tests require maximizing both weighted and unweighted log-likelihoods.

The maximum likelihood estimates for the unweighted log-likelihood are

$$\vec{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \vec{Y}$$

$$\hat{\sigma}^2 = N^{-1} \sum_{k \in S} (Y_k - \vec{X}_k \hat{\beta})^2,$$

and, according to Lohr (2022), the maximum likelihood estimates for the weighted log-likelihood are

$$\vec{\beta} = \frac{\frac{\sum_{k \in S} W_k Y_k \cdot \sum_{k \in S} W_k \vec{X}_k}{\sum_{k \in S} W_k \vec{X}_k W_k} - \sum_{k \in S} W_k \vec{X}_k Y_k}{\frac{\sum_{k \in S} W_k \vec{X}_k \cdot \sum_{k \in S} W_k \vec{X}_k}{\sum_{k \in S} W_k \vec{X}_k W_k} - \sum_{k \in S} W_k \vec{X}_k^2} = \frac{\sum_{k \in S} W_k \frac{1}{\hat{\sigma}_k^2} \vec{X}_k Y_k}{\sum_{k \in S} W_k \frac{1}{\hat{\sigma}_k^2} \vec{X}_k \vec{X}_k^\top}$$

$$\hat{\sigma}^2 = \frac{\sum_{k\in S} W_k (Y_k - \vec{X}_k \vec{\beta})^2}{\sum_{k\in S} W_k}.$$

Let the information matrices be denoted as $J_u = \sum_{k\in S} \mathcal{I}(\vec{X}_k; \theta_0) = \mathcal{I}(\mathbf{X}; \theta_0)$, $J_w = \sum_{k\in S} W_k \mathcal{I}(\vec{X}_k; \theta_0)$, and $K_w = \sum_{k\in S} W_k^2 \mathcal{I}(\vec{X}_k; \theta_0)$ where $\mathcal{I}(\vec{X}_k; \theta_0)$ is the Fisher information for the kth sample observation with the true parameters $\vec{\theta}_0$.

Under the null hypothesis of noninformative weights

$$\sqrt{n}(\hat{\theta}_w - \hat{\theta}_u) \xrightarrow{\mathcal{L}} \mathcal{N}(0, -J_u^{-1} + J_w^{-1} K_w J_w^{-1}).$$

The asymptotic distribution of T_u is $T_u \xrightarrow{\mathcal{L}} \sum_{j=1}^{q} \lambda_{u,j} Z_j^2$ where λ_u are the eigenvalues of

$$(-J_u^{-1} + J_w^{-1} K_w J_w^{-1})^{T/2} J_u (-J_u^{-1} + J_w^{-1} K_w J_w^{-1})^{1/2}$$

and $Z_j, j = 1, \ldots, p$, are independent standard Normal random variables.

The specifications above are denoted T_u as empirically shown to have larger power in Wang *et al.* (2023) simulations. The limiting distribution is a linear combination of chi-square random variables with coefficients being the eigenvalues of the matrix (Breidt *et al.*, 2013). The test requires a distributional specification on the regression errors where the test may lose power if the distribution is misspecified (Wang *et al.*, 2023).

Steps for performing the Bredit Likelihood Ratio Test for T_u, given $\{Y_k, \vec{X}_k, W_k\}_{k\in S}$:

1. Determine the maximum likelihood estimates $(\vec{\theta}_u, \vec{\theta}_w)$ for the unweighted and weighted log likelihoods for $\hat{\beta}$ and $\hat{\sigma}^2$ where

$$\log L\left(\vec{\beta}, \sigma^2 \mid \vec{Y}, \mathbf{X}, \vec{W}\right) = -\frac{1}{2}\log(2\pi\sigma^2)\sum_{k\in S} W_k - \frac{1}{2\sigma^2}\sum_{k\in S} W_k (Y_k - \vec{X}_k \vec{\beta})^2.$$

2. With maximum likelihood estimates $\vec{\theta}_u = [\hat{\vec{\beta}}_u, \hat{\sigma}_u^2]$ and $\vec{\theta}_w = [\hat{\vec{\beta}}_w, \hat{\sigma}_w^2]$, calculate the log-likelihood of $l(\hat{\theta}_u; \vec{\omega}_u)$ and $l(\hat{\theta}_w; \vec{\omega}_u)$. Compute test statistic $T_u = 2(l(\hat{\theta}_u; \vec{\omega}_u) - l(\hat{\theta}_w; \vec{\omega}_u))$.
3. Calculate the information matrices:

$$J_u = \operatorname{diag}\left(\sum_{k\in S} \frac{\vec{X}_k \vec{X}_k^\top}{\hat{\sigma_u}^2}, \sum_{k\in S} \frac{1}{2n\hat{\sigma_u}^4}\right), \; J_w = \operatorname{diag}\left(\sum_{k\in S} \frac{\vec{X}_k W_k \vec{X}_k^\top}{\hat{\sigma_u}^2}, \sum_{k\in S} \frac{W_k}{2n\hat{\sigma_u}^4}\right)$$

$$K_w = \operatorname{diag}\left(\sum_{k\in S} \frac{\vec{X}_k W_k^2 \vec{X}_k^\top}{\hat{\sigma_u}^2}, \sum_{k\in S} \frac{W_k^2}{2n\hat{\sigma_u}^4}\right).$$

4. Compute eigenvalues $\vec{\lambda}$ of $(-J_u^{-1} + J_w^{-1} K_w J_w^{-1})^{T/2} J_u (-J_u^{-1} + J_w^{-1} K_w J_w^{-1})^{1/2}$.
5. Calculate the linear combination of χ_1^2 scaled by $\vec{\lambda}$ to generate empirical distribution to determine p-value using the test statistic $T_u \xrightarrow{\mathcal{L}} \sum_{j=1}^{q} \lambda_{u,j} Z_j^2$.

3

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

In an attempt to compare the plethora of survey weight diagnostic tests, Wang *et al.* (2023) ran two large simulation studies, each determining the robustness of the tests in various circumstances. This first simulation study aims to reproduce the empirical results from Wang *et al.* (2023) and to determine possible reasons for the difference of results.

Within the simulation studies, eight unique formal diagnostic tests were included. With some tests allowing for specified functions $f(\mathbf{X};\eta)$, some tests include quadratic terms, which are indicated with a "q" to address any possible non-linearity. To align with the notation in Wang *et al.* (2023), the tests were abbreviated as follows:

- **DD**: DuMouchel-Duncan WA Test
- **PN**: Pfeffermann-Nathan Predictive Power Test
- **HP**: Hausman-Pfeffermann DC Test
- **PS1**: Pfeffermann-Sverchkov (1999) WA Test
- **PS1q**: Pfeffermann-Sverchkov (1999) WA Test, with quadratic terms
- **PS2**: Pfeffermann-Sverchkov (2007) WA Test
- **PS2q**: Pfeffermann-Sverchkov (2007) WA Test, with quadratic terms
- **PS3**: Pfeffermann-Sverchkov Estimation Test
- **WF**: Wu-Fuller WA Test
- **LR**: Breidt Likelihood-Ratio Test

3.1 Study 1: Pfeffermann & Sverchkov (1999) Adaptation

Wang *et al.* (2023)'s first study is an adaptation of Pfeffermann & Sverchkov (1999)'s simulation study. A population size of $N = 3000$ was generated for (Y_i, X_i) with 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 $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$ with $\sigma \in \{0.1, 0.2\}$. The sample sizes $n \in \{100, 200\}$ were drawn from the population with the probability proportional to the weight as defined by

$$W_i = \alpha Y_i + 0.3X_i + \delta U_i,$$

where $\alpha \in \{0, 0.2, 0.4, 0.6\}$ 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 $\alpha = 0$ and informative when $\alpha \neq 0$ (Wang *et al.*, 2023).

Simulation Setup — Study 1

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

1. For each generated population unit $i = 1, 2, \ldots, N$:
   (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) Generate $Y_i = 1 + X_i + \varepsilon_i$.
   (c) Generate the weights $W_i = \alpha Y_i + 0.3X_i + \delta U_i$.
2. Using Probability Proportional to Size (PPS), sample n sized sample set S from the population. Subsequently, redefine $W_k = 1/\pi_k$ where π_i are generated from PPS for $k \in S$.
3. Perform all the aforementioned tests on the generated data with sample data $\{Y_k, X_k, W_k\}_{k \in S}$.
4. Record the corresponding p-values.

The simulation has $2 \times 2 \times 2 \times 4 = 32$ case scenarios. With the linear weight-generating function from Pfeffermann & Sverchkov (1999), the cases vary by sample sizes n, noise amplifier δ, noise factor σ, and weight informative factor α. The power of the tests is expected to increase with large sample sizes n, small noise amplifiers δ, large variation factors σ, and large weight informative factors a.

Cases:

1. Sample Size: $n \in \{100, 200\}$
2. Noise Amplifier: $\delta \in \{1, 1.5\}$
3. Variation factor: $\sigma \in \{0.1, 0.2\}$
4. Weight Informativeness: $\alpha \in \{0, 0.2, 0.4, 0.6\}$

Constants:

- Iterations: $B = 1000$
- Population per iteration: $N = 3000$
- Significance level: $\alpha = 0.05$

Table 3.1: *Wang et al. (2023) study 1 empirical rejection rates of ten tests with $\vec{W}$ is linear in $\vec{Y}$ based on 1000 replicates and 32 case scenarios.*

n	σ	δ	α	*DD*	*PN*	*HP*	*PS1*	*PS1q*	*PS2*	*PS2q*	*PS3*	*WF*	*LR*
100	0.1	1.5	0.0	5.9	8.3	5.6	5.2	4.9	5.4	6.0	4.3	5.8	6.2
			0.2	5.9	6.8	5.4	4.6	5.8	5.6	5.4	4.1	5.7	6.9
			0.4	9.6	9.1	9.2	8.8	8.8	11.6	10.6	6.4	9.6	8.6
			0.6	21.2	12.2	21.0	17.4	16.9	27.1	19.8	13.6	21.2	16.5
		1.0	0.0	4.6	9.5	4.5	4.9	4.6	5.9	3.8	4.0	4.7	5.4
			0.2	7.2	8.9	6.9	6.7	6.8	9.0	7.2	5.3	7.4	7.1
			0.4	21.1	11.0	21.1	16.1	18.9	28.6	21.2	14.0	21.2	14.6
			0.6	41.6	12.4	40.7	28.4	34.9	51.2	40.4	28.0	40.6	25.9
	0.2	1.5	0.0	5.7	5.9	5.5	4.9	3.9	5.3	4.9	3.2	5.0	5.1
			0.2	9.6	8.0	9.3	11.2	10.1	13.3	10.5	7.7	10.0	10.3
			0.4	31.5	11.5	30.9	33.7	27.5	41.6	31.1	19.8	31.3	24.8
			0.6	64.7	16.1	63.9	65.9	58.0	75.3	64.4	47.1	63.9	48.9
		1.0	0.0	6.0	8.1	5.8	4.1	5.1	4.6	5.9	4.7	6.2	5.8
			0.2	16.4	9.5	16.2	17.3	14.8	23.2	16.4	9.9	16.4	12.8
			0.4	63.3	15.8	62.9	59.0	55.1	73.3	62.6	44.4	62.7	46.1
			0.6	94.6	25.5	94.3	90.2	92.0	97.6	94.2	85.8	94.1	81.7
200	0.1	1.5	0.0	4.5	7.3	4.4	3.9	4.3	4.2	4.0	4.5	4.1	4.8
			0.2	9.0	8.4	8.9	8.1	8.9	9.9	9.0	8.4	9.6	8.6
			0.4	17.8	11.4	17.6	17.7	14.8	22.0	16.7	13.0	17.9	14.4
			0.6	39.6	12.4	39.4	36.6	33.4	48.1	38.8	28.5	38.9	28.0
		1.0	0.0	4.8	7.2	4.7	3.2	4.5	4.3	4.5	4.7	5.1	5.5
			0.2	10.5	10.8	10.4	9.8	11.9	14.5	11.3	9.2	11.8	9.6
			0.4	36.1	14.6	35.6	29.4	31.4	46.2	36.0	27.2	35.7	23.9
			0.6	70.4	19.5	70.1	58.4	64.2	80.5	71.2	57.1	70.8	47.3
	0.2	1.5	0.0	4.4	8.3	4.3	4.5	4.5	4.7	4.7	4.5	4.5	5.0
			0.2	18.4	10.2	18.0	19.6	15.6	21.5	18.7	14.1	18.0	15.8
			0.4	57.4	14.7	57.1	61.2	50.0	67.8	57.1	45.7	56.7	47.4
			0.6	91.7	25.2	91.5	91.8	89.0	96.1	92.1	86.3	91.8	83.1
		1.0	0.0	4.4	8.3	4.4	3.2	4.3	4.4	4.2	5.5	4.7	4.2
			0.2	35.0	13.9	34.8	35.4	31.3	44.2	34.9	26.9	35.0	27.5
			0.4	92.2	26.6	92.0	92.1	87.2	96.4	91.7	85.7	91.8	81.1
			0.6	100.0	49.6	100.0	99.8	99.9	100.0	100.0	99.7	100.0	98.8

Note: Rejection rates were determined at the $\alpha = 0.05$ significance level where rates are the percentage of tests rejecting the null hypothesis of noninformative weights.

Table 3.2: *Replication of Wang et al. (2023) study 1 empirical rejection rates of ten tests with $\vec{W}$ is linear in $\vec{Y}$ based on 1000 replicates and 32 case scenarios.*

n	σ	δ	α	DD	PN	HP	PS1	PS1q	PS2	PS2q	PS3	WF	LR
100	0.1	1.5	0.0	4.6	38.4	4.1	7.1	7.2	4.6	6.1	3.8	3.6	51.5
			0.2	5.2	33.4	5.0	9.0	9.2	9.7	9.1	5.2	6.0	49.7
			0.4	10.3	34.4	10.0	11.9	13.3	15.2	13.6	9.7	11.6	52.5
			0.6	19.3	34.7	18.7	16.5	19.6	26.0	21.2	22.3	23.0	52.9
		1.0	0.0	5.3	33.4	5.1	7.5	6.7	6.2	7.5	4.6	5.1	52.3
			0.2	7.4	35.8	7.2	10.8	11.3	12.2	12.0	7.1	7.6	51.8
			0.4	18.0	33.9	17.6	17.4	22.8	26.9	20.4	19.2	21.2	49.6
			0.6	35.3	33.3	34.5	29.4	40.0	47.0	35.6	37.1	39.9	52.5
	0.2	1.5	0.0	4.7	34.7	4.2	6.4	6.6	4.4	4.5	3.6	5.3	48.9
			0.2	9.7	35.5	9.5	10.7	11.7	13.3	11.6	9.5	12.1	52.3
			0.4	28.0	33.6	27.2	24.1	23.9	29.7	27.7	29.1	32.4	47.5
			0.6	55.6	35.6	54.4	48.2	47.6	55.7	54.7	57.0	61.9	51.2
		1.0	0.0	5.0	35.7	4.6	6.1	8.5	6.0	7.5	4.1	4.0	50.0
			0.2	19.3	35.9	18.8	17.4	18.8	21.6	20.0	18.2	21.5	51.7
			0.4	58.0	36.2	56.7	48.1	49.2	58.2	54.4	60.2	62.3	53.4
			0.6	92.4	33.7	92.1	84.4	87.7	90.6	88.4	92.2	94.2	53.4
200	0.1	1.5	0.0	5.1	37.3	4.8	7.9	7.8	5.2	7.2	3.7	3.9	43.2
			0.2	6.3	33.0	5.9	9.3	10.6	9.8	9.3	8.3	9.3	45.7
			0.4	15.9	34.6	15.7	16.3	18.5	22.0	16.8	18.5	18.4	47.4
			0.6	34.4	34.3	34.1	31.7	36.6	41.7	35.2	37.0	38.6	46.5
		1.0	0.0	5.0	34.4	4.9	7.2	8.2	7.0	7.9	3.8	3.9	47.6
			0.2	10.3	34.5	9.9	13.3	17.2	17.8	13.8	11.4	12.7	47.9
			0.4	35.0	34.6	34.7	28.7	38.9	46.0	32.8	37.1	40.3	48.3
			0.6	70.0	32.4	69.7	58.6	69.9	77.7	64.8	70.1	73.3	47.0
	0.2	1.5	0.0	4.2	35.7	3.9	6.7	6.9	5.3	6.2	4.2	4.5	47.0
			0.2	14.3	33.5	14.1	13.5	15.4	17.5	15.7	15.4	16.8	48.3
			0.4	54.9	33.7	54.0	46.4	46.1	54.6	51.9	56.4	58.1	47.3
			0.6	91.1	36.8	91.1	83.0	82.6	88.0	86.3	91.3	92.7	49.8
		1.0	0.0	3.8	35.7	3.4	6.8	7.9	6.7	6.2	4.1	3.9	45.5
			0.2	33.3	31.8	32.3	26.2	29.2	33.4	29.8	35.8	38.0	48.7
			0.4	91.2	35.4	90.9	80.5	83.4	88.0	85.6	90.9	93.2	48.6
			0.6	100.0	35.8	100.0	99.5	99.5	99.8	99.8	99.9	99.9	46.2

Note: Rejection rates were determined at the $\alpha = 0.05$ significance level where rates are the percentage of tests rejecting the null hypothesis of noninformative weights.

Results

Table 3.1 and Table 3.2 are the empirical rejection rates of the ten tests under the $\vec{W}$ linear generating function with $\vec{Y}$ of Wang *et al.* (2023) and the replication attempt, respectively. For a well-performing test, the rejection rates should scale from 5.0 to 100.0 steadily as the weight informativeness α increases. As noted in Wang *et al.* (2023) and in the replication simulation, PN is repeatedly above the nominal 5.0 size which is believed to be caused by the dependence of the prediction errors on the estimates of similar coefficients. Since PN has much less variability and power than other tests — likely due to the division of the sample into estimation sets E and validation sets V — PN will be excluded from future test power comparisons.

As anticipated, larger values of α and n translate into power of the tests increasing. Also, holding all other variables constant, larger δ values increase noise in the weight models which hinders the tests' ability to determine weight informativeness. Surprisingly, σ leads to higher rejection rates as σ adds more variation on $\vec{Y}$, possibly by increasing the signal-to-noise ratio (Wang *et al.*, 2023).

With the replication simulation study in Table 3.2, PS2 and DD performed the best in rejecting the null hypothesis of noninformative weights as α and n increased with each test performing better than each other periodically. This contrasts with Wang *et al.* (2023) since their results suggested that PS2 performed the best in all cases with DD trailing slightly behind. PS1q has more power than PS1 when $\sigma = 0.1$ but are similar when $\sigma = 0.2$ which departs from Wang *et al.* (2023) that has PS1q performing worse than PS1. In contrast, PS2q is a bit less powerful than PS2. Noticeably, DD and HP perform almost identical in all 32 cases. PS1 is the least powerful test among the 10 tests.

3.2 Study 2: Quadratic Weight Generating Function

Wang *et al.* (2023) were also interested in the performance of diagnostic tests when weights are generated from a quadratic function of $\mathbf{X}$ and $\vec{Y}$ and thus proposed an alteration to Study 1 by the following weight generation model:

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

where $U_i \overset{iid}{\sim} \text{Unif}(0,1)$ and $\alpha \in \{0, 0.5, 1.0, 1.5\}$. The quadratic function was designed with characteristics similar to the linear weight generation function with the additional characteristic that for $\alpha = 1$, the partial correlation between W_i and Y_i is zero. Wang *et al.* (2023) claim that this makes it difficult for diagnostic tests based on linear regression to determine the importance of W_i on Y_i.

