Probability Weighted Nonparametric Conditional PDF Estimation

Luc Clair

November 11, 2016

Abstract

Conditional probability distribution function (PDF) estimation provides a broader view of the relationship between an outcome variable y and a vector of predictor variables x than does conditional expectation estimation. Conditional PDF estimation allows one to extract multiple quantities of interest, including the expectation, modes, and prediction intervals. Many variables derived from survey responses are discrete and when the outcome variable in a regression model is binary, the conditional mean coincides with the conditional probability that $y = 1$, but alternatively one could model the conditional PDF directly. The mathematical definition for the conditional PDF is the joint PDF of the outcome and predictor variables divided by the PDF of predictor variables. I propose estimating the conditional PDF by replacing the unknown population density functions by their respective probability weighted kernel estimators. I derive the asymptotic properties of this estimator and show that it is asymptotically normal under the familiar assumptions in the kernel estimation and survey statistic literatures. I run Monte Carlo simulations to compare the finite sample properties of the proposed estimator to traditional parametric and nonparametric binary choice model estimators. I conclude with an empirical example where I estimate the effect of prescription drug insurance on the use of mental health pharmaceuticals.

1 Introduction

Conditional probability distribution function (PDF) estimation provides a broader view of the relationship between an outcome variable y and a vector of predictor variables x than does conditional expectation estimation. Conditional PDF estimation allows one to extract multiple quantities of interest, including the expectation, modes, and prediction intervals. These methods have many applications in econometrics including modelling count data, discrete choice, and propensity score matching.

Parametric and semi-parametric estimation of conditional PDFs can be thought of as *indirect*. In estimating binary choice models, the discrete outcome variable is often viewed as binary transformation of an unobserved regression,

$$y^* = x'\beta + \epsilon,$$

known as a latent variable regression. The outcome variable y then takes a value of one or zero according to a rule about the latent variable y^*, i.e. $y = \mathbf{1}(y* > 0)$, where $\mathbf{1}(\cdot)$ is an indicator function taking a value of one if the logical argument in brackets is true and zero otherwise. The probability that $y = 1$ conditional on x is then given by a single-index form: $Pr(y = 1|x) = F(x'\beta)$, where $F(\cdot)$ is the cumulative distribution function (CDF) of ϵ and $Pr(y = 1|x)$ is a function of x only through the index function $x'\beta$. The set of parameters β reflects the impact of changes in x on the probability $Pr(y = 1|x)$. The parametric methods for estimating β require one to specify $F(\cdot)$; the most popular choices are the standard normal CDF (Probit) and the logistic CDF (Logit). By selecting a CDF for the single-index function, the conditional probabilities are constrained between zero and one. These functions are nonlinear and often estimated using maximum likelihood methods. The fitted values from estimating the single-index model give the estimated conditional probability that $y = 1$.

While the Logit and Probit models correct the issue of improper probabilities of the linear probability model, the restriction placed on the data generating process (DGP) of the dependent variable is problematic. For example, the Logit model assumes that the logistic distribution is the correct DGP. If this is the true DGP, then the model is efficient and unbiased (McLeod 2011). However, if the underlying DGP does not follow the logistic distribution, then the model will be

misspecified and estimates will no longer be consistent. This raises the appeal of semiparametric and nonparametric estimators as they relax the functional form assumption about the DGP. In semiparametric single-index models, $F(x'\beta)$ is not specified and must be estimated (Klein & Spady 1993). Like parametric binary single-index models, semiparametric single-index models estimate the parameter of interest using maximum likelihood estimation, where the fitted values give the estimated conditional probability that $y = 1$.

Letting $f(y|x)$, $f(x, y)$, and $f(x)$ denote the conditional PDF of y given x, the joint probability of x and y and the marginal density of x, respectively, then:

$$f(y|x) = \frac{f(y, x)}{f(x)}. \tag{1.1}$$

By replacing the unknown population quantities on the right-hand-side (RHS) of (1.1) by their kernel density estimators, $\tilde{f}(y, x)$ and $\tilde{f}(x)$, one can obtain a *direct* method for estimating conditional PDFs nonparametrically:

$$\tilde{g}(y|x) = \frac{\tilde{f}(y, x)}{\tilde{f}(x)} = \frac{\frac{1}{h_0}\sum_{i=1}^{n} K_\gamma(x, x_i) w\left(\frac{y-y_i}{h_0}\right)}{\sum_{i=1}^{n} K_\gamma(x, x_i)}, \tag{1.2}$$

where K_γ is a multivariate mixed-data product kernel, $w(\cdot)$ is a univariate kernel, and h_0 is the bandwidth for the outcome variable (Hall et al., 2004).

In an applied setting, researchers often rely on survey data to estimate their models. The sampling scheme for many of these surveys is complex with a multi-stage design, using a mix of stratification and clustering. These complex sampling plans lead to unequal inclusion probabilities for the units in the sample and an unrepresentative sample for the population. If one is attempting to estimate descriptive statistics, it is highly recommended that observations be weighted by the inverse of their inclusion probabilities to return precise estimates. Solon et al. (2013) compared a sample of units with unequal probabilities of inclusion to viewing a representative sample through a 'funhouse mirror,' where oversampled subgroups will be exaggerated. Estimates using weighted observations better represent the true population parameters. If the goal is to estimate marginal effects of x on y, then the inclusion of survey weights depends on the sampling criterion. The sampling criterion is the variable by which the sampling scheme is designed. If analysts stratify the population by income cohort in order to increase the precision of estimates for low income

individuals, then income is the sampling criterion. If the model is set up so that the sample criterion is correlated with the error term, then sampling is said to be endogenous. As with most forms of endogeneity in econometric models, ignoring sampling endogeneity will lead to inconsistent estimates. If the sampling criterion was omitted as a predictor variable, adding this variable on the right-hand side (RHS) will correct for the resulting endogeneity (Cameron & Trivedi 2009). If the sampling criterion is the outcome variable, then each observation in the sample must be weighted by the inverse of its inclusion probability to ensure consistent results (Magee, Robb & Burbidge 1998).

Estimators that include a model structure and weight observations by inverse probability weights are called *model-assisted* estimators. These estimators are a compromise between *model-based* and *purely design-based* estimators. Model-based estimators ignore the sampling design and assume the data (y, x) is generated according to a given model. These estimators hold the strong assumption that the model applies to all units in the sample. Survey data contains a limited number of variables, therefore it is probable that a theoretical model that holds for all observations does not exist. To use asymptotic results in finite population sampling, one assumes that the population is a subset of nested superpopulations. The parametric and semi-parametric single-index models and the nonparametric estimator proposed by Hall et al. (2004) are examples of model based estimators. Conversely, purely design-based estimators make no assumption of a model structure between y and explanatory variables x thereby ignoring any explanatory power x has for y. An example of this would be the probability weighted proportion, denoted by $\bar{y}_w$:

$$\bar{y}_w = \frac{\sum_{i=1}^{n} \pi_i^{-1} y_i}{\sum_{i=1}^{n} \pi_i^{-1}},$$

where π_i is the probability that unit i is in the sample. This type of estimator provides no information on the marginal affect some variable x has on y. Model-assisted estimators provide an option for including a model structure while accounting for the design of the sample. Inverse probability weighting can be used to account for missing data by inflating the weights for observed subjects who are under-represented due to unequal probability sampling (Li & Yang 2016).

The purpose of this paper is to develop a probability weighted nonparametric conditional PDF estimator. I propose replacing the unknown population densities from (1.1) by their probabil-

ity weighted kernel estimators to develop a nonparametric conditional PDF estimator that takes into account the sampling design. There is a growing literature on the estimation of nonparametric models using complex survey data. Breidt and Opsomer (2000) introduced local polynomial estimators for estimating population totals, showing that their estimator is design unbiased and consistent. Buskirk and Lohr (2005) studied the finite sample and asymptotic properties of a probability weighted kernel density estimator introduced by Bellhouse and Stafford (1999). Breunig (2001) and (2008) developed methods for kernel density estimation using data from clustered and stratified samples, respectively. Sánchez-Borregeo et al. (2014) extend the estimator of Breidt and Opsomer (2000) to include mixed-data types for the local constant. Clair (2015) derived the asymptotic properties of the modified local constant and found that efficiency gains can be made by including survey weights under endogenous sampling. To the best of my knowledge, there have been no papers that have looked at probability weighted nonparametric conditional PDF estimators.

Following this introduction, in Section 2, I review methods for estimating discrete choice models with complex survey data, focusing on binary choice models. Next, in Section 3, I provide an overview of the proposed estimator where I derive its asymptotic properties under the combined framework outlined in Pfefferman (1993). I show that the estimator is consistent under standard assumptions in both the kernel estimation literature and survey statistics literature. I further show that the estimator is asymptotically normal by applying Liapunov's Central Limit Theorem. I also propose methods for selecting the smoothing parameters for the probability weighted kernel estimator. I examine a least squares cross-validation approach to selecting the smoothing parameters, however, I elect to use likelihood cross-validation for my simulations and application due to its computational advantages. In Section 4, I run Monte Carlo simulations in order to assess the finite sample properties of the estimator under different sampling plans, where performance is based on the mean squared error (MSE) criterion. Section 5 applies the model-assisted nonparametric PDF estimator to the estimation of the effect of private supplementary health insurance on the use of mental health care pharmaceuticals in Canada. Results show that while insurance has a positive effect on prescription drug usage, weighted estimates reduce the likelihood of anti-psychotics use for insured individuals and enhances the effect of insurance on the use anti-depressants and benzodiazepines compared to unweighted estimates. Section 6 concludes.

2 Conditional PDF Estimation

2.1 Sample Design

Consider a finite population $U = \{1, ..., N\}$ of N units. For each $j \in U$ a binary outcome variable y_j and auxiliary variables x_j are observed, where x_j is a $p \times 1$ vector with a mixture of continuous and discrete variables. Next, a sample S of size n_s is drawn based on a sampling plan $p_N(\cdot)$, where $p_N(\mathcal{S})$ is the probability of drawing the sample S. The sampling rate is $Q = n_s/N$, with first order inclusion probabilities $\pi_j = Pr(j \in \mathcal{S}) = \sum_{j \in \mathcal{S}} p_N(\mathcal{S})$ and second order inclusion probabilities $\pi_{ji} = Pr(j, i \in \mathcal{S}) = \sum_{j,i \in s} p_N(\mathcal{S})$. The variable n_s may be fixed (as in simple random sampling (SRS)) or random; however, no sampling plan is specified. The first and second order probabilities are the probabilities of obtaining the unit j and units j and i, respectively, while sampling from the population according to the complex sampling design.

2.2 Parametric Estimators

The use of ordinary least squares (OLS) to estimate $E(y|x)$ when y is binary is known as the linear probability model (LPM). The issues relating to the LPM are well-known to econometricians; review of this model should be viewed as a pedagogical warm up for those unfamiliar or lacking recent experience with binary choice models. The LPM is simply estimating the relationship between y and x by the linear equation:

$$y = x\beta + \epsilon, \tag{2.1}$$

where β is a $p \times 1$ vector of parameters and ϵ is the residual. It is also assumed that $E(\epsilon|x) = 0$. Taking the conditional expectation of (2.1) with respect to x gives $E(y|x) = x\beta$. In addition, note that $E(y|x) = Pr(y = 1|x) \Rightarrow Pr(y = 1|x) = x\beta$. In this model, the fitted values, $\tilde{y} = x\tilde{\beta}_{OLS}$ are interpreted as the probability that the outcome variable equals 1 given x and the estimated coefficients, $\tilde{\beta}_{OLS}$, give the marginal effects of the predictor variables.

While this model provides convenient interpretation, the LPM is critically flawed for two reasons. The first is there is no mechanism which restrict the fitted values to the interval [0,1], i.e. estimated probabilities greater than one or less than zero are possible. Furthermore, the assumption that the probability is linearly related to a regressor variable for all possible values is problematic. If it were,

then continually increasing a continuous predictor variable would eventually drive $Pr(y = 1|x)$ above one or below zero. The second is that the model will display heteroskedastic errors. The binary nature of the dependent variable implies a binary error term, i.e. $\epsilon_i = -x\beta$ or $1 - x\beta$. The conditional variance of this error term, $var(\epsilon|x) = x\beta[1 - x\beta]$, will take on different values when these probabilities are closer to the extreme values, 0 and 1, than when they are closer to 0.5 (and potentially negative) (Greene 2012). The OLS estimator is still unbiased, but the conventional formula for estimating the standard errors, and the t-statistics, will be wrong. The easiest way of solving this problem is to obtain estimates of the standard errors that are robust to heteroskedasticity.

The parametric solution is to specify the distribution of the error term ϵ, $F(\cdot)$, such that $P(y = 1|x) = 1 - F(-x'\beta)$. For the Probit estimator, $F(\cdot)$ is specified as the CDF for the standard normal distribution, denoted by Φ,

$$\Phi(x\beta) \equiv \int_{-\infty}^{x\beta} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right),$$

and for the Logit estimator $F(\cdot)$ is specified as the logistic distribution CDF, denoted by Λ,

$$\Lambda(x\beta) = \frac{\exp(x\beta)}{1 + \exp(x\beta)}.$$

These functions are both monotonically increasing in $x\beta$ and have varying partial effects.

Because Logit and Probit models are non-linear, one must employ maximum likelihood (ML) methods to estimate β. The ML estimate of β is the particular vector $\tilde{\beta}_{ML}$ that gives the greatest likelihood of observing the sample $\{y_1, y_2, ..., y_n\}$, conditional on the explanatory variables x. Since y is a Bernoulli random variable, the likelihood function is given by:

$$L(y|x) = \prod_{i=1}^{n} F(x_i'\beta)^{y_i}[1 - F(x_i'\beta)]^{(1-y_i)}.$$

To solve for $\tilde{\beta}_{ML}$, one maximizes the log-likelihood function given by:

$$\max_{\beta} \ln L = \sum_{i=1}^{n} \{y_i \ln F(x_i'\beta) + (1 - y_i) \ln[1 - F(x_i'\beta)]\}$$

(Greene 2012).

The LPM, Logit, and Probit estimators are considered model-based estimators as they do not take into account the sample-design. If one chooses a linear specification (i.e. LPM), a design-based estimator would simply be to estimate the model's parameters, β, by weighted least squares:

$$\hat{\beta}_{WLS} = (x^T W x)^{-1} x^T W y,$$

where $W = \text{diag}(\pi_i^{-1})$. Magee et al. (1998) showed this estimator was consistent under the combined inference framework. Like the LPM, the binary choice model estimated via weighted least squares suffers from the same issues that arise when restricting the conditional PDF to be a linear function. To incorporate survey weights into a Logit or Probit specification, one would choose β to maximize the following log-likelihood function:

$$\ln L = \sum_{i=1}^{n} \pi_i^{-1} \{ y_i \ln F(x_i'\beta) + (1 - y_i) \ln[1 - F(x_i'\beta)] \},$$

where each observation is weighted by the inverse of the probability of inclusion.

While the Logit and Probit methods correct the issue of invalid probabilities of the LPM, the restriction placed on the DGP of the dependent variable is problematic. These estimators assume that $F(\cdot)$ is the *correct* DGP. If this is the true DGP, then the model is efficient and unbiased. However, if the underlying DGP does not follow the specified distribution, then the model will be misspecified and no longer efficient and unbiased. While these methods are still popular today, researchers using these methods ought to recognize their limitations..

2.3 Semiparametric Estimation

In the case where $F(x'\beta)$ is assumed to be unknown, one can apply semiparametric methods for estimating β. Klein and Spady (1993) proposed estimating β by maximizing the following log-likelihood function:

$$L(\beta) = \sum_i (1 - y_i) \ln(1 - \tilde{l}(x_i'\beta)) + \sum_i y_i \ln(\tilde{l}(x_i'\beta)), \tag{2.2}$$

where

$$\tilde{l}(x_i'\beta) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x_i'\tilde{\beta} - x'\tilde{\beta}}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i'\tilde{\beta} - x'\tilde{\beta}}{h}\right)}$$

is the local constant estimator of $F(x_i'\beta)$. In order for β to be identified, the semiparametric index model must meet certain underlying conditions. The first is that x cannot contain a location parameter and must contain at least one continuous variable. If x contains a constant term, $l(\cdot)$ is not unique and β is not identified. Li and Racine (2007, p.252) provided the following example: "for any nonzero constants α_1 and α_2 and for any $l(\cdot)$ function and fixed vector β_α, we can always find another function $l_2(\cdot)$, such that $l_2(\alpha_1 + \alpha_2 x'\beta) = g(x'\beta)$". If x only includes discrete variables there will be an infinite number of choices for $l(\cdot)$ and β to satisfy a finite number of restrictions imposed by $P(y = 1|x) = l(x'\beta)$. Second, looking at (2.2), it is obvious that if $l(\cdot)$ is constant, then β is not identified. It is apparent then, that $l(\cdot)$ must be differentiable on the support of $x'\beta$. Furthermore, it is important that x be full rank, i.e. no perfect multicollinearity.

The semiparametric single-index model allows one to relax functional form assumptions and avoid the problem of distribution misspecification while still mitigating the effects arising from the curse of dimensionality.

Recent research has looked at inverse probability weighting in semiparametric single-index models. Li and Yang (2016) looked at estimating a single-index model of the form:

$$y = g(Z'\beta) + \epsilon, \tag{2.3}$$

where some of the l covariates Z are missing. The authors estimate the model using a modified local linear estimator of the form:

$$\hat{g}_{SIM}(x'\beta) = \sum_{i=1}^{n} \frac{V_i}{\pi_i}[y_i - a - b(Z_i'\beta - u)]K_h(Z_i'\beta), \tag{2.4}$$

where V_i is an indicator function taking a value of 1 if Z is missing and 0 otherwise, and $K_h(\cdot)$ is a product kernel for continuous variables. The authors only considered the continuous variable case and used a rule-of-thumb method for selecting the bandwidth. To the best of my knowledge, there is no research looking at a probability weighted single-index model in the form of Klein and Spady

(1993). I do not investigate it here and leave it to future research.

