Diagnostic Tests for the Necessity of Weight in Regression with Survey Data

1

2

3

Feng Wang¹, HaiYing Wang², Jun Yan²

School of Statistics, Shanxi University of Finance and Economics
 Department of Statistics, University of Connecticut

Abstract

To weight or not to weight in regression analyses with survey data has been debated in the 5 literature. The problem is essentially a tradeoff between the bias and the variance of the 6 regression coefficient estimator. An array of diagnostic tests for informative weights have 7 been developed. Nonetheless, studies comparing the performance of the tests, especially 8 for finite samples, are scarce, and the theoretical equivalence of some tests has not been 9 investigated. Focusing on the linear regression setting, we review a collection of such tests 10 and propose enhanced versions of some of them that require an auxiliary regression model for 11 the weight. Further, the equivalence of two popular tests is established which has not been 12 reported before. In contrast to existing reviews with no empirical comparison, we compare 13 the sizes and powers of the tests in simulation studies. The reviewed tests are applied to an 14 regression analysis of the family expenditure using the data from the China Family Panel 15 Study. 16

¹⁷ Keywords: bias-variance tradeoff; complex survey; hypothesis test; weighted regression

4

18 1 Introduction

To weight or not to weight in analyses of survey data is a long standing question for survey 19 methodologists, dating back to Smith (1988). The same question keeps coming back in 20 statistics (e.g., Bertolet, 2008) as well as application fields such as epidemiology (Frohlich 21 et al., 2001; Tchetgen et al., 2012), economics (Nguyen and Murphy, 2015; Gluschenko, 22 2018), and social and behavioral studies (Hsieh, 2004). Survey data are often released 23 with a weight for each observation. "Contrary to what is assumed by many theoretical 24 statisticians, survey weights are not in general equal to inverse probabilities of selection 25 but rather are typically constructed based on a combination of probability calculations and 26 nonresponse adjustments" (Gelman, 2007, p.153). There is a general consensus that weights 27 should be used for descriptive statistics such as means and ratios (e.g., Kish and Frankel, 28 1974). For regression models, however, it has been debated on whether weights should be 29 used (Winship and Radbill, 1994; Gelman, 2007; Solon et al., 2015). When weights are quite 30 different, especially when they represent different probabilities of being selected, weighting 31 corrects biases in inferences about the population. If weights are ingorable in the sense 32 that the inference is valid without them, not weighting may be preferred for lower variance 33 than otherwise. Given the fundamental importance of linear regression in practice and the 34 extensibility of the concepts beyond linear regression, we limit our scope to diagnostic tests 35 of informative weight in linear regression. 36

A recent review by Bollen et al. (2016) classifies the tests for the necessity of weights 37 in regression analysis into two groups. Tests in the first group are difference-in-coefficients 38 (DC) tests, which examine whether the difference between the weighted and unweighted co-39 efficients estimates is different from zero (Kott, 1991; Pfeffermann, 1993). Tests in the second 40 group are weight-association (WA) tests, which examine whether the weight is informative 41 about the response variable after conditioning on the covariates (Dumouchel and Duncan, 42 1983; Pfeffermann and Sverchkov, 1999; Wu and Fuller, 2005; Pfeffermann and Sverchkov, 43 2007). Bollen et al. (2016) conceptually reviewed the assumptions and properties of the 44

tests, and noted that Monte Carlo simulation studies on the finite sample performance of these tests are quite limited, most of which were designed to illustrate a new test with a small simulation study to demonstrate its potential. Unaddressed questions remain that are important for guiding the practitioners. For example, do these tests hold their size? Which tests have higher power? Are some of the tests equivalent to each other? Are there software implementations for the tests?

There are tests that belong to neither the DC nor the WA groups. Some are reviewed 51 as tests of informative sampling process (Pfeffermann and Sverchkov, 2003, Section 12.2.2) 52 or sampling ignorability (Pfeffermann and Sverchkov, 2010, Section 7). Pfeffermann and 53 Nathan (1985) proposed a test based comparing the out-of-sample prediction power between 54 the weighted and unweighted fits. A large difference in squared prediction errors indicates 55 non-ignorable weights. Pfeffermann and Sverchkov (2003) proposed a test that compares 56 the estimating equations with and without the weights. The estimating equations could 57 be score equations if likelihood is specified, but can be more general without distributional 58 Eideh and Nathan (2006) proposed to test based on the Kullback–Leibler assumption. 59 information against exponential or linear inclusion probability models. It was claimed that 60 the testing statistic follows a chi-squared distribution with one degree of freedom. In their 61 simulation study, however, the null distribution of the test statistics appears to be far different 62 from chi-squared with one degree of freedom (Eideh and Nathan, 2006, Table 3). This test 63 thus needs a rigorous further investigation. Finally, Breidt et al. (2013) proposed a likelihood 64 ratio (LR) test that compares two weighted log-likelihoods with different weights. The null 65 distribution of the test statistic is a mixture of chi-squared distributions with one degree of 66 freedom, The performances of these tests in comparison with those reviewed by Bollen et al. 67 (2016) would be a great practical value. 68

This paper revisits an array of diagnostic tests on ignorable weights in linear regression with survey data. We focus on linear regression as this is the arena where most of the widely used model in survey data analysis in many fields. Only unclustered and homoskedastic

scenarios are considered to simplify the presentation and to remain consistent with the 72 literature as there is a lot to summarize. Our contribution is three-fold. First, we conduct 73 a comprehensive numerical study to compare the size and power of a few commonly used 74 tests and their variations in several scenarios. Such comparison has been long missing in the 75 literature. Some tests need an auxiliary linear model, which may not pick up the nonlinear 76 associations. Some tests were found to performed almost identically in the study, which 77 led to our second contribution — we establish the equivalence of two powerful tests. The 78 test statistics of the DC test of Pfeffermann (1993) and the WA test of Dumouchel and 79 Duncan (1983) are 1-to-1 maps of each other. Finally, the tests are applied to a regression 80 analysis of family expenditures with data from the China Family Panel Studies (CFPS) and 81 its subsamples of different sizes. 82

⁸³ 2 Tests for Necessity of Weight in Regression

⁸⁴ Consider a regression analysis arising from a survey data obtained without clustered sam-⁸⁵ pling. Suppose that the survey consists of a sample S from a finite population U of size N. ⁸⁶ The linear regression model for the population U is assumed to be

$$Y_j = X_j \beta + \epsilon_j, \quad j \in U, \tag{1}$$

⁸⁷ where Y_j is the response variable, X_j is a $p \times 1$ covariate vector (including a component of 1 ⁸⁸ for intercept), β is a $p \times 1$ vector of regression coefficients, and the regression error ϵ_j has ⁸⁹ mean zero and variance σ^2 . The observed survey data S of sample size n is $\{(Y_i, X_i, W_i) :$ ⁹⁰ $i = 1, ..., n\}$, where W_i is the survey weight associated with the *i*th observation. Each ⁹¹ weight W_i may or may not be the inverse probability of selection. Let $Y = (Y_1, ..., Y_n)^{\top}$, ⁹² $X = (X_1^{\top}, ..., X_p^{\top})^{\top}$, $\epsilon = (\epsilon_1, ..., \epsilon_n)^{\top}$, and $W = (W_1, ..., W_n)^{\top}$. A working linear regression ⁹³ of Y on X for the survey data is

$$Y = X\beta + \epsilon. \tag{2}$$

⁹⁴ We are interested in testing the necessity of weighting in fitting (2) to the observed data in ⁹⁵ estimating β , that is, testing whether an unweighted estimator for the β in (2) is unbiased for ⁹⁶ the population parameter β in regression (1). Based on the observed survey data. the least ⁹⁷ squares estimators of β are $\hat{\beta}_u = (X^T X)^{-1} XY$ without weight and $\hat{\beta}_w = (X^T H X)^{-1} X H Y$ ⁹⁸ with weight matrix H = diag(W). Tests for the necessity of weight attempt to answer the ⁹⁹ question whether or not to weight. We review six such tests in approximately chronological ¹⁰⁰ order as follows.

