DISSERTATION

TESTING AND ADJUSTING FOR INFORMATIVE SAMPLING IN SURVEY DATA

Submitted by

Wade Wilson Herndon

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2014

Doctoral Committee:

Advisor: F. Jay Breidt
Co-advisor: Jean Opsomer

Dan Cooley
Mary Meyer
Paul Doherty

ABSTRACT

TESTING AND ADJUSTING FOR
INFORMATIVE SAMPLING IN SURVEY DATA

Fitting models to survey data can be problematic due to the potentially complex sampling mechanism through which the observed data are selected. Survey weights have traditionally been used to adjust for unequal inclusion probabilities under the design-based paradigm of inference, however, this limits the ability of analysts to make inference of a more general kind, such as to characteristics of a superpopulation. The problems induced by the presence of a complex sampling design can be generally contained under the heading of *informative sampling*. To say that the sampling is informative is to say that the distribution of the data in the sample is different from the distribution of the data in the population. Two major topics relating to analyzing survey data with (potentially) informative sampling are addressed: testing for informativeness, and model building in the presence of informative sampling.

Chapter 2 addresses the problem of running formal tests for informative sampling in survey data. The major contribution contained here is to detail a new test for informative sampling. The test is shown to be widely applicable and straight-forward to implement in practice, and also useful compared to existing tests. The test is illustrated through a variety of empirical studies as well. These applications include a censored regression problem, linear regression, logistic regression, and fitting a gamma mixture model. Results from the analogous bootstrap test are also presented; these results agree with the analytic versions of the test. Alternative tests for informative sampling do in fact exist, however, the existing methods each have significant drawbacks and limitations which may be resolved in some situation with this new methodology, and overall the literature is quite sparse in this area. In a simulation study, the test is shown to have many desirable properties and

maintains high power compared to alternative tests. Also included is discussion about the limiting distribution of the test statistic under a sequence of local alternative hypotheses, and some extensions that are useful in connecting the work contained here with some of the previous work in the area. These extensions also help motivate the semiparametric methods considered in chapter 3.

In chapter 3, semiparametric methods are introduced for including design information in a regression model while staying within a model-based inferential framework. The ideas explored here attempt to exploit relationships between design variables (such as the sample inclusion probabilities) and model covariates. In order to account for the complex sampling design and (potential) bias in estimating model parameters, design variables are included as covariates and considered to be functions of the model covariates that can then be estimated in a design-based paradigm using nonparametric methods. The nonparametric method explored here is kernel smoothing with degree zero. In principle, other (and more complex) kinds of estimators could be used to estimate the functions of the design variables conditional on the model covariates, but the framework presented here provides asymptotic results for only the more simple case of kernel smoothing. The method is illustrated via empirical applications and also through a simulation study in which confidence band coverage rates from the semiparametric method are compared to those obtained through regular linear regression. The semiparametric estimator soundly outperforms the regression estimator.

ACKNOWLEDGEMENTS

This research was supported in part by the US National Science Foundation (SES-0922142).

DEDICATION

For John, Brice, Jacob, and Johnny Rad

TABLE OF CONTENTS

INDEX OF THEOREMS AND COROLLARIES

CHAPTER 2

CHAPTER 3

CHAPTER 1

INTRODUCTION

1.1 Overview

This paper explores two aspects of model building for survey data: testing for informative sampling in survey data, and semiparametric approaches to model building in the presence of informative sampling. Chapter 2 is primarily concerned with the issue of testing for design informativeness, and Chapter 3 proposes methods for fitting predictive models to survey data under informative selection. The material in Chapter 2 is joint work with my advisers, Jay Breidt and Jean Opsomer, along with Ricardo Cao and Mario Francisco-Fernández from the University of Corũna in Spain; as of this writing, much of the material from this introduction and Chapter 2 is intended to be submitted in condensed form to Biometrika. The chapters are structured so that the motivation of the problems comes first, followed by theoretical results and then simulation studies and applications to empirical data are presented. In this introductory chapter, I will discuss some ideas relating to missing data, both due to the sampling mechanism and also due to the response mechanism, and I will also develop the general concept of informative sampling and its presence in statistical literature. In particular, I will discuss testing for informative sampling in survey data and the literature that exists on this topic.

1.2 Modes of Inference for Survey Data

Surveys generate large quantities of data in a wide range of disciplines. A survey is typically designed to estimate characteristics of the particular finite population from which the sample is drawn. This context is referred to as *descriptive inference* for surveys. For large-scale surveys, a combination of statistical efficiency and cost considerations often results

in a complex sampling design that includes unequal inclusion probabilities, stratification and clustering. An extensive literature exists on how to incorporate these design complexities into appropriate descriptive inference methods. So-called *design-based* methods are the standard approach to construct estimates and perform inference in this context

It is also common for analysts to use survey data to answer scientific questions that are applicable more widely than for one particular finite population. In such situations, the questions concern characteristics of a statistical model describing relationships among variables, and the finite population is viewed as representing a realization from that model. This is referred to as *analytic inference* for surveys. Statisticians have long been aware of the fact that it is not appropriate to ignore survey considerations when doing analytic inference for survey data. Both design-based and model-based methods can be applied in this context, and there is currently still some disagreement as to which of these approaches is most appropriate. See Little (2004) and Pfeffermann (2011) for recent discussions of this topic.

1.3 Missing Data and Informative Sampling

There are two primary ways in which missing data can enter an analysis: through the selection mechanism, i.e. a unit is not sampled and measured, and the response mechanism, where a sampled unit may have partially or fully missing information. I would like to discuss a class of problems that arises from the presence of informative sampling in survey data: Informative sampling occurs when, due to the design complexities, the model that is true for the data given that they are included in the sample is not the same as the model for the population as a whole. It is also possible to have an informative non-response mechanism, and while this is a related and interesting problem, the focus here will be on informativeness coming from the selection mechanism. This is referred to variably as informative sampling, informative selection, and design informativeness.

Noninformative selection occurs when the underlying process that generates the population values is independent of the sample selection process. To establish a more rigorous definition, let $\boldsymbol{X}$ and $\boldsymbol{y}$ denote observed data (I will typically refer to these as "Model" variables), along with design information contained in $\boldsymbol{Z}$ and sample membership indicators, $\boldsymbol{I}$. This is standard notation for survey data; $\boldsymbol{Z}$ may contain design information such as strata, clusters, or modified inclusion probabilities, and $\boldsymbol{I}$ is a vector of zeros and ones indicating whether or not a unit is sampled. With our data structured this way, Chambers, Steel, Wang, and Welsh (2012) state that noninformative selection occurs when

$$f(\boldsymbol{I} \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z}) = f(\boldsymbol{I} \mid \boldsymbol{Z}). \tag{1}$$

For individual units indexed by k we can alternatively write

$$f(y_k \mid \boldsymbol{x}_k, \boldsymbol{z}_k, I_k = 1) = f(y_k \mid \boldsymbol{x}_k, \boldsymbol{z}_k), \tag{2}$$

which is a convenient and intuitive way to think about informative sampling especially in the context of regression or a similar conditional problem; it says that the distribution of the data given that they were sampled is the same as the distribution of the data in general.

There is one more concept relating to missingness which should be discussed here – the ideas of Missing at Random (MAR) and Missing Completely at Random (MCAR). These terms are typically applied in the context of nonresponse, but they are used analogously for the sampling mechanism and so will be defined here. The general approach (e.g. Chambers, Steel, Wang, and Welsh (2012)) is to say that the data are MCAR if the probability of being sampled is independent of all the other design and model information, that is

$$f(I_k \mid \boldsymbol{x}_k, y_k, \boldsymbol{z}_k) = f(I_k), \tag{3}$$

and missing at random if inclusion in the sample is independent of the response variable,

that is

$$f(I_k \mid \boldsymbol{x}_k, y_k, \boldsymbol{z}_k) = f(I_k \mid \boldsymbol{x}_k, \boldsymbol{z}_k). \tag{4}$$

One important distinction is that $\boldsymbol{Z}$ contains design information and $\boldsymbol{X}$ contains model information that comes from measurements on our sampled units. Later, especially in Chapter 3, I will make the distinction between a sampling process that is noninformative given $\boldsymbol{X}$ and noninformative given $\boldsymbol{X}$ and $\boldsymbol{Z}$ together.

1.4 Testing for Informative Sampling

Testing for *informative sampling* is a crucial component in choosing a suitable approach for performing analytic inference. If the design can be determined to be non-informative with respect to a particular postulated model, then it is reasonable to ignore the design in subsequent model fitting and analysis. On the other hand, if informativeness cannot be rejected, the analysis will need to account explicitly for the design complexities, which can be done either by staying within a design-based framework or by adjusting the model to incorporate design effects.

In this paper, I introduce a new method for testing the hypothesis of no design informativeness. I focus mostly on the application to the regression setting since that is the most common type of analytic inference for survey data, but the method is applicable to any likelihood-based analysis. While informativeness could in principle be assessed by directly comparing the population and sample distributions of model variables, this is almost never possible in practice because the analyst only has access to sample data, supplemented by survey weights and summary information about the sampling design such as stratum and cluster indicators. Hence, I will consider testing for the case in which only sample level information is available, and therefore we will rely on the weights and summary survey information.

A number of authors have previously considered testing for informativeness, but overall, the literature on this existing topic is quite sparse. An important class of tests is based on assessing the significance of the difference between weighted and unweighted estimates of model parameters. This idea forms the basis of the procedures proposed by DuMouchel and Duncan (1983) and Fuller (1984) for the coefficients in linear regression. Pfeffermann (1993) extended this to general likelihood-based problems with explicit estimators, and Pfeffermann and Sverchkov (2003) to estimators that are defined as the solutions to estimating equations. The procedure I will present most closely relates to these types of tests but is connected more directly with the model likelihood. I will return to a comparison of the new procedure with these others in Chapter 2.

When the postulated model is a linear regression model, the test based on the difference between weighted and unweighted estimated coefficients is equivalent to an F-test for the significance of the parameters of an *expanded* linear model, with the extension composed of the interactions between the covariates of the original model and the weights. See Fuller (2009, Section 6.3.1) for a derivation of this equivalence. Testing based on comparing the postulated model with an extended version of the model was also used by Nordberg (1989) for logistic regression. In chapter 2 I present results that justify this approach by generalizing the F-test result to any test of submodels for linear parameters. In logistic regression, for example, the classical likelihood ratio test is used to test a full vs. a reduced model, and this is shown to be equivalent to testing for no difference between the weighted vs. unweighted parameter estimates, under certain conditions. These conditions are analogous to those in the original result by DuMouchel and Duncan (1983).

Another class of tests targets the moments of the postulated model rather than the model parameters, and tries to evaluate whether they are equal to the moments of the model that holds for the sample data. This is generally done in the regression context, so that the relevant moments are conditional on model covariates. Pfeffermann and Sverchkov (1999) show that the hypothesis of equal conditional moments for both models is equivalent to

lack of correlation between the model errors and the sampling weights, and uses classical correlation test statistics to test this hypothesis. This testing procedure is easy to apply but is not exact, in the sense that it is generally not clear how many moments should be compared. Pfeffermann and Sverchkov (1999) noted that "in practice, it would normally suffice to test the first 2–3 correlations." A more serious problem is the difficulty of having to interpret multiple tests simultaneously, so that the overall confidence level of the procedure is typically unknown.

A final class of tests is based on an identity in Pfeffermann and Sverchkov (1999), which shows that the difference between the postulated model and the sample model can be assessed through a regression of the survey weights on the model variables. This class of tests targets the informativeness directly, but requires that a model relating the weights and the model variables be defined, so that it is subject to its own possible model specification bias.

As an illustration of analytic inference and the effect of informativeness, consider the following "textbook" example. Korn and Graubard (1999, Example 4.3-1) describe an analysis of data relating gestational age to birthweight in the 1988 National Maternal and Infant Health Survey (NMIHS). The NMIHS was conducted by the US National Center for Health Statistics with a goal of studying factors that are related to poor pregnancy outcomes. The study used a nationally-representative stratified sample from birth records, with deliberate oversampling of low-birthweight infants. Fuller (2009, Example 6.3.1) simulates data to mimic properties of NMIHS. The simulated data are a stratified simple random sample in 18 strata, with five observations per stratum. They reflect key properties of the real data: a strong functional relationship between birthweight and gestational age (in weeks) and an informative design.

Let $I_k = 1$ if birth record k is selected, and 0 otherwise, where $k \in U = \{1, \ldots, N\}$, the finite population of all birth records. Using the terminology traditional in survey statistics, suppose an analyst is interested in fitting a "superpopulation" model: a stochastic model

assumed to have generated the measurements in the finite population. Initially, the analyst considers simple linear regression of y_k = gestational age on x_k = birthweight, with normal errors. An immediate concern is that the selection may have been "informative" in the sense of distorting the regression relationship between the variables, so that the conditional distribution of gestational age given birthweight in the superpopulation model, $f(y \,|\, x; \boldsymbol{\theta})$, is different from the relationship in the sample, $f(y \,|\, x, I = 1; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ contains the linear model coefficients and variance term. Indeed, with $\pi_k = \Pr[I_k = 1 \,|\, x_k, y_k]$, we have via Bayes' rule

$$\begin{aligned} f(y \,|\, x, I = 1; \boldsymbol{\theta}) &= \frac{\Pr[I = 1 \,|\, x, y]}{\int \Pr[I = 1 \,|\, x, y]\, f(y \mid x; \boldsymbol{\theta})\, dy} f(y \mid x; \boldsymbol{\theta}) \\ &= \frac{\mathrm{E}[\pi \mid x, y]}{\int \mathrm{E}[\pi \,|\, x, y]\, f(y \mid x; \boldsymbol{\theta})\, dy} f(y \mid x; \boldsymbol{\theta}). \end{aligned}$$

The leading factor depends on $\boldsymbol{\theta}$, and in general cannot be ignored for inference on $\boldsymbol{\theta}$. However, if in fact $\mathrm{E}[\pi \mid x, y]$ is independent of y, the leading factor cancels and we have a non-informative design.

As noted in the previous section, a standard approach to estimation and inference under possibly informative selection is to use sampling weights w_k provided with the survey data set. These weights are typically adjusted versions of the inverse inclusion probabilities, π_k^{-1}, and have the property that the weighted sample quantity $\sum_{k \in U} w_k I_k g(x_k, y_k) / \sum_{k \in U} w_k I_k$ is design consistent (with respect to the random selection mechanism) for the corresponding finite population quantity $\sum_{k \in U} g(x_k, y_k)/N$. There are no selection effects in the latter quantity, so it will in turn be model consistent (with respect to the superpopulation model) for $\mathrm{E}[g(x, y) \mid x]$. The same consistency properties hold for solutions of weighted estimating equations. For the NMIHS superpopulation model, the solution comes from maximizing the

weighted Gaussian log-likelihood,

$$l_w(\boldsymbol{\theta}) = l_w(\boldsymbol{\beta}, \sigma^2) = \sum_{k \in U} w_k \ln \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_k - (1, x_k)\boldsymbol{\beta})^2}{2\sigma^2} \right) \right\} I_k,$$

which yields the weighted least squares (WLS) estimator of the regression coefficients,

$$\hat{\boldsymbol{\beta}}_w = \underset{\boldsymbol{\beta} \in \mathbb{R}^2}{\operatorname{argmin}} \sum_{k \in U} w_k \left(y_k - (1, x_k)\boldsymbol{\beta}\right)^2 I_k. \tag{5}$$

The WLS estimator is consistent for the finite population quantity,

$$\hat{\boldsymbol{\beta}}_N = \underset{\boldsymbol{\beta} \in \mathbb{R}^2}{\operatorname{argmin}} \sum_{k \in U} \left(y_k - (1, x_k)\boldsymbol{\beta}\right)^2,$$

which is in turn consistent for $\boldsymbol{\beta}$. For the NMIHS data, $\hat{\boldsymbol{\beta}}_w = (28.974, 0.297)^T$ with corresponding design-based standard errors $(0.426, 0.013)^T$. These standard errors reflect the sample-to-sample variability of $\hat{\boldsymbol{\beta}}_w$ as an estimator of $\hat{\boldsymbol{\beta}}_N$. If the difference between $\hat{\boldsymbol{\beta}}_N$ and $\boldsymbol{\beta}$ is negligible (a common assumption when the sampling fraction is small), then these standard errors can also be interpreted as being valid for the difference between $\hat{\boldsymbol{\beta}}_w$ and $\boldsymbol{\beta}$, albeit under a different mode of inference.

While this design-based approach ensures that a valid estimator is available for model parameters of interest, it has a few major drawbacks. First, it requires the use of specialized software; for example, even though the estimator (5) has the form of a WLS estimator, the inferential framework is based on the design, not on a heteroskedastic linear model. Using non-survey software will result in the same point estimates but incorrect standard errors and tests. Second, if the design is in fact non-informative, the estimator $\hat{\boldsymbol{\theta}}_w = (\hat{\boldsymbol{\beta}}_w, \hat{\sigma}^2_w)$ is inefficient compared to the (unweighted) maximum likelihood estimator,

$$\hat{\boldsymbol{\theta}}_s \;=\; \underset{\boldsymbol{\theta} \in \mathbb{R}^3}{\operatorname{argmax}}\, l_s(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta} \in \mathbb{R}^3}{\operatorname{argmax}} \sum_{k \in U} \ln \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_k - (1, x_k)\boldsymbol{\beta})^2}{2\sigma^2} \right) \right\} I_k, \tag{6}$$

which yields the ordinary least squares (OLS) estimator of the regression coefficients,

$$\hat{\boldsymbol{\beta}}_s = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{k\in U} (y_k - (1, x_k)\boldsymbol{\beta})^2 \, I_k. \tag{7}$$

For the NMIHS data, $\hat{\boldsymbol{\beta}}_s = (25.765, 0.370)^T$, with standard errors $(0.389, 0.012)^T$.

But perhaps the most fundamental issue is that, while the design-based approach provides estimators of the model parameters and (asymptotic) inference tools associated with the estimators, *it does not actually provide a model for the sample data.* Hence, many methods that rely on the model and are in common use among data analysts do not apply. These include model selection methods, residual diagnostic tests and plots, and prediction methods, to name a few. While the design-based approach is a possible solution to the problem of informativeness, there is a clear desire for alternative solutions that account for the informative selection yet stay within a model-based mode of estimation and inference. In particular, analysts often want to assess whether they are allowed to "ignore" the design when fitting a model. In the NMIHS example, they would like to use $\hat{\boldsymbol{\beta}}_s$ to estimate $\boldsymbol{\beta}$ and use the traditional OLS variance estimator for inference.

In this linear regression setting, DuMouchel and Duncan (1983) recommended an F-test procedure to determine the informativeness of the design with respect to the estimation of $\boldsymbol{\beta}$, in the sense of determining whether $\mathrm{E}\left[\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_s\right] = \mathbf{0}$. This is the standard F-test of the full model

$$H_a: \quad \mathrm{E}\left[y_k \mid x_k, w_k\right] = \beta_0 + \beta_1 x_k + \gamma_0 w_k + \gamma_1 w_k x_k \tag{8}$$

versus the reduced model

$$H_0: \quad \mathrm{E}\left[y_k \mid x_k, w_k\right] = \beta_0 + \beta_1 x_k. \tag{9}$$

For the NMIHS data, the test statistic is $F = 55.591$ on 2 numerator and 88 denominator degrees of freedom, with a p-value much less than 0.001, strongly rejecting the null hypothesis of non-informative selection.

This test is simple and efficient when the effects of informative selection can be described with an expanded mean structure. But the effects of informative selection may appear elsewhere in the model structure, since any informativeness with respect to other model parameters (σ^2 in particular) is not captured by this test. We therefore present a new test for informative selection based on comparing the log-likelihood at the weighted maximum likelihood estimates, $\hat{\boldsymbol{\theta}}_w$, to the log-likelihood at the unweighted maximum likelihood estimates, $\hat{\boldsymbol{\theta}}_s$. We also derive the asymptotic properties of the test and develop a bootstrap version, and I will then return to the NMIHS data for an empirical example, and also illustrate with a Tobit regression on data from the National Health and Nutrition Examination Survey (NHANES). Simulation experiments compare Dumouchel and Duncan type tests to the new proposal, illustrating its size and power properties in both its asymptotic and bootstrap versions.

CHAPTER 2

TESTING FOR INFORMATIVE SAMPLING IN SURVEY DATA

2.1 Theoretical Results Under the Null Hypothesis

Here theoretical arguments for the new test for informative sampling will be presented, and in what follows, a number of potential uses will be presented, along with simulation results and empirical studies. To establish the theoretical results for the new test, consider a sequence of finite populations indexed by population size, N. Throughout, we condition on $\boldsymbol{X}_N = [\boldsymbol{x}_k^T]_{k\in U}, \boldsymbol{I}_N = [I_k]_{k\in U}$, corresponding to the standard regression setting in which only the conditional distributions of (selected) responses y_k given $\boldsymbol{x}_k$ are of interest. The marginal distribution of $(\boldsymbol{X}_N, \boldsymbol{I}_N)$ may become important in the case of informative selection, but here we are deriving properties of the test statistics under the null hypothesis of non-informative selection. We consider a non-negative weighting sequence $\{w_k\}_{k\in U}$ that is completely determined by $(\boldsymbol{X}_N, \boldsymbol{I}_N)$, such as design weights $w_k = \pi_k^{-1}$ or truncated regression weights,

$$w_k = \max\left\{\frac{1}{\pi_k} + \left(\sum_{k\in U}\boldsymbol{z}_k^T - \sum_{k\in U}\frac{\boldsymbol{z}_k^T I_k}{\pi_k}\right)\left(\sum_{k\in U}\frac{\boldsymbol{z}_k\boldsymbol{z}_k^T I_k}{\pi_k}\right)^{-1}\frac{\boldsymbol{z}_k}{\pi_k}, \delta\right\}$$

for some subvector $\boldsymbol{z}_k$ of $\boldsymbol{x}_k$ and some $\delta \geq 0$. In what follows, a is used as generic notation for either the unweighted case, with $a = 1$ denoting $\{a_k\} \equiv 1$, or the weighted case, with $a = w$ denoting $\{a_k\} = \{w_k\}$. We introduce notation to allow consideration of both original data and parametric bootstrap samples. For every N, we consider independent random variables with probability density functions $f(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_\dagger)$ satisfying the following regularity conditions:

A1. $f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}\right)$ is log-concave in $\boldsymbol{\theta}$.

A2. Under $f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_\dagger\right)$, the Fisher information for the kth observation is given by

$$\begin{aligned}\boldsymbol{\mathcal{I}}(\boldsymbol{x}_k; \boldsymbol{\theta}_\dagger) &= \mathrm{Var}\left(\left. \frac{\partial \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\dagger} \middle| \boldsymbol{x}_k; \boldsymbol{\theta}_\dagger \right) \\ &= \mathrm{E}\left[\left. \frac{-\partial^2 \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\dagger} \middle| \boldsymbol{x}_k; \boldsymbol{\theta}_\dagger \right].\end{aligned}$$

A3. $\ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}\right)$ admits the expansion

$$\begin{aligned}& -\ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_\dagger + N^{-1/2}\boldsymbol{u}\right) + \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_\dagger\right) \\ &= \frac{\boldsymbol{u}^T}{N^{1/2}} \left. \frac{-\partial \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\dagger} \\ &\quad + \left\{ \frac{\boldsymbol{u}^T}{2N} \left. \frac{-\partial^2 \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\dagger} \boldsymbol{u} + r_k\left(y_k, \frac{\boldsymbol{u}}{N^{1/2}}\right) \right\} \\ &= \frac{\boldsymbol{u}^T}{N^{1/2}} \boldsymbol{D}(y_k, \boldsymbol{x}_k; \boldsymbol{\theta}_\dagger) + \boldsymbol{R}\left(y_k, \boldsymbol{x}_k, N^{-1/2}\boldsymbol{u}; \boldsymbol{\theta}_\dagger\right) \qquad (10)\end{aligned}$$

where

$$\begin{aligned}\mathrm{E}\left[\boldsymbol{D}(y_k, \boldsymbol{x}_k; \boldsymbol{\theta}_\dagger) \mid \boldsymbol{X}_N, \boldsymbol{I}_N; \boldsymbol{\theta}_\dagger\right] &= \mathbf{0}, \\ \mathrm{E}\left[\boldsymbol{R}\left(y_k, \boldsymbol{x}_k, N^{-1/2}\boldsymbol{u}; \boldsymbol{\theta}_\dagger\right) \middle| \boldsymbol{X}_N, \boldsymbol{I}_N; \boldsymbol{\theta}_\dagger\right] &= \frac{\boldsymbol{u}^T}{2N^{1/2}} \boldsymbol{\mathcal{I}}(\boldsymbol{x}_k; \boldsymbol{\theta}_\dagger) \frac{\boldsymbol{u}}{N^{1/2}} \\ &\quad + v_{k,0}\left(\frac{\boldsymbol{u}}{N^{1/2}}; \boldsymbol{\theta}_\dagger\right), \\ \mathrm{Var}\left(\boldsymbol{R}\left(y_k, \boldsymbol{x}_k, N^{-1/2}\boldsymbol{u}; \boldsymbol{\theta}_\dagger\right) \mid \boldsymbol{X}_N, \boldsymbol{I}_N; \boldsymbol{\theta}_\dagger\right) &= v_k\left(N^{-1/2}\boldsymbol{u}; \boldsymbol{\theta}_\dagger\right) \\ \sup_{\boldsymbol{\theta}: \|\boldsymbol{\theta}-\boldsymbol{\theta}_\dagger\| \le \eta} \sum_{k \in U} a_k I_k v_{k,0}\left(N^{-1/2}\boldsymbol{u}; \boldsymbol{\theta}\right) &= o(1) \\ \sup_{\boldsymbol{\theta}: \|\boldsymbol{\theta}-\boldsymbol{\theta}_\dagger\| \le \eta} \sum_{k \in U} a_k^2 I_k v_k\left(N^{-1/2}\boldsymbol{u}; \boldsymbol{\theta}\right) &= o(1)\end{aligned}$$

for some $\eta > 0$, for both $a_k \equiv 1$ and $a_k = w_k$, and for all $\boldsymbol{u} \in \mathbb{R}^p$.

A4. As $N \to \infty$,

$$\widehat{\boldsymbol{K}_a}(\boldsymbol{\theta}_\dagger) = \frac{1}{N}\sum_{k\in U} a_k^2 I_k \boldsymbol{\mathcal{I}}(\boldsymbol{x}_k;\boldsymbol{\theta}_\dagger) \to \boldsymbol{K}_a(\boldsymbol{\theta}_\dagger)$$

and

$$\widehat{\boldsymbol{J}_a}(\boldsymbol{\theta}_\dagger) = \frac{1}{N}\sum_{k\in U} a_k I_k \boldsymbol{\mathcal{I}}(\boldsymbol{x}_k;\boldsymbol{\theta}_\dagger) \to \boldsymbol{J}_a(\boldsymbol{\theta}_\dagger)$$

where $\boldsymbol{J}_a(\boldsymbol{\theta}_\dagger)$ is positive definite and $\boldsymbol{K}_1(\boldsymbol{\theta}_\dagger) = \boldsymbol{J}_1(\boldsymbol{\theta}_\dagger)$.

A5. Under $f\left(y_k \mid \boldsymbol{x}_k;\boldsymbol{\theta}_\dagger\right)$,

$$\frac{1}{N^{1/2}}\sum_{k\in U}\begin{bmatrix} \boldsymbol{D}(y_k,\boldsymbol{x}_k;\boldsymbol{\theta}_\dagger) \\ w_k\boldsymbol{D}(y_k,\boldsymbol{x}_k;\boldsymbol{\theta}_\dagger)\end{bmatrix} I_k \xrightarrow{\mathcal{L}} \mathcal{N}\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{K}_1(\boldsymbol{\theta}_\dagger) & \boldsymbol{J}_w(\boldsymbol{\theta}_\dagger)\\ \boldsymbol{J}_w(\boldsymbol{\theta}_\dagger) & \boldsymbol{K}_w(\boldsymbol{\theta}_\dagger)\end{bmatrix}\right)$$

as $N \to \infty$.

A6. If $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_\dagger$ under $\boldsymbol{\theta}_\dagger$, then $\widehat{\boldsymbol{K}_a}\left(\widehat{\boldsymbol{\theta}}\right) \xrightarrow{P} \boldsymbol{K}_a\left(\boldsymbol{\theta}_\dagger\right)$ and $\widehat{\boldsymbol{J}_a}\left(\widehat{\boldsymbol{\theta}}\right) \xrightarrow{P} \boldsymbol{J}_a\left(\boldsymbol{\theta}_\dagger\right)$.

Theorem 1. *Suppose $\{y_k\}_{k\in U}$ are independent random variables with $y_k \sim f(\cdot \mid \boldsymbol{x}_k;\boldsymbol{\theta}_0)$. Let $\widehat{\boldsymbol{\theta}}_a$ be the maximizer of the log-likelihood criterion*

$$l_a(\boldsymbol{\theta}) = \sum_{k\in U} a_k I_k \ln f\left(y_k \mid \boldsymbol{x}_k;\boldsymbol{\theta}\right). \tag{11}$$

Then, under A1–A4,

$$N^{1/2}\left(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_0\right) = \boldsymbol{J}_a^{-1}\frac{1}{N^{1/2}}\sum_{k\in U} a_k I_k \boldsymbol{D}(y_k,\boldsymbol{x}_k;\boldsymbol{\theta}_0) + o_P(1). \tag{12}$$

If A5 also holds, then

$$N^{1/2}\begin{pmatrix}\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0\\ \widehat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_0\end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{J}_1^{-1} & \boldsymbol{J}_1^{-1}\\ \boldsymbol{J}_1^{-1} & \boldsymbol{J}_w^{-1}\boldsymbol{K}_w\boldsymbol{J}_w^{-1}\end{bmatrix}\right), \tag{13}$$

where $\boldsymbol{J}_a = \boldsymbol{J}_a(\boldsymbol{\theta}_0)$ and $\boldsymbol{K}_a = \boldsymbol{K}_a(\boldsymbol{\theta}_0)$.

