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

Feng Wang ${ }^{1}$, HaiYing Wang ${ }^{2}$, Jun Yan ${ }^{2}$

1. School of Statistics, Shanxi University of Finance and Economics
2. Department of Statistics, University of Connecticut


#### Abstract

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


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

## 1 Introduction

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

A recent review by Bollen et al. (2016) classifies the tests for the necessity of weights in regression analysis into two groups. Tests in the first group are difference-in-coefficients (DC) tests, which examine whether the difference between the weighted and unweighted coefficients estimates is different from zero (Kott, 1991; Pfeffermann, 1993). Tests in the second group are weight-association (WA) tests, which examine whether the weight is informative about the response variable after conditioning on the covariates (Dumouchel and Duncan, 1983; Pfeffermann and Sverchkov, 1999; Wu and Fuller, 2005; Pfeffermann and Sverchkov, 2007). Bollen et al. (2016) conceptually reviewed the assumptions and properties of the
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 as tests of informative sampling process (Pfeffermann and Sverchkov, 2003, Section 12.2.2) or sampling ignorability (Pfeffermann and Sverchkov, 2010, Section 7). Pfeffermann and Nathan (1985) proposed a test based comparing the out-of-sample prediction power between the weighted and unweighted fits. A large difference in squared prediction errors indicates non-ignorable weights. Pfeffermann and Sverchkov (2003) proposed a test that compares the estimating equations with and without the weights. The estimating equations could be score equations if likelihood is specified, but can be more general without distributional assumption. Eideh and Nathan (2006) proposed to test based on the Kullback-Leibler information against exponential or linear inclusion probability models. It was claimed that the testing statistic follows a chi-squared distribution with one degree of freedom. In their simulation study, however, the null distribution of the test statistics appears to be far different from chi-squared with one degree of freedom (Eideh and Nathan, 2006, Table 3). This test thus needs a rigorous further investigation. Finally, Breidt et al. (2013) proposed a likelihood ratio (LR) test that compares two weighted log-likelihoods with different weights. The null distribution of the test statistic is a mixture of chi-squared distributions with one degree of freedom, The performances of these tests in comparison with those reviewed by Bollen et al. (2016) would be a great practical value.

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 literature as there is a lot to summarize. Our contribution is three-fold. First, we conduct a comprehensive numerical study to compare the size and power of a few commonly used tests and their variations in several scenarios. Such comparison has been long missing in the literature. Some tests need an auxiliary linear model, which may not pick up the nonlinear associations. Some tests were found to performed almost identically in the study, which led to our second contribution - we establish the equivalence of two powerful tests. 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. Finally, the tests are applied to a regression analysis of family expenditures with data from the China Family Panel Studies (CFPS) and its subsamples of different sizes.

## 2 Tests for Necessity of Weight in Regression

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

$$
\begin{equation*}
Y_{j}=X_{j} \beta+\epsilon_{j}, \quad j \in U \tag{1}
\end{equation*}
$$

where $Y_{j}$ is the response variable, $X_{j}$ is a $p \times 1$ covariate vector (including a component of 1 for intercept), $\beta$ is a $p \times 1$ vector of regression coefficients, and the regression error $\epsilon_{j}$ has mean zero and variance $\sigma^{2}$. The observed survey data $S$ of sample size $n$ is $\left\{\left(Y_{i}, X_{i}, W_{i}\right)\right.$ : $i=1, \ldots, n\}$, where $W_{i}$ is the survey weight associated with the $i$ th observation. Each weight $W_{i}$ may or may not be the inverse probability of selection. Let $Y=\left(Y_{1}, \ldots, Y_{n}\right)^{\top}$, $X=\left(X_{1}^{\top}, \ldots, X_{p}^{\top}\right)^{\top}, \epsilon=\left(\epsilon_{1}, \ldots, \epsilon_{n}\right)^{\top}$, and $W=\left(W_{1}, \ldots, W_{n}\right)^{\top}$. A working linear regression of $Y$ on $X$ for the survey data is

$$
\begin{equation*}
Y=X \beta+\epsilon \tag{2}
\end{equation*}
$$

We are interested in testing the necessity of weighting in fitting (2) to the observed data in estimating $\beta$, that is, testing whether an unweighted estimator for the $\beta$ in (2) is unbiased for the population parameter $\beta$ in regression (1). Based on the observed survey data. the least squares estimators of $\beta$ are $\hat{\beta}_{u}=\left(X^{\top} X\right)^{-1} X Y$ without weight and $\hat{\beta}_{w}=\left(X^{\top} H X\right)^{-1} X H Y$ with weight matrix $H=\operatorname{diag}(W)$. Tests for the necessity of weight attempt to answer the question whether or not to weight. We review six such tests in approximately chronological order as follows.