2.4 Nonparametric Estimation

The fitted values for the LPM and single-index models described above are the conditional probability densities of y given x, i.e. $f(y|x)$. These methods are semiparametric in nature as they use a parametric specification for the indexs function but a nonparametric specification for the unknown distribution function. As mentioned above, one can take a more direct approach by estimating $f(y|x)$ nonparametrically. The estimator in (1.2) proposed by Hall et al. (2004) is for a univariate dependent variable. For a general form with multivariate mixed data dependent variables, first define the p-variate vector z as $z = (y, x)$. Letting the superscripts c and d denote continuous and discrete variables, partition the p variables into q continuous variables and r discrete variables, i.e. $z = \{z^c, z^d\} = \{z_1^c, ..., z_q^c, z_1^d, ..., z_r^d\}$. I use z_{is}^c to denote the sth component of z_i^c and z_{is}^d for the sth component of z_i^d and assume that z_s^d takes $c_s \geq 2$ different values in $\mathcal{D}_s = \{0, 1, ..., c_s - 1\}$, $s = 1, ..., r$. Write $y = \{y^c, y^d\}$ and $x = \{x^c, x^d\}$ and assume that y^c contain the first q_y continuous components of z^c while y^d contains the first r_y discrete components of z^d. Thus, $y^c \in \mathbf{R}^{q_y}$, $y^d \in \prod_{s=1}^{r_y}\{0, 1, ..., c_s - 1\}$, and $x^c \in \mathbf{R}^{q-q_y}$. Similarly, $x^d \in \prod_{s=r_y+1}^{r}\{0, 1, ..., c_s - 1\}$.

To estimate $f(y|x) = f(z)/f(x)$ in this case, define a univariate kernel function for discrete variables as

$$l(z_{is}^d, z_{js}^d, \lambda_s) = \lambda_s^{1-\mathbf{1}(z_{is}^d = z_{js}^d)}, \tag{2.5}$$

which takes a value of 1 if $z_{is}^d = z_{js}^d$ and λ_s otherwise, where λ_s is the smoothing parameter for z_s^d. The discrete kernel function in (2.5) is a variation of the Aitchison and Aiken (1976) kernel described in Li and Racine (2007, p.131). This specification holds the property that when $\lambda_s = 1$, the kernel for x_s^d is constant and the variable gets smoothed out, i.e. z_s^d is an irrelevant variable. When $\lambda_s = 0$, the kernel for z_j^d is an indicator function. The product kernel for discrete variables z^d is

$$L(z_i^d, z_j^d, \lambda) = \prod_{s=1}^{r} l(z_{is}^d, z_{js}^d, \lambda_s). \tag{2.6}$$

Letting $W(\cdot)$ denote the product kernel function for z^c, write

$$W(z_{is}^c, z_{js}^c) = \prod_{s=1}^{q} \frac{1}{h_s} k\left(\frac{z_{is}^c - z_{js}^c}{h_s}\right), \tag{2.7}$$

where $k(\cdot)$ is a univariate kernel function and h_s is the bandwidth for z^c. There are many choices of $k(\cdot)$; the most common are the Gaussian and Epinechnikov kernels (Silverman 1986). A mixed-data product kernel function can then be written as $K_{\gamma,iz} = W(z_{is}^c, z_{js}^c)L(z_i^d, z_j^d, \lambda)$, leading to a kernel density estimator for $f(z)$ given by:

$$\tilde{f}(z) = \sum_{i=1}^{n} K_{\gamma,iz}, \tag{2.8}$$

where $\gamma = (h, \lambda)$. Using similar notation, the product kernels for y^d and y^c are $L(y_i^d, y_j^d, \lambda) = \prod_{s=1}^{r_y} l(y_{is}^d, y_{js}^d, \lambda_s)$ and $W(y_{is}^c, y_{js}^c) = \prod_{s=1}^{q_y} h_s^{-1} k((y_{is}^c - y_{js}^c)/h_s)$, respectively. Also, define $L(x_i^d, x_j^d, \lambda) = \prod_{s=1}^{r-r_y} l(x_{is}^d, x_{js}^d, \lambda_s)$ and $W(x_{is}^c, x_{js}^c) = \prod_{s=1}^{q-q_y} h_s^{-1} k((x_{is}^c - x_{js}^c)/h_s)$. Using these definitions, a kernel estimator for $f(y|x)$ for the multivariate dependent variable case is given by (Racine, Li & Zhu 2004):

$$\tilde{g}(y|x) = \frac{\sum_{i=1}^{n} K_{\gamma,iz}}{\sum_{i=1}^{n} K_{\gamma,ix}}, \tag{2.9}$$

with $K_{\gamma,ix} = W(x_{is}^c, x_{js}^c)L(x_i^d, x_j^d, \lambda)$.

3 Model-Assisted Nonparametric Conditional PDF Estimation

An advantage of the Hall et al. (2004) method is that while it assumes the existence of a differentiable DGP, it does not assume anything about its functional form other than smoothness. The issue in a complex survey setting is that the estimator in (2.9) is a model-based estimator as it does not take the sampling scheme into account. In order to incorporate sample design into conditional PDF estimation, I propose replacing the unknown densities $f(z)$ and $f(x)$ on the RHS of (1.1) with probability weighted kernel density estimators $\hat{f}(z) = \sum_{i=1}^{n} \pi_i^{-1} K_{\gamma,iz}$ and $\hat{f}(x) = \sum_{i=1}^{n} \pi_i^{-1} K_{\gamma,ix}$, respectively:

$$\hat{g}(y|x) = \frac{\sum_{i=1}^{n} \pi_i^{-1} K_{\gamma,iz}}{\sum_{i=1}^{n} \pi_i^{-1} K_{\gamma,ix}}. \tag{3.1}$$

Note that if the sample design is a simple random sample (SRS), $\pi_i = N/n$ is constant for all $i = 1, ..., n$ and (3.1) reduces to (2.9).

3.1 Asymptotic Properties

The inference method used in this paper is the combined inference framework outlined in Pfefferman (1993) and used in Clair (2015), Harms and Duchesne (2010), and Buskirk and Lohr (2005). This mode of inference has two stages: a model stage and a design stage. First a finite population U is generated according to a superpopulation model, denoted ξ, where elements in the finite population are presumed to be realizations of random variables with a joint probability distribution. A model is selected based on the belief that it has generated the population. For each j in the population, the realization (x_j, y_j) is obtained such that (x, y) follows the joint density $f(x, y)$. For the analysis that follows, it is assumed that the x_j's, $j \in U$, are independently and identically distributed (i.i.d.). This is a popular assumption when working in the combined inference framework (Bellhouse & Stafford 1999). This represents the model stage. The conditional distribution is estimated in the design stage: a sample $\mathcal{S}$ of size n_s is drawn according to a specified sampling plan and the corresponding weights are included in the model. Let the subscripts ξ, P, and C denote the conditional mathematical expectation under model-based inference, design-based inference, and combined inference, respectively. The conditional expectation under the combined inference framework is calculated as $E_C = E_\xi\{E_P\{\cdot|\pi\}|x\}$, where $x = \{x^c, x^d\}$. For a basic application of the combined inference method, see Clair (2015, Section 4). Before deriving the asymptotic properties, I presume the following assumptions hold.

Assumption 3.1. *Denote* $\mathbf{S} = \mathbf{S}^c \times \mathbf{S}^d$ *as the compact support of* x*. The densities* $f(z)$ *and* $f(x)$ *have two continuous derivatives as functions of* z^c *and* x^c*, respectively;* $f(x)$ *is bounded away from 0 for* $x = (x^c, x^d) \in \mathbf{S}$*, and* $\sup_{x \in \mathbf{S}} f(z)$ *vanishes outside a compact set of values* y*.*

Assumption 3.2 (Kernel function)**.** *The kernel function* $k(\cdot)$*:*$\mathbf{R} \to \mathbf{R}$ *is symmetric with* $k(v) \geq 0$ *with* $v \in \mathbf{R}$*, and bounded by finite constant* z *so that* $k(v) \leq z$*.* $k(\cdot)$ *is* m *times differentiable with* $\int k(v)v^4 dv < \infty$*.* $k(\cdot)$ *is a second order kernel and define* $\kappa_2 = \int v^2 k(v) dv$ *and* $\kappa = \int k^2(v) dv$*.*

Assumption 3.3. *(*$h_1, ..., h_{q_y}, h_{q_y+1}, .., h_q, \lambda_1, ..., \lambda_{r_y}, ..., \lambda_r) \in [0, \eta]^{q+r}$ *lies in a shrinking set and* $\eta = \eta_N$ *is a positive sequence that converges to zero at a rate slower than the inverse of any*

polynomial in N. $Nh_1...h_q \geq t_N$ with $t_N \to \infty$ as $N \to \infty$.

Assumption 3.4 (Sample Design)**.** *The sampling plan $p_N(S)$ is such that as $i \to \infty$, the sampling rate $n_{s,i}/N_i$ converges with probability one to a finite constant $1 \geq Q > 0$. It is further assumed the design expectation of n_s is $E_P(n_s) = n$. The first order inclusion probabilities are such that for all N, $\min_{j \in U} \pi_j \geq \epsilon > 0$, with probability one. The second order inclusion probabilities satisfy $\min_{i,j \in U} \pi_{ij} \geq \epsilon^* > 0$ and*

$$\lim_{i \to \infty} \sup n_{s,i} \max_{j, i \in U: j \neq i} |\pi_{ij} - \pi_i \pi_j| \leq \infty,$$

with probability one.

Assumptions 3.1 to 3.3 are standard assumptions in the kernel statistics literature. Assumption 3.3 simply ensures that as $N \to \infty$, the bandwidths h_s for $s = 1, ..., q$ shrink to zero and the kernel function remains smooth. Assumption 3.4 is standard in the survey statistics literature and states that the first and second order inclusion probabilities are non-zero for all i and j.

Theorem 3.1 (Asymptotic Pointwise MSE of $\hat{g}(y|x)$)**.** *If Assumptions* 3.1-3.4 *are satisfied, then the conditional pointwise MSE of the model-assisted conditional PDF estimator $\hat{g}(y|x)$ under the combined inference mode is given by:*

$$\begin{aligned} & MSE(\hat{g}(y|x)) \\ &= \left[f^{-1}(x) \left(\frac{\kappa_2}{2} \sum_{s=1}^{q} B_{1s} h_s^2 + \sum_{s=1}^{r} \sum_{t^d \in \mathcal{D}} B_{2s} \lambda_s \right) \right]^2 \\ &\quad + \frac{1}{nh_1...h_q} (\Delta + Q) \frac{\kappa^q g(y|x)}{f(x)} + o_p \left((nh_1...h_q)^{-1} \left(\sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{r} \lambda_s \right) + \left(\sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{r} \lambda_s \right)^2 \right) \end{aligned} \tag{3.2}$$

where $\Delta = N^{-2} n \sum_{i=1}^{N} (\pi_i^{-1} - 1)$, $Q = n/N$ and

$$B_{1s}(z) = \begin{cases} (1/2)\kappa_2 f_{ss}(y,x)/f(x) & \text{if } s = 1, ..., q_y \\ (1/2)\kappa_2 [f_{ss}(y,x) - f_{ss}(x) g(y|x)]/f(x) & s = q_y + 1, ..., q \end{cases} \tag{3.3}$$

$$B_{2s}(z) = \begin{cases} \sum\limits_{t^d \in \mathcal{D}} \mathbf{1}(t^d, z^d) f(z^c, t^d) & \text{if } s = 1, ..., r_y \\ \sum\limits_{u^d \in \mathcal{D}_x} \mathbf{1}(u^d, x^d)(f(z^c, y^d, u^d) - g(y|x) f(x^c, u^d)) & s = r_y + 1, ..., r \end{cases} \tag{3.4}$$

Theorem 3.1 outlines the asymptotic pointwise MSE for the weighted nonparametric conditional PDF estimator. The first term on the RHS of (3.2) is the square of the pointwise bias of $\hat{g}(y|x)$ and the second term is the asymptotic pointwise variance. It has the same form as the MSE for the estimator developed by Hall, Li, and Racine (2004) except for the adjustment term $\Delta + Q$ on the variance. This adjustment factor is present because of the sample design. Under SRS $\Delta + Q = 1$ and the MSE for $\hat{g}(y|x)$ is equal to the MSE of $\tilde{g}(y|x)$. However, under sampling plans with unequal inclusion probabilities, this adjustment factor is greater than one. Therefore, the variance of $\hat{g}(y|x)$ is greater or equal to the variance of $\tilde{g}(y|x)$. The adjustment factor can be interpreted as the ratio of the average of the sampling weights over the sampling weight assuming the sampling plan is SRS. A direct corollary of the result is that $\hat{g}(y|x)$ is consistent under assumptions 3.1 to 3.4.

The following theorem is proved in Section 7.2 and describes the asymptotic normality of $\hat{g}(y|x)$.

Theorem 3.2 (Asymptotic Normality of $\hat{g}(y|x)$)**.** *If assumptions* 3.1-3.4 *are satisfied, the asymptotic normality of* $\hat{g}(x)$ *is defined by:*

$$\sqrt{Nh_1...h_q}\left(\hat{g}(y|x) - g(y|x) - \sum_{s=1}^{q} h_s^2 - \sum_{s=1}^{q} \lambda_s\right) \xrightarrow{d} N(0, (\Delta + Q)\kappa^q g(y|x)/f(x)). \tag{3.5}$$

3.2 Bandwidth Selection

In this section I examine cross-validation methods for selecting the bandwidths in the probability weighted conditional density estimator. While nonparametric kernel estimation has been shown to be relatively insensitive to the choice of kernel function, the choice of bandwidth is known to drive resulting behaviour. One may not be concerned if they are simply using these methods for exploratory purposes; however, for sound analysis a selection method with known optimality properties must be adopted. Therefore, it is generally preferred that a fully-automated method that balances the squared bias and variance be selected. As in Li and Racine (2007, p.157), one could adopt a least-squares cross-validation approach for selecting smoothing parameters. The following

criterion is based on a weighted integrated square error:

$$ISE = \sum_{x^d} \int \{\hat{g}(y|x) - g(y|x)\}^2 f(x) M(x^c) dx^c dy \tag{3.6}$$

$$= I_{1n} - 2I_{2n} + I_{3n}, \tag{3.7}$$

where $M(\cdot)$ is a weight function,

$$I_{1n} = \sum_{x^d} \int \hat{g}^2(y|x) f(x) M(x^c) dx^c dy, \; I_{2n} = \sum_{x^d} \int \hat{g}(y|x) f(z) M(x^c) dx^c dy,$$

and $I_{3n} = \sum_{y^d} \sum_{x^d} \int g^2(y|x) f(x) M(x^c) dx^c dy^c$ does not depend on the smoothing parameters used to compute $\hat{f}$. Therefore, I_{1n} and I_{2n} are the items of interest. Note that these quantities can be written in expectation notation:

$$I_{2n} = E_z \left[\frac{\hat{f}(z) M(x^c)}{\hat{f}(x)} \right], \tag{3.8}$$

$$I_{1n} = E_X \left[\sum_{y^d} \int \hat{f}^2(z) \frac{f(x)}{\hat{f}^2(x)} M(x^c) dy^c \right], \tag{3.9}$$

where the expectation in (3.9) is taken with respect to x, not the random observations $\{X_i, Y_i\}_{i=1}^N$ and the expectation in (3.8) is taken with respect to $z = (x, y)$. The cross-validation objective function is derived using approximations for I_{1n} and I_{2n}, $\hat{I}_{1n}$ and $\hat{I}_{2n}$, respectively:

$$\hat{I}_{2n} = \sum_{i \in \mathcal{S}} \frac{\hat{f}_{-i}(z_i) M(x_i^c)}{\hat{f}_{-i}(x_i)}, \tag{3.10}$$

$$\hat{I}_{1n} = \sum_{i \in \mathcal{S}} \sum_{y^d} \int \hat{f}_{-i}^2(z_i) \frac{1}{\hat{f}_{-i}^2(x_i)} M(x_i^c) dy^c, \tag{3.11}$$

where the subscript $-i$ denotes the leave-one-out kernel estimator and

$$\sum_{y^d}\int \hat{f}^2_{-i}(z_i)dy^c = \frac{1}{(N-1)^2}\sum_{i_1=1,i_1\neq i}^{N}\sum_{i_2=1,i_2\neq i}^{N}\pi_{i_1}^{-1}\pi_{i_2}^{-1}\mathbf{1}(i_1\in\mathcal{S})\mathbf{1}(i_2\in\mathcal{S})K(x_i,x_{i_1})K(x_i,x_{i_2}) \times\sum_{y^d}\int K(y_i,y_{i_1})K(y_i,y_{i_2})dy^c. \tag{3.12}$$

Therefore, one chooses $(h,\lambda)=(h_{y,1},...,h_{y,q_y},h_{x,1},...,h_{x,q_x},\lambda_{y,1},...,\lambda_{y,r_y},\lambda_{x,1},...,\lambda_{x,r_x})$ to minimize the cross-validation objective function defined as:

$$CV(h,y)=\hat{I}_{1n}-2\hat{I}_{2n}. \tag{3.13}$$

The following describe smoothing parameters that, in asymptotic terms, are optimal for minimizing the mean integrated squared error (MISE) defined by taking the expected value of 3.6:

$$\text{MISE}(h,\lambda)=\sum_{y^d}\sum_{x^d}\int E_C\{\hat{g}(y|x)-g(y|x)\}^2 f(x)M(x^c)dx^c dy. \tag{3.14}$$

Using results from the bias and variance from section 3.1, the leading term of $CV(h,\lambda)$ is:

$$\begin{aligned} CV_0 =& \sum_{z^d}\int\left(\left[\sum_{s=1}^{q}h_sB_{1s}+\sum_{s=1}^{r}\lambda_sB_{2s}\right]^2+(\Delta+Q)\frac{\kappa^q g(y|x)}{nh_1...h_q}\right)\frac{M(x^c)}{f(x)}dz^c, \\ =& n^{-q/(q+4)}\chi_g(a,b) \end{aligned} \tag{3.15}$$

with

$$\chi_g(a,b)=\sum_{z^d}\int\left(\left[\sum_{s=1}^{q}a_sB_{1s}+\sum_{s=1}^{r}b_sB_{2s}\right]^2+(\Delta+Q)\frac{\kappa^q g(y|x)}{a_1...a_q}\right)\frac{M(x^c)}{f(x)}dz^c,$$

such that the a_s's and b_s's are defined as $h_s=n^{-1/(q+5)}a_s$ and $\lambda_s=n^{-2/(q+5)}b_s$.

