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
MARCH 2024

This page intentionally left blank.

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Keywords: Keyword A, Keyword B, Keyword C.

This page intentionally left blank.

Acknowledgements

This page intentionally left blank.

Contents

1 Introduction

Welcome to the introduction of your dissertation. The introduction of a dissertation serves as a critical component, setting the tone and laying the foundation for the entire research endeavour. It is tasked with providing a clear and concise overview of the research topic, elucidating the context and significance of the study within the broader academic landscape. A well-crafted dissertation introduction should delineate the research problem or question, offering a rationale for its relevance and addressing any existing gaps in knowledge. Furthermore, it typically outlines the objectives and aims of the study, guiding the reader through the anticipated contributions and outcomes. In addition, the introduction often encapsulates the methodology employed, presenting the chosen approach and rationale behind it. Lastly, it functions as a road-map, offering a brief glimpse into the structure and organisation of the dissertation, thereby orienting the reader and facilitating comprehension of the subsequent chapters. Overall, a dissertation introduction should engage the reader's interest, provide a clear framework for the research, and justify its importance in the academic realm. For a clearer and more accessible readability in referencing chapters, refer to the chapter titled **??** (referred to as **??**).

Chapter 1

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 errors 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: $\text{Cov}(\varepsilon_i, \varepsilon_j \mid \mathbf{X_i}, \mathbf{X_j}) = 0$, 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 follow all four main assumptions. Specifically, observations may have different inclusion probabilities π_i as 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 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 N. The observed survey data S is $\{Y_i, X_i, W_i\}_{i \in S}$ where W_i is the survey weight associated with the ith observation unit which does not necessarily have to be the inverse of the selection probability. A model for the sample S is

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

where $\vec{Y} = (Y_1, \dots, Y_n)^\top$ is a vector of response variables $n \times 1$, $\mathbf{X} = (X_1^\top, \dots, X_p^\top)^\top$ is a $n \times p$ matrix of the explanatory variables (including component 1 for calculating the intercept), β is a $p \times 1$ vector of regression coefficients, and ε is a $1 \times n$ vector of regression errors.

For the observed survey data, the least squares estimators for β are

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

$$\hat{\beta}_w = \frac{\sum_{i \in S} w_i \vec{x}_i y_i}{\sum_{i \in S} w_i \vec{x}_i^\top \vec{x}_i} = \frac{\mathbf{X}^\top \mathbf{H} \vec{Y}}{\mathbf{X}^\top \mathbf{H} \mathbf{X}}, \text{ where } H = \text{diag}(\vec{W}).$$

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 β. 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.

> Survey weight diagnostic tests are only meant to be used as a determinant of 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 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

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

Hausman (1978) create the basis of the DC test as a test for general misspecifications. Hausman proposed 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).

2.2.1 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 $\hat{V} = \hat{V}(\hat{\beta}_w) - \hat{V}(\hat{\beta}_u)$.

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, $\text{Cov}(\hat{\beta}_u, \hat{\beta}_w - \hat{\beta}_u) = 0$. Unfortunately, this estimator is not necessarily positive-definite, especially for small and moderate sample sizes when $\hat{\beta}_w$ will inflate as noted within the literature.

TO-DO: For the Hausman-Pfeffermann DC test rate to obtain a negative variance estimate, visit Appendix A.

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

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

where C is an estimator of the covariance matrix of the two estimators as

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

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

The proposed estimator of V is positive definite, even for small sample sizes (Asparouhov & Muthen, 2007). However, C can be difficult to compute if the standard linear regression assumptions do not hold for some 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 (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 to compute β_u and the other with unequal inclusion probability weights β_w. Then, the unweighted copy covariates $\mathbf{x_i}^\top$ are replaced by $\mathbf{x_i}^\top \mathbf{x_i}^\top$ and the weighted copy is $\mathbf{x_i}^\top \mathbf{0}^\top$. Finally, running a linear regression to obtain the regression coefficients $\mathbf{d} = (\beta_u, \beta_w - \beta_u)^\top$ is simple with design-based statistical software (Kott, 2018).