Additionally, the finite sample performance of the tests may depend on the distribution of the regression errors. To test this, Wang *et al.* (2023) considered four distributions of ε_i: Gamma, Normal, Uniform, and Student-t. The distribution parameters were selected — and scaled as necessary — to have $E(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$. Although this simulation study is not replicated here, Wang *et al.* (2023) showed that nearly all tests were robust to the regression error distribution, excluding the LR test, which fails under the heavily right-skewed Student-t distribution. Under the null hypothesis, the tests' distributions are asymptotically correctly specified such that the error distribution is inconsequential.

Simulation Setup — Study 2

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

1. For each generated population unit $i = 1, 2, \ldots, N$:
   (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) Generate $Y_i = 1 + X_i + \varepsilon_i$.
   (c) Generate the weights $W_i = \alpha(Y_i - 1.5\alpha)^2 + 0.3X_i - 0.3X_i^2 + \delta U_i$.
2. Using Probability Proportional to Size (PPS), sample n sized sample set S from the population. Subsequently, redefine $W_k = 1/\pi_k$ where π_i are generated from PPS for $k \in S$.
3. Perform all the aforementioned tests on the generated data with sample data $\{Y_k, X_k, W_k\}_{k \in S}$.
4. Record the corresponding p-values.

The simulation has $2 \times 4 = 8$ case scenarios. With the quadratic weight-generating function from Pfeffermann & Sverchkov (1999), the cases vary by sample size n and weight informative factor α. The power of the tests is expected to increase with large sample sizes n, small noise amplifiers δ, large variation factors σ, and large weight informative factors a. Weights W_k are expected to be noninformative in Y_k when $\alpha = 0$. For $\alpha = 1$, partial correlation between W_k and Y_k is zero, which can cause diagnostic tests with linear auxiliary regressions to have issues with power.

Cases:

1. Sample Size: $n \in \{100, 200\}$
2. Weight Informativeness: $\alpha \in \{0, 0.2, 0.4, 0.6\}$

Constants:

- Iterations: $B = 1000$
- Population per iteration: $N = 3000$
- $\sigma = 0.1$
- Significance level: $\alpha = 0.05$

Table 3.3: *Wang et al. (2023) study 2 empirical rejection rates of ten tests with $\vec{W}$ is quadratic in $\vec{Y}$ based on 1000 replicates and 8 case scenarios.*

n	α	*DD*	*PN*	*HP*	*PS1*	*PS1q*	*PS2*	*PS2q*	*PS3*	*WF*	*LR*
100	0.0	7.8	7.1	7.5	6.1	6.4	6.0	6.3	6.1	7.6	7.6
	0.5	69.5	15.2	69.0	60.9	66.0	77.0	72.5	53.0	70.8	43.5
	1.0	33.9	8.2	33.5	7.7	35.7	7.7	40.2	17.4	33.4	29.4
	1.5	100.0	77.1	100.0	99.8	100.0	100.0	100.0	100.0	100.0	98.1
200	0.0	4.7	10.5	4.7	5.0	5.1	5.0	5.1	4.5	4.9	5.6
	0.5	94.0	27.2	93.8	91.2	93.5	96.6	95.9	90.7	95.2	79.8
	1.0	66.7	6.5	66.4	6.9	66.0	6.9	72.5	50.1	66.6	58.9
	1.5	100.0	97.3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Note: Rejection rates were determined at the $\alpha = 0.05$ significance level where rates are the percentage of tests rejecting the null hypothesis of noninformative weights.

Table 3.4: *Replication of Wang et al. (2023) study 2 empirical rejection rates of ten tests with $\vec{W}$ is quadratic in $\vec{Y}$ based on 1000 replicates and 8 case scenarios.*

n	α	*DD*	*PN*	*HP*	*PS1*	*PS1q*	*PS2*	*PS2q*	*PS3*	*WF*	*LR*
100	0.0	4.4	38.3	4.4	7.3	9.5	8.3	7.1	3.3	5.2	50.4
	0.5	54.8	36.4	53.3	28.3	29.6	34.9	32.1	65.6	70.9	56.9
	1.0	18.0	35.9	17.3	11.7	16.2	25.8	12.2	5.7	8.1	62.5
	1.5	100.0	36.7	100.0	86.2	98.9	98.6	92.5	86.7	92.7	56.2
200	0.0	5.2	37.1	4.9	5.9	8.2	7.2	5.5	4.1	4.2	42.2
	0.5	86.7	36.1	86.3	47.1	53.2	60.8	55.0	94.6	95.2	55.9
	1.0	39.1	37.8	38.6	22.7	43.5	61.9	30.6	10.8	14.8	63.1
	1.5	100.0	40.8	100.0	98.4	100.0	100.0	99.5	98.4	99.7	49.5

Note: Rejection rates were determined at the $\alpha = 0.05$ significance level where rates are the percentage of tests rejecting the null hypothesis of noninformative weights.

Results

Table 3.2 and Table 3.2 are the empirical rejection rates of the ten tests under the $\vec{W}$ quadratic generating function with $\vec{Y}$ of Wang *et al*. (2023) and the replication attempt, respectively. For a well-performing test, it should scale from 5.0 to 100.0 steadily as the weight informativeness α increases from 0 to 0.5 and 1 to 1.5.

As anticipated, values of $\alpha = 0.5, 1.5$ and n translate into higher test power. With the replication simulation study in Table 3.2, not all tests necessarily hold their power of 5.0 when $\alpha = 0$ in contrast to Wang *et al*. (2023) as PS1q and PS2 depart significantly from 5.0. Likely the most important difference is probably the rejection rates between the tests when $\alpha = 0.5, 1.0$. In Table 3.2, Wang *et al*. (2023) shows a significant drop in rejection rates from $\alpha = 0.5$ to $\alpha = 1.0$, while the replication in Table 3.2 shows a smaller drop in the rejection rates. This is mainly due to the smaller magnitudes of rejection rates for $\alpha = 0.5$.

With regards to tests' performances, PS3 and WF performed well except when $\alpha = 1.0$ while DD generally performed the best. This also contrasts with the results from Wang *et al*. (2023) that show that the modified tests PS1q and PS2q turn out to be the most powerful. In the replication results, PS1q is consistently more powerful than PS1 while PS2q is significantly less powerful than PS2.

3.3 Study 3: Wu & Fuller (2005) Adaptation

Wang *et al*. (2023) last simulation study adapts Wu & Fuller (2005)'s simulation study of their proposed test by exploring the robustness of nonlinear weight associations by generating selection probabilities for the ith population unit. Population data (Y_i, X_i) were generated from a linear regression model

$$Y_i = 0.5 + X_i + \varepsilon_i, \ i = 1, \dots, N,$$

where $X_i, \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, 0.5)$. W_i, initially defined as the selection probability for the population unit i, is generated by

$$W_i = \alpha \cdot \eta(X_i) + \beta \cdot \eta(\psi \cdot \varepsilon_i + (1 - \psi) \cdot Z_i)$$

with scalars (α, β, ψ) are scalars, $\alpha + \beta = 2$, and $Z_i \overset{iid}{\sim} \mathcal{N}(0, 0.5)$. The function $\eta(x)$ is constructed to have a monotonically increasing W_i for an increase in X_i and to ensure $W_i \in (0, 1]$:

$$\eta(x) = \begin{cases} 0.025, & x < 0.2 \\ 0.475(x - 0.2) + 0.025, & 0.2 \le x \le 1.2 \\ 0.5, & 1.2 < x. \end{cases}$$

Wang *et al*. (2023) claim that the expectation of W_i is 0.221. However, $E(W_i)$ is a function of the scalars (α, β, ψ) and the random variables $(X_i, Z_i, \varepsilon_i)$. The derivation of $E(W_i)$

is denoted in Appendix B and shows how $E(W_i)$ changes between the cases set-up by Wang *et al*. (2023). For example, when $\psi = 0.0$ and $\alpha = 1.0$, $E(W_i) = 0.221$ while if $\psi = 0.3$ and $\alpha = 0.25$, $E(W_i) = 0.177$.

When adapting the simulation study from Wu & Fuller (2005), Wang *et al*. (2023) used Poisson sampling such that for all $i \in N$, a population unit i was selected if $U_i < W_i$ where $U_i \overset{iid}{\sim} \text{Unif}(0, 1)$ (Lohr, 2022). Given that the sampling of a unit i is random conditional on its selection probability, the size of the sample set S is random. Wang *et al*. (2023) selected their desired sample size by sampling if $U_i < W_i$ until they got their desired sample size. This departs from Wu & Fuller (2005) since their simulation design aimed to select an expected sample size of 250. For this replication, the sample was set to have the expected value of the fixed sample sizes of Wang *et al*. (2023).

Simulation Setup — Study 3

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

1. For each generated population unit $i = 1, 2, \ldots, N$:
   (a) Sample $X_i, Z_i, \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, 0.5)$.
   (b) Generate $Y_i = 0.5 + X_i + \varepsilon_i$.
   (c) Generate the inclusion probabilities

$$W_i = \alpha \cdot \eta(X_i) + \beta \cdot \eta(\psi \cdot \varepsilon_i + (1 - \psi) \cdot Z_i),$$

$$\text{with } \eta(x) = \begin{cases} 0.025, & x < 0.2 \\ 0.475(x - 0.2) + 0.025, & 0.2 \le x \le 1.2 \\ 0.5, & 1.2 < x. \end{cases}$$

2. Using Poisson sampling of given $\vec{W}$, draw $U_i \overset{iid}{\sim} \text{Unif}(0, b)$ and select population unit i if $U_i < W_i$. To obtain an expected value of the desired sample size n, set $b = n^{-1} \sum_i^N W_i$. See Appendix B for explanation. Subsequently, redefine W_i to be the inverse of the selection probabilities where $W_i \to \frac{1}{W_i}$.
3. Perform all the aforementioned tests on the generated data with sample data $\{Y_k, X_k, W_k\}_{k \in S}$.
4. Record the corresponding p-values.

Cases:

1. Sample Size: $E(n) \in \{100, 200\}$
2. Correlation factor: $\alpha \in \{0.25, 0.5, 0.75, 1\}$
3. Weight Informativeness: $\psi \in \{0, 0.2, 0.4, 0.6\}$

Constants:

- Iterations: $B = 1000$
- Population per iteration: $N = 3000$
- $\sigma^2 = 0.5$
- Significance level: $\alpha = 0.05$

The simulation has $2 \times 4 \times 4 = 32$ case scenarios. With the selection probability function W_i from Wu & Fuller (2005), the cases vary by expected sample size $E(n)$, weight informative factor ψ, and correlation factor α. As α increases, the correlation between W_i and X_i increases, while the correlation between W_i and ε_i decreases. Lastly, a higher ψ implies more informativeness of W_i on Y_i (Wang *et al.*, 2023).

Results

Table 3.5 and Table 3.6 are the empirical rejection rates of the ten tests with the adapted simulation design of Wu & Fuller (2005) from Wang *et al.* (2023) and the replication attempt, respectively. For a well-performing test, rejection rates should increase from 5.0 to 100.0 steadily as weight informativeness ψ increases and sample size n increases.

As anticipated, the powers of the tests increased as ψ increased, but, concerningly, not all tests held their power of approximately 5.0 when $\psi = 0$ for the significance level of 0.05. As shown in Table 3.6, PN, PS1q, and LR failed consistently to maintain their power when $\psi = 0$. As shown in Wang *et al.* (2023) results in Table 3.5, rejection rates increased as α decreased. However, the replication results indicate that DD, PS2, and PS2q performed the best while Wang *et al.* (2023) depicted ambiguity in the tests' performance. Other differences between the replication and Wang *et al.* (2023) results will be addressed hereafter.

3.4 Review

The contrasting results between the replication attempts and the simulation studies in Wang *et al.* (2023) are significantly different where as the differences cannot be explained by the randomness of the data generation process. While Wang *et al.* (2023) provided a general framework for their simulation studies, it is possible that some details were not clearly conveyed. With no ability to compare simulation code, note that the following theories are speculations on how the differences of the studies' results were caused.

Table 3.5: *Wang et al. (2023) study 3 empirical rejection rates of ten tests based on 1000 replicates and 32 case scenarios.*

n	α	ψ	*DD*	*PN*	*HP*	*PS1*	*PS1q*	*PS2*	*PS2q*	*PS3*	*WF*	*LR*
100	1.00	0.0	4.3	6.7	4.2	1.5	4.6	4.3	5.0	3.6	4.2	5.5
		0.1	11.1	9.4	10.9	5.6	10.6	11.4	12.0	6.4	10.0	7.9
		0.2	33.1	10.5	33.1	14.7	34.8	31.4	38.0	15.2	24.2	22.6
		0.3	66.7	10.7	66.5	25.9	66.0	51.9	70.2	26.1	42.1	38.3
	0.75	0.0	5.5	7.3	5.3	3.7	4.8	4.7	4.6	5.6	5.4	5.8
		0.1	13.0	8.8	12.8	12.1	11.8	15.5	12.5	10.9	11.9	11.1
		0.2	36.7	11.3	36.1	34.9	35.4	42.2	40.9	23.0	33.3	27.6
		0.3	78.9	16.7	78.8	66.1	76.4	76.6	83.2	48.2	66.7	64.5
	0.50	0.0	6.4	6.7	6.2	4.4	5.1	4.5	4.1	6.1	5.6	6.0
		0.1	14.5	9.0	14.3	16.7	12.1	17.5	14.2	10.7	14.1	12.7
		0.2	45.4	12.6	45.1	54.8	42.7	56.9	46.4	36.4	45.4	37.2
		0.3	86.4	22.0	86.2	90.3	82.0	91.2	87.8	72.7	85.5	75.9
	0.25	0.0	4.5	7.2	4.4	6.1	5.0	6.2	5.4	6.9	4.2	4.8
		0.1	13.2	8.8	13.1	17.5	11.9	17.8	13.9	11.8	13.6	10.8
		0.2	50.6	15.7	50.3	60.1	42.6	60.8	48.3	42.7	51.0	41.1
		0.3	91.0	24.6	90.8	94.1	85.9	94.2	90.5	83.0	91.0	82.6
200	1.00	0.0	5.0	6.3	4.7	2.4	5.4	5.8	5.1	3.5	4.4	5.9
		0.1	16.8	9.7	16.7	9.0	15.6	19.6	19.5	10.9	14.6	12.3
		0.2	61.7	14.0	61.5	31.4	61.2	51.7	66.4	31.2	42.2	39.1
		0.3	93.7	18.9	93.6	56.1	94.2	81.6	96.3	58.8	73.5	70.6
	0.75	0.0	4.8	7.3	4.8	3.8	5.1	4.6	4.1	7.2	5.9	5.4
		0.1	19.4	9.6	19.0	20.1	18.4	24.9	20.8	18.2	18.1	15.6
		0.2	68.4	17.5	68.3	66.5	64.0	72.7	71.0	53.4	63.6	57.0
		0.3	98.1	29.4	98.1	95.1	97.8	97.8	98.6	88.3	95.2	91.3
	0.50	0.0	6.3	8.3	6.2	5.3	4.4	5.4	5.0	6.3	6.1	6.7
		0.1	23.8	12.6	23.7	30.4	19.9	31.2	24.0	21.0	24.1	19.3
		0.2	76.8	22.1	76.8	84.0	72.1	85.0	78.3	69.8	75.4	69.2
		0.3	99.3	37.4	99.3	99.5	98.6	99.6	99.4	98.0	98.9	97.6
	0.25	0.0	4.7	7.3	4.6	6.6	5.1	6.4	5.3	7.1	5.1	5.8
		0.1	25.9	10.4	25.7	35.4	22.7	35.0	26.8	26.1	26.3	20.5
		0.2	83.3	21.6	82.9	89.8	77.7	90.0	82.6	77.1	83.1	75.7
		0.3	99.4	44.4	99.4	99.6	99.2	99.5	99.4	98.9	99.4	99.1

Note: Rejection rates were determined at the $\alpha = 0.05$ significance level where rates are the percentage of tests rejecting the null hypothesis of noninformative weights.

Table 3.6: *Replication of Wang et al. (2023) study 3 empirical rejection rates of ten tests based on 1000 replicates and 32 case scenarios.*

$E(n)$	α	ψ	*DD*	*PN*	*HP*	*PS1*	*PS1q*	*PS2*	*PS2q*	*PS3*	*WF*	*LR*
100	1.00	0.0	3.6	35.0	3.4	10.3	12.2	8.0	7.5	2.0	4.1	2.4
		0.1	9.2	40.0	8.7	17.1	21.0	15.1	16.5	2.2	4.6	3.1
		0.2	24.6	39.9	23.8	35.3	38.1	31.3	35.0	1.5	4.2	3.1
		0.3	57.1	41.0	55.7	67.3	69.5	62.8	68.2	1.9	4.6	7.5
	0.75	0.0	4.9	37.8	4.7	7.1	9.2	8.7	8.6	2.7	4.8	2.0
		0.1	10.1	38.2	9.5	13.6	16.4	11.0	13.9	2.5	5.8	3.2
		0.2	28.4	41.8	27.1	33.2	37.7	30.0	37.0	3.2	5.4	6.1
		0.3	70.5	43.6	69.6	69.9	72.9	68.0	74.2	3.4	5.7	11.3
	0.50	0.0	4.1	39.3	3.7	4.7	6.1	4.9	4.9	2.6	4.9	1.3
		0.1	8.9	39.3	8.5	9.3	11.2	8.9	11.4	3.1	5.0	3.2
		0.2	36.4	41.3	34.5	32.7	35.7	33.7	39.1	4.5	4.9	6.1
		0.3	80.3	46.6	79.8	74.8	75.6	76.3	80.4	4.1	4.5	14.5
	0.25	0.0	4.1	41.2	3.9	4.5	4.5	4.8	4.5	4.6	4.3	2.4
		0.1	13.3	39.3	12.8	11.1	10.6	11.9	12.8	4.2	6.1	4.0
		0.2	42.2	46.0	40.8	33.1	32.1	38.2	40.1	5.6	6.5	7.8
		0.3	87.9	50.9	87.3	79.2	77.8	84.6	84.1	3.6	5.2	17.1
200	1.00	0.0	5.7	37.5	5.5	9.3	15.4	9.6	9.8	2.3	4.1	5.4
		0.1	12.3	38.7	12.2	19.0	24.6	15.2	19.3	1.9	4.2	8.6
		0.2	44.5	43.6	43.8	54.3	59.9	48.4	55.9	2.0	3.7	10.9
		0.3	87.5	45.8	87.2	92.5	93.7	89.5	93.3	1.6	5.1	19.6
	0.75	0.0	6.1	38.2	6.1	8.6	15.3	10.5	8.3	3.9	4.2	6.0
		0.1	16.7	40.5	16.3	19.9	28.7	16.5	21.5	3.8	4.9	9.1
		0.2	58.8	43.4	58.1	59.2	65.3	55.7	63.4	1.9	4.4	15.6
		0.3	96.5	53.7	96.3	95.7	96.8	95.5	96.9	2.5	4.1	26.2
	0.50	0.0	6.0	41.4	6.0	7.8	11.3	8.2	7.5	5.0	5.1	7.5
		0.1	17.4	39.4	17.0	18.1	22.9	17.5	20.6	3.1	4.2	9.4
		0.2	64.9	46.1	64.2	61.5	66.4	62.4	67.5	4.0	6.7	15.7
		0.3	98.9	58.6	98.9	98.0	98.0	98.2	99.0	3.9	6.0	32.8
	0.25	0.0	4.6	40.0	4.6	4.6	5.4	4.6	4.1	4.6	4.8	7.2
		0.1	15.3	40.4	15.2	14.0	13.9	15.6	15.2	5.0	6.0	8.7
		0.2	71.4	50.5	70.9	62.5	61.1	67.9	68.6	4.0	5.4	16.6
		0.3	99.3	61.3	99.3	98.3	98.1	99.2	99.4	3.6	5.5	34.0

Note: Rejection rates were determined at the $\alpha = 0.05$ significance level where rates are the percentage of tests rejecting the null hypothesis of noninformative weights.