¹⁰¹ 2.1 Dumouchel–Duncan's WA Test

Dumouchel and Duncan (1983) proposed the first WA test for testing informative weights.
A WA test checks whether it holds that

$$H_0: \mathbb{E}(Y|X, W) = \mathbb{E}(Y|X). \tag{3}$$

For linear regression (2), the null hypothesis (3) is equivalent to that the coefficients of the interactions between X and the weight are zero in an extended linear model (e.g., Fuller, 2009, Section 6.3.1). The latter can be easily tested by an F-test. Specifically, consider the extended regression model

$$\mathbb{E}(Y|X,W) = X\beta + HX\gamma,\tag{4}$$

where γ is a $p \times 1$ coefficient vector of HX. Ignorable weight is tested by an F-test for $H_0: \gamma = 0$ with testing statistic

$$F = \frac{(\text{SSE}_r - \text{SSE}_f)/p}{\text{SSE}_f/(n-2p)},\tag{5}$$

where SSE_r and SSE_f are the residual sum of squares under the reduced model (2) and under the full model (4), respectively. Under $\gamma = 0$ and normality assumption of the regression errors, F follows an F(p, n - 2p) distribution. Without the normality assumption, the null distribution is asymptotically F(p, n-2p) for large n. Rejection of $\gamma = 0$ implies that weights are informative; otherwise, there is no sufficient evidence against the unweighted analysis.

¹¹⁵ 2.2 Pfeffermann–Nathan's Test Based on Predictive Power

¹¹⁶ Pfeffermann and Nathan (1985) proposed a simple test based on comparing the out-of-sample ¹¹⁷ predictive power between the weighted and unweighted estimation. Let S = E + V define ¹¹⁸ a split of the sample into two mutually exclusive subsamples E for estimation and V for ¹¹⁹ validation. Weighted and unweighted regressions fitted with the estimation set E are used ¹²⁰ to make predictions for observations in the validation set V. Let v_{ui} and v_{wi} , $i \in V$, denote ¹²¹ the prediction errors under the unweighted fit and weighted fit, respectively. Uninformative ¹²² weight implies

$$H_0: \mathbb{E}(v_{ui}^2 - v_{wi}^2) = 0, \quad i \in V.$$

This hypothesis can be tested by standard Z-test with $Z = \overline{D}/S_D$, where \overline{D} and S_D^2 are the sample mean and sample variance of D_i 's, $i \in V$, with $D_i = v_{ui}^2 - v_{wi}^2$.

Implementation of this prescription requires a random splitting of the sample, so the result is subject to the random split. The prediction errors are only independent conditional on the estimation set E, but not unconditionally independent because they are calculated based on the same $\hat{\beta}_u$ or $\hat{\beta}_w$. There has been no study of the size and power of the test. The dependence among the prediction errors may render the test to have empirical sizes exceeding its nominal sizes. The reduced sample size by half in the construction of the Zmay drastically reduce its power. Both conjectures are observed in our numerical studies.

¹³² 2.3 Hausman–Pfeffermann's DC Test

¹³³ Pfeffermann (1993) proposed a DC test which directly compares $\hat{\beta}_u$ and $\hat{\beta}_w$ using a model ¹³⁴ specification test in econometrics studied by Hausman (1978). Hausman's test can be used ¹³⁵ to detect omitted variables, incorrect functional forms, and other model misspecifications. If the weight W is noninformative about Y conditional on X, then $\hat{\beta}_u$ and $\hat{\beta}_w$ converge to the same target β as the sample size n increases. A DC test checks whether it holds that

$$H_0: \mathbb{E}(\hat{\beta}_u) = \mathbb{E}(\hat{\beta}_w), \tag{6}$$

¹³⁸ The test statistic is

$$T = (\hat{\beta}_u - \hat{\beta}_w)^\top \hat{V}^{-1} (\hat{\beta}_u - \hat{\beta}_w), \tag{7}$$

where \hat{V} is an estimate of $V = \mathbb{V}(\hat{\beta}_u - \hat{\beta}_w)$. The asymptotic null distribution of T is χ_p^2 . When the null hypothesis is rejected, it may be of interest to identify which coefficients are causing the rejection. This can be done by considering statistic \hat{d}_i^2/\hat{V}_{ii} , $i = 1, \ldots, p$, where \hat{d}_i is the *i*th component of $\hat{\beta}_u - \hat{\beta}_w$ and \hat{V}_{Ii} is the *i*th component of the diagonal of \hat{V} . This statistics has asymptotic null distribution of χ_1^2 .

In implementation, the estimate \hat{V} of V needs some care. Hausman (1978) suggested 144 $\hat{V} = \hat{\mathbb{V}}(\hat{\beta}_w) - \hat{\mathbb{V}}(\hat{\beta}_u)$ because $\operatorname{Cov}(\hat{\beta}_u, \hat{\beta}_w - \hat{\beta}_u) = 0$. Unfortunately, this estimator is not 145 necessarily positive definite for small to moderate sample sizes. Asparouhov and Muthen 146 (2007) extended the test to compare the estimators from two different weights and proposed 147 an estimator for V that is always positive definite. Specifically, they suggested $\hat{V}_{\rm AM}$ = 148 $[\hat{\mathbb{V}}(\hat{\beta}_w) + \hat{\mathbb{V}}(\hat{\beta}_u) - 2C]$, where C is an estimator of the covariance matrix of the two estimators. 149 This estimator C is not straightforward to obtain. An explicit variance estimator can be 150 obtained by fitting a regression model with augmented data including weight (Kott, 2018) 151 using a regression routine that allows "design-based" variance estimator. An additional 152 advantage is that the resulting test is heteroscedastic-resistant. We propose a more direct 153 estimator $\hat{V} = \hat{\sigma}^2 A A^{\top}$, where $A = (X^{\top} H X)^{-1} X^{\top} H - (X^{\top} X)^{-1} X^{\top}$, and $\hat{\sigma}^2$ is an estimator 154 of the σ^2 from least squares under the null hypothesis of noninformative weight. This \hat{V} is 155 different from $\hat{\mathbb{V}}(\hat{\beta}_w) - \hat{\mathbb{V}}(\hat{\beta}_u)$ in that $\hat{\sigma}^2$ in $\hat{\mathbb{V}}(\hat{\beta}_w)$ is obtained without weight. 156

The test statistics of the DC test of Pfeffermann (1993) and the WA test of Dumouchel and Duncan (1983) are 1-to-1 maps of each other. **Theorem 1.** Under the null hypothesis of noninformative weight for the linear model (2), the Hausman–Pfeffermann test and the Dumouchel–Duncan test are asymptotically equivalent. If the σ^2 for the test in (7) is estimated with the mean squared error from the model in (4), then the statistics T in (7) and F in (5) are 1-to-1 maps of each other via T = pF.

¹⁶³ The result does not appear to have been noted in the literature. The proof is in the ¹⁶⁴ Appendix.

¹⁶⁵ 2.4 Pfeffermann–Sverchkov's WA Tests

Pfeffermann and Sverchkov proposed multiple WA tests in a sequence of works. Pfeffermann and Sverchkov (1999) checked the association between the residuals from the unweighted regression and weights. Let $\hat{\epsilon}_u = Y - X\hat{\beta}_u$. Pfeffermann and Sverchkov (1999) considered hypotheses H_{0k} : Corr $(\hat{\epsilon}_u^k, W) = 0$, k = 1, 2, 3. For a given k, the sample correlation after the Fisher transformation follows a normal distribution asymptotically under the null hypothesis. Alternatively, Pfeffermann and Sverchkov (1999) suggested considering regressing W on $\hat{\epsilon}_u^k$:

$$\mathbb{E}(W|\hat{\epsilon}_u^k) = \alpha + \beta^{(k)}\hat{\epsilon}_u^k, \quad k = 1, 2, 3,$$
(8)

where α and $\beta^{(k)}$ are the intercept and slope coefficient, respectively. Then, for a given k, a *t*-test H_{0k} : $\beta^{(k)} = 0$ is conducted. The two methods were reported to have similar performance.