Wang-Wang-Yan Estimator Adjustment

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

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

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 non-informative weights.

Steps for performing the Hausman-Pfeffermann DC Test with Wang-Wang-Yan variance estimator, given $\{Y_i, \vec{X}_i, W_i\}_{i \in S}$:

1. Calculate $\beta_u = (\mathbf{X}^\top \mathbf{X})^{-1}(\mathbf{X}^\top \vec{Y})$.
2. With $\mathbf{H} = \text{diag}(\vec{W})$, calculate $\beta_w = (\mathbf{X}^\top \mathbf{H}\mathbf{X})^{-1}(\mathbf{X}^\top \mathbf{H}\vec{Y})$.
3. Compute $\hat{\sigma}^2 = (n - p + 1)^{-1} \sum_{i=1}^{n} \varepsilon_i$ where $\varepsilon_i = Y_i - \vec{X}_i^\top \hat{\beta}_u$.
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 $T = (\hat{\beta}_w - \hat{\beta}_u)^\top \hat{V}^{-1}(\hat{\beta}_w - \hat{\beta}_u)$.
6. Determine p-value with $T \sim \chi_p^2$.

2.3 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 β_M in equation

$$Y = X\beta + X_M \beta_M + \varepsilon$$

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. Many WA tests require the normality assumption for ε_i to perform F tests, which is not assumed as in DC tests.

2.3.1 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. With regard to weights, a WA test checks whether

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

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

Within this context, consider the regression

$$\vec{Y} = \mathbf{X}\beta_u + \mathbf{X}_w\beta_w + \vec{\varepsilon}.$$

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

Steps for performing the DuMouchel-Duncan WA Test, given $\{Y_i, \vec{X}_i, W_i\}_{i \in S}$:

1. Create $\tilde{\mathbf{X}} = \mathbf{H}\mathbf{X}$, 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 as

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

5. Calculate p-value with

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

2.3.2 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

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

For a given k, the sample correlation after Fisher transformation follows a Normal distribution asymptotically. Although the range of k is not specified, the first 2-3 correlations are sufficient to test the null hypothesis.

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

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

with intercept α and slope coefficient $\beta^{(k)}$. For a given k, perform a t-test with $H_{0,k}$: $\beta^{(k)} = 0$. For any of k t-tests, a statistically significant p-value is sufficient to reject the null hypothesis of non-informative weights for the model. Pfeffermann & Sverchkov (1999) report that the two variations of the WA test have similar performance.

Wang-Wang-Yan Adjustment

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

$$E(\vec{W} \mid \hat{\varepsilon}_u) = f(\mathbf{X};\eta) + \sum_{k=1}^{2}\beta^{(k)}\hat{\varepsilon}_u^k + \mathrm{diag}(\hat{\varepsilon}_u)X\gamma,$$

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 $\hat{\varepsilon}$. 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 $\{Y_i, \vec{X}_i, W_i\}_{i\in S}$:

1. Compute the unweighted regression $E\left(\vec{Y} \mid \mathbf{X}\right)$ and calculate the residuals $\hat{\varepsilon}_u = \vec{Y} - \mathbf{X}\hat{\beta}_u$.
2. Construct the full model matrix $\mathbf{X}_{full} = [\mathbf{X}, \hat{\varepsilon}, \hat{\varepsilon}^2, \tilde{\mathbf{X}}]$ with $\tilde{\mathbf{X}} = \mathrm{diag}(\varepsilon)\mathbf{X}$. 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 summing the squared differences between $\hat{\vec{W}}$ and $\vec{W}$.
5. Compute test statistic T as
$$T = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/(p_{\text{full}} - p_{\text{reduced}})}{SSE_{\text{full}}/(n - p_{\text{full}} - 1)}.$$
6. Calculate p-value with
$$T \sim F_{df_{\text{reduced}} - df_{\text{full}},\, df_{\text{full}}}.$$

2.3.3 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 whether the weight is informative 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

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

where $f(\mathbf{X}; \eta)$ is a function of $\mathbf{X}$ with parameter η, coefficient γ_k of $\vec{Y}^k$. 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 $\{Y_i, \vec{X}_i, W_i\}_{i \in S}$:

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 as
$$T = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/(p_{\text{full}} - p_{\text{reduced}})}{SSE_{\text{full}}/(n - p_{\text{full}} - 1)}.$$
5. Calculate p-value with
$$T \sim F_{df_{\text{reduced}} - df_{\text{full}},\, df_{\text{full}}}.$$

2.3.4 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 $\widetilde{\mathbf{X}} = \mathbf{QX}$, $\mathbf{Q} = \text{diag}(q_1, q_2, \ldots, q_n)$, and $q_i = w_i \hat{w}_i^{-1}(x_i)$ where $\hat{w}_i$ is estimated by regressing of w_i on $f(x_i; \eta)$.

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 $H_0 : \gamma = 0$ determines the impact of $\vec{W}$ on $\vec{Y}$ after removing the information contained in $\mathbf{X}$ as q_i are the predictable factors of weight W_i by 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 $\{Y_i, \vec{X}_i, W_i\}_{i \in S}$:

1. Compute the regression of $E(\vec{W} \mid \mathbf{X}) = f(\mathbf{X}; \eta)$ and estimate $\hat{w}_i$ for $i \in S$.
   - Reasonable choices for $f(\mathbf{X}; \eta)$ may include linear and quadratic relationships.
2. With $\mathbf{Q} = \text{diag}(\vec{q})$, create $\widetilde{\mathbf{X}} = \mathbf{QX}$.
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 as
$$T = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/(p_{\text{full}} - p_{\text{reduced}})}{SSE_{\text{full}}/(n - p_{\text{full}} - 1)}.$$
7. Calculate p-value with
$$T \sim F_{df_{\text{reduced}} - df_{\text{full}},\, df_{\text{full}}}.$$

2.4 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 rely

on. 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 overlap is a crude diagnostic test. Schenker & Gentleman (2001) recommend to use confidence intervals only when more formal DC tests are not available.

2.4.1 Pfeffermann-Sverchkov Estimation Test

Pfeffermann & Sverchkov (2003) propose a test that uses the 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

$$\delta_i(\beta) = \vec{X}_i(Y_i - \vec{X}_i^{\top}\beta), i \in S.$$

Define $\hat{W}_i$ as the fitted value of the regression, $q_i = W_i/\hat{W}_i$, and $R(\vec{X}_i;\beta) = \delta_i(\beta) - q_i\delta_i(\beta)$. Thus, the null hypothesis is $H_0 : E(R(\vec{X}_i;\beta)) = 0$. The sampling weight means $E(R(\vec{X}_i;\beta))$ 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 $R(\vec{X}_i;\hat{\beta}_u)$ with $i \in S$. 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 $f(\mathbf{X};\eta)$ 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 $\{Y_i, \vec{X}_i, W_i\}_{i \in S}$:

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 $\hat{W}_i$ from the estimates fitted with the auxiliary regression and calculate $q_i = W_i / \hat{W}_i$.
3. Estimate β from regressing $\vec{Y}$ on $\mathbf{X}$ and estimate the fitted $\hat{Y}_i$.
4. Use the unweighted estimation function $\delta_i(\hat{\beta})$ for $i \in S$ to compute $R(\vec{X}_i; \hat{\beta}) = \delta_i(\hat{\beta}) - q_i \delta_i(\hat{\beta})$.
5. Compute test statistic T as
$$\frac{n-p}{p} \bar{R}_n^{-\intercal} \hat{\Sigma}_{R,n}^{-1} \bar{R}_n.$$
6. Calculate p-value with $T \sim F_{p,\, n-p}$.

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

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

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 $D_i = v_{u,i}^2 - v_{w,i}^2$.

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 a "training" 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 $\{Y_i, \vec{X}_i, W_i\}_{i \in S}$:

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 $E(Y_i \mid \vec{X}_i), i \in E$ to obtain $\hat{\beta}_u$. With the regression coefficient estimates, fit the unweighted regression onto the validation set V and compute the prediction errors $v_{u,i} = Y_i - \hat{Y}_i, i \in V$.
3. Compute the weighted linear regression of $E(Y_i \mid \vec{X}_i, W_i), i \in E$ 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 $v_{w,i} = Y_i - \hat{Y}_i, i \in V$.
4. With $D_i = v_{u,i}^2 - v_{w,i}^2$, 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 $E(D) = 0$.