Weights and Inclusion Probabilities

By definition, survey weights $\vec{W}$ are generally defined as the inverse selection probabilities $\vec{\pi}$ such that $\vec{W} = \frac{1}{\vec{\pi}}$. Within the replications, the generated weights — unless otherwise specified — were interpreted as the inverse selection probabilities that were computed with the generation process. Within the simulation procedures in Wang *et al.* (2023), $\vec{W}$ was not defined — ex-ante or ex-post sampling — as the inverse of the selection probabilities. To see whether the replication results match the results in Wang *et al.* (2023), studies 1, 2, and 3 were performed again without the presumption of weights being the inverse of the inclusion probabilities. Refer to Appendix C for the replication rejection rates of studies 1, 2, and 3 without the presumption that $\vec{W} = 1/\vec{\pi}$.

- Study 1: Pfeffermann & Sverchkov (1999) Adaptation: Weights $\vec{W}$ were interpreted to be the vector of generated data to be used to compute the inclusion probabilities $\vec{\pi}$ of the PPS procedure. The replication simulation design assumed that Wang *et al.* (2023) redefined W_k as $W_k = 1/\pi_k$ for all k elements in the sample. Comparing the Wang *et al.* (2023) results in Table 3.1 and replication results without assuming $\vec{W} = 1/\vec{\pi}$ in Table C, it appears that the results are nearly identical.
- Study 2: Quadratic Weight Generating Function: Like Study 1, weights $\vec{W}$ were interpreted to be the vector of generated data to be used to compute the inclusion probabilities $\vec{\pi}$ of the PPS procedure. The replication simulation design assumed that Wang *et al.* (2023) redefined W_k as $W_k = 1/\pi_k$ for all k elements in the sample. Comparing the Wang *et al.* (2023) results in Table 3.2 and replication results without assuming $\vec{W} = 1/\vec{\pi}$ in Table C, it appears that the results are similar with the exception of PS1 and PS2 still having substantial power compared to Wang *et al.* (2023), in addition to LR and PN tests are insensitive to the cases.
- Study 3: Wu & Fuller (2005) Adaptation: Weights $\vec{W}$ for the population served as the inclusion probabilities of selecting the ith population unit for the sample. For the sampling procedure, Wang *et al.* (2023) utilized the Poisson sampling procedure, where the population unit i will be selected if $U_i < W_i$ where $U_i \overset{iid}{\sim} \text{Unif}(0, 1)$ and stated to stop sampling when the desired sample size n was obtained. Getting a predetermined sample size for Poisson sampling is difficult without causing some dependence of a population unit being selected with others. The replication simulation design sought to instead strive to obtain the sample sizes as its expected value. This was done by setting $U_i \overset{iid}{\sim} \text{Unif}(0, n^{-1}\sum_i^N W_i)$. After selecting K units for the sample S , the weights were redefined to be $W_k \to W_k^{-1}$. Comparing the Wang *et al.* (2023) results in Table 3.5 and replication results without assuming $\vec{W} = 1/\vec{\pi}$ in Table C, it appears that the results are quite different. This could be explained as weights being reciprocal of themselves, which likely shows a similar degree of informativeness within the tests.

Limited Iterations

For all three studies, Wang *et al.* (2023) set the simulated iterations $B = 1000$. While B may be high enough to determine performance within and between diagnostic tests, the difference between the replicated results and Wang *et al.* (2023) results could be determined by the randomness of the data-generating functions that is not trival. As $B \to \infty$, the simulated rejection rates should define the true properties of the diagnostic tests given the simulation design. To determine the convergence rejection rates, B was increased to 10000. Refer to Appendix D for the replication rejection rates of studies 1, 2, and 3 when $B = 10000$. There are no significant differences between the replication results when $B = 1000$ and when $B = 10000$.

4

Simulation Study 2: CE Sampling

In contrast to simulation studies in Wang *et al.* (2023), it is necessary to test the weight diagnostic of the survey on complex survey data to legitimize the empirical utility of the tests. As such, 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 (Toth, 2021). 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 on methods and weighting (U.S. Bureau of Labor Statistics, 2023).

Performing simulations on existing survey data has the advantage of testing diagnostic tests on complex survey designs. Replicating complex survey designs is difficult with generated data with multi-level factors like primary sampling levels (psus) and secondary sampling levels (ssus). For the CE data, it contains 68,415 observations on 47 variables with respect to sample design, location, housing and transportation, family, earner characteristics, labor status, income, assets, and expenditure information. In CE data, observation unit weights are not necessary the inverse of the selection probability, since the Bureau of Labor Statistics adjusts the base weights with calibration methods to adjust for nonresponse and known population characteristics to account for frame undercoverage (King *et al.*, 2021).

Suppose a researcher wanted to predict an individual's income before taxes based on their total expenditures and wanted to utilize the consumer expenditure data to model the relationship. Within the data, the researcher has access to consumer characteristics, expenditure information, income and personal taxes, and other financial information, as shown in Table 4.1. As the researcher knows about CE's complex survey design, the researcher would like to determine whether they should incorporate survey weights within their regression analysis. With the results in Table 4.2 of the aforementioned survey weight diagnostic tests (excluding PN and LR), the researcher has sufficient evidence to necessitate survey weights in their regression analyzes.

Table 4.1: *Variable descrptions for* `rpms`*' 2015 Consumer Expenditure dataset (Toth, 2021).*

Variable	Description
NEWID	The consumer unit (CU) identifying variable, constructed using the first seven digits of NEWID as derived by BLS.
CID	Cluster Identifier for all clusters created using PSU, REGION, STATE, and POPSIZE.
FINLWT21	BLS final sample weight to make inference to total population.
STATE	State FIPS code.
REGION	Region code: 1 Northeast; 2 Midwest; 3 South; 4 West.
BLS_URBN	Urban: 1; Rural: 2.
POPSIZE	Population size class of PSU: 1-biggest through 5-smallest.
CUTENURE	Housing tenure classifications.
ROOMSQ	Number of rooms, including finished living areas and excluding all baths.
BATHRMQ	Number of bathrooms.
BEDROOMQ	Number of bedrooms.
VEHQ	Number of owned vehicles.
FAM_TYPE	CU code based on relationship of members to the interviewed reference person.
FAM_SIZE	Number of members in CU.
PERSLT18	Number of people younger than 18 years old.
PERSOT64	Number of people older than 64 years old.
NO_EARNR	Number of earners.
AGE	Age of primary earner in CU.
EDUCA	Coded education level spanning from none to advanced degree.
SEX	Gender code of F for female and M for male.
MARITAL	Martial status code for primary earner.
MEMBRACE	Race code of primary earner.
HORIGIN	Coded Y or N for whether primary earner is hispanic, latino, or of spanish origin.
ARM_FORC	Coded Y or N for whether primary earner is a member of the armed forces.
IN_COLL	Coded for whether primary earner is enrolled in college.
EARNTYPE	Code for primary earners' worker status.
OCCUCODE	Occupation code for primary earner.
INCOMEY	Type of employment with regard to the institution.
FINCBTAX	Amount of CU income before taxes in past 12 months.
SALARYX	Amount of wage or salary income received in past 12 months before deductions.
SOCRRX	Amount of income received from Social Security and Railroad Retirement in past 12 months.
TOTEXPCQ	Total expenditures for current quarter.
EHOUSNGC	Total expenditures for housing paid during current quarter.
HEALTHCQ	Total expenditures on health care during current quarter.
FOODCQ	Total expenditures on food during current quarter.

For more information on the variables' characteristics and definitions, see U.S. Bureau of Labor Statistics (2015). Table 4.1 only contains a portion of `rpms` dataset varibles. See Appendix E for justifications regarding transformations and data wrangling decisions.

Table 4.2: *Survey Weight Diagnostic Test p-value results on Consumer Expenditure Data*

	DD	*HP*	*PS1*	*PS1q*	*PS2*	*PS2q*	*PS3*	*WF*
p-values	0.03403	0.03404	0.03617	0.07303	0.04080	0.04296	0.01182	0.01047

Diagnostic tests were performed on transformed CE data based on the dataset provided in the `rpms` package. See Appendix E for more information. Tests used a regression of `FINCBTAX` on `TOTEXPCQ` with weights `FINLWT21`. While PS1q failed to reject the null hypothesis, its original version PS1 rejected the null with the significance level $\alpha = 0.05$.

4.1 Sampling

Since the significance of the survey weights for the dataset indicates a sufficiently complex sampling design, it is reasonable to use the CE data to justify performing survey weight diagnostic tests within complex surveys. For CE data, the variable of interest is the amount of CU income `FINCBTAX`. Recall that in Wang *et al*. (2023), their sampling designs were simple unequal probability sampling with no stages or levels. The following sampling methods are proposed to mimic reasonable survey designs that survey administrators may implement.

4.1.1 Grouping

Grouping is a sampling technique that groups a continuous variable X into groups based on whether the observation x_i is within a specified percentile group of X such that x_i is in some group h if $x_i \in (a, b]$ where a and b are scalars with $\min(X) \leq a < b \leq \max(X)$. Grouping by the percentiles of observations is a variation of stratified sampling, where your percentile groups are the stratum and sampling within each group across all groups. This approach acknowledges that different segments of the continuous variable X may have varying associations with the variable of interest Y.

For example, the Bureau of Labor Statistics employs a stratified sampling method with optimum allocation for the Current Employment Statistics (CES) survey which publishes detailed industry estimates of employment, earnings, and hours for nonfarm institutions. The Bureau of Labor Statistics assigns a firm to class codes determined by their number of employees (U.S. Bureau of Labor Statistics, 2024). Since larger firms generally have more variability in quantities like payroll and total hours works, optimum allocation will disproportionately sample more larger firms than smaller firms to minimize variances at a fixed cost.

With regards to calculating the inclusion probabilities, let n be the sample size, N be the population size, and p_h be the probability of selecting a population unit from a group h within the stratum set H. After determining the groups according to X, the inclusion probabilities are those of the stratum in a stratified sampling method such

Table 4.3: *Decile mean and standard deviation values for total expenditure in the current quarter* `TOTEXPCQ`*.*

Deciles	μ_h	σ_h	n_h
1	23,296.49	18,969.98	1909
2	37,599.08	26,642.66	1909
3	48,320.67	30,765.61	1909
4	56,673.46	35,820.43	1909
5	65,551.51	43,975.78	1909
6	73,859.07	46,106.41	1909
7	83,972.32	52,185.13	1909
8	90,792.47	48,947.72	1909
9	109,386.76	56,402.22	1909
10	128,754.53	67,593.16	1909

Values for μ_h and σ_h were computed based on the transformations and data wrangling as noted in Appendix E with the exception of `FINCBTAX` and `TOTEXPCQ` not being transformed by the natural logarithm function.

that the inclusion probabilities are

$$\pi_{h,i} = \frac{n \cdot p_h}{N} = \frac{n_h}{N},$$

where weights for the ith population unit in group h are $w_{h,i} = \pi_h^{-1}$.

For the grouping variable X in the CE dataset, quantitative variables would have to show significant heteroskedacity between stratum groups $h \in H$. The variable representing current total expenditures `TOTEXPCQ` shows signs of heteroskedacity of `FINCBTAX` when grouping by `TOTEXPCQ` as shown in Table 4.3.

4.1.2 Probability Proportional to Size

Probability proportional to size (PPS) is a sampling design in which each population unit has an inclusion probability proportional to a size metric X, where $X_i \in \mathbb{R}^+, \forall i$. When selecting the observation for a single sample, the probability of selecting the ith population unit p_i of being selected for this particular sample is

$$p_i = \frac{x_i}{\sum_{i \in \mathcal{U}} x_i},$$

and the inclusion probability of the ith being selected for the sample S with size n is

$$\pi_i = \frac{n \cdot p_i}{N}, \text{ with } w_i = \pi_i^{-1}.$$

Table 4.4: *Transforming* `TOTEXPCQ` *with added noise* ε *with varying degrees of noise.*

	SD(ε_i)			
X_i	0.025	0.050	0.075	0.100
2966.96	2930.53	2964.18	2855.67	3071.99
1617.65	1634.22	1622.40	1617.33	1698.22
8980.5	9028.17	9124.46	9452.63	8691.05
3205.90	3205.54	3299.39	3274.06	3217.04
2533.41	2553.28	2560.15	2511.51	2308.99
6137.75	6068.20	5884.79	6236.20	6201.16
4607.41	4616.80	4653.49	4478.87	4371.68
8302.08	8249.23	7983.75	8262.80	8139.59
19,123.55	19,045.25	19,799.64	20,080.02	19,187.37
8444.08	8574.95	8572.92	8201.33	8213.46

Values X_i as obtained using the CE data and transformations and wrangling as noted in Appendix E. Data are generated where $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ with σ varying.

Since survey administrators rarely have complete certainty about their size metric for their target population, it is necessary to add a randomness element to account for uncertainty during the survey design process. An additive noise variable is problematic since the size variable X must be positive-definite to ensure positive inclusion probabilities. Thus, let ε_i be the multiplicative noise variable for the signal variable X_i to get the new size metric Z_i where, for all ith population units,

$$Z_i = X_i \cdot (1 + \varepsilon_i),$$

with $E(\varepsilon_i) = 0$ and $\vec{\varepsilon}_i$ are independent and identically distributed. Without specifying the distribution of ε_i, $E(Z_i) = X_i, \forall i$. The noise of ε_i is dependent on its variance σ^2. In a simple model, let $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ which leads to the variance expression of Z_i of $\text{Var}(Z_i) = X_i^2 \text{Var}(\varepsilon_i) = X_i^2 \sigma^2$. See Appendix F for details on the derivation.

For the CE data, let `TOTEXPCQ` be X and Z be the transformed values of `TOTEXPCQ` with added noise of $SD(\varepsilon_i)$. As depicted in Table 4.4, larger magnitudes of σ translate to more variation of X. The larger values of X are more likely to have larger changes in magnitude than the smaller values.

4.1.3 Stratified Sampling

Stratified random sampling is a sampling method that divides N population units into H strata, where N_h is the population size within stratum h. As a common — and often desirable — sampling technique for estimator efficiency, stratified sampling takes a specified sample size from each stratum n_h which ensures that each stratum population

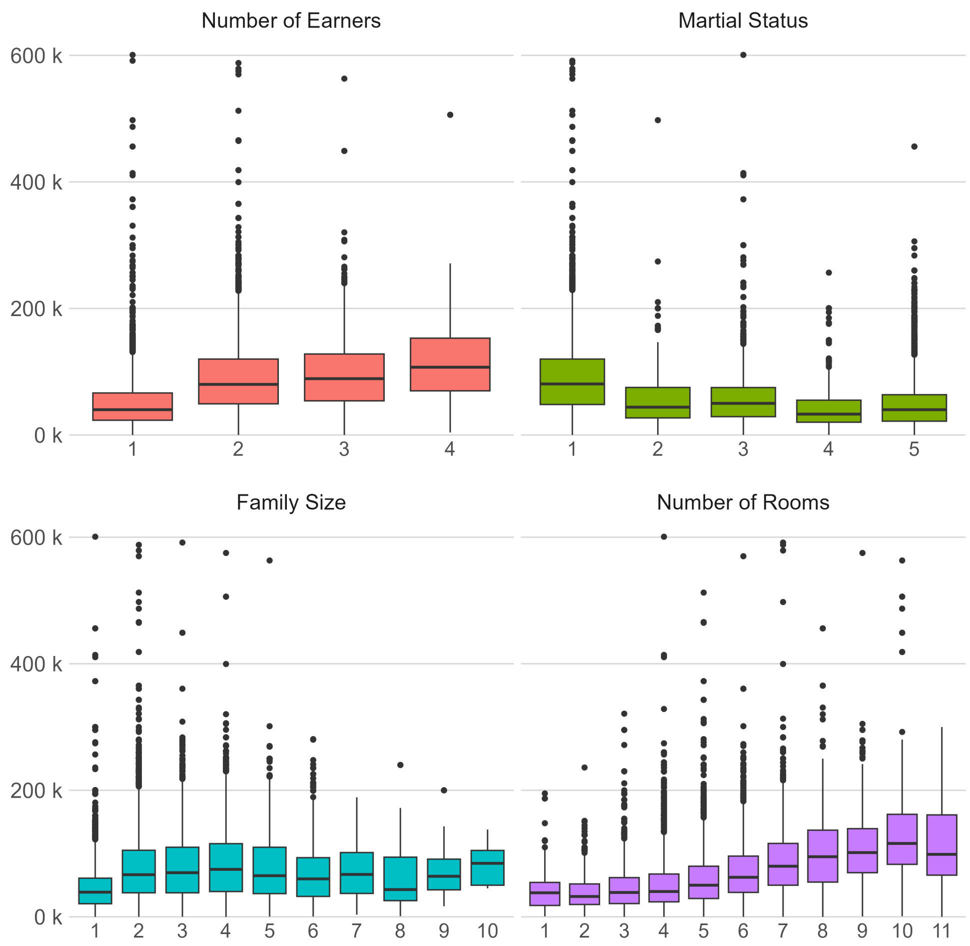

Figure 4.1: *Spread of variable* `FINCBTAX` *across earner characteristics for determining reasonable stratifying variables.*

has representation in the sample in contrast to simple random sampling (SRS). The most simple form of stratified random sampling is to take an SRS within each stratum with sample sizes n_h for a stratum h where the inclusion probability that a population unit i in stratum h will be included in the sample S is

$$\pi_{h,i} = \frac{n_h}{N_h}, \text{ and } w_{h,i} = \pi_{h,i}^{-1} = \frac{N_h}{n_h}.$$

Stratified random sampling is preferable when the strata have differences with each other to ensure that different population groups are represented. Variables that are known to have differences between strata are generally characteristic-based. In the case of CE data, candidate variables to act as the stratifying variable include `NO_EARNR`, `MARITAL`, `FAM_SIZE`, and `ROOMSQ`. As shown in Figure 4.1, the variables show a significant spread across their levels where `NO_EARNR` depict the most significant spread of `FINCBTAX` across the number of earner levels.