Theorem 2. *Let $l_a(\cdot)$ and $\widehat{\boldsymbol{\theta}}_a$ be as defined in Theorem 1. Under A1–A4,*

$$T_1 = 2\left\{l_1\left(\widehat{\boldsymbol{\theta}}_1\right) - l_1\left(\widehat{\boldsymbol{\theta}}_w\right)\right\} = N\left(\widehat{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_w\right)^T \boldsymbol{J}_1\left(\widehat{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_w\right) + o_P(1)$$

and

$$T_w = 2\left\{l_w\left(\widehat{\boldsymbol{\theta}}_w\right) - l_w\left(\widehat{\boldsymbol{\theta}}_1\right)\right\} = N\left(\widehat{\boldsymbol{\theta}}_w - \widehat{\boldsymbol{\theta}}_1\right)^T \boldsymbol{J}_w\left(\widehat{\boldsymbol{\theta}}_w - \widehat{\boldsymbol{\theta}}_1\right) + o_P(1),$$

where the probability is with respect to the parametric distribution indexed by $\boldsymbol{\theta}_0$. Under the additional assumption A5,

$$N^{1/2}\left(\widehat{\boldsymbol{\theta}}_w - \widehat{\boldsymbol{\theta}}_1\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, -\boldsymbol{J}_1^{-1} + \boldsymbol{J}_w^{-1}\boldsymbol{K}_w\boldsymbol{J}_w^{-1}\right) = \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Gamma}\right), \tag{14}$$

so that

$$T_a \xrightarrow{\mathcal{L}} \sum_{j=1}^{p} \lambda_{aj} Z_j^2 \tag{15}$$

where λ_{aj} are the eigenvalues of $\boldsymbol{\Gamma}^{T/2}\boldsymbol{J}_a\boldsymbol{\Gamma}^{1/2}$, and $\{Z_j\}_{j=1}^p$ are independent and identically distributed $\mathcal{N}(0,1)$.

A closely-related test statistic to T_a is the quadratic form

$$N^{-1/2}\left(\widehat{\boldsymbol{\theta}}_w - \widehat{\boldsymbol{\theta}}_1\right)^T \left\{-\widehat{\boldsymbol{J}_1}^{-1} + \widehat{\boldsymbol{J}_w}^{-1}\widehat{\boldsymbol{K}_w}\widehat{\boldsymbol{J}_w}^{-1}\right\}^{-1} N^{-1/2}\left(\widehat{\boldsymbol{\theta}}_w - \widehat{\boldsymbol{\theta}}_1\right),$$

which is given in equation (4.3) of Pfeffermann (1993), along with its limiting chi-squared distribution under the null hypothesis that $\mathrm{E}\left[\widehat{\boldsymbol{\theta}}_w - \widehat{\boldsymbol{\theta}}_1; \boldsymbol{\theta}_0\right] = \mathbf{0}$. The statement in that paper that "The V–C [variance-covariance] matrices ... can be obtained by estimating the corresponding randomization V–C matrices" is ambiguous. With variance-covariance matrices estimated by the plug-in methods of A6, the limiting behavior of Pfeffermann's test statistic is an immediate corollary of (14) in Theorem 2.

Corollary 3. *Under A1–A6,*

$$\left(\widehat{\boldsymbol{\theta}}_w - \widehat{\boldsymbol{\theta}}_1\right)^T \left\{-\widehat{\boldsymbol{J}_1}^{-1}\left(\widehat{\boldsymbol{\theta}}_a\right) + \widehat{\boldsymbol{J}_w}^{-1}\left(\widehat{\boldsymbol{\theta}}_a\right)\widehat{\boldsymbol{K}_w}\left(\widehat{\boldsymbol{\theta}}_a\right)\widehat{\boldsymbol{J}_w}^{-1}\left(\widehat{\boldsymbol{\theta}}_a\right)\right\}^{-1}\left(\widehat{\boldsymbol{\theta}}_w - \widehat{\boldsymbol{\theta}}_1\right) \xrightarrow{\mathcal{L}} \chi^2_p, \quad (16)$$

a chi-squared distribution with p degrees of freedom.

Both Pfeffermann's test statistic and the asymptotic distribution of T_a require consistent estimation of $\boldsymbol{J}_a$ and $\boldsymbol{K}_a$, which is possible via plug-in methods under A6.

2.2 Testing Using Bootstrap Methods

Alternatively, the distribution of T_a may be approximated via parametric bootstrap, which does not require estimation of the covariance matrices. Our parametric bootstrap consists of sampling $\{y_k^*\}_{k\in U}$ as independent random variables with $y_k^* \sim f(\cdot \mid \boldsymbol{x}_k; \widehat{\boldsymbol{\theta}}_w)$ or $y_k^* \sim f(\cdot \mid \boldsymbol{x}_k; \widehat{\boldsymbol{\theta}}_1)$; both are possible because the bootstrap distribution of interest is computed under the null hypothesis. With $\boldsymbol{\theta}_\dagger = \widehat{\boldsymbol{\theta}}_a$, we then have immediately the following bootstrap analogues of Theorems 1 and 2:

Theorem 4. *Suppose $\{y_k^*\}_{k\in U}$ are independent random variables with $y_k^* \sim f(\cdot \mid \boldsymbol{x}_k; \widehat{\boldsymbol{\theta}}_a)$. Let $\widehat{\boldsymbol{\theta}}_a^*$ be the maximizer of the log-likelihood criterion*

$$l_a^*(\boldsymbol{\theta}) = \sum_{k\in U} a_k I_k \ln f\left(y_k^* \mid \boldsymbol{x}_k; \boldsymbol{\theta}\right). \quad (17)$$

Then, under A1–A4,

$$N^{1/2}\left(\widehat{\boldsymbol{\theta}}_a^* - \widehat{\boldsymbol{\theta}}_a\right) = -\boldsymbol{J}_a^{-1}\left(\widehat{\boldsymbol{\theta}}_a\right)\frac{1}{N^{1/2}}\sum_{k\in U} a_k I_k \boldsymbol{D}(y_k, \boldsymbol{x}_k; \widehat{\boldsymbol{\theta}}_a) + o_{P*}(1), \quad (18)$$

where the probability is with respect to the parametric distribution at the fixed value $\widehat{\boldsymbol{\theta}}_a$. If

A5 also holds, then

$$N^{1/2}\begin{pmatrix}\widehat{\boldsymbol{\theta}}_1^* - \widehat{\boldsymbol{\theta}}_a \\ \widehat{\boldsymbol{\theta}}_w^* - \widehat{\boldsymbol{\theta}}_a\end{pmatrix} \overset{\mathcal{L}_*}{\to} \mathcal{N}\left(\begin{bmatrix}\mathbf{0} \\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{J}_1^{-1}\left(\widehat{\boldsymbol{\theta}}_a\right) & \boldsymbol{J}_1^{-1}\left(\widehat{\boldsymbol{\theta}}_a\right) \\ \boldsymbol{J}_1^{-1}\left(\widehat{\boldsymbol{\theta}}_a\right) & \boldsymbol{J}_w^{-1}\left(\widehat{\boldsymbol{\theta}}_a\right)\boldsymbol{K}_w\left(\widehat{\boldsymbol{\theta}}_a\right)\boldsymbol{J}_w^{-1}\left(\widehat{\boldsymbol{\theta}}_a\right)\end{bmatrix}\right). \quad (19)$$

Theorem 5. *Let $l_a^*(\cdot)$ and $\widehat{\boldsymbol{\theta}}_a^*$ be as defined in Theorem 4. Under A1–A4,*

$$T_1^*\left(\widehat{\boldsymbol{\theta}}_a\right) = 2\left\{l_1^*\left(\widehat{\boldsymbol{\theta}}_1^*\right) - l_1^*\left(\widehat{\boldsymbol{\theta}}_w^*\right)\right\} = N\left(\widehat{\boldsymbol{\theta}}_1^* - \widehat{\boldsymbol{\theta}}_w^*\right)^T \boldsymbol{J}_1\left(\widehat{\boldsymbol{\theta}}_a\right)\left(\widehat{\boldsymbol{\theta}}_1^* - \widehat{\boldsymbol{\theta}}_w^*\right) + o_{P*}(1)$$

and

$$T_w^*\left(\widehat{\boldsymbol{\theta}}_a\right) = 2\left\{l_w^*\left(\widehat{\boldsymbol{\theta}}_w^*\right) - l_w^*\left(\widehat{\boldsymbol{\theta}}_1^*\right)\right\} = N\left(\widehat{\boldsymbol{\theta}}_w^* - \widehat{\boldsymbol{\theta}}_1^*\right)^T \boldsymbol{J}_w\left(\widehat{\boldsymbol{\theta}}_a\right)\left(\widehat{\boldsymbol{\theta}}_w^* - \widehat{\boldsymbol{\theta}}_1^*\right) + o_{P*}(1),$$

where the probability is with respect to the parametric distribution indexed by $\widehat{\boldsymbol{\theta}}_a$. Under the additional assumption A5,

$$N^{1/2}\left(\widehat{\boldsymbol{\theta}}_w^* - \widehat{\boldsymbol{\theta}}_1^*\right) \overset{\mathcal{L}_*}{\to} \mathcal{N}\left(\mathbf{0}, -\boldsymbol{J}_1^{-1}\left(\widehat{\boldsymbol{\theta}}_a\right) + \boldsymbol{J}_w^{-1}\left(\widehat{\boldsymbol{\theta}}_a\right)\boldsymbol{K}_w\left(\widehat{\boldsymbol{\theta}}_a\right)\boldsymbol{J}_w^{-1}\left(\widehat{\boldsymbol{\theta}}_a\right)\right) = \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Gamma}\left(\widehat{\boldsymbol{\theta}}_a\right)\right),$$

so that

$$T_1^*\left(\widehat{\boldsymbol{\theta}}_a\right) \overset{\mathcal{L}_*}{\to} \sum_{j=1}^{p} \lambda_{1j}\left(\widehat{\boldsymbol{\theta}}_a\right) Z_j^2 \quad (20)$$

where $\lambda_{1j}\left(\widehat{\boldsymbol{\theta}}_a\right)$ are the eigenvalues of $\boldsymbol{\Gamma}^{T/2}\left(\widehat{\boldsymbol{\theta}}_a\right)\boldsymbol{J}_1\left(\widehat{\boldsymbol{\theta}}_a\right)\boldsymbol{\Gamma}^{1/2}\left(\widehat{\boldsymbol{\theta}}_a\right)$,

$$T_w^*\left(\widehat{\boldsymbol{\theta}}_a\right) \overset{\mathcal{L}_*}{\to} \sum_{j=1}^{p} \lambda_{wj}\left(\widehat{\boldsymbol{\theta}}_a\right) Z_j^2 \quad (21)$$

where $\lambda_{wj}\left(\widehat{\boldsymbol{\theta}}_a\right)$ are the eigenvalues of $\boldsymbol{\Gamma}^{T/2}\left(\widehat{\boldsymbol{\theta}}_a\right)\boldsymbol{J}_w\left(\widehat{\boldsymbol{\theta}}_a\right)\boldsymbol{\Gamma}^{1/2}\left(\widehat{\boldsymbol{\theta}}_a\right)$, and $\{Z_j\}_{j=1}^p$ are independent and identically distributed $\mathcal{N}(0,1)$.

Theorem 6. *Assume the conditions of Theorem 5. For $z > 0$ and $\{Z_j\}_{j=1}^p$ independently*

and identically distributed $\mathcal{N}(0,1)$*, define*

$$G_{bN}(z;\boldsymbol{\theta}) = Pr[T_b(\boldsymbol{\theta}) \leq z;\boldsymbol{\theta}] \text{ and } L_b(z;\boldsymbol{\theta}) = Pr\left[\sum_{j=1}^{p} \lambda_{bj}(\boldsymbol{\theta}) Z_j^2 \leq z;\boldsymbol{\theta}\right]$$

for $b = 1$ *or* $b = w$*. Further, assume there exists* $\delta > 0$ *such that*

$$\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\delta} |G_{bN}(z;\boldsymbol{\theta}) - L_b(z;\boldsymbol{\theta})| \to 0 \tag{22}$$

as $N \to \infty$*. Then the bootstrap test statistics are consistent in the sense that for all* $z > 0$ *and for all* $\epsilon > 0$*,*

$$Pr\left[\left|G_{bN}\left(z;\widehat{\boldsymbol{\theta}}_a\right) - L_b(z;\boldsymbol{\theta}_0)\right| > \epsilon;\boldsymbol{\theta}_0\right] \to 0 \tag{23}$$

as $N \to \infty$*, for* $a = w$ *or* $a = 1$ *and for* $b = w$ *or* $b = 1$*.*

Theorem 6 implies that for either choice of a and either choice of b, the empirical distribution function of independent copies of $T_b^*\left(\widehat{\boldsymbol{\theta}}_a\right)$ can be used to approximate the distribution function of T_b.

2.3 Theoretical Results Under the Alternative Hypothesis

Under the alternative hypothesis of informative selection, we proceed by assuming that the distribution of the sample data is in the same parametric family as the distribution holding for the superpopulation. For parametric families and sampling designs for which this is true, see Pfeffermann, Krieger, and Rinott (1998). This sort of conjugacy arises when sampling from exponential families via Poisson sampling, and since this is quite broad we will consider this situation here. The distribution under the superpopulation is

$$f(y_k \mid \boldsymbol{x}_k;\boldsymbol{\theta}_0),$$

and under the alternative hypothesis of informative selection this is not, in general, equal to the distribution of the data given that they are included in the sample, which is

$$f(y_k \mid \boldsymbol{x}_k, I_k = 1; \boldsymbol{\theta}_s).$$

Here we are using $\boldsymbol{\theta}_0$ to denote the parameter values holding at the population level, and $\boldsymbol{\theta}_s$ to denote the parameter values holding at the sample level. We will suppose that assumptions analogous to A1–A5 hold under the sample distribution as well, that is, at the parameter value $\boldsymbol{\theta}_s$; this will ensure that we have well behaved information matrices under the sample likelihood and allow us to use classical maximum likelihood results. Suppose we are sampling from an exponential family of the form

$$f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}_0) = a_i(\boldsymbol{\theta}_0) \exp\left[\sum_{k=1}^{K} \theta_{0k} b_k(y_i) + c_i(y_i)\right],$$

where $\boldsymbol{\theta}$ defines the $K-$dimensional natural parameterization of the family, and $b_k(\cdot)$ and $c_i(\cdot)$ are known functions. Further suppose that the inclusion probabilities have expectations

$$\mathrm{E}\left[\pi_i \mid y_i, \boldsymbol{x}_i\right] = r_i \exp\left[\sum_{k=1}^{K} \Delta_k b_k(y_i)\right],$$

where r_i and $\{\Delta_k\}$ are constants which may depend on $\boldsymbol{x}_i$ but not y_i. Then we have that $\theta_{sk} = \theta_{0k} + \Delta_k$ (Pfeffermann, Krieger, and Rinott (1998)). The constants Δ_k function as an offset for the parameters in the sample distribution versus the parameters in the population distribution, and in Theorem 7 we will define a sequence of local alternatives by allowing that offset to go to zero.

Theorem 7. *Let $l_a(\cdot)$ and $\widehat{\boldsymbol{\theta}}_a$ be as defined in Theorem 1. Define a sequence of local alternatives*

$$H_{1N} : \boldsymbol{\theta}_s = \boldsymbol{\theta}_0 + \frac{\boldsymbol{d}}{\sqrt{N}}.$$

Under this sequence of alternatives and the assumptions holding under the alternative hypothesis,

$$T_1 = 2\left\{l_1\left(\widehat{\boldsymbol{\theta}}_1\right) - l_1\left(\widehat{\boldsymbol{\theta}}_w\right)\right\} = N\left(\widehat{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_w\right)^T \boldsymbol{J}_1\left(\widehat{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_w\right) + o_P(1)$$

and

$$T_w = 2\left\{l_w\left(\widehat{\boldsymbol{\theta}}_w\right) - l_w\left(\widehat{\boldsymbol{\theta}}_1\right)\right\} = N\left(\widehat{\boldsymbol{\theta}}_w - \widehat{\boldsymbol{\theta}}_1\right)^T \boldsymbol{J}_w\left(\widehat{\boldsymbol{\theta}}_w - \widehat{\boldsymbol{\theta}}_1\right) + o_P(1),$$

where the probability is with respect to the parametric distribution indexed by $\boldsymbol{\theta}_0$*. Furthermore,*

$$N^{1/2}\left[\left(\widehat{\boldsymbol{\theta}}_w - \widehat{\boldsymbol{\theta}}_1\right) - (\boldsymbol{\theta}_0 - \boldsymbol{\theta}_s)\right] \xrightarrow{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, -\boldsymbol{J}_1^{-1} + \boldsymbol{J}_w^{-1}\boldsymbol{K}_w\boldsymbol{J}_w^{-1}\right) = \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Gamma}\right), \tag{24}$$

so that

$$T_a \xrightarrow{\mathcal{L}} \sum_{j=1}^{p} \lambda_{aj}\chi^2(1;\delta_j), \tag{25}$$

where λ_{aj} *are the eigenvalues of* $\boldsymbol{\Gamma}^{T/2}\boldsymbol{J}_a\boldsymbol{\Gamma}^{1/2}$*, and* $\chi^2(1;\delta_j)$ *is a non-central chi-squared random variable with* 1 *degree of freedom and non-centrality parameter* δ_j *defined as*

$$\delta_j = \frac{[\boldsymbol{P}\boldsymbol{\Gamma}^{-1/2}\boldsymbol{d}]_j^2}{2},$$

that is, the j^{th} *element of the resulting vector in brackets, where* $\boldsymbol{P}$ *is a matrix of eigenvectors of* $\boldsymbol{\Gamma}^{T/2}\boldsymbol{J}_a\boldsymbol{\Gamma}^{1/2}$*. With the non-centrality parameter defined in this way, the non-central chi-squared distribution is of the form*

$$f(q) = \sum_{i=0}^{\infty}\left(e^{-\delta}\frac{\delta^i}{i!}\right)Q(q: 1+2i),$$

where $Q(q:k)$ *is the cumulative distribution function of* χ^2 *distribution with* k *degrees of freedom.*

This theorem can be used to perform local power calculations under certain conditions. The assumptions made for the theoretical results shown in this section were made because they allow for some relatively general asymptotic results. In particular, allowing the data to come from an exponential family covers many potential cases. The major complications in trying to compare sample distributions to population distributions under the alternative hypothesis of informative selection come from the fact that under an arbitrary sampling design, the sample likelihood is not guaranteed to be in the same family as the population, and so direct comparisons of parameter values are tricky, in fact it is not immediately clear how to even proceed. Pfeffermann, Krieger, and Rinott (1998) provides some specific sampling designs and parametric family pairs under which the sort of conjugacy needed holds (that is to say, the sample and population distributions are in the same parametric family), and so this has been the starting point. In practice it may be easier to perform power calculations via simulation.

2.4 Theoretical Extensions

Using the framework presented above, we wish to establish results analogous to those of DuMouchel and Duncan (1983), in which effects of informative selection are tested by comparing a linear mean model to the linear model extended to include weighted covariates,. We extend those results to a more general framework that can be applied to a wide range of models. We begin by partitioning the parameters into mean parameters, $\boldsymbol{\beta}$, which are the coefficients of model covariates $\boldsymbol{x}_k$, and nuisance parameters, $\boldsymbol{\xi}$. Then we will extend the mean structure to incorporate weighted covariates $w_k\boldsymbol{x}_k$, which have coefficients $\boldsymbol{\gamma}$, and finally link the limiting distribution of $N^{1/2}(\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_1)$ with that of $N^{1/2}(\hat{\boldsymbol{\gamma}}_1 - 0)$. Under this partitioning of parameters, the following assumptions are needed:

A7. The following relations between the original and extended models hold:

$$\left.\frac{\partial \ln f(y_k|\boldsymbol{x}_k, w_k\boldsymbol{x}_k; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi})}{\partial \boldsymbol{\gamma}}\right|_{\boldsymbol{\gamma}=0} = w_k \frac{\partial \ln f(y_k|\boldsymbol{x}_k; \boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}},$$

$$\left.\frac{\partial^2 \ln f(y_k|\boldsymbol{x_k}, w_k\boldsymbol{x}_k; \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\xi}}\right|_{\boldsymbol{\gamma}=0} = w_k \frac{\partial^2 \ln f(y_k|\boldsymbol{x_k}; \boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\xi}},$$

$$\left.\frac{\partial^2 \ln f(y_k|\boldsymbol{x}_k, w_k\boldsymbol{x}_k; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}}\right|_{\boldsymbol{\gamma}=0} = w_k \frac{\partial^2 \ln f(y_k|\boldsymbol{x}_k; \boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T},$$

$$\left.\frac{\partial^2 \ln f(y_k|\boldsymbol{x}_k, w_k\boldsymbol{x}_k; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T}\right|_{\boldsymbol{\gamma}=0} = w_k^2 \frac{\partial^2 \ln f(y_k|\boldsymbol{x}_k; \boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}.$$

It will be shown later that A7 holds in many models of potential interest.

Theorem 8. *Let $\boldsymbol{\theta}_0$ be partitioned into linear parameters, $\boldsymbol{\beta}$, and other parameters, $\boldsymbol{\xi}$. The parameters $\boldsymbol{\beta}$ are linear in the sense that they enter the model as coefficients for known covariates $\boldsymbol{x}_k$ in the probability density function $f(y_k|\boldsymbol{x}_k; \boldsymbol{\beta}, \boldsymbol{\xi})$. Let $f(y_k|\boldsymbol{x}_k, w_k\boldsymbol{x}_k; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi})$ be the probability density function under the model extended to include weighted covariates, $w_k\boldsymbol{x}_k$, with coefficients $\boldsymbol{\gamma}$. Under the additional assumption A7,*

$$N^{1/2}\boldsymbol{M}^{1/2}\boldsymbol{\Gamma}_{11}^{-1/2}(\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_1) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{M}), \tag{26}$$

and

$$N^{1/2}(\hat{\boldsymbol{\gamma}}_1 - 0) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{M}), \tag{27}$$

where

$$\boldsymbol{M} = \{a_{w^2} - a_w V_{11}^{-1} a_w - 2b_w d_1^{-1} b_1^T V_{11}^{-1} a_w + b_w V_{21}^{-1} b_w^T\}^{-1},$$

$$\boldsymbol{\Gamma}_{11} = V_{1w}^{-1}[a_{w^2} - 2b_w d_w^{-1} b_{w^2}^T + b_w d_w^{-1} d_{w^2} d_w^{-1} b_w^T] V_{1w}^{-1} - V_{11}^{-1},$$

where $V_{1r} = a_r - b_r d_r^{-1} b_r^T$, $V_{2r} = d_r - b_r^T a_r^{-1} b_r$, $a_r = \lim_{n,N\to\infty} -\frac{1}{N}\sum_{k\in U} I_k r_k \frac{\partial^2 \ln l(\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$,
$b_r = \lim_{n,N\to\infty} -\frac{1}{N}\sum_{k\in U} I_k r_k \frac{\partial^2 \ln l(\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\xi}}$, *and*
$d_r = \lim_{n,N\to\infty} -\frac{1}{N}\sum_{k\in U} I_k r_k \frac{\partial^2 \ln l(\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^T}$; *for* $r = 1$ *and* $r = w$.

In the case in which the maximum likelihood estimates for the mean parameters, $\boldsymbol{\beta}$, and the nuisance parameters, $\boldsymbol{\xi}$ are asymptotically uncorrelated, the following corollary arises.

Corollary 9. *Let* $\boldsymbol{\theta}_0$ *be partitioned as in Theorem 8 and let* $f(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})$ *and* $f(y_k|\boldsymbol{x}_k, w_k\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\xi})$ *be defined as in Theorem 8. If the maximum likelihood estimates for the mean parameters,* $\boldsymbol{\beta}$*, and the nuisance parameters,* $\boldsymbol{\xi}$ *are asymptotically uncorrelated, then*

$$N^{1/2}\{a_{w^2} - a_w a_1^{-1} a_w\}^{-1} a_w(\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_1) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \{a_{w^2} - a_w a_1^{-1} a_w\}^{-1}), \tag{28}$$

and

$$N^{1/2}(\hat{\boldsymbol{\gamma}}_1 - 0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \{a_{w^2} - a_w a_1^{-1} a_w\}^{-1}), \tag{29}$$

where a_{w^2}, a_w, *and* a_1 *are defined as in Theorem 8.*

Before continuing on to simulation studies and applications of the theory discussed, I would like to take some time to look in detail at some parametric families that satisfy the assumptions needed to apply Theorem 8 and Corollary 9.