Note that the objective function (3.13) involves three summations. This means that least squares cross-validation, in the context of conditional PDFs, is computationally costly. As there are no rule-of-thumb bandwidth selection algorithms for discrete variables in kernel estimation methods, plug-in bandwidths are not an option for the mixed data case. To alleviate the computational burden of bandwidth selection, one can apply likelihood cross-validation. The likelihood cross-validation

method involves choosing (h, λ) by maximizing the log-likelihood function

$$\mathcal{L} = \sum_{i=1}^{n} \ln \hat{g}_{-i}(y_i|x_i) = \sum_{i=1}^{n} \ln \frac{\hat{f}_{-i}(z_i)}{\hat{f}_{-i}(x_i)}. \tag{3.16}$$

As before, $\hat{f}_{-i}(z_i)$ and $\hat{f}_{-i}(x_i)$ are the leave-one-out kernel estimators of $f(y, x)$ and $f(x)$, respectively. The objective function (3.16) involves one less summation than that for least squares cross-validation, reducing the computational requirements for selecting $\gamma = (h, \lambda)$. It is for this reason I apply likelihood cross-validation in my Monte Carlo simulations below.

4 Simulations

In this section I compare the finite-sample performance of the probability weighted nonparametric conditional PDF estimator to other nonparametric and parametric estimators. I assess the performance of each estimator under different sampling plans by computing the MSE for each Monte Carlo replication and plot its distribution. Each Monte Carlo replication follows six steps:

1. I set $N = 10,000$ and I generate two variables $y \sim \mathcal{N}(0, 1)$ and $x \sim \mathcal{N}(0, 1)$ that follow a multivariate normal distribution, i.e $z = (y, x) \sim \mathcal{N}(\mu, \mathbf{\Sigma})$, with $\mu = (0, 0)$ and $\Sigma = \begin{bmatrix} 1 & \sigma_{yx} \\ \sigma_{yx} & 1 \end{bmatrix}$;
2. I then compute the joint PDF, $f(y, x)$, and the conditional PDF, $f(y|x)$. Letting $\pi_0 = 3.1415$, and $|A|$ denote the determinant of any matrix A, the joint PDF $f(y, x)$ is given by:

$$f(y, x) = (2\pi_0)^{-1} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)};$$

3. Next, I compute the conditional CDF $F(Y \leq y|X = x) = \int_{-\infty}^{y} f(t, x)/f(x)dt$ and generate a variable y^* which takes a value of 1 if $F(Y \leq y|X = x) > 0.5$ and 0 otherwise. A typical sample of $F(y|x)$ is reported in Figure 1;
4. I take a sample of size n based on a specified sampling plan. The sampling schemes I consider are SRS, stratified sampling on the y^* variable, and stratified sampling on the x variable. The

strata borders are listed in Table 1;[1]

5. I then estimate the conditional PDF $f(y|x)$ using the probability weighted kernel estimator given in (3.1), the kernel estimator from Hall, Li, and Racine (2004), a weighted Probit, and unweighted Probit, denoted by WKPDF, KPDF, WP, and UP, respectively:

$$\text{WKPDF} = \frac{\sum_{i=1}^{n} \pi_i^{-1} K_{\gamma,iz}}{\sum_{i=1}^{n} \pi_i^{-1} K_{\gamma,ix}},$$

$$\text{KPDF} = \frac{\sum_{i=1}^{n} K_{\gamma,iz}}{\sum_{i=1}^{n} K_{\gamma,ix}},$$

$$\text{WP}: \max_{\beta} \ln L = \sum_{i=1}^{n} \pi_i^{-1} \{y_i \ln \Phi(x_i'\beta) + (1-y_i) \ln[1-\Phi(x_i'\beta)]\}, \text{ and}$$

$$\text{UP}: \max_{\beta} \ln L = \sum_{i=1}^{n} \{y_i \ln \Phi(x_i'\beta) + (1-y_i) \ln[1-\Phi(x_i'\beta)]\}.$$

6. Finally, I compute the sample MSE using the following formula:

$$MSE(\hat{f}(y|x)) = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}(y_i|x_i)_i - f(y_i|x_i))^2, \tag{4.1}$$

where $\hat{f}(y|x)$ is one of WKPDF, KPDF, WP, or UP.

Bandwidths for WKPDF and KPDF are computed using the likelihood cross-validation methods from the previous section and Li and Racine (2007, p. 160), respectively. I set the number of Monte Carlo replications to 1000 and I vary the sample size $n = 200, 500, 1000$ and covariance term $\sigma_{yx} = 0.25, 0.50, 0.75$.

4.1 Simple Random Sampling

Results from SRS are presented in Table 2. Columns three to six present the median values of the mean squared errors derived from Monte Carlo simulations. The values in brackets are the

[1]The sample sizes taken from each strata are arbitrary; I recreated this exercise varying the sample sizes and obtained the same results.

median absolute deviation (MAD). The MAD is calculated by median($|MSE_m - \text{median}(MSE)|$) with $m = 1, ..., 1000$.

For all combinations of n and σ_{yx}, the median values for the MSE for the four estimators are very close. The difference between estimates is due to simulation noise. Under SRS, $\pi_i = \pi$ for all $i = 1, ..., n$. Therefore, WKPDF=KPDF and the results are as expected. The equivalence of the median MSE values for both nonparametric estimators and the Probit are encouraging as WKPDF and KPDF make no assumptions about the underlying DGP. Furthermore, as n increases and keeping σ_{yx} constant, the median MSE of WKPDF decreases, indicating it is a consistent estimator.

Figures 2 to 4 present the boxplots for MSE of WKPDF, KPDF, and WP for all combinations of n and σ_{yx}. While the median values are comparable and stable across estimators, the MSEs for the Probit specification are more variable. This can also be seen in Table 2, as the MADs for the Probit are larger than that for the nonparametric estimators. As σ_{yx} increases, the MADs decrease.

4.2 Stratification on Binary Outcome Variable

Table 3 presents the results from simulation for the $y*$ variable, where $y*$ is the binary transformation of y. In this case, sampling endogeneity is present in the model. Comparing WKPDF to KPDF, it is clear that the weighted estimator outperforms the unweighted estimator for all combinations of n and σ_{yx}. The median values of the MSE for WKPDF are smaller than those for KPDF, and with smaller MADs show the results for the weighted estimator are less variable. Figures 5 to 7 show the boxplots of the MSEs for the three selected estimators. There is a clear preference for weighted estimators as there is a stochastic dominance relationship evident for the weighted estimators over the non-weighted estimators along with less variability of the former.

The results for the WKPDF estimator and the WP estimator are quite striking. Since the underlying link function is the Gaussian CDF, the Probit model is perfectly specified. However, with no assumptions about the underlying DGP, WKPDF performs as well as the WP model, with comparable median and MAD values for the MSEs.

Another interesting result is that under endogenous sampling, the unweighted kernel conditional PDF estimator, KPDF, is inconsistent. Keeping σ_{yx} constant, as n increases the median and mean MSEs increase. Looking at the 95 percent confidence intervals in Table 4, the lower bounds for

$n = 1000$ are larger than the upper bounds for $n = 200$ and $n = 500$ (highlighted in bold in columns five and six). The weighted kernel estimator, however, is consistent. The median values are decreasing incrementally and the 95% confidence intervals do not intersect. These results reinforce the importance of using weighted estimators when sampling endogeneity is present in the model.

4.3 Stratification on Exogenous Predictor Variable

The results for sampling on the x variable presented in Table 5 are similar to those found for SRS (Table 2). The median results are similar across estimators while the MADs for the nonparametric estimators are smaller than the MADs for the Probit estimators. The difference between the median values at the fourth decimal place are due to simulation noise.

These results are similar to that of Clair (2015), where inverse probability weights do not come with a cost in terms of increased MSE or variability of the MSE when weighting is unnecessary (under SRS and exogenous sampling) but may provide a large benefit by reducing MSE under endogenous sampling. This result is encouraging as it is recommended to weight observations when the objective is to estimate population totals under unequal probability sampling.

5 Application: Private Prescription Drug Insurance and the Use of Mental Health Pharmaceuticals

Mental health issues have transitioned from being a private matter to becoming a matter of public policy. Bringing these issues into public light is the first step to changing the stigma of mental illness that exacerbates the effects of the illness on the individual sufferer. Furthermore, these discussions have brought forth the true impacts of mental health on society. The Mental Health Commission of Canada (MHCC) approximates that one in five Canadians experiences a mental health problem yearly, costing the economy over $50 billion (MHCC 2012). Due to the early onset of many of these conditions, the economic impacts of mental illnesses are felt long-term. The global acceptance of mental illness as a public health issue has helped to raise awareness; however, policy makers require evidence that investing in mental health is worthwhile (WHO 2006).

Psychological wellbeing influences many aspects of an individual's life. A person's ability to study, work, and make daily decisions depends on his or her mental health. The largest impact of

mental illness is being felt in the workplace as the majority of mental illness sufferers are among working aged people (MHCC 2012). Mental health problems lead to absenteeism, presenteeism, and turnover (Andres 2004); typically, close to one-third of long-term disability claims are for mental health reasons. The 'World Health Organization Mental Health Declaration and Action Plan for Europe' specifies improving mental health as a determinant for the social and economic prosperity of Europe (Barry 2009).

Individuals with mental disorders create pressure on society to provide a range of health care and welfare services. Lim et al. (2008) showed people living with a mental illness utilized more physician visits, specialist visits, and hospital days, on average, compared to those without a mental illness. Individuals exhibiting symptoms of depression are at a higher risk of hospital admission for non-psychiatric conditions and are more likely to have longer hospital stays with worse hospital outcomes, compared with non-depressed patients (Prina et al. 2013).

The heterogeneous nature of mental illness leads to individual treatment plans that may consist of community-based non-physician mental health services (e.g. psychologists and social workers) and prescription medications. While Canada's universal system of public health insurance fully covers the cost of medically necessary hospital and physician services, the public plan generally does not cover the use of prescription drugs. In this case, even if physician services are fully insured, the fact that complementary services are not may compromise access to necessary medical services. Despite the risk of *ex-ante* and *ex-post* moral hazard, supplementary insurance has been associated with increased preventative care and reduces ambulatory care sensitive hospital admissions (ACSC)[2] (Devlin, Sarma, and Zhang, 2011).

The majority of Canadians who hold supplementary health insurance receive their coverage through employer-provided insurance plans. Approximately 61 percent of supplemental insurance holdings in Canada come from employers, 26 percent from public plans, 7 percent from private insurance, and 6 percent hold a combination of the three types (Statistics Canada 2014). Employer insurance is part of a benefit plan that may include coverage of non-physician health care services, where health care fees are subject to cost-sharing. Public insurance plans are typically available for vulnerable groups in the population including senior citizens or recipients of other forms of social

[2]ASCS's are conditions that, given the appropriate treatment, are preventable and do not require hospitalization. E.g. hypertension.

assistance. Private supplemental insurance is an option for those who can afford to pay premiums. Therefore, in Canada, the distribution of supplemental insurance for prescription drugs tends to be skewed to higher-income, employed Canadians, leading some to question the universality of the public insurance plan to cover important complementary health care services.

The purpose of this application is to investigate the role of supplementary insurance plans in the utilization of prescription drugs for mental illness using the probability weighted nonparametric conditional density estimator (WKPDF) described above. I use data from the 2012 Canadian Community Health Survey Mental Health Component (CCHS-MH) public use file, which contains information on health status, health care utilization, socioeconomic status, and an individual's social support system.

5.1 Data and Sampling Method

The CCHS-MH was designed to provide a detailed look at mental health with respect to who is affected by selected mental health disorders as well as positive mental health (Statistics Canada 2013). Individuals aged fifteen years and older living in one of the ten provinces were selected based on a complex, two-staged stratified design with each stratum formed of clusters. In the first stage, a group of clusters is selected according to a sampling method with a probability proportional to size. In the second stage, a list of households is collected for each cluster. A sample of households is then chosen with only one individual per household (selected at random) responding to the survey. The sample weights included in the survey not only reflect the sampling plan, but also non-response of individuals. The total number of selected households was 36,443. The overall household and person response rate was 68.9 percent resulting in a total of 25,113 respondents.

5.2 Methodology

To examine the effect of supplementary insurance on prescription drug usage, utilization of medication is modelled as a function of insurance status, controlling for the individual's health status, socioeconomic status, demographic characteristics, and social support system. I estimate these models using the WKPDF, KPDF, and Logit estimators and compare results. The smoothing parameters for WKPDF and KPDF are selected using their respective maximum likelihood cross-validation methods.

5.3 Variables

5.3.1 Prescription Pharmaceutical Utilization

I look at the utilization of four categories of prescription pharmaceuticals for mental illness: any medication, anti-depressants, anti-psychotics, and benzodiazepines (anxiety medications), denoted by Y_{any}, Y_{adep}, Y_{apsy}, and Y_{benzo}, respectively. Y_{any} contains information on the usage of the latter three categories, other types of anxiety medications, and medications for alcohol and illicit drug use. In each case, utilization is measured dichotomously as use/no-use in the previous two days. Y_i takes a value of one if the individual has taken medication i in the last two days and two otherwise, $i = \{any, adep, apsy, benzo\}$.

5.3.2 Independent Variables

Insurance coverage is measured by a binary variable taking a value of one if the individual has insurance that pays for all or part of prescription medications and two otherwise. The socioeconomic variables I consider for this application are age, gender, education, and income. Age is a grouped variable with fourteen categories ranging from '15 to 19 years' to '80 years or more'. Gender is equal to one if the respondent is male and two if they are female. Education is measured as the highest level of education attained by the respondent based on a four point scale and income is a discretized variable measuring the total income of the household. Finally, I include one variable measuring the individual's self-assessed health (SAH) and one variable measuring their social support system. Social support is measured by the social provision scale (SPS), which ranges from 10 to 40, where a score of 40 means the individual has a strong social support system. Variable category descriptions are given in Table 6.

Y_{any}, Y_{adep}, Y_{apsy}, Y_{benzo}, insurance status, and gender are unordered binary discrete variables while the variables representing education, household income, and SAH are ordered discrete variables. Age and social provision score are treated as continuous variables. Let the superscripts d, (d,u), (d,o), and c denote discrete, unordered discrete, ordered discrete, and continuous variables, respectively. Then $X = \{X^d, X^c\}$ where $X^d = \{X^{(d,u)}, X^{(d,o)}\}$, $X^{(d,u)} = \{\text{insurance, gender}\}$, $X^{(d,o)} = \{\text{education, income, SAH}\}$, and $X^c = \{\text{age, SPS}\}$. For the nonparametric analysis that follows, I use the second order Gaussian kernel for continuous variables and the variation of the

Aitchison and Aiken (1979) kernel in (2.5) for unordered discrete variables. For the ordered discrete variables, I use the Wang and van Ryzin (1981) kernel given by $l(x_{is}^{d,o}, x_s^{d,o}, \lambda_s^o) = 1 - \lambda_s^o$ if $|x_{is}^{d,o} - x_s^{d,o}| = 0$ and $((1-\lambda_s^o)/2)(\lambda_s^o)^{|x_{is}^{d,o} - x_s^{d,o}|}$ if $|x_{is}^{d,o} - x_s^{d,o}| \geq 1$. $0 \leq \lambda_s^o \leq 1$ is the smoothing parameter for the ordered discrete variable s, s={education, income, SAH}.

5.4 Results

5.4.1 Descriptive Findings

Given that the sample is community-based, I observe generally low levels of the use of mental health medications (Table 6). 6.7 percent of the sample reported taking 'any medication' for mental illness, drug use, and/or alcohol use. The type of medication used most frequently was anti-depressants with 5.3 percent of respondents having taken at least one in the previous two days. 1.4 percent reported taking benzodiazepines and 1 percent reported using anti-psychotics. Approximately 78 percent of individuals surveyed claimed to have insurance that covered all or part of their prescription drug costs. The median value for age is 7 (45 to 49 years old), with a median absolute deviation of 3. There are approximately 1.4 percent (2347) more females than males in the sample. The majority of respondents hold a post-secondary degree or diploma at 56.5 percent and 55.6 percent earn a household income of $60,000 or more. 89.8 percent of those surveyed reported to be in good or better health with a median SAH of 2 (very good) and MAD of 1. The majority of respondents reported having a strong social support system with median social provision scale of 37 (max 40) and MAD of 3.

5.4.2 Estimated Conditional Probabilities

Table 7 presents the bandwidths computed for WKPDF and KPDF using maximum likelihood cross-validation. Insurance, gender, and SAH are relevant variables in the estimation of utilization of all types of medication as the bandwidths lie between zero and one for these three discrete variables. Income is irrelevant in the estimation of $Pr(Y_{adep}|X)$ using WKPDF, as the smoothing parameter equals one and each observation carries the same weight. Similarly, education is irrelevant in the estimation of $Pr(Y_{apsy}|X)$ and $Pr(Y_{benzo}|X)$ using KPDF and $Pr(Y_{any}|X)$ using both estimators.

Estimated probabilities of medication use conditional on X versus X^d are presented in Table 8

with the remaining predictors held constant at their medians/modes. The values in the table are the predicted $Pr(Y = 1|X)$, i.e the conditional probability that an individual used the medication within the last two days (presented as percents) and the percentage change (denoted $\%\Delta$) from base categories. Results for 'any medication', anti-depressants, anti-psychotics, and benzodiazepines are listed in columns (1) to (6), (7) to (12), (13) to (18), and (19) to (24), respectively. Results from weighted nonparametric estimation show that the probability of consuming 'any medication' for mental illness is 9.19 percent for those who have insurance compared to 7.61 percent for those who do not. The low probability of use simply reflects the low baseline prevalence of use. In relative terms, a person with drug insurance is nearly 21 percent more likely to use medication for mental illness. This result is higher than for unweighted nonparametric estimation (11.05 percent) and more conservative than Logit estimation (65.95 percent). Similar results are found for anti-depressants as the three models estimate that having insurance increases the likelihood of using this type of drug at 24.54 percent, 19.46 percent, and 67.19 percent for WKPDF, KPDF, and Logit, respectively. The predicted probability that someone uses anti-psychotics is low at 1.15 percent for those who have insurance. Relative to uninsured individuals, those who have insurance are 69 percent more likely to use anti-psychotics. While this relative measure is high, it is lower than those derived from KPDF estimation (155.5 percent) and logistic regression (130.5 percent). For benzodiazepines, both WKPDF and Logit estimate a positive impact of insurance, while KPDF estimates a negative relationship.