### 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

$$
\begin{equation*}
H_{0}: \mathbb{E}(Y \mid X, W)=\mathbb{E}(Y \mid X) \tag{3}
\end{equation*}
$$

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

$$
\begin{equation*}
\mathbb{E}(Y \mid X, W)=X \beta+H X \gamma \tag{4}
\end{equation*}
$$

where $\gamma$ is a $p \times 1$ coefficient vector of $H X$. Ignorable weight is tested by an $F$-test for $H_{0}: \gamma=0$ with testing statistic

$$
\begin{equation*}
F=\frac{\left(\mathrm{SSE}_{r}-\mathrm{SSE}_{f}\right) / p}{\operatorname{SSE}_{f} /(n-2 p)} \tag{5}
\end{equation*}
$$

where $\mathrm{SSE}_{r}$ and $\mathrm{SSE}_{f}$ 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 $F(p, n-2 p)$ distribution. Without the normality assumption, the null
distribution is asymptotically $F(p, n-2 p)$ for large $n$. Rejection of $\gamma=0$ implies that weights are informative; otherwise, there is no sufficient evidence against the unweighted analysis.

### 2.2 Pfeffermann-Nathan's Test Based on Predictive Power

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

$$
H_{0}: \mathbb{E}\left(v_{u i}^{2}-v_{w i}^{2}\right)=0, \quad i \in V .
$$

This hypothesis can be tested by standard $Z$-test with $Z=\bar{D} / S_{D}$, where $\bar{D}$ and $S_{D}^{2}$ are the sample mean and sample variance of $D_{i}$ 's, $i \in V$, with $D_{i}=v_{u i}^{2}-v_{w i}^{2}$.

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 $Z$ may drastically reduce its power. Both conjectures are observed in our numerical studies.

### 2.3 Hausman-Pfeffermann's DC Test

Pfeffermann (1993) proposed a DC test which directly compares $\hat{\beta}_{u}$ and $\hat{\beta}_{w}$ using a model specification test in econometrics studied by Hausman (1978). Hausman's test can be used 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 $\beta$ as the sample size $n$ increases. A DC test checks whether it holds that

$$
\begin{equation*}
H_{0}: \mathbb{E}\left(\hat{\beta}_{u}\right)=\mathbb{E}\left(\hat{\beta}_{w}\right), \tag{6}
\end{equation*}
$$

The test statistic is

$$
\begin{equation*}
T=\left(\hat{\beta}_{u}-\hat{\beta}_{w}\right)^{\top} \hat{V}^{-1}\left(\hat{\beta}_{u}-\hat{\beta}_{w}\right), \tag{7}
\end{equation*}
$$

where $\hat{V}$ is an estimate of $V=\mathbb{V}\left(\hat{\beta}_{u}-\hat{\beta}_{w}\right)$. The asymptotic null distribution of $T$ is $\chi_{p}^{2}$. 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 $\hat{d}_{i}^{2} / \hat{V}_{i i}, i=1, \ldots, p$, where $\hat{d}_{i}$ is the $i$ th component of $\hat{\beta}_{u}-\hat{\beta}_{w}$ and $\hat{V}_{I i}$ is the $i$ th component of the diagonal of $\hat{V}$. This statistics has asymptotic null distribution of $\chi_{1}^{2}$.

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

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 $\sigma^{2}$ 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 $T=p F$.

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

### 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 $H_{0 k}: \operatorname{Corr}\left(\hat{\epsilon}_{u}^{k}, W\right)=0, k=1,2,3$. 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 $\hat{\epsilon}_{u}^{k}$ :

$$
\begin{equation*}
\mathbb{E}\left(W \mid \hat{\epsilon}_{u}^{k}\right)=\alpha+\beta^{(k)} \hat{\epsilon}_{u}^{k}, \quad k=1,2,3, \tag{8}
\end{equation*}
$$