2.4.3 Breidt Likelihood-Ratio Test

Breidt *et al.* (2013) formally proposed 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 $\vec{\omega}$ is

$$l(\theta; \vec{\omega}) = \sum_{i \in S} \omega_i \log(f(Y_i \mid \vec{X}_i; \theta)).$$

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 U and n is the size of sample S. (Herndon, 2014)

Let $\hat{\theta}_u = \operatorname{argmin}_\theta l(\theta; \vec{\omega}_u)$ and $\hat{\theta}_w = \operatorname{argmin}_\theta l(\theta; \vec{\omega}_w)$. 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

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

$$\hat{\sigma}^2 = N^{-1} \sum_{i \in S} (Y_i - \vec{\mathbf{X}}_i \hat{\beta})^2,$$

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

$$\vec{\beta} = \frac{\frac{\sum_{i \in S} W_i Y_i \cdot \sum_{i \in S} W_i \vec{X}_i}{\sum_{i \in S} W_i \vec{X}_i W_i} - \sum_{i \in S} W_i \vec{X}_i Y_i}{\frac{\sum_{i \in S} W_i \vec{X}_i \cdot \sum_{i \in S} W_i \vec{X}_i}{\sum_{i \in S} W_i \vec{X}_i W_i} - \sum_{i \in S} W_i \vec{X}_i^2} = \frac{\sum_{i \in S} W_i \frac{1}{\hat{\sigma}_i^2} \vec{X}_i Y_i}{\sum_{i \in S} W_i \frac{1}{\hat{\sigma}_i^2} \vec{X}_i \vec{X}_i^\intercal}$$

$$\hat{\sigma}^2 = \frac{\sum_{i \in S} W_i (Y_i - \vec{X}_i \vec{\beta})^2}{\sum_{i \in S} W_i}.$$

Let the information matrices be denoted as $J_u = \sum_{i \in S} \mathcal{I}(\vec{X}_i; \theta_0) = \mathcal{I}(\mathbf{X}; \theta_0)$, $J_w = \sum_{i \in S} W_i \mathcal{I}(\vec{X}_i; \theta_0)$, and $K_w = \sum_{i \in S} W_i^2 \mathcal{I}(\vec{X}_i; \theta_0)$ where $\mathcal{I}(\vec{X}_i; \theta_0)$ is the Fisher information for the ith observation with the true parameter θ_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 $\{Y_i, \vec{X}_i, W_i\}_{i \in S}$:

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

$$\log L\left(\vec{\beta}, \sigma^2 \mid \vec{Y}, \mathbf{X}, \vec{W}\right) = -\frac{1}{2} \log(2\pi\sigma^2) \sum_{i \in S} W_i - \frac{1}{2\sigma^2} \sum_{i \in S} W_i (Y_i - \vec{X}_i \vec{\beta})^2.$$

2. With maximum likelihood estimates $\vec{\theta}_u$ and $\vec{\theta}_w$, 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:

$$J_u = \text{diag}\left(\sum_{i \in S} \frac{\vec{X}_i \vec{X}_i^\top}{\hat{\sigma}^2}, \sum_{i \in S} \frac{1}{2n\hat{\sigma}^4}\right), \ J_w = \text{diag}\left(\sum_{i \in S} \frac{\vec{X}_i W_i \vec{X}_i^\top}{\hat{\sigma}^2}, \sum_{i \in S} \frac{W_i}{2n\hat{\sigma}^4}\right)$$

$$K_w = \text{diag}\left(\sum_{i \in S} \frac{\vec{X}_i W_i^2 \vec{X}_i^\top}{\hat{\sigma}^2}, \sum_{i \in S} \frac{W_i^2}{2n\hat{\sigma}^4}\right).$$

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.