Females had a higher estimated probability of use for all drug types. WKPDF estimated higher percent changes for 'any medication' and anti-depressants compared to KPDF and Logit estimates. Both nonparametric estimators reported small changes in the use of 'any medication' for mental illness across education levels relative to the base category 'less than secondary'. However, results from logistic regression show a positive relationship between probability of use and education level. Results for all three estimators show that individuals holding a post-secondary degree are more likely to use anti-psychotics than those without a high school diploma. Individuals in the lowest income cohort ($0 to $19,999) have a higher probability of using all categories of medication compared to individuals in higher income brackets. Relative changes in these groups compared to the baseline category ($0 to $19,999) are smaller for WKPDF. SAH was found to have a clear gradient for all drug categories, with worsening SAH associated with higher likelihood of use compared to 'excellent'

SAH.

Figures 11 to 14 present the predicted $Pr(Y = 1|X)$ versus age group for 'any medication', anti-depressants, anti-psychotics, and benzodiazepines, respectively, with all other covariates held constant at their medians/modes. In each figure, the red dashed line represents the estimates from WKPDF, the solid black line represents the estimates from KPDF, and the dotted blue line represents estimates from the logistic regression. In Figure 11, both WKPDF and KPDF estimate a positive relationship between age and probability of use of 'any medication' between age groups 2 (15 to 19 years old) and 7 (45 to 49 years old). While KPDF estimated probabilities begin to decrease, WKPDF estimates continue to increase. Results from logistic regression simply return a negative relationship between age group and the probability of use, reporting a probability of use approximately four percentage points higher than both nonparametric estimators for the '15 to 19 years old' cohort. A similar pattern is observed for anti-depressant use in Figure 12. For anti-psychotics, WKPDF and KPDF show a positive relationship between age and probability of use up to age group 5 (35 to 39 years old), a negative relationship between groups 5 and 10 (60 to 64 years old), and an increase between age groups 10 and 14 (80 years or more) (Figure 13). However, the estimated probabilities for WKPDF are higher than estimates from KPDF, with WKPDF reporting that individuals aged 80 and over are 547 percent more likely to use anti-psychotics than individuals aged 60 to 64 years old. Again, the Logit estimator reports a near-linear negative relationship. Nonparametric estimation shows that the probability of benzodiazepine use conditional on age increases with age group until group 10 for WKPDF estimates and group 11 (65 to 69 years) for KPDF estimates, then decreases for both. Logistic regression reports a positive relationship between age group and probability of use.

Figures 15 to 18 present the predicted $Pr(Y = 1|X)$ versus SPS for 'any medication', anti-depressants, anti-psychotics, and benzodiazepines, respectively, with all other covariates held constant at their medians/modes. In figures 15, 16, and 18 KPDF estimates are negatively related to social provision score, with a steep decrease between SPS scores of 10 and 18. Estimates from WKPDF increase between SPS scores of 10 and 15, then begin to decrease between scores of 15 and 20. These figures also show that WKPDF estimates have a second peak at SPS scores of 24, then decrease to similar levels of KPDF. Figure 17 shows an increase in predicted probabilities estimated by WKPDF between SPS scores of 10 and 12, then show a steep decrease between scores of 12 and

20 and then flattening. KPDF estimates are relatively flat compared to WKPDF estimates.

Table 9 looks at relative differences in insurance status by discrete covariates. The values in the table are the percent changes in insurance status for each category in X_i^d, with i = gender, education, income, and S Results from WKPDF estimation show that insured males are 7.3 percent less likely to use 'any medication' for mental illness than uninsured males, while insured females are 35.7 percent more likely to use these medications than uninsured females. Similarly, insured females are more likely than uninsured females to use anti-depressants, anti-psychotics, and benzodiazepines. WKPDF estimates show that insured individuals who have a total household income between $0 and $19,999 are 56.2 percent more likely to use 'any medication' than uninsured individuals in the same income cohort, while those with a household income of $80,000 or more are 31.6 percent more likely to use 'any medication'. Logit estimates report a positive relationship between income level and percent change in insurance status. Furthermore, insured individuals from households in the lowest income cohort are 428.6 percent more likely to use anti-psychotics than individuals in the same income cohort without insurance based on WKPDF estimates. Insured individuals in the highest income group are 44.1 percent more likely to use anti-psychotics than uninsured individuals in households earning $80,000 or more. Note, the percent changes derived from logistic regression are nearly equal for all income levels for consumption of anti-psychotics. Individuals who reported being in excellent health are more likely to use all types of medication if they are insured.

Figures 19 to 22 display the predicted $Pr(Y = 1|X)$ from WKPDF estimation versus age group for both insured and uninsured individuals. In each figure, the red dashed line is the estimated $Pr(Y = 1|X)$ for insured individuals and the blue dotted line is the estimated probability for uninsured individuals. In Figure 19, both curves rise between age groups 1 (15 to 19 years) and 8 (50 to 54 years) and decrease from age groups 8 to 10 (60 to 64 years), with the red (insured) curve lying above the blue (uninsured) curve. While the estimated probabilities for uninsured individuals continue to decrease, the probabilities of use for insured individuals rise from age group 10 to 14 (80 years or more). Figure 20 displays a similar relationship as the gap between the two curves widens as individuals get older. As in Figure 13, the probabilities of using anti-psychotics in Figure 21 increase from age group 1 to 5 (35 to 39 years), decrease from 5 to 10, and increase from 10 to 14. The red curve lies above the blue curve indicating that insured individuals are more likely to use anti-psychotics than uninsured individuals. Again, there is a widening gap between the two curves;

insured individuals aged 80 years or more are 300 percent more likely to use anti-psychotics than uninsured individuals in the same age group compared to a difference of 150 percent for individuals aged 60 to 64 years old. Figure 22 displays the predicted $Pr(Y_{benzo} = 1|X)$ versus age group. Both curves display the same pattern: an increase in the probability of using benzodiazepines between age groups '15 to 19 years' and '60 to 64 years', and a decrease between age groups '60 to 64 years' and '80 years and over'. The probabilities of using benzodiazepines are higher for insured individuals for each age group.

Finally, Figures 23 to 26 present the predicted $Pr(Y = 1|X)$ from WKPDF estimation versus social provision score for both insured and uninsured individuals. In Figure 23, the probabilities of consuming 'any medication' for mental illness are higher than those for uninsured individuals for overall social provision scores between 10 and 16. Uninsured individuals actually have a higher predicted probability of usage for SPS scores between 18 and 21. Similar results are found for anti-depressants in Figure 24. The probabilities of using anti-psychotics and benzodiazepines for individuals with insurance lie above those who do not in Figures 25 and 26, respectively.

5.5 Discussion

WKPDF VS KPDF VS Logit These results highlight two important points: the flexibility of nonparametric estimators and the effect of weighting observations in conditional PDF estimation. Figures 11 to 18 best display the benefit of allowing for flexibility in the DGP as the nonparametric estimators pick up variations that would otherwise go unobserved using a linear model. For income, the percentage change estimated in insurance status on the use of any medication, using either nonparametric estimator, is higher for low income households than it is for high income households. Results from logistic regression show a reverse relationship where the percentage change is lower for low income households than for higher income households.

Using a weighted estimator helped to mitigate or enhance the effects of the predictor variables on medication use to more accurately reflect usage in the population. The percentage change in insurance status on the usage of 'any medication' was 20.88 percent for WKPDF and 11.05 percent for KPDF. The percentage change in insurance status on the usage of anti-psychotics was 69.02 percent for WKPDF and 115.51 percent for KPDF. Therefore, using the weighted estimator increased the impact of insurance on using 'any medication' and lessened the impact of insurance on

anti-psychotic medication usage that would have been obtained from using an unweighted estimator.

Insurance and Prescription Medication Usage Results from WKPDF estimation in Tables 8 and 9 and Figures 19 to 26 show that insurance has a positive effect on the use of all medication types. Positive values in Table 9 mean that insured individuals within the same category are more likely to use the selected prescription medication. The highest percentage change in insurance status within income categories was observed in the lowest income group. That is, insurance status has a higher impact on prescription pharmaceutical utilization for individuals living in households earning between \$0 and \$19,999, than it does on households earning \$80,000 or more. This is true for all medication types. Also, the fact that the probabilities of use for insured individuals lie above the curves for uninsured individuals in Figures 19 to 22 imply that insurance has a positive effect on drug usage. The widening gap between the curves in 19 and 20 suggest that insurance status has a greater impact on usage of 'any medication' and anti-depressants as individuals get older.

6 Conclusion

This paper introduced an inverse probability weighted nonparametric kernel conditional density estimator for estimating models with multivariate outcome variables. This estimator addresses two issues with traditional parametric methods for estimating discrete choice models. First, it relaxes the functional form assumption of the underlying DGP. If the link function in the parametric estimation of a binary choice model is incorrectly specified, the estimator will be inconsistent. By estimating the conditional probability density function nonparametrically, one allows for the data to "speak for itself". The second issue the proposed estimator solves is the issue of endogenous sampling. If the error term is correlated with the sampling criterion, estimates will be inconsistent. The parametric and semiparametric solutions are to weight the log-likelihood function by inverse probability weights. These methods, however, are indirect. By weighting a nonparametric conditional density estimator, I developed a direct method for estimating these conditional probabilities while correcting for endogenous sampling.

Simulation results provided further evidence that weighting is important when endogenous sampling is present and the probabilities of inclusion are unequal across observations. Estimation under

stratified sampling on the outcome variable showed that the probability weighted nonparametric estimator outperformed unweighted estimators based on the MSE criterion and performed as well as the correctly specified weighted parametric estimator. Under SRS, the proposed estimator reduces to the estimator from Hall, Li, and Racine (2004) and weighting has no effect. A similar result was found when the sample design was based on stratification of the predictor variable, i.e. when sampling is exogenous. In this case, results were similar for both nonparametric estimators. As there is no loss of efficiency when using the model-assisted conditional density estimator and potentially lower MSEs when sampling is endogenous, I recommend using this estimator when working with survey data collected via unequal sampling methods.

References

Bellhouse, D. & Stafford, J. (1999), 'Density estimation from complex surveys', *Statistica Sinica* **9**, 407–424.

Breidt, F. & Opsomer, J. (2000), 'Local polynomial regression estimators in survey sampling', *The annals of Statistics* (28), 1026–1053.

Breunig, R. V. (2001), 'Density estimation for clustered data', *Econometric Reviews* **20**(3), 353–367.

Breunig, R. V. (2008), 'Nonparametric density estimation for stratified samples', *Statistics and Probability Letters* **78**, 2194–2200.

Buskirk, T. D. & Lohr, S. L. (2005), 'Asymptotic properties of kernel density estimation with complex survey data', *Journal of Statistical Planning and Inference* (128), 165.

Cameron, C. & Trivedi, P. (2009), *Microeconometrics: Methods and applications*, Cambridge University Press, 32 Avenue of the Americas, New York, NY 10013-2473, USA.

Clair, L. (2015), Local constant regression with mixed data types in complex survey data, PhD thesis, McMaster University, 1280 Main St. West, Hamilton Ontario.

Greene, W. H. (2012), *Econometric Analysis: Seventh Edition*, Prentice Hall, Saddle River, NJ.

Klein, R. W. & Spady, R. (1993), 'An efficient semiparametric estimator for binary response models', *Econometrica* **61**, 387–421.

Li, T. & Yang, H. (2016), 'Inverse probability weighted estimators for single-index models with missing covariates', *Communications in Statistics-theory and Methods* **45**(5), 1199–1214.

Magee, L., Robb, A. & Burbidge, J. (1998), 'On the use of sampling weights when estimating regression models with survey data', *Journal of Econometrics* **84**, 251–271.

McLeod, L. (2011), 'A nonparametric vs. latent class model of general practitioner utilization: Evidence from canada', *Journal of Health Economics* **30**(6), 1261–1279.

Racine, J. S., Li, Q. & Zhu, X. (2004), 'Kernel estimation of multivariate conditional distributions', *Annals of Economics and Finance* **5**, 211–235.

Sánchez-Borrego, I., Opsomer, J., Rueda, M. & Arcos, A. (2014), 'Nonparametric estimation with mixed data types in survey sampling', *Rev Mat Complut* (27), 685–700.

Silverman, B. (1986), *Density Estimation for Statistics and Data analysis*, Chapman and Hall, London, England.

Solon, G., Haider, S. J. & Wooldridge, J. (2013), What are we weighting for?, Working Paper 18859, National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138.

7 Proofs

7.1 Proof of Theorem 3.1

The following proof follows similar steps as in Li and Racine (2007, p.157). To simplify notation, I use (s.o.) to denote terms of smaller orders, or terms independent of γ.

To derive the asymptotic MSE of $\hat{g}(y|x)$, first look at the difference $\hat{g}(y|x) - g(y|x)$:

$$\begin{aligned} \hat{g}(y|x) - g(y|x) &= \frac{[\hat{g}(y|x) - g(y|x)]\hat{f}(y|x)}{\hat{f(y|x)}} \\ &= \frac{\hat{m}(y,x)}{\hat{f}(x)} \end{aligned} \tag{7.1}$$

where $\hat{m}(y,x) = [\hat{g}(y|x) - g(y|x)]\hat{f}(y|x) = \hat{f}(y,x) - g(y|x)\hat{f}(x)$. Denoting $z = (y,x)$ such that $f(z) = f(y,x)$, take the expectation under the combined inference method of the numerator in (7.1):

$$E_C(\hat{m}(y|x)) = E_C(\hat{f}(z)) - g(y|x)E_C(\hat{f}(x)). \tag{7.2}$$

It is easiest to look at the two expectations on the right hand side of (7.2) separately. First, look

at $E[\hat{f}(z)]$:

$$
\begin{aligned}
E_C(\hat{f}(z)) &= E_\xi\{E_P(\hat{f}(z))|\pi)|z\} = E_M\left\{E_P\left(\sum_{i\in S}\pi^{-1}K_{h,iz}\right)\Big|z\right\} \\
&= E_M\left\{E_P\left(N^{-1}\sum_{i=1}^{N}\pi^{-1}\mathbf{1}(i\in\mathcal{S})K_{h,iz}\right)\Big|z\right\} \\
&= E_M\left\{N^{-1}\sum_{i=1}^{N}\pi^{-1}E(\mathbf{1}(i\in\mathcal{S})|\pi)K_{h,iz}\Big|z\right\} \\
&= E_M\left(N^{-1}\sum_{i=1}^{N}\pi_i^{-1}\pi_i K_{h,iz}\right) \\
&= E_M\left(N^{-1}\sum_{i=1}^{N}K_{h,iz}\right) \\
&= E_M(K_{\gamma,z}) \text{ (because of i.i.d. in the population)} \\
&= \sum_{t^d\in\mathcal{D}}\int_{\mathbf{R}^q}\prod_{s=1}^{q}h_s^{-1}w\left(\frac{t_s-z_s}{h_s}\right)\prod_{s=1}^{r}\lambda_s^{\mathbf{1}(t_s\neq z_s)}f(t^c,t^d)dt^c \\
&= \int_{\mathbf{R}^q}\prod_{s=1}^{q}w(v_s)f(z^c+hv_s,z^d)dv_s + \sum_{t^d\in\mathcal{D}}\mathbf{1}(t^d,z^d)\lambda_s\int_{\mathbf{R}^q}\prod_{s=1}^{q}w(v_s)f(z^c+hv_s,t^d)dv_s \\
&= f(z)+\frac{\kappa_2}{2}\sum_{s=1}^{q}h_s^2 f_{ss}(z)+\sum_{s=1}^{r}\sum_{t^d\in\mathcal{D}}\mathbf{1}(t^d,z^d)f(z^c,t^d)\lambda_s + O\left(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s\right), \qquad (7.3)
\end{aligned}
$$

where $K_{\gamma z}=\prod_{s=1}^{q}h_s^{-1}k((t_s-z_s)/h_s)\prod_{s=1}^{r}\lambda_s^{\mathbf{1}(t_s\neq z_s)}f(t^c,t^d)$ and $\mathbf{1}_s(z^d,t^d)=\mathbf{1}_s(z_s^d\neq t_s^d)\prod_{s'\neq s}^{r}\mathbf{1}_s(z_s^d=t_s^d)$. Using a derivation similar to that in Li and Racine (2007, p. 157), it can be shown that:

$$
E[\hat{f}(x)] = f(x)+\frac{\kappa_2}{2}\sum_{s=q_y+1}^{q}h_s^2 f_{ss}(x)+\sum_{s=r_y+1}^{r}\sum_{u^d\in\mathcal{D}_x}\mathbf{1}(u^d,x^d)f(x^c,u^d)\lambda_s + O\left(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s\right). \qquad (7.4)
$$

Combining (7.1), (7.3), and (7.4), and using the fact that $\hat{f}(x)=f(x)+o_P(1)$, the bias of $\hat{g}(y|x)$ is given by:

$$
Bias[\hat{g}(y|x)] = \sum_{s=1}^{q}h_s^2 B_{1s}(z)+\sum_{s=1}^{r}\lambda_s^2 B_{1s}(z)+O\left(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s\right), \qquad (7.5)
$$

where

$$
B_{1s}(z)=\begin{cases}(1/2)\kappa_2 f_{ss}(y,x)/f(x) & \text{if } s=1,...,q_y \\ (1/2)\kappa_2[f_{ss}(y,x)-f_{ss}(x)g(y|x)]/f(x) & s=q_y+1,...,q\end{cases} \qquad (7.6)
$$

and

$$B_{2s}(z) = \begin{cases} \sum\limits_{t^d \in \mathcal{D}} \mathbf{1}(t^d, z^d) f(z^c, t^d) & \text{if } s = 1, ..., r_y \\ \sum\limits_{u^d \in \mathcal{D}_\S} \mathbf{1}(u^d, x^d)(f(z^c, y^d, u^d) - g(y|x) f(x^c, u^d)) & s = r_y + 1, ..., r \end{cases}. \tag{7.7}$$

Next, I need to derive the variance of $\hat{g}(y|x)$ under the combined inference famework. Considering again the numerator in (7.1), I use the formula for the variance in Harms and Duchesne (2010):

$$\text{var}_C(\hat{m}(y,x)) = \text{var}_\xi(E_P[\hat{m}(y,x)|\pi]|z) + E_\xi(\text{var}_P[\hat{m}(y,x)|\pi]|z). \tag{7.8}$$