2.4.1 Gamma Mixture Model

To illustrate a situation in which Theorem 8 is applicable, consider the mixture model

$$y_k = \{z_k \times 0\} + \{(1 - z_k) \times v_k\},$$

where $z_k \sim$ Bernoulli(δ) and $v_k \sim$ Gamma($e^{\boldsymbol{x_k\beta}}$, τ), with mean $\exp(\boldsymbol{x}_k\boldsymbol{\beta})\tau$ and variance $\exp(\boldsymbol{x}_k\boldsymbol{\beta})\tau^2$. y_k is zero with probability δ and is positive following a gamma distribution with probability $1 - \delta$. In the original model, the probability density is

$$f(y_k|\boldsymbol{x_k}; \delta, \tau, \boldsymbol{\beta}) = \delta^{z_k} \left\{ (1 - \delta) \frac{y_k^{e^{\boldsymbol{x_k\beta}}-1} e^{-y_k/\tau}}{\tau^{e^{\boldsymbol{x_k\beta}}} \Gamma(e^{\boldsymbol{x_k\beta}})} \right\}^{1-z_k}.$$

This is similar to a model used to model fishery data in section 2.6.1. We are interested in the scores and information with respect to $\boldsymbol{\beta}$ since this is where we will be extending the model. In the following, the derivative of the natural logarithm of the gamma function is called the digamma function and is denoted $\psi(\cdot)$; its derivative is called the trigamma function and is denoted $\psi_1(\cdot)$. The derivatives are

$$\begin{aligned} &\frac{\partial \ln f(y_k|\boldsymbol{x_k}; \delta, \tau, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= \left(e^{\boldsymbol{x_k\beta}} \boldsymbol{x_k} \ln(y_k) - e^{\boldsymbol{x_k\beta}} \boldsymbol{x_k} \ln(\tau) - \left[\psi(e^{\boldsymbol{x_k\beta}}) e^{\boldsymbol{x_k\beta}} \boldsymbol{x_k} \right] \right) (1 - z_k), \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^2 \ln f(y_k|\boldsymbol{x_k}; \delta, \tau, \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= \Big\{ \boldsymbol{x}_k^T e^{\boldsymbol{x}_k\boldsymbol{\beta}} \ln(y_k) \boldsymbol{x}_k - \boldsymbol{x}_k^T e^{\boldsymbol{x}_k\boldsymbol{\beta}} \ln(\tau) \boldsymbol{x}_k \\ &\quad - \left[\psi_1(e^{\boldsymbol{x_k\beta}}) \boldsymbol{x}_k^T e^{2\boldsymbol{x_k\beta}} \boldsymbol{x}_k + \psi(e^{\boldsymbol{x_k\beta}}) \boldsymbol{x}_k e^{\boldsymbol{x}_k^T\boldsymbol{\beta}} \boldsymbol{x}_k \right] \Big\} (1 - z_k). \end{aligned}$$

The double partial derivatives with respect to $\boldsymbol{\beta}$ and δ are zero, as are the double partial derivatives with respect to τ and δ. The double partial derivative with respect to $\boldsymbol{\beta}$ and τ is

$$\frac{\partial^2 \ln f(y_k|\boldsymbol{x_k};\delta,\tau,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\tau} = -\frac{\boldsymbol{x}_k e^{\boldsymbol{x}_k\boldsymbol{\beta}}}{\tau}(1-z_k)$$

Extending the model to include weighted covariates,

$$f(y_k|\boldsymbol{x_k},w_k\boldsymbol{x_k};\delta,\tau,\boldsymbol{\beta},\boldsymbol{\gamma}) = \delta^{z_k}\left\{(1-\delta)\frac{y_k^{\exp(\boldsymbol{x}_k\boldsymbol{\beta}+w_k\boldsymbol{x}_k\gamma)-1}e^{-y_k/\tau}}{\tau^{\exp(\boldsymbol{x}_k\boldsymbol{\beta}+w_k\boldsymbol{x}_k\gamma)}\Gamma(e^{\boldsymbol{x}_k\boldsymbol{\beta}+w_k\boldsymbol{x}_k\gamma})}\right\}^{1-z_k}.$$

The relationships in A7 hold since

$$\begin{aligned}
&\frac{\partial \ln f(y_k|\boldsymbol{x_k},w_k\boldsymbol{x_k};\delta,\tau,\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\bigg|_{\boldsymbol{\gamma}=0} \\
&= \Big(e^{\boldsymbol{x}_k\boldsymbol{\beta}+w_k\boldsymbol{x}_k\gamma}w_k\boldsymbol{x_k}\ln(y_k) - e^{\boldsymbol{x}_k\boldsymbol{\beta}+w_k\boldsymbol{x}_k\gamma}w_k\boldsymbol{x_k}\ln(\tau) \\
&\quad - \psi(e^{\boldsymbol{x}_k\boldsymbol{\beta}+w_k\boldsymbol{x}_k\gamma})e^{\boldsymbol{x}_k\boldsymbol{\beta}+w_k\boldsymbol{x}_k\gamma}w_k\boldsymbol{x_k}\Big)(1-z_k)\Big|_{\boldsymbol{\gamma}=0} \\
&= w_k\frac{\partial \ln f(y_k|\boldsymbol{x_k};\delta,\tau,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial^2 \ln f(y_k|\boldsymbol{x_k},w_k\boldsymbol{x}_k;\delta,\tau,\boldsymbol{\beta}),\boldsymbol{\gamma}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}}\bigg|_{\boldsymbol{\gamma}=0} \\
&= \big\{w_k\boldsymbol{x}_k^T e^{\boldsymbol{x}_k\beta+w_k\boldsymbol{x}_k\gamma}\ln(y_k)\boldsymbol{x}_k - w_k\boldsymbol{x}_k^T e^{\boldsymbol{x}_k\beta+w_k\boldsymbol{x}_k\gamma}\ln(\tau)\boldsymbol{x}_k \\
&- \Big[\psi_1(e^{\boldsymbol{x}_k\boldsymbol{\beta}+w_k\boldsymbol{x}_k\gamma})\boldsymbol{x}_k^T e^{2(\boldsymbol{x}_k\boldsymbol{\beta}+w_k\boldsymbol{x}_k\gamma)}w_k\boldsymbol{x}_k \\
&\quad + \psi(e^{\boldsymbol{x}_k\boldsymbol{\beta}+w_k\boldsymbol{x}_k\gamma})\boldsymbol{x}_k^T e^{\boldsymbol{x}_k\boldsymbol{\beta}+w_k\boldsymbol{x}_k\gamma}w_k\boldsymbol{x}_k\Big]\Big\}(1-z_k)\Big|_{\boldsymbol{\gamma}=0} \\
&= w_k\frac{\partial^2 \ln f(y_k|\boldsymbol{x_k};\delta,\tau,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T},
\end{aligned}$$

$$
\begin{aligned}
&\frac{\partial^2 \ln f(y_k|\boldsymbol{x_k}, w_k\boldsymbol{x}_k; \delta, \tau, \boldsymbol{\beta}), \boldsymbol{\gamma}}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T}\bigg|_{\boldsymbol{\gamma}=0} \\
&= \left\{ w_k^2 \boldsymbol{x}_k^T e^{\boldsymbol{x}_k\beta + w_k\boldsymbol{x}_k\gamma} \ln(y_k)\boldsymbol{x}_k - w_k^2 \boldsymbol{x}_k^T e^{\boldsymbol{x}_k\beta + w_k\boldsymbol{x}_k\gamma} \ln(\tau)\boldsymbol{x}_k \right. \\
&- \left[\psi_1(e^{\boldsymbol{x}_k\boldsymbol{\beta} + w_k\boldsymbol{x}_k\gamma})\boldsymbol{x}_k^T e^{2(\boldsymbol{x}_k\boldsymbol{\beta} + w_k\boldsymbol{x}_k\gamma)} w_k^2 \boldsymbol{x}_k \right. \\
&\quad \left. \left. + \ \psi(e^{\boldsymbol{x}_k\boldsymbol{\beta} + w_k\boldsymbol{x}_k\gamma})\boldsymbol{x}_k^T e^{\boldsymbol{x}_k\boldsymbol{\beta} + w_k\boldsymbol{x}_k\gamma} w_k^2 \boldsymbol{x}_k \right] \right\} (1 - z_k)\Big|_{\boldsymbol{\gamma}=0} \\
&= w_k^2 \frac{\partial^2 \ln f(y_k|\boldsymbol{x_k}; \delta, \tau, \boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T},
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial^2 \ln f(y_k|\boldsymbol{x_k}, w_k\boldsymbol{x}_k; \delta, \tau, \boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\partial\tau}\bigg|_{\boldsymbol{\gamma}=0} &= -\frac{w_k\boldsymbol{x}_k e^{\boldsymbol{x}_k\boldsymbol{\beta} + w_k\boldsymbol{x}_k\boldsymbol{\gamma}}}{\tau}(1 - z_k)\bigg|_{\boldsymbol{\gamma}=0} \\
&= w_k \frac{\partial^2 \ln f(y_k|\boldsymbol{x_k}; \delta, \tau, \boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\tau},
\end{aligned}
$$

so we see that the theory of this section is applicable in this case.

2.4.2 Linear Regression Applications

Corollary (9) is widely applicable and deserves some special attention here. It can be applied in the regular linear model case, where

$$
f(y_k|\boldsymbol{x}_k; \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_k - \boldsymbol{x}_k\boldsymbol{\beta}) \right]
$$

and

$$
f(y_k|\boldsymbol{x}_k, w_k\boldsymbol{x}_k; \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} \left(y_k - \begin{bmatrix} \boldsymbol{x}_k & w_k\boldsymbol{x}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} \right) \right].
$$

The mean and variance estimates are uncorrelated, and the second derivatives needed for the information matrix in the extended model are

$$\frac{\partial^2 \ln f(y_k|\boldsymbol{x}_k, w_k\boldsymbol{x}_k; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = -\frac{\boldsymbol{x}_k^T\boldsymbol{x}_k}{\sigma^2},$$

$$\frac{\partial^2 \ln f(y_k|\boldsymbol{x}_k, w_k\boldsymbol{x}_k; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}} = -\frac{\boldsymbol{x}_k^T w_k\boldsymbol{x}_k}{\sigma^2},$$

and

$$\frac{\partial^2 \ln f(y_k|\boldsymbol{x}_k, w_k\boldsymbol{x}_k; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T} = -\frac{\boldsymbol{x}_k^T w_k^2\boldsymbol{x}_k}{\sigma^2},$$

thus the relationships given in A7 hold. The result is effectively established in this case since this gives us the limiting variance of $N^{1/2}(\hat{\boldsymbol{\gamma}} - 0)$ and also that of $N^{1/2}(\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_1)$, and the transformation that links the two. This is exactly the result from DuMouchel and Duncan (1983) but presented in the framework of this paper; in this framework we can find analogous results for many analyses besides regular linear regression. An immediate extension is for any normal linear mixed model, the only difference being that the nuisance parameter $\boldsymbol{\xi}$ will be of higher dimension than one (in linear regression with only fixed covariates its dimension is one, containing only the model variance, σ^2). For normal linear mixed models the information matrix will still have a block diagonal form (one block containing fixed covariates, the other containing random effects and the random error variance), regardless of the density of the random effects, suggesting the asymptotic independence between the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ that is needed.

2.4.3 Applications to Exponential Families

The assumptions are in fact satisfied for all exponential families. This is a simple result since the form of an s-dimensional exponential family, indexed by the parameter $\boldsymbol{\theta}$, is

$$p_\theta(y) = \exp\left[\sum_{i=1}^{s} \eta_i(\boldsymbol{\theta}) T_i(y) - B(\boldsymbol{\theta})\right] h(y),$$

which, since $\boldsymbol{\theta}$ contains linear parameters $\boldsymbol{\beta}$ and nuisance parameters $\boldsymbol{\xi}$, can be written as

$$p_\theta(y) = \exp\left[\sum_{i=1}^{s} \eta_i(\boldsymbol{x}_k^T\boldsymbol{\beta}, \boldsymbol{\xi}) T_i(y) - B(\boldsymbol{x}_k^T\boldsymbol{\beta}, \boldsymbol{\xi})\right] h(y).$$

The partial derivative of $B(\boldsymbol{\theta})$ with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial}{\partial \boldsymbol{\beta}} B\left(\boldsymbol{x}_k^T\boldsymbol{\beta}, \boldsymbol{\xi}\right) = \boldsymbol{x}_k^T B'\left(\boldsymbol{x}_k^T\boldsymbol{\beta}, \boldsymbol{\xi}\right),$$

and extending the linear structure to contain weighted covariates the derivative (evaluated at $\boldsymbol{\gamma} = 0$) is

$$\frac{\partial}{\partial \boldsymbol{\gamma}} B\left(\boldsymbol{x}_k^T\boldsymbol{\beta} + w_k\boldsymbol{x}_k\boldsymbol{\gamma}, \boldsymbol{\xi}\right)\big|_{\boldsymbol{\gamma}=0} = w_k\boldsymbol{x}_k^T B'\left(\boldsymbol{x}_k^T\boldsymbol{\beta}, \boldsymbol{\xi}\right) = w_k \frac{\partial}{\partial \boldsymbol{\beta}} B\left(\boldsymbol{x}_k^T\boldsymbol{\beta}, \boldsymbol{\xi}\right).$$

This is true whether or not $B(\cdot)$ is a function of the linear parameters; if it is not, then the derivative is 0. The argument is identical for the $\eta_i(\cdot)$, and analogous for the necessary second derivatives.

It is also possible to carry out the test for informativeness in exponential families using the natural parameterization. This is intuitively obvious because of the invariance property of maximum likelihood estimates. Because of invariance the $\hat{\boldsymbol{\eta}}$ that maximize the likelihood function will be $\eta(\hat{\boldsymbol{\theta}})$, thus the maximum values obtained are the same under the $\boldsymbol{\theta}$ or $\boldsymbol{\eta}$

parameterization, and the same is true for the test statistics T_1 and T_w. Since the test statistics are identical, their asymptotic distributions are also the same and can be obtained either by computing the information with respect to $\boldsymbol{\theta}$ and following Theorem 2 exactly, or by computing information with respect to $\boldsymbol{\eta}$ and following Theorem 2 in terms of $\boldsymbol{\eta}$.

We can verify experimentally that the results agree under either parameterization. Recall the simple linear regression example relating infant birth weight to gestational age. The probability density for y_k is

$$\begin{aligned}
&\exp\left[\frac{\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma^2}y_k - \frac{1}{2\sigma^2}y_k^2 - \frac{(\boldsymbol{x}_k^T\boldsymbol{\beta})^2}{2\sigma^2} - \frac{1}{2}\ln(\sigma^2)\right]\frac{1}{\sqrt{2\pi}} \\
&= \quad \exp\left[\boldsymbol{x}_k^T\boldsymbol{\eta}_1 y_k + \eta_2 y_k^2 - \left(-\frac{1}{4}(\boldsymbol{x}_k^T\boldsymbol{\eta}_1)^2\eta_2^{-1} - \frac{1}{2}\ln(-\eta_2) + \frac{1}{2}\ln(2)\right)\right]\frac{1}{\sqrt{2\pi}}.
\end{aligned}$$

The information with respect to $\boldsymbol{\eta}$ is given by the second and double partial derivatives of $A(\boldsymbol{\eta})$, the part of the exponent not containing y_k. These derivatives are

$$\frac{\partial^2}{\partial\boldsymbol{\eta}_1\partial\boldsymbol{\eta}_1^T}A(\boldsymbol{\eta}) = -\frac{1}{2}\boldsymbol{x}_k\boldsymbol{x}_k^T\eta_2^{-1},$$

$$\frac{\partial^2}{\partial\eta_2^2}A(\boldsymbol{\eta}) = -\frac{1}{2}(\boldsymbol{x}_k^T\boldsymbol{\eta}_1)^2\eta_2^{-3} + \frac{1}{2}\eta_2^{-2},$$

and

$$\frac{\partial^2}{\partial\boldsymbol{\eta}_1\partial\eta_2}A(\boldsymbol{\eta}) = \frac{1}{2}\boldsymbol{x}_k\boldsymbol{x}_k^T\boldsymbol{\eta}_1\eta_2^{-2}.$$

These derivatives are used to construct the matrices $\hat{\boldsymbol{J}}_1$, $\hat{\boldsymbol{J}}_w$, and $\hat{\boldsymbol{K}}_w$, by summing over the sample, summing with weights, or summing with weights squared respectively. In short, the test is carried out exactly as before. From these three matrices, $\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{J}}_w^{-1}\hat{\boldsymbol{K}}_w\hat{\boldsymbol{J}}_w^{-1} - \hat{\boldsymbol{J}}_1^{-1}$ is computed, and the eigenvalues of $\hat{\boldsymbol{\Gamma}}^{T/2}\hat{\boldsymbol{J}}_1\hat{\boldsymbol{\Gamma}}^{1/2}$ give the weights for the linear combination of χ_1^2 random variables. The process has not changed, the only difference is that the information has been computed in terms of $\boldsymbol{\eta}$; the end result is the same. For the birth weight data, using the natural parameterization will result in the same test statistic, $2(l_1(\hat{\boldsymbol{\eta}}_1) - l_1(\hat{\boldsymbol{\eta}}_w)) = 132.88$,

and the eigenvalues are again $\boldsymbol{\lambda}^T = (1.967, 0.352, 0.021)$. Using the natural parameters, we have obtained identical results to the test under the $\boldsymbol{\theta}$ parameterization.

2.4.4 Logistic Regression Applications

Another case of interest is in logistic regression analyses. Logistic regression fits perfectly into the framework of Corollary 9 because there are linear parameters, and no associated nuisance parameters. Nordberg (1989) uses this idea to test for informativeness by expanding the covariate structure in a logistic regression model to include weighted covariates and then applying the deviance test for nested models. This would be an exact analogue of the Dumouchel and Duncan F-test for informative sampling in linear regression, and Corollary 9 provides the justification for this application. In section 2.5.2 I will present some power calculations and compare the classical likelihood ratio test (deviance test) to the new test and also make comparisons to the Wald-type test proposed by Pfeffermann (1993).

For logistic regression the original and extended models are

$$f(y_k|\boldsymbol{x}_k;\boldsymbol{\beta}) = \left(\frac{e^{\boldsymbol{x}_k\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_k\boldsymbol{\beta}}}\right)^{y_k}\left(1-\frac{e^{\boldsymbol{x}_k\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_k\boldsymbol{\beta}}}\right)^{(1-y_k)}$$

and

$$f(y_k|\boldsymbol{x}_k, w_k\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\gamma}) = \left(\frac{e^{\boldsymbol{x}_k\boldsymbol{\beta}+w_k\boldsymbol{x}_k\boldsymbol{\gamma}}}{1+e^{\boldsymbol{x}_k\boldsymbol{\beta}+w_k\boldsymbol{x}_k\boldsymbol{\gamma}}}\right)^{y_k}\left(1-\frac{e^{\boldsymbol{x}_k\boldsymbol{\beta}+w_k\boldsymbol{x}_k\boldsymbol{\gamma}}}{1+e^{\boldsymbol{x}_k\boldsymbol{\beta}+w_k\boldsymbol{x}_k\boldsymbol{\gamma}}}\right)^{(1-y_k)}.$$

The information in the extended model is

$$\mathcal{I}(\boldsymbol{X},\boldsymbol{W}\boldsymbol{X};\boldsymbol{\beta},\boldsymbol{\gamma})|_{\boldsymbol{\gamma}=0} = \frac{1}{N}\begin{bmatrix} \sum_{k\in U}\boldsymbol{x}_k^T\boldsymbol{x}_k p_k(1-p_k) & \sum_{k\in U}\boldsymbol{x}_k^T w_k\boldsymbol{x}_k p_k(1-p_k) \\ \sum_{k\in U}\boldsymbol{x}_k^T w_k\boldsymbol{x}_k p_k(1-p_k) & \sum_{k\in U}\boldsymbol{x}_k^T w_k^2\boldsymbol{x}_k p_k(1-p_k) \end{bmatrix},$$

where

$$p_k = \frac{\exp(\boldsymbol{x_k}\boldsymbol{\beta})}{1+\exp(\boldsymbol{x_k}\boldsymbol{\beta})}.$$

This is analogous to the previous examples; the relationships in A7 hold, and the relevant limiting distributions are known.

Furthermore, the test in Theorem 2 can be easily adapted to test for informativeness in a subspace of $\boldsymbol{\theta}_0$ whenever $\boldsymbol{\theta}_0$ can be partitioned into uncorrelated pieces. To test for informativeness in the mean structure, the test statistics T_1 and T_w would only look at the differences in log-likelihood due to estimation of $\boldsymbol{\beta}$, and because of the uncorrelated maximum likelihood estimates this can be done by evaluating at any point in the $\boldsymbol{\xi}$ space. For example, one could use the statistic $T_1 = 2\{l_1(\hat{\boldsymbol{\xi}}_1, \hat{\boldsymbol{\beta}}_1) - l_1(\hat{\boldsymbol{\xi}}_1, \hat{\boldsymbol{\beta}}_w)\}$. The necessary eigenvalues could be computed from the relevant portion of $\boldsymbol{\Gamma}^{T/1} J_a \boldsymbol{\Gamma}^{1/2}$.

2.4.5 Relating the Likelihood Ratio Test to the F-test

There is one final idea that I would like to discuss before moving on to the simulations studies and applications. For linear regression, the estimate of σ^2 is independent of the estimates for $\boldsymbol{\beta}$. As mentioned before, this suggests that Corollary 9 can be applied. In this setting, Corollary 9 gives the asymptotic distributions for $(\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_1)$ and $(\hat{\boldsymbol{\gamma}}_1 - \mathbf{0})$ as

$$N^{1/2}\{a_{w^2} - a_w a_1^{-1} a_w\}^{-1} a_w (\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_1) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \{a_{w^2} - a_w a_1^{-1} a_w\}^{-1}),$$

and

$$N^{1/2}(\hat{\boldsymbol{\gamma}}_1 - 0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \{a_{w^2} - a_w a_1^{-1} a_w\}^{-1}),$$

where $\{a_{w^2} - a_w a_1^{-1} a_w\}^{-1} a_w = \left\{ \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{W} \boldsymbol{X} - \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \right\}^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}$.

The justification for using the F-test to test for design informativeness is based on the fact that $\hat{\boldsymbol{\gamma}}_1 = \left\{ \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{W} \boldsymbol{X} - \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \right\}^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} (\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_1)$. Thus the one-to-one transformation that justifies the use of the F-test for design informativeness is precisely the one-to-one transformation that links the limiting distributions of $(\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_1)$ and $(\hat{\boldsymbol{\gamma}}_1 - \mathbf{0})$ in Corollary 9. The advantages provided by the new theory are obvious. Not only can the F-test only detect informativeness in the mean structure, it also does not provide a

framework that can justify the use of tests for submodels in testing for design informativeness in other situations. For example, the deviance test for submodels in the logistic regression setting is analogous to the F-test in the linear model setting, and Corollary 9 provides the justification for the use of the deviance test in testing for design informativeness.

2.5 Simulation Studies

This section contains simulation studies and will be followed by empirical applications to survey data from recent years in the section that follows. First, the likelihood ratio test for informativeness will be compared to the DuMouchel and Duncan test when the informativeness enters the model through the variance structure, and a logistic regression study shows that the likelihood ratio test has comparable power to the Wald test and is more robust and well behaved. The likelihood ratio test will then be applied in the regular linear regression case; this will also provide a connection between the likelihood ratio test and the F-test for submodels proposed by DuMouchel and Duncan.

2.5.1 Student's t Simulation

This section contains a simulation study in which informative selection is present in the variance structure of a linear model. Since the DuMouchel and Duncan test looks for informativeness in the mean structure of a linear model it does not perform well in this situation, in fact it is not able to detect design informativeness almost at all. For $k = 1, \ldots, N$ the true model is

$$y_k = \mu + \sigma \frac{z_k}{\sqrt{v_k/\nu}} \sqrt{\frac{\nu - 2}{\nu}} = \mu + \sigma_k z_k,$$

$\{z_k\}$ iid N(0,1), independent of $\{v_k\}$ iid χ^2_ν. The error terms here are distributed as scaled t_ν, with mean zero and variance σ^2 for $\nu > 2$. Given v_k, $y_k \sim \mathrm{N}(0, \sigma_k^2)$.

Selection is via Poisson sampling with inclusion probabilities:

$$\pi_k = \frac{n\sigma_k}{\sum_{k\in U}\sigma_k}.$$

The selection is informative and selected elements have large conditional variances given σ_k, compared to non-selected elements. This is an interesting scenario because designs with $\pi_k \propto \sigma_k$ minimize the unconditional variance (with respect to design and model randomness) of the Horvitz-Thompson estimator. As $\nu \to \infty$, σ_k converges in probability to σ, and the design becomes noninformative. But the w_k's then all converge to n/N, and the critical value of the test converges to zero, so that the test is not defined.

In this setting, the Dumouchel and Duncan test is the test of significance of the slope coefficient in simple linear regression of y_k on the intercept and $w_k = \pi_k^{-1}$. This test is not defined for $w_k \equiv$ constant, since the design matrix is singular.

The following table shows empirical rejection frequencies based on 1000 replicate samples with $\mu = 2$ and $\sigma = 1$:

Table 1: Empirical Rejection Frequencies

ν	Test	$n = 50$	$n = 100$	$n = 200$
5	T_1	0.822	0.989	1.000
	DD	0.121	0.119	0.125
20	T_1	0.303	0.527	0.835
	DD	0.073	0.062	0.059
80	T_1	0.093	0.158	0.263
	DD	0.043	0.053	0.056

Not surprisingly, the likelihood ratio test based on T_1 has better power in all cases than the Dumouchel and Duncan test, DD. DD has almost no power to detect the informativeness in the variance, since it looks in the mean. At $\nu = 5$ degrees of freedom, the DD test has a small amount of power. This seems to be due to the fact that the variance of the unweighted estimator is much larger (approximately 27% larger) than that of the weighted estimator at all sample sizes considered. The weighted and unweighted estimates will therefore differ by chance, and the DD test will (correctly) reject by (incorrectly) interpreting this difference as bias in the mean estimate. Note that this "lucky" power does not increase with increasing sample size.

Once the degrees of freedom increase, the weights stabilize substantially and the weighted and unweighted estimators have similar variances; e.g., unweighted has only about 6% higher variance than weighted at $\nu = 20$, and 2% higher variance at $\nu = 80$. For these cases, the DD test has essentially no power, rejecting about as often as would be expected under non-informativeness. Again, the rejection frequency does not increase with increasing sample size. The likelihood ratio test, on the other hand, continues to have power to detect the informative selection, and this power increases with sample size.

2.5.2 Logistic Regression Simulation

The following simulation study is based on a data set from Nordberg (1989) involving a population of 12195 milk producing farms in Sweden. The goal of the analysis is to fit a logistic regression model to predict $P(Y = 1)$ where the response variable Y is a binary varible indicating whether or not the farms that had milk cows in 1983 still had milk cows in 1984. Farms that did not have milk cows in 1984 were given a value $y = 0$ and a value of $y = 1$ otherwise. The predictor variables used were

- Region ($R1$, $R2$, $R3$; coded as $0 - 1$ indicator variables)
- Farm Size (Large S=0, and small S=1)
- Farm Type (Primarily milk producing T=1, and Other T=0)
- Age of farmer in three categories ($A_1 = 1$ if Age ≤ 49 and 0 otherwise; $A_2 = 1$ if $50 \leq$ Age ≤ 59, and 0 otherwise; and $A_3 = 1$ if Age $geq60$, and 0 otherwise.)

The model used to generate the data is

$$P(Y=1) = \frac{e^{-2.5+\delta 1.6S-0.3A_2+0.8A_3-\delta 0.8A_3\times S+\delta 1.0T-\delta 0.3R_2\times S-\delta 0.5R_3\times S}}{1+e^{-2.5+\delta 1.6S-0.3A_2+0.8A_3-\delta 0.8A_3\times S+\delta 1.0T-\delta 0.3R_2\times S-\delta 0.5R_3\times S}}, \tag{30}$$

where δ is a constant that controls the level of design informativeness.

The population consisting of 12195 elements was created from the model given in equation (30), where the values for the predictors were taken from the original study. These probabilities were then used to generate a population of y values. From this finite population a stratified sample was taken where the strata are defined by the farm size (S) and type (T), and thus there are four strata based on the four size/type combinations. The number of elements drawn from each stratum was 840, 521, 920, and 720, which corresponds to inclusion probabilities of 0.10, 1.00, 0.60, and 0.42 respectively.

Since the variables on which the data are stratified (farm size and farm type) hold predictive information for the response, it is clear that the sampling is *informative* in the sense that $f(y_k|\boldsymbol{x}_k, I_k = 1) \neq f(y_k|\boldsymbol{x}_k)$, and as such their exclusion from the model should lead to some level of model bias. The level of design informativeness can be controlled by δ in that choosing $\delta = 0$ will correspond to the case in which the sampling is non-informative, and values further from 0 in magnitude reflect a higher degree of informativeness. Choosing $\delta = 1$ corresponds to the model found in the Nordberg paper.

For the following power calculations, the Wald test (Pfeffermann (1993)) given in Corollary 3 will be compared to the likelihood ratio test. Each power calculation is based on 1000 simulations and uses an α of 0.05. "Weighted" and "Unweighted" refers to whether the weighted or unweighted maximum likelihood estimates were be used when computing the expected information in the sample.

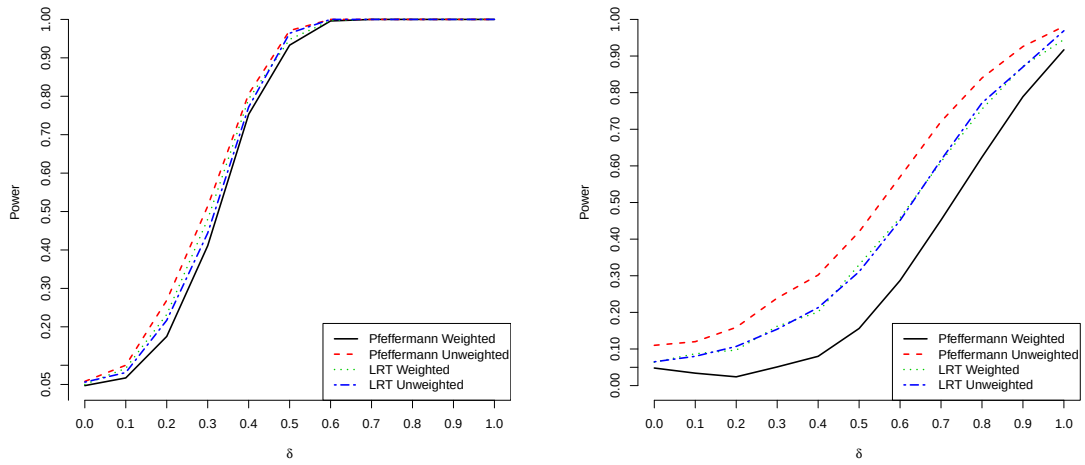

Figure 1: Power curves for Wald test vs. likelihood ratio test using two sample sizes, $n = 3001$ (left), and $n = 601$ (right)

The power curves using the larger sample size show the Wald test proposed by Pfeffermann (1993) performing the best when the information is evaluated at the unweighted maximum likelihood estimator, the worst when the information is evaluated at the weighted maximum likelihood estimates, and the likelihood ratio test falls in between the power curves for the Wald test when using both the weighted and unweighted estimates. However, the power curves using the smaller sample size show some very poor behavior from the Wald test. First, when using the unweighted estimates, the Wald test has the wrong size. The size of the test should be 0.05, and is actually higher than 0.10, so naturally it "wins" for other values of δ as well.

Second, and perhaps even more interestingly, when the weighted estimates are used the Wald test does not even produce a monotone increasing power curve. This is very strange (and undesirable) behavior especially considering that the hypotheses are "equal" vs. "not equal". The likelihood ratio test, in contrast, has very good behavior. The size is correct, or approximately so, for both sample sizes considered, and the test seems to be unaffected by the choice of maximum likelihood estimate used to compute expected information.

From Corollary 9 we can also test for design informativeness by extending the model to include weighted covariates with parameters $\boldsymbol{\gamma}$ and testing that $\boldsymbol{\gamma} = \mathbf{0}$. The classical likelihood ratio test (referred to as the deviance test here for clarity) for submodels in logistic regression compares twice the ratio of log-likelihoods from the reduced and full models to a χ^2_p distribution (p being the difference in the number of parameters for the two models; for our purposes this is the number of parameters in the original model). Below are power curves from the deviance test compared to the the new likelihood ratio test.

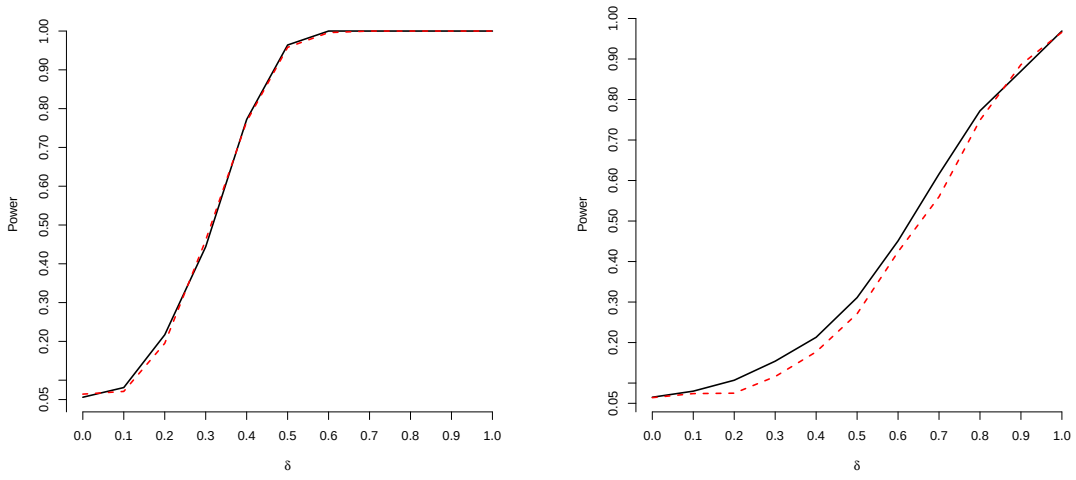

Figure 2: Power curves for deviance test (dotted line) vs. likelihood ratio test (solid line) using two sample sizes, $n = 3001$ (left), and $n = 601$ (right)

The deviance and likelihood ratio tests have almost identical power curves when using the larger sample size ($n = 3001$) and the new likelihood ratio test performs better when the sample size is reduced ($n = 601$).