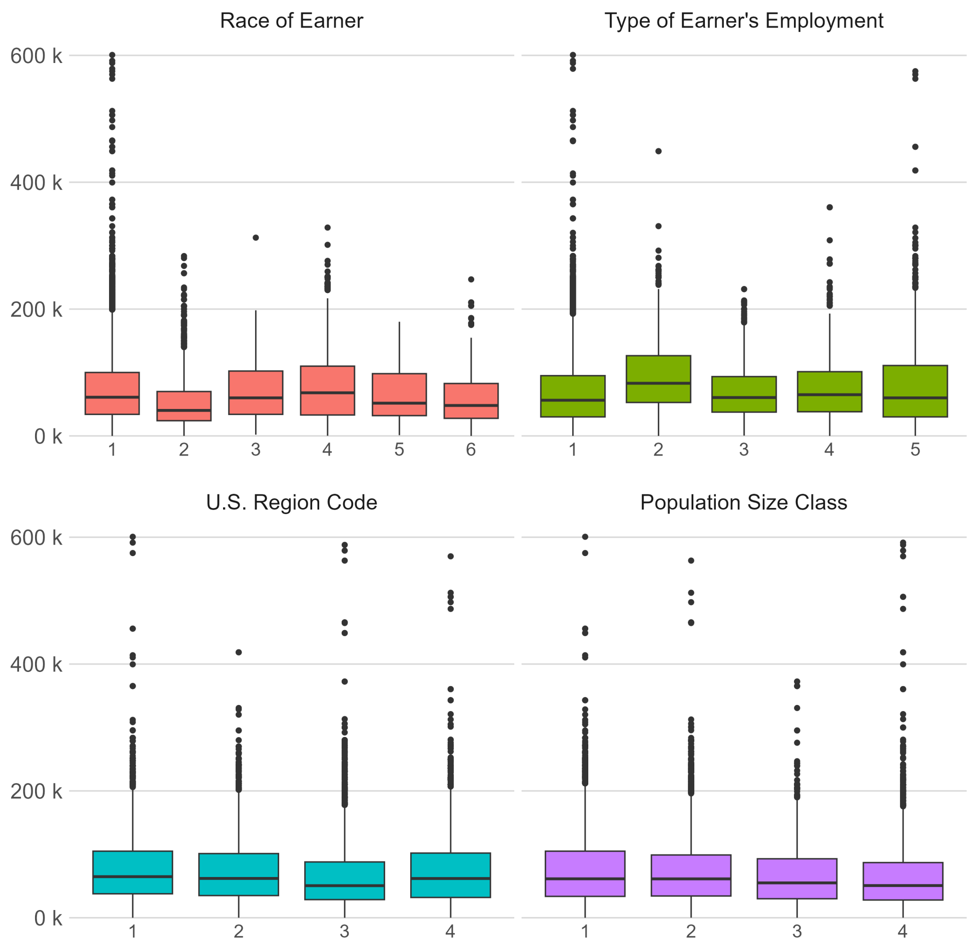

Figure 4.2: *Spread of variable* **`FINCBTAX`** *across location and earner characteristics for determining reasonable clustering psu variables.*

4.1.4 Cluster Sampling

Cluster sampling is a sampling method that selects n primary sampling units (psu) from the psu population with size N. For one-stage cluster sampling, all population units within a psu are selected. Alternatively, two-stage cluster sampling performs an SRS of m_i secondary sampling elements within each selected psu where M_i is the ssu population size within the ith psu. In contrast with stratified random sampling, cluster sampling deliberately excludes sampling for some psus since cluster sampling only samples ssu elements within the sampled psu units. Although cluster sampling is generally not optimal for estimator efficiency in comparison with other sampling methods, it is generally preferable when sampling psus is costly and can typically compensate for poor efficiency when the sample size is increased (Lohr, 2022).

The inclusion probability for the jth ssu of psu i is

$$\pi_{i,j} = \frac{n}{N}\frac{m_i}{M_i}, \text{ with } w_{i,j} = \pi_{i,j}^{-1} = \frac{N}{n}\frac{M_i}{m_i}.$$

For cluster sampling, the inclusion probability for ssu j in psu i is the product of the probability of the ith psu is selected (n/N) and the probability of the jth ssu given that the ith psu is selected (m_i/M_i).

Cluster sampling is preferable when the cluster psus is homogeneous throughout the psus and heterogeneous within to minimize the possibility of ignoring population groups. Variables that are homogeneous across and heterogeneous within the cluster psus are generally location-based. In the case of CE data, candidate variables to act as psus include `CID`, `STATE`, `REGION`, and `POPSIZE`. Interestingly, `CID` and `STATE` had significant heterogeneity across the psus and were therefore not considered further. Figure 4.2 shows the spread of `FINCBTAX` within possible cluster psu variables where `REGION` and `POPSIZE` depict homogeneity across psu levels.

4.1.5 Two-Stage Clustering and Stratified Sampling

A two-stage sampling design adds an additional layer of complexity to the sampling design to mimic the complex survey designs used by large modern surveys. Generally, the first layer of a complex survey design is to use cluster sampling to select n psus from the population of N psus. After cluster sampling, the second layer of the design is to stratify the ssus to then perform simple random sampling to obtain k tertiary sampling units (tsus) for the sample set S.

Determining the inclusion probabilities of tsus is based on the inclusion probability expressions of cluster and stratified random sampling, as mentioned above. The inclusion probability of the kth tsu in a three-stage clustering and stratifying sampling design is defined as

$$\begin{aligned}\pi_{k,h} &= P(k_h \in S)\\ &= P(i \in S_I) \cdot P(k_h \in S \mid i \in S_I)\\ &= \frac{n_I}{N_I}\frac{n_h}{N_h}.\end{aligned}$$

The inclusion probability $\pi_{k,h}$ for the kth tsu depends on the probability that its cluster psu is sampled, $P(i \in S_I)$, where S_I is the set of indices of the sampled psu groups and the probability that the tsu is sampled within the stratum h, $P(k_h \in S \mid i \in S_I)$, where S is the set of indices of the sampled tsu elements. Furthermore, n_I is the size of the sampled psu clusters S_I, N_I is the population size of the psus, N_h is the tsu population size within stratum h, and n_h is the sample size of tsus within stratum h.

4.2 Simulation Design

The purpose of this simulation is to determine the robustness of the survey weight diagnostic tests in rejecting the non-informative weight null hypothesis under complex survey designs. This simulation will sample from the Consumer Expenditure dataset using the five proposed sampling designs and compute inclusion probabilities according to the sampling design. Using the Consumer Expenditure data as the finite population to select samples, the population size is 18966 individuals after performing data wrangling as recorded in Appendix E. Suppose a researcher is interested in modeling the relationship between income and expenditures for the finite population where they decide to regress `FINCBTAX` on `TOTEXPCQ`. With this motivation, consider the following simulation setup.

Simulation Setup

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

1. With the given sampling method, calculate inclusion probabilities and corresponding weights for each population element.
2. Sample n observations from N population elements with computed inclusion probabilities.
   - Note that some sampling methods, notably clustering, may not guarantee a fixed sample size if clusters are not of equal size and may depend on the sampling allocation method. If a fixed sample size cannot be guaranteed, then sample $E(n)$ observations.
3. Perform all the aforementioned tests on the generated data with sample data $\{Y_k, X_k, W_k\}_{k \in S}$.
4. Record the corresponding p-values.

The simulation has a 5 factorial design with 25 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 (Bollen *et al.*, 2016). The power of diagnostic tests is expected to increase with larger sample sizes.

With regards to determining which variables to use for each sampling method, the choice of the variable is a significant determining factor in the performance of the weight tests, as it determines the inclusion probabilities and thus the distribution of the weights. Although the simulation design is not constructed to necessarily make an estimator efficient, the simulation design was made to mimic a reasonable sampling design. Furthermore, the distribution of the weights was made as extreme as possible to maximize the likelihood that the diagnostic tests capture the degree of informative weights.

For the grouping sampling method, `TOTEXPCQ` was determined to be a reasonable variable for grouping observations. Recall that the motivation of grouping is to segment a continuous variable into groups with heterogeneous means and variances. Then, to reduce the variance of the estimates, sample disproportionately more observations within more variable groups. Suppose that the survey administrator approximately knew which population units were in each group and knew that groups with larger values of `TOTEXPCQ` tended to be more variable than smaller values from pilot surveys. Thus, grouping would be a reasonable sampling method.

Cases:

1. Sampling Method:
   (a) Grouping: $\pi_{h,i} = \frac{n \cdot p_h}{N} = \frac{n_h}{N}$ with $w_{h,i} = \pi_h^{-1}$ using `TOTEXPCQ` as the grouping variable. For the allocation, let $n_g = n \cdot p_g$ be the sample size for the g group stratum and the strata sampling proportions for the G strata are $\vec{p}_{g \in G} = [0.15, 0.20, 0.25, 0.40]$.
   (b) Probability Proportional to Size: $\pi_i = \frac{n \cdot p_i}{N}$ with $w_i = \pi_i^{-1}$ using `TOTEXPCQ` as the signal variable.
   (c) Stratified Sampling: $\pi_{h,i} = \frac{n_h}{N_h}$ with $w_{h,i} = \pi_{h,i}^{-1}$ using `NO_EARNR` as the stratifying variable. For the allocation, let $n_h = n \cdot p_h$ be the sample size for the h stratum and the strata sampling proportions for the H strata are $\vec{p}_{h \in H} = [0.40, 0.35, 0.15, 0.10]$.
   (d) Cluster Sampling: $\pi_{i,j} = \frac{n}{N} \frac{m_i}{M_i}$, with $w_{i,j} = \pi_{i,j}^{-1} = \frac{N}{n} \frac{M_i}{m_i}$ using `INCOMEY` as the clustering variable. For the psu sample size, sample 3 clusters using equal allocation where m_i = sample size/3 is the sample size for the ith psu cluster.
   (e) Two-Stage Clustering and Stratified Sampling: $\pi_{k,h} = \frac{n_I}{N_I} \frac{n_h}{N_h}$ with $w_k = \pi_{k,h}^{-1}$ using `REGION` as clustering variable and `MARITAL` as stratifying variable. For psu sample size, sample 2 clusters using equal allocation in each stratum if $n \in \{50, 100\}$ or sample 3 clusters if $n \in \{250, 500, 1000\}$ to ensure that a sufficient amount of observations were available to be sampled within each sample level.
2. Sample Size: $n \in \{50, 100, 250, 500, 1000\}$

Constants:

- Iterations: $B = 10000$
- Sampling Population: Rows of the wrangled Consumer Expenditure dataset.
- Significance level: $\alpha = 0.05$

For Probability Proportional to Size, suppose that the survey administrator had past census data for the population on `TOTEXPCQ` where they had a continuous value instead of a categorical value. To mimic uncertainty between the time of the census and the current sampling, PPS adds noise to `TOTEXPCQ`. Note that for grouping and PPS, a possible concern is that the weights generated from `TOTEXPCQ` may be highly correlated with the target variable `FINCBTAX`. A consistent estimator of β for a linear regression of

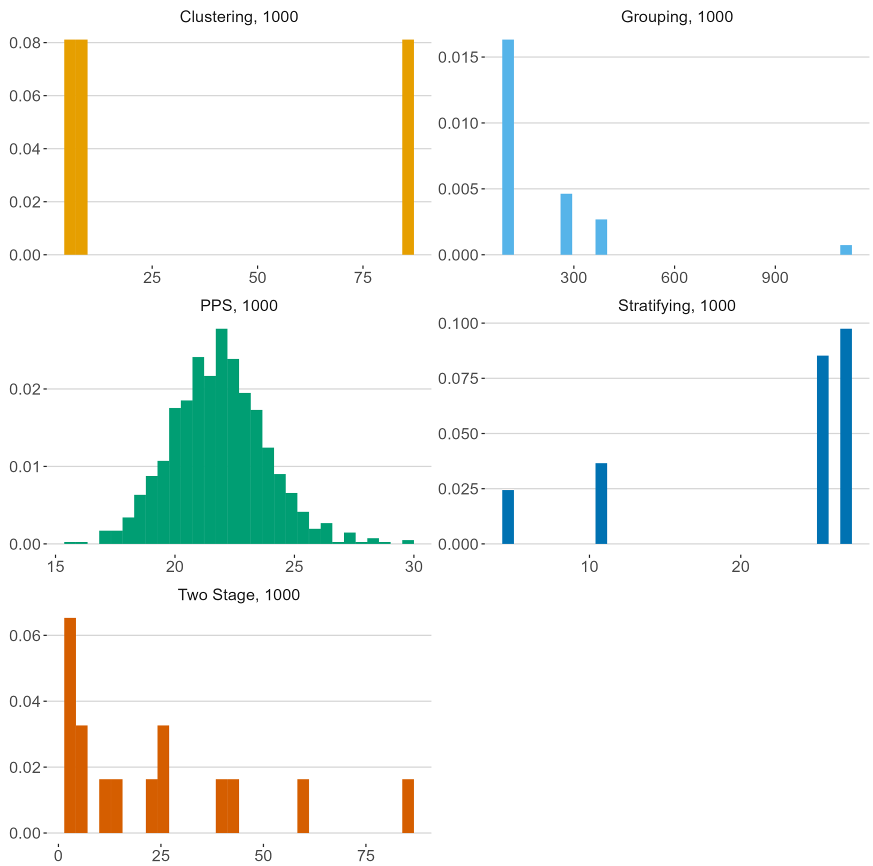

Figure 4.3: *Distribution of weights across sampling methods for showcasing sensitivity of the survey weight diagnostic tests on variable choice. Data was sampled from CE using the specified sampling method and sample size. x-axis refers to the magnitude of the weights and y-axis refers to the proportion of the population within the bins.*

`FINCBTAX` on `TOTEXPCQ` is consistent if

$$\mathrm{Cov}(W_i, Y_i \mid X_i) = 0, i \in \mathcal{U}.$$

While this is difficult to verify, the magnitude of the covariance between weights $\vec{W}$ and $\vec{Y}$ may substantially decrease conditional on $\mathbf{X}$.

For the stratification and cluster sampling methods, the stratification and cluster variables were chosen by Figure 4.1 and Figure 4.2, respectively. Cluster variables are ideal when the group means and variances are similar, and dissimilar for stratifying variables. The important factors for the variables are the number of factors within the variables and the number of elements within the factors. Factors with moderate size variation will increase the likelihood that diagnostic tests detect informative weights. See Figure 4.3 for the weight distributions of the sampling methods with their corresponding variables. Thus, weights computed from the sampling methods are highly deterministic for the survey weight diagnostic tests, which is the motivation of this simulation study.

4.3 Results

Table 4.5: *Rejection rates of survey weight diagnostic tests of eight tests based on 5000 iterations and 25 case scenarios.*

Sampling Method	n	*DD*	*HP*	*PS1*	*PS1q*	*PS2*	*PS2q*	*PS3*	*WF*
Grouping	50	22.9	22.1	38.2	86.9	17.0	14.8	5.7	13.5
	100	22.8	22.1	38.8	86.0	18.4	14.6	5.8	15.1
	250	23.0	22.3	38.6	85.7	19.3	15.6	5.5	14.9
	500	22.6	22.0	38.1	86.7	18.6	14.5	5.6	15.3
	1000	22.8	22.1	38.4	86.6	17.4	14.2	5.6	14.2
PPS	50	13.3	12.3	9.1	56.1	13.6	5.1	5.5	4.3
	100	20.7	20.0	11.0	89.8	18.2	5.7	7.8	4.9
	250	43.0	42.6	15.8	100.0	31.3	5.1	14.3	4.5
	500	68.8	68.7	24.1	100.0	48.7	6.5	25.7	5.6
	1000	93.0	93.0	40.6	100.0	74.4	6.5	49.7	6.3
Stratifying	50	17.2	15.7	12.8	10.5	13.9	12.9	18.4	12.2
	100	30.1	28.9	18.2	15.7	20.3	20.1	25.5	18.8
	250	65.2	64.8	46.3	40.6	49.2	47.8	48.6	42.1
	500	92.2	92.1	78.6	73.6	81.7	80.5	76.5	73.0
	1000	99.9	99.9	98.1	97.4	98.3	98.0	96.5	95.6
Clustering	50	10.5	9.3	6.0	5.6	7.4	7.1	10.1	7.9
	100	14.6	14.2	10.5	10.0	12.9	12.6	15.5	13.2
	250	29.5	29.2	26.7	25.3	30.8	29.6	30.4	28.7
	500	47.3	47.1	45.7	43.9	48.2	46.9	43.6	42.5
	1000	61.3	61.3	61.7	61.4	61.9	60.0	55.0	54.1
Two Stage	50	12.3	11.0	20.4	20.0	25.4	23.7	17.2	14.3
	100	20.7	20.0	37.1	36.6	43.3	41.9	26.7	23.9
	250	44.9	44.4	69.8	70.7	78.3	76.8	50.4	50.3
	500	75.5	75.4	95.1	95.8	98.2	97.9	75.6	80.4
	1000	97.3	97.3	99.9	99.9	100.0	100.0	95.9	97.9

Note: Rejection rates were determined at the $\alpha = 0.05$ significance level where rates are the percentage of tests rejecting the null hypothesis of noninformative weights.

As shown in Table 4.3, the survey weight diagnostic tests are considerably influenced by the sampling method that determines that distribution of the weights. Eight tests were considered (all previous tests excluding PN and LR due to nonconformity) which showed notable performance differences between tests and across cases. Recall that as the sample sizes n increase, the tests are expected to reject the null hypothesis more often given informative weighting. All sampling methods were designed to produce reasonably extreme weights that did not impede the performance of the tests' regression models. Note that as weights are designed to be informative for all cases, the rejection rates do not necessarily have to start from 0.05 given the significance level $\alpha = 0.05$ as was the case in Simulation Study 1: Wang *et al.* (2023).

The rejection rates for Grouping for the set of sample sizes were significantly variable across diagnostic tests. Interestingly, the rejection rates did not change significantly as n increased. PS1q was able to identify the informativeness of the weights considerably regardless of the sample size with the seven other tests performing poorly in terms of the magnitude of their rejection rates. Recall that Section 2.4.2 identifies if there is any correlation between the residuals of the unweighted regression $\hat{\varepsilon}$ and $\vec{W}$ where PS1q tries to identify whether the sample distribution of the residuals is the different from the population distribution of the errors.

The rejection rates for PPS was not uniform among the tests with some tests significantly outperforming others. Similar to Grouping, PS1q was able to reject the null hypothesis of noninformative weights considerably at a small sample size of $n = 50$ and always rejected the null hypothesis at $n \in \{250, 500, 1000\}$. HP and DD performed almost identical. Since HP is a difference-in-coefficients test and DD is a weight association test, a hypothesis for the similar rejection rates is caused by the distribution of the weights, which is approximately a Normal distribution. Bollen *et al*. (2016) notes the asymptotic equivalence between WA and DC tests where a WA test statistic has an F-distribution and a DC test statistic has a chi-square distribution. Lastly, PS2q and WF performed poorly, while PS1 and PS3 performed moderately well.

The rejection rates for Section 4.1.3 had all tests increase their rejection rates to almost 100.0% as n increased. DD and HP performed the best with a quicker convergence compared to the other tests. Although PS1q performed the worst, all tests were able to reject the null hypothesis sufficiently as n increased. For Clustering, all tests performed similar, with DD and HP performing slightly better at small sample sizes. Notably, the tests did not converge to 100.0% like other sampling methods which is likely caused by the bimodal weight distribution as depicted in Figure 4.3. Lastly, Two Stage sampling cases saw all tests converge to nearly 100.0% where PS1, PS1q, PS2, and PS2q rejecting more often at smaller sample sizes.