The tests of Pfeffermann and Sverchkov (1999) has two limitations. First, for k = 1, 2, 3together, a multiple testing issue arises and needs to be appropriately taken care of. Second, the regression model for W in Equation (8) does not condition on X so that a high correlation between W and $\hat{\epsilon}_u$ could be due to X. Here we propose a simple modification by regressing W on the first two moments of $\hat{\epsilon}_u$ and its interaction with X in addition to X:

$$\mathbb{E}(W|\hat{\epsilon}_u) = f(X;\eta) + \sum_{k=1}^2 \beta^{(k)} \hat{\epsilon}_u^k + \operatorname{diag}(\hat{\epsilon}_u) X\gamma, \qquad (9)$$

where $f(X;\eta)$ is some function of X with parameter η , $\beta^{(1)}$ and $\beta^{(2)}$ are scalars, δ is a $p \times 1$ coefficient vector for X, and γ is a $p \times 1$ coefficient vector for the interaction between X and $\hat{\epsilon}$. The simplest forms of $f(X;\eta)$ are linear and quadratic in X. Then we test the hypothesis $H_0: \beta^{(1)} = \beta^{(2)} = 0, \ \gamma = 0$ by a standard F-test.

Pfeffermann and Sverchkov (2007) suggested another WA test based on regressing W on both X and Y:

$$\mathbb{E}(W|X,Y) = X\eta + Y\gamma.$$
⁽¹⁰⁾

Then a *t*-test is conducted for the hypothesis $H_0: \gamma = 0$. Rejecting the hypothesis implies that the weight is informative for Y. This test was studied in the context of small area estimation, where the same test was conducted in multiple areas.

The regression model (10) only captures the linear relationship between W and (X, Y). To capture possible nonlinear relationships, here we propose a simple modification by considering regression model

$$\mathbb{E}(W|X,Y) = f(X;\eta) + \sum_{k=1}^{2} Y^k \gamma_k, \qquad (11)$$

where $f(X;\eta)$ is some function of X with parameter η , γ_k is the coefficient of Y^k , k = 1,2. The simplest forms of $f(X;\eta)$ are linear and quadratic. An *F*-test for hypothesis $H_0: \gamma_1 = \gamma_2 = 0$ can then be used to determines whether W and Y are associated given X. Misspecification of f may have serious consequences; in some scenarios we have experimented, the size of the test can be completely ruined.

¹⁹⁷ 2.5 Pfeffermann–Sverchkov's Test Based on Estimating Equations

¹⁹⁸ Pfeffermann and Sverchkov (2003) proposed a test that uses the estimating equations to ¹⁹⁹ estimate β . This test requires an auxiliary regression model for W, $\mathbb{E}(W|X) = f(X;\eta)$, ²⁰⁰ which is some function of X with parameter η . The unweighted estimating function $\delta_i(\beta) =$ ²⁰¹ $X_i(Y_i - X_i^{\top}\beta), i \in S$. Let \hat{W}_i be the fitted value of this regression. Define $q_i = W_i/\hat{W}_i$. Let 202 $R(X_i;\beta) = \delta_i(\beta) - q_i\delta_i(\beta)$. Ignorable sampling weight means

$$H_0: \mathbb{E}[R(X_i;\beta)] = 0.$$

²⁰³ This hypothesis can be tested by a Hotelling statistic

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

where \bar{R}_n is the sample mean and $\hat{\Sigma}_{R,n}$ is the sample variance matrix of $R(X_i; \hat{\beta}_u)$'s, $i \in S$. The statistic follows approximately an F distribution with degrees of freedom (p, n - p)under the null hypothesis.

Implementation of this test can use any valid estimating equations. If likelihood is specified, for example, it can be the score equations as Pfeffermann and Sverchkov (2003) suggested. The simplest form of $f(X;\eta)$ is a linear regression, but a more flexible form accommodating non-linearity could improve the power of the test at the cost of a model building process for W.

212 2.6 Wu–Fuller's WA Test

²¹³ Wu and Fuller (2005) proposes a WA test which takes a slightly different extended model ²¹⁴ than that in Dumouchel and Duncan (1983). Similar to Pfeffermann and Sverchkov (2003), ²¹⁵ this test also requires an auxiliary regression model for W, $\mathbb{E}(W|X) = f(X;\eta)$. Let Q =²¹⁶ diag (q_1, \ldots, q_n) , where q_i 's are the same as defined in the last subsection. Consider an ²¹⁷ extended regression

$$\mathbb{E}(Y|X,W) = X\beta + QX\gamma.$$

This regression was suggested by Pfeffermann and Sverchkov (1999) for estimating regression models with survey data. Wu and Fuller (2005) used it to test for informative weight by testing $H_0: \gamma = 0$ with a standard *F*-test as in Wu and Fuller (2005).

The rational of this test is to check the impact of W on Y after removing the information 221 contained in X. The definition of q_i 's factors out the part in the weight W_i that is predictable 222 by X_i . If weight is informative for Y after conditioning on X, then QX is expected to have 223 a significantly nonzero coefficient γ in the extended regression. Otherwise, one would expect 224 $\gamma = 0$. Implementing this test requires an auxiliary regression of W on X, $f(x; \eta)$. As 225 for Pfeffermann and Sverchkov (2003), a model building process may be beneficial. Poor 226 approximation for the relation between W and X might lead to incorrect size and poor 227 power of the test. 228

229 2.7 LR Test

Breidt et al. (2013) proposed an LR test, which is neither a DC nor a WA test. A super-230 population model is assumed that have generated the finite population U. Suppose that 231 the conditional distribution Y_i given X_i in the superpopulation has density $f(\cdot|X_i;\theta)$ with 232 parameter vector θ of dimension q with true value θ_0 . Here θ contains β as a subset. For 233 example, if the distribution is normal, there is a variance parameter in addition to β in θ . 234 Note that $\ln f(Y_i|X_i;\theta)$ is the log-likelihood for the superpopulation distribution, but it may 235 not be the log-likelihood for an observation in the sampled data. For convenience, we still 236 call it log-likelihood as in Breidt et al. (2013). 237

A weighted log-likelihood with a general weight vector $\omega = (\omega_1, \dots, \omega_n)^\top$ is

$$l(\theta; \omega) = \sum_{i=1}^{n} \omega_i \ln f(Y_i | X_i; \theta).$$

²³⁹ Let $\hat{\theta}_U = \arg \min_{\theta} l(\theta; U)$, where $U = (1, \ldots, 1)^{\top}$, and $\hat{\theta}_W = \arg \min_{\theta} l(\theta; W)$. Two LR ²⁴⁰ statistics are considered:

$$T_{U} = 2\left\{ l(\hat{\theta}_{U}; U) - l(\hat{\theta}_{W}; U) \right\} = n(\hat{\theta}_{U} - \hat{\theta}_{W})^{T} J_{U}(\hat{\theta}_{U} - \hat{\theta}_{W}) + o_{p}(1),$$

$$T_{W} = 2\left\{ l(\hat{\theta}_{W}; W) - l(\hat{\theta}_{U}; W) \right\} = n(\hat{\theta}_{W} - \hat{\theta}_{U})^{T} J_{W}(\hat{\theta}_{W} - \hat{\theta}_{U}) + o_{p}(1),$$

where $J_{\omega} = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \omega_i \mathcal{I}(x_i; \theta_0), \ \omega \in \{U, W\}$, and $\mathcal{I}(x_i; \theta_0)$ is the Fisher Information for the *i*th observation. Under the null hypothesis of noninformative weight, $n^{1/2}(\hat{\theta}_W - \hat{\theta}_U) \xrightarrow{\mathcal{L}} \mathcal{N}(0, -J_U^{-1} + J_W^{-1}K_W J_W^{-1})$, where $K_W = \lim_{n \to \infty} \frac{1}{n} \sum_{i \in S} W_i^2 \mathcal{I}(x_i; \theta_0)$. The asymptotic distribution of $T_{\omega}, \ \omega \in \{U, W\}$, is $T_{\omega} \xrightarrow{\mathcal{L}} \sum_{j=1}^{q} \lambda_{\omega j} Z_j^2$, where λ_{ω} is the vector of eigenvalues of

$$(-J_U^{-1} + J_W^{-1} K_W J_W^{-1})^{T/2} J_\omega (-J_U^{-1} + J_W^{-1} K_W J_W^{-1})^{1/2}$$

and Z_j 's, $j = 1, \ldots, p$, are independent $\mathcal{N}(0, 1)$ variables.