2.5.3 Linear Model

Recall the National Maternal and Infant Health Survey example in which infant gestational age is regressed on infant birthweight. For linear regression, the likelihood ratio test is performed as follows: Use the expected information to construct $\hat{\boldsymbol{J}}_1$, $\hat{\boldsymbol{J}}_w$, and $\hat{\boldsymbol{K}}_w$; Use these matrices to obtain the necessary covariance matrix, $\hat{\boldsymbol{\Gamma}}$, then compute the eigenvalues of $\hat{\boldsymbol{\Gamma}}^{T/2}\hat{\boldsymbol{J}}_1\hat{\boldsymbol{\Gamma}}^{1/2}$ which are used as the weights for the linear combination of χ^2_1 random variables. The expected information is

$$\text{blockdiag}\left\{\sum_{k\in U} I_k \frac{\boldsymbol{x}_k\boldsymbol{x}_k^T}{\hat{\sigma}^2}, \sum_{k\in U} I_k \frac{1}{2n\hat{\sigma}^4}\right\},$$

where $\boldsymbol{x}_k$ is a column vector containing a 1 and the k^{th} birthweight. $\hat{\boldsymbol{J}}_1$ is computed directly from the sum over the sample of the expected information, $\hat{\boldsymbol{J}}_w$ and $\hat{\boldsymbol{K}}_w$ are computed by summing over the sample with weights and squared weights respectively. The eigenvalues are $\boldsymbol{\lambda}^T = (1.967, 0.352, 0.021)$, and the 0.05 critical value from the asymptotic distribution (computed from 1,000,000 simulated linear combinations of χ^2_1 random variables) is 7.98. The test statistic is

$$2(l_1(\hat{\boldsymbol{\theta}}_1) - l_1(\hat{\boldsymbol{\theta}}_w)) = 132.88,$$

which is highly significant. This agrees with the F-test which produced a p-value much less than 0.001.

Again, it should be pointed out that the F-test will be generally more powerful at detecting informativeness in the mean structure, however it will completely fail to detect informativeness in the variance structure, while the new test can detect informativeness within any parameter.

2.6 Empirical Applications

Here I will present some empirical applications, first to fishery data involving a Gamma mixture model, and then to national health data which involves a censored regression estimation problem. These applications show the wide applicability of the test in practice. Both of these models are quite complex, leading to involved calculations in order to obtain the analytic distributions of the test statistics. Therefore, I will present an application of the bootstrap version of the test because the usefulness of simulation methods for conducting the test is made obvious when the likelihood function for the proposed model is very complex.

2.6.1 Gamma Mixture Model for American Plaice Data

In this section the likelihood ratio test for design informativeness will be applied to fitting a Gamma mixture model for biomass of hauls of American Plaice fish in the southern Gulf of St. Lawrence. For $y =$ Biomass, consider the mixture model

$$y_k = \{z_k \times 0\}\{(1 - z_k) \times x_k\},$$

where $z_k \sim$ Bernoulli(δ), and $x_k \sim$ Gamma(α, τ). Biomass is represented by a random variable that takes a value of 0 with probability δ and is positive following a Gamma distribution with probability $(1 - \delta)$. The probability density function for y_k is

$$f(y_k) = \delta^{z_k} \left\{ (1-\delta) \frac{y_k^{\alpha-1} e^{-y_k/\tau}}{\tau^\alpha \Gamma(\alpha)} \right\}^{1-z_k}$$

The log-likelihood function at the population level is

$$\begin{aligned} l(\delta, \alpha, \tau) = \sum_{k \in U} \ln f(y_k) &= \sum_{k \in U} \ln \left[\delta^{z_k} \left\{ (1-\delta) \frac{y_k^{\alpha-1} e^{-y_k/\tau}}{\tau^\alpha \Gamma(\alpha)} \right\}^{1-z_k} \right] \\ &= \ln \delta \sum_{k \in U} z_k + \{\ln(1-\delta) - \alpha \ln \tau - \ln \Gamma(\alpha)\} \left\{ \sum_{k \in U} (1 - z_k) \right\} \\ &\quad + (\alpha - 1) \sum_{k \in U} (1 - z_k) \ln y_k - \frac{1}{\tau} \sum_{k \in U} y_k (1 - z_k). \end{aligned}$$

Eventually, when obtaining maximum likelihood estimates, or applying tests for informativeness, all sums will be over the sample and may contain weights as well (for obtaining weighted parameter estimates for example).

To obtain unweighted maximum likelihood estimates, $\hat{\boldsymbol{\theta}}_a = (\hat{\delta}_a, \hat{\alpha}_a, \hat{\tau}_a)$, for $a = 1$ or $a = w$, maximize the sample-level log-likelihood function with respect to δ, α, and τ. The sample-level log-likelihood is

$$\begin{aligned} l_a(\delta, \alpha, \tau) = \ln \delta \sum_{k \in s} a_k z_k &+ \{\ln(1-\delta) - \alpha \ln \tau - \ln \Gamma(\alpha)\} \left\{ \sum_{k \in s} a_k (1 - z_k) \right\} \\ &+ (\alpha - 1) \sum_{k \in s} a_k (1 - z_k) \ln y_k - \frac{1}{\tau} \sum_{k \in s} a_k y_k (1 - z_k). \end{aligned}$$

This can be done using software and maximizing the three-dimensional likelihood function above, or by setting the partial derivative of $l_a(\delta, \alpha, \tau)$ with respect to δ equal to zero to obtain

$$\hat{\delta}_a = \left(\sum_{k \in s} a_k \right)^{-1} \sum_{k \in s} a_k z_k,$$

setting the derivative of $l_a(\hat{\delta}_a, \alpha, \tau)$ with respect to τ equal to zero to obtain

$$\hat{\tau}_a(\alpha) = \frac{1}{\alpha} \left(\sum_{k \in s} a_k (1 - z_k) \right)^{-1} \sum_{k \in s} a_k y_k (1 - z_k),$$

and then maximizing the one-dimensional log-likelihood as a function of α only:

$$\begin{aligned} l_a(\hat{\delta}_a, \alpha, \hat{\tau}_a(\alpha)) &= \text{constant} - \ln\{\hat{\tau}_1(\alpha)^\alpha \Gamma(\alpha)\} \sum_{k \in s} a_k (1 - z_k) \\ &\quad + (\alpha - 1) \sum_{k \in s} a_k (1 - z_k) \ln y_k - \sum_{k \in s} a_k (1 - z_k). \end{aligned}$$

The estimate for δ is the (weighted or unweighted) proportion of zero hauls in the sample, and the estimate for τ is the (weighted or unweighted) mean of the non-zero hauls divided by α, which makes sense because the theoretical mean of the non-zero hauls is $\alpha\tau$.

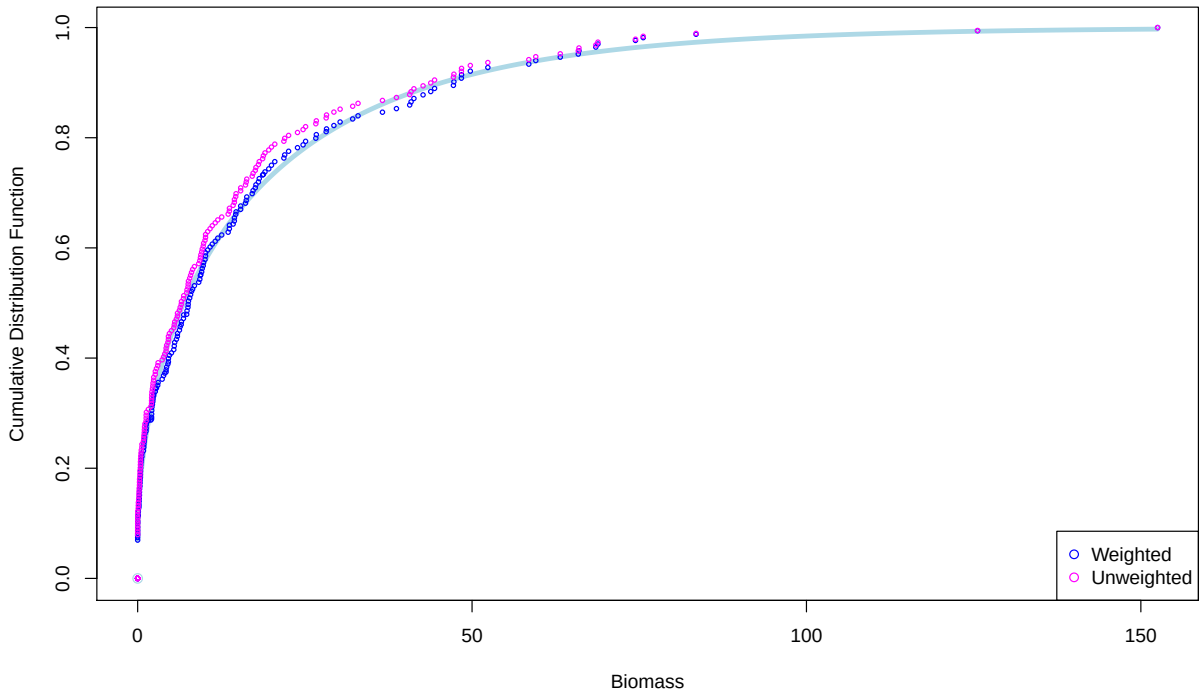

Figure 3: Weighted vs unweighted data with weighted mixture model fit

Figure (3) shows that the weighted data do in fact differ slightly from the unweighted data. This difference is found to be significant. The weighted and unweighted parameter estimates are

$$(\hat{\delta}_w, \hat{\alpha}_w, \hat{\tau}_w) = (0.070, 0.514, 34.342),$$

and

$$(\hat{\delta}_1, \hat{\alpha}_1, \hat{\tau}_1) = (0.0794, 0.497, 32.465).$$

These estimates yield a test statistic of

$$T_1 = 2(l_1(\hat{\delta}_1, \hat{\alpha}_1, \hat{\tau}_1) - l_1(\hat{\delta}_w, \hat{\alpha}_w, \hat{\tau}_w)) = 1.11.$$

The second derivatives are needed for computing the expected information matrix, which is needed to find the asymptotic distribution of the test statistic, T_1. The derivative of $\ln \Gamma(\cdot)$ is called the digamma function and is denoted $\psi(\cdot)$;

its derivative is called the trigamma function and is denoted $\psi_1(\cdot)$ (see section 2.4.1 for more on applications to a gamma mixture model). The non-zero elements of the information matrix are

$$\mathrm{E}\left[-\sum_{k\in U} I_k \frac{\partial^2 l_a(\delta,\alpha,\tau)}{\partial\delta^2}\right] = \frac{\sum_{k\in U} I_k a_k}{\delta(1-\delta)},$$

$$\mathrm{E}\left[-\sum_{k\in U} I_k \frac{\partial^2 l_a(\delta,\alpha,\tau)}{\partial\alpha^2}\right] = \psi_1(\alpha)(1-\delta)\sum_{k\in U} I_k a_k,$$

$$\mathrm{E}\left[-\sum_{k\in U} I_k \frac{\partial^2 l_a(\delta,\alpha,\tau)}{\partial\tau^2}\right] = \frac{1-\delta}{\tau}\sum_{k\in U} I_k a_k,$$

and

$$\mathrm{E}\left[-\sum_{k\in U} I_k \frac{\partial^2 l_a(\delta,\alpha,\tau)}{\partial\alpha\partial\tau^2}\right] = \frac{\alpha(1-\delta)}{\tau^2}\sum_{k\in U} I_k a_k.$$

$$\frac{\partial^2 l(\delta,\alpha,\tau)}{\partial\delta^2} = -\frac{1}{\delta^2}\sum_{k\in U} z_k - \frac{1}{(1-\delta)^2}\sum_{k\in U}(1-z_k),$$

$$\frac{\partial^2 l(\delta,\alpha,\tau)}{\partial\alpha^2} = -\psi_1(\alpha)\sum_{k\in U}(1-z_k),$$

and

$$\frac{\partial^2 l(\delta,\alpha,\tau)}{\partial\tau^2} = \frac{\alpha}{\tau^2}\sum_{k\in U}(1-z_k) - \frac{2}{\tau^3}\sum_{k\in U} y_k(1-z_k).$$

The double partial derivatives with respect to α and δ are zero, as are the double partial derivatives with respect to τ and δ. The double partial derivative with respect to α and τ is

$$\frac{\partial^2 l(\delta,\alpha,\tau)}{\partial\alpha\partial\tau} = -\frac{1}{\tau}\sum_{k\in U}(1-z_k)$$

The information is given by the expectation of the negative of the second and double partial

derivatives above. Taking expectations we have

$$\mathrm{E}\left[\frac{1}{\delta^2}\sum_{k\in U} z_k + \frac{1}{(1-\delta)^2}\sum_{k\in U}(1-z_k)\right] = \frac{\sum_{k\in U} 1}{\delta(1-\delta)},$$

$$\mathrm{E}\left[\psi_1(\alpha)\sum_{k\in U}(1-z_k)\right] = \psi_1(\alpha)(1-\delta)\sum_{k\in U} 1,$$

$$\mathrm{E}\left[\frac{1}{\tau}\sum_{k\in U}(1-z_k)\right] = \frac{1-\delta}{\tau}\sum_{k\in U} 1,$$

and

$$\mathrm{E}\left[\frac{\alpha}{\tau^2}\sum_{k\in U}(1-z_k)\right] + \mathrm{E}\left[\frac{2}{\tau^3}\sum_{k\in U}(1-z_k)y_k\right] = \frac{\alpha(1-\delta)}{\tau^2}\sum_{k\in U} 1.$$

The information is used to construct $\hat{\boldsymbol{J}}_a$ and $\hat{\boldsymbol{K}}_a$; $\hat{\boldsymbol{J}}_1$ is obtained by summing the information over the sample, $\hat{\boldsymbol{J}}_w$ by weighting the sum over the sample by w_k, and $\hat{\boldsymbol{K}}_w$ by weighting the sum over the sample by w_k^2. The test proceeds by computing $\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{J}}_w^{-1}\hat{\boldsymbol{K}}_w\hat{\boldsymbol{J}}_w^{-1} - \hat{\boldsymbol{J}}_1^{-1}$ and then getting the eigenvalues from $\hat{\boldsymbol{\Gamma}}^{T/2}\hat{\boldsymbol{J}}_1\hat{\boldsymbol{\Gamma}}^{1/2}$. These eigenvalues provide the weights for the linear combination of χ_1^2 random variables that constitutes the asymptotic distribution of T_1. The eigenvalues are $\lambda^T = (0.0554, 0.0554, 0.0554)$, and the 0.05 critival value is 0.433. Recalling that the value of the test statistic was $T_1 = 1.11$, we strongly reject the null hypothesis of non-informative selection.

2.6.2 Tobit Regression

The National Health and Nutrition Examination Survey is a yearly survey conducted by the National Center for Health Statistics. For this example the likelihood ratio test for informative sampling is applied to a Tobit model, which is a censored regression model. The response variable is $y = \ln(Cotinine)$. Cotinine is the primary metabolite of nicotine in cigarette smoke and it is of interest how various economic and housing factors affect cotanine levels in children (see Wilson et. al. (2011)). Consider the following two potential models:

$$\begin{aligned}\ln(Cotinine) &= \beta_0 + \beta_1 I_{\{Age<12\}} + \beta_2 Poverty \\ &\quad + \beta_3 I_{\{AttachedHousing\}} + \beta_4 I_{\{Apartment\}} + \epsilon,\end{aligned} \tag{31}$$

and

$$\begin{aligned} \ln(Cotinine) &= \beta_0 + \beta_1 I_{\{Age<12\}} + \beta_2 Poverty \\ &\quad + \beta_3 I_{\{AttachedHousing\}} + \beta_4 I_{\{Apartment\}} + \beta_5 I_{\{Hispanic\}} + \epsilon, \end{aligned} \tag{32}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. If the test for informativeness is applied when fitting model (31), highly significant design informativeness will be found. However, it is known that the survey was designed to oversample Hispanics in the population, and in fact if Hispanic origin is taken into account, as in model (32), significant design informativeness is no longer found.

The lower detection limit for cotinine is 0.015, so instead of observing every response y we observe

$$y* = \max(y, \tau),$$

where $\tau = \ln(0.015)$.

The likelihood function for the proposed model is

$$L(y_k|\boldsymbol{x}_k; \boldsymbol{\beta}, \sigma) = \prod_{k \notin c} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_k - \boldsymbol{x}_k^T\boldsymbol{\beta})^2} \prod_{k \in c} \Phi\left(\frac{\tau - \boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right),$$

where $\{k \in c\}$ indicates censored observations, $\{k \notin c\}$ indicates uncensored observations, and $\Phi(\cdot)$ is the standard normal cumulative distribution function. The log-likelihood is

$$l(y_k|\boldsymbol{x}_k; \boldsymbol{\beta}, \sigma) = \sum_{k \notin c} \left[-\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2\sigma^2}(y_k - \boldsymbol{x}_k^T\boldsymbol{\beta})^2\right] + \sum_{k \in c} \log \Phi\left(\frac{\tau - \boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right).$$

$$\begin{aligned} \frac{\partial l(y_k|\boldsymbol{x}_k; \boldsymbol{\beta}, \sigma)}{\partial \sigma} &= \sum_{k \notin c} \left[\frac{1}{\sigma^3}(y_k - \boldsymbol{x}_k^T\boldsymbol{\beta})^2 - \frac{1}{\sigma}\right] \\ &\quad - \sum_{k \in c} \frac{1}{\Phi\left(\frac{\tau - \boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)} \phi\left(\frac{\tau - \boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\tau - \boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma^2}\right), \end{aligned}$$

and

$$\frac{\partial l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\sigma)}{\partial \boldsymbol{\beta}} = \sum_{k\notin c} \frac{\boldsymbol{x}_k^T}{\sigma^2}(y_k - \boldsymbol{x}_k^T\boldsymbol{\beta}) - \sum_{k\in c} \frac{1}{\Phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)}\phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\boldsymbol{x}_k^T}{\sigma}\right).$$

The second partial derivatives are

$$\begin{aligned}\frac{\partial^2 l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\sigma)}{\partial \boldsymbol{\beta}^T\boldsymbol{\beta}} &= -\sum_{k\notin c}\frac{1}{\sigma^2}\boldsymbol{x}_k^T\boldsymbol{x}_k \\ &\quad -\sum_{k\in c}\left[\frac{1}{\Phi^2\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)}\phi^2\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\frac{\boldsymbol{x}_k^T\boldsymbol{x}_k}{\sigma^2}\right. \\ &\quad \left. +\frac{1}{\Phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)}\phi'_{\boldsymbol{\beta}}\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\boldsymbol{x}_k^T}{\sigma}\right)\right],\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\sigma)}{\partial \sigma^2} &= \sum_{k\notin c}\left[\frac{1}{\sigma^2}-\frac{3}{\sigma^4}(y_k-\boldsymbol{x}_k^T\boldsymbol{\beta})^2\right] \\ &\quad -\sum_{k\in c}\left[\frac{1}{\Phi^2\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)}\phi^2\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma^2}\right)^2\right. \\ &\quad +\frac{1}{\Phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)}\left(\phi'_{\sigma}\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma^2}\right)\right. \\ &\quad \left.\left. -\ 2\phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma^3}\right)\right)\right],\end{aligned}$$

and

$$
\begin{aligned}
\frac{\partial^2 l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\sigma)}{\partial\sigma\partial\boldsymbol{\beta}} &= \sum_{k\notin c} -\frac{2\boldsymbol{x}_k^T}{\sigma^3}(y_k-\boldsymbol{x}_k^T\boldsymbol{\beta}) \\
&\quad -\sum_{k\in c}\left\{\frac{1}{\Phi^2\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)}\phi^2\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma^2}\right)\left(\frac{\boldsymbol{x}_k^T}{\sigma}\right)\right. \\
&\quad -\frac{1}{\Phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)}\left[\phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\boldsymbol{x}_k^T}{\sigma^2}\right)\right. \\
&\quad \left.\left. -\ \phi'_{\boldsymbol{\beta}}\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma^2}\right)\right]\right\},
\end{aligned}
$$

where

$$
\phi'_{\boldsymbol{\beta}}\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)=\phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\boldsymbol{x}_k^T}{\sigma^2}(\tau-\boldsymbol{x}_k^T\boldsymbol{\beta})\right),
$$

and

$$
\phi'_{\sigma}\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)=\phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{1}{\sigma^3}(\tau-\boldsymbol{x}_k^T\boldsymbol{\beta})^2\right).
$$

Now we need to take expectations conditioning on $I_{\{k\in s\}}$ and $\boldsymbol{x}_k^T$. For all sums over $k\in c$, since the y_k only enter through the censoring indicator,$I_{y_k\leq\tau}$, this will amount to adding the term $\mathrm{E}\left[I_{\{Y_k\leq\tau\}}\right]=P(Y_k\leq\tau)=\Phi(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma})$. Some additional work is necessary however for the three sums over $\{k\notin c\}$. For the following assume we are conditioning on $I_{\{k\in s\}}$ and $\boldsymbol{x}_k$. Starting with the expectation for the uncensored observations in the second derivative with respect to $\boldsymbol{\beta}$ we have

$$
\sum_{k\in s}\frac{\boldsymbol{x}_k^T\boldsymbol{x}_k}{\sigma^2}\mathrm{E}\left[I_{\{Y_k>\tau\}}\right]=\sum_{k\in s}\frac{\boldsymbol{x}_k^T\boldsymbol{x}_k}{\sigma^2}\left(1-\Phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\right),
$$

and the expectation for the uncensored observations in the second derivative with respect to

σ is

$$\begin{aligned}
&\sum_{k\in s}\left[\frac{1}{\sigma^2}\mathrm{E}\left[I_{\{y_k>\tau\}}\right]-\frac{3}{\sigma^4}\mathrm{E}\left[I_{\{Y_k>\tau\}}(y_k-\boldsymbol{x}_k^T\boldsymbol{\beta})^2\right]\right]\\
&\quad=\sum_{k\in s}\left[-\frac{2}{\sigma^2}\left(1-\Phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\right)-\frac{3}{\sigma^3}(\tau-\boldsymbol{x}_k^T\boldsymbol{\beta})\phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\right].
\end{aligned}$$

In the equations above, $\mathrm{E}\left[I_{\{y_k>\tau\}}\right]$ is clearly $P(y_k>\tau)=P\left(Z>\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)=1-\Phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)$, and $\mathrm{E}\left[I_{\{Y_k>\tau\}}(y_k-\boldsymbol{x}_k^T\boldsymbol{\beta})^2\right]$ is computed using integration by parts as follows:

$$\begin{aligned}
&\mathrm{E}\left[I_{\{y>\tau\}}(y-\boldsymbol{x}^T\boldsymbol{\beta})^2\right]=\int_\tau^\infty\frac{(y-\boldsymbol{x}^T\boldsymbol{\beta})^2}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y-\boldsymbol{x}^T\boldsymbol{\beta})^2}dy\\
&\quad=-\frac{\sigma}{\sqrt{2\pi}}(y-\boldsymbol{x}^T\boldsymbol{\beta})e^{-\frac{1}{2\sigma^2}(y-\boldsymbol{x}^T\boldsymbol{\beta})^2}\Big|_\tau^\infty+\sigma^2\int_\tau^\infty\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y-\boldsymbol{x}^T\boldsymbol{\beta})^2}dy\\
&\quad=\sigma(\tau-\boldsymbol{x}^T\boldsymbol{\beta})\phi\left(\frac{\tau-\boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right)+\sigma^2\left(1-\Phi\left(\frac{\tau-\boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right)\right).
\end{aligned}$$

Finally, the expectation of the double partial with respect to σ and $\boldsymbol{\beta}$ (for uncensored observations) is

$$\sum_{k\in s}-\frac{2\boldsymbol{x}_k}{\sigma^3}\mathrm{E}\left[(y_k-\boldsymbol{x}_k^T\boldsymbol{\beta})I_{\{Y_k>\tau\}}\right]=\sum_{k\in s}-\frac{2\boldsymbol{x}_k^T}{\sigma^2}\phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right),$$

since

$$\begin{aligned}
&\mathrm{E}\left[I_{\{y>\tau\}}(y-\boldsymbol{x}^T\boldsymbol{\beta})\right]=\int_\tau^\infty\frac{(y-\boldsymbol{x}^T\boldsymbol{\beta})}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y-\boldsymbol{x}^T\boldsymbol{\beta})^2}dy\\
&\quad=-\frac{\sigma}{\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(y-\boldsymbol{x}^T\boldsymbol{\beta})^2}\Big|_\tau^\infty=\sigma\phi\left(\frac{\tau-\boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right).
\end{aligned}$$

Putting everything together, we have

$$\begin{aligned}
\mathrm{E}\left[\frac{\partial^2 l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\sigma)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right] &= \sum_{k\in s}\left\{-\frac{\boldsymbol{x}_k^T\boldsymbol{x}_k}{\sigma^2}\left(1-\Phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\right)\right. \\
&\quad -\left[\frac{1}{\Phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)}\phi^2\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\frac{\boldsymbol{x}_k^T\boldsymbol{x}_k}{\sigma^2}\right. \\
&\quad \left.\left.+\,\phi'_{\boldsymbol{\beta}}\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\boldsymbol{x}_k}{\sigma}\right)\right]\right\},
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\left[\frac{\partial^2 l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\sigma)}{\partial\sigma^2}\right] &= \sum_{k\in s}\left\{\left[-\frac{2}{\sigma^2}\left(1-\Phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\right)\right.\right. \\
&\quad \left.-\,\frac{3}{\sigma^3}(\tau-\boldsymbol{x}_k^T\boldsymbol{\beta})\phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\right] \\
&\quad -\left[\frac{1}{\Phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)}\phi^2\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma^2}\right)^2\right. \\
&\quad +\phi'_{\sigma}\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma^2}\right) \\
&\quad \left.\left.-\,2\phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma^3}\right)\right]\right\},
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left[\frac{\partial^2 l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\sigma)}{\partial\sigma\partial\boldsymbol{\beta}}\right] &= \sum_{k\in s}\left\{-\frac{2\boldsymbol{x}_k^T}{\sigma^2}\phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\right. \\
&\quad -\left(\frac{\boldsymbol{x}_k^T}{\Phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)}\phi^2\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma^3}\right)\right. \\
&\quad -\left[\phi\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\boldsymbol{x}_k^T}{\sigma^2}\right)\right. \\
&\quad \left.\left.\left.-\,\phi'_{\boldsymbol{\beta}}\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\tau-\boldsymbol{x}_k^T\boldsymbol{\beta}}{\sigma^2}\right)\right]\right)\right\}
\end{aligned}$$

Once the expected information is computed all that is left is to construct the estimated covariance matrix, $\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{J}}_w^{-1}\hat{\boldsymbol{K}}_w\hat{\boldsymbol{J}}_w^{-1} - \hat{\boldsymbol{J}}_1^{-1}$, and find the eigenvalues of $\hat{\boldsymbol{\Gamma}}^{T/2}\hat{\boldsymbol{J}}_1\hat{\boldsymbol{\Gamma}}^{1/2}$. This will

give us the weights for the linear combination of χ^2_1 random variables that are needed for the test.

If the test is applied to the smaller model (no effect for Hispanics), then the eigenvalues are $\boldsymbol{\lambda}^T = (1.154, 1.070, 0.973, 0.945, 0.729, 0.535)$; the 0.05 critical value for the asymptotic distribution computed from 10,000 simulated linear combinations of χ^2_1 random variables is 11.58. Letting $\boldsymbol{\theta} = (\sigma, \boldsymbol{\beta})$, the likelihood ratio test statistic is

$$T_1 = 2(l_1(\hat{\boldsymbol{\theta}}_1) - l_1(\hat{\boldsymbol{\theta}}_w)) = 19.38,$$

and so we strongly reject the null hypothesis of non-informativeness. If an indicator for people of Hispanic origins is included in the model and the test is repeated, the relevant eigenvalues are $\boldsymbol{\lambda}^T = (1.154, 1.061, 0.920, 0.912, 0.862, 0.587, 0.461)$; the 0.05 critical value is 12.14 and the likelihood ratio test statistic is

$$T_1 = 2(l_1(\hat{\boldsymbol{\theta}}_1) - l_1(\hat{\boldsymbol{\theta}}_w)) = 5.19.$$

Thus the design is no longer found to be informative. This indicates that all of the information held in the design (with respect to estimation of model parameters), in addition to model covariates, was contained in knowing whether a person is of Hispanic origins or not. The inclusion of this additional design information in the model has removed the model and design bias simultaneously.