It is best to derive $\text{var}_C(\hat{m}(y,x))$ by examining the items on the right hand side of (7.15) separately. First, look at $\text{var}_\xi(E_P[\hat{m}(y,x)|\pi]|z)$:

$$\begin{aligned} \text{var}_\xi(E_P[\hat{m}(y,x)|\pi]) &= \frac{1}{N^2}\text{var}_M\left(E_P\left[\left(\sum_{i=1}^N \pi_i \mathbf{1}(i \in \mathcal{S}) K_{\gamma z} - g(y|x)\sum_{i=1}^N \pi_i \mathbf{1}(i \in \mathcal{S}) K_{\gamma x}\right)|\pi\right]|z\right) \\ &= \frac{1}{N^2}\text{var}_M\left(\sum_{i=1}^N \left[\left(K_{\gamma i,z} - g(y|x)\right) K_{\gamma i,x}\right]\right) \\ &= \frac{1}{N}\{E_M[(K_{\gamma z} - g(y|x))K_{\gamma x}]^2 - [E_\xi((K_{\gamma z} - g(y|x))K_{\gamma x})]^2\}, \end{aligned} \tag{7.9}$$

where $K_{\gamma x} = \prod\limits_{s=1}^q h_s^{-1} k((u_s - x_s)/h_s) \prod\limits_{s=1}^r \lambda_s^{\mathbf{1}(t_s \neq x_s)} f(u^c, t^d)$. Isolating the second term on the right hand side of (7.9):

$$\begin{aligned} [E_\xi((K_{\gamma z} - g(y|x))K_{\gamma x})]^2 =& \frac{1}{N}[E_\xi(K_{\gamma z}) - g(y|x)E_\xi(K_{\gamma x})]^2 \\ =& \frac{1}{N}\left[\sum_{s=1}^q h_s^2 B_{1s}(z) + \sum_{s=1}^r \lambda_s^2 B_{1s}(z) + O\left(\sum_{s=1}^q h_s^2 + \sum_{s=1}^r \lambda_s\right)\right]^2 \qquad (7.10) \\ =& O\left(\left(\sum_{s=1}^q h_s^2\right)^2 + \left(\sum_{s=1}^r \lambda_s\right)^2\right). \qquad (7.11) \end{aligned}$$

The RHS of equation (7.9) then becomes:

$$
\begin{aligned}
&\frac{1}{N}E_M[(K_{\gamma y}-g(y|x))K_{\gamma x}]^2-[E_\xi((K_{\gamma y}-g(y|x))K_{\gamma x})]^2\\
&\qquad=\frac{1}{N}E_M\left[\prod_{s=1}^{q}\frac{1}{h_s}w\left(\frac{t_s^c-z_s^c}{h_s}\right)\prod_{s=1}^{r}\lambda_s^{\mathbf{1}(t_s^d\neq z_s^d)}-g(y|x)\prod_{s=1}^{q_x}\frac{1}{h_{x,s}}w\left(\frac{u_s^c-x_s^c}{h_{x,s}}\right)\prod_{s=1}^{r_x}\lambda_s^{\mathbf{1}(u_s^d\neq x_s^d)}\right]^2\\
&\qquad\quad+O\left(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s\right)\\
&\qquad=\frac{1}{N}E_\xi(K_{\gamma z}^2-[g(y|x)]^2K_{\gamma x}^2)+O\left(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s\right)\\
&\qquad=\frac{1}{N}\sum_{t^d\in D}\int K_{\gamma z}^2f(t^c,t^d)dt^c-[g(y|x)]^2\sum_{u^d\in D_x}\int K_{\gamma x}^2f(u^c,u^d)du^c+O\left(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s\right)\\
&\qquad=\frac{1}{N}\int_{\mathbf{R}^q}\prod_{s=1}^{q}\frac{1}{h_s}w^2(v_s)f(z^c+hv,z^d)dv_s+O\left(\left(N\prod_{s=1}^{q}h_s\right)^{-1}\left(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s\right)\right).
\end{aligned}
\tag{7.12}
$$

Therefore, the first element on the RHS of (7.15) is:

$$
\mathrm{var}_\xi(E_P[\hat{m}(y,x)|\pi]|z)=\frac{\kappa^q f(y,x)}{N\prod_{s=1}^{q}h_s}+(s.o.). \tag{7.13}
$$

Looking now at the second term on the RHS of (7.15):

$$\begin{aligned}
E_\xi\{\text{var}_P[\hat{m}(y,x)|\pi]|z\} =& E_M\left\{\frac{1}{N^2}\text{var}_P\left\{\sum_{i=1}^{N}[(K_{\gamma i,y}-g(y|x))\pi_i\mathbf{1}(i\in\mathcal{S})K_{\gamma i,x}]\Big|\pi\right\}\Big|z\right\}\\
=& E_M\left\{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(K_{\gamma i,y}-g(y|x))(K_{\gamma j,y}-g(y|x))\right.\\
&\left.\pi_i^{-1}\pi_j^{-1}\text{var}_P\{\mathbf{1}(i\in\mathcal{S})\}K_{\gamma i,x}K_{\gamma j,x}\right\}\\
=& \frac{1}{N^2}E_M\left\{\sum_{i=1}^{N}(K_{\gamma i,y}-g(y|x))^2\frac{1-\pi_i}{\pi_i}\right.\\
&\left.+\sum_{i=1}^{N}\sum_{j\neq i}(K_{\gamma i,y}-g(y|x))(K_{\gamma j,y}-g(y|x))\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}K_{\gamma i,x}K_{\gamma j,x}\right\}\\
=& \frac{1}{N^2}E_M\left\{\sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}K_{\gamma i,z}^2\right\}+(s.o.)\\
=& \frac{\kappa^q f(y,x)}{N^2\prod_{s=1}^{q}h_s}\sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}+(s.o.). \qquad (7.14)
\end{aligned}$$

Combining (7.13) and (7.14) gives the asymptotic pointwise variance of $\hat{m}(y,x)$:

$$\text{var}_C(\hat{m}(y,x)) = \frac{1}{nh_1...h_q}(\Delta+Q)\kappa^q f(y,x). \qquad (7.15)$$

Consider now the variance of the difference $\hat{g}(y|x)-g(y|x)$ under the combined framework:

$$\begin{aligned}
\text{var}_C(\hat{g}(y|x)-g(y|x)) &= \frac{1}{[f(x)]^2}\text{var}_C(\hat{m}(y,x))\\
&= \frac{1}{nh_1...h_q}(\Delta+Q)\kappa^q g(y|x)[f(x)]^{-1}. \qquad (7.16)
\end{aligned}$$

Equations (7.5) and (7.16) complete the proof.

7.2 Proof of Theorem 3.2

In order to prove the aysymptotic normality of $\hat{g}(x)$, I make use of the following theorem taken from the statistical appendix in Li and Racine (2007).

Theorem 7.1 (Liapunov Double Array Central Limit Theorem)**.** *Let $\{Z_{n,i}\}$ be a sequence of inde-*

pendent (double array) random variables with $E|Z_{n,i}|^{2+\delta} < \infty$ *for some* $\delta > 0$. *Let* $S_n = \sum_{i=1}^n Z_{n,i}$, *and* $\sigma_n^2 = var(S_n) = \sum_{i=1}^n \sigma_{n,i}$. *If* $\sigma_n^2 = \sigma^2 + o(1)$, *and*

$$\lim_{n\to\infty} \sum_{i=1}^{n} E|(Z_{n,i} - E(Z_{n,i}))|^{2+q} = 0 \text{ for some } \delta > 0, \tag{7.17}$$

then

$$\sigma_n^{-1}(S_n - E(S_n)) = \sigma_n^{-1} \sum_{i-1}^{n} [(Z_{n,i} - E(Z_{n,i})] \xrightarrow{d} N(0,1). \tag{7.18}$$

I first show the asymptotic normality of $\hat{m}(y,x)$ which leads to the result for $\hat{g}(y|x)$. Using (7.3) and (7.4):

$$\begin{aligned}
&\sqrt{Nh_1...h_q}\left\{\hat{m}(y,x) - m(y,x) - \sum_{s=1}^{q} h_s^2 B_{1s}(z) - \sum_{s=1}^{r} \lambda_s^2 B_{1s}(z)\right\} \\
&= \sqrt{Nh_1...h_q}\{\hat{m}(y,x) - E[\hat{m}(y,x)]\} \\
&\quad + \sqrt{Nh_1...h_q}\left\{E[\hat{m}(y,x)] - m(y,x) - \sum_{s=1}^{q} h_s^2 B_{1s}(z) - \sum_{s=1}^{r} \lambda_s^2 B_{1s}(z)\right\} \\
&= \sqrt{Nh_1...h_q}\{\hat{m}(y,x) - E[\hat{m}(y,x)]\} + O\left(\sqrt{Nh_1...h_q}\left(\sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{q} \lambda_s\right)\right) \\
&= \sqrt{Nh_1...h_q}[\hat{m}(y,x) - E(\hat{m}(y,x))] + o(1) \\
&= \sum_{i=1}^{N} Z_{N,i} + o(1),
\end{aligned} \tag{7.19}$$

where $Z_{N,i} = (\sqrt{Nh_1...h_q})^{-1}[\pi^{-1}\mathbf{1}(i \in S)(K_{\gamma,iy} - g(y|x))K_{\gamma,ix} - E(\pi^{-1}\mathbf{1}(i \in S)(K_{\gamma,iy} - g(y|x))K_{\gamma,ix}]$. To apply the Liapunov Central Limit Theorem, take the expectation of of the absolute value of $Z_{N,i}$ raised to the power of $2 + \delta$, where δ is some constant and $\delta > 0$:

$$\begin{aligned}
E|Z_{N,i}|^{2+q} =& (\sqrt{Nh_1...h_q})^{-(2+q)} E[\pi^{-1}\mathbf{1}(i \in S)(K_{\gamma,iy} - g(y|x))K_{\gamma,ix} \\
&- E(\pi^{-1}\mathbf{1}(i \in S)(K_{\gamma,iy} - g(y|x))K_{\gamma,ix})]^{2+q}.
\end{aligned} \tag{7.20}$$

Applying the Cr inequality and using the result from (7.15):

$$E|Z_{N,i}|^{2+q} \leq \frac{2^{1+q}E[\pi^{-1}\mathbf{1}(i \in S)(K_{\gamma,iy} - g(y|x))K_{\gamma,ix}]^{2+q}}{(\sqrt{Nh_1...h_q})^{2+q}}$$
$$= o(1) \tag{7.21}$$

and

$$\sum_{i=1}^{N} Z_{N,i} \xrightarrow{d} N(0, (\Delta + Q)\kappa^q f(z)). \tag{7.22}$$

Combining this result with $\hat{f}(x) - f(x) = o_P(1)$:

$$\begin{aligned}
&\sqrt{Nh_1...h_q}\left\{\hat{g}(y|x) - g(y|x) - \sum_{s=1}^{q} h_s^2 B_{1s}(z) - \sum_{s=1}^{r} \lambda_s^2 B_{1s}(z)\right\} \\
&\quad \equiv \sqrt{Nh_1...h_q}\left\{\hat{m}(y,x) - \hat{f}(x)\left[\sum_{s=1}^{q} h_s^2 B_{1s}(z) + \sum_{s=1}^{r} \lambda_s^2 B_{1s}(z)\right]\right\}/\hat{f}(x) \\
&\quad = \sqrt{Nh_1...h_q}\left\{\hat{m}(y,x) - f(x)\left[\sum_{s=1}^{q} h_s^2 B_{1s}(z) + \sum_{s=1}^{r} \lambda_s^2 B_{1s}(z)\right]\right\}/f(x) + o_P(1) \\
&\quad \xrightarrow{d} \frac{1}{f(x)} N(0, (\Delta + Q)\kappa^q f(z)) = N(0, (\Delta + Q)\kappa^q g(y|x)/f(x))
\end{aligned} \tag{7.23}$$

8 Tables

Table 1: Strata Borders for Endogeneous and Exogeneous Sampling Schemes

Sample Criterion	Strata Border	Sample Size
y^*	$y^* = 0$	$n/4$
	$y^* = 1$	$3n/4$
x	$x \leq$ %15 quantile	$5n/10$
	%15 quantile $< x \leq$ %85 quantile	$2n/10$
	$x >$ %85 quantile	$3n/10$

Table 2: Median and MAD MSE values for WKPDF, KPDF, Weighted Probit, and Unweighted Probit Models Under Simple Random Sampling

n	σ_{yx}	MSE[$WKPDF$]	MSE[$KPDF$]	MSE[WP]	MSE[UP]
200	0.25	0.0587	0.0601	0.0583	0.0583
		(0.0036)	(0.0044)	(0.0104)	(0.0104)
200	0.50	0.0481	0.0497	0.0481	0.0481
		(0.0034)	(0.0044)	(0.0090)	(0.0090)
200	0.75	0.0349	0.0363	0.0349	0.0349
		(0.0031)	(0.0037)	(0.0042)	(0.0042)
500	0.25	0.0574	0.0583	0.0570	0.0570
		(0.0020)	(0.0025)	(0.0066)	(0.0066)
500	0.50	0.0474	0.0483	0.0473	0.0473
		(0.0021)	(0.0025)	(0.0053)	(0.0053)
500	0.75	0.0341	0.0345	0.0341	0.0341
		(0.0018)	(0.0020)	(0.0027)	(0.0027)
1000	0.25	0.0572	0.0579	0.0572	0.0572
		(0.0015)	(0.0017)	(0.0047)	(0.0047)
1000	0.50	0.0472	0.0477	0.0472	0.0472
		(0.0014)	(0.0017)	(0.0039)	(0.0039)
1000	0.75	0.0339	0.0343	0.0338	0.0338
		(0.0012)	(0.0014)	(0.0020)	(0.0020)

Table 3: Median and MAD MSE values for WKPDF, KPDF, Weighted Probit, and Unweighted Probit Models Under Endogeneous Sampling

n	σ_{yx}	MSE[$WKPDF$]	MSE[$KPDF$]	MSE[WP]	MSE[UP]
200	0.25	0.0584	0.1680	0.0587	0.2246
		(0.0037)	(0.0056)	(0.0038)	(0.0063)
200	0.50	0.0485	0.1506	0.0484	0.1971
		(0.0034)	(0.0056)	(0.0036)	(0.0063)
200	0.75	0.0347	0.1119	0.0351	0.1331
		(0.0030)	(0.0059)	(0.0032)	(0.0065)
500	0.25	0.0579	0.1703	0.0576	0.2242
		(0.0021)	(0.0033)	(0.0025)	(0.0037)
500	0.50	0.0479	0.1517	0.0476	0.1966
		(0.0022)	(0.0034)	(0.0024)	(0.0037)
500	0.75	0.0340	0.1131	0.0343	0.1332
		(0.0018)	(0.0037)	(0.0019)	(0.0040)
1000	0.25	0.0575	0.1710	0.0571	0.2238
		(0.0016)	(0.0023)	(0.0021)	(0.0027)
1000	0.50	0.0474	0.1520	0.0472	0.1967
		(0.0015)	(0.0025)	(0.0017)	(0.0026)
1000	0.75	0.0339	0.1138	0.0339	0.1332
		(0.0013)	(0.0025)	(0.0013)	(0.0029)

Table 4: 95 Percent Confidence Interval of MSE for WKPDF, KPDF, Weighted Probit, and Unweighted Probit Models Under Endogeneous Sampling

		MSE[$WKPDF$]		MSE[$KPDF$]		MSE[WP]		MSE[UP]	
σ_{yx}	n	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
0.25	200	0.059126	0.060076	0.16814	**0.16925**	0.058834	0.059587	0.22426	0.22535
0.25	500	0.057918	0.058376	0.16992	**0.17053**	0.057486	0.057954	0.22375	0.22441
0.25	1000	0.057451	0.057740	**0.17062**	0.17104	0.057033	0.057418	0.22359	0.22407
0.50	200	0.048984	0.049805	0.15070	0.15182	0.048482	0.049167	0.19667	0.19778
0.50	500	0.048017	0.048466	0.15154	0.15216	0.047480	0.047916	0.19646	0.19715
0.50	1000	0.047392	0.047675	0.15197	0.15242	0.047092	0.047431	0.19642	0.19691
0.75	200	0.034972	0.035723	0.11157	0.11273	0.035111	0.035726	0.13272	0.13393
0.75	500	0.034078	0.034444	0.11295	0.11363	0.034201	0.034549	0.13294	0.13367
0.75	1000	0.033962	0.034196	0.11376	0.11423	0.033857	0.034110	0.13288	0.13343

Table 5: Median and MAD MSE values for WKPDF, KPDF, Weighted Probit, and Unweighted Probit Models Under Exogeneous Sampling

n	σ_{yx}	MSE[$WKPDF$]	MSE[$KPDF$]	MSE[WP]	MSE[UP]
200	0.25	0.0579	0.0604	0.0591	0.0581
		(0.0036)	(0.0047)	(0.0160)	(0.0102)
200	0.50	0.0481	0.0499	0.0500	0.0480
		(0.0030)	(0.0041)	(0.0137)	(0.0084)
200	0.75	0.0347	0.0362	0.0366	0.0352
		(0.0029)	(0.0038)	(0.0059)	(0.0043)
500	0.25	0.0574	0.0586	0.0576	0.0573
		(0.0019)	(0.0025)	(0.0106)	(0.0064)
500	0.50	0.0474	0.0484	0.0479	0.0475
		(0.0019)	(0.0025)	(0.0090)	(0.0054)
500	0.75	0.0338	0.0345	0.0349	0.0340
		(0.0017)	(0.0019)	(0.0039)	(0.0028)
1000	0.25	0.0572	0.0577	0.0565	0.0568
		(0.0013)	(0.0017)	(0.0072)	(0.0045)
1000	0.50	0.0471	0.0477	0.0471	0.0470
		(0.0014)	(0.0016)	(0.0064)	(0.0040)
1000	0.75	0.0339	0.0343	0.0342	0.0338
		(0.0013)	(0.0015)	(0.0028)	(0.0020)