Implementation of the LR tests require maximizing both the weighted and unweighted log-likelihood. The limiting distribution is not chi-square as in the commonly encountered situations. Instead, it is a linear combination of chi-square random variables with coefficients being the eigenvalues of a certain matrix. This matrix depends on the true parameter θ_0 , which has to be evaluated at an estimate $\hat{\theta}_U$. This method is limited in that it requires distributional specification of the regression errors. The validity of the test may be undermined if the distribution is misspecified.

²⁵⁴ **3** Simulation Studies

Two simulation studies were conducted to compare the performances of the reviewed tests. 255 Eight tests were included in the comparison with the following abbreviations : DD (Du-256 mouchel and Duncan, 1983); PN (Pfeffermann and Nathan, 1985); HP (Hausman, 1978; 257 Pfeffermann, 1993); PS1 (Pfeffermann and Sverchkov, 1999); PS2 (Pfeffermann and Sver-258 chkov, 2007); PS3 (Pfeffermann and Sverchkov, 2003); WF (Wu and Fuller, 2005); LR (Breidt 259 et al., 2013). The LR test used the one based on T_U because it performed better than T_W 260 in our studies. For PS1 and PS2 that requires regressing W on residuals of Y or Y itself, 261 we also used versions that uses quadratic terms to model possible nonlinearity. They are 262 abbreviated as PS1q and PS2q, respectively, and also included in the comparison study. 263

²⁶⁴ 3.1 Study 1

The first study was adapted from Pfeffermann and Sverchkov (1999), A population of size N = 3,000 was generated for (Y_i, X_i) with a linear regression model

$$Y_i = 1 + X_i + \varepsilon_i, \qquad i = 1, \dots, N, \tag{12}$$

where X_i 's were independently generated from the standard uniform distribution $\mathcal{U}(0, 1)$ and ε_i 's were independently generated from $\mathcal{N}(0, \sigma^2)$ with $\sigma \in \{0.1, 0.2\}$. The levels of σ here are lower than that used in Pfeffermann and Sverchkov (1999) so that the differences in power are visible. Samples of size $n \in \{100, 200\}$ were drawn from the population with probability proportional to weight defined by

$$W_i = aY_i + 0.3X_i + \delta U_i,\tag{13}$$

where U_i 's are independently drawn from $\mathcal{U}(0,1)$, δ has two levels (1,1.5), and a has four levels (0, 0.2, 0.4, 0.6). When a = 0, the weight W_i is not informative about Y_i conditioning on X_i . This design led to $2 \times 2 \times 2 \times 4 = 32$ configurations. For each configuration, we generated 1,000 samples, and applied the nine tests to each sample.

Table 1 shows the empirical rejection rates of the ten tests with significance level 0.05 as a function of a. In all the settings for a = 0, the rejection rates are close to 0.05 except for the PN test, indicating that these tests maintain their sizes in this study. The empirical size of the PN test is repetitively above the nominal size 0.05, which may be explained by the dependence among the prediction errors introduced by the shared coefficient estimates. Despite being liberal, PN has power that is much lower than other tests due to halved sample size. Therefore, PN is excluded in the discussions in the sequel.

Next we compare the powers of the tests with PN excluded. As a deviates from zero further or sample size n increases, the power of all tests in all settings increases. Other factors

n	σ	δ	a	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	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	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	4.8	7.2	4.7	3.2	4.5	4.3	4.5	4.7	5.1	5.5
			0.2	10.5	10.8	10.4	9.8	11.9	14.5	11.3	9.2	11.8	9.6
			0.4	36.1	14.6	35.6	29.4	31.4	46.2	36.0	27.2	35.7	23.9
			0.6	70.4	19.5	70.1	58.4	64.2	80.5	71.2	57.1	70.8	47.3
	0.2	1.5	0.0	4.4	8.3	4.3	4.5	4.5	4.7	4.7	4.5	4.5	5.0
			0.2	18.4	10.2	18.0	19.6	15.6	21.5	18.7	14.1	18.0	15.8
			0.4	57.4	14.7	57.1	61.2	50.0	67.8	57.1	45.7	56.7	47.4
			0.6	91.7	25.2	91.5	91.8	89.0	96.1	92.1	86.3	91.8	83.1
		1	0.0	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

Table 1: Empirical rejection percentages of ten tests in Study 1 with W linear in Y based on 1000 replicaties for normal regression error and sample size $n \in \{100, 200\}$. The rejection rates are sizes when a = 0 and powers otherwise.

held constant, higher δ leads to lower power because of more noise in the weight model (13). In contrast, higher σ leads to higher power, which is expected as higher σ means higher variation of Y_i and, hence, higher signal-to-noise ratio in the weight model (13). Among all the tests, PS2 appears to have the highest power in all the settings, followed by DD, HP, and WF which are very similar. PS3 and LR appears to have the lowest power in all the
settings. The modified versions PS1q and PS2q are a bit less powerful than PS1 and PS2,
respectively. PS3 is not better than the DD or HP.

The finite sample performance of the tests, especially the LR test, may depend on the 292 distribution of the regression error. To investigate this issue, we considered three additional 293 distributions of ϵ_i in Equation (12): 1) gamma with shape 10 and scale $\sqrt{10/\sigma^2}$; 2) student 294 t with 5-degrees of freedom and scale $\sqrt{5/3}\sigma^2$; 3) uniform $(0, \sqrt{12/\sigma^2})$. These distributions 295 were centered by their means so that they have mean zero and variance σ^2 , matching the first 296 two moments of $\mathcal{N}(0, \sigma^2)$. Table 2 shows the empirical rejection percentage of the tests with 297 $\sigma=0.1$ and $\delta=1$ under different error distributions. The LR test does not hold its size in 298 the case of heavy-tailed regression error, t distribution; under other distributions, it appears 299 to hold its size. The performances of all other tests are robust to the error distribution. 300 which is expected because their null distributions are asymptotically valid regardless of the 301 error distribution. The relative performances of these tests remain in the same order as those 302 under the normal regression error. 303

Now we change the weight generation model from a linear function in X and Y to a quadratic function in X and Y:

$$W_i = a(Y_i - 1.5a)^2 + 0.3X_i - 0.3X_i^2 + U_i,$$
(14)

where U_i 's are independent $\mathcal{U}(0,1)$ variables, and the scalar parameter a controlling the informativeness of W for Y has four levels $\{0, 0.5, 1.0, 1.5\}$. This design has interesting features. When a = 0, the weight is obviously noninformative. When $a \neq 0$, the weight is informative, but for a = 1, the partial correlation between W_i and Y_i is zero, which makes it hard to tests based on an auxiliary linear regression for W_i to detect the informativeness of W_i . Table 3 summarizes the empirical powers in percentage of the tests with $\sigma = 0.1$ based on 1000 replicates. All tests reported here hold their sizes when a = 0. When a = 0.5, all