2.6.3 Bootstrap Results

The bootstrap analogue to the LRT test is easy to apply; the distribution of the test statistic T_1 can be bootstrapped by simulating from the proposed model under the null hypothesis of non-informativeness, and then computing the value of the test statistic at each simulation step. This is a very nice feature since it allows for the possible avoidance of tedious information calculations, like for a Tobit regression model. Comparisons of bootstrapped and analytic distributions are shown below for both the model with an indicator for Hispanics (32), and the model without (31).

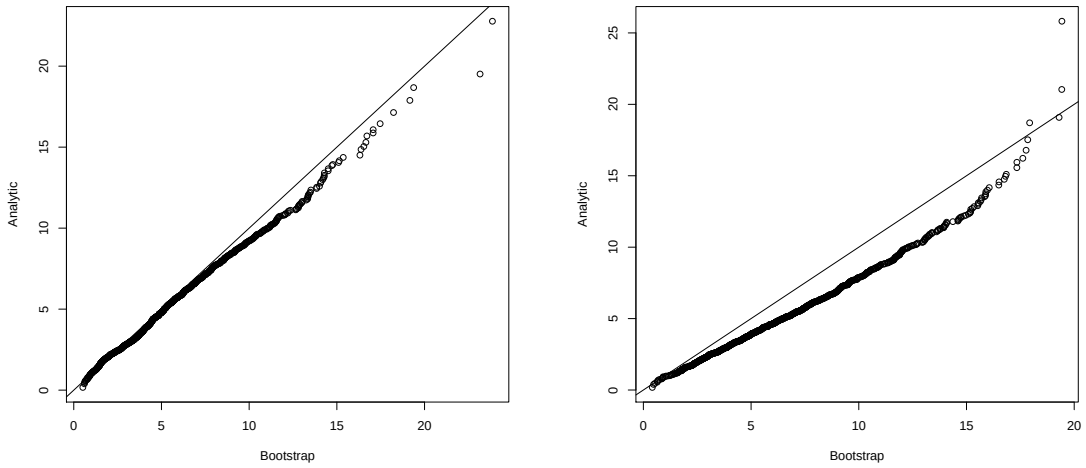

Figure 4: QQ plots for the analytic distribution vs. the bootstrapped test statistic for the model without an indicator for Hispanics (31) (left) and the model with the indicator (32) (right)

For the model without an indicator for Hispanic origins the bootstrap 0.05 critical value is 12.68 compared to a critical value of 11.44 from the analytic distribution; In the model which includes the indicator the bootstrap 0.05 critical value is 13.94 compared to 11.04 obtained from the analytic distribution. The asymptotic distributions closely mirror those distribution that are obtained via bootstrapping, however the asymptotic and bootstrap distributions agree more closely using the smaller model. In both cases our inference from either method is in agreement.

CHAPTER 3

SEMIPARAMETRIC APPROACHES TO MODEL BUILDING IN THE PRESENCE OF INFORMATIVE SAMPLING

3.1 Introduction

Consider the problem of estimating the conditional distribution of a variable y given x. Informative sampling occurs when

$$f(y_k \mid x_k, I_k = 1) \neq f(y_k \mid x_k).$$

If it is not possible to extend the covariate structure to account for the informativeness in the design then it may be possible to include extra design information in the model specification and then "integrate out" the additional design information later. In order to do this we need to have access to design variables, z, such that

$$f(y_k \mid x_k, z_k, I_k = 1) = f(y_k \mid x_k, z_k). \tag{33}$$

That is, the design is noninformative after conditioning on the information in z_k. Chambers defines noninformativeness as in equation (33): conditional independence of the population generating process and the sample selection process given z. Here we only wish to make the distinction that variables of scientific interest are considered to be valid model variables, x, and design variables that are not scientifically interesting (such as an inclusion probability) are contained in z. The auxillary information in z is necessary for performing unbiased inference, but is not otherwise interesting. In short, we consider the scenario in which the design is informative given only x, but noninformative given x and z.

Assuming (33) holds, we can estimate $f(y \mid x, z)$ in an unbiased way using the sample data. This is important because in the presence of design informativeness the model that holds at the sample level is different from the model that holds at the population level, and since our inferential goal is the latter, the sample data will provide biased results in general. Furthermore, if our inferential goal is $f(y \mid x)$ then we would like to investigate ways of extracting $f(y \mid x)$, which we can not estimate directly from the sample data, from

$f(y \mid x, z)$, which we can estimate directly from the sample data. The following sections discuss integrating out design effects to obtain $f(y \mid x)$ via

$$f(y \mid x) = \int f(y \mid x, z) f(z \mid x) dz.$$

In the context of regression (or any conditional expectation problem), notice that

$$\mathrm{E}\left[\mathrm{E}\left[y \mid x, z\right] \mid x\right] = \int \int y f(y \mid x, z) dy dz = \int y \int f(y \mid x, z) dz dy = \mathrm{E}\left[y \mid x\right],$$

so design variables can be integrated out of the mean structure via iterating expectations.

In the next section, a semiparametric approach to this problem will be proposed. In subsequent sections further asymptotic properties will be derived, and detailed applications will be made.

3.2 A Semiparametric Model

Suppose $y = \boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{f}^T(\boldsymbol{x}, \boldsymbol{z})\boldsymbol{\gamma} + \epsilon$, where $\mathrm{E}\left[\epsilon \mid \boldsymbol{x}, \boldsymbol{z}\right] = 0$. Here $\boldsymbol{x}$ contains model variables and $\boldsymbol{z}$ contains design variables (such as weights, strata, clusters). Writing $\mathrm{E}\left[\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{z}) \mid \boldsymbol{x}\right] = \Gamma(\boldsymbol{x})$, we have $\mathrm{E}\left[y \mid \boldsymbol{x}\right] = \mu(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} + \Gamma^T(\boldsymbol{x})\boldsymbol{\gamma}$. Our goal is to integrate the design effects out of $f(y|\boldsymbol{x}, \boldsymbol{z})$ and find the mean function as a function of model variables only. The proposed estimator is

$$\widehat{\mu}(\boldsymbol{x}) = \boldsymbol{x}^T\widehat{\boldsymbol{\beta}} + \widehat{\Gamma}_\pi^T(\boldsymbol{x})\widehat{\boldsymbol{\gamma}}, \tag{34}$$

where $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)$ come from the regression of y on $(\boldsymbol{x}^T, \boldsymbol{f}^T(\boldsymbol{x}, \boldsymbol{z}))$, and $\widehat{\Gamma}_\pi(\boldsymbol{x})$ is a design-based estimator of the finite population quantity corresponding to $\Gamma(\boldsymbol{x})$. Since the regression coefficients are estimated via ordinary least squares they can be written as

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{pmatrix} = \left(\sum_{k \in U} \begin{pmatrix} \boldsymbol{x}_k \\ \boldsymbol{z}_k \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_k & \boldsymbol{z}_k \end{pmatrix} I_k \right)^{-1} \left(\sum_{k \in U} \begin{pmatrix} \boldsymbol{x}_k \\ \boldsymbol{z}_k \end{pmatrix} y_k I_k \right). \tag{35}$$

Let the design-based estimator for $\Gamma(x)$ be of the form

$$\widehat{\Gamma}_{\pi}(x) = \left(\sum_{k \in U} K\left(\frac{x - x_i}{h} \right) \frac{I_k}{\pi_k} \right)^{-1} \left(\sum_{k \in U} K\left(\frac{x - x_i}{h} \right) z_k \frac{I_k}{\pi_k} \right), \tag{36}$$

where $K(\cdot)$ is a kernel function, assumptions for which will be given in a following section. The model is semiparametric in the sense that the model parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, will be estimated using standard parameteric regression methods, and $\Gamma(\boldsymbol{x})$ will be estimated non-parametrically under a classical design-based framework; this accomplishes the task of integrating out the design effects. The model would be fit in three steps: first $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ would be estimated via ordinary least squares, then $\Gamma(\boldsymbol{x})$ would be estimated by an appropriate design-based estimator, and then the two pieces would be combined.

Conditions for the joint asymptotic normality of the design and model-based pieces will be addressed in the following sections, and asymptotic results will follow.

3.3 Notation and Assumptions

Standard survey sampling notation will be used throughout. For design-based components of the estimation problems that follow we will consider a finite population of N elements contained in the set $U = \{1, 2, \ldots, N\}$. The subset of elements contained in the sample will be denoted $s \subseteq U$. Sample membership indicators are defined as

$$I_k = \begin{cases} 1, \, k \in s, \\ 0, \, k \notin s, \end{cases},$$

where $\mathrm{E}[I_k] = \pi_k$. Here π_k is the probability that the k^{th} element will be included in the sample, and $w_k = \pi_k^{-1}$ are the sampling weights. Totals will be denoted by a lower case t and will represent sums over U; subscripts will indicate what is being summed up. For example,

$$t_y = \sum_{k \in U} y_k$$

is the sum of all y in the finite population, an estimator for which is given by

$$\widehat{t}_y = \sum_{k \in U} y_k \frac{I_k}{\pi_k}.$$

This is the Horvitz-Thompson estimator (Horvitz and Thompson (1952)) of the finite population total, and will be used to estimate totals for the remainder of the paper, unless otherwise specified.

For analytic inference we will consider a sequence of finite populations U_N, and focus inference on the underlying process (superpopulation) that generates the population of observations from which we are sampling. Asymptotic results will rely on the population size N going to infinity, and thus the sample size n and the bandwidth h can be written as n_N and h_N indicating that as $N \to \infty$, $n \to \infty$ and $h \to 0$ at rates that will be specified in the assumptions section. The N subscript will be suppressed in much of what follows.

The following assumptions will be made to prove the theoretical results and follow closely those assumptions made by Breidt and Opsomer (2000).

B1. For each N, the x_i are considered fixed with respect to the superpopulation model. The x_i are independent and identically distributed $F(x) = \int_{-\infty}^{x} f(t)\,dt$, where $f(\cdot)$ is a density with compact support $[a_x, b_x]$ with $f(x) > 0$, $\forall\, x \in [a_x, b_x]$.
In addition, the z_i are uniformly bounded with compact support $[a_z, b_z]$.
B2. As $N \to \infty$, $n_N N^{-1} \to \pi \in (0,1)$, $h_N \to 0$, and $N h_N^3 \to \infty$.
B3. For all N, $\min_{i \in U_N} \pi_i \geq \lambda > 0$, $\min_{i,j \in U_N} \pi_{ij} \geq \lambda > 0$, and

$$\limsup_{N\to\infty} n_N \max_{i \neq j} |\pi_{ij} - \pi_i \pi_j| < \infty,$$

and

$$\lim_{N\to\infty} \max_{(i_1,i_2,i_3,i_4)\in D_4} |\mathrm{E}\left[(I_{i_1} I_{i_2} - \pi_{i_1}\pi_{i_2})(I_{i_3} I_{i_4} - \pi_{i_3}\pi_{i_4})\right]| = O(N^{-1}),$$

where D_t is the set of all distinct t-tuples from U_N.
B4. The kernel function $K(\cdot)$ is symmetric, continuous and bounded, and has compact support.
B5. The function $\Gamma(x)$ is a continuous and differentiable function.
B6. For $\boldsymbol{\lambda} \neq \mathbf{0}$,

$$\sum_{k\in U} \left[\boldsymbol{\lambda}^T \begin{pmatrix} \boldsymbol{x}_k \\ \boldsymbol{z}_k \end{pmatrix} \right]^4 = O(N),$$

and the limit

$$\lim_{N\to\infty} N^{-1} \sum_{k\in U} \boldsymbol{\lambda}^T \begin{pmatrix} \boldsymbol{x}_k \\ \boldsymbol{z}_k \end{pmatrix} \sigma^2 I_k = \lim_{N\to\infty} N^{-1} \sum_{k\in U} \boldsymbol{\lambda}^T \begin{pmatrix} \boldsymbol{x}_k \\ \boldsymbol{z}_k \end{pmatrix} \sigma^2 \pi_k$$

exists, where σ^2 is the error variance for the model. Additionally, for $\lambda \neq 0$,

$$N^{-1/2}\lambda \sum_{i\in U} \frac{1}{\sum_{j\in U} K\left(\frac{x-x_i}{h}\right)} \left[z_i - \frac{\sum_{j\in U} K\left(\frac{x-x_i}{h}\right) z_j}{\sum_{j\in U} K\left(\frac{x-x_i}{h}\right)} \right] K\left(\frac{x-x_i}{h}\right) \left(\frac{I_i}{\pi_i} - 1\right)$$

is asymptotically normally distributed with mean 0 and variance Σ_d, where Σ_d is the asymptotic, design-based variance of the expression on the left.
B7. $\lim_{N\to\infty} N^{-1} \sum_{k\in U} \begin{pmatrix} \boldsymbol{x}_k \\ \boldsymbol{z}_k \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_k & \boldsymbol{z}_k \end{pmatrix} I_k = \boldsymbol{A}$, a positive definite matrix.

Assumptions B5 and B6 are used to establish a central limit theorem for the semiparametric estimator. It is common to make a central limit theorem assumption for the design, and conditions can be checked on a design by design basis (see for example Fuller (2009)). The inclusion of $\lambda \neq 0$ is for the proof which relies on the Cramér-Wold device which says that a vector of random variables is jointly normally distributed if any linear combination of the variables is univariate normal.

3.4 Limiting Distribution of $\widehat{\mu}(\boldsymbol{x})$

This section introduces a central limit theorem for $\widehat{\mu}(\boldsymbol{x})$. The asymptotic variances for the estimators given in equations (35) and (36), which will be needed for obtaining the joint limiting distribution of the parameter estimates, are obtained as follows. The model-based and design-based components of the estimator are in fact uncorrelated (as shown in Lemma 1 in the Appendix), hence the variance components for the estimated model coefficients and design-based estimator can be found separately. Denote these variance components as $\boldsymbol{\Sigma}_m$ and Σ_d. The matrix $\boldsymbol{\Sigma}_m$ is obtained using standard least squares regression results, so we immediately have that

$$n\boldsymbol{\Sigma}_m \equiv n\text{Var}\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{pmatrix} = \boldsymbol{A}^{-1}\sigma^2, \tag{37}$$

where $\boldsymbol{A}$ is defined as in assumption B6. Next define

$$\Sigma_d \equiv n\text{Var}\left(\widehat{\Gamma}(x) \mid \boldsymbol{X}, \boldsymbol{Z}\right) = \lim_{N\to\infty} n \sum_{i,j\in U}\sum (\pi_{ij} - \pi_i\pi_j)\frac{d_i}{\pi_i}\frac{d_j}{\pi_j}, \tag{38}$$

where

$$d_i = \frac{1}{\sum_{j\in U} K\left(\frac{x-x_j}{h}\right)}\left[z_i - \frac{\sum_{j\in U} K\left(\frac{x-x_j}{h}\right) z_j}{\sum_{j\in U} K\left(\frac{x-x_j}{h}\right)}\right] K\left(\frac{x-x_j}{h}\right). \tag{39}$$

This is a standard design-based variance calculation, which is obtained by taking the variance of the linearized form of $\widehat{\Gamma}(x)$ from Lemma 5 in the Appendix.

We may now state and prove the following theorem on the joint asymptotic normality of

the model and design pieces for the estimator given in (34).

Theorem 10. *Under B5–B6,*

$$n^{1/2}\begin{bmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \\ \widehat{\Gamma}_\pi(\boldsymbol{x}) - \Gamma(\boldsymbol{x}) \end{bmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{\Sigma}_m & \mathbf{0} \\ \mathbf{0} & \Sigma_d \end{bmatrix}\right),$$

where $\boldsymbol{\Sigma}_m$ and Σ_d are defined as in equations (37) and (38).

3.5 Variance Estimation

The asymptotic variances given by (37) and (38) can be estimated as follows:

$$\widehat{\mathrm{V}}\begin{bmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \end{bmatrix} = \left[\begin{pmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{pmatrix}\right]^{-1} \widehat{\sigma}^2, \tag{40}$$

where the model variance, σ^2, can be estimated by the model mean square error as usual:

$$\widehat{\sigma}^2 = \frac{\sum_{i\in U}\left[y_i - \begin{pmatrix} \boldsymbol{x}_i^T & \boldsymbol{z}_i^T \end{pmatrix}\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{pmatrix}\right]^2 I_i}{n-p-q}.$$

The design component can be estimated via a linearization argument as follows:

$$\widehat{\mathrm{V}}(\widehat{\Gamma}_\pi(x)) = \sum_{i\in U}\sum_{j\in U}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{\widehat{d}_i}{\pi_i}\frac{\widehat{d}_j}{\pi_j}I_iI_j, \tag{41}$$

where $\widehat{d}_i$ corresponds to the d_i defined in equation (39) with sums over the population replaced by sums over the sample, weighted by inverse inclusion probabilities. These estimates are in fact design-consistent for their targets under mild assumptions.

Theorem 11. *Under B1–B4,*

$$\lim_{N\to\infty} E\left[|\widehat{\Sigma}_d - \Sigma_d|\right] = 0,$$

where

$$\Sigma_d = \frac{n}{\left(\sum_{i\in U} K\left(\frac{x-x_i}{h}\right)\right)^2} \sum_{i,j\in U} \left[z_i - \Gamma(x)\right] K\left(\frac{x-x_i}{h}\right) \left[z_j - \Gamma(x)\right] K\left(\frac{x-x_j}{h}\right) \frac{\Delta_{ij}}{\pi_i \pi_j},$$

and

$$\begin{aligned} \widehat{\Sigma}_d &= \frac{n}{\left(\sum_{j\in U} K\left(\frac{x-x_j}{h}\right)\frac{I_j}{\pi_j}\right)^2} \sum_{i,j\in U} \left[z_i - \widehat{\Gamma}(x)\right] K\left(\frac{x-x_i}{h}\right) \left[z_j - \widehat{\Gamma}(x)\right] \\ &\quad \times K\left(\frac{x-x_j}{h}\right) \frac{\Delta_{ij}}{\pi_i \pi_j} \frac{I_i I_j}{\pi_{ij}}. \end{aligned}$$

3.5.1 Consistency of $\widehat{\mu}(\boldsymbol{x})$

In this section consistency of the estimator will be established along with variance estimates so that we may apply error bounds to the semiparametric fits.

The following result shows that (34) is a mean square consistent estimator for $\mu(\boldsymbol{x})$.

Theorem 12. *Assume B1–B4, then*

$$\widehat{\mu}(\boldsymbol{x}) = \boldsymbol{x}^T\widehat{\boldsymbol{\beta}} + \widehat{\Gamma}_\pi^T(\boldsymbol{x})\widehat{\boldsymbol{\gamma}}$$

is mean square consistent in the sense that

$$\lim_{N\to\infty} E\left[\{\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})\}^2\right] = 0.$$

3.5.2 Variance Estimation for $\widehat{\mu}(\boldsymbol{x})$

To derive the asymptotic variance of the semiparametric estimator given by equation (34), begin by writing

$$\begin{aligned}
\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x}) &= \boldsymbol{x}^T\widehat{\boldsymbol{\beta}} + \widehat{\Gamma}_\pi^T(\boldsymbol{x})\widehat{\boldsymbol{\gamma}} - \boldsymbol{x}^T\boldsymbol{\beta} - \Gamma^T(\boldsymbol{x})\boldsymbol{\gamma} \\
&= \boldsymbol{x}^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \Gamma^T(\boldsymbol{x})(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + (\widehat{\Gamma}_\pi(\boldsymbol{x}) - \Gamma(\boldsymbol{x}))^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} + \boldsymbol{\gamma}) \\
&= \boldsymbol{x}^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \Gamma^T(\boldsymbol{x})(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + \boldsymbol{\gamma}^T(\widehat{\Gamma}_\pi(\boldsymbol{x}) - \boldsymbol{\Gamma}(\boldsymbol{x})) \\
&\quad +(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T(\widehat{\Gamma}_\pi(\boldsymbol{x}) - \Gamma(\boldsymbol{x})) \\
&= \boldsymbol{x}^T\boldsymbol{a} + \Gamma^T(\boldsymbol{x})\boldsymbol{b} + \boldsymbol{\gamma}^T\boldsymbol{c} + \boldsymbol{b}^T\boldsymbol{c}.
\end{aligned}$$

As noted in the Appendix, $\boldsymbol{a}$ and $\boldsymbol{b}$ are $O_p(n^{-1/2})$, and $\boldsymbol{c}$ is $O_p((nh)^{-1/2})$, so $\boldsymbol{b}^T\boldsymbol{c}$ has a smaller order of $O_p(nh^{-1/2})$. It will then suffice to estimate the variance of $\boldsymbol{x}^T\boldsymbol{a} + \Gamma^T(\boldsymbol{x})\boldsymbol{b} + \boldsymbol{\gamma}^T\boldsymbol{c}$ conditional on $\boldsymbol{x}$. Thus for purposes of application, we will calculate the variance as

$$\begin{aligned}
\mathrm{Var}\left(\boldsymbol{x}^T\boldsymbol{a} + \Gamma^T(\boldsymbol{x})\boldsymbol{b} + \boldsymbol{\gamma}^T\boldsymbol{c} \mid \boldsymbol{x}\right) &= \mathrm{Var}\left(\boldsymbol{x}^T\boldsymbol{a} + \Gamma^T(\boldsymbol{x})\boldsymbol{b} \mid \boldsymbol{x}\right) + \mathrm{Var}\left(\boldsymbol{\gamma}^T\boldsymbol{c} \mid \boldsymbol{x}\right) \\
&\quad +2\mathrm{Cov}\left(\boldsymbol{x}^T\boldsymbol{a} + \Gamma^T(\boldsymbol{x}), \boldsymbol{\gamma}^T\boldsymbol{c} \mid \boldsymbol{x}\right). \qquad (42)
\end{aligned}$$

In (42), $\boldsymbol{a}$ and $\boldsymbol{b}$ represent model pieces, and $\boldsymbol{c}$ represents a design piece. As in Lemma 1 from the Appendix, the model and design pieces will be exactly uncorrelated and so the remaining variance components can be estimated individually and added. For the model

piece, the conditional variance formula will be applied to obtain

$$\begin{aligned}
&\mathrm{Var}\left(\boldsymbol{x}^T\boldsymbol{a} + \Gamma^T(\boldsymbol{x})\boldsymbol{b} \mid \boldsymbol{x}\right) \\
&= \mathrm{Var}\left(\mathrm{E}\left[\boldsymbol{x}^T\boldsymbol{a} + \Gamma^T(\boldsymbol{x})\boldsymbol{b} \mid \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{I}\right] \mid \boldsymbol{x}\right) \\
&\quad + \mathrm{E}\left[\mathrm{Var}\left(\boldsymbol{x}^T\boldsymbol{a} + \Gamma^T(\boldsymbol{x})\boldsymbol{b} \mid \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{I}\right) \mid \boldsymbol{x}\right] \\
&= \mathrm{E}\left[\begin{pmatrix} \boldsymbol{x} & \Gamma(\boldsymbol{x}) \end{pmatrix}^T \left[\begin{pmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{pmatrix} \right]^{-1} \sigma_\epsilon^2 \begin{pmatrix} \boldsymbol{x} \\ \Gamma(\boldsymbol{x}) \end{pmatrix} \right],
\end{aligned}$$

where $\boldsymbol{X}$ is a matrix of model covariates and $\boldsymbol{Z}$ is a matrix of $f(\boldsymbol{x}, \boldsymbol{z})$ values.

For the design piece, we again begin by conditioning to obtain

$$\begin{aligned}
\mathrm{Var}\left(\boldsymbol{\gamma}^T\boldsymbol{c} \mid \boldsymbol{x}\right) &= \mathrm{Var}\left(\mathrm{E}\left[\boldsymbol{\gamma}^T\boldsymbol{c} \mid \boldsymbol{x}, \boldsymbol{z}\right] \mid \boldsymbol{x}\right) + \mathrm{E}\left[\mathrm{Var}\left(\boldsymbol{\gamma}^T\boldsymbol{c} \mid \boldsymbol{x}, \boldsymbol{z}\right) \mid \boldsymbol{x}\right] \\
&= \mathrm{E}\left[\mathrm{Var}\left(\boldsymbol{\gamma}^T\boldsymbol{c} \mid \boldsymbol{x}, \boldsymbol{z}\right) \mid \boldsymbol{x}\right],
\end{aligned}$$

since $\mathrm{E}\left[\boldsymbol{c} \mid \boldsymbol{x}, \boldsymbol{z}\right] = 0$. Now, $\mathrm{Var}\left(\boldsymbol{\gamma}^T\boldsymbol{c} \mid \boldsymbol{x}, \boldsymbol{z}\right)$ is approximated as in (41) (via a linearization argument), and the model and design pieces are exactly uncorrelated (by Lemma 1 in the Appendix).

3.6 Empirical Applications

This section details applications of the proposed estimator in the context of regular linear regression with one predictor. One such data set can be found in Chapter 6 of Fuller (2009). This example (Example 6.3.3) describes data simulated to approximate data from the Canadian Workplace and Employee Survey.

3.6.1 Canadian Workplace and Employee Survey Data

The original survey is described by Patak, Hidiroglou, and Lavallee (1998), where the data come from a stratified simple random sample of workplaces in which three strata are defined based on a function of existing tax records that is highly correlated with payroll. The strata have highly variable sampling rates which leads to informative sampling. The model of interest relates payroll to employment through the function

$$\ln(\text{payroll}) = \beta_0 + \beta_1 \ln(\text{total employment}) + \varepsilon, \tag{43}$$

where the error terms are $iid(0, \sigma^2)$. The data consist of employment and payroll numbers along with sampling weights. Throughout, denote ln(payroll) by y, ln(total employment) by x, and the sample weights by w. Then fitting model (43) via ordinary least squares (OLS) yields

$$\widehat{y} = 10.019 + 0.907x,$$

with $\widehat{\sigma}^2 = 0.320$. The probability weighted regression yields

$$\widehat{y} = 9.745 + 0.931x.$$

The second estimator is weighted to adjust for potential bias that has entered through the sample selection process. If there is no bias present then the weighted estimates will be inefficient compared to the unweighted. To test the hypothesis that these procedures are in fact estimating the same quantities, we extend the model to include sampling weights and a weight by log of employment interaction term then use the classic F test for sub-models (e.g. DuMouchel and Duncan (1983)) to test whether or not the smaller model is adequate. This is a practical test for informativeness in this case because it is powerful for detecting informativeness in the mean structure for regular linear regression models and very easy to apply in practice. The OLS estimator for the extended model is

$$\widehat{y} = 10.888 + 0.722x - 0.0004w + 0.000016wx,$$

with $\widehat{\sigma}^2 = 0.2663$. The F test gives a p-value < 0.001, and so we reject the null hypothesis of non-informativeness. To be thorough, the results from the test proposed in Chapter 1 should be mentioned as well. In this case the two versions of the test statistic produce different results; using the unweighted log likelihood function and the test statistic T_1 we obtain a p-value of 0.16, and fail to detect informativeness in the design, however, T_w gives a p-value of 0.002, which is highly significant. It is unclear why there is a discrepancy in this case, and this kind of problem was not observed in any other simulations or power calculations.

To continue the analysis, we could extend the model further with a quadratic term for the weights and conduct the F-test again to test for additional informativeness; this test is not significant, and so we will work with the full interaction model above. Our goal now is to integrate the weights out of the regression model and obtain a model that is a function

of x only. In this context, the proposed semiparametric model is

$$y = \beta_0 + \beta_1 x + \gamma_1 f_1(x, z) + \gamma_2 f_2(x, z) + \varepsilon, \tag{44}$$

where $z = w$, $f_1(x, z) = w$ and $f_2(x, z) = wx$, and $\mathrm{E}[\varepsilon \mid x, w] = 0$. Taking expectation we have

$$\mu(x) = \mathrm{E}[y \mid x] = \beta_0 + \beta_1 x + \gamma_1 \Gamma(x) + \gamma_2 x \Gamma(x),$$

where $\Gamma(x) = \mathrm{E}[w \mid x]$.

Figure 5 shows scatterplots of the relationships of interest.

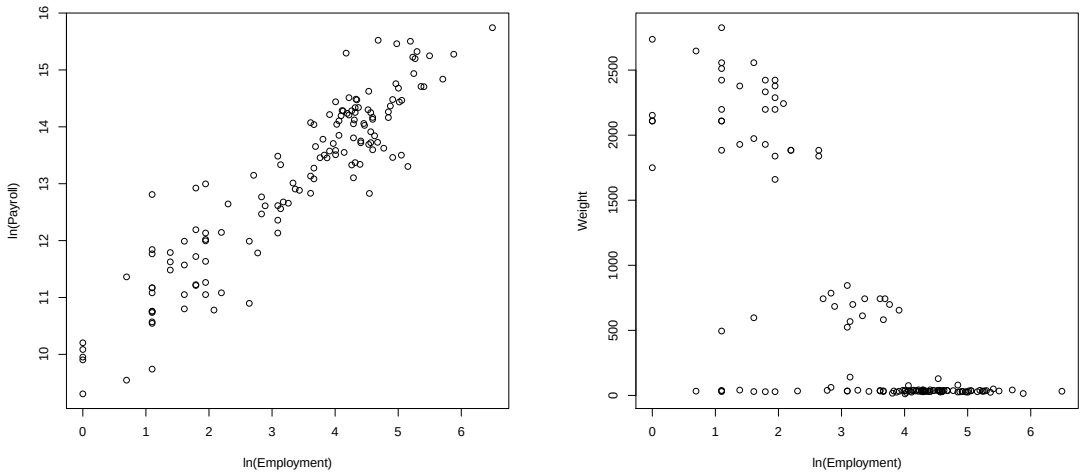

Figure 5: Scatterplots for log of payroll vs. log of employment (left), and sample weight vs. log of employment (right).