5

Simulation Study 3: Permutation Tests

Whereas test statistics are often bounded by strong assumptions — as noted in the descriptions of the diagnostic tests — resampling methods can provide flexible test conditions under relaxed assumptions. Permutation tests are a subset of non-parametric statistical hypothesis tests that shuffle data to determine whether all possible samples come from the same distribution. For the purpose of determining informative weights in a regression model, the survey weights are resampled to the sample data $(Y_k, \vec{X}_k)_{k \in S}$ and applied to test statistics that determined whether the original sample implied informative weighting.

When the weights are shuffled to the sample data, the test statistics will be distributed by the null hypothesis distribution such that the weights are not informative. Traditional permutation tests compare the distributions of groups, whereas for testing weight informative, the continuous survey weights act as categorical factors with the number of groups $N_g \in [1, n]$. Often, for sampling methods that use levels, the weight distributions will be similar within groups and heterogeneous between groups.

Let the sample data $(Y_1, \vec{X}_1), \ldots, (Y_n, \vec{X}_n) \overset{iid}{\sim} F_{Y,X}$ follow the arbitrary joint CDF $F_{Y,X}$ and $W_1, \ldots, W_n \overset{iid}{\sim} F_W$ follow the CDF F_W. A permutation test for

$$H_0 : F_{Y,X,W} = F_{Y,X} F_W \quad \text{vs} \quad H_A : F_{Y,X,W} \neq F_{Y,X} F_W,$$

can be conducted as follows.

1. Let T_H be a predetermined test statistic before resampling the weights. T_H should be chosen where large values of T are evidence against H_0.
2. Compute the observed value t_0 of T_H from the sample data.
3. Generate a large number of B random permutations of the weights with simple random sampling without replacement. For notably expensive computational calculations, a simple random sample with replacement may be convenient. Typical choices are $B = 1,000$ or $B = 10,000$ which is not dependent on the number of possible permutations $n!$ as the factorial function spikes with increasing sample size n.

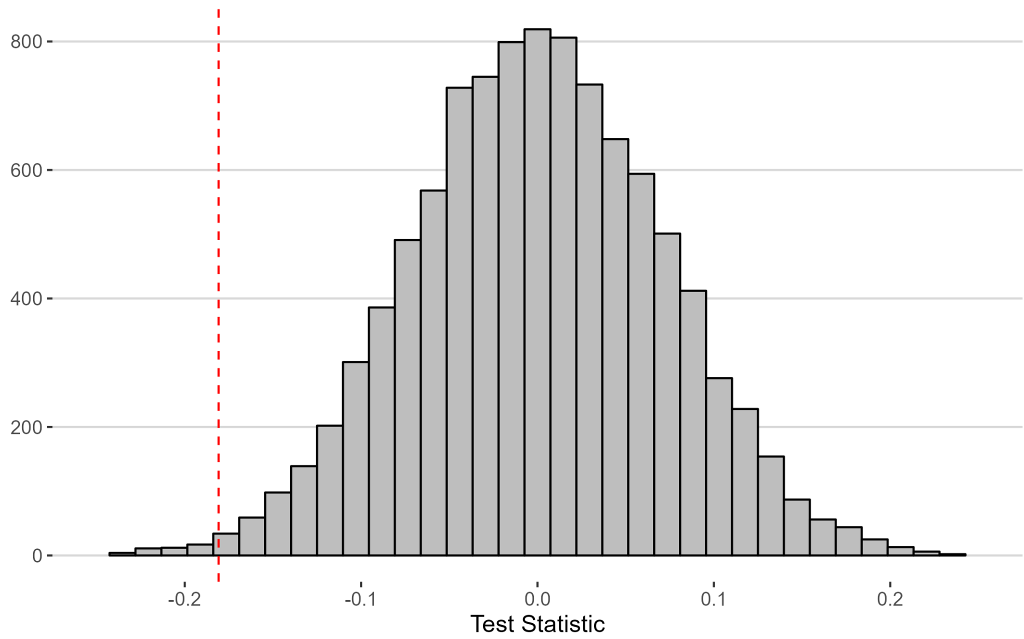

Figure 5.1: *Histogram of test statistics for the permutation version of PS1 with* $B = 10,000$. *The p-value is 0.0102 with* $t_0 = -0.1808$ *as denoted with dashed red line.*

4. For each B of random permutations, calculate the test statistic T_H as $t_1, \ldots, t_B$.
5. The p-value for the permutation test is

$$P_0(T_H \geq t_0) \approx B^{-1} \sum_{j=1}^{B} \mathcal{I}(t_j \geq t_0),$$

where P_0 denotes the probability under the permutation distribution of T of random weight shuffles. P_0 denotes a one-sided test in which two-sided tests can be computed as

$$P_{0,\ \text{two-sided}} = 2 * \min(P_0, 1 - P_0).$$

Intuitively, a permutation test under the null hypothesis suggests that weights $\vec{W}$ are interchangeable with sample data $(Y_k, \vec{X}_k)_{k \in S}$ — suggesting noninformative weights. Under the null, it would be surprising if the observed value t_0 of the test statistic were extreme compared with the simulated values of the test statistic obtained by randomly shuffling the weights.

5.1 Permutation Tests for Weight Informativeness

I propose two non-parametric permutation tests to assess the informativeness of survey weights in linear regression models. These tests are grounded in the principle that, under the null hypothesis of non-informative weights, permuting the weights across observations should not systematically alter the distribution of a test statistic that captures the effect of weighting.

5.1.1 Setup and Null Hypothesis

Let $(Y_K, \vec{X}_k, W_k)_{k \in S}$ denote observed outcomes, outcomes, and survey weights for observation k in sample S with n observations. Define $X \in \mathbb{R}^{n \times p}$ as the design matrix and $y \in \mathbb{R}^n$ as the outcome vector. The diagonal weight matrix is $W = \operatorname{diag}(w_1, \ldots, w_n)$. Let's consider two estimators:

$$\hat{\beta}_U = (X^\top X)^{-1} X^\top y \quad \text{(unweighted OLS)},$$
$$\hat{\beta}_W = (X^\top W X)^{-1} X^\top W y \quad \text{(weighted LS)}.$$

The null hypothesis is that the weights are non-informative given X:

$$H_0 : F(w \mid X, y) = F(w \mid X).$$

Under H_0, the distribution of any statistic measuring the effect of weighting should be invariant to permutations of the weights. Next, we consider two statistics:

- Predicted mean difference: $T_{\mathrm{PM}} = \hat{y}_w - \hat{y}_u$, where $\hat{y}_\cdot = n^{-1} 1^\top X \hat{\beta}_\cdot$.
- Coefficient Mahalnobis distance: $T_{\mathrm{CM}} = (\hat{\beta}_w - \hat{\beta}_u)^\top (X^\top X)(\hat{\beta}_w - \hat{\beta}_u)$.

The permutation test computes the observed statistic T_{obs}, then generates B random permutations P_b of the weight vector (optionally within blocks), forming $w^{*(b)} = P_b w$, and recomputes $T^{*(b)}$. The two-sided p-value is

$$p_B = \frac{1 + \sum_{b=1}^{B} \mathbb{I}\left\{|T^{*(b)} - T_0| \geq |T_{\mathrm{obs}} - T_0|\right\}}{B+1},$$

where T_0 is the baseline statistic under equal weights. For formalizing the notion of validity, consider the two regimes:

- Exact finite-sample conditional validity under exchangeability.
- Asymptotic validity under mild regularity (no exact exchangeability but sufficient smoothness).

Throughout, we will condition on X (and, for finite-sample arguments, treat y as fixed), consistent with the permutation paradigm that randomizes only the weights.

5.1.2 Test Statistics 1: Predicted Mean Difference

Consider the following assumptions:

- A1 (Design nonrandom for testing): Treat X as fixed; condition on the realized y.
- A2 (Null independence): Under H_0, $F(w \mid X, y) = F(w \mid X)$.
- A3 (Exchangeability): Conditional on X, the joint law of w is invariant under any permutation operator P (or within-block permutation in the blocked case).
- A4 (Well-posedness): $X^\top X$ is nonsingular; $X^\top W X$ is approximately nonsingular.

For considering invariance and exactness, define

$$T_{\mathrm{PM}} = n^{-1}\mathbb{I}^\top X\left[(X^\top WX)^{-1}X^\top Wy - (X^\top X)^{-1}X^\top y\right].$$

For any permutation matrix P that preserves blocks (if present), set $w' = Pw$ and $W' = PWP^\top$. Then

$$T_{\mathrm{PM}}(Pw \mid X, y) = T_{\mathrm{PM}}(w \mid X, y), \text{ in distribution under } H_0,$$

because by A2-A3, $F(Pw \mid X, y) = F(w \mid X, y)$. Hence, conditional on (X, y), the multiset $\{T_{\mathrm{PM}}(Pw) : P \in G\}$ has a distribution equal to that of $T_{\mathrm{PM}}(w)$ where G is the group of all admissible permutations (full or blockwise). Let $O = \{T_{\mathrm{PM}}(w), T_{\mathrm{PM}}(P_1 w), \ldots, T_{\mathrm{PM}}(P_B w)\}$. Under A3, the rank of $T_{\mathrm{PM}}(w)$ among O is uniformly distributed conditional on (X, y). Therefore, the permutation p-value p_B satisfies

$$P(p_b \le \alpha \mid X, y) \le \alpha, \forall \alpha \in [0, 1].$$

This is exact condition finite-sample validity. Now for consistency under informative weights, consider that under an informative alternative H_1 where $F(w \mid X, y) \neq F(w \mid X)$, the weighted fit responds systematically to the dependence structure. Write

$$\hat{y}_w - \hat{y}_u = \underbrace{n^{-1}\mathbb{I}^\top X(X^\top X)^{-1}X^\top(Wy - y)}_{\text{first-order constrast}} + R_n,$$

where R_n captures higher-order terms from the nonlinearity in $W \to (X^\top WX)^{-1}$. If $E(Wy \mid X) \neq E(y \mid X)E(W \mid X)$, the contrast has nonzero expectation and, under standard moment bounds, $|T_{\mathrm{PM}}(w)|$ diverges from the center T_0 at rate $n^{-1/2}$, while the permutation distribution, computed by disassociating w from y, remains centered near T_0. Hence the test is consistent:

$$p_B \xrightarrow{P} 0, \text{ under } H_1. \quad \checkmark$$

5.1.3 Test Statistics 2: Coefficient Mahalanobis Distance

Consider the following assumptions:

- B1 (Design nonrandom for testing): Design matrix X is fixed and of full column rank.
- B2 (Null independence): Under H_0, $F(w \mid X, y) = F(w \mid X)$.
- B3 (Exchangeability): Conditional on X, the joint law of w is invariant under any permutation operator P (or within-block permutation in the blocked case).
- B4 (Well-posedness): $X^\top X$ is nonsingular; $X^\top WX$ is approximately nonsingular.

Define the coefficient distance $d = \hat{\beta}_w - \hat{\beta}_u$ with

$$\hat{\beta}_u = (X^\top X)^{-1}X^\top y \quad \text{and} \quad \hat{\beta}_w = (X^\top WX)^{-1}X^\top Wy,$$

and the Mahalanobis statistic $T_{\mathrm{CM}} = d^\top(X^\top X)d$. For invariance and exactness, take any permutaiton matrix P that preserves blocks (if present), set $w' = Pw$ and $W' = PWP^\top$.

Under H_0, by null independence and exchangeability, $F(Pw \mid X, y) = F(w \mid X, y)$, so the induced distributions of $\hat{\beta}_w$ and thus d are invariant in law under $w \to Pw$. Consequentially,

$$T_{\text{CM}}(Pw \mid X, y) = T_{\text{CM}}(w \mid X, y) \text{ under } H_0.$$

Let G denote the group of admissible permutations, and consider the multiset $\{T_{\text{CM}}(Pw) : P \in G\}$. Conditional on (X, y), the rank of the observed $T_{\text{CM}}(w)$ among the permuted values is uniformly distributed. Hence, for the usual two-sided centered permutation p-value p_B formed from B random permutations,

$$P(p_B \leq \alpha \mid X, y) \leq \alpha, \forall \alpha \in [0, 1],$$

establishing exact conditional finite-sample validity.

For consistency under informative weights, suppose an alternative H_1 where $F(w \mid X, y) \neq F(w \mid X)$ induces a nonzero systematic coefficient shift such that $E(d \mid X) \neq 0$. Because $X^\top X$ is positive definite, $v \top v^\top (X^\top X) v$ is convex, and thus Jensen's inequality yields

$$E(T_{\text{CM}} \mid X) = E(d^\top (X^\top X) d \mid X) \geq E(d \mid X)^\top (X^\top X) E(d \mid X) > 0.$$

Under permutations, the dependence between w and y is broken. Thus, conditional on (X, y), the randomized statistics $T_{\text{CM}}(Pw)$ concentrate around the baseline determined by the design, lacking the systematic contribution from $E(d \mid X) \neq 0$. Therefore, the observed $T_{\text{CM}}(w)$ separates from its permutation distribution. Under standard moment bounds and eigenvalue conditions ensuring concentration, this separation implies

$$p_B \xrightarrow{P} 0, \text{ under } H_1. \quad \checkmark$$

5.2 Simulation Design

Resampling procedures are typically used to determine the sampling distribution of some statistic when distributional assumptions are not guaranteed. Permutation tests might be useful for regression models when one of the following applies:

1. Residuals are non-Normal
2. Residuals do not have constant variance
3. Sample size is relatively small.

Although permutation tests are often very flexible with linear regression assumptions, it does not help with other assumptions such as linearity or independence. Permutation tests are notably powerful because the law of large numbers ensures that the sampling CDF approximation is valid in the limit as more random permutations are performed (Blitzstein & Hwang, 2015). In a similar fashion to previous simulation studies, the tests are abbreviated as follows.

- **DD**: DuMouchel-Duncan WA Test
- **HP**: Hausman-Pfeffermann DC Test
- **PS1**: Pfeffermann-Sverchkov (1999) WA Test
- **PS2**: Pfeffermann-Sverchkov (2007) WA Test
- **PS3**: Pfeffermann-Sverchkov Estimation Test
- **PM**: Test Statistics 1: Predicted Mean Difference
- **CM**: Test Statistics 2: Coefficient Mahalanobis Distance

5.3 Study 1 Design

As discussed in Wang *et al*. (2023) and Bollen *et al*. (2016), the sensitivity of diagnostic tests to heteroskedastic data has limited research in past simulation studies. As linear regression inferential models generally assume homoskedasticity among the residuals, whether the diagnostic tests can still capture informative weights is important for determining the tests' flexibility. This first simulation study of the permutation tests recycles the weight informative setting of Study 1: Pfeffermann & Sverchkov (1999) Adaptation while incorporating heteroskedastic residuals.

Simulation Setup — Study 1

For each iteration d in D total simulation iterations, $d = 1, 2, \ldots, D$:

1. For each generated population unit $i = 1, 2, \ldots, N$:
   (a) Sample $X_i \overset{iid}{\sim} \text{Unif}(0, 1)$, $\sigma_i \overset{iid}{\sim} \mathcal{N}(\sigma, \sigma^2/9)$, $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_i^2)$, 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 $W_i = \alpha Y_i + 0.3X_i + \delta U_i$.
2. Using Probability Proportional to Size (PPS), sample n sized sample set S from the population. Subsequently, redefine $W_k = 1/\pi_k$ where π_i are generated from PPS for $k \in S$.
3. Perform all the aforementioned formal diagnostic tests on the generated data with sample data $\{Y_k, X_k, W_k\}_{k \in S}$.
4. Perform all permutation tests by computing a test statistic using the sample data $\{Y_k, X_k, W_k\}_{k \in S}$ for each b in B permutations.
5. Record the corresponding p-values.

The simulation has $4 \times 4 = 16$ case scenarios. With the linear weight-generating function from Pfeffermann & Sverchkov (1999), the cases will vary by sample sizes n and weight informative factor α. The power of the tests is expected to increase with larger sample sizes n and larger weight informative factors α. The simulation study is designed to test how the permutation tests detect informative weights at considerably low sample sizes and heteroskedastic data. Under these conditions, the permutation tests have the potential to outperform more formal diagnostic tests.

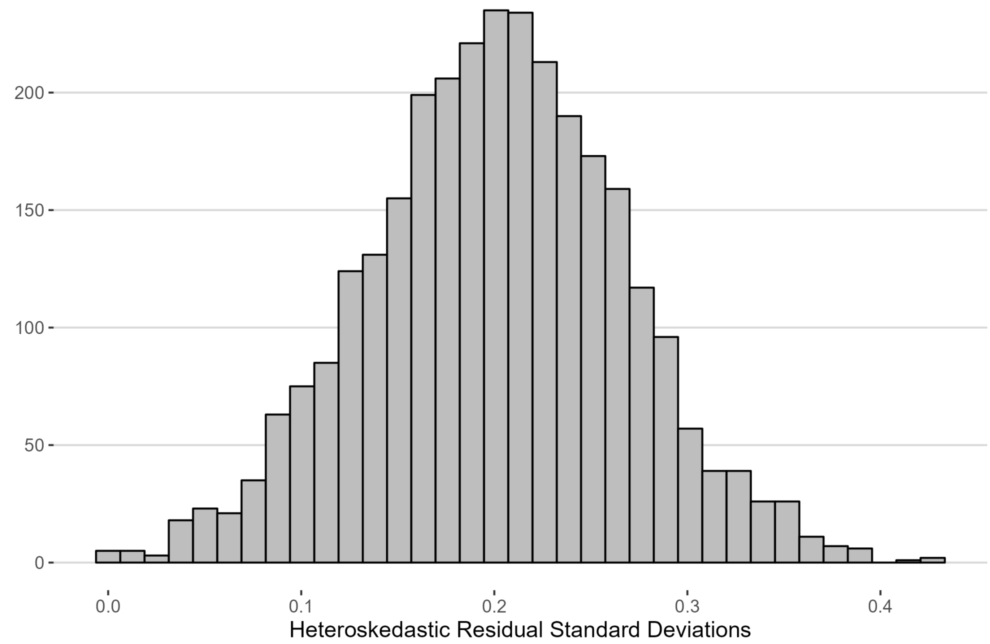

Figure 5.2: *Histogram of showcasing the heteroskedasticity of the residuals' standard deviation to further determine the robustness of diagnostic tests in non-homoscedastic data.*

Cases:

1. Sample Size: $n \in \{50, 100, 150, 200\}$
2. Weight Informativeness: $\alpha \in \{0, 0.2, 0.4, 0.6\}$

Constants:

- Simulation iterations: $D = 1000$
- Residual Expected Standard Deviation: $E(\sigma_i) = 0.2$
- Noise amplifier: $\delta = 1$
- Permutation iterations: $B = 1000$
- Population per iteration: $N = 3000$
- Significance level: $\alpha = 0.05$

5.4 Study 2 Design

A key assumption for many diagnostic tests is for the residuals being Normally distributed with a constant variance σ^2. Study 2 seeks to determine the robustness of the diagnostic tests and potential out-performance of the permutation tests given non-Normally distributed residuals. Especially in small-to-moderate sample sizes, the Central Limit Theorem might not be sufficient to meet the Normally distributed assumption for the residuals. Study 2 adds to Study 1 by incorporating various sampling distributions to generate the residuals.