Distribution	n	a	DD	PN	HP	PS1	PS1q	PS2	PS2q	PS3	WF	LR
normal	100	0.0	5.1	8.0	5.0	4.6	4.2	5.7	4.8	3.7	4.7	5.3
		0.2	8.9	6.2	8.8	6.7	8.4	10.1	8.7	5.2	8.2	7.5
		0.4	19.3	11.8	19.1	13.4	16.7	23.9	18.3	12.9	19.4	12.0
		0.6	42.3	12.0	41.8	28.9	36.5	52.4	42.7	26.6	42.3	24.1
	200	0.0	5.0	10.2	4.8	3.4	5.0	3.9	4.3	4.8	4.8	5.7
		0.2	11.7	10.1	11.6	10.4	10.1	14.5	10.5	7.7	10.8	11.2
		0.4	36.8	11.4	36.4	29.3	29.7	44.6	34.8	27.3	35.0	24.6
		0.6	72.1	19.9	71.8	61.1	65.1	81.6	71.9	59.9	71.3	49.3
unif	100	0.0	5.3	6.7	5.1	3.2	4.6	3.8	4.4	4.5	4.7	3.0
		0.2	10.3	7.8	10.1	8.3	8.2	11.5	9.0	7.6	10.1	4.8
		0.4	15.9	10.2	15.4	12.7	14.3	22.3	16.2	13.9	17.3	9.5
		0.6	39.0	14.7	38.8	26.6	33.7	49.0	39.7	29.5	40.1	18.7
	200	0.0	6.3	7.9	6.3	4.8	5.0	5.1	5.9	5.0	6.2	3.4
		0.2	13.1	8.1	12.9	10.6	10.9	15.3	12.7	9.7	12.0	7.9
		0.4	35.1	12.5	34.9	31.8	29.8	48.4	35.7	27.9	35.8	17.8
		0.6	71.3	21.9	71.1	58.3	64.9	82.0	70.2	62.8	70.1	41.9
gamma	100	0.0	4.8	8.5	4.7	5.4	3.8	5.7	4.5	4.2	5.0	6.8
		0.2	9.0	8.7	8.8	7.4	8.1	10.6	8.2	6.2	8.9	11.2
		0.4	19.9	8.4	19.2	14.8	14.8	26.5	19.8	11.3	19.3	15.8
		0.6	41.5	13.4	40.6	27.9	36.9	51.8	41.5	27.1	40.6	28.5
	200	0.0	4.2	9.3	4.1	3.9	4.2	4.9	4.2	4.2	4.4	6.1
		0.2	11.3	9.5	10.9	11.4	9.8	15.3	11.3	9.3	11.8	11.9
		0.4	38.1	13.4	37.8	32.2	30.9	48.6	37.5	27.1	37.4	30.7
		0.6	74.2	18.4	74.0	63.6	70.1	81.8	74.5	62.6	74.5	56.8
\mathbf{t}	100	0.0	5.6	8.6	5.5	4.6	4.9	5.5	4.5	3.4	5.6	13.6
		0.2	12.7	8.9	12.3	11.1	9.8	14.4	11.3	7.2	12.4	17.7
		0.4	34.8	8.6	33.9	29.0	29.1	42.9	33.4	16.7	34.4	32.6
		0.6	59.9	13.1	59.1	45.4	52.4	70.9	59.4	34.8	59.2	47.0
	200	0.0	5.6	7.8	5.6	3.4	4.6	4.3	4.6	4.1	5.1	16.3
		0.2	19.8	11.4	19.8	19.5	15.4	25.5	18.5	11.9	20.7	27.1
		0.4	59.9	13.8	59.4	54.2	52.2	69.2	59.0	40.7	59.8	50.9
		0.6	91.5	21.0	91.4	86.3	87.8	95.5	91.5	80.8	91.7	79.2

Table 2: Empirical rejection percentages of ten tests in Study 1 with W linear in Y based on 1000 replicaties for different error distributions and sample size $n \in \{100, 200\}$. The rejection rates are sizes when a = 0 and powers otherwise.

tests have decent powers with PS2 being, again, the most powerful, followed by WF, DD, HP and PS1. When a = 1, however, tests PS1 and PS2 appear to be powerless. The modified tests PS1q and PS2q turn out to be powerful, with PS2q being the most competitive.

					1						
\overline{n}	a	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

Table 3: Empirical rejection percentages of ten tests in Study 1 with W quadratic in Y based on 1000 replicaties for normal regression error and sample size $n \in \{100, 200\}$. The rejection rates are sizes when a = 0 and powers otherwise.

316 3.2 Study 2

The second study was adapted from Wu and Fuller (2005). The population data (Y_i, X_i) 's were generated from a linear regression model

$$Y_i = 0.5 + X_i + e_i$$
 $i = 1, 2, \dots,$

where X_i 's and e_i 's were independently generated from $\mathcal{N}(0, 0.5)$. The selection probability W_i for subject $i, i = 1, 2, \ldots$, was set to be

$$W_i = a\eta(X_i) + b\eta\left(\psi e_i + (1 - \psi)z_i\right),$$

where z_i was generated from $\mathcal{N}(0, 0.5)$ independent of e_i , and

$$\eta(x) = \begin{cases} 0.025, & x < 0.2, \\ 0.475(x - 0.20) + 0.025, & 0.2 \le x \le 1.2, \\ 0.5, & x > 1.2. \end{cases}$$

with parameters (a, b, ψ) . Function $\eta(\cdot)$ controls the nonlinear association between W_i and Y_i through ψe_i . The weight is noninformative when $\psi = 0$.

The simulation was designed with the following settings. Following Wu and Fuller (2005), 324 the sum of a and b was fixed at 2 to ensure that $W_i \in [0, 1]$. The expectation of W_i was 0.221. 325 Four levels of a were considered: $\{0.25, 0.5, 0.75, 1\}$. As a increases, the correlation between 326 W_i and X_i increases while the correlation between W_i and e_i decreases. Four levels of ψ were 327 considered, $\{0, 0.1, 0.2, 0.3\}$; higher ψ implies that W_i is more informative for Y_i . Two sample 328 sizes, $n \in \{100, 200\}$, were attained by a Poisson sampling. That is, subject i, i = 1, 2, ...,329 is selected in the sample if $U_i < W_i$, where U_i 's are independent $\mathcal{U}(0,1)$ variables, until the 330 desired sample size is reached. In each configuration, 1000 replicates were generated. In each 331 replicate, the population was regenerated before the sample was drawn. 332

Table 4 summarizes the empirical rejection percentage of the tests with significance level 333 0.05 based 1000 replicates for all the settings. When $\psi = 0$, the powers of all the tests 334 are about 5%, suggesting that they all, including the likelihood ratio tests maintain their 335 sizes. This is expected as the residuals were normally distributed. Nonetheless, if the 336 quadratic term of X in $f(X;\eta)$ in the PS2q test were dropped, the test would become 337 extremely liberal (not shown), which is why we always included the quadratic form in all 338 the simulation studies. The powers increase as ψ increases or n increases when other factors 339 are held constant. Increases in a reduces the power in general, the powers are highest when 340 a = 0.25 and lowest when a = 1. Nonetheless, in this specific design, the effect is not 341 monotone; the powers of all the tests increased slightly but noticeably when a increases from 342 0.5 to 0.75. Due to the complexity in the design, no single test is uniformly the best. When 343 $a \in \{0.25, 0.75\}$, PS1 and PS2 have the highest power, followed by HP, DD, and WF which 344 are very close. The differences are about 10% when they are distinguishable. When a = 0.5, 345 PS2 has the highest power, followed closely by HP, DD, and WF. When a = 1, PS2q has 346 the highest power, followed by PS1q, HP, DD, PS2, WF, LR, and PS1. The edge of PS2q 347 over PS2 suggests the importance of capturing the nonlinear relationship between W and Y348 in the auxiliary regression in a situation like here. PS3 ranks the lowest in all the scenarios 349 among all tests except PN. The LR test ranks the second lowest in all scenarios except in 350

n	a	ψ	DD	PN	ΗP	PS1	PS1q	PS2	PS2q	PS3	WF	LR
100	1.0	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	33.0	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	36.0	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
	0.5	0.0	6.4	6.7	6.2	4.4	5.1	4.5	4.1	6.0	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	1.0	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
		0.3	98.1	29.4	98.1	95.1	97.8	97.8	98.6	88.3	95.2	91.3
	0.5	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

Table 4: Empirical rejection percentages of ten tests in Study 2 based on 1000 replicaties for sample size $n \in \{100, 200\}$. The rejection rates are sizes when $\psi = 0$ and powers otherwise.

the case of a = 1. The results suggest that the each test may have its own favorable settings.