Table 6: Variable Descriptions and Summary Statistics

Variable	Description	Percent
Medication		
Any	Individual took any medication for mental illness in the past 2 days	
	=1 if Yes	6.7
	=2 if No	93.3
Anti-depressant	Individual took anti-depressant in the past 2 days	
	=1 if Yes	5.3
	=2 if No	94.7
Anti-psychotic	Individual took anti-psychotic in the past 2 days	
	=1 if Yes	1.0
	=2 if No	99.0
Benzodiazepine	Individual took benzodiazepine in the past 2 days	
	=1 if Yes	1.4
	=2 if No	98.6
Insurance	Respondent has supplementary insurance	
	=1 if Yes	77.9
	=2 if No	22.1
Age	Respondent's age	
	=1 if 15 to 19 years	7.80
	=2 if 20 to 24 years	8.04
	=3 if 25 to 29 years	6.62
	=4 if 30 to 34 years	7.71
	=5 if 35 to 39 years	7.03
	=6 if 40 to 44 years	6.87
	=7 if 45 to 49 years	6.78
	=8 if 50 to 54 years	7.89
	=9 if 55 to 59 years	8.99
	=10 if 60 to 64 years	8.78
	=11 if 65 to 69 years	7.58
	=12 if 70 to 74 years	5.57
	=13 if 75 to 79 years	4.57
	=14 if 80 years or more	5.73
Gender	Respondent's gender	
	=1 if Male	49.3
	=2 if Female	50.7
Education	Respondent's highest level of education attained	

Continued on next page

Table6-*Continued from previous page*

Variable	Description	Percent
	=1 if < Secondary	20.6
	=2 if Secondary Graduate	16.2
	=3 if Some Post-Secondary	6.6
	=4 if Post-Secondary Graduate	56.6
Income	Total household income	
	=1 if $0 to $19,999	6.4
	=2 if $20,000 to $39,999	17.0
	=3 if $40,000 to $59,999	21.0
	=4 if $60,000 to $79,999	16.8
	=5 if $80,000 or more	38.8
SAH	Respondent's self-assessed health	
	=1 if Excellent	22.3
	=2 if Very good	38.3
	=3 if Good	29.1
	=4 if Fair	8.0
	=5 if Poor	2.2
SPS	Social Provision Score	35.82 (sd=4.39)

Table 7: Maximum Likelihood Cross-Validated Bandwidths for Insurance for WKPDF and KPDF

	Medication							
	Any		Anti-Depressant		Anti-Psychotic		Benzodiazepine	
Variable	WKPDF	KDPF	WKPDF	KDPF	WKPDF	KDPF	WKPDF	KDPF
Insurance	0.2626	0.1125	0.2354	0.133	0.1301	0.075	0.2467	0.3398
Gender	0.077	0.0903	0.162	0.0743	0.2691	0.4867	0.1597	0.1026
SAH	0.1011	0.1363	0.1393	0.1228	0.3183	0.3182	0.1631	0.2006
Income	0.775	0.5401	0.9998	0.8082	0.2865	0.3565	0.8506	0.6213
Education	0.9999	1	0.8006	0.8832	0.5924	1	0.9039	1
Age	1.327	1.487	1.7784	1.4864	1.1552	1.3878	1.3268	2.5951
SPS	1.9613	4.3503	1.1371	4.4318	1.4051	4.9749	1.315	4.3956

9 Figures

Table 8: Predicted $Pr(Y = 1|X)$ Versus X^d (all other variables held at their medians/modes)

Variable	Any						Anti-Depressants					
	WKPDF		KPDF		Logit		WKPDF		KPDF		Logit	
	Prob	%Δ	Prob	%Δ	Prob	%Δ	Prob	%Δ	Prob	%Δ	Prob	%Δ
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Insurance												
No	7.61	-	7.69	-	4.81	-	4.96	-	6.24	-	4.22	-
Yes	9.19	20.88	8.55	11.05	7.99	65.95	6.18	24.54	7.46	19.46	7.06	67.19
Gender												
Male	4.29	-	4.71	-	4.03	-	2.73	-	3.97	-	3.18	-
Female	9.19	114.22	8.55	81.35	7.99	98.27	6.18	126.25	7.46	87.87	7.06	121.81
Education												
< Secondary	8.45	-	8.23	-	4.94	-	4.97	-	7.07	-	3.8	-
Secondary	7.86	-6.96	7.9	-3.99	6.26	26.87	4.54	-8.69	6.86	-2.95	5.56	46.05
Some Post Sec.	8.42	-0.34	8.49	3.13	7.8	57.92	5.08	2.06	7.46	5.55	6.7	76.14
Pst Sec. Grad.	9.19	8.82	8.55	3.82	7.99	61.79	6.18	24.15	7.46	5.49	7.06	85.47
Income												
0 to 19,999	9.6	-	12.38	-	14.39	-	6.48	-	8.15	-	11.59	-
20,000 to 39,999	9.17	-4.43	8.84	-28.65	9.01	-37.41	6.48	-0.09	7.55	-7.32	6.78	-41.47
40,000 to 59,999	9.09	-5.29	9.05	-26.9	8.08	-43.84	6.52	0.56	7.71	-5.33	6.48	-44.06
60,000 to 79,999	9.19	-4.21	8.55	-30.99	7.99	-44.51	6.18	-4.76	7.46	-8.46	7.06	-39.11
80,000 and over	9.65	0.5	8.54	-31	7.85	-45.45	6.41	-1.11	7.73	-8.46	6.9	-40.42
SAH												
Excellent	6.76	-	6.76	-	4.28	-	4.07	-	5.05	-	3.7	-
Very Good	9.19	35.91	8.55	26.41	7.99	86.55	6.18	51.88	7.46	47.81	7.06	90.48
Good	14.34	112.04	14.87	120.03	13.33	211.29	13.65	235.81	14.14	180.16	11.78	217.89
Fair	20.36	200.97	25.89	283.08	22.79	432.19	18.89	364.7	22.89	353.62	21.04	468.06
Poor	22.44	231.76	24.49	262.31	29.66	592.76	18.03	343.34	24.89	353.62	25.74	594.84

Continued on next page

Table8-*Continued from previous page*

Variable	Anti-Psychotics						Benzodiazeines					
	WKPDF		KPDF		Logit		WKPDF		KPDF		Logit	
	Prob	%Δ	Prob	%Δ	Prob	%Δ	Prob	%Δ	Prob	%Δ	Prob	%Δ
	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
Insurance												
No	0.68	-	0.36	-	0.27	-	0.96	-	1.45	-	0.73	-
Yes	1.15	69.02	0.92	155.51	0.61	130.54	1.22	26.44	1.39	-4.05	1.15	56.56
Gender												
Male	0.97	-	0.62	-	0.53	-	0.71	-	0.48	-	0.51	-
Female	1.15	19.25	0.92	48.24	0.61	15.88	1.22	72.27	1.39	188.58	1.15	127.34
Education												
< Secondary	0.74	-	0.79	-	0.34	-	1.08	-	1.41	-	0.88	-
Secondary	0.77	4.59	0.75	-5.65	0.4	18.19	0.99	-8.44	1.19	-15.4	0.95	8.1
Some Post Sec.	0.92	24.05	0.81	2.58	0.44	32.34	1.14	5.15	1.35	-4.13	1.2	36.64
Pst Sec. Grad.	1.15	55.38	0.92	15.37	0.61	82.7	1.22	12.34	1.39	-1.22	1.15	30.55
Income												
0 to 19,999	3.42	-	4.65	-	2.04	-	1.26	-	2.01	-	2.11	-
20,000 to 39,999	1.33	-61.16	1.28	-72.43	1.14	-44.07	1.06	-15.35	1.6	-20.51	1.62	-23.46
40,000 to 59,999	0.68	-80	0.61	-86.82	0.6	-70.42	1.04	-17.6	1.57	-21.68	1.42	-32.87
60,000 to 79,999	1.15	-66.37	0.92	-80.31	0.61	-69.83	1.22	-3.3	1.39	-30.74	1.15	-45.54
80,000 and over	1.22	-64.4	0.46	-90.18	0.36	-82.23	0.86	-31.8	0.95	-30.74	1.06	-50.04
SAH												
Excellent	0.73	-	0.74	-	0.44	-	1.13	-	1.03	-	0.62	-
Very Good	1.15	57.33	0.92	23.67	0.61	40.43	1.22	7.55	1.39	35.68	1.15	85.79
Good	1.8	145.46	1.51	103.65	1.6	266.09	1.49	32.11	2.51	144.31	2.42	291.26
Fair	3.19	336	2.98	302.89	3.14	617.85	4.33	283.35	6.28	512.04	4.75	667.41
Poor	2.29	212.39	2.92	294.43	4.85	1008.38	3.46	206.32	7.25	512.04	8.64	1295.33

Table 9: Percent Change in Insurance by X^d (All Other Variables Held Constant at Their Medians/Modes)

	Percent Change in Insurance Status											
	Any			Anti-Depresants			Anti-Psychotics			Benzodiazepines		
Variable	WKPDF	KPDF	Logit	WKPDF	KPDF	Logit	WKPDF	KPDF	Logit	WKPDF	KPDF	Logit
Gender												
Male	-7.3	14.4	68.8	-23.2	-3.2	70	74.7	73.4	130.7	24	-1.9	56.9
Female	35.7	37.5	66	24.5	22.4	67.2	69	154.8	130.5	26.4	4.8	56.6
Education												
< Secondary	31.3	32.4	68.1	9.2	15.9	69.5	41.3	120.9	130.9	21.9	6.1	56.7
Secondary	23.1	27.1	67.2	-4.8	13.1	68.3	52.7	108.4	130.8	17.2	-10.2	56.7
Some Post Sec.	30.2	36.6	66.1	10.8	22.5	67.4	54.2	126.6	130.8	22.2	1.7	56.5
Pst Sec. Grad.	35.7	37.5	66	24.5	22.4	67.2	69	154.8	130.5	26.4	4.8	56.6
Income												
0 to 19,999	52.6	123.1	61.4	31.9	38.5	63.9	428.6	104.2	128.7	38.4	19.6	56
20,000 to 39,999	41.7	63	65.2	24.9	28.8	67.4	64.9	-23.3	129.9	28.5	3.9	56.3
40,000 to 59,999	41.7	60	65.9	28.1	30	67.6	54.3	0.8	130.6	25.2	10.5	56.4
60,000 to 79,999	35.7	37.5	66	24.5	22.4	67.2	69	154.8	130.5	26.4	4.8	56.6
80,000 and over	31.6	41.3	66.1	25.1	27	67.3	44.1	99.6	130.9	17.1	-12.9	56.6
SAH												
Excellent	64.1	51.5	68.6	59.8	9.4	69.6	59.4	167.6	130.8	41.8	18.7	56.9
Very Good	35.7	37.5	66	24.5	22.4	67.2	69	154.8	130.5	26.4	4.8	56.6
Good	36.8	53.8	62.1	31.1	41.7	63.8	81	152.4	129.2	18.9	5.3	55.8
Fair	26.3	80.2	55.3	29.1	50.1	57.1	59.9	115.5	127.2	8.8	26.4	54.5
Poor	-30.8	18.4	50.4	-30.9	11	53.7	-56.3	-8.7	125	22.2	2.8	52.3

Figure 1: A Typical Draw from the Conditional CDF $F(y|x)$

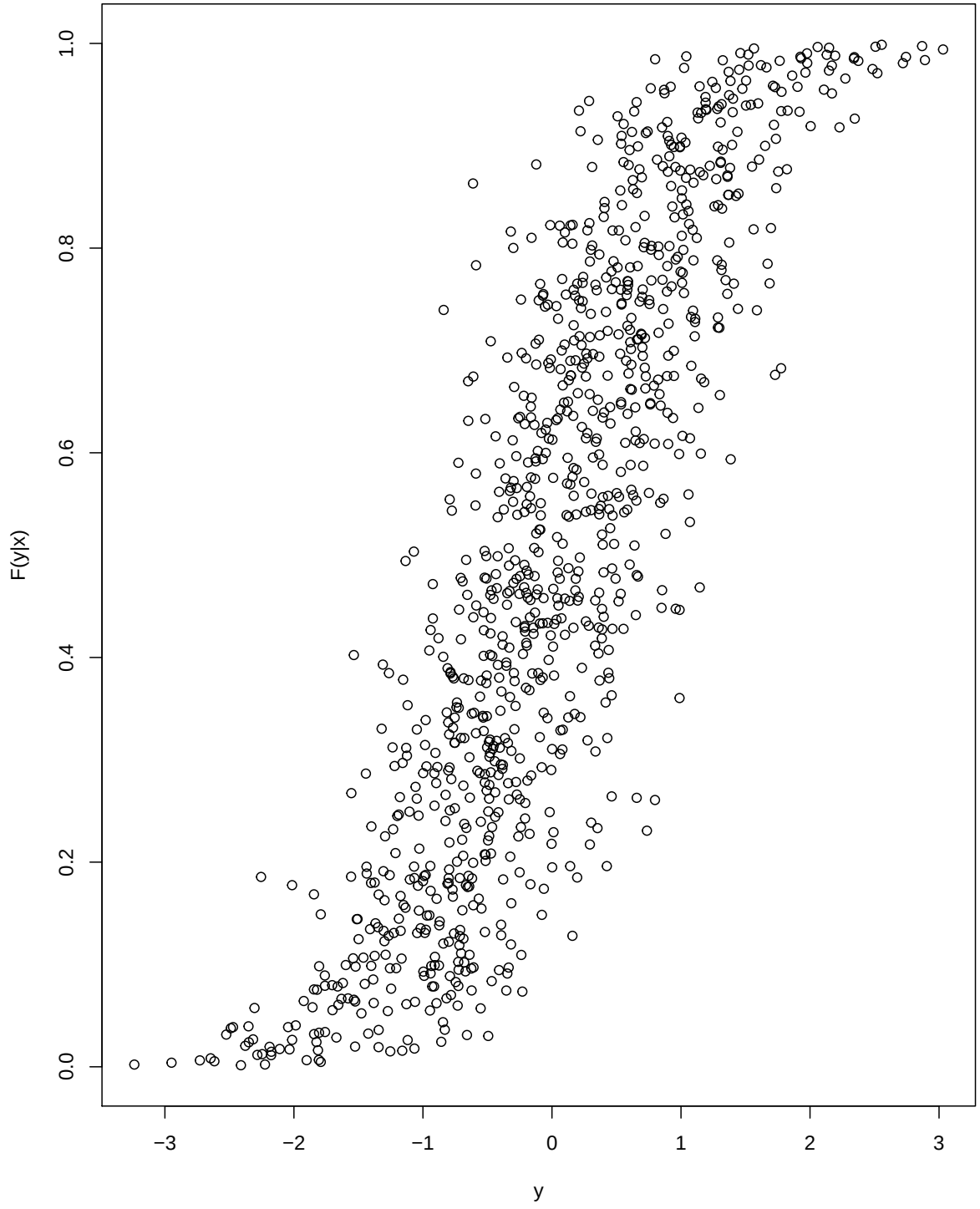
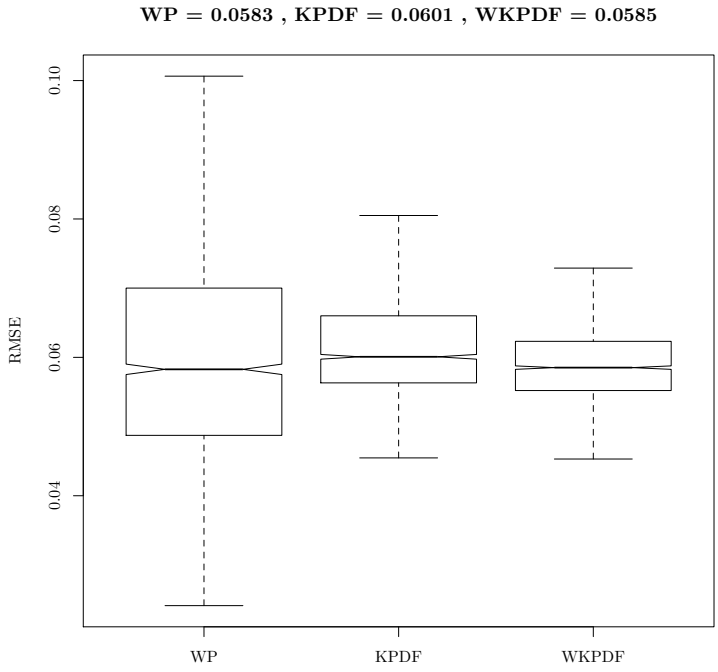
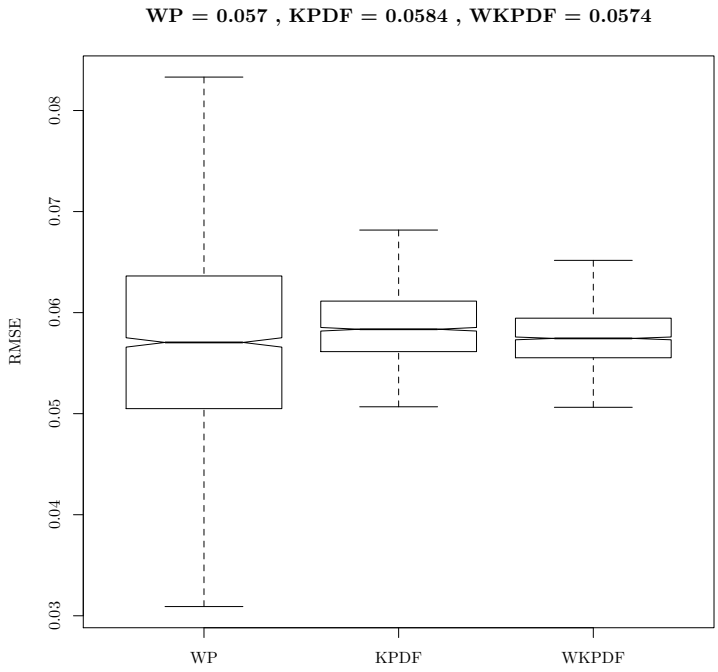


Figure 2: Boxplot of MSE for KPDF, WKPDF, Weighted Probit under SRS for $\sigma_{yx} = 0.25$