Simulation Setup — Study 2

For each iteration d in D total simulation iterations, $d = 1, 2, \dots, D$:

1. For each generated population unit $i = 1, 2, \dots, N$:
   (a) Sample $X_i \stackrel{iid}{\sim} \text{Unif}(0,1)$, $\sigma_i \sim \mathcal{N}(\sigma, \sigma^2/9)$, and $U_i \stackrel{iid}{\sim} \text{Unif}(0,1)$.
   - If distributed by the Normal distribution, then $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$.
   - If distributed by the Uniform distribution, then
   $$\varepsilon_i \sim \text{Unif}\left(-\sqrt{3\sigma_i^2}, \sqrt{3\sigma_i^2}\right).$$
   - If distributed by the Gamma distribution, then
   $$\varepsilon_i \sim \text{Gamma}\left(\alpha = 10, \lambda = \sqrt{10/\sigma_i^2}\right) - \sqrt{10\sigma_i^2}.$$
   - If distributed by the Student-t distribution, then
   $$\varepsilon_i \stackrel{iid}{\sim} t_5 * \sqrt{\frac{\sigma_i^2 * 3}{5}}.$$
   (b) For all i, generate $Y_i = 1 + X_i + \varepsilon_i$.
   (c) For all i, generate the weights $W_i = \alpha Y_i + 0.3X_i + \delta U_i$.
2. Using Probability Proportional to Size (PPS), sample n sized sample set S from the population. Subsequently, redefine $W_k = 1/\pi_k$ where π_i are generated from PPS for $k \in S$.
3. Perform all the aforementioned formal diagnostic tests on the generated data with sample data $\{Y_k, X_k, W_k\}_{k \in S}$.
4. Perform all permutation tests by computing a test statistic using the sample data $\{Y_k, X_k, W_k\}_{k \in S}$ for each b in B permutations.
5. Record the corresponding p-values.

The simulation has $4 \times 2 \times 4 = 32$ case scenarios. With the linear weight-generating function from Pfeffermann & Sverchkov (1999), the cases will vary by sample sizes n, distribution, and weight-informative factor α. The power of the tests is expected to increase with larger sample sizes n and larger weight informative factors α. In a similar fashion to Wang *et al.* (2023) simulation study regarding error distributions, the distributions were scaled to center them by their means so that their mean is zero with variance σ^2 which matches the first two moments of $\mathcal{N}(0, \sigma^2)$. The simulation study is designed to test how the permutation tests detect informative weights at moderately low sample sizes with Non-Normal, heteroskedastic data. The cases were chosen to test the robustness of the diagnostic tests despite the data-generating process and small sample size violating the tests' assumptions. Given this, permutation tests have the potential to outperform formal diagnostic tests.

Cases:

1. Sample Size: $n \in \{50, 100\}$
2. Distribution (zero mean with variance σ^2):
    - Normal as $\mathcal{N}\left(0, \sigma_i^2\right)$
    - Uniform as Unif $\left(-\sqrt{3\sigma_i^2}, \sqrt{3\sigma_i^2}\right)$
    - Gamma as Gamma $\left(\alpha = 10, \lambda = \sqrt{10/\sigma_i^2}\right) - \sqrt{10\sigma_i^2}$
    - Student-t as $t_5 * \sqrt{\frac{\sigma_i^2 * 3}{5}}$
3. Weight Informativeness: $\alpha \in \{0, 0.2, 0.4, 0.6\}$

Constants:

- Simulation iterations: $D = 1000$
- Residual Expected Standard Deviation: $E(\sigma_i) = 0.2$
- Noise amplifier: $\delta = 1$
- Permutation iterations: $B = 1000$
- Population per iteration: $N = 3000$
- Significance level: $\alpha = 0.05$

5.5 Results

Study 1 and Study 2 were constructed to showcase potential advantages of permutation tests over the formal survey weight diagnostic tests given non-Normal, heteroskedastic data under small-to-moderate sample sizes. As shown in Table 5.1, the survey weight permutation tests and diagnostic tests show considerable differences between sample sizes and degrees of informative weights. Fortunately, the permutation tests generally maintained their sizes at the significance level 0.05 when $\alpha = 0.0$ for noninformative weighting.

Across the tests, the permutation tests PM and CM did not perform better than the parametric tests, suggesting that the diagnostic tests are robust to heteroskedastic data conditions. When comparing Table 3.2 to the cases in Table 5.1 for $\delta = 1.0, n = 100$, and $\sigma = 0.2$ for HP, DD, PS1, and PS2, rejection rates were not significantly different, further suggesting robustness to heteroskedastic data.

For Study 2, Table 5.2 shows that while the permutation and formal diagnostic tests generally show higher rejection rates as the sample sizes and the degree of informative weighting increase, they show a clear gap where the formal diagnostic tests reject the null hypothesis more often than the permutation tests. Recall that Study 2 generates data with heteroskedastic residuals under various distributions with small sample sizes, which directly conflicts with the assumptions of the formal diagnostic tests.

Across the four distributions, the tests perform similarly between the tests, but some distributions show higher rejection rates at the same n and α. For example, the Gamma distributed residuals have high test rejection rates while Student-t distributed residuals have considerably lower test rejection rates. This contrasts with the Wang *et al.* (2023) simulation study showing generally similar rejection rates across distributions, with Student-t rejection rates being the highest among the four distributions.

Table 5.1: *Study 1 empirical rejection rates of four permutation tests and four corresponding diagnostic tests with heteroskedastic residuals based on 1000 simulation replicates, 1000 permutations, and 16 case scenarios.*

n	α	*DD*	*HP*	*PS1*	*PS2*	*PS3*	*PM*	*CM*
50	0.0	4.3	4.0	6.7	6.4	2.4	3.5	3.5
	0.2	10.0	9.6	13.1	14.2	6.5	7.8	7.7
	0.4	28.6	27.3	26.2	34.1	18.5	21.7	16.3
	0.6	62.6	61.2	50.8	61.3	45.8	45.1	33.8
100	0.0	2.9	2.9	6.3	6.4	3.1	2.6	3.5
	0.2	15.5	15.0	15.1	19.8	10.3	10.9	10.0
	0.4	59.1	58.3	46.4	58.3	50.4	49.2	36.4
	0.6	92.3	92.1	84.4	89.7	88.8	83.7	73.0
150	0.0	4.6	4.4	7.9	6.6	3.1	2.7	3.7
	0.2	24.7	24.2	20.6	28.3	19.8	20.3	17.3
	0.4	81.3	81.2	69.9	79.4	74.0	74.2	59.4
	0.6	99.1	99.1	96.6	98.6	98.8	97.0	92.2
200	0.0	6.8	6.7	8.2	7.2	5.3	3.3	5.0
	0.2	29.6	29.3	25.0	33.6	26.5	25.4	20.2
	0.4	89.5	89.5	82.9	87.9	88.0	86.2	74.0
	0.6	99.9	99.9	99.4	99.8	99.7	99.5	98.0

Note: Rejection rates were determined at the $\alpha = 0.05$ significance level where rates are the percentage of tests rejecting the null hypothesis of noninformative weights.

Table 5.2: *Study 2 empirical rejection rates of four permutation tests and four corresponding diagnostic tests with heteroskedastic residuals under various distributions based on 1000 simulation replicates, 1000 permutations, and 32 case scenarios.*

Distribution	n	α	DD	HP	PS1	PS2	PS3	PM	CM
Gamma	50	0.0	5.9	5.7	8.8	7.4	3.0	4.4	5.3
		0.2	12.4	11.6	11.4	17.6	10.5	12.1	6.5
		0.4	60.8	59.1	49.5	58.6	50.1	52.6	37.3
		0.6	93.2	92.9	84.6	91.0	85.9	84.5	69.0
	100	0.0	4.6	4.3	6.6	6.1	3.1	3.5	4.2
		0.2	26.5	26.1	21.5	28.6	28.8	27.7	17.3
		0.4	90.5	90.3	79.5	87.3	90.7	86.5	72.1
		0.6	99.9	99.9	99.7	99.9	99.8	99.4	98.4
Normal	50	0.0	5.8	5.7	7.0	6.5	4.3	4.1	4.9
		0.2	9.8	9.2	11.4	13.7	6.0	7.4	7.2
		0.4	29.7	28.8	25.8	33.4	18.6	22.7	16.4
		0.6	68.4	67.2	54.9	66.3	48.7	51.5	36.1
	100	0.0	5.3	5.0	7.5	6.5	2.6	2.8	4.7
		0.2	18.3	18.1	17.3	22.0	15.2	16.1	12.5
		0.4	62.6	62.4	52.4	61.9	52.2	52.4	40.1
		0.6	94.7	94.5	88.8	92.7	88.8	88.2	74.0
Uniform	50	0.0	4.9	3.8	7.6	6.9	5.1	3.8	4.5
		0.2	9.9	9.4	12.9	14.7	6.2	7.2	8.6
		0.4	34.5	32.8	28.0	35.3	24.2	25.3	19.3
		0.6	68.1	66.9	54.9	65.9	55.0	49.5	36.4
	100	0.0	4.2	4.1	7.1	5.5	2.8	2.6	3.4
		0.2	19.1	18.8	17.8	22.8	15.3	15.8	12.2
		0.4	62.7	62.4	52.7	62.1	56.7	55.2	42.0
		0.6	95.1	94.8	89.1	94.6	92.6	89.3	79.2
Student-t	50	0.0	05.2	05.2	06.3	04.5	04.7	04.2	04.2
		0.2	9.6	9.0	10.6	12.1	5.4	8.0	5.7
		0.4	32.5	31.1	29.4	33.7	18.2	24.4	17.7
		0.6	63.1	62.1	50.9	61.4	38.5	46.9	32.6
	100	0.0	6.3	6.2	7.2	5.6	2.6	4.4	4.6
		0.2	16.3	16.2	14.3	18.6	09.8	13.8	10.9
		0.4	60.9	60.8	48.4	57.9	43.3	49.6	38.1
		0.6	93.0	92.8	84.7	90.9	78.8	84.2	71.1

Note: Rejection rates were determined at the $\alpha = 0.05$ significance level where rates are the percentage of tests rejecting the null hypothesis of noninformative weights.

6

Conclusion

After three grueling, multifaceted simulation studies, this researcher can confidently declare: 'to weight or not to weight' — that is no longer the question in weighted linear regressions. Determining whether the survey weights are informative for the regression model of the response variable $\vec{Y}$ and the explanatory variables $\mathbf{X}$ can be adequately performed using the survey weight diagnostic tests. The three simulation studies showed that the diagnostic tests can 1) sufficiently detect informative weights given a reasonable sample size, 2) be used with complex survey data such that the tests can detect informative weights with various survey weight distributions, and 3) reasonably robust to small sample sizes, heteroskedasticity of the data with regards to the residuals distributions and variances.

6.1 Simulation Overview

In Chapter 3, the simulation studies in Wang *et al.* (2023) were replicated to verify their findings. Generally, the replication results matched the results in Wang *et al.* (2023), but with some notable exceptions. First, the simulation instructions did not explicitly redefine the survey weights as the inverse of the selection probabilities, but assumed them for the replication study designs. This discrepancy is arguably a key reason for the differences in results between Wang *et al.* (2023) and the replications. Note the most of the results were preserved since the generated weights and inverse selection probability weights are highly correlated. For example, Study 1: Pfeffermann & Sverchkov (1999) Adaptation had nearly identical results when the weights were not redefined as the inverse selection probability rates. Furthermore, the construction of the PN and LR tests may have been performed differently than Wang *et al.* (2023), which is likely the reason the tests performed differently but cannot be determined without comparing the tests' code.

Continuing for Simulation Study 1, Study 2: Quadratic Weight Generating Function contrasted with Wang *et al.* (2023) results such that the replication results showed that most tests were able to detect informative weighting within the non-linear weight-generating function. Again, this is likely due to the redefining of the survey weights as the inverse selection probabilities, where Wang *et al*. (2023) did not explicitly redefine the weights after generating them. This distinction is significant because it might advocate the use of diagnostic tests in non-linear circumstances instead of creating variations to specifically be used in non-linear cases. Study 3: Wu & Fuller (2005) Adaptation was designed to showcase the many circumstances where some tests outperformed others given different correlation factors α and degree of informative weighting ψ. While the general patterns matched the results of Wang *et al*. (2023), the redefining of the weights and random sample size is likely the cause of this difference.

Figure 6.1: *Scatterplot of 500 sampled CE observations using stratified random sampling to showcase the effect of weighting on linear regressions. The blue and red best-fit lines represent the nonweighted and weighted linear regressions, respectively. Weights are depicted by radius of the circles centered by observation points.*

In Simulation Study 2: CE Sampling, the diagnostic tests were evaluated within real complex survey data to determine their performance with sampling methods instead of simple unequal probability sampling. Arguably the most important simulation study of the three, the results indicate that the diagnostic tests are considerably sensitive to the distribution of the weights (see Figure 4.3 for a visualization) but sufficiently able to detect informative weights within the regressions. Fortunately, complex sampling designs such as stratifying, clustering, and two-stage sampling designs did not disrupt the performances of the tests despite rigid distributions of the weights, which indicates the tests' usefulness within complex survey data.

Furthermore, Figure 6.1 and Table 6.1 demonstrate the importance of diagnostic tests when performing linear regressions with survey data. Figure 6.1 showcases the sampled Consumer Expenditure sampled data from stratified sampling with sample size of 500. Although the unweighted and weighted regressions look nearly identical and well within their 95% confidence intervals, the tests indicate that the weights are highly informative to the regression model by the p-values in Table 6.1. The unweighted regression coefficient estimate is $\hat{\beta}_u = 0.7597$ with $\mathrm{CI}(\hat{\beta}_u) = [0.6576, 0.8617]$ and the weighted regression coefficient estimate is $\hat{\beta}_w = 0.7627$ with $\mathrm{CI}(\hat{\beta}_w) = [0.6425, 0.8829]$. The regression coefficients are well within their 95% confidence intervals, despite the diagnostic tests strongly rejecting the null hypothesis of noninformative weights. This indicates that diagnostic tests are somewhat necessary regardless of informal methods such as visualizing the regression's fit or determining overlap with confidence intervals.

Table 6.1: *Survey Weight Diagnostic Test p-value results on sampled Consumer Expenditure Data as shown in Figure 6.1.*

	DD	*HP*	*PS1*	*PS1q*
p-values	2.41×10^{-5}	2.95×10^{-5}	1.02×10^{-5}	2.41×10^{-5}
	PS2	*PS2q*	*PS3*	*WF*
p-values	6.79×10^{-6}	1.01×10^{-5}	2.73×10^{-5}	2.20×10^{-4}

Diagnostic tests were performed on sampled CE data based on the the simulation design in Simulation Study 2. The tests used a regression of `FINCBTAX` on `TOTEXPCQ` with generated weights from stratified sampling.

Lastly, Simulation Study 3 tests the potential usefulness of permutation tests when the assumptions of the formal diagnostic tests are violated. Within the standard linear regression assumptions, diagnostic tests assume that the residuals are distributed as $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ where violations can underestimate the standard errors of the coefficient — increasing the probability of Type I errors and likelihood of stating invalid inferences. Several permutation tests were proposed to incorporate elements of the formal diagnostic tests while remaining flexible to the tests' assumptions. The simulation design was constructed to allow the permutation tests to outperform the diagnostic tests given non-Normal, heteroskedastic residuals within small sample sizes. Remarkably, diagnostic tests seemed to be unrestrained despite violations of their residuals' assumptions as they widely outperformed their permutation test counterparts across sample sizes and distributions.

6.2 Diagnostic Test Review

Although simulation studies indicated that diagnostic tests can detect informative weighting within many circumstances, the performances of individual tests are highly informative to determine which tests to perform for a regression model. The following are aggregated test descriptions as a short-hand review of the tests in addition to the tables made by Bollen *et al.* (2016).

Table 6.2: *Difference-in-Coefficients Tests and Other Diagnostic Tests Overview*

Test Name	Test Statistic	Simulation Notes
Hausman-Pfeffermann Test HP — Hausman (1978)	$T_H = (\hat{\beta}_1 - \hat{\beta}_2)' \hat{V}_H^{-1} (\hat{\beta}_1 - \hat{\beta}_2)$ $\hat{V}_H = \hat{\sigma}^2 \mathbf{A}\mathbf{A}^\top$ $T_H \sim \chi^2_k$, with $k = \dim(\hat{\beta})$	As the only difference-in-coefficients test, HP consistently performed well across the simulation studies. The proposed permutation test version of HP failed to detect informative weighting in Simulation 3. Because of its performance and lone DC test, HP should be considered within the portfolio of diagnostic tests at the disposal of researchers.
Pfeffermann-Sverchkov Test PS3 — Pfeffermann & Sverchkov (2003)	$T_H = \frac{n-p}{p} \bar{R}_n^{-\top} \hat{\Sigma}_{R,n}^{-1} \bar{R}_n$ $T_H \sim F_{p,n-p}$	In contrast to the results in Wang *et al.* (2023), PS3 performed moderately well along with PS1 and PS2. However, in terms of sampling, PS3 performed considerably worse in detecting informative weights for complex sampling, such as clustering and PPS.
Pfefferman-Nathan Test — PN — Pfeffermann & Nathan (1985)	$D_i = v_{u,i}^2 - v_{w,i}^2$ $T_H = \bar{D}/S_D \sim \mathcal{N}(0,1)$	Throughout the simulation studies, PN continuous performed poorly with no noticeable differences regardless of cases. This is likely due to splitting the sample size into smaller sets and shows dependence among the prediction errors by shared coefficient estimates.
Breidt Likelihood Ratio Test LR — Breidt *et al.* (2013)	$T_U = 2(l(\hat{\theta}_u; \vec{\omega}_u) - l(\hat{\theta}_w; \vec{\omega}_u))$ $T_u \xrightarrow{\mathcal{L}} \sum_{j=1}^{q} \lambda_{u,j} Z_j^2$	Similar to PN, LR performed poorly throughout the simulation studies — unable to detect various levels of informative weighting throughout the cases. While it may be infer to other tests for linear regressions, the LR test may be key for testing weights for generalized linear models.

Table 6.3: *Weight Association Tests Overview*

Test Name	Test Statistic	Simulation Notes
DuMouchel-Duncan Test — DD — DuMouchel & Duncan (1983)	$\vec{Y} = \mathbf{X}\beta_u + \mathbf{X}_w\beta_w + \vec{\varepsilon}$ F-test of $H_0 : \beta_w = 0$	Across the simulation studies, DD performed very well with sometimes being the best performing test. As a relatively simple weight association test, DD should be utilized to test informative weights.
Pfefferman-Sverchkov Test — PS1 — Pfeffermann & Sverchkov (1999)	$E(\vec{W} \mid \hat{\varepsilon}_u) = f(\mathbf{X};\eta) + \sum_{k=1}^{2} \beta^{(k)} \hat{\varepsilon}_u^k + \hat{\varepsilon}_u X\gamma$ F-test of $H_0 : \gamma = 0$	Together with its quadratic variant PS1q, PS1 performed well throughout the studies except for performs higher Type 1 errors for Study 3 in the replication simulation studies. For sampling, PS1 and PS1q performed very well, especially in grouping and PPS.
Pfefferman-Sverchkov Test — PS2 — Pfeffermann & Sverchkov (2007)	$E\left(\vec{W} \mid \mathbf{X}, \vec{Y}\right) = f(\mathbf{X};\eta) + \sum_{k=1}^{2} \vec{Y}^k \gamma_k$ F-test of $H_0 : \gamma_1 = \gamma_2 = 0$	Similar to PS1, PS2 also performs well across the sampling methods and within the studies. For their permutation variants, pPS1 and pPS2 performed the best out of the permutation tests, but trailing the performance of the diagnostic tests.
Wu-Fuller Test — WF — Wu & Fuller (2005)	$\vec{Y} = \mathbf{X}^\top \beta + \widetilde{\mathbf{X}}\widetilde{\beta} + \vec{\varepsilon}$, where $\widetilde{\mathbf{X}} = \mathbf{Q}\mathbf{X}$ $\mathbf{Q} = \mathrm{diag}(q_1, q_2, \ldots, q_n)$ F-test of $H_0 : \tilde{\beta} = 0$	Out of the weight association tests, WF likely performs the most poorly. This was shown in poor rejection rates across the sampling methods and the replication of Wang *et al*. (2023)'s Study 3.