³⁵² 4 Consumption Expenditure of Chinese Families

We apply the tests to a study on Chinese household consumption expenditure using the 353 CFPS data (Institute of Social Science Survey, Peking University, 2015; Xie and Hu, 2014). 354 The CFPS is a nearly nationwide, comprehensive, longitudinal social survey that is intended 355 to serve research needs on a large variety of social phenomena in contemporary China. A 356 multi-stage probability strategy was used in CFPS to reduce operation costs, with implicit 357 stratification to increase efficiency (Xie and Lu, 2015). The 2014 data contains 13,946 house-358 holds, each with a weight representing the inverse sampling probability. The data has been 359 used in many studies on Chinese families, such as the properties of household wealth (Xie 360 and Jin, 2015) and reduction of catastrophic health expenditures (Ma et al., 2019). 361

Our focus is the impact of householder's education level on household consumption ex-362 penditure. For better data quality on household expenditure, we screened the households by 363 two conditions: 1) the householder was the one who responded to the questionare; and 2) 364 the householder was the principal of family expenditure decisions. After removing cases with 365 missing values, we ended up with n = 4,834 householders. The data needed for regression 366 modeling was obtained by joining the household table and householder table. The response 367 variable is the log-transformed household consumption expenditure. The householder ed-368 ucation level is a factor with five levels: junior high or lower, high school, junior college, 369 bachelor, and master or higher. Control variables include: log-transformed family income in 370 Chinese Yuan; proportion of asset-based income in total family income; family size; house-371 holder age; and householder gender. The continuous variable (log family income, property 372 income proportion, and age) were centered by their means; family size was centralized by 3, 373 which was the mode. Of the 4,384 householders, 2,863 (59%) were male; the proportion of 374 householders with different education levels were 76.50%, 14.90%, 5.25%, 3.00%, and 0.35%, 375 respectively, for junior high or lower, high school, junior college, bachelor, and master or 376 higher. Obviously, householders with a master degree or higher are oversampled. We expect 377 to reject that the weight is noninformative. 378

	Unweigh	nted	Weight	P-value for	
	coefficient	SE	coefficient	SE	difference
Intercept	10.259	0.020	10.375	0.032	0.388
log family income	0.250	0.009	0.280	0.023	0.924
asset-based income proportion	0.506	0.116	0.662	0.178	0.000
family size	0.121	0.009	0.101	0.015	0.949
family size, quadratic	-0.011	0.002	-0.009	0.003	0.999
age	-0.083	0.008	-0.040	0.012	0.899
age, quadratic	-0.004	0.005	-0.003	0.008	0.999
male	-0.091	0.022	-0.184	0.033	0.460
high school	0.253	0.030	0.268	0.049	0.872
junior college	0.476	0.049	0.418	0.060	0.339
bachelor	0.614	0.064	0.617	0.083	0.958
master or higher	0.938	0.179	0.433	0.150	0.000

Table 5: Estimated coefficients and their standard errors (SE) from unweighted and weighted regression. Each p-value is for testing the null hypotheses that there is no difference in expectation between the two versions of the corresponding coefficient.

Table 5 summarizes the estimated coefficients and their standard errors from both un-379 weighted and weighted regression. The results of weighted regression were obtained with 380 R package survey (Lumley, 2004). All the reviewed tests rejected the hypothesis that the 381 weight was noninformative strongly with extremely small p-values (below 0.001). Therefore, 382 the analyses should be based on the results from the weighted regression. All the coefficients 383 are significantly nonzero except the quadratic term of householder age. As expected, fami-384 lies with higher income and higher proportion of asset-based income consumed more; bigger 385 families consumed more, but the rate of increase slowed as family size increased as indi-386 cated by the negative quadratic effect. From the householder's perspective, older and male 387 householders spent less. With junior high or lower as reference, householders with higher 388 education level tend to spend more, but the increasing trend stopped at the bachelor's level. 389 Householders with a master degree or higher consumed less on average than those with a 390 bachelor's degree; the opposite conclusion was obtained in the unweighted regression. 391

³⁹² If the weight were incorrectly ignored, the results from the unweighted regression would ³⁹³ be misleading. To tell which coefficients have been estimated significantly differently in the

1000	•									
m	DD	PN	ΗP	PS1	PS1q	PS2	PS2q	PS3	WF	LR
300	33.7	17.9	31.9	47.4	32.6	49.2	51.8	35.6	37.0	38.4
500	55.0	18.8	54.2	74.5	64.1	76.8	81.1	70.6	61.3	64.7
1000	91.9	23.3	91.5	97.4	97.3	97.8	99.6	98.6	94.2	96.8
	m 300 500 1000	m DD 300 33.7 500 55.0 1000 91.9	m DD PN 300 33.7 17.9 500 55.0 18.8 1000 91.9 23.3	m DD PN HP 300 33.7 17.9 31.9 500 55.0 18.8 54.2 1000 91.9 23.3 91.5	m DD PN HP PS1 300 33.7 17.9 31.9 47.4 500 55.0 18.8 54.2 74.5 1000 91.9 23.3 91.5 97.4	m DD PN HP PS1 PS1q 300 33.7 17.9 31.9 47.4 32.6 500 55.0 18.8 54.2 74.5 64.1 1000 91.9 23.3 91.5 97.4 97.3	m DD PN HP PS1 PS1q PS2 300 33.7 17.9 31.9 47.4 32.6 49.2 500 55.0 18.8 54.2 74.5 64.1 76.8 1000 91.9 23.3 91.5 97.4 97.3 97.8	m DD PN HP PS1 PS1q PS2 PS2q 300 33.7 17.9 31.9 47.4 32.6 49.2 51.8 500 55.0 18.8 54.2 74.5 64.1 76.8 81.1 1000 91.9 23.3 91.5 97.4 97.3 97.8 99.6	m DD PN HP PS1 PS1q PS2 PS2q PS3 300 33.7 17.9 31.9 47.4 32.6 49.2 51.8 35.6 500 55.0 18.8 54.2 74.5 64.1 76.8 81.1 70.6 1000 91.9 23.3 91.5 97.4 97.3 97.8 99.6 98.6	m DD PN HP PS1 PS1q PS2 PS2q PS3 WF 300 33.7 17.9 31.9 47.4 32.6 49.2 51.8 35.6 37.0 500 55.0 18.8 54.2 74.5 64.1 76.8 81.1 70.6 61.3 1000 91.9 23.3 91.5 97.4 97.3 97.8 99.6 98.6 94.2

Table 6: Percentage of rejecting the null hypothesis of noninformative weight in the study of Chinese household consumption expenditure from 1000 valid subsamples of size $m \in \{300, 500, 1000\}$.

weighted regression, an individual test can be performed on each regression coefficient. The 394 p-values of such tests reported in Table 5 suggest that two coefficients were estimated with 395 significant differences. One is that the effect of asset-based income proportion is higher from 396 the weighted regression than that from the unweighted regression. The other is the effect of 397 householders with a master degree or higher with junior higher or lower as reference, which 398 is of primary interest. The unweighted regression suggests that, other factors held constant, 399 families whose householder had a master degree or higher had the highest consumption 400 expenditure; the weighted regression, however, suggests families whose householders had a 401 bachelor's degree has the highest. The drastic difference shows the impact of the correctly 402 incorporating weight in this analysis. 403

The large sample size of this application provides an opportunity to compare the tests in 404 a realistic setting by treating the sample as a population. Using the weight to resample from 405 the data, we obtained subsamples of size $m = \{300, 500, 1000\}$. Because of the categorical 406 nature of the education level, not all subsamples had a full-rank design matrix. We kept 407 resampling until 1000 valid subsamples were obtained. The acceptance rates were 42.4%, 408 70.6%, and 95.2%, respectively, for subsample size 300, 500, and 1000. For each subsample, 400 we tested for noninformative weight using the tests compared in the simulation studies. 410 Table 6 summarizes the percentages of rejection with significance level 0.05 based on the 411 1000 replicates. For this application, PS2q, PS3, PS2, and PS1 turns out to have the highest 412 power; WF comes next, followed by DD, and HP. LR based on normal errors cannot be 413 trusted because diagnostics show that the residuals are unlikely to be normally distributed. 414

⁴¹⁵ PN is not recommended for its not holding its size and low power.