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

The tests of Pfeffermann and Sverchkov (1999) has two limitations. First, for $k=1,2,3$ together, 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 $X$ :

$$
\begin{equation*}
\mathbb{E}\left(W \mid \hat{\epsilon}_{u}\right)=f(X ; \eta)+\sum_{k=1}^{2} \beta^{(k)} \hat{\epsilon}_{u}^{k}+\operatorname{diag}\left(\hat{\epsilon}_{u}\right) X \gamma \tag{9}
\end{equation*}
$$

where $f(X ; \eta)$ is some function of $X$ with parameter $\eta, \beta^{(1)}$ and $\beta^{(2)}$ are scalars, $\delta$ is a $p \times 1$ coefficient vector for $X$, and $\gamma$ 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 $Y$ :

$$
\begin{equation*}
\mathbb{E}(W \mid X, Y)=X \eta+Y \gamma \tag{10}
\end{equation*}
$$

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

$$
\begin{equation*}
\mathbb{E}(W \mid X, Y)=f(X ; \eta)+\sum_{k=1}^{2} Y^{k} \gamma_{k} \tag{11}
\end{equation*}
$$

where $f(X ; \eta)$ is some function of $X$ with parameter $\eta, \gamma_{k}$ 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.

### 2.5 Pfeffermann-Sverchkov's Test Based on Estimating Equations

Pfeffermann and Sverchkov (2003) proposed a test that uses the estimating equations to estimate $\beta$. This test requires an auxiliary regression model for $W, \mathbb{E}(W \mid X)=f(X ; \eta)$, which is some function of $X$ with parameter $\eta$. The unweighted estimating function $\delta_{i}(\beta)=$ $X_{i}\left(Y_{i}-X_{i}^{\top} \beta\right), i \in S$. Let $\hat{W}_{i}$ be the fitted value of this regression. Define $q_{i}=W_{i} / \hat{W}_{i}$. Let
$R\left(X_{i} ; \beta\right)=\delta_{i}(\beta)-q_{i} \delta_{i}(\beta)$. Ignorable sampling weight means

$$
H_{0}: \mathbb{E}\left[R\left(X_{i} ; \beta\right)\right]=0
$$

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 $R\left(X_{i} ; \hat{\beta}_{u}\right)$ '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 buiding process for $W$.

### 2.6 Wu-Fuller's WA Test

Wu and Fuller (2005) proposes a WA test which takes a slightly different extended model than that in Dumouchel and Duncan (1983). Similar to Pfeffermann and Sverchkov (2003), this test also requires an auxiliary regression model for $W, \mathbb{E}(W \mid X)=f(X ; \eta)$. Let $Q=$ $\operatorname{diag}\left(q_{1}, \ldots, q_{n}\right)$, where $q_{i}$ 's are the same as defined in the last subsection. Consider an extended regression

$$
\mathbb{E}(Y \mid X, W)=X \beta+Q X \gamma
$$

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 contained in $X$. The definition of $q_{i}$ 's factors out the part in the weight $W_{i}$ that is predictable by $X_{i}$. If weight is informative for $Y$ after conditioning on $X$, then $Q X$ is expected to have a significantly nonzero coefficient $\gamma$ in the extended regression. Otherwise, one would expect $\gamma=0$. Implementing this test requires an auxiliary regression of $W$ on $X, f(x ; \eta)$. As for Pfeffermann and Sverchkov (2003), a model building process may be beneficial. Poor approximation for the relation between $W$ and $X$ might lead to incorrect size and poor power of the test.

### 2.7 LR Test

Breidt et al. (2013) proposed an LR test, which is neither a DC nor a WA test. A superpopulation model is assumed that have generated the finite population $U$. Suppose that the conditional distribution $Y_{i}$ given $X_{i}$ in the superpopulation has density $f\left(\cdot \mid X_{i} ; \theta\right)$ with parameter vector $\theta$ of dimension $q$ with true value $\theta_{0}$. Here $\theta$ contains $\beta$ as a subset. For example, if the distribution is normal, there is a variance parameter in addition to $\beta$ in $\theta$. Note that $\ln f\left(Y_{i} \mid X_{i} ; \theta\right)$ is the log-likelihood for the superpopulation distribution, but it may not be the log-likelihood for an observation in the sampled data. For convenience, we still call it log-likelihood as in Breidt et al. (2013).