3

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

As the first attempt to compare the phletora 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 is to reproduce the empirical results from Wang *et al.* (2023) and to suggest alterations to the simulation design to draw additional conclusions.

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 (Wang *et al.*, 2023). 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 Set Up — 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) 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 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 will 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: $N = 3000$

Results

Table 3.1 and Table C.1 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,

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	1 1.8	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.

respectively. For a well-performing a test, it 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 variable and lower power than other tests — likely due to dividing 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 C.1, 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 nearly identical across the 32 cases. PS1 is the least powerful test among the 10 tests. **TO-DO: Address LR issue in critique section**.

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 Set Up — 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) For all i, generate $Y_i = 1 + X_i + \varepsilon_i$.
   (c) For all i, 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: $N = 3000$
- $\sigma = 0.1$

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.

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.

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 are 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 between $\alpha = 0.5$ to 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 (denoted as study 2), 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

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

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 $W_i \in (0, 1]$:

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

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 A 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 $U_i \overset{iid}{\sim}$ 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 Set Up — 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) For all i, generate $Y_i = 0.5 + X_i + \varepsilon_i$.
   (c) For all i, generate the inclusion probabilities
$$W_i = \alpha \cdot \eta(X_i) + \beta \cdot \eta(\psi \cdot \varepsilon_i + (1-\psi) \cdot Z_i),$$
$$\text{with } \eta(x) = \begin{cases} 0.025, & x < 0.2 \\ 0.475(x-0.2) + 0.025, & 0.2 \leq x \leq 1.2 \\ 0.5, & 1.2 < x. \end{cases}$$
2. Using Poisson sampling of given $\vec{W}$, draw $U_i \overset{iid}{\sim}$ 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 A for explanation. Subsequently, redefine W_i to be the inverse of the selection probabilities where $W_i \longrightarrow \frac{1}{W_i}$.
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: $n \in \{100, 200\}$
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: $N = 3000$
- $\sigma^2 = 0.5$

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 sample size 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 3.3 and Table 3.3 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 3.3, PN, PS1q, and LR failed consistently to maintain their power when $\psi = 0$. As shown in Wang *et al.* (2023) results in Table 3.3, rejection rates increased as α decreased. However, the replication results hint 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 different results between the replication attempts and the simulation studies in Wang *et al.* (2023) are significantly different where 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, the following are speculations on how the differences of the studies were created.

Weights and Inclusion Probabilities

As noted in literature and statistical practice, survey weights $\vec{W}$ are generally the inverse of the selection probabilities $\vec{\pi}$ such that $W_i = \frac{1}{\pi_i}$. Within the replications, generated weights, unless otherwise specified, were interpreted as the inverse selection probabilities that were computed with the generation process.

- Study 1: Pfeffermann & Sverchkov (1999) Adaptation and Study 2: Quadratic Weight Generating Function: The generated weights W_i for $i = \{1, \ldots, N\}$ were interpreted to be a vector of generated data to then utilize the Probability Proportional to Size (PPS) procedure to get the inclusion probabilities π_i. With the inclusion probabilities, the sample was selected by using unequal probability selection. Subsequently, for elements k in the sample set S, weights were redefined to be $W_k = \pi_k^{-1}$. Although it was assumed that Wang *et al.* (2023) redefined W_k, it is not clear about this procedure and may be a factor in the difference between

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.

results for the adaption of the study in Pfeffermann & Sverchkov (1999) and the attempted replication.

- Study 3: Wu & Fuller (2005) Adaptation: The generated weights W_i for the population served as the inclusion probabilities of a population unit i being selected for the sample. For the sampling procedure, Wang *et al.* (2023) utilized the Poisson sampling procedure where population unit i will be selected if $U_i < W_i$ where $U_i \stackrel{iid}{\sim} \text{Unif}(0, 1)$ 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 $U_i \stackrel{iid}{\sim} \text{Unif}(0, n^{-1} \sum_i^N W_i)$. After selecting K units for the sample S , the weights were redefined to be $W_k \to W_k^{-1}$. Again, whether weights W_k used for the diagnostic tests are the inverse probabilities of the sampled units is unclear and may be a determinant in the difference of results.