(a) $n = 200$



(b) $n = 500$



(c) $n = 1000$

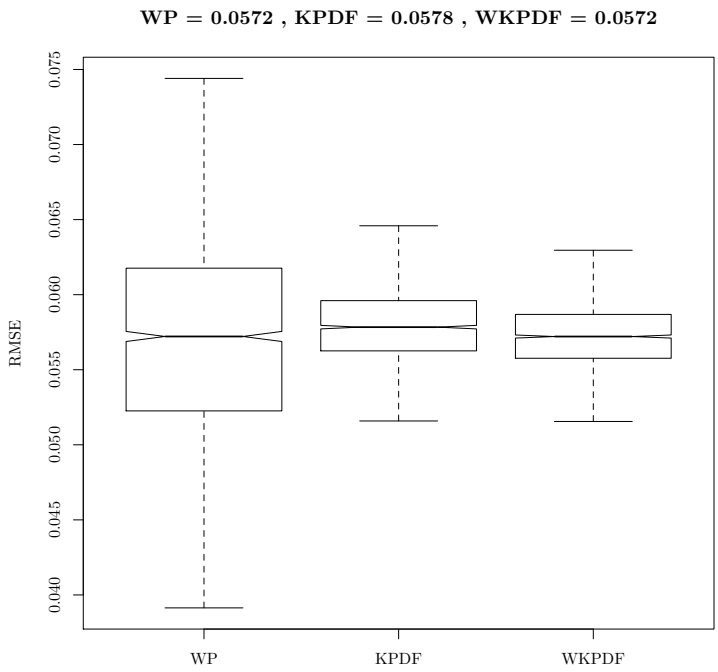
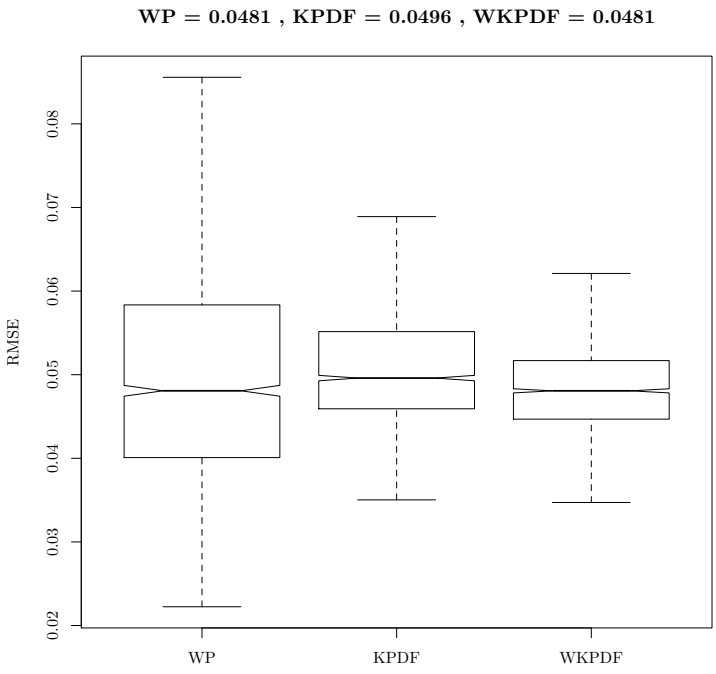
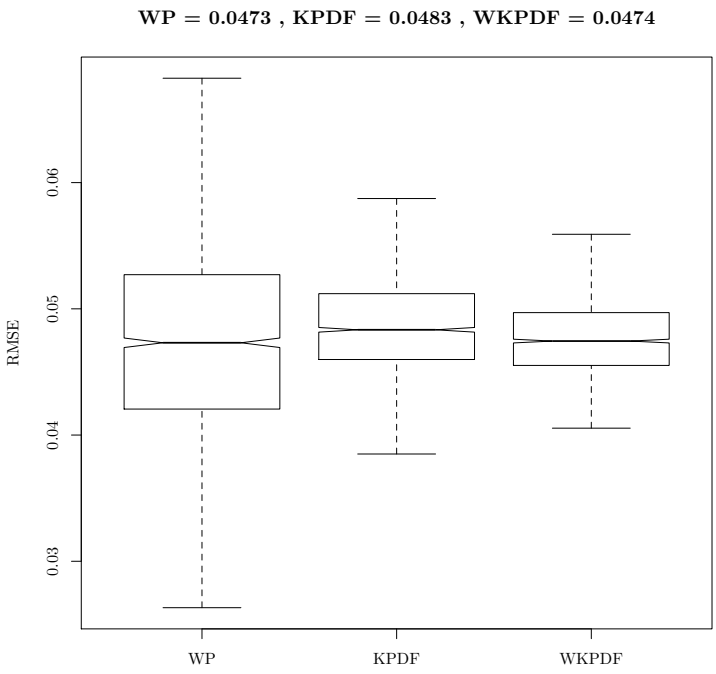

WP = 0.0572 , KPDF = 0.0578 , WKPDF = 0.0572
RMSE
0.040
0.045
0.050
0.055
0.060
0.065
0.070
0.075
WP
KPDF
WKPDF

Figure 3: Boxplot of MSE for KPDF, WKPDF, Weighted Probit under SRS for $\sigma_{yx} = 0.50$

(a) $n = 200$



(b) $n = 500$



(c) $n = 1000$

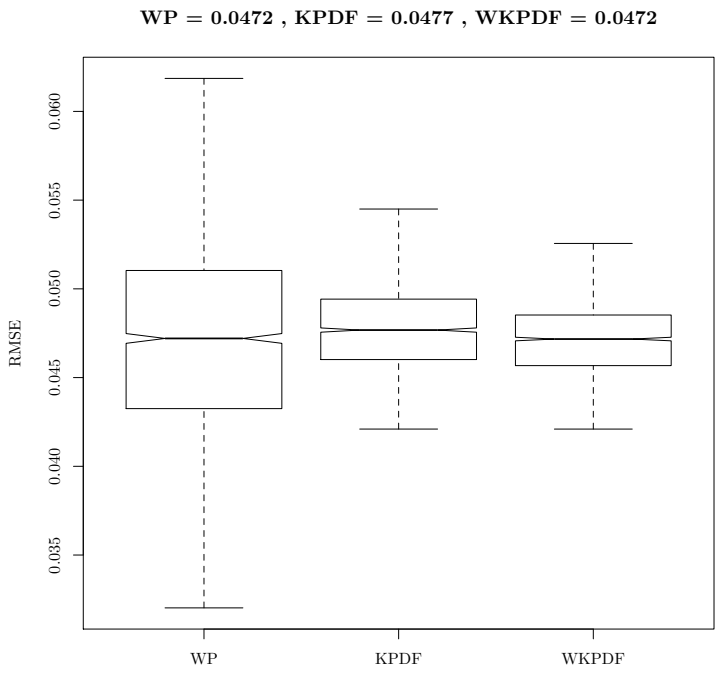
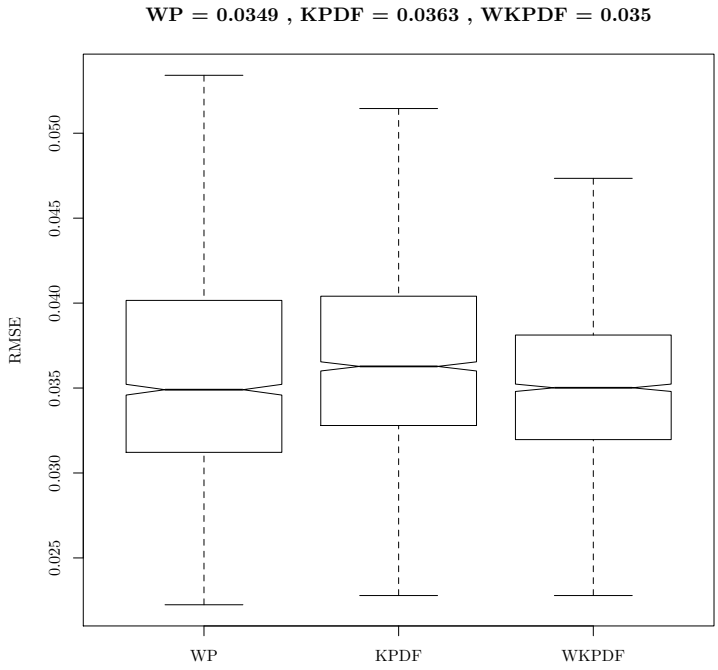
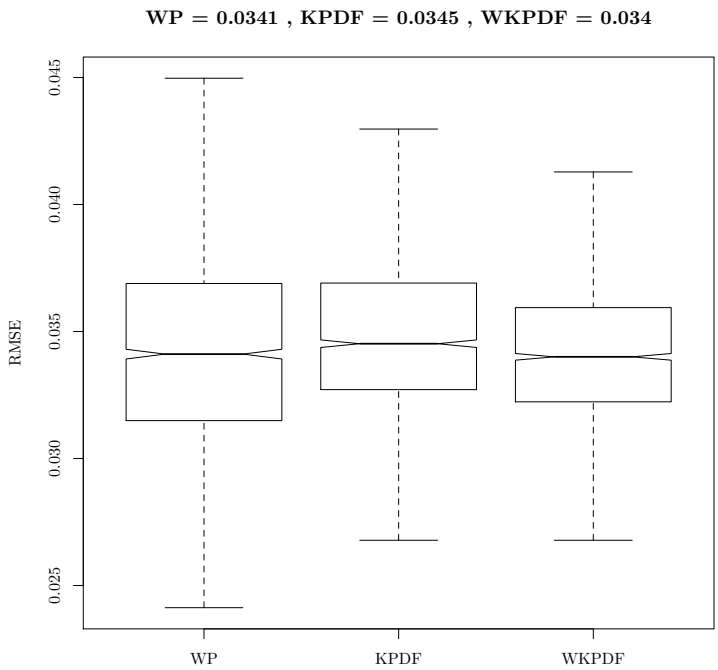

WP = 0.0472 , KPDF = 0.0477 , WKPDF = 0.0472
RMSE
0.035
0.040
0.045
0.050
0.055
0.060
WP
KPDF
WKPDF

Figure 4: Boxplot of MSE for KPDF, WKPDF, Weighted Probit under SRS for $\sigma_{yx} = 0.75$

(a) $n = 200$



(b) $n = 500$



(c) $n = 1000$

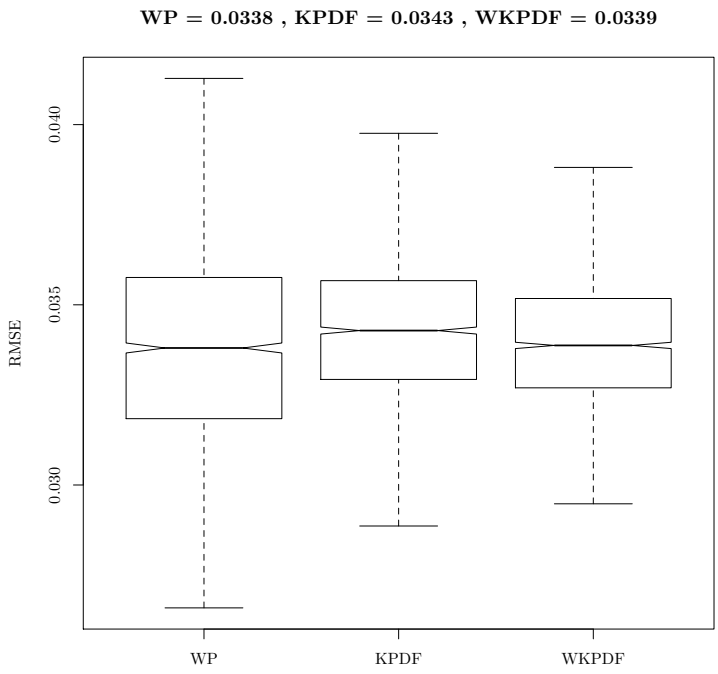
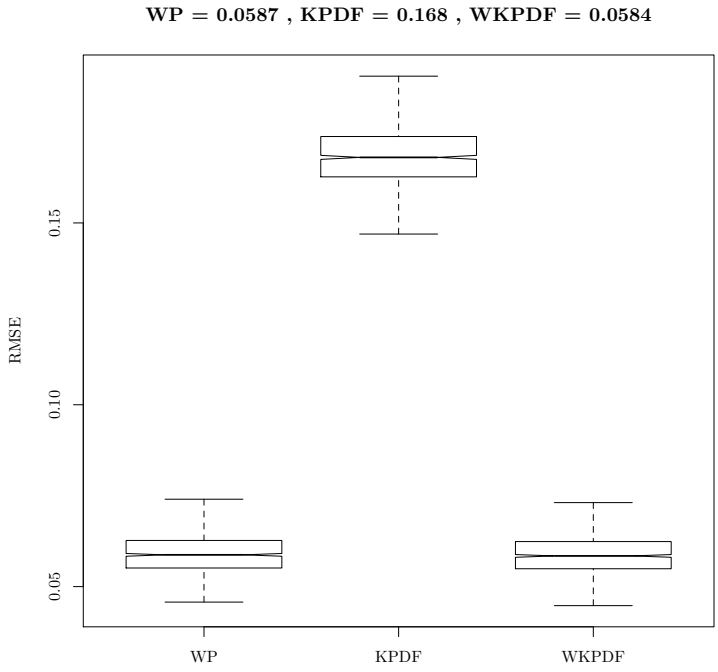
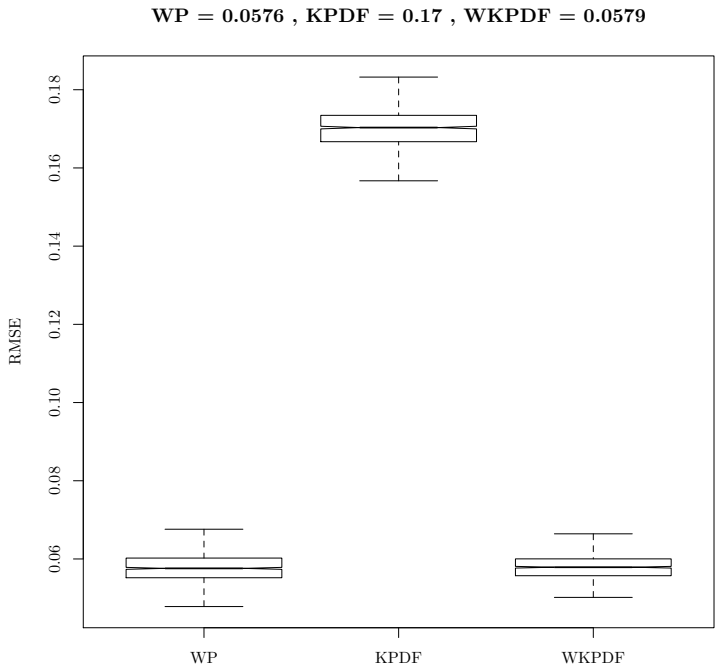

WP = 0.0338 , KPDF = 0.0343 , WKPDF = 0.0339
RMSE
0.030
0.035
0.040
WP
KPDF
WKPDF

Figure 5: Boxplot of MSE for KPDF, WKPDF, Weighted Probit under endog for $\sigma_{yx} = 0.25$

(a) $n = 200$



(b) $n = 500$



(c) $n = 1000$

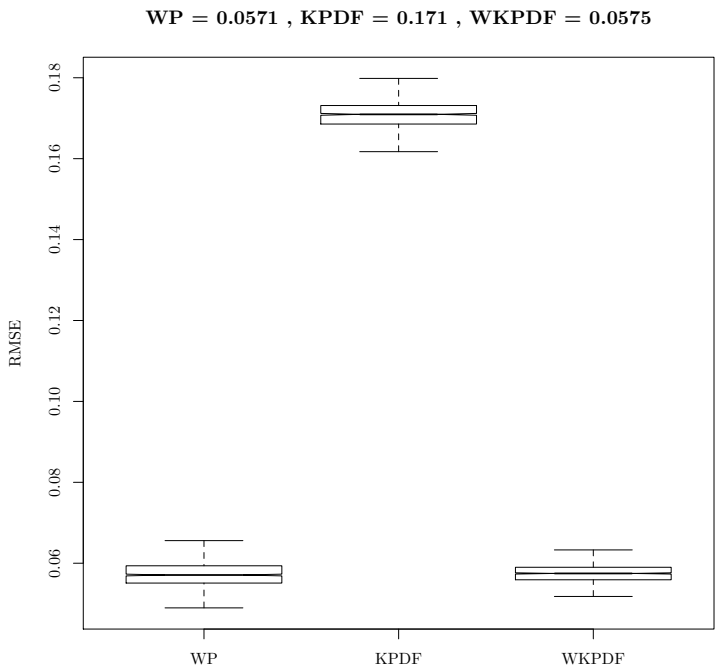
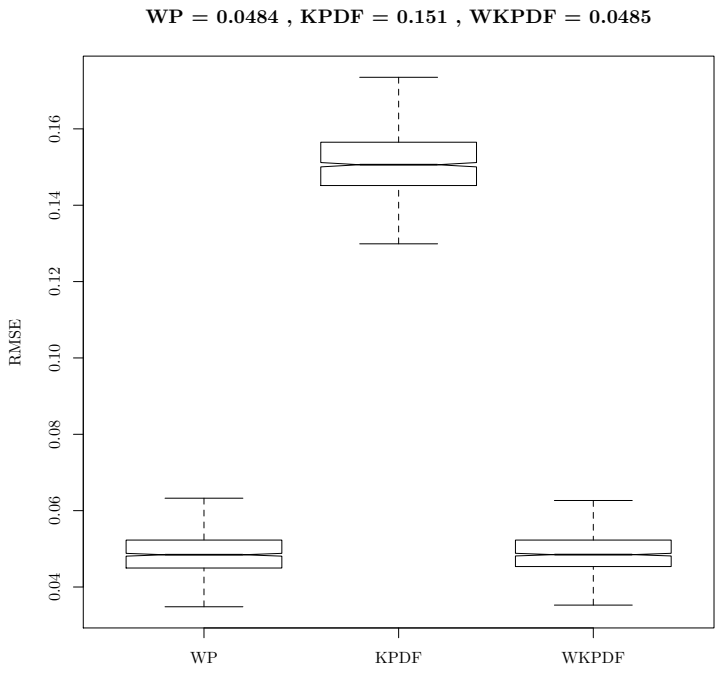
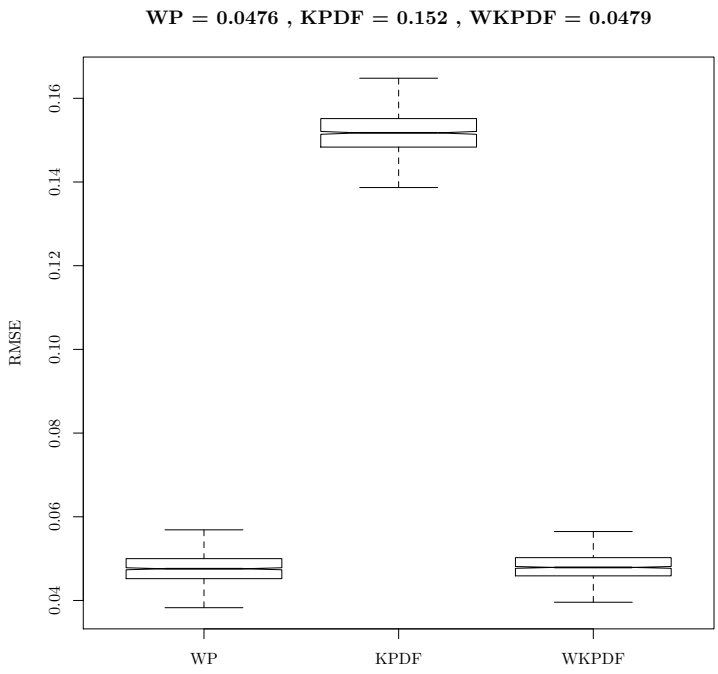