A weighted $\log$-likelihood with a general weight vector $\omega=\left(\omega_{1}, \ldots, \omega_{n}\right)^{\top}$ is

$$
l(\theta ; \omega)=\sum_{i=1}^{n} \omega_{i} \ln f\left(Y_{i} \mid X_{i} ; \theta\right)
$$

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 statistics are considered:

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

where $J_{\omega}=\lim _{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^{n} \omega_{i} \mathcal{I}\left(x_{i} ; \theta_{0}\right), \omega \in\{U, W\}$, and $\mathcal{I}\left(x_{i} ; \theta_{0}\right)$ is the Fisher Information for the $i$ th observation. Under the null hypothesis of noninformative weight, $n^{1 / 2}\left(\hat{\theta}_{W}-\right.$ $\left.\hat{\theta}_{U}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0,-J_{U}^{-1}+J_{W}^{-1} K_{W} J_{W}^{-1}\right)$, where $K_{W}=\lim _{n \rightarrow \infty} \frac{1}{n} \sum_{i \in S} W_{i}^{2} \mathcal{I}\left(x_{i} ; \theta_{0}\right)$. 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 $\lambda_{\omega}$ is the vector of eigenvalues of

$$
\left(-J_{U}^{-1}+J_{W}^{-1} K_{W} J_{W}^{-1}\right)^{T / 2} J_{\omega}\left(-J_{U}^{-1}+J_{W}^{-1} K_{W} J_{W}^{-1}\right)^{1 / 2}
$$

and $Z_{j}{ }^{\prime} \mathrm{s}, j=1, \ldots, p$, 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 $\theta_{0}$, 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.

## 3 Simulation Studies

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

### 3.1 Study 1

The first study was adapted from Pfeffermann and Sverchkov (1999), A population of size $N=3,000$ was generated for $\left(Y_{i}, X_{i}\right)$ with a linear regression model

$$
\begin{equation*}
Y_{i}=1+X_{i}+\varepsilon_{i}, \quad i=1, \ldots, N, \tag{12}
\end{equation*}
$$

where $X_{i}$ 's were independently generated from the standard uniform distribution $\mathcal{U}(0,1)$ and $\varepsilon_{i}$ 's were independently generated from $\mathcal{N}\left(0, \sigma^{2}\right)$ with $\sigma \in\{0.1,0.2\}$. The levels of $\sigma$ 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

$$
\begin{equation*}
W_{i}=a Y_{i}+0.3 X_{i}+\delta U_{i}, \tag{13}
\end{equation*}
$$

where $U_{i}$ 's are independently drawn from $\mathcal{U}(0,1), \delta$ 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

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.

| $n$ | $\sigma$ | $\delta$ | $a$ | 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 |

held constant, higher $\delta$ leads to lower power because of more noise in the weight model (13). In contrast, higher $\sigma$ leads to higher power, which is expected as higher $\sigma$ 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 distribution of the regression error. To investigate this issue, we considered three additional distributions of $\epsilon_{i}$ in Equation (12): 1) gamma with shape 10 and scale $\sqrt{10 / \sigma^{2}} ; 2$ ) student $t$ with 5-degrees of freedom and scale $\left.\sqrt{5 / 3} \sigma^{2} ; 3\right)$ uniform $\left(0, \sqrt{12 / \sigma^{2}}\right)$. These distributions were centered by their means so that they have mean zero and variance $\sigma^{2}$, matching the first two moments of $\mathcal{N}\left(0, \sigma^{2}\right)$. Table 2 shows the empirical rejection percentage of the tests with $\sigma=0.1$ and $\delta=1$ under different error distributions. The LR test does not hold its size in the case of heavy-tailed regression error, $t$ distribution; under other distributions, it appears to hold its size. The performances of all other tests are robust to the error distribution, which is expected because their null distributions are asymptotically valid regardless of the error distribution. The relative performances of these tests remain in the same order as those under the normal regression error.

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

$$
\begin{equation*}
W_{i}=a\left(Y_{i}-1.5 a\right)^{2}+0.3 X_{i}-0.3 X_{i}^{2}+U_{i}, \tag{14}
\end{equation*}
$$

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

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.

| 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 |
| 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 |

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.

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.

| $n$ | $a$ | 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 |

### 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

$$
Y_{i}=0.5+X_{i}+e_{i} \quad i=1,2, \ldots,
$$

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

$$
W_{i}=a \eta\left(X_{i}\right)+b \eta\left(\psi e_{i}+(1-\psi) z_{i}\right),
$$

where $z_{i}$ was generated from $\mathcal{N}(0,0.5)$ independent of $e_{i}$, and

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

with parameters $(a, b, \psi)$. Function $\eta(\cdot)$ controls the nonlinear association between $W_{i}$ and $Y_{i}$ through $\psi e_{i}$. The weight is noninformative when $\psi=0$.