6.3 Discussion

While the three simulation studies provided much needed insight into diagnostic tests, additional simulation studies are needed to determine their feasibility in other circumstances. While marginally addressed in the small simulation studies that were largely designed to illustrate a new test's potential, determining the performances of the tests with ordinal scale or categorical variables is needed since many survey variables of interest are measured in discrete terms.

Furthermore, the simulation studies only used a simple linear regression within the tests. Although it is likely analogous between simple and multiple linear regression, confirming the insensitivity to the complexity of the linear regression model would be assuring. Wang *et al.* (2023) did a limited simulation of multiple linear regressions by sampling observations from the 2015 Chinese household consumption expenditure data and found no noticeable performance changes. This is likely the case because adding more predictor variables does not significantly change the informativeness of the weights on the response variables.

While the simulation studies were testing linear regressions, the necessity of weighting in other models — such as logistic regressions within the larger group of generalized linear models — is uncertain for researchers. The LR test proposed in Breidt *et al.* (2013) has the general structure to test generalized linear models by likelihood ratio tests of models with and without survey weights. Sadly, its linear regression form did not perform well in replication studies or in Wang *et al.* (2023), but its performance for generalized linear models is of great interest.

In conclusion, the survey weight diagnostic tests are relatively robust, which shows promise in their use in practice. More effort needs to be made to make these tests widely accessible to researchers through packages or guides that utilize existing functions to test informative weights. Fortunately, the performance of the tests in the simulation studies makes an argument for their adaptation by researchers who use complex survey data.

References

Asparouhov, T. & B. Muthen (2007). "Testing for informative weights and weights trimming in multivariate modeling with survey data". In: 2, pp. 3394–99. URL: https://api.semanticscholar.org/CorpusID:4506846.

Blitzstein, Joseph K. & Jessica Hwang (2015). *Introduction to probability*. eng. 2nd edition. Texts in Statistical Science. Boca Raton: CRC Press/Taylor & Francis Group. ISBN: 1-4665-7559-X.

Bollen, K. A., P. P. Biemer, F. A. Karr, S. Tueller & M. E. Berzofsky (2016). "Are Survey Weights Needed? A Review of Diagnostic Tests in Regression Analysis". In: *Annual Review of Statistics and Its Applications* 3, pp. 375–392. DOI: 10.1146/annurev-statistics-011516-012958.

Breidt, F. Jay, Jean D. Opsomer, Wade Herndon, Ricardo Cao & Mario Francisco-Fern (2013). "Testing for informativeness in analytic inference from complex surveys". In: *Proceedings 59th isi world statistics congress*. Hong Kong, pp. 889–893.

DuMouchel, William H. & Greg J. Duncan (1983). "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples". In: *Journal of the American Statistical Association* 78, pp. 535–543.

Gelman, Andrew (2007). "Struggles with Survey Weighting and Regression Modeling". In: *Statistical Science* 22.2, pp. 153–164.

Hausman, J.A. (1978). "Specification Tests in Econometrics". In: *Econometrica* 46.6, pp. 1251–1271.

Herndon, Wade Wilson (2014). *Testing and adjusting for informative sampling in survey data*. eng.

Horvitz, D. G. & D. J. Thompson (1952). "A Generalization of Sampling Without Replacement from a Finite Universe". In: *Journal of the American Statistical Association* 47.260, pp. 663–685. ISSN: 0162-1459.

King, Susan, Taylor Wilson & Sharon Krieger (Feb. 2021). *An Overview of the State-Level Weighting Procedure for the Consumer Expenditure Survey*. Tech. rep. Consumer Expenditure Surveys Program.

Kish, Leslie & Martin Richard Frankel (1974). "Inference from Complex Samples". In: *Journal of the Royal Statistical Society* 36.1, pp. 1–37.

Kott, Phillip S. (2018). "A design-sensitive approach to fitting regression models with complex survey data". eng. In: *Statistics surveys* 12.none. ISSN: 1935-7516.

Lohr, Sharon L. (2022). *Sampling: Design and Analysis*. 3rd ed. Boca Raton: CRC Press.

Lusinchi, Dominic (2014). "Straw Poll Journalism and Quantitative Data: The case ofThe Literary Digest". In: *Journalism studies (London, England)* 16, pp. 417–432. ISSN: 1461-670X.

Minnesota Libraries, University of (2016). *American Government and Politics in the Informtaion Age: Polling the Public*. URL: https://open.lib.umn.edu/americangovernment/chapter/7-3-polling-the-public/ (visited on 03/26/2024).

Pfeffermann, Danny (1993). "The Role of Sampling Weights When Modeling Survey Data". In: *International Statistical Review* 61.2, pp. 317–337.

Pfeffermann, Danny & Gideon Nathan (1985). "Problems in model identification based on data from complex sample surveys". In: *Bulletin of the International Statistical Institute* 51.12.2, pp. 1–12.

Pfeffermann, Danny & Michail Sverchkov (1999). "Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data". In: *Indian Statistical Institute* 61.1, pp. 166–186.

— (2003). "Fitting generalized linear models under informative sampling". In: Chichester, UK: John Wiley & Sons, Ltd. Chap. 12, pp. 175–195.

— (2007). "Small area estimation under informative probability sampling of areas and within the selected areas". In: *Journal of the American Statistical Association* 102.480, pp. 1427–1439.

Särndal, Carl-Erik, Bengt Swensson & Jan Wretman (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Schenker, Nathaniel & Jane F Gentleman (2001). "On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals". eng. In: *The American statistician* 55.3, pp. 182–186. ISSN: 0003-1305.

Si, Yajuan, Rob Trangucci, Jonah Sol Gabry & Andrew Gelman (2020). "Bayesian hierarchical weighting adjustment and survey inference". In: *Survey Methodology* 46.2, pp. 181–214.

Squire, Peverill (1988). "Why the 1936 Literary Digest Poll Failed". In: *Public opinion quarterly* 52.1, pp. 125–133. ISSN: 0033-362X.

Toth, Daniell (2021). *rpms: Recursive Partitioning for Modeling Survey Data*. R package version 0.5.1. URL: https://CRAN.R-project.org/package=rpms.

U.S. Bureau of Labor Statistics (2015). *Consumer Expenditure Surveys - Glossary*. URL: https://www.bls.gov/cex/csxgloss.htm (visited on 03/20/2024).

— (2023). *Consumer Expenditure Surveys*. URL: https://www.bls.gov/cex/ (visited on 01/09/2024).

— (Feb. 2024). *Business Employment Dynamics: Design*. URL: https://www.bls.gov/opub/hom/ces/design.htm#selection-weights (visited on 03/21/2024).

Wang, Feng, HaiYing Wang & Yan Jun (2023). "Diagnostic Tests for the Necessity of Weight in Regression With Survey Data". In: *International Statistical Review* 91.1, pp. 55–71.

Wu, Yuehua & Wayne A. Fuller (2005). "Preliminary testing procedures for regression with survey samples". In: *Proceedings of the joint statistical meetings, survey research methods section*, pp. 3683–88.

List of Figures

List of Tables

A

Simulation Code Directory

Providing the code for the tests and simulation set-ups is key for ensuring reproducibility and comparing results with other researchers. Any code used for creating tables and figures are documented below.

Visit https://github.com/cnlubianski/survey-weight/tree/main for the main page of the github repository.

- Diagnostic Test Function Code
- Permutation Test Function Code
- Sampling Method Function Code
- Simulation 1 Replication Functions and Case Set-up
- Simulation 2 Sampling CE Function and Case Set-Up
- Simulation 3 Permutation Test Function and Case Set-Up
- Images
- Datasets from Simulations

Although not finalized, the functions from this thesis will accumulate into an R package called "weitest" that will make the diagnostic tests more accessible to researchers testing survey weights in their regression models.

Figure A.1: *Proposed R package Hex Sticker.*

B

Simulation 1 Derivations

Wu & Fuller (2008) $E(W_i)$ Derivation

The claim that $E(W_i) = 0.221$ in Wang *et al.* (2023) for study 3 is not contextualized for the parameters (α, β, ψ) when $E(W_i)$ has its function. Below is the derivation of its expectation and a table of $E(W_i)$ by the simulation cases ψ and α. W_i has the random components X_i, ε_i, and Z_i where they are all distributed $\mathcal{N}(\mu = 0, \sigma^2 = 0.5)$.

$$W_i = \alpha \cdot \eta(X_i) + \beta\eta(\psi \cdot \varepsilon_i + (1-\psi) \cdot Z_i)$$

$$E(W_i) = E(\alpha \cdot \eta(X_i) + \beta\eta(\psi \cdot \varepsilon_i + (1-\psi) \cdot Z_i)) = \alpha E(\eta(X_i)) + \beta E(\eta(\psi \cdot \varepsilon_i + (1-\psi) \cdot Z_i))$$

$$\text{Recall that } \eta(x) = \begin{cases} 0.025, & x < 0.2 \\ 0.475(x-0.2) + 0.025, & 0.2 \le x \le 1.2 \\ 0.5, & 1.2 < x. \end{cases}$$

$$\begin{aligned}
E(\eta(X_i)) &= E(0.025 \cdot \mathcal{I}(X_i < 0.2) + (0.475(X_i - 0.2) + 0.025) \cdot \mathcal{I}(X_i \in [0.2, 1.2]) + 0.5 \cdot \mathcal{I}(1.2 < X_i)) \\
&= 0.025 \cdot P(X_i < 0.2) + 0.475 \cdot E(X_i \cdot I(X_i \in [0.2, 1.2])) \\
&\quad - 0.07 \cdot P(0.2 \le X_i \le 1.2) + 0.5 \cdot P(1.2 < X_i) \\
&= 0.025 \cdot \Phi\left(\frac{0.2}{\sigma}\right) + 0.475 \int_{0.2}^{1.2} \frac{X_i}{\sigma} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(X_i/\sigma)^2}{2}\right) dx \\
&\quad - 0.07\left(\Phi\left(\frac{1.2}{\sigma}\right) - \Phi\left(\frac{0.2}{\sigma}\right)\right) + 0.5\left(1 - \Phi\left(\frac{1.2}{\sigma}\right)\right) \\
&\approx 0.025 \cdot 0.61135 + 0.475 \cdot 0.20420 - 0.07 \cdot (0.95516 - 0.61135) + 0.5 \cdot 0.04484 \\
&\approx 0.110637
\end{aligned}$$

For $\eta(\psi \cdot \varepsilon_i + (1-\psi) \cdot Z_i)$, let $V_i = \psi \cdot \varepsilon_i + (1-\psi) \cdot Z_i$, such that

$$V_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_V^2 = 0.5(\psi^2 + (1-\psi)^2)).$$

$$
\begin{aligned}
E[\eta(V_i)] &= E[0.025 \cdot \mathcal{I}(V_i < 0.2) + (0.475(V_i - 0.2) + 0.025) \cdot \mathcal{I}(V_i \in [0.2, 1.2]) + 0.5 \cdot \mathcal{I}(1.2 < V_i)] \\
&= 0.025 \cdot P(V_i < 0.2) + 0.475 \cdot E(V_i \cdot I(V_i \in [0.2, 1.2])) \\
&\quad - 0.07 \cdot P(0.2 \leq V_i \leq 1.2) + 0.5 \cdot P(1.2 < V_i) \\
&= 0.025 \cdot \Phi\left(\frac{0.2}{\sigma_V}\right) + 0.475 \int_{0.2}^{1.2} \frac{V_i}{\sigma_V} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(V_i/\sigma_V)^2}{2}\right) dx \\
&\quad - 0.07 \left[\Phi\left(\frac{1.2}{\sigma_V}\right) - \Phi\left(\frac{0.2}{\sigma_V}\right)\right] + 0.5 \left[1 - \Phi\left(\frac{1.2}{\sigma_V}\right)\right].
\end{aligned}
$$

$$
\boxed{E(W_i) \approx 0.110637 + E(\eta(V_i))}.
$$

Table B.1: *Theoretical Expectations of W_i under Study 3 cases showcasing deviation from $E(W_i) = 0.221$.*

	$\psi = 0.0$	$\psi = 0.1$	$\psi = 0.2$	$\psi = 0.3$
$\alpha = 1.0$	0.221	0.221	0.203	0.196
$\alpha = 0.75$	0.221	0.209	0.198	0.189
$\alpha = 0.50$	0.221	0.207	0.198	0.183
$\alpha = 0.25$	0.221	0.205	0.189	0.177

Poisson Expected Sample Size for Uniform Parameters

To ensure that the expected sample size from the sampling design in Simulation 1 Study 3, the Uniform distribution parameters need to be specified per sample size. With Unif(a, b), set $a = 0$ and b to depend on the desired sample size n. By definition of Poisson sampling, the sample size is determined by

$$
n = \sum_{i \in U} \mathcal{I}(\text{Unif}(0, b) < W_i).
$$

We can solve for b as follows, per probability theory (Blitzstein & Hwang, 2015):

$$
E(n) = \sum_{i \in U} P(\text{Unif}(0, b) < W_i) = \sum_{i \in U} \frac{W_i}{b} = \frac{1}{b} \sum_{i \in U} W_i
$$

$$
E(n) = \frac{1}{b} \sum_{i \in U} W_i \Rightarrow \boxed{b = \frac{1}{E(n)} \sum_{i \in U} W_i}.
$$

C

Simulation 1 Revised Weight Function

The procedure for defining weights W_k in Wang *et al.* (2023) after obtaining their sample S is not entirely clear. For Study 1: Pfeffermann & Sverchkov (1999) Adaptation and Study 2: Quadratic Weight Generating Function, weights W_i were generated as a continuous variable to be used for Probability Proportional to Size. In Study 3: Wu & Fuller (2005) Adaptation, weights W_i were initially defined as the generated inclusion probabilities for the ith population unit to be used for Poisson sampling. Since their generated weights were not presumed to be defined as the number of observations that the ith population unit represents in the population, W_i was redefined as the inverse of the inclusion probabilities, yet Wang *et al.* (2023) does not explicitly denote this step.

The following tables denote the initial definitions of W_i as Wang *et al.* (2023) denote where $\vec{W}$ are not explicitly the inverse of the inclusion probabilities for the sampled units.

Table C.1: *Replication of Wang et al. (2023) study 2 empirical rejection rates of ten tests with $\vec{W}$ is quadratic in $\vec{Y}$ based on 1000 replicates and 8 case scenarios presuming weights are not explicitly inverse of inclusion probabilities.*

n	α	*DD*	*PN*	*HP*	*PS1*	*PS1q*	*PS2*	*PS2q*	*PS3*	*WF*	*LR*
100	0.0	4.6	34.2	4.0	4.0	6.0	8.6	5.1	4.9	4.5	0.0
	0.5	66.1	34.9	65.2	63.0	61.9	74.6	70.4	76.2	77.5	1.4
	1.0	34.9	34.9	33.9	33.5	55.3	64.5	41.1	11.0	14.0	0.0
	1.5	100.0	39.7	100.0	100.0	100.0	100.0	100.0	83.0	98.6	3.3
200	0.0	5.7	33.2	5.5	5.6	9.7	9.2	5.2	5.7	5.4	0.4
	0.5	94.4	35.7	94.1	94.4	93.3	96.5	96.3	98.1	98.2	1.6
	1.0	66.1	35.9	65.1	64.3	87.0	92.9	71.8	23.1	25.6	0.0
	1.5	100.0	39.0	100.0	100.0	100.0	100.0	100.0	98.5	100.0	13.7

Note: Rejection rates were determined at the $\alpha = 0.05$ significance level where rates are the percentage of tests rejecting the null hypothesis of noninformative weights.

Table C.2: *Replication of Wang et al. (2023) study 1 empirical rejection rates of ten tests with $\vec{W}$ is linear in $\vec{Y}$ based on 1000 replicates and 32 case scenarios presuming weights are not explicitly inverse of inclusion probabilities.*

n	σ	δ	α	*DD*	*PN*	*HP*	*PS1*	*PS1q*	*PS2*	*PS2q*	*PS3*	*WF*	*LR*
100	0.1	1.5	0.0	5.3	34.4	5.1	5.0	5.1	4.4	5.2	5.4	4.5	0.8
			0.2	5.6	34.4	4.9	5.5	4.7	7.1	5.7	5.3	5.6	2.9
			0.4	10.2	33.3	9.3	8.6	7.6	11.4	10.2	11.4	12.0	8.9
			0.6	21.6	35.1	20.8	18.9	16.7	20.5	20.0	20.8	22.0	13.5
		1.0	0.0	4.9	31.1	4.6	4.1	4.5	4.5	4.6	5.2	5.0	0.1
			0.2	8.1	32.8	7.4	6.9	5.3	7.8	7.4	8.1	9.2	1.2
			0.4	17.7	34.0	16.9	13.6	12.4	17.3	16.2	18.7	20.3	4.9
			0.6	40.0	33.1	39.2	34.9	29.4	39.5	39.2	42.4	44.6	11.6
	0.2	1.5	0.0	3.9	35.5	3.7	5.0	4.3	4.4	3.9	4.2	4.4	1.2
			0.2	9.5	33.2	9.4	8.3	6.8	9.1	9.8	10.7	11.5	2.5
			0.4	30.5	34.4	30.0	26.0	23.2	30.4	30.5	33.3	35.0	7.4
			0.6	64.6	33.1	63.6	58.3	50.9	65.0	65.0	65.9	66.4	14.4
		1.0	0.0	5.0	33.0	4.9	4.7	4.3	4.5	5.2	5.3	6.1	0.0
			0.2	20.1	37.7	19.4	16.1	14.0	20.0	19.7	22.4	23.5	1.2
			0.4	62.6	33.9	61.3	56.5	49.7	63.3	62.3	62.7	64.0	6.8
			0.6	94.7	34.7	94.4	93.0	91.2	94.7	94.2	94.4	94.6	16.4
200	0.1	1.5	0.0	5.0	35.5	4.9	4.7	4.5	4.8	5.4	4.4	4.3	2.5
			0.2	8.6	32.5	8.4	6.6	6.1	8.7	8.3	9.5	9.3	7.8
			0.4	19.2	32.6	18.5	17.1	15.7	18.8	18.2	20.5	20.7	14.5
			0.6	39.9	30.3	39.1	34.9	30.0	40.0	40.0	40.5	41.8	21.8
		1.0	0.0	4.5	36.4	4.4	3.6	3.5	4.3	4.2	4.3	4.2	0.1
			0.2	11.9	34.7	11.5	10.0	8.9	11.7	12.5	12.3	12.2	4.5
			0.4	36.1	33.0	35.4	30.1	26.0	36.9	36.8	39.1	39.0	10.6
			0.6	73.0	35.4	72.3	64.9	60.3	73.7	72.8	73.4	74.6	22.8
	0.2	1.5	0.0	5.6	34.5	5.4	5.1	5.5	5.0	4.7	5.1	5.1	3.9
			0.2	16.3	33.2	16.1	14.7	13.8	16.5	16.6	18.0	18.8	8.5
			0.4	57.2	32.7	57.1	51.3	47.2	57.0	57.5	59.8	61.0	17.2
			0.6	90.0	32.9	89.8	86.8	84.9	90.3	89.9	91.1	91.4	25.6
		1.0	0.0	4.9	33.8	4.7	4.5	4.3	5.1	5.6	3.9	4.9	0.4
			0.2	35.4	36.9	34.9	29.4	25.5	34.4	34.5	38.1	38.2	5.6
			0.4	93.0	35.3	92.8	90.2	86.8	93.4	94.1	93.7	93.6	13.6
			0.6	99.8	34.2	99.8	99.6	99.5	99.8	99.8	99.9	99.9	26.8

Note: Rejection rates were determined at the $\alpha = 0.05$ significance level where rates are the percentage of tests rejecting the null hypothesis of noninformative weights.