WP = 0.0571 , KPDF = 0.171 , WKPDF = 0.0575
RMSE
0.06
0.08
0.10
0.12
0.14
0.16
0.18
WP
KPDF
WKPDF

Figure 6: Boxplot of MSE for KPDF, WKPDF, Weighted Probit under endog for $\sigma_{yx} = 0.50$

(a) $n = 200$



(b) $n = 500$



(c) $n = 1000$

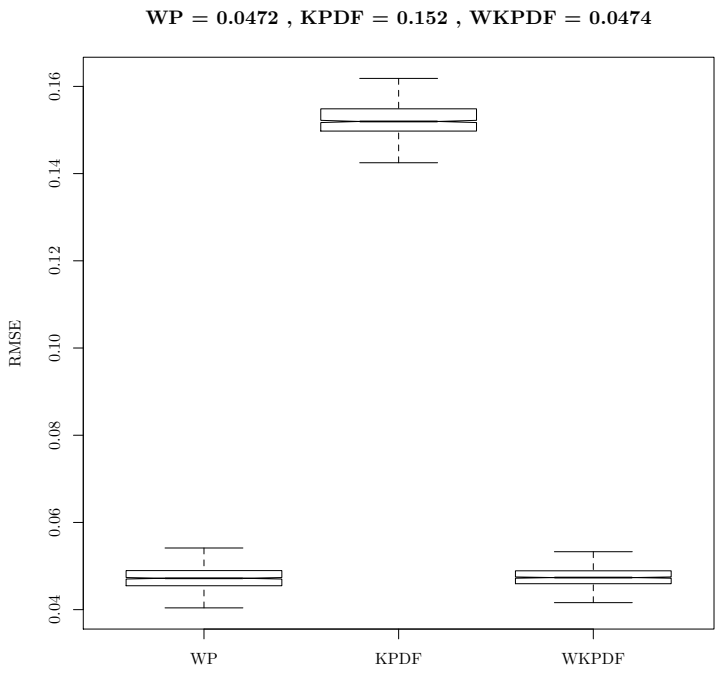
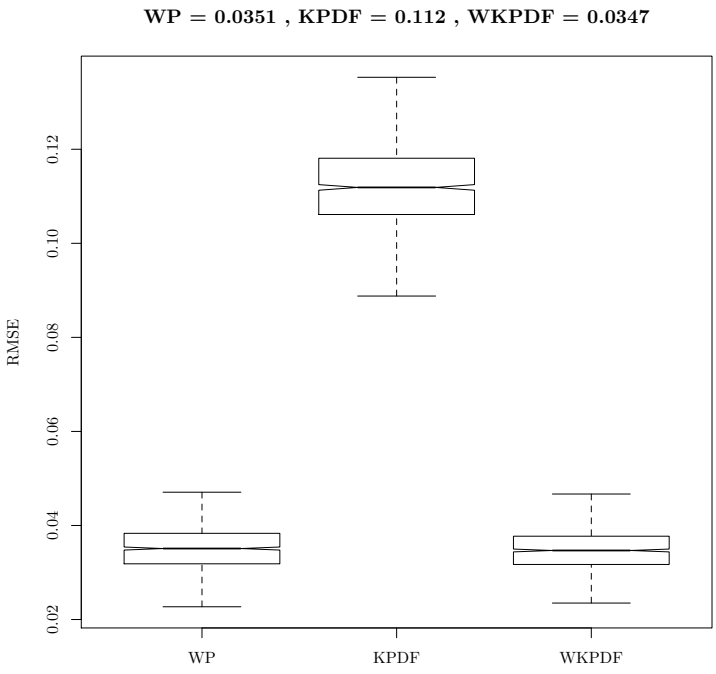
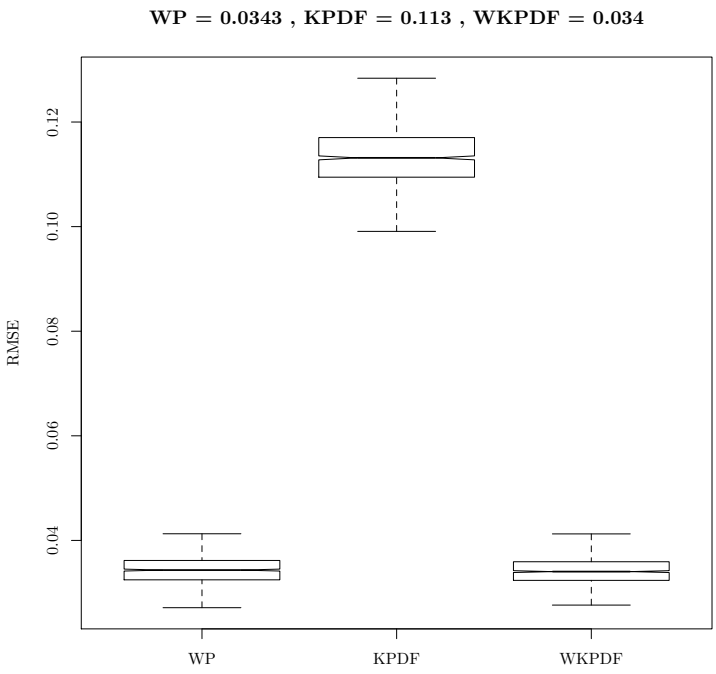

WP = 0.0472 , KPDF = 0.152 , WKPDF = 0.0474
RMSE
0.04
0.06
0.08
0.10
0.12
0.14
0.16
WP
KPDF
WKPDF

Figure 7: Boxplot of MSE for KPDF, WKPDF, Weighted Probit under endog for $\sigma_{yx} = 0.75$

(a) $n = 200$



(b) $n = 500$



(c) $n = 1000$

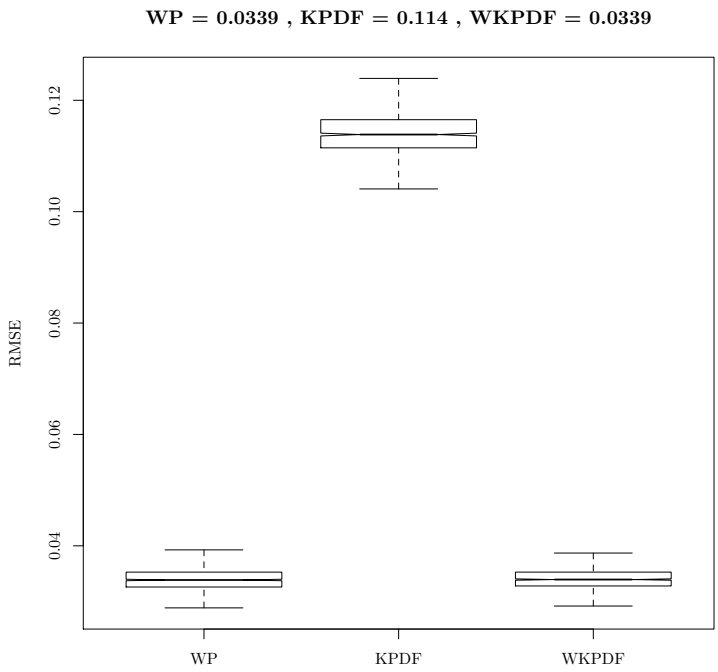
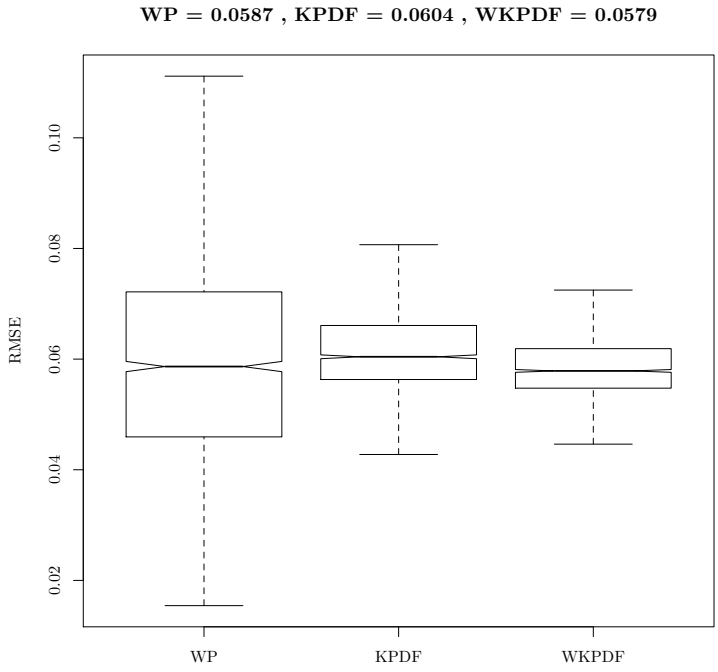
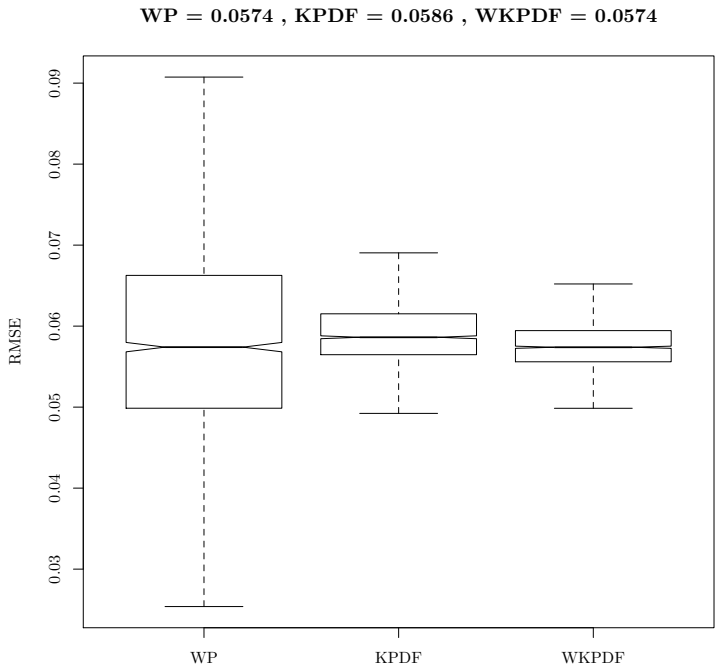

WP = 0.0339 , KPDF = 0.114 , WKPDF = 0.0339
RMSE
0.04
0.06
0.08
0.10
0.12
WP
KPDF
WKPDF

Figure 8: Boxplot of MSE for KPDF, WKPDF, Weighted Probit under exog for $\sigma_{yx} = 0.25$

(a) $n = 200$



(b) $n = 500$



(c) $n = 1000$

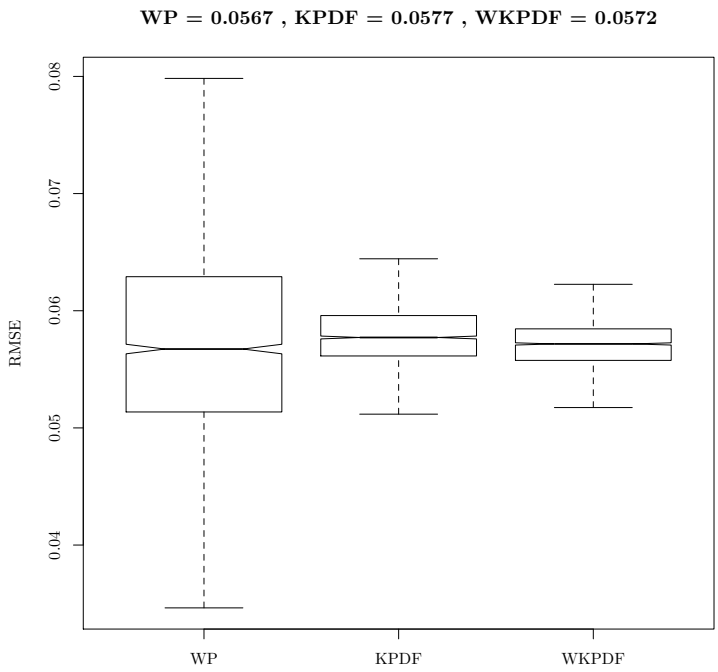
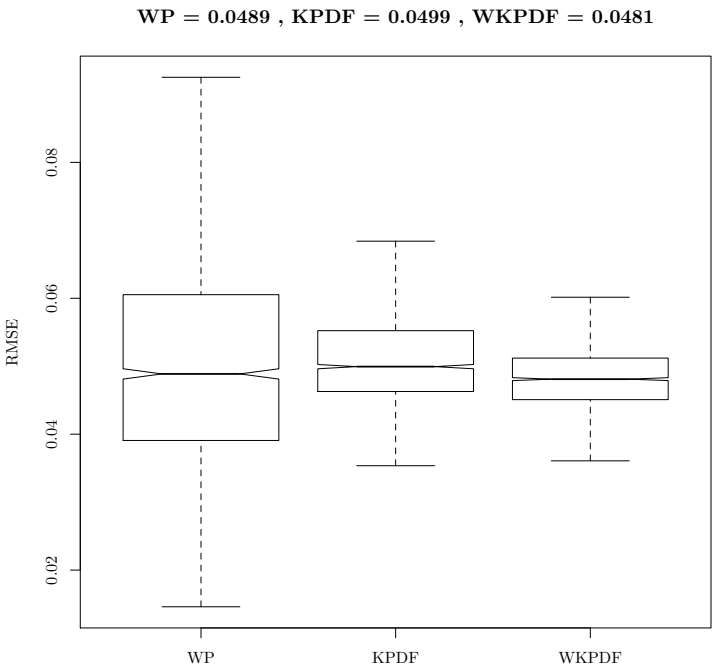
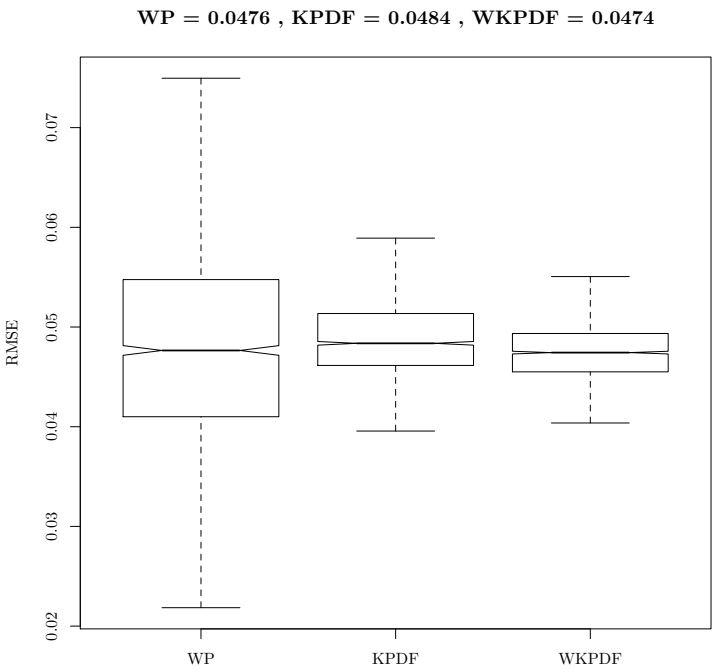

WP = 0.0567 , KPDF = 0.0577 , WKPDF = 0.0572
RMSE
0.04
0.05
0.06
0.07
0.08
WP
KPDF
WKPDF

Figure 9: Boxplot of MSE for KPDF, WKPDF, Weighted Probit under exog for $\sigma_{yx} = 0.50$

(a) $n = 200$



(b) $n = 500$



(c) $n = 1000$

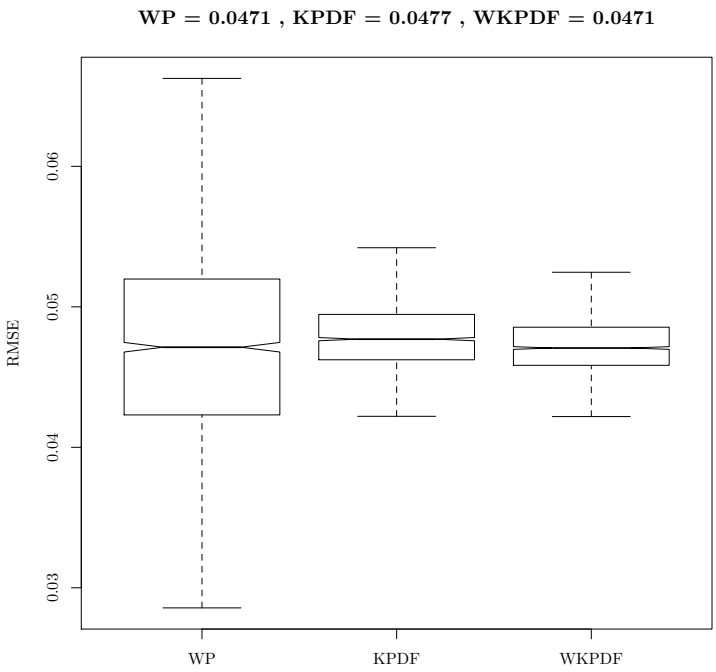