The simulation was designed with the following settings. Following Wu and Fuller (2005), 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 . Four levels of $a$ were considered: $\{0.25,0.5,0.75,1\}$. As $a$ increases, the correlation between $W_{i}$ and $X_{i}$ increases while the correlation between $W_{i}$ and $e_{i}$ decreases. Four levels of $\psi$ were considered, $\{0,0.1,0.2,0.3\}$; higher $\psi$ implies that $W_{i}$ is more informative for $Y_{i}$. Two sample sizes, $n \in\{100,200\}$, were attained by a Poisson sampling. That is, subject $i, i=1,2, \ldots$, is selected in the sample if $U_{i}<W_{i}$, where $U_{i}$ 's are independent $\mathcal{U}(0,1)$ variables, until the desired sample size is reached. In each configuration, 1000 replicates were generated. In each replicate, the population was regenerated before the sample was drawn.

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

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.

| $n$ | $a$ | $\psi$ | DD | PN | HP | 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 |  |
|  | 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 | 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 |  |
|  | 0.0 | 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 |  |

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

## 4 Consumption Expenditure of Chinese Families

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

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

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.

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

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

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 $m \in$ $\{300,500,1000\}$.

| $m$ | DD | PN | HP | 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 |

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

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

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

## 5 Discussion

Testing for necessity of weight in regression models arises frequently in practical analyses of survey data. Reviews on such tests exist (Bollen et al., 2016) but none compares their sizes and powers in simulation studies. We conducted a comprehensive numerical study to compare the sizes and powers of a few commonly used weight tests under various configurations. The results show that the test of Pfeffermann and Sverchkov (2007) is the most competitive overall in the settings considered. Nonetheless, it is easy to construct scenarios where this test completely looses its power; this happens when, for example, the weight has zero correlation with the regression error but have strong association with the squared regression error. For tests that require an auxiliary regression model for the weight, the size and power are affected by the specification of the auxiliary model. Most tests are robust to the distribution of the regression error except the likelihood ratio test, which has inflated size under a heavy-tailed error distribution. An interesting theoretical result is that the DC test of Pfeffermann (1993) and the WA test of Dumouchel and Duncan (1983) are equivalent if they use the estimate for the variance of the regression error. In addition, unlike those tests that rely an auxiliary regression whose misspecification may affect their performances, they have no additional model specification burden but give very competitive powers in our simulation study. These findings provide recommendations for choosing the tests in practice.

Our review suggests several future research directions. Whether or not to use weight is a general question applicable to all kinds analyses. This review only focuses on linear regression analyses. Similar diagnostic tests for generalized linear models (Nordberg, 1989; Lumley and Scott, 2017), survival models with censored data (Boudreau and Lawless, 2006), or exploratory data analysis and nonparametric regression (Chambers et al., 2003) merits further research. For tests based on correlations (Pfeffermann and Sverchkov, 1999), a new
measure of correlation that better distinguishes independence from zero correlation has the potential to perform better where linear correlation fails (Chatterjee, 2020). Most tests in the literature assumed independent homoscedastic data. In practice, however, many complex survey data have a clustered or nested data structure with possible heteroscedasticity. The dependence structure in such data, sometimes co-present with heteroscedasticity, adds considerable complexity to the estimation problem (e.g., Rabe-Hesketh and Skrondal, 2006; Kott, 2018) and, hence, diagnostic tests. The DC test can be extended to handle clustered data in the general framework of generalized estimating equations (Yan et al., 2013). The likelihood ratio test did not perform well in our study because of its dependence on correct distributional specifications. The derivation is likely to hold for M-estimation (Stefanski and Boos, 2002) where the likelihood specification is replaced with moment specifications. More efforts are needed to research on these immediate questions.

## 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).

## 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 $T=p F$ if the $\hat{\sigma}^{2}$ in $T$ is set to be the $\operatorname{SSE}_{f} /(n-2 p)$.