The scatterplot on the left represents the relationship we want to model; the scatterplot on the right shows the relationship we will use to integrate out the design weights. Our approach will be to estimate model (44) using OLS and then estimate $\Gamma(x)$ using a design-based estimator. The first step is easy and requires little discussion. The second part will require a little more discussion and multiple ideas will be considered. The estimated regression coefficients are $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\gamma}_1, \widehat{\gamma}_2) = (10.888, 0.722, -0.00043, 0.000016)$. The last two estimated coefficients are very small, but they are attached to weights which can be quite large in this example. Partial output from `R` is included below.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     10.8881203  0.1943763  56.016  < 2e-16 ***
lnemploy         0.7218792  0.0461403  15.645  < 2e-16 ***
Weight          -0.0004263  0.0001142  -3.731 0.000277 ***
lnemploy:Weight  0.0000160  0.0000534   0.300 0.764948
```

Next we will estimate $\Gamma(x)$ using a design-based method. Fuller noted that when the weights are plotted against x they fall roughly into three intervals. Define the groups by the intervals $(0, 2.67)$, $[2.67, 3.95)$, and $[3.95, \infty)$. With intervals defined in this way we will investigate two possible estimators for $\Gamma(x)$. First, recall that $\Gamma(x) = \mathrm{E}[w \mid x]$, and so consider estimating the means within each interval with a sample-weighted estimator. For this problem, a reasonable estimate for $\Gamma(x)$ is then

$$\begin{aligned}\widehat{\Gamma}_{\pi}(x) &= \frac{\sum_{x<2.67} w_k^2}{\sum_{x<2.67} w_k} I_{\{x<2.67\}} \\ &\quad + \frac{\sum_{2.67\le x<3.95} w_k^2}{\sum_{2.67\le x<3.95} w_k} I_{\{2.67\le x<3.95\}} + \frac{\sum_{x\ge 3.95} w_k^2}{\sum_{x\ge 3.95} w_k} I_{\{x\ge 3.95\}}. \end{aligned} \tag{45}$$

A second approach would be to use the kernel regression method proposed in this chapter. To apply this method I chose a Nadaraya-Watson smoother with a bandwidth of 1/2 to smooth the sample weights on log of total employment. The two fits can be seen in the graphs below.

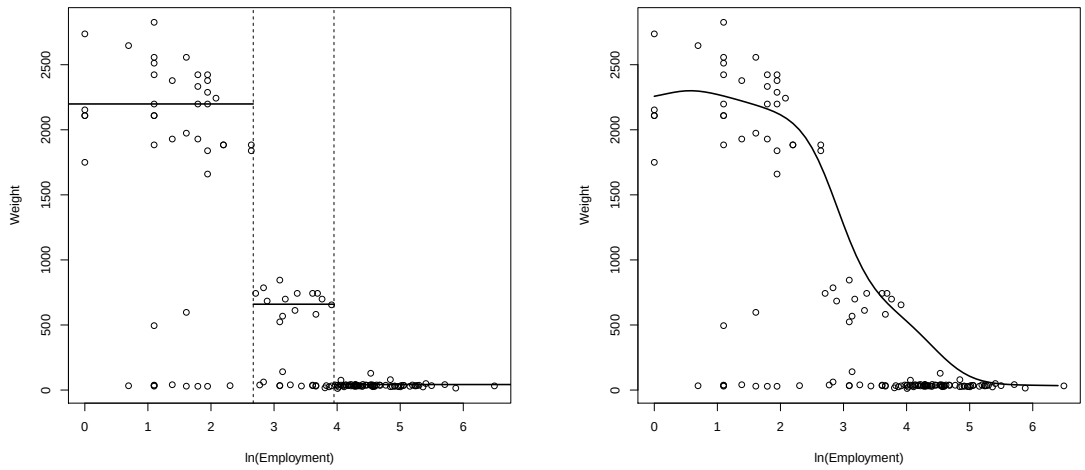

Figure 6: $\Gamma(x)$ estimated by weighted means within groups (left) and weighted kernel regression (right).

Combining these design-based fits for $\Gamma(x)$ with the estimated model coefficients produces the estimates for $\mu(x)$ seen in Figure 7.

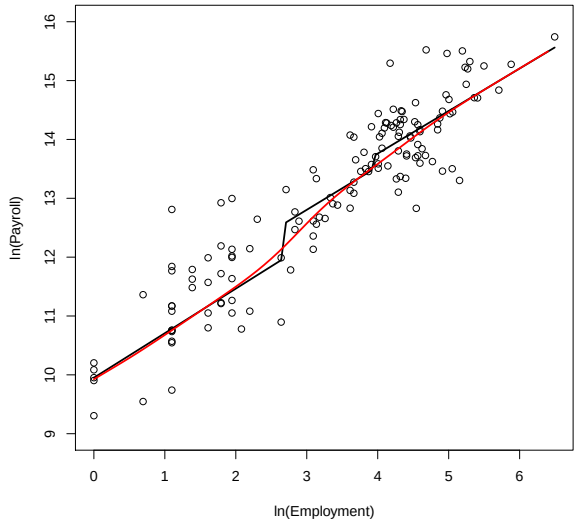

Figure 7: Regression fits using weighted sample weight means for $\widehat{\Gamma}(x)$ (black), and weighted kernel regression of sample weights on ln(total employment) for $\widehat{\Gamma}(x)$ (red).

Estimating $\mathrm{E}[w \mid x]$ by a within-groups mean estimate seems like a very logical thing to do in this situation, and indeed it accomplishes our goal of integrating out the sample weights. The semiparametric fit, however, obtains the same thing with a smooth function. The jumps and angles in the fit that estimates the weights by within-strata design-based mean estimates are artifacts of the zero-one indicator variables that are present. The semiparametric fit follows the same trend with a smooth curve. Next we would like to apply error bounds to $\widehat{\mu}(x)$.

Since the sample is drawn via stratified sampling within three strata we can approximate the variance for the design piece by applying a form appropriate for stratified sampling (e.g. Sarndal, Swennson, and Wretman, 1992, Section 3.7). Applying the error bounds as derived in equation (40) to the semiparametric fit we obtain the band shown in Figure 8.

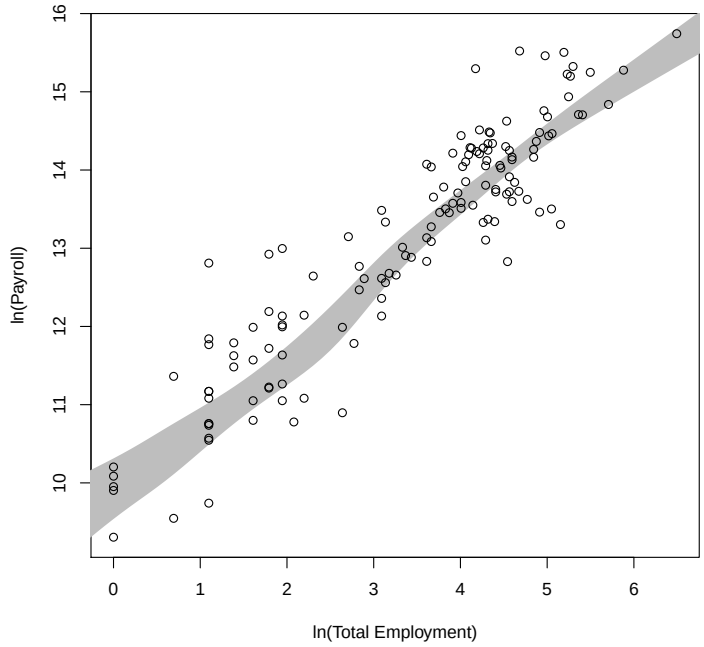

Figure 8: Confidence bands applied to the semiparametric estimate of $\mu(\boldsymbol{x})$

3.7 Simulation Study

This section details simulation results comparing the proposed semiparametric model to the ordinary least-squares model and the weighted ordinary least-squares model, in a scenario roughly mimicking the data from the Canadian Workplace data set. The data are simulated by generating a population of 86514 y variables from a linear model that is a function of x and z, where x and z are jointly normal random variables. The population was then stratified based on quantiles of z and sampled proportional to z within each of three strata. Three different sampling rates were used within each strata, and so the sampling is indeed informative. Since z is related to x, x will also hold explanatory power for the weights.

The simulation was performed according to the following steps:

- Simulate $x \sim \mathcal{N}(100, 10^2)$
- Get z from the model $z_k = 8x_k + \eta$, where $\eta \sim \mathcal{N}(0, 20^2)$
- Get y from the model $y_k = 100 + 0.5x_k + z_k + 2z_k x_k + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 7^2)$
- Stratify the population into three strata defined by the 40^{th} and 60^{th} percentiles of z
- Take samples of size $n_1 = 30$, $n_2 = 25$, $n_3 = 90$ with probabilities proportional to z within each stratum

Notes: The Canadian Workplace data comes from a stratified sample in which the strata are defined using previous tax records, which were not available at the data analysis step, only inclusion probabilities and a covariate are available; this is the function that the z variable is playing here — it is used to stratify the data and then it is lost to the analyst so that we must rely on the relationship between the covariate and the weights to integrate out the design information.

For illustrative purposes, a single realization of the simulation will be presented, and then summary tables will follow. The tabled results include coverage rates for (nominally) 95% confidence intervals and mean-square prediction error for prediction across the finite population of values. To begin, Figure 9 is a scatterplot of y vs. x.

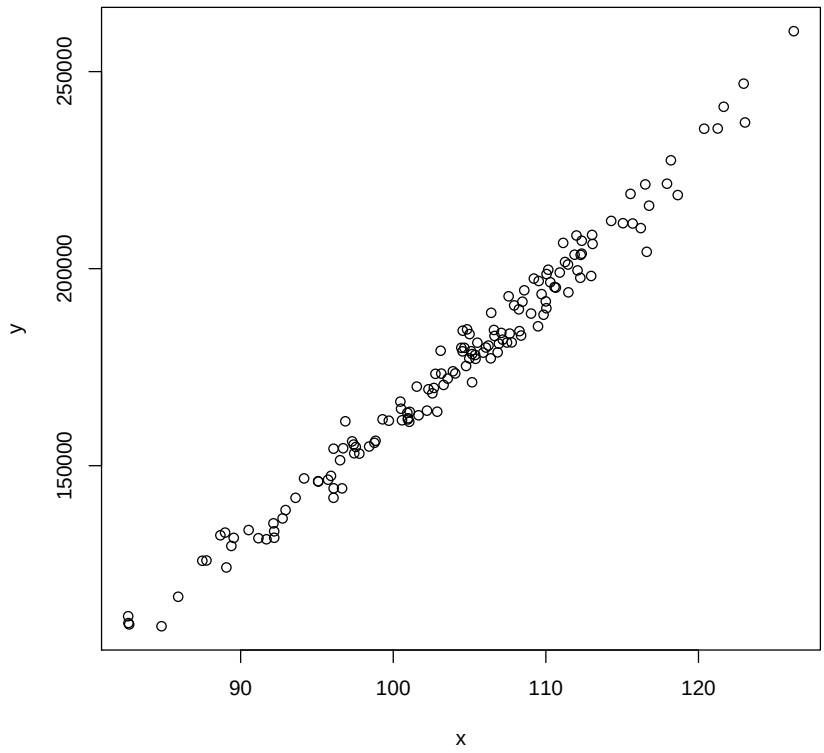

Figure 9: Scatterplot of y vs. x.

If one were to fit an ordinary least-squares regression line to this scatterplot, the following fit would result. Compared with the weighted OLS fit (dotted line), there is visually very little difference between the two fits.

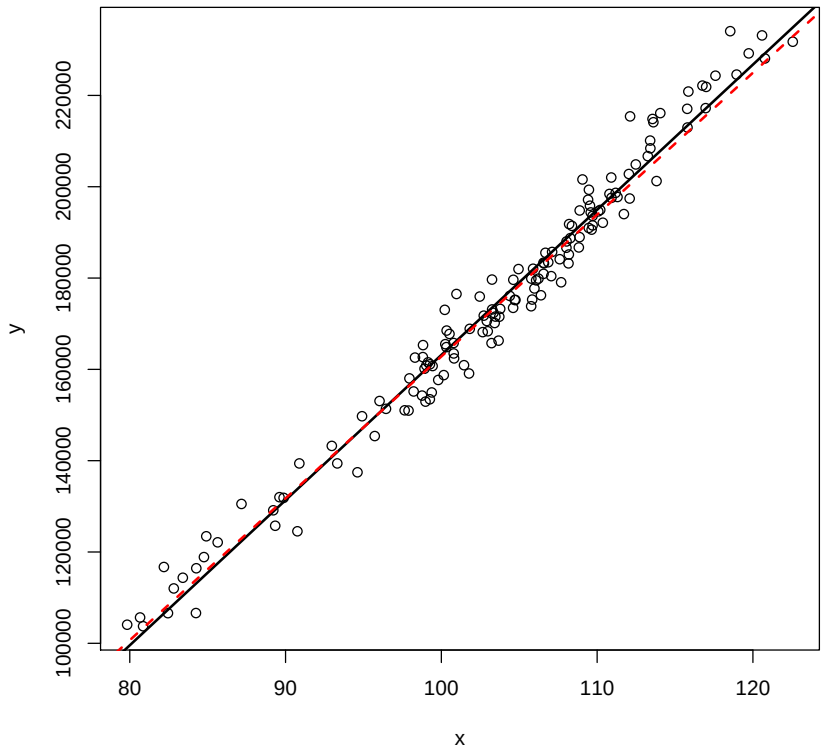

Figure 10: Scatterplot with weighted (red) and unweighted (black) fits.

Furthermore, standard residual diagnostics do not appear to raise any concerns about standard model assumptions such as constant variance (notice the random scatter of points seen in the residuals vs fits plot in Figure 11) and normally distributed errors (Notice that in the normal probability plot in Figure 11 the theoretical and sample quantiles match almost perfectly).

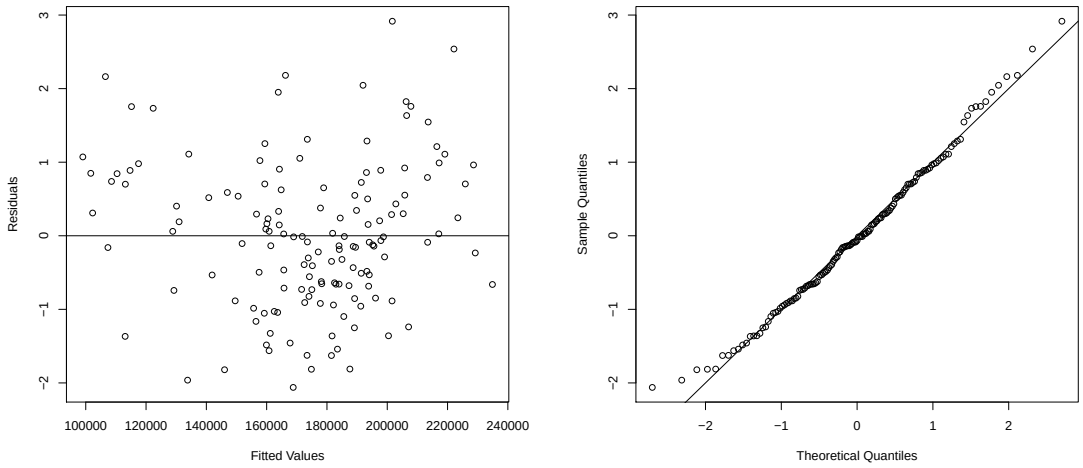

Figure 11: Standardized residuals vs. fitted values (left) and normal probability plot of standardized residuals (right)

However, the F-test for informative sampling strongly rejects the null hypothesis of non-informativeness, as can be seen in the following partial output from R.