416 5 Discussion

Testing for necessity of weight in regression models arises frequently in practical analyses 417 of survey data. Reviews on such tests exist (Bollen et al., 2016) but none compares their 418 sizes and powers in simulation studies. We conducted a comprehensive numerical study to 419 compare the sizes and powers of a few commonly used weight tests under various configu-420 rations. The results show that the test of Pfeffermann and Sverchkov (2007) is the most 421 competitive overall in the settings considered. Nonetheless, it is easy to construct scenarios 422 where this test completely looses its power; this happens when, for example, the weight 423 has zero correlation with the regression error but have strong association with the squared 424 regression error. For tests that require an auxiliary regression model for the weight, the size 425 and power are affected by the specification of the auxiliary model. Most tests are robust to 426 the distribution of the regression error except the likelihood ratio test, which has inflated 427 size under a heavy-tailed error distribution. An interesting theoretical result is that the DC 428 test of Pfeffermann (1993) and the WA test of Dumouchel and Duncan (1983) are equivalent 429 if they use the estimate for the variance of the regression error. In addition, unlike those 430 tests that rely an auxiliary regression whose misspecification may affect their performances, 431 they have no additional model specification burden but give very competitive powers in our 432 simulation study. These findings provide recommendations for choosing the tests in practice. 433 Our review suggests several future research directions. Whether or not to use weight 434 is a general question applicable to all kinds analyses. This review only focuses on linear 435 regression analyses. Similar diagnostic tests for generalized linear models (Nordberg, 1989; 436 Lumley and Scott, 2017), survival models with censored data (Boudreau and Lawless, 2006), 437 or exploratory data analysis and nonparametric regression (Chambers et al., 2003) merits 438 further research. For tests based on correlations (Pfeffermann and Sverchkov, 1999), a new 430

measure of correlation that better distinguishes independence from zero correlation has the 440 potential to perform better where linear correlation fails (Chatterjee, 2020). Most tests in 441 the literature assumed independent homoscedastic data. In practice, however, many com-442 plex survey data have a clustered or nested data structure with possible heteroscedasticity. 443 The dependence structure in such data, sometimes co-present with heteroscedasticity, adds 444 considerable complexity to the estimation problem (e.g., Rabe-Hesketh and Skrondal, 2006; 445 Kott, 2018) and, hence, diagnostic tests. The DC test can be extended to handle clustered 446 data in the general framework of generalized estimating equations (Yan et al., 2013). The 447 likelihood ratio test did not perform well in our study because of its dependence on correct 448 distributional specifications. The derivation is likely to hold for M-estimation (Stefanski and 449 Boos, 2002) where the likelihood specification is replaced with moment specifications. More 450 efforts are needed to research on these immediate questions. 451

452 Acknowledgement

The authors thank the editor and two anonymous referees' for their constructive suggestions that led to a much improved paper. In particular, we are very grateful for a simplification of the proof of the equivalence between the HP and the DD tests suggested by a referee, which has been adopted. FW's research was partially supported by Humanities and Social Science Fund of Ministry of Education of China (20YJA910005) and by Shanxi University of Finance and Economics (2020205).

⁴⁵⁹ A Equivalence between the HP Test and the DD Test

We show that the statistics T in (7) and F in (5) are 1-to-1 maps of each other via T = pFif the $\hat{\sigma}^2$ in T is set to be the $\text{SSE}_f/(n-2p)$.

⁴⁶² *Proof.* We first express the coefficient estimator from the extended regression model (4) in

terms of $\hat{\beta}_u$ and $\hat{\beta}_w$. Let $\hat{\beta}$ and $\hat{\gamma}$ be the least squares estimator of β and γ under the extended regression model (4). They satisfy the following normal equations:

$$X^{\top} X \hat{\beta} + X^{\top} H X \hat{\gamma} = X^{\top} Y, \tag{15}$$

$$X^{\top}HX\hat{\beta} + X^{\top}H^{2}X\hat{\gamma} = X^{\top}HY.$$
(16)

465 Multiplying (15) by $(X^{\top}X)^{-1}$ and (16) by $(X^{\top}HX)^{-1}$, and on subtraction, we get

$$\hat{\gamma} = (X^{\top} H X)^{-1} G^{-1} (\hat{\beta}_w - \hat{\beta}_u), \tag{17}$$

where $G = (X^{\top}HX)^{-1}(X^{\top}H^2X)(X^{\top}HX)^{-1} - (X^{\top}X)^{-1}$. Putting $\hat{\gamma}$ back into Equation (15) 467 gives

$$\hat{\beta} = \hat{\beta}_u - (X^{\top} X)^{-1} G^{-1} (\hat{\beta}_w - \hat{\beta}_u).$$
(18)

468 The SSE_r and SSE_f are, respectively,

$$SSE_r = Y^{\top}Y - \hat{\beta}_u^{\top}X^{\top}Y,$$

$$SSE_f = Y^{\top}Y - \hat{\beta}^{\top}X^{\top}Y - \hat{\gamma}^{\top}X^{\top}HY.$$

469 Their difference is

$$SSE_r - SSE_f = (\hat{\beta} - \hat{\beta}_u)^\top X^\top Y + \hat{\gamma}^\top X^\top HY$$
$$= -(\hat{\beta}_w - \hat{\beta}_u)^\top G^{-1} (X^\top X)^{-1} X^\top Y + (\hat{\beta}_w - \hat{\beta}_u)^\top G^{-1} (X^\top HX)^{-1} X^\top HY$$
$$= -(\hat{\beta}_w - \hat{\beta}_u)^\top G^{-1} \hat{\beta}_u + (\hat{\beta}_w - \hat{\beta}_u)^\top G^{-1} \hat{\beta}_w$$
$$= (\hat{\beta}_u - \hat{\beta}_w)^\top G^{-1} (\hat{\beta}_u - \hat{\beta}_w),$$

where the second equality is by inserting the expressions of $\hat{\beta}$ and $\hat{\gamma}$ in (17) and (18), respectively. Since $G = AA^{\top}$ and $\hat{V} = \hat{\sigma}^2 AA^{\top}$, where $A = (X^{\top}HX)^{-1}X^{\top}H - (X^{\top}X)^{-1}X^{\top}$, 472 we have

$$\frac{\mathrm{SSE}_r - \mathrm{SSE}_f}{\hat{\sigma}^2} = T.$$

 $_{473}$ Combined with the DD test statistic (5), we have

$$\frac{T}{F} = \frac{p}{\hat{\sigma}^2} \frac{\text{SSE}_f}{(n-2p)}$$

⁴⁷⁴ Note that under H_0 , $SSE_f/(n-2p)$ is a consistent estimator of σ^2 , which has the same limit ⁴⁷⁵ as $\hat{\sigma}^2$. Therefore, as $n \to \infty$, $T/F \to p$ in probability. If the two estimators of σ^2 are taken ⁴⁷⁶ to be the same, the map between F and T is established.

When the null hypothesis is true, the two estimator of σ^2 should be similar, so the two statistics gives similar p-values. Under the alternative hypothesis, the two estimator of σ^2 may differ; the p-values of the two statistics may not be very close.