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 $\beta$ and $\gamma$ under the extended regression model (4). They satisfy the following normal equations:

$$
\begin{align*}
X^{\top} X \hat{\beta}+X^{\top} H X \hat{\gamma} & =X^{\top} Y,  \tag{15}\\
X^{\top} H X \hat{\beta}+X^{\top} H^{2} X \hat{\gamma} & =X^{\top} H Y . \tag{16}
\end{align*}
$$

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

$$
\begin{equation*}
\hat{\gamma}=\left(X^{\top} H X\right)^{-1} G^{-1}\left(\hat{\beta}_{w}-\hat{\beta}_{u}\right), \tag{17}
\end{equation*}
$$

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

$$
\begin{equation*}
\hat{\beta}=\hat{\beta}_{u}-\left(X^{\top} X\right)^{-1} G^{-1}\left(\hat{\beta}_{w}-\hat{\beta}_{u}\right) . \tag{18}
\end{equation*}
$$

The $\mathrm{SSE}_{r}$ and $\mathrm{SSE}_{f}$ are, respectively,

$$
\begin{aligned}
& \operatorname{SSE}_{r}=Y^{\top} Y-\hat{\beta}_{u}^{\top} X^{\top} Y, \\
& \operatorname{SSE}_{f}=Y^{\top} Y-\hat{\beta}^{\top} X^{\top} Y-\hat{\gamma}^{\top} X^{\top} H Y .
\end{aligned}
$$

Their difference is

$$
\begin{aligned}
\mathrm{SSE}_{r}-\mathrm{SSE}_{f} & =\left(\hat{\beta}-\hat{\beta}_{u}\right)^{\top} X^{\top} Y+\hat{\gamma}^{\top} X^{\top} H Y \\
& =-\left(\hat{\beta}_{w}-\hat{\beta}_{u}\right)^{\top} G^{-1}\left(X^{\top} X\right)^{-1} X^{\top} Y+\left(\hat{\beta}_{w}-\hat{\beta}_{u}\right)^{\top} G^{-1}\left(X^{\top} H X\right)^{-1} X^{\top} H Y \\
& =-\left(\hat{\beta}_{w}-\hat{\beta}_{u}\right)^{\top} G^{-1} \hat{\beta}_{u}+\left(\hat{\beta}_{w}-\hat{\beta}_{u}\right)^{\top} G^{-1} \hat{\beta}_{w} \\
& =\left(\hat{\beta}_{u}-\hat{\beta}_{w}\right)^{\top} G^{-1}\left(\hat{\beta}_{u}-\hat{\beta}_{w}\right),
\end{aligned}
$$

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

$$
\frac{\mathrm{SSE}_{r}-\mathrm{SSE}_{f}}{\hat{\sigma}^{2}}=T
$$

Combined with the DD test statistic (5), we have

$$
\frac{T}{F}=\frac{p}{\hat{\sigma}^{2}} \frac{\mathrm{SSE}_{f}}{(n-2 p)} .
$$

Note that under $H_{0}, \mathrm{SSE}_{f} /(n-2 p)$ is a consistent estimator of $\sigma^{2}$, which has the same limit as $\hat{\sigma}^{2}$. Therefore, as $n \rightarrow \infty, T / F \rightarrow p$ in probability. If the two estimators of $\sigma^{2}$ are taken to be the same, the map between $F$ and $T$ is established.

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

## 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.

Bollen, K. A., P. P. Biemer, A. F. Karr, S. Tueller, and M. E. Berzofsky (2016). Are survey weights needed? A review of diagnostic tests in regression analysis. Annual Review of 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.

Breidt, F. J., J. D. Opsomer, W. Herndon, R. Cao, and M. Francisco Fernandez (2013). Testing for informativeness in analytic inference from complex surveys. In Proceedings 59th ISI World Statistics Congress, Hong Kong, pp. 889-893.

Chambers, R. L., A. H. Dorfman, and M. Sverchkov (2003). Nonparametric regression with complex survey data. In R. L. Chambers and C. J. Skinner (Eds.), Analysis of Survey Data, Wiley series in survey methodology, Chapter 11, pp. 151-174. Chichester: Wiley.

Chatterjee, S. (2020). A new coefficient of correlation. Journal of the American Statistical Association. Forthcoming.

Dumouchel, W. H. and G. J. Duncan (1983). Using sample survey weights in multiple regression analyses of stratified samples. Journal of the American Statistical Association 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.

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.

Lumley, T. (2004). Analysis of complex survey samples. Journal of Statistical Software 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.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. International 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.

Winship, C. and L. Radbill (1994). Sampling weights and regression analysis. Sociological Methods $\mathcal{E B}^{2}$ Research 23(2), 230-257.

Wu, Y. and W. A. Fuller (2005). Preliminary testing procedures for regression with survey samples. In Proceedings of the Joint Statistical Meetings, Survey Research Methods Section, 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.

Yan, J., R. H. Aseltine Jr, and O. Harel (2013). Comparing regression coefficients between nested linear models for clustered data with generalized estimating equations. Journal of Educational and Behavioral Statistics 38(2), 172-189.