```
Model 1: y ~ x
Model 2: y ~ x * w
  Res.Df        RSS Df Sum of Sq      F    Pr(>F)
1    143 3171356015
2    141 2307974435  2 863381580 26.373 1.862e-10 ***
```

Following the analysis steps outlined in the applications section, we expand the model to include sample weights and a sample weights by x interaction term, then smooth the sample weights on x with a (weighted) Nadaraya-Watson smooth and combine the two fits. In Figure 12 the nonparametric smooth is shown, along with the final semiparametric fit.

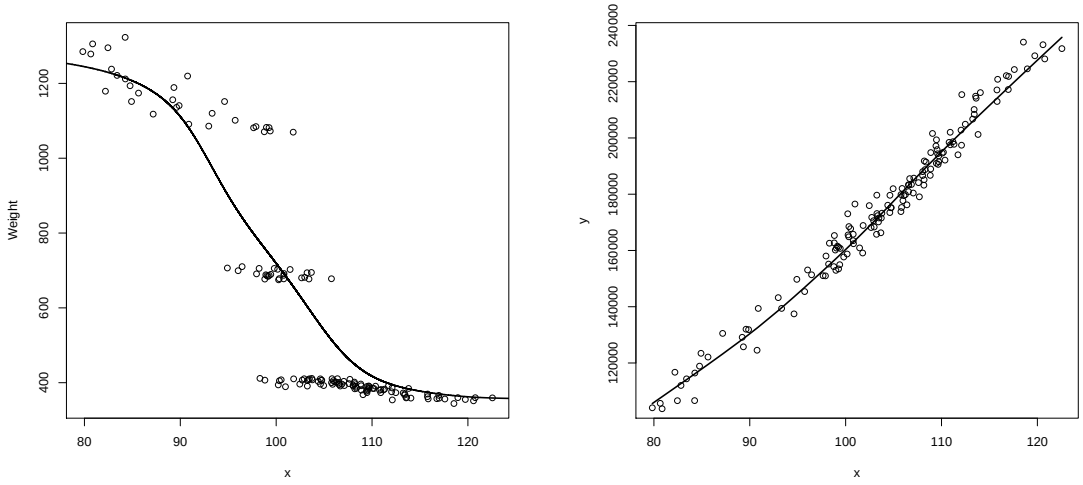

Figure 12: Weighted Nadaraya-Watson smooth of sample weights on x (left), and the resulting semiparametric fit (right).

In the simulation results that follow, prediction will be evaluated at the 0.01%, 1%, 10%, 25%, 50%, 75%, 90%, 99%, and 99.99% quantiles of the distribution of the x_k. Below are 95% confidence bounds applied to the OLS fit and the semiparametric fit seen in Figures 10 and 12, with the nine prediction points marked by triangles.

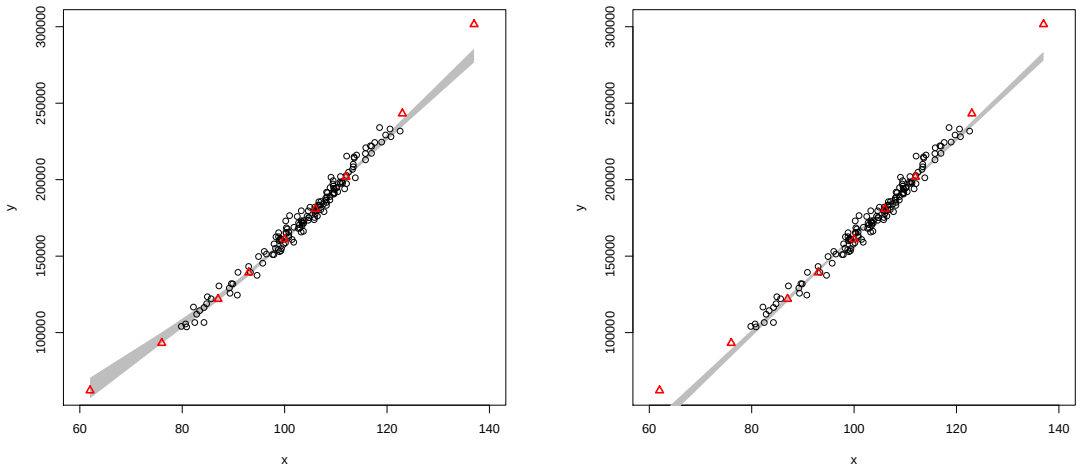

Figure 13: Confidence bounds applied to the semiparametric fit (left) and the linear model fit (right), with prediction goals in red triangles.

The linear model confidence bounds fail to capture all of the nine goals except for the third and seventh points in red; the semiparametric confidence bounds successfully capture the first seven points. It will be seen later that the coverage rates for the highest two points are very low, and this is not surprising as the points are on the extreme end of the x scale and are far beyond the scope of the sample data for this realization.

The process described above is repeated 1000 times and the simulation results are compiled into the following tables. The prediction points, x_p, are chosen to correspond to the 0.0001, 0.01, 0.10, 0.25, 0.50, 0.75, 0.90, 0.99, 0.9999 quantiles of the x distribution. The mean square prediction errors are expressed as a ratio of the predictor to the semiparametric method; thus values larger than one represent worse prediction.

Table 2: Coverage Rates for 95% Confidence Bounds

x_p	Semiparametric Model	OLS
62	0.975	0.000
76	0.876	0.000
87	0.974	0.318
93	0.756	0.851
100	0.693	0.014
106	0.677	0.011
112	0.630	0.813
123	0.114	0.000
137	0.000	0.000

Table 3: Mean Square Prediction Errors

x_p	Semiparametric Model	OLS	WOLS
62	1.00	41.507	32.738
76	1.00	7.091	4.860
87	1.00	4.679	2.241
93	1.00	0.403	0.688
100	1.00	2.735	2.255
106	1.00	4.656	1.978
112	1.00	0.649	1.030
123	1.00	1.456	2.847
137	1.00	1.216	1.690

We can see from Table 3 that the semiparametric model shows significant gains in accuracy at almost all points along the x domain. Furthermore, the confidence intervals have much better coverage under the semiparametric model as seen in Table 2. Some noteworthy items are that both models have very bad coverage at the 0.99 and 0.9999 percentiles, they *never* capture the highest point, but this is much beyond the scope of a typical sample data set, and also, even though the semiparametric model has 11% coverage for the 0.99 percentile, the OLS fit has no coverage, and this is still a difficult point to predict as it is very high in the x range. Rather remarkably, the semiparametric model is able to capture the points on the

very low end of the x range with very good accurace, and coverage rates not very far from 95%, while the OLS fit again has no coverage at these x values. This may be because for large values of x we see small values of the sample weight, and because of this the fit is unable to adjust upward enough to capture those points. Regardless, for points inside the range of the data we see reasonable coverage rates, although there is room for improvement, and overall we see a huge gain in predictive accuracy.

CHAPTER 4

CONCLUSION

Data from surveys can present unique challenges to analysts because of the sampling design. Much of the traditional results seen in statistics rely on the data coming from a random sample, which is almost never true in survey data by design. Due to this complication, reliable testing procedures for informative sampling are very important in practice. The test proposed in this dissertation is widely applicable in practice, to effectively any problem involving a likelihood function, and shows robust behavior and high power compared to competing tests. Under an informative design, there are two promising ways of adjusting the analysis appropriately. Additional predictors may be included in the model that contain the relevant design information, such as in the NHANES example, or we may be able to expand the model with survey weights and integrate the weights out by modeling them as a smooth function of model covariates. The second option is far from ideal. If we are able to include variables that are of scientific interest and remove the design informativeness then our job is made much easier. When this is not possible, our final option for estimation within the model-based paradigm, is to integrate the weights out of the model. If we succeed in this we can significantly reduce the design-induced bias in the model and improve our predictive abilities.

REFERENCES

Horvitz, D. G., and Thompson D. J. (1952). A Generalization of Sampling Without Replacement from a Finite Population. *Journal of the American Statistical Association*. Volume 47, Issue 260, 663-685.

DuMouchel, W. H. and Duncan, G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*. 8, 535-534.

Fuller, W. A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology* 10, 97118.

Fuller, W. A (2009). *Sampling Statistics*. Hoboken, NJ: Wiley & Sons.

Hjort, N., and Pollard, D. (1993). Asymptotics for minimisers of convex processes. *Arxiv preprint* arXiv:1107.3806.

Korn, E. L. and Graubard, B. I. (1999). *Analysis of Health Surveys*. New york: John Wiley & Sons, Inc.

Little, R. J. A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association* 99(466), 546-556.

Nordberg, L (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics* 5, 223-239.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review* 61, 317-337.

Pfeffermann, D. and Sverchkov, M (2003). Fitting generalized linear models under informative sampling. *Analysis of Survey Data* 175-195.a (2003): 175-195.

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya, Series B* 61, 166-186.

Pfeffermann, D. (2011) Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology* 37, 115-136.

Pfeffermann, D., Krieger, A., and Rinott, Y. (1998) Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica* 8, 1087-1114.

Hocking, R. R (2003). *Methods and Applications of Linear Models.* Hoboken, NJ: Wiley & Sons.

Breidt, F.J., and Opsomer, J.D. (2000) Local Polynomial Regression Estimators in Survey Sampling. *The Annals of Statistics* 28, 1026-1053

Chambers, R. L., Steel, D. G., Wang, S., and Welsh, A. H. (2012) *Maximum Likelihood Estimation for Sample Surveys* Boca Raton, FL: Taylor and Francis Group.

Billingsley, P. (1995). *Probability and Measure.* Hoboken, NJ: Wiley & Sons.

Wilson, K. M., Klein, J. D., Blumkin, A. K., Gottlieb, M., and Winickoff, J. P. (2011) Tobacco-Smoking Exposure in Children Who Live in Multiunit Housing. *Pediatrics* 127, 85-92

APPENDICES

Appendix A: Lemmas and Proofs

Lemma 1. *If* $E[\varepsilon \mid \boldsymbol{x}, \boldsymbol{z}] = 0$ *then*

$$Cov\left(\left[\begin{array}{c}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{\gamma}}\end{array}\right], \widehat{\Gamma}_{\pi}(\boldsymbol{x})\right) = 0.$$

Proof of Lemma 1. The condition that $\mathrm{E}[\varepsilon \mid \boldsymbol{x}, \boldsymbol{z}] = 0$ ensures that the estimated regression model coefficents are unbiased given $\boldsymbol{x}$ and $\boldsymbol{z}$, so using the conditional covariance formula and noting that $\widehat{\Gamma}_{\pi}(\boldsymbol{x})$ is constant given $\boldsymbol{Z}$ and $\boldsymbol{I}$ we have

$$\begin{aligned}
\mathrm{Cov}\left(\left[\begin{array}{c}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{\gamma}}\end{array}\right], \widehat{\Gamma}_{\pi}(\boldsymbol{x})\right) &= \mathrm{Cov}\left(\mathrm{E}\left[\left(\begin{array}{c}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{\gamma}}\end{array}\right) \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{I}\right], \mathrm{E}\left[\widehat{\Gamma}_{\pi}(\boldsymbol{x}) \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{I}\right]\right)\\
&\quad + \mathrm{E}\left[\mathrm{Cov}\left(\left[\begin{array}{c}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{\gamma}}\end{array}\right], \widehat{\Gamma}_{\pi}(\boldsymbol{x}) \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{I}\right)\right]\\
&= \mathrm{E}\left\{\mathrm{E}\left[\left(\begin{array}{c}\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\end{array}\right) \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{I}\right]\left[\widehat{\Gamma}_{\pi}(\boldsymbol{x}) - \Gamma(\boldsymbol{x})\right]\right\} + 0\\
&= \mathrm{E}\left\{\mathrm{E}\left[\left(\sum_{k\in U}\left(\begin{array}{c}\boldsymbol{x}_k\\ \boldsymbol{z}_k\end{array}\right)\left(\begin{array}{cc}\boldsymbol{x}_k & \boldsymbol{z}_k\end{array}\right) I_k\right)^{-1}\right.\right.\\
&\quad \left.\left. \times \left(\sum_{k\in U}\left(\begin{array}{c}\boldsymbol{x}_k\\ \boldsymbol{z}_k\end{array}\right)\varepsilon_k I_k\right) \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{I}\right]\left[\widehat{\Gamma}_{\pi}(\boldsymbol{x}) - \Gamma(\boldsymbol{x})\right]\right\} = 0.
\end{aligned}$$

■

Billingsley (1995) treats the following inequality as a trivial fact, but to be thorough here it will be stated as a lemma and proved.

Lemma 2. *For a random variable, X, with finite second moment,*

$$E\left[\min(|tx|^2, |tx|^3)\right] \leq \int_{\{|x|<\delta\}} |tx|^3 dP + \int_{\{|x|\geq\delta\}} |tx|^2 dP. \tag{46}$$

Proof of Lemma 2. In equation (46), equality holds when $\delta = t^{-1}$, since below this point $|tx|^3$ is the minimum, and above this point $|tx|^2$ is the minimum. That is,

$$\mathrm{E}\left[\min(|tx|^2, |tx|^3)\right] = \int_{\{|x|<t^{-1}\}} |tx|^3 dP + \int_{\{|x|\geq t^{-1}\}} |tx|^2 dP.$$

Thus the sum on the right hand side is minimized by splitting the integrals at the point $\delta = t^{-1}$ and the sum can only grow larger for any other choice of δ.

■

Lemma 3. *Under B1–B5, for $r \geq 1$,*

$$\frac{1}{Nh}\sum_{i\in U} K\left(\frac{x-x_i}{h}\right)^r \Gamma(x_i)) = \Gamma(x)f(x)\int K(u)^r du + O\left(h^2 + \frac{1}{N}\right).$$

Proof of Lemma 3. For $r \geq 1$

$$\begin{aligned}
\frac{1}{Nh}\sum_{i\in U} K\left(\frac{x-x_i}{h}\right)^r \Gamma(x_i) &= \frac{1}{h}\int K\left(\frac{x-y}{h}\right)^r \Gamma(y)f(y)dy + O\left(\frac{1}{N}\right) \\
&= \int K(u)^r \Gamma(uh+x)f(uh+x)du + O\left(\frac{1}{N}\right) \\
&= \int K(u)^r \{\Gamma(x)f(x) + [\Gamma'(x)f(x) + \Gamma(x)f'(x)]hu \\
&\quad + \frac{1}{2}[2\Gamma'(x)f'(x) + \Gamma''(x)f(x) + \Gamma(x)f''(x)]h^2u^2 + \ldots\}du \\
&\quad + O\left(\frac{1}{N}\right) \\
&= \Gamma(x)f(x)\int K(u)^r du + O\left(h^2 + \frac{1}{N}\right).
\end{aligned}$$

■

Lemma 4. *Under B1–B4*

$$\frac{1}{Nh}\sum_{i\in U} K\left(\frac{x-x_i}{h}\right) z_i \left(\frac{I_i}{\pi_i} - 1\right) = O_p\left(\frac{1}{\sqrt{Nh}}\right)$$

Proof of Lemma 4. Write

$$\begin{aligned}
&\mathrm{E}\left[\left(\frac{1}{Nh}\sum_{i\in U} K\left(\frac{x-x_i}{h}\right) z_i \left(\frac{I_i}{\pi_i} - 1\right)\right)^2\right] \\
&= \frac{1}{N^2h^2}\sum_{i,j\in U} K\left(\frac{x-x_i}{h}\right) K\left(\frac{x-x_j}{h}\right) z_i z_j \frac{\Delta_{ij}}{\pi_i\pi_j} \\
&= \frac{1}{N^2h^2}\sum_{i\in U} K^2\left(\frac{x-x_i}{h}\right) z_i^2 \frac{1-\pi_i}{\pi_i} + \sum_{i\neq j} K\left(\frac{x-x_i}{h}\right) K\left(\frac{x-x_j}{h}\right) z_i z_j \frac{\Delta_{ij}}{\pi_i\pi_j} \\
&\leq \frac{1}{N^2h^2}\frac{b_z^2}{\lambda}\sum_{i\in U} K^2\left(\frac{x-x_i}{h}\right) + \frac{b_z^2 \max_{i\neq j}|\Delta_{ij}|}{Nh^2\lambda^2}\sum_{i\in U} K^2\left(\frac{x-x_i}{h}\right) = O\left(\frac{1}{Nh}\right).
\end{aligned}$$

Lemma 5. *Under B1–B4,*

$$\begin{aligned}\widehat{\Gamma}_\pi(x) &= \Gamma(x) + \frac{1}{\sum_{j\in U} K\left(\frac{x-x_j}{h}\right)} \sum_{i\in U} \left[z_i - \frac{\sum_{j\in U} K\left(\frac{x-x_j}{h}\right) z_j}{\sum_{j\in U} K\left(\frac{x-x_j}{h}\right)} \right] K\left(\frac{x-x_i}{h}\right) \frac{I_i}{\pi_i} \\ &\quad + O\left(\frac{1}{Nh}\right).\end{aligned}$$

Proof of Lemma 5. Using a Taylor series approximation we have

$$\begin{aligned}\widehat{\Gamma}_\pi(x) = \frac{\sum_{i\in U} K\left(\frac{x-x_i}{h}\right) z_i \frac{I_i}{\pi_i}}{\sum_{i\in U} K\left(\frac{x-x_i}{h}\right) \frac{I_i}{\pi_i}} &\equiv \frac{\widehat{t}_1}{\widehat{t}_2} \\ &= \Gamma(x) + \frac{1}{t_2}(\widehat{t}_1 - t_1) - \frac{t_1}{t_2^2}(\widehat{t}_2 - t_2) + O\left(\frac{1}{Nh}\right) \\ &= \frac{1}{t_2} \sum_{i\in U} \left[z_i - \Gamma(x) \right] K\left(\frac{x-x_i}{h}\right) \frac{I_i}{\pi_i} + O\left(\frac{1}{Nh}\right).\end{aligned}$$

■

Lemma 6. *Under B1–B4,*

$$E\left[\widehat{\Gamma}_{\pi}^{2}\right] = \Gamma^{2} + 2\Gamma E\left[\widehat{\Gamma}_{\pi} - \Gamma\right] + O\left(\frac{1}{Nh}\right).$$

Proof of Lemma 6. Define

$$\boldsymbol{x} = \begin{pmatrix} \frac{1}{Nh}(\widehat{t}_1 - t_1) \\ \frac{1}{Nh}(\widehat{t}_2 - t_2) \end{pmatrix}$$

and

$$f_N(\boldsymbol{x}) = \left(\frac{x_1 + \frac{1}{Nh}t_1}{x_2 + \frac{1}{Nh}t_2}\right)^2.$$

Then

$$\frac{1}{Nh}\mathrm{E}\left[\widehat{t}_a - t_a)^2\right] = O\left(\frac{1}{Nh}\right)$$

for $a = 1, 2$, by Lemma 4. Furthermore, $f_N(\mathbf{0})$ is bounded and continuous and has bounded and continuous first derivatives by Lemma 3, and $f_N(\boldsymbol{x})$ has continuous and bounded first derivatives by Lemma 4. Thus the conditions of Theorem 5.4.3 of Fuller (1996) hold for $\alpha = 1$ and $s = 2$ and the result is immediate.

■

Lemma 7. *Assume B1 - B4, then*

$$E\left[\left(\Gamma_N(x) - \widehat{\Gamma}_N(x)\right)^2\right] = O\left(\frac{1}{Nh}\right).$$

Proof of Lemma 7. From Lemma 6 we can write

$$\begin{aligned}
\mathrm{E}\left[\left(\Gamma(x) - \widehat{\Gamma}_\pi(x)\right)^2\right] &= \mathrm{E}\left[\widehat{\Gamma}_\pi^2 - 2\widehat{\Gamma}_\pi\Gamma + \Gamma^2\right] \\
&= \mathrm{E}\left[\widehat{\Gamma}_\pi^2\right] - 2\Gamma\mathrm{E}\left[\widehat{\Gamma}_\pi\right] + \Gamma^2 \\
&= \Gamma^2 + 2\Gamma\mathrm{E}\left[\widehat{\Gamma}_\pi + O\left(\frac{1}{Nh}\right) - \Gamma\right] - 2\Gamma\mathrm{E}\left[\widehat{\Gamma}_\pi\right] + \Gamma^2 \\
&= O\left(\frac{1}{Nh}\right).
\end{aligned}$$

■

5.1 Appendix B1: Proof of Theorem 1

Proof of Theorem 1. Maximizing $l_a(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is equivalent to minimizing the convex function

$$-\sum_{k\in U} a_k I_k \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta_0} + N^{-1/2}\boldsymbol{u}\right) + \sum_{k\in U} a_k I_k \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta_0}\right), \tag{47}$$

which is minimized at $\boldsymbol{u} = N^{1/2}\left(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_0\right)$. The results then follow from Theorem 2.2 of Hjort and Pollard (1993) and the given assumptions. Expanding the function about $\boldsymbol{u} = 0$, we have

$$\begin{aligned}
&-\sum_{k\in U} a_k I_k \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{u}\right) + \sum_{k\in U} a_k I_k \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0\right) \\
&= -\sum_{k\in U} a_k I_k \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0\right) - \sum_{k\in U} \frac{\partial \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{u}\right)}{\partial \boldsymbol{u}} I_k a_k \Bigg|_{\boldsymbol{u}=0} \frac{\boldsymbol{u}}{\sqrt{N}} \\
&\quad -\frac{\boldsymbol{u}^T}{2\sqrt{N}} \sum_{k\in U} \frac{\partial^2 \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{u}\right)}{\partial \boldsymbol{u}\partial \boldsymbol{u}^T} I_k a_k \Bigg|_{\boldsymbol{u}=0} \frac{\boldsymbol{u}}{\sqrt{N}} \\
&\quad + \sum_{k\in U} a_k I_k \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0\right) + o_p(1).
\end{aligned}$$

Now taking derivatives with respect to $\boldsymbol{u}$ and setting to zero we have

$$\begin{aligned}
0 \;\equiv\; & -\frac{1}{\sqrt{N}} \sum_{k\in U} a_k I_k \frac{\partial \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{u}\right)}{\partial \boldsymbol{u}} \Bigg|_{\boldsymbol{u}=0} \\
& -\frac{1}{N} \sum_{k\in U} a_k I_k \frac{\partial^2 \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{u}\right)}{\partial \boldsymbol{u}\partial \boldsymbol{u}^T} \Bigg|_{\boldsymbol{u}=0} \boldsymbol{u} + o_p(1)
\end{aligned}$$

and thus by assumptions A4 and A5

$$\boldsymbol{u} \to \mathcal{N}(0, \boldsymbol{J}_a^{-1}\boldsymbol{K}_a\boldsymbol{J}_a^{-1})$$

in distribution, and

$$\boldsymbol{u} = \sqrt{N}(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_0) = \boldsymbol{J}_a^{-1} \frac{1}{\sqrt{N}} \sum_{k\in U} a_k I_k D(y_k, \boldsymbol{x}_k; \boldsymbol{\theta}_0) + o_p(1).$$

5.2 Appendix B2: Proof of Theorem 2

Proof of Theorem 2. Write $c = w$ when $a = 1$ and $c = 1$ when $a = w$. Then by the expansion of Theorem 1,

$$
\begin{aligned}
T_a &= 2\Bigg\{ l_a\left(\widehat{\boldsymbol{\theta}}_a\right) - l_a\left(\boldsymbol{\theta}_0\right) - \left(\widehat{\boldsymbol{\theta}}_c - \widehat{\boldsymbol{\theta}}_a\right) \sum_{k\in U} a_k I_k \boldsymbol{D}(y_k, \boldsymbol{x}_k; \boldsymbol{\theta}_0) \\
&\quad - \left(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_0\right) \sum_{k\in U} a_k I_k \boldsymbol{D}(y_k, \boldsymbol{x}_k; \boldsymbol{\theta}_0) \\
&\quad - \frac{1}{2}\left(\widehat{\boldsymbol{\theta}}_c - \widehat{\boldsymbol{\theta}}_a\right)^T \sum_{k\in U} a_k I_k \left.\frac{\partial^2 \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \left(\widehat{\boldsymbol{\theta}}_c - \widehat{\boldsymbol{\theta}}_a\right) \\
&\quad - \left(\widehat{\boldsymbol{\theta}}_c - \widehat{\boldsymbol{\theta}}_a\right)^T \sum_{k\in U} a_k I_k \left.\frac{\partial^2 \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \left(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_0\right) \\
&\quad - \frac{1}{2}\left(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_0\right)^T \sum_{k\in U} a_k I_k \left.\frac{\partial^2 \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \left(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_0\right) \\
&\quad + o_P(1)\Bigg\} \\
&= -2N^{1/2}\left(\widehat{\boldsymbol{\theta}}_c - \widehat{\boldsymbol{\theta}}_a\right) \boldsymbol{J}_a N^{1/2}\left(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_0\right) \\
&\quad - N^{1/2}\left(\widehat{\boldsymbol{\theta}}_c - \widehat{\boldsymbol{\theta}}_a\right)^T \left(\boldsymbol{J}_a + \frac{1}{N}\sum_{k\in U} a_k I_k \left.\frac{\partial^2 \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} - \boldsymbol{J}_a\right) N^{1/2}\left(\widehat{\boldsymbol{\theta}}_c - \widehat{\boldsymbol{\theta}}_a\right) \\
&\quad - 2N^{1/2}\left(\widehat{\boldsymbol{\theta}}_c - \widehat{\boldsymbol{\theta}}_a\right) \left(\boldsymbol{J}_a + \frac{1}{N}\sum_{k\in U} a_k I_k \left.\frac{\partial^2 \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} - \boldsymbol{J}_a\right) N^{1/2}\left(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_0\right) \\
&\quad + o_P(1) \\
&= N^{1/2}\left(\widehat{\boldsymbol{\theta}}_a - \widehat{\boldsymbol{\theta}}_c\right)^T \boldsymbol{J}_a N^{1/2}\left(\widehat{\boldsymbol{\theta}}_a - \widehat{\boldsymbol{\theta}}_c\right) + o_P(1).
\end{aligned}
$$

Under the additional assumption A5, the remaining results (15) are immediate from (13) and the distribution of quadratic forms in asymptotically normal random variables.

■

5.3 Appendix B3: Proof of Theorem 4 and Theorem 5

Proofs of Theorem 4 and Theorem 5. The proofs are identical to those of Theorem 1 and Theorem 2, noting that as $N \to \infty$, $\widehat{\boldsymbol{\theta}}_a$ remains fixed and the vectors $\boldsymbol{X}_N = [\boldsymbol{x}_k^T]_{k \in U}, \boldsymbol{I}_N = [I_k]_{k \in U}$ grow exactly as in the non-bootstrap setting.

■

5.4 Appendix B4: Proof of Theorem 6

Proof of Theorem 6. Let $z > 0$ and $\epsilon > 0$ be given. Then

$$
\begin{aligned}
&\Pr\left[\left|G_{bN}\left(z;\widehat{\boldsymbol{\theta}}_a\right) - L_b\left(z;\boldsymbol{\theta}_0\right)\right| > \epsilon; \boldsymbol{\theta}_0\right] \\
&= \Pr\left[\left|G_{bN}\left(z;\widehat{\boldsymbol{\theta}}_a\right) - L_b\left(z;\boldsymbol{\theta}_0\right)\right| > \epsilon, \|\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_0\| \leq \delta; \boldsymbol{\theta}_0\right] \\
&\quad + \Pr\left[\left|G_{bN}\left(z;\widehat{\boldsymbol{\theta}}_a\right) - L_b\left(z;\boldsymbol{\theta}_0\right)\right| > \epsilon, \|\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_0\| > \delta; \boldsymbol{\theta}_0\right] \\
&\leq \Pr\left[\left|G_{bN}\left(z;\widehat{\boldsymbol{\theta}}_a\right) - L_b\left(z;\widehat{\boldsymbol{\theta}}_a\right)\right| > \epsilon/2, \|\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_0\| \leq \delta; \boldsymbol{\theta}_0\right] \\
&\quad + \Pr\left[\left|L_b\left(z;\widehat{\boldsymbol{\theta}}_a\right) - L_b\left(z;\boldsymbol{\theta}_0\right)\right| > \epsilon/2, \|\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_0\| \leq \delta; \boldsymbol{\theta}_0\right] + \Pr\left[\|\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_0\| > \delta; \boldsymbol{\theta}_0\right] \\
&\leq \Pr\left[\sup_{\boldsymbol{\theta}: \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|} \left|G_{bN}\left(z;\boldsymbol{\theta}\right) - L_b\left(z;\boldsymbol{\theta}\right)\right| > \epsilon/2; \boldsymbol{\theta}_0\right] \\
&\quad + \Pr\left[\left|L_b\left(z;\widehat{\boldsymbol{\theta}}_a\right) - L_b\left(z;\boldsymbol{\theta}_0\right)\right| > \epsilon/2; \boldsymbol{\theta}_0\right] + \Pr\left[\|\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_0\| > \delta; \boldsymbol{\theta}_0\right].
\end{aligned}
$$

By hypothesis, the first probability is zero for all N sufficiently large. The third term goes to zero as $N \to \infty$ because $\widehat{\boldsymbol{\theta}}_a \xrightarrow{P} \boldsymbol{\theta}_0$. The second term goes to zero by consistency of $\widehat{\boldsymbol{\theta}}_a$ and continuity of $L_b(z;\boldsymbol{\theta})$ in $\boldsymbol{\theta}$, since the probabilities depend only on the eigenvalues, which are continuous functions of $\boldsymbol{\theta}$.

5.5 Appendix B5: Proof of Theorem 7

Proof of Theorem 7. Under the sequence of alternatives defined by

$$H_{1n} : \boldsymbol{\theta}_s = \boldsymbol{\theta}_0 + \frac{\boldsymbol{d}}{\sqrt{N}}$$

the results in Theorem 1 will be identical for the "w" subscripts; for the "1" subscript, the Hjort and Pollard argument is expanded to allow the sequence of alternatives to converge to the null:

$$\begin{aligned}
&-\sum_{k\in U} I_k \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{u} + N^{-1/2}\boldsymbol{d}\right) + \sum_{k\in U} I_k \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{d}\right) \\
&= -\sum_{k\in U} I_k \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{d}\right) \\
&\quad -\sum_{k\in U} \frac{\partial \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{u} + N^{-1/2}\boldsymbol{d}\right)}{\partial \boldsymbol{u}} I_k \Big|_{\boldsymbol{u}=0} \frac{\boldsymbol{u}}{\sqrt{N}} \\
&\quad -\frac{\boldsymbol{u}^T}{2\sqrt{N}} \sum_{k\in U} \frac{\partial^2 \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{u} + N^{-1/2}\boldsymbol{d}\right)}{\partial \boldsymbol{u}\partial \boldsymbol{u}^T} I_k \Big|_{\boldsymbol{u}=0} \frac{\boldsymbol{u}}{\sqrt{N}} \\
&\quad +\sum_{k\in U} I_k \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{d}\right) + o_p(N^{-1}) \\
&= -\sum_{k\in U} \frac{\partial \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{u} + N^{-1/2}\boldsymbol{d}\right)}{\partial \boldsymbol{u}} I_k \Big|_{\boldsymbol{u}=0} \frac{\boldsymbol{u}}{\sqrt{N}} \\
&\quad -\frac{\boldsymbol{u}^T}{2\sqrt{N}} \sum_{k\in U} \frac{\partial^2 \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{u} + N^{-1/2}\boldsymbol{d}\right)}{\partial \boldsymbol{u}\partial \boldsymbol{u}^T} I_k \Big|_{\boldsymbol{u}=0} \frac{\boldsymbol{u}}{\sqrt{N}} + o_p(N^{-1}).
\end{aligned}$$

Now set the derivative with respect to $\boldsymbol{u}$ to zero, and expand both terms about $\boldsymbol{d} = 0$ to get

$$\begin{aligned}
0 \equiv & -\frac{1}{\sqrt{N}} \left\{ \sum_{k \in U} \frac{\partial \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{u}\right)}{\partial \boldsymbol{u}} I_k \Bigg|_{\boldsymbol{u}=0} \right. \\
& \left. + \sum_{k \in U} I_k \frac{\partial}{\partial \boldsymbol{d}} \frac{\partial \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{u} + N^{-1/2}\boldsymbol{d}\right)}{\partial \boldsymbol{u}} \Bigg|_{(\ \boldsymbol{u} \quad \boldsymbol{d}\ )=\mathbf{0}} \frac{\boldsymbol{d}}{\sqrt{N}} \right\} \\
& - \frac{1}{N} \sum_{k \in U} I_k \frac{\partial^2 \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{u}\right)}{\partial \boldsymbol{u} \partial \boldsymbol{u}^T} \Bigg|_{\boldsymbol{u}=0} \boldsymbol{u} \\
& - \frac{\partial}{\partial \boldsymbol{d}} \left\{ \sum_{k \in U} I_k \frac{\partial^2 \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{u} + N^{-1/2}\boldsymbol{d}\right)}{\partial \boldsymbol{u} \partial \boldsymbol{u}^T} \right\} \Bigg|_{\boldsymbol{d}=0} \frac{\boldsymbol{d}}{\sqrt{N}} + o_p(1) \\
= & -\frac{1}{\sqrt{N}} \sum_{k \in U} I_k D(y_k, \boldsymbol{x}_k; \boldsymbol{\theta}_0) + \boldsymbol{J}_1 \boldsymbol{d} + \boldsymbol{J}_1 \boldsymbol{u} + o_p(1),
\end{aligned}$$

so

$$\boldsymbol{u} = \boldsymbol{J}_1^{-1} \left(\frac{1}{\sqrt{N}} \sum_{k \in U} I_k D(y_k, \boldsymbol{x}_k; \boldsymbol{\theta}_0) - \boldsymbol{J}_1 \boldsymbol{d} \right) + o_p(1),$$

Thus, since $\boldsymbol{u} = \sqrt{N}(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_s) = \sqrt{N}(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0) - \boldsymbol{d}$, the linearization from equation (12) becomes

$$N^{1/2}(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0) = \boldsymbol{J}_1^{-1} \frac{1}{\sqrt{N}} \sum_{k \in U} I_k D(y_k, \boldsymbol{x}_k; \boldsymbol{\theta}_0) + o_p(1).$$