For the rejection rates presuming that W_k was not redefined to be $W_k = \pi_k^{-1}$, refer to Appendix B for the replication rejection rates of studies 1, 2, and 3. **TO-DO**

Limited Iterations

For all three studies, Wang *et al.* (2023) set the simulated iterations $B = 1000$. While B may be sufficiently high to determine performances within and across 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 completely nontrival. As $B \to \infty$, the simulated rejection rates should define the true properties of the diagnostic tests given the simulation design. To determine the converging rejection rates, B was increased to 10000. Refer to Appendix C for the replication rejection rates of studies 1, 2, and 3 when $B = 10000$.

4 Simulation Study 2: CE Sampling

In contrast to the simulation studies in Wang *et al*. (2023), testing survey weight diagnostic on complex survey data is needed 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 regarding methods and weighting (U.S. Bureau of Labor Statistics, 2023).

Performing simulations on existing survey data has the advantage of testing the diagnostic tests on the 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, as 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 D 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 D 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 D 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

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

and the inclusion probability of the ith 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.*

	$SD(\varepsilon_i)$			
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 D. Data are generated where $\varepsilon_i \overset{iid}{\sim} \mathbb{N}(0, \sigma^2)$ 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 $\text{Var}(Z_i) = X_i^2 \text{Var}(\varepsilon_i)$. See Appendix E 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

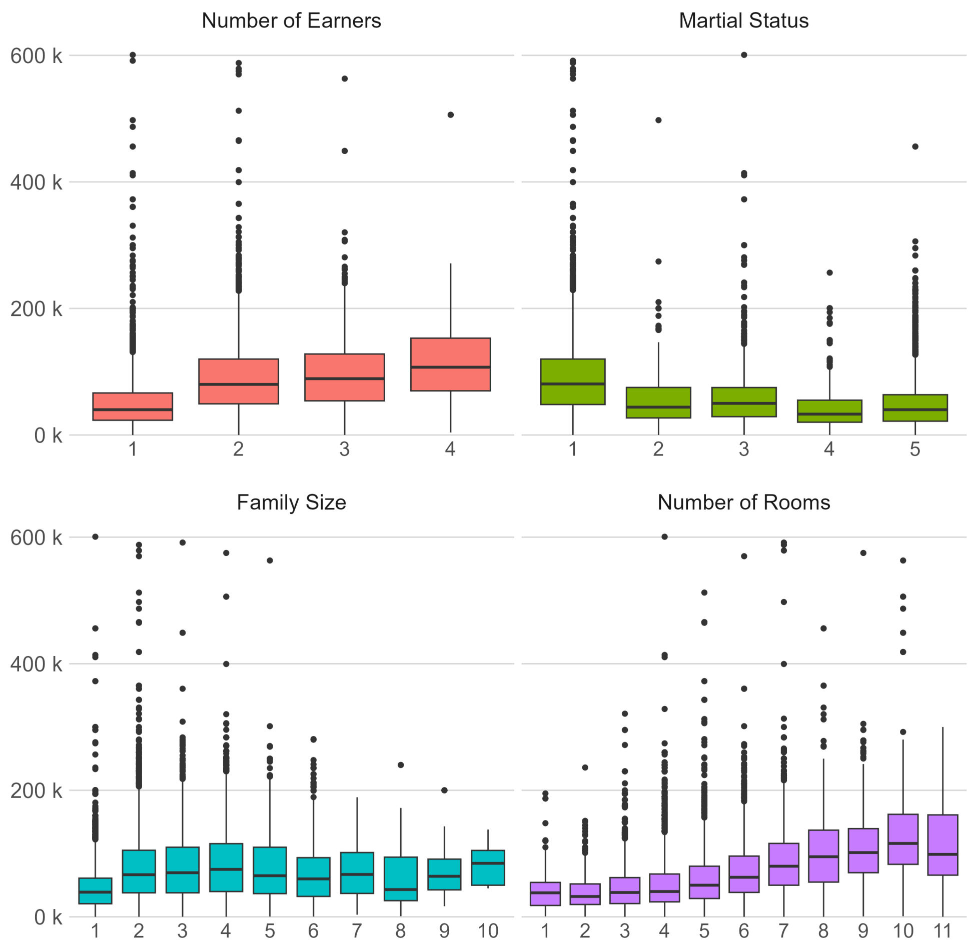

Figure 4.1: *Spread of variable* **FINCBTAX** *across earner characteristics for determining reasonable stratifying variables.*

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.