480 References

- Asparouhov, T. and B. Muthen (2007). Testing for informative weights and weights trimming
 in multivariate modeling with survey data. In *Proceedings of the Joint Statistical Meetings*2007, Volume 2, Salt Lake City, pp. 3394–99.
- Bertolet, M. M. (2008). To weight or not to weight? Incorporating sampling designs into
 model-based analyses. Ph. D. thesis, Carnegie Mellon University.
- ⁴⁸⁶ Bollen, K. A., P. P. Biemer, A. F. Karr, S. Tueller, and M. E. Berzofsky (2016). Are survey
 ⁴⁸⁷ weights needed? A review of diagnostic tests in regression analysis. *Annual Review of*⁴⁸⁸ Statistics and Its Application 3(1), 375–392.
- Boudreau, C. and J. F. Lawless (2006). Survival analysis based on the proportional hazards
 model and survey data. *Canadian Journal of Statistics* 34 (2), 203–216.

- ⁴⁹¹ Breidt, F. J., J. D. Opsomer, W. Herndon, R. Cao, and M. Francisco Fernandez (2013).
 ⁴⁹² Testing for informativeness in analytic inference from complex surveys. In *Proceedings*⁴⁹³ 59th ISI World Statistics Congress, Hong Kong, pp. 889–893.
- ⁴⁹⁴ Chambers, R. L., A. H. Dorfman, and M. Sverchkov (2003). Nonparametric regression with
 ⁴⁹⁵ complex survey data. In R. L. Chambers and C. J. Skinner (Eds.), *Analysis of Survey*⁴⁹⁶ Data, Wiley series in survey methodology, Chapter 11, pp. 151–174. Chichester: Wiley.
- ⁴⁹⁷ Chatterjee, S. (2020). A new coefficient of correlation. Journal of the American Statistical
 ⁴⁹⁸ Association. Forthcoming.
- ⁴⁹⁹ Dumouchel, W. H. and G. J. Duncan (1983). Using sample survey weights in multiple
 ⁵⁰⁰ regression analyses of stratified samples. *Journal of the American Statistical Associa-*⁵⁰¹ tion 78(383), 535–543.
- Eideh, A. A. H. and G. Nathan (2006). Fitting time series models for longitudinal survey
 data under informative sampling. *Journal of Statistical Planning and Inference 136*(9),
 3052–3069.
- Frohlich, N., K. C. Carriere, L. Potvin, and C. Black (2001). Assessing socioeconomic effects
 on different sized populations: To weight or not to weight? *Journal of Epidemiology & Community Health* 55(12), 913–920.
- ⁵⁰⁸ Fuller, W. A. (2009). *Sampling Statistics*. Hoboken, New Jersey: John Wiley & Sons Inc.
- Gelman, A. (2007). Struggles with survey weighting and regression modeling. Statistical
 Science 22(2), 153–164.
- Gluschenko, K. (2018). Measuring regional inequality: To weight or not to weight? Spatial
 Economic Analysis 13(1), 36–59.
- Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica* 46(6), 1251–1271.

- Hsieh, C.-m. (2004). To weight or not to weight: The role of domain importance in quality
 of life measurement. Social Indicators Research 68(2), 163–174.
- Institute of Social Science Survey, Peking University (2015). China Family Panel Studies
 (CFPS). https://doi.org/10.18170/DVN/45LCSO.
- Kish, L. and M. R. Frankel (1974). Inference from complex samples. Journal of the Royal
 Statistical Society: Series B (Methodological) 36(1), 1–22.
- Kott, P. S. (1991). What does performing linear regression on sample survey data mean?
 Journal of Agricultural Economics Research 43(1), 30–33.
- Kott, P. S. (2018). A design-sensitive approach to fitting regression models with complex
 survey data. *Statistics Surveys 12*, 1–17.
- ⁵²⁴ Lumley, T. (2004). Analysis of complex survey samples. Journal of Statistical Soft-⁵²⁵ ware 9(i08), 1–19.
- Lumley, T. and A. Scott (2017). Fitting regression models to survey data. *Statistical Science* 32(2), 265–278.
- Ma, X., Z. Wang, and X. Liu (2019). Progress on catastrophic health expenditure in China:
 Evidence from China Family Panel Studies (CFPS) 2010 to 2016. International Journal
 of Environmental Research and Public Health 16(23), 4775.
- Nguyen, N. D. and P. Murphy (2015). To weight or not to weight? A statistical analysis of how weights affect the reliability of the quarterly national household survey for
 immigration research in ireland. *The Economic and Social Review* 46(4), 567–603.
- Nordberg, L. (1989). Generalized linear modeling of sample survey data. Journal of Official
 Statistics 5(3), 223–239.
- ⁵³⁶ Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. Interna ⁵³⁷ tional Statistical Review 61, 317–337.

- Pfeffermann, D. and G. Nathan (1985). Problems in model identification based on data from
 complex sample surveys. *Bulletin of the International Statistical Institute 51*(12.2), 1–12.
- Pfeffermann, D. and M. Sverchkov (1999). Parametric and semi-parametric estimation of
 regression models fitted to survey data. Sankhyā: The Indian Journal of Statistics, Series
 B 61(1), 166–186.
- Pfeffermann, D. and M. Sverchkov (2003). Fitting generalized linear models under informative sampling. In R. L. Chambers and C. J. Skinner (Eds.), *Analysis of Survey Data*,
 Wiley series in survey methodology, Chapter 12, pp. 175–196. Chichester: Wiley.
- Pfeffermann, D. and M. Sverchkov (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical*Association 102(480), 1427–1439.
- Pfeffermann, D. and M. Sverchkov (2010). Inference under informative sampling. In D. Pfeffermann and C. R. Rao (Eds.), *Sample Surveys: Inference and Analysis* (digital printing
 ed.)., Handbook of Statistics. Vol. 29 B:, Chapter 39, pp. 455–488. Amsterdam: Elsevier.
- Rabe-Hesketh, S. and A. Skrondal (2006). Multilevel modelling of complex survey data.
 Journal of the Royal Statistical Society: Series A (Statistics in Society) 169(4), 805–827.
- Smith, T. M. F. (1988). To weight or not to weight, that is the question. In J. M. Bernardo,
 M. H. Degroot, D. V. Lindley, and A. Smith (Eds.), *Bayesian statistics*, Volume 3, pp.
 437–451. Oxford: Oxford University Press.
- Solon, G., S. J. Haider, and J. M. Wooldridge (2015). What are we weighting for? *Journal*of Human resources 50(2), 301–316.
- Stefanski, L. A. and D. D. Boos (2002). The calculus of M-estimation. The American
 Statistician 56(1), 29–38.

- Tchetgen, E. J., M. M. Glymour, I. Shpitser, and J. Weuve (2012). Rejoinder: To weight or not to weight? On the relation between inverse-probability weighting and principal stratification for truncation by death. *Epidemiology* 23(1), 132–137.
- ⁵⁶⁴ Winship, C. and L. Radbill (1994). Sampling weights and regression analysis. Sociological
 ⁵⁶⁵ Methods & Research 23(2), 230-257.
- ⁵⁶⁶ Wu, Y. and W. A. Fuller (2005). Preliminary testing procedures for regression with survey
 ⁵⁶⁷ samples. In *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section*,
 ⁵⁶⁸ pp. 3683–3688. American Statistical Association.
- Xie, Y. and J. Hu (2014). An introduction to the China Family Panel Studies (CFPS). *Chinese Sociological Review* 47(1), 3–29.
- Xie, Y. and Y. Jin (2015). Household wealth in China. *Chinese Sociological Review* 47(3),
 203–229.
- Xie, Y. and P. Lu (2015). The sampling design of the China Family Panel Studies (CFPS). *Chinese Journal of Sociology* 1(4), 471–484.
- ⁵⁷⁵ Yan, J., R. H. Aseltine Jr, and O. Harel (2013). Comparing regression coefficients between
- ⁵⁷⁶ nested linear models for clustered data with generalized estimating equations. *Journal of*
- 577 Educational and Behavioral Statistics 38(2), 172–189.