The test statistics can be expanded in the same way as under the null hypothesis of non-informative selection, so

$$T_a = N^{1/2}(\widehat{\boldsymbol{\theta}}_c - \widehat{\boldsymbol{\theta}}_a)^T J_a N^{1/2}(\widehat{\boldsymbol{\theta}}_c - \widehat{\boldsymbol{\theta}}_a) + o_p(1),$$

where $c = w$ when $a = 1$ and $c = 1$ when $a = w$ as before. Under the alternative we have that $N^{1/2}(\widehat{\boldsymbol{\theta}}_c - \widehat{\boldsymbol{\theta}}_a)$ is a non-zero mean, normally distributed random variable. Specifically it is mean $\boldsymbol{d}$, and so using results for quadratic forms and positive definite matrices along with limiting distributions of quadratic forms of non-central asymptotically normal random variables (e.g. Hocking (2003)) we can establish the limiting distribution here. For notational convenience, let $T_a = \boldsymbol{q}^T \boldsymbol{J}_a \boldsymbol{q}$ in what follows. We know that $\boldsymbol{q}$ is asymptotically normal with

mean $\boldsymbol{d}$ and covariance $\boldsymbol{\Gamma}$. Next write

$$\boldsymbol{q}^T \boldsymbol{J}_a \boldsymbol{q} = \boldsymbol{q}^{*T} \boldsymbol{\Gamma}^{T/2} \boldsymbol{J}_a \boldsymbol{\Gamma}^{1/2} \boldsymbol{q}^*,$$

where $\boldsymbol{q}^* \sim \mathcal{AN}(\boldsymbol{\Gamma}^{-1/2}\boldsymbol{d}, \boldsymbol{I})$. Then

$$\boldsymbol{q}^{*T} \boldsymbol{\Gamma}^{T/2} \boldsymbol{J}_a \boldsymbol{\Gamma}^{1/2} \boldsymbol{q}^* = \boldsymbol{q}^{**T} \boldsymbol{P}^T \boldsymbol{\Gamma}^{T/2} \boldsymbol{J}_a \boldsymbol{\Gamma}^{1/2} \boldsymbol{P} \boldsymbol{q}^{**},$$

where $\boldsymbol{P}$ is an orthogonal matrix of eigenvectors of $\boldsymbol{\Gamma}^{T/2} \boldsymbol{J}_a \boldsymbol{\Gamma}^{1/2}$, and $\boldsymbol{q}^{**} \sim \mathcal{AN}(\boldsymbol{P}\boldsymbol{\Gamma}^{-1/2}\boldsymbol{d}, \boldsymbol{I})$. Finally, letting $\boldsymbol{\Lambda}$ be a diagonal matrix of eigenvalues of $\boldsymbol{\Gamma}^{T/2} \boldsymbol{J}_a \boldsymbol{\Gamma}^{1/2}$, we have

$$\boldsymbol{q}^T \boldsymbol{J}_a \boldsymbol{q} = \boldsymbol{q}^{**T} \boldsymbol{\Lambda} \boldsymbol{q}^{**},$$

and by definition 16.2 of Hocking (2003), and recalling that $\boldsymbol{P}$ is orthogonal, we have that

$$T_a \xrightarrow{\mathcal{L}} \sum_{j=1}^{p} \lambda_{aj} \chi^2(1; \delta_j),$$

where λ_{aj} are the eigenvalues of $\boldsymbol{\Gamma}^{T/2} \boldsymbol{J}_a \boldsymbol{\Gamma}^{1/2}$, and $\delta_j = [\boldsymbol{P}\boldsymbol{\Gamma}^{-1/2}\boldsymbol{d}]_j$, the j^{th} element of $\boldsymbol{P}\boldsymbol{\Gamma}^{-1/2}\boldsymbol{d}$.

■

5.6 Appendix B6: Proof of Theorem 8

Proof of Theorem 8. The information in the unextended model is

$$\mathcal{I}(\boldsymbol{X}; \boldsymbol{\beta}, \boldsymbol{\xi}) = \begin{bmatrix} -\frac{1}{N}\sum_{k \in U} \frac{\partial^2 \ln l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} & -\frac{1}{N}\sum_{k \in U} \frac{\partial^2 \ln l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\xi}} \\ -\frac{1}{N}\sum_{k \in U} \frac{\partial^2 \ln l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial \boldsymbol{\xi} \partial \boldsymbol{\beta}} & -\frac{1}{N}\sum_{k \in U} \frac{\partial^2 \ln l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^T} \end{bmatrix} = \boldsymbol{J}_1.$$

Under the assumption that the derivatives in the extended model with respect to $\boldsymbol{\gamma}$ correspond to weighted versions of the derivatives with respect to $\boldsymbol{\beta}$, (A7), we can write the

information in the extended model as

$$\begin{aligned}
&\mathcal{I}(\boldsymbol{X}, \boldsymbol{W}\boldsymbol{X}; \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=0} \\
&= -\frac{1}{N}\begin{bmatrix}
\sum_{k\in U} \frac{\partial^2 \ln l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} & \sum_{k\in U} \frac{\partial^2 \ln l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\xi}} & \sum_{k\in U} w_k \frac{\partial^2 \ln l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \\
\sum_{k\in U} \frac{\partial^2 \ln l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial\boldsymbol{\xi}\partial\boldsymbol{\beta}} & \sum_{k\in U} \frac{\partial^2 \ln l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^T} & \sum_{k\in U} w_k \frac{\partial^2 \ln l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial\boldsymbol{\xi}\partial\boldsymbol{\beta}^T} \\
\sum_{k\in U} w_k \frac{\partial^2 \ln l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} & \sum_{k\in U} w_k \frac{\partial^2 \ln l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\xi}^T} & \sum_{k\in U} w_k^2 \frac{\partial^2 \ln l(y_k|\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}
\end{bmatrix}.
\end{aligned}$$

Letting $a_r = -\frac{1}{N}\sum_{k\in U} r_k \frac{\partial^2 \ln l(\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})Q}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$, $b_r = -\frac{1}{N}\sum_{k\in U} r_k \frac{\partial^2 \ln l(\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\xi}}$, and $d_r = -\frac{1}{N}\sum_{k\in U} r_k \frac{\partial^2 \ln l(\boldsymbol{x}_k;\boldsymbol{\beta},\boldsymbol{\xi})}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^T}$, we can write the extended model information as

$$\mathcal{I}(\boldsymbol{X}, \boldsymbol{W}\boldsymbol{X}; \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=0} = \begin{bmatrix} a_1 & b_1 & a_w \\ b_1^T & d_1 & b_w^T \\ a_w & b_w & a_{w^2} \end{bmatrix}.$$

To get the limiting variance of $\sqrt{N}(\hat{\boldsymbol{\gamma}} - 0)$ we will need the $(3,3)$ element of $\mathcal{I}^{-1}(\boldsymbol{X}, \boldsymbol{W}\boldsymbol{X}; \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=0}$. Denote the $(3,3)$ element of the inverse by $\boldsymbol{M}_4$. To obtain this inverse we will treat the matrix as a 2×2 block matrix and use the standard block matrix inverse result. Recalling that the top left 2×2 block is $\boldsymbol{J}_1$, we have

$$\boldsymbol{M}_4 = \left\{ a_{w^2} - \begin{bmatrix} a_w & b_w \end{bmatrix} \boldsymbol{J}_1^{-1} \begin{bmatrix} a_w \\ b_w^T \end{bmatrix} \right\}^{-1},$$

where

$$\boldsymbol{J}_1^{-1} = \begin{bmatrix} \{a_1 - b_1 d_1^{-1} b_1^T\}^{-1} & -\{a_1 - b_1 d_1^{-1} b_1^T\}^{-1} b_1 d_1^{-1} \\ -d_1^{-1} b_1^T \{a_1 - b_1 d_1^{-1} b_1^T\}^{-1} & \{d_1 - b_1 a_1^{-1} b_1^T\}^{-1} \end{bmatrix}.$$

So,

$$\begin{aligned}
\boldsymbol{M}_4 &= \{a_{w^2} - a_w(a_1 - b_1 d_1^{-1} b_1^T)^{-1} a_w - b_w d_1^{-1} b_1^T (a_1 - b_1 d_1^{-1} b_1^T)^{-1} a_w \\
&- a_w(a_1 - b_1 d_1^{-1} b_1^T)^{-1} b_1 d_1^{-1} b_w^T + b_w(d_1 - b_1^T a_1^{-1} b_1)^{-1} b_w^T)\}^{-1} \\
&= \{a_{w^2} - a_w V_{11}^{-1} a_w - 2 b_w d_1^{-1} b_1^T V_{11}^{-1} a_w + b_w V_{21}^{-1} b_w^T\}^{-1}
\end{aligned}$$

Now the limiting variance of $\sqrt{N}(\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_1)$ is needed. To obtain this we will need the $(1,1)$ element of $\boldsymbol{J}_w^{-1}\boldsymbol{K}_w\boldsymbol{J}_w^{-1} - \boldsymbol{J}_1^{-1}$ computed from the original model. This results in

$$\begin{aligned}
\{\boldsymbol{J}_w^{-1}\boldsymbol{K}_w\boldsymbol{J}_w^{-1} - \boldsymbol{J}_1^{-1}\}_{[1,1]} &= \{a_w - b_w d_w^{-1} b_w^T\}^{-1} a_{w^2} \{a_w - b_w d_w^{-1} b_w^T\}^{-1} \\
&\quad - \{a_w - b_w d_w^{-1} b_w^T\}^{-1} b_w d_w^{-1} b_{w^2}^T \{a_w - b_w d_w^{-1} b_w^T\}^{-1} \\
&\quad - \{a_w - b_w d_w^{-1} b_w^T\}^{-1} b_{w^2}^T d_w^{-1} b_w^T \{a_w - b_w d_w^{-1} b_w^T\}^{-1} \\
&\quad + \{a_w - b_w d_w^{-1} b_w^T\}^{-1} b_w d_w^{-1} d_{w^2} d_w^{-1} b_w^T \{a_w - b_w d_w^{-1} b_w^T\}^{-1} \\
&\quad - \{a_1 - b_1 d_1^{-1} b_1^T\}^{-1} \\
&= \{a_w - b_w d_w^{-1} b_w^T\}^{-1} [a_{w^2} - 2 b_w d_w^{-1} b_{w^2}^T + b_w d_w^{-1} d_{w^2} d_w^{-1} b_w^T] \\
&\quad \times \{a_w - b_w d_w^{-1} b_w^T\}^{-1} - \{a_1 - b_1 d_1^{-1} b_1^T\}^{-1} \equiv \boldsymbol{\Gamma}_{11}.
\end{aligned}$$

It remains to find a 1-to-1 relationship between the limiting distributions of $\sqrt{N}(\hat{\boldsymbol{\gamma}}_1 - 0)$ and $\sqrt{N}(\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_1)$, which are

$$\sqrt{N}(\hat{\boldsymbol{\gamma}}_1 - 0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{M}_4)$$

and

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_1) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_{11}).$$

In general, there is no obvious or simple transformation linking these two distributions. However, since $\boldsymbol{M}_4$ and $\boldsymbol{\Gamma}_{11}$ are symmetric and positive definite, in general, the transformation $\boldsymbol{M}_4^{1/2}\boldsymbol{\Gamma}_{11}^{-1/2}$ can be used since

$$\lim_{N\to\infty} \text{Var}\left(\boldsymbol{M}_4^{1/2}\boldsymbol{\Gamma}_{11}^{-1/2}\sqrt{N}(\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_1)\right) = \boldsymbol{M}_4^{T/2}\boldsymbol{\Gamma}_{11}^{-T/2}\boldsymbol{\Gamma}_{11}\boldsymbol{\Gamma}_{11}^{-1/2}\boldsymbol{M}_4^{1/2} = \boldsymbol{M}_4$$

■

5.7 Appendix B7: Proof of Corollary 9

Proof of Corollary 9. If the maximum likelihood estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ are asymptotically uncorrelated then the expected double partial derivatives with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ are zero, thus $b_r = \mathbf{0}$. The limiting variances from Theorem 8 become

$$\boldsymbol{M}_4 = \{a_{w^2} - a_w a_1^{-1} a_w\}^{-1}, \tag{48}$$

and

$$\boldsymbol{\Gamma}_{11} = a_w^{-1} a_{w^2} a_w^{-1} - a_1^{-1}. \tag{49}$$

The one-to-one transformation linking these distributions is straightforward in this case. The transformation is $\{a_{w^2} - a_w a_1^{-1} a_w\}^{-1} a_w$ since

$$\begin{aligned}
&\lim_{N\to\infty} \mathrm{Var}\left(\{a_{w^2} - a_w a_1^{-1} a_w\}^{-1} a_w (\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_1)\right) \\
&= \{a_{w^2} - a_w a_1^{-1} a_w\}^{-1} a_w \{a_w^{-1} a_{w^2} a_w^{-1} - a_1^{-1}\} a_w \{a_{w^2} - a_w a_1^{-1} a_w\}^{-1} \\
&= \{a_{w^2} - a_w a_1^{-1} a_w\}^{-1} \{a_{w^2} - a_w a_1^{-1} a_w\} \{a_{w^2} - a_w a_1^{-1} a_w\}^{-1} \\
&= \{a_{w^2} - a_w a_1^{-1} a_w\}^{-1}.
\end{aligned}$$

■

5.8 Appendix B8: Proof of Theorem 10

Proof of Theorem 10. It follows from the Cramér-Wold Device that if $\lambda_1(\widehat{\Gamma}_\pi(\boldsymbol{x}) - \Gamma(\boldsymbol{x})) + \boldsymbol{\lambda}_2^T \begin{pmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}, & \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \end{pmatrix}^T$ converges in distribution to a univariate normal distribution for every λ_1, $\boldsymbol{\lambda}_2$ such that at least one is non-zero, then the desired joint asymptotic normality holds.

To establish joint normality, we will show that the characteristic function for

$$\frac{1}{\sqrt{N}} \left[\sum_{i\in U} \lambda_1 \frac{1}{t_2} \left[z_i - \frac{t_1}{t_2} \right] K\left(\frac{x - x_i}{h}\right) \left(\frac{I_i}{\pi_i} - 1\right) + \sum_{i\in U} \boldsymbol{\lambda}_2^T \begin{pmatrix} \boldsymbol{x}_i \\ \boldsymbol{z}_i \end{pmatrix} \varepsilon_i I_i \right]$$

converges to the characteristic function of a normally distributed random variable. The first

term in the sum above comes from the linearization of $\widehat{\Gamma}_\pi(x)$ from Lemma 5

Denoting the characteristic function for the above expression with the subscript †, and letting $\lambda_1 t_2^{-1}\left[z_i - t_2^{-1}t_1\right] K\left(\frac{x-x_i}{h}\right) = g_{1i}$ and $\boldsymbol{\lambda}_2^T \begin{pmatrix} \boldsymbol{x}_i, & \boldsymbol{z}_i \end{pmatrix}^T = g_{2i}$, we have

$$\begin{aligned}
\varphi_\dagger(t) &= \mathrm{E}\left[\exp\left\{\frac{it}{\sqrt{N}}\sum_{k\in U} g_{1k}\left(\frac{I_k}{\pi_k}-1\right)\right\}\exp\left\{\frac{it}{\sqrt{N}}\sum_{k\in U} g_{2k}\varepsilon_k I_k\right\}\right] \\
&= \mathrm{E}\left[\exp\left\{\frac{it}{\sqrt{N}}\sum_{k\in U} g_{1k}\left(\frac{I_k}{\pi_k}-1\right)\right\}\mathrm{E}\left[\exp\left\{\frac{it}{\sqrt{N}}\sum_{k\in U} g_{2k}\varepsilon_k I_k\right\}\middle| \boldsymbol{x}_k, \boldsymbol{z}_k, I_k\right]\right] \\
&\equiv \mathrm{E}\left[A\mathrm{E}\left[B|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{I}\right]\right] = \mathrm{E}\left[A\varphi_N^B(t;\boldsymbol{I}_N)\right] = \mathrm{E}\left[A\{\varphi_N^B(t;\boldsymbol{I}_N)-\varphi_B(t)+\varphi_B(t)\}\right] \\
&= \mathrm{E}\left[A\{\varphi_N^B(t;\boldsymbol{I}_N)-\varphi_B(t)\}\right] + EA\varphi_B(t).
\end{aligned}$$

Here, φ_B denotes a normal characteristic function. It remains to prove that $\varphi_N^B(t;\boldsymbol{I}_N) \to \varphi_B(t)$, and then application of the dominated convergence theorem will yield the desired result.

The following argument closely follows those in section 27 of Billingsley (1995) on asymptotic normality of sums of independent but not identically distributed random variables. Per the Billingsley arguments, the following three inequalities will be needed:
(i) For complex $\{z_n\}$ and $\{w_n\}$ of modulus at most 1

$$\left|\prod_{i=1}^n z_i - \prod_{i=1}^n w_i\right| \le \sum_{i=1}^n |z_i - w_i|$$

(ii) $|e^{ix} - (1 + ix - \frac{1}{2}x^2)| \le \min\{|x|^2, \frac{1}{6}|x|^3\}$
(iii) For real x such that $|x| \le 1/2$,

$$e^x - 1 - x \le x^2$$

The following Lindeberg condition is also needed:

$$\frac{1}{\sum_{k\in U} N^{-1}g_{2k}^2\sigma^2 I_k}\sum_{k\in U}\mathrm{E}\left[(N^{-1/2}g_{2k}\varepsilon_k I_k)^2 1_{\{|N^{-1/2}g_{2k}\varepsilon_k I_k| > t\sqrt{\sum N^{-1}g_{2k}^2\sigma^2 I_k}\}}|\boldsymbol{x}_k, \boldsymbol{z}_k, I_k\right] \to 0$$

as $n, N \to \infty$, for all $t > 0$.

Now, $\varphi_N^B(t; \boldsymbol{I}_N) = \prod_{k \in U} \varphi_\varepsilon(tN^{-1/2} g_{2k} I_k) = \prod_{k \in U} \mathrm{E}\left[\exp\{itN^{-1/2} g_{2k} \varepsilon_k I_k | \boldsymbol{x}_k, \boldsymbol{z}_k, I_k\}\right]$, and the goal is to show that this product of characteristic functions converges to a normal characteristic function. For each k,

$$\varphi_\varepsilon(tN^{-1/2} g_{2k} I_k) = 1 + it\mathrm{E}\left[N^{-1/2} g_{2k} \varepsilon_k I_k | \boldsymbol{x}_k, \boldsymbol{z}_k, I_k\right] - \frac{t^2}{2}\mathrm{E}\left[N^{-1} g_{2k}^2 \varepsilon_k^2 I_k | \boldsymbol{x}_k, \boldsymbol{z}_k, I_k\right] + o(t^2),$$

and by inequality (ii),

$$\begin{aligned} &\left|\exp\{itN^{-1/2} g_{2k} \varepsilon_k I_k\} - \left(1 + itN^{-1/2} g_{2k} \varepsilon_k I_k - \frac{t^2}{2} N^{-1} g_{2k}^2 \varepsilon_k^2 I_k\right)\right| \\ &\quad \leq \quad \min(|tN^{-1/2} g_{2k} \varepsilon_k I_k|^2, |tN^{-1/2} g_{2k} \varepsilon_k I_k|^3). \end{aligned}$$

Thus, by dominated convergence and Lemma 2, we have for $\delta > 0$

$$\begin{aligned} &\left|\varphi_\varepsilon(tN^{-1/2} g_{2k} I_k) - \left(1 - \frac{t^2}{2} N^{-1} g_{2k}^2 \sigma^2 I_k\right)\right| \\ &\quad \leq \quad \mathrm{E}\left[\min(|tN^{-1/2} g_{2k} \varepsilon_k I_k|^2, |tN^{-1/2} g_{2k} \varepsilon_k I_k|^3) | \boldsymbol{x}_k, \boldsymbol{z}_k, I_k\right] \\ &\quad \leq \quad \mathrm{E}\left[|N^{-1/2} t g_{2k} \varepsilon_k I_k|^3 1_{\{|\varepsilon_k I_k| < \delta\}} | \boldsymbol{x}_k, \boldsymbol{z}_k, I_k\right] \\ &\qquad + \mathrm{E}\left[|N^{-1/2} t g_{2k} \varepsilon_k I_k|^2 1_{\{|\varepsilon_k I_k| \geq \delta\}} | \boldsymbol{x}_k, \boldsymbol{z}_k, I_k\right] \\ &\quad \leq \quad \delta |t|^3 N^{-1} g_{2k}^2 \sigma^2 I_k + t^2 \mathrm{E}\left[N^{-1} t g_{2k}^2 \varepsilon_k^2 I_k^2 1_{\{|\varepsilon_k I_k| \geq \delta\}} | \boldsymbol{x}_k, \boldsymbol{z}_k, I_k\right]. \end{aligned}$$

It follows by our assumed Lindeberg condition that

$$\begin{aligned} &\sum_{k \in U} \left|\varphi_\varepsilon(tN^{-1/2} g_{2k} I_k) - \left(1 - \frac{t^2}{2} N^{-1} g_{2k}^2 \sigma_k^2 I_k\right)\right| \\ &\quad \leq \quad \sum_{k \in U} \left\{\delta |t|^3 N^{-1} g_{2k}^2 \sigma^2 I_k + t^2 \mathrm{E}\left[N^{-1} t g_{2k}^2 \varepsilon_k^2 I_k^2 1_{\{|\varepsilon_k I_k| \geq \delta\}}\right]\right\} \to 0. \end{aligned}$$

This fact, combined with inequality (i) gives us the following relationship:

$$\begin{aligned} &\left|\prod_{k \in U} \varphi_\varepsilon(tN^{-1/2} g_{2k} I_k) - \prod_{k \in U} \left(1 - \frac{t^2}{2} N^{-1} g_{2k}^2 \sigma_k^2 I_k\right)\right| \\ &\quad \leq \quad \sum_{k \in U} \left|\varphi_\varepsilon(tN^{-1/2} g_{2k} I_k) - \left(1 - \frac{t^2}{2} N^{-1} g_{2k}^2 \sigma_k^2 I_k\right)\right| = o(1). \end{aligned} \tag{50}$$

For N sufficiently large, inequalities (i) and (iii) along with B5 also imply that

$$\begin{aligned}
&\left|\prod_{k\in U}\exp\left\{-\frac{t^2}{2}N^{-1}g_{2k}^2\sigma^2 I_k\right\}-\prod_{k\in U}\left(1-\frac{t^2}{2}N^{-1}g_{2k}^2\sigma_k^2 I_k\right)\right| \\
&\quad\leq\ \sum_{k\in U}\left|\exp\left\{-\frac{t^2}{2}N^{-1}g_{2k}^2\sigma^2 I_k\right\}-\left(1-\frac{t^2}{2}N^{-1}g_{2k}^2\sigma_k^2 I_k\right)\right| \\
&\quad\leq\ \sum_{k\in U}\frac{t^4}{4}N^{-2}g_{2k}^4\sigma^4 I_k\leq\frac{t^4\sigma^4}{4}\frac{\sum_{k\in U}g_{2k}^4}{N^2}=o(1). \qquad (51)
\end{aligned}$$

Combining (50) and (51) we have

$$\begin{aligned}
\prod_{k\in U}\varphi_\varepsilon(tN^{-1/2}g_{2k}I_k) &= \prod_{k\in U}\left(1-\frac{t^2}{2}N^{-1}g_{2k}^2\sigma_k^2 I_k\right)+o(1) \\
&= \prod_{k\in U}\exp\left\{-\frac{t^2}{2}N^{-1}g_{2k}^2\sigma^2 I_k\right\}+o(1) \\
&\Rightarrow\ \prod_{k\in U}\varphi_\varepsilon(tN^{-1/2}g_{2k}I_k)\to\exp\left\{-\frac{t^2}{2}\lim_{N\to\infty}N^{-1}\sum_{k\in U}g_{2k}^2\sigma^2\pi_k\right\},
\end{aligned}$$

as $N\to\infty$. This establishes the convergence of $\varphi_N^B(t;\boldsymbol{I}_N)$ to $\varphi_B(t)$, and so by dominated convergence and assumption B5, we have that

$$\mathrm{E}\left[A\{\varphi_N^B(t;\boldsymbol{I}_N)-\varphi_B(t)\}\right]+\mathrm{E}\left[A\varphi_B(t)\right]\to\varphi_A(t)\varphi_B(t),$$

where $\varphi_A(t)$ is the characteristic function of a normal random variable. This follows because $|A\{\varphi_N^B(t;\boldsymbol{I}_N)-\varphi_B(t)\}|\leq|A||\{\varphi_N^B(t;\boldsymbol{I}_N)-\varphi_B(t)\}|\leq|\{\varphi_N^B(t;\boldsymbol{I}_N)-\varphi_B(t)\}|\leq 2$ implies that $\mathrm{E}\left[A\{\varphi_N^B(t;\boldsymbol{I}_N)-\varphi_B(t)\}\right]\to 0$, and assumption B5 implies that $\mathrm{E}\left[A\right]\varphi_B(t)\to\varphi_A(t)\varphi_B(t)$. Thus since $\varphi_A(t)\varphi_B(t)$ is the product of two normal characteristic functions, the result is proved.

■

5.9 Appendix B9: Proof of Theorem 11

Proof of Theorem 11. Since the leading term

$$\frac{1}{\frac{1}{N^2h^2}\left(\sum_{j\in U} K\left(\frac{x-x_i}{h}\right)\frac{I_j}{\pi_j}\right)^2} \to \frac{1}{\frac{1}{N^2h^2}\left(\sum_{j\in U} K\left(\frac{x-x_i}{h}\right)\right)^2},$$

as $N \to \infty$, we focus on the remaining sum, which is

$$\frac{1}{N^2h^2}\sum_{i,j\in U}\left[y_i - \widehat{\Gamma}(x)\right] K\left(\frac{x-x_i}{h}\right)\left[y_j - \widehat{\Gamma}(x)\right] K\left(\frac{x-x_j}{h}\right)\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\frac{I_iI_j}{\pi_{ij}},$$

and show that this converges to its target. For compactness of notation, let $K\left(\frac{x-x_i}{h}\right) = K_i$ in what follows. Write

$$\begin{aligned}
&\frac{n}{N^2h^2}\mathrm{E}\left|\sum_{i,j\in U}\left[z_i - \widehat{\Gamma}(x)\right] K_i\left[z_j - \widehat{\Gamma}(x)\right] K_j\frac{\Delta_{ij}}{\pi_i\pi_j}\frac{I_iI_j}{\pi_{ij}}\right.\\
&\qquad\left. - \sum_{i,j\in U}\left[z_i - \Gamma(x)\right] K_i\left[z_j - \Gamma(x)\right] K_j\frac{\Delta_{ij}}{\pi_i\pi_j}\right|\\
=\;& \frac{n}{N^2h^2}\mathrm{E}\left|\sum_{i,j\in U}\left[z_i - \widehat{\Gamma}(x)\right] K_i\left[z_j - \widehat{\Gamma}(x)\right] K_j\frac{\Delta_{ij}}{\pi_i\pi_j}\frac{I_iI_j}{\pi_{ij}}\right.\\
&\qquad + \sum_{i,j\in U}\left[z_i - \Gamma(x)\right] K_i\left[z_j - \Gamma(x)\right] K_j\frac{\Delta_{ij}}{\pi_i\pi_j}\left(\frac{I_iI_j-\pi_{ij}}{\pi_{ij}}\right)\\
&\qquad\left. - \sum_{i,j\in U}\left[z_i - \Gamma(x)\right] K_i\left[z_j - \Gamma(x)\right] K_j\frac{\Delta_{ij}}{\pi_i\pi_j}\frac{I_iI_j}{\pi_{ij}}\right|\\
=\;& \frac{n}{N^2h^2}\mathrm{E}\left|\sum_{i,j\in U}\left\{\left[z_i - \widehat{\Gamma}(x)\right] K_i\left[z_j - \widehat{\Gamma}(x)\right] K_j - \left[z_i - \Gamma(x)\right] K_i\left[z_j - \Gamma(x)\right] K_j\right\}\right.\\
&\qquad\times\frac{\Delta_{ij}}{\pi_i\pi_j}\frac{I_iI_j}{\pi_{ij}}\\
&\qquad\left. - \sum_{i,j\in U}\left[z_i - \Gamma(x)\right] K_i\left[z_j - \Gamma(x)\right] K_j\frac{\Delta_{ij}}{\pi_i\pi_j}\left(\frac{I_iI_j-\pi_{ij}}{\pi_{ij}}\right)\right|\\
=\;& \frac{n}{N^2h^2}\mathrm{E}\left|A_N - B_N\right| \le \frac{n}{N^2h^2}\mathrm{E}\left|A_N\right| + \frac{n}{N^2h^2}\mathrm{E}\left|B_N\right|.
\end{aligned}$$

Now

$$\begin{aligned}
\frac{n}{N^2h^2}\mathrm{E}\left|A_N\right| &= \frac{n}{N^2h^2}\mathrm{E}\left|\sum_{i,j\in U}\left\{2\left[z_i-\Gamma(x)\right]K_i\left[\Gamma(x)-\widehat{\Gamma}(x)\right]K_j\right.\right. \\
&\quad \left.\left.+\left[\Gamma(x)-\widehat{\Gamma}(x)\right]K_i\left[\Gamma(x)-\widehat{\Gamma}(x)\right]K_j\right\}\frac{\Delta_{ij}}{\pi_i\pi_j}\frac{I_iI_j}{\pi_{ij}}\right| \\
&\leq \frac{1}{h}\left(\frac{2n\max_{i\neq j}\left|\Delta_{ij}\right|}{\lambda^2\lambda^*}+\frac{2n}{\lambda^2N}\right) \\
&\quad \times\left\{\frac{\sum_{i\in U}(z_i-\Gamma(x))^2K_i^2}{Nh}\frac{\sum_{i\in U}\mathrm{E}\left[(\Gamma(x)-\widehat{\Gamma}(x))^2\right]K_i^2}{Nh}\right\} \\
&\quad +\frac{1}{h}\left(\frac{2n\max_{i\neq j}\left|\Delta_{ij}\right|}{\lambda^2\lambda^*}+\frac{2n}{\lambda^2N}\right)\frac{\sum_{i\in U}\mathrm{E}\left[(\Gamma(x)-\widehat{\Gamma}(x))^2\right]K_i^2}{Nh} \\
&= O(1)O\left(\frac{1}{\sqrt{Nh^3}}\right)+O(1)O\left(\frac{1}{Nh^2}\right)=O\left(\frac{1}{\sqrt{Nh^3}}\right),
\end{aligned}$$

by Lemma 7. Next write

$$\begin{aligned}
\frac{n^2}{N^4}\mathrm{E}\left[B_N^2\right] &= \frac{n^2}{N^4h^4}\mathrm{E}\left[\left\{\sum_{i,j\in U}\left[z_i-\Gamma(x)\right]K_i\left[z_j-\Gamma(x)\right]K_j\frac{\Delta_{ij}}{\pi_i\pi_j}\frac{I_iI_j-\pi_{ij}}{\pi_i\pi_j}\right\}^2\right] \\
&= \frac{N^2}{N^4h^4}\sum_{i,k\in U}\frac{1-\pi_i}{\pi_i}\frac{1-\pi_k}{\pi_k}\left[z_i-\Gamma(x)\right]^2K_i^2\left[z_k-\Gamma(x)\right]^2K_k^2\frac{\Delta_{ik}}{\pi_i\pi_k} \\
&\quad +\frac{2n^2}{N^4h^4}\sum_{i\in U}\sum_{k\neq l}\left[z_i-\Gamma(x)\right]^2K_i^2\left[z_k-\Gamma(x)\right]K_k\left[z_l-\Gamma(x)\right]K_l\frac{\Delta_{kl}}{\pi_k\pi_l} \\
&\quad \times\mathrm{E}\left[\frac{I_i-\pi_i}{\pi_i}\frac{I_kI_l-\pi_{kl}}{\pi_{kl}}\right] \\
&\quad +\frac{n^2}{N^4h^4}\sum_{i\neq j}\sum_{k\neq l}\frac{\Delta_{ij}}{\pi_i\pi_j}\frac{\Delta_{kl}}{\pi_k\pi_l}\left[z_i-\Gamma(x)\right]K_i\left[z_j-\Gamma(x)\right]K_j\left[z_k-\Gamma(x)\right] \\
&\quad \times K_k\left[z_l-\Gamma(x)\right]K_l\mathrm{E}\left[\frac{I_iI_j-\pi_{ij}}{\pi_{ij}}-\frac{I_kI_l-\pi_{kl}}{\pi_{kl}}\right] \\
&= b_{1N}+b_{2N}+b_{3N}.
\end{aligned}$$

Addressing the individual pieces we have

$$\begin{aligned} b_{1N} &\leq \frac{n^2}{N^4h^4\lambda^3}\sum_{i\in U}(z_i-\Gamma(x))^4K_i^4 \\ &\quad +\frac{n^2\max_{i\neq j}|\Delta_{ij}|}{N^4h^4\lambda^4}\sum_{i\neq j}(z_i-\Gamma(x))^2K_i^2(z_j-\Gamma(x))^2K_j^2 \\ &\leq \left(\frac{n^2}{N^4h^4\lambda^3}+\frac{n^2\max_{i\neq j}|\Delta_{ij}|}{N^4h^4\lambda^4}\right)\sum_{i\in U}(z_i-\Gamma(x))^4K_i^4 \\ &= O\left(\frac{1}{Nh^3}\right)+O\left(\frac{1}{N^2h^3}\right)\to 0 \end{aligned}$$

as $N\to\infty$ by Lemma 3. Next write

$$\begin{aligned} b_{3N} &\leq \frac{n^2(\max_{i\neq j}|\Delta_{ij}|)^2}{h^3\lambda^4\lambda^{*2}}\max_{i,j,k,l\in D_4}|\mathrm{E}\left[(I_iI_j-\pi_{ij})(I_kI_l-\pi_{kl})\right]|\sum_{i\in U}\frac{(z_i-\Gamma(x))^4K_i^4}{Nh} \\ &= O\left(\frac{1}{Nh^3}\right)\to 0 \end{aligned}$$

as $N\to\infty$ by assumption A3 and Lemma (3). By the Cauchy-Schwarz inequality, $b_{2N}\to 0$ as $N\to\infty$, so $nN^{-2}|B_N|\to 0$ as N goes to infinity.

■

5.10 Appendix B10: Proof of Theorem 12

Proof of Theorem 12. Write

$$\begin{aligned} \mathrm{E}\left[\{\widehat{\mu}(\boldsymbol{x})-\mu(\boldsymbol{x})\}^2\right]\eta^{-2} &= \mathrm{Var}\left(\widehat{\mu}(\boldsymbol{x})\right)=\mathrm{Var}\left(\boldsymbol{x}^T\widehat{\boldsymbol{\beta}}+\widehat{\Gamma}_\pi^T\widehat{\boldsymbol{\gamma}}\right) \\ &= \left[\mathrm{Var}\left(\boldsymbol{x}^T\widehat{\boldsymbol{\beta}}\right)+\mathrm{Var}\left(\widehat{\Gamma}_\pi^T\widehat{\boldsymbol{\gamma}}\right)+2\mathrm{Cov}\left(\boldsymbol{x}^T\widehat{\boldsymbol{\beta}},\widehat{\Gamma}_\pi^T\widehat{\boldsymbol{\gamma}}\right)\right] \\ &= \left\{O(n^{-1})+O((Nh)^{-1})+O((nNh)^{-1/2})\right\}\to 0, \end{aligned}$$

as $N\to\infty$. Further justification for the final line above is as follows: from standard least-squares regression results, $(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ and $(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})$, are both $O_p(n^{-1/2})$, and thus, $\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right)=$

$O(n^{-1})$. Furthermore, $(\widehat{\Gamma}_\pi(x) - \Gamma(x))$ is $O_p((nh)^{-1/2})$ by Lemma 7, so

$$\begin{aligned}
\mathrm{Var}\left(\widehat{\Gamma}_\pi^T\widehat{\boldsymbol{\gamma}}\right) &= \mathrm{E}\left[(\widehat{\Gamma}_\pi^T\widehat{\boldsymbol{\gamma}})^2\right] - \mathrm{E}\left[\widehat{\Gamma}_\pi\right]^T \mathrm{E}\left[\widehat{\boldsymbol{\gamma}}\right] \\
&= \left(\mathrm{Var}\left(\widehat{\Gamma}_\pi\right) + \mathrm{E}\left[\widehat{\Gamma}_\pi\right]\right)\left(\mathrm{Var}\left(\widehat{\boldsymbol{\gamma}}\right) + \mathrm{E}\left[\widehat{\boldsymbol{\gamma}}\right]\right) + \mathrm{E}\left[\widehat{\Gamma}_\pi\right]^T \mathrm{E}\left[\widehat{\boldsymbol{\gamma}}\right] \\
&= \mathrm{Var}\left(\widehat{\Gamma}_\pi\right)\mathrm{Var}\left(\widehat{\boldsymbol{\gamma}}\right) + \mathrm{E}\left[\widehat{\Gamma}_\pi\right]^T \mathrm{Var}\left(\widehat{\boldsymbol{\gamma}}\right) + \mathrm{Var}\left(\widehat{\Gamma}_\pi\right)\mathrm{E}\left[\widehat{\boldsymbol{\gamma}}\right] \\
&= O\left(\frac{1}{nNh}\right) + O\left(\frac{1}{n}\right) + O\left(\frac{1}{Nh}\right) = O\left(\frac{1}{Nh}\right).
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathrm{Cov}\left(\boldsymbol{x}^T\widehat{\boldsymbol{\beta}}, \widehat{\Gamma}_\pi^T\widehat{\boldsymbol{\gamma}}\right) &\leq \mathrm{Var}^{1/2}(\boldsymbol{x}^T\widehat{\boldsymbol{\beta}})\mathrm{Var}^{1/2}(\widehat{\Gamma}_\pi^T\widehat{\boldsymbol{\gamma}}) = O\left(\frac{1}{\sqrt{n}}\right) O\left(\frac{1}{\sqrt{Nh}}\right) \\
&= O\left(\frac{1}{\sqrt{nNh}}\right).
\end{aligned}$$

■