Table C.3: *Replication of Wang et al. (2023) study 3 empirical rejection rates of ten tests based on 1000 replicates and 32 case scenarios presuming weights are not explicitly inverse of inclusion probabilities.*

$E(n)$	α	ψ	*DD*	*PN*	*HP*	*PS1*	*PS1q*	*PS2*	*PS2q*	*PS3*	*WF*	*LR*
100	1.00	0.0	4.4	34.9	4.1	4.5	55.2	20.3	4.9	2.0	4.3	0.0
		0.1	13.0	37.1	11.9	11.3	66.4	31.3	15.2	2.9	7.0	0.0
		0.2	41.8	34.0	41.0	34.3	84.0	57.9	44.1	3.0	6.4	0.0
		0.3	81.4	39.1	80.7	75.3	97.1	88.4	87.2	2.7	10.0	0.0
	0.75	0.0	5.1	34.5	4.7	3.0	17.6	9.3	4.4	4.5	4.7	0.0
		0.1	12.3	33.8	11.7	12.0	29.4	19.4	14.2	5.4	5.7	0.0
		0.2	44.6	39.4	43.5	37.0	59.1	50.7	47.3	8.0	8.7	0.0
		0.3	90.8	39.8	90.3	86.0	93.0	93.1	91.9	8.8	11.8	0.0
	0.50	0.0	3.6	36.3	3.4	2.7	4.3	3.8	3.1	4.4	4.5	0.0
		0.1	11.3	37.6	10.4	9.7	11.8	13.0	10.9	5.7	5.9	0.0
		0.2	49.8	41.3	48.6	41.0	43.3	50.9	52.5	7.2	8.4	0.0
		0.3	94.1	41.1	93.8	91.0	87.6	93.9	93.8	10.4	13.9	0.0
	0.25	0.0	4.1	35.7	4.1	4.6	4.6	4.7	3.9	4.7	4.6	0.0
		0.1	13.7	33.4	13.0	11.4	10.6	13.4	13.7	4.8	5.7	0.0
		0.2	49.6	39.2	48.2	41.3	38.0	49.5	49.3	8.2	10.1	0.0
		0.3	93.8	41.8	93.7	90.9	87.2	93.9	93.4	8.9	12.5	0.0
200	1.00	0.0	4.6	34.0	4.3	3.5	89.0	38.7	5.7	2.3	6.6	0.0
		0.1	18.9	35.7	18.8	15.8	92.1	54.6	21.8	1.4	4.8	0.0
		0.2	73.5	38.0	72.7	64.3	99.3	87.7	76.6	3.0	7.6	0.0
		0.3	99.2	42.1	99.2	98.0	100.0	99.3	99.4	2.0	9.4	0.0
	0.75	0.0	6.0	35.6	5.6	5.8	38.5	15.9	5.5	4.0	5.1	0.0
		0.1	23.7	37.4	23.4	18.8	55.9	34.0	23.5	5.3	6.8	0.0
		0.2	80.3	38.3	79.9	73.3	90.4	86.8	82.0	7.0	10.5	0.0
		0.3	99.8	41.8	99.8	99.9	100.0	99.9	99.9	10.1	18.3	0.0
	0.50	0.0	5.6	34.2	5.3	4.8	8.8	6.6	5.0	6.3	5.3	0.0
		0.1	22.3	35.8	22.1	19.2	24.3	25.9	24.2	5.9	6.3	0.0
		0.2	81.4	39.9	80.9	74.7	75.8	82.8	82.5	9.9	10.4	0.0
		0.3	100.0	46.3	99.9	99.9	99.9	99.8	100.0	17.7	21.3	0.0
	0.25	0.0	4.3	36.7	4.1	5.0	5.8	5.3	4.4	5.7	5.5	0.0
		0.1	20.8	38.8	20.7	17.2	17.5	20.9	21.0	7.7	8.3	0.0
		0.2	80.8	39.3	80.4	73.8	70.1	81.0	80.1	8.4	9.2	0.0
		0.3	100.0	49.9	100.0	99.8	99.8	100.0	99.9	14.8	19.0	0.0

Note: Rejection rates were determined at the $\alpha = 0.05$ significance level where rates are the percentage of tests rejecting the null hypothesis of noninformative weights.

D

Simulation 1 Increased Iterations

The convergence of the simulation results in Wang *et al.* (2023) is important for comparing the tests so increasing the number of simulation iterations will provide more concrete comparisons between the tests. Since the simulations are not too computationally expensive, the following tables are the simulation studies in Simulation Study 1: Wang *et al.* (2023) with $B = 10000$ instead of 1000.

Table D.1: *Replication of Wang et al. (2023) study 2 empirical rejection rates of ten tests with $\vec{W}$ is quadratic in $\vec{Y}$ based on 10000 replicates and 8 case scenarios.*

n	α	*DD*	*PN*	*HP*	*PS1*	*PS1q*	*PS2*	*PS2q*	*PS3*	*WF*	*LR*
100	0.0	5.0	36.2	4.7	7.1	8.0	6.9	6.7	3.5	5.0	51.8
	0.5	53.7	34.8	52.6	27.0	29.9	36.0	32.6	66.2	70.2	56.4
	1.0	19.0	35.0	18.2	11.7	16.8	26.1	13.1	6.2	8.8	63.8
	1.5	100.0	37.1	100.0	87.9	99.1	98.3	92.5	86.8	92.0	57.7
200	0.0	5.3	36.4	5.1	7.3	9.4	8.5	7.6	4.2	5.4	43.0
	0.5	86.5	35.6	86.2	46.9	52.3	58.2	54.5	95.2	95.8	53.5
	1.0	40.5	36.3	39.9	23.0	46.1	64.5	29.2	11.3	14.9	63.2
	1.5	100.0	39.4	100.0	98.6	100.0	100.0	99.5	97.8	99.6	52.7

Note: Rejection rates were determined at the $\alpha = 0.05$ significance level where rates are the percentage of tests rejecting the null hypothesis of noninformative weights.

Table D.2: *Replication of Wang et al. (2023) study 1 empirical rejection rates of ten tests with $\vec{W}$ is linear in $\vec{Y}$ based on 10000 replicates and 32 case scenarios.*

n	σ	δ	α	*DD*	*PN*	*HP*	*PS1*	*PS1q*	*PS2*	*PS2q*	*PS3*	*WF*	*LR*
100	0.1	1.5	0.0	4.8	35.3	4.5	6.9	7.0	5.3	6.4	4.0	4.9	51.3
			0.2	6.2	34.2	5.9	8.0	8.9	8.8	8.6	5.6	6.4	52.2
			0.4	10.3	33.8	9.8	11.3	13.0	14.3	13.0	10.7	11.6	50.9
			0.6	18.4	34.0	17.7	18.2	20.8	24.7	21.0	19.2	21.6	50.9
		1.0	0.0	5.3	35.5	5.0	7.6	7.9	6.5	7.6	4.5	5.1	51.7
			0.2	7.6	33.5	7.3	10.1	12.4	12.9	11.1	7.8	9.0	51.1
			0.4	18.8	34.2	18.1	18.5	22.8	26.7	21.2	19.8	21.9	51.0
			0.6	38.4	33.4	37.2	32.8	41.1	48.1	37.8	39.4	43.1	51.8
	0.2	1.5	0.0	5.0	35.7	4.7	6.9	7.1	6.1	6.3	4.0	5.3	51.3
			0.2	9.3	34.1	8.9	10.7	11.4	12.3	11.7	9.2	10.7	52.0
			0.4	28.4	33.9	27.3	24.4	24.6	29.9	28.0	29.2	32.7	52.1
			0.6	57.8	34.1	56.6	48.2	47.9	57.1	54.1	57.8	61.6	51.3
		1.0	0.0	4.9	35.7	4.6	7.4	8.0	6.7	6.8	4.3	5.2	51.7
			0.2	17.1	34.3	16.4	16.5	18.2	20.8	18.7	18.1	20.7	53.2
			0.4	58.6	34.5	57.6	48.8	50.1	57.9	54.4	60.5	64.4	51.3
			0.6	93.1	33.9	92.8	85.2	86.2	91.1	89.5	92.7	93.7	50.6
200	0.1	1.5	0.0	5.2	35.9	5.0	6.8	7.0	5.8	6.8	4.7	4.9	44.9
			0.2	7.4	34.7	7.2	8.9	10.4	10.5	9.8	7.5	8.1	47.4
			0.4	16.3	34.0	16.0	15.6	18.4	22.0	18.3	17.5	18.9	47.9
			0.6	33.6	33.6	33.2	28.9	34.2	41.4	33.7	36.2	38.0	47.7
		1.0	0.0	4.8	35.4	4.7	7.7	8.7	7.0	7.9	4.6	4.7	44.9
			0.2	10.8	34.0	10.6	12.8	16.9	18.2	14.4	12.2	13.0	48.5
			0.4	34.6	34.4	34.1	29.9	38.5	44.8	33.7	36.8	39.1	47.5
			0.6	69.5	33.3	69.0	58.3	70.3	76.5	64.9	70.8	73.1	48.5
	0.2	1.5	0.0	4.8	35.9	4.7	6.8	7.1	5.9	6.3	5.0	5.1	44.9
			0.2	14.8	34.2	14.5	14.3	14.9	17.5	16.4	15.9	17.9	46.7
			0.4	53.0	34.9	52.4	44.0	44.1	51.8	48.9	55.6	58.2	46.8
			0.6	90.1	35.0	89.9	81.8	81.6	87.2	86.3	90.3	91.4	47.1
		1.0	0.0	4.9	36.3	4.8	7.4	8.3	7.1	7.2	5.1	5.4	45.3
			0.2	31.3	34.4	30.9	26.5	29.8	34.1	30.3	33.9	36.7	48.1
			0.4	90.0	34.7	89.8	80.7	82.9	87.2	85.8	91.0	91.9	46.9
			0.6	99.9	35.4	99.9	99.5	99.7	99.8	99.8	99.9	99.9	46.8

Note: Rejection rates were determined at the $\alpha = 0.05$ significance level where rates are the percentage of tests rejecting the null hypothesis of noninformative weights.

Table D.3: *Replication of Wang et al. (2023) study 3 empirical rejection rates of ten tests based on 10000 replicates and 32 case scenarios.*

$E(n)$	α	ψ	*DD*	*PN*	*HP*	*PS1*	*PS1q*	*PS2*	*PS2q*	*PS3*	*WF*	*LR*
100	1.00	0	5.0	37.2	4.6	9.6	12.3	8.6	8.6	2.1	4.9	2.0
		0.1	8.6	38.8	8.2	15.3	18.2	12.9	15.9	1.8	4.8	2.9
		0.2	25.0	40.2	24.1	35.4	38.6	30.9	36.5	2.0	4.7	4.2
		0.3	58.4	41.5	57.1	68.5	70.4	63.2	69.8	1.6	4.7	6.9
	0.75	0	5.0	38.7	4.7	7.5	9.5	7.2	7.0	3.0	4.8	2.0
		0.1	9.1	38.5	8.7	12.8	16.1	10.9	13.6	3.0	5.0	3.2
		0.2	30.9	41.0	29.9	34.8	38.6	32.1	38.0	2.8	5.2	5.7
		0.3	70.5	44.2	69.4	72.3	74.8	70.3	75.9	2.6	5.1	10.8
	0.50	0	4.9	39.6	4.6	5.9	6.6	5.3	5.7	3.5	5.1	2.2
		0.1	10.8	39.7	10.2	11.2	12.8	10.4	12.8	3.8	5.1	3.7
		0.2	36.1	42.5	35.2	33.9	35.6	34.5	39.6	3.8	4.7	6.7
		0.3	81.3	47.6	80.6	75.5	76.7	77.9	81.4	4.4	5.8	14.0
	0.25	0	5.0	39.5	4.7	5.0	5.0	4.3	4.9	4.1	5.0	2.7
		0.1	11.3	41.0	10.7	9.6	9.3	10.0	11.2	4.4	5.4	3.9
		0.2	41.9	43.9	40.9	33.3	32.2	38.3	40.0	4.1	5.2	8.1
		0.3	87.8	51.1	87.3	79.5	77.5	85.3	85.6	5.0	6.2	16.5
200	1.00	0	5.2	38.2	5.0	9.4	15.7	9.5	8.5	2.0	5.2	6.8
		0.1	13.4	38.0	13.1	20.6	27.0	16.4	21.2	1.9	4.8	8.5
		0.2	47.2	41.3	46.7	57.6	62.0	50.9	58.7	1.8	5.1	11.7
		0.3	88.6	47.4	88.3	91.8	93.1	89.8	92.8	1.8	4.7	21.0
	0.75	0	4.9	38.5	4.7	7.6	14.0	8.4	7.4	3.1	5.0	6.5
		0.1	14.3	40.5	14.0	18.1	25.5	15.2	19.7	3.0	4.7	9.4
		0.2	56.6	43.7	56.0	58.2	64.9	55.2	62.3	2.5	5.0	14.4
		0.3	95.2	50.6	95.1	94.9	96.0	94.8	96.4	2.7	4.4	26.0
	0.50	0	4.7	39.0	4.5	5.9	9.8	6.4	5.5	3.7	5.1	6.9
		0.1	16.8	40.0	16.5	16.5	22.7	16.1	19.0	4.0	5.0	9.8
		0.2	65.8	45.3	65.3	59.7	65.0	61.9	65.9	4.2	4.9	16.1
		0.3	98.6	54.2	98.5	96.9	97.6	98.0	98.3	4.0	5.3	31.8
	0.25	0	5.1	40.3	4.9	4.9	5.2	4.8	4.9	4.6	5.3	6.4
		0.1	17.9	41.3	17.5	14.9	15.4	16.1	17.7	4.4	5.4	9.6
		0.2	73.8	48.5	73.4	63.3	62.8	70.2	70.8	4.8	5.5	18.0
		0.3	99.6	61.3	99.6	98.7	98.6	99.3	99.4	4.5	5.7	34.8

Note: Rejection rates were determined at the $\alpha = 0.05$ significance level where rates are the percentage of tests rejecting the null hypothesis of noninformative weights.

E

Consumer Expenditure Wrangling

The dataset CE from the rpms R package by Toth (2021) was revised to optimize the performance of the simulation and to build a reasonable simulation design to sample and perform the survey weight diagnostic tests.

```
library(tidyverse)
library(rpms)

ce = rpms::CE %>%
  filter(TOTEXPCQ > 0, FINCBTAX > 10, SALARYX > 0, !is.na(REGION),
         FAM_SIZE %in% factor(1:10), ROOMSQ %in% factor(1:11),
         NO_EARNR %in% factor(1:4)) %>%
  mutate(TOTEXPCQ = log(TOTEXPCQ), FINCBTAX = log(FINCBTAX)) %>%
  select(-c(QINTRVMO, PSU, INCNONWK, IRAX, LIQUIDX, STOCKX, STUDNTX,
            FOOTWRCQ, TOBACCCQ, TOTXEST, VEHQL, EARNER)) %>%
  group_by(REGION, MARITAL) %>%
  filter(n() >= 70) %>%
  ungroup()
```

Listing E.1: *Data wranling of* **rpms** *Consumer Expenditure* **CE** *dataset.*

- **Filtering Observations**:
  - Ensured sufficient data for each group for clustering/stratifying variables FAM_SIZE, REGION, ROOMSQ, NO_EARNR. For Two-Stage Clustering and Stratified Sampling, ssus were filtered out if there were not at least 70 observations.
  - Bounded the range of quantitative variables TOTEXPCQ, FINCBTAX, and SALARYX for reasonable regression estimates during simulation.
- **Transformations**:
  - Natural logarithm transformed FINCBTAX and TOTEXPCQ to get an empirical Normal distribution, since the data are highly skewed to the right.
- **Dropping Variables**:
  - Dropped variables with no relevance to simulation design, mostly missing data, or constructions of other variables.

F

PPS Scaled Z_i Derivation

Expectation of Z_i

Knowing $E(\varepsilon_i) = 0$, values of $\vec{X}_{i \in U}$, and $X_i \perp\!\!\!\perp \varepsilon_i$, we get from $Z_i = X_i \cdot (1 + \varepsilon_i)$ to

$$\begin{aligned}
E(Z_i) &= E(X_i + X_i \varepsilon_i) \\
&= E(X_i) + E(X_i \varepsilon_i) \\
&= E(X_i) + E(X_i)E(\varepsilon_i), \text{ by } X_i \perp\!\!\!\perp \varepsilon_i \\
&= E(X_i), \text{ by } E(\varepsilon_i) = 0 \\
&= X_i, \text{ since } X_i \text{ is known.}
\end{aligned}$$

Variance of Z_i

$$\begin{aligned}
\mathrm{Var}(Z_i) &= \mathrm{Var}(X_i \cdot (1 + \varepsilon_i)) \\
&= \mathrm{Var}(X_i) + \mathrm{Var}(X_i \varepsilon_i) + 2\mathrm{Cov}(X_i, X_i \varepsilon_i) \\
&= 0 + \mathrm{Var}(X_i \varepsilon_i) + 2(0) \\
&= \mathrm{Var}(X_i \varepsilon_i) \\
&= \mathrm{Var}(E(X_i \varepsilon_i \mid X_i)) + E(\mathrm{Var}(X_i \varepsilon \mid X_i)) \\
&= \mathrm{Var}(X_i E(\varepsilon_i \mid X_i)) + E(X_i^2 \mathrm{Var}(\varepsilon_i \mid X_i)) \\
&= \mathrm{Var}(X_i E(\varepsilon_i)) + E(X_i^2 \mathrm{Var}(\varepsilon_i)) \\
&= E(\varepsilon_i)^2 \mathrm{Var}(X_i) + E(X_i^2)\mathrm{Var}(\varepsilon_i) \\
&= 0 + E(X_i^2)\mathrm{Var}(\varepsilon_i) \\
&= X_i^2 \mathrm{Var}(\varepsilon_i).
\end{aligned}$$

This page intentionally left blank.

Harvard University

Evaluation of Survey Weight Diagnostic Tests in Regressions with Complex Survey Sampling

Corbin Craig Lubianski

Harvard College
Cambridge, Massachusetts

April 2024