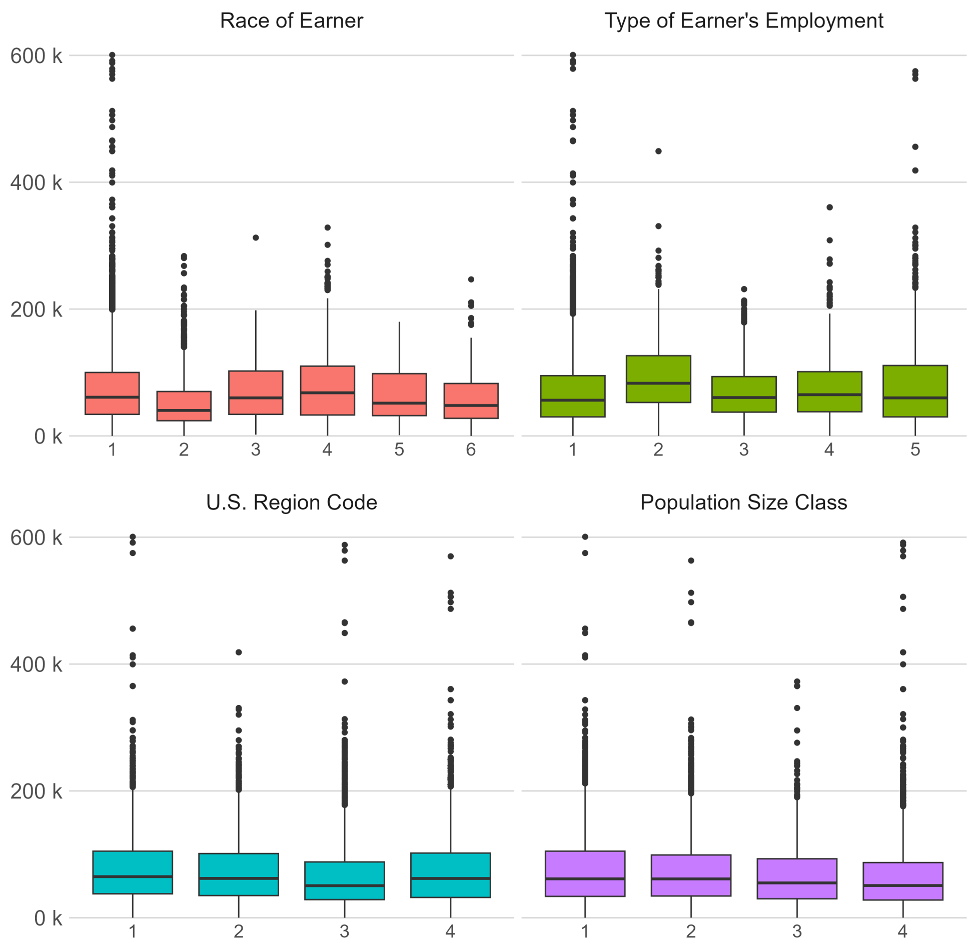

Figure 4.2: *Spread of variable `FINCBTAX` across location and earner characteristics for determining reasonable clustering psu variables.*

4.1.4 Two-Stage 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 usually 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 Three-Stage Clustering and Stratifying Sampling

A three-stage sampling design adds an additional layer of complexity to the sampling design to mimic the complex survey designs used by modern large 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 D. 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`.

Simulation Set Up

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

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

The simulation has a 4 factorial design with 20 scenarios. Varying based on sampling methods will test how each survey weight diagnostic test performs in complex sampling. Additionally, the robustness of the tests in different sample sizes is of great interest given many of the tests are asymptotically correct. Bollen *et al.*, 2016

Cases:

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

Constants:

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

n	*methods*	*DD*	*PN*	*HP*	*PS1*	*PS1q*	*PS2*	*PS2q*	*PS3*	*WF*	*LR*
25	grouping	100.0	40.0	100.0	100.0	100.0	100.0	100.0	99.4	99.8	3.8
25	pps	99.2	23.9	99.0	100.0	100.0	99.9	98.8	67.0	98.9	15.4
25	strata	7.7	23.4	7.4	11.3	17.4	13.9	4.3	4.2	3.8	38.9
50	grouping	100.0	43.9	100.0	100.0	100.0	100.0	100.0	99.3	99.6	4.2
50	pps	100.0	25.3	100.0	100.0	100.0	100.0	100.0	83.3	100.0	6.3
50	strata	7.4	20.7	7.1	10.9	18.6	12.9	4.1	3.9	4.1	38.0
100	grouping	100.0	43.7	100.0	100.0	100.0	100.0	100.0	98.3	99.7	4.7
100	pps	100.0	23.5	100.0	100.0	100.0	100.0	100.0	96.5	100.0	1.9
100	strata	5.8	21.8	5.5	11.1	18.5	14.1	3.3	3.8	3.7	39.5
500	grouping	100.0	40.7	100.0	100.0	100.0	100.0	100.0	99.5	99.7	3.6
500	pps	100.0	27.3	100.0	100.0	100.0	100.0	100.0	99.7	100.0	0.9
500	strata	7.2	19.6	6.9	11.0	18.7	13.6	4.6	4.0	4.8	38.9
1000	grouping	100.0	40.2	100.0	100.0	100.0	100.0	100.0	98.9	99.5	3.5
1000	pps	100.0	31.1	100.0	100.0	100.0	100.0	100.0	99.8	100.0	0.9
1000	strata	7.1	20.7	6.7	10.6	17.6	13.1	3.2	3.9	3.5	41.0

Table 4.5: *Wang et al data*

5 Conclusion

To-DO

This page intentionally left blank.

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.

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.

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.

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.

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.

A

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 $\mathcal{N}(\mu = 0, \sigma^2 = 0.5)$.

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

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

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

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

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

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

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

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

Table A.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):

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

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

B

How the weights W_k in the sample after using Probability Proportional to Size sampling procedure is not clear. The uncertainty of structure of the weights $\vec{W}$ is a possible cause of the difference between the replication results and the results shown in Wang *et al*. (2023).

TO-DO

C

Wang *et al.* (2023) Increased Replications

How the weights W_k in the sample after using Probability Proportional to Size sampling procedure is not clear. The uncertainty of structure of the weights $\vec{W}$ is a possible cause of the difference between the replication results and the results shown in Wang *et al*. (2023).

Table C.1: *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.

D

Consumer Expenditure Wrangling Justification

E

PPS Transformation Derivation

Recall that probability proportional to size (PPS) is a sampling design in which each population unit has the inclusion probability proportional

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

To incorporate randomness to replicate uncertainty during the survey design process, a transformed variable Z is proposed that has added noise from the variable X. 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, $E(Z_i) = X_i, \forall i$.

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

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

Variation of Z_i

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

$$\begin{aligned}
\mathrm{Var}(X_i \varepsilon_i) &= \mathrm{Var}(E(X_i \varepsilon_i \mid X_i)) + E(\mathrm{Var}(X_i \varepsilon \mid X_i)) \\
&= \mathrm{Var}(X_i E(\varepsilon_i \mid X_i)) + E(X_i^2 \mathrm{Var}(\varepsilon_i \mid X_i)) \\
&= \mathrm{Var}(X_i E(\varepsilon_i)) + E(X_i^2 \mathrm{Var}(\varepsilon_i)) \\
&= E(\varepsilon_i)^2 \mathrm{Var}(X_i) + E(X_i^2) \mathrm{Var}(\varepsilon_i) \\
&= E(X_i^2) \mathrm{Var}(\varepsilon_i) \\
&= X_i^2 \mathrm{Var}(\varepsilon_i).
\end{aligned}$$

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
March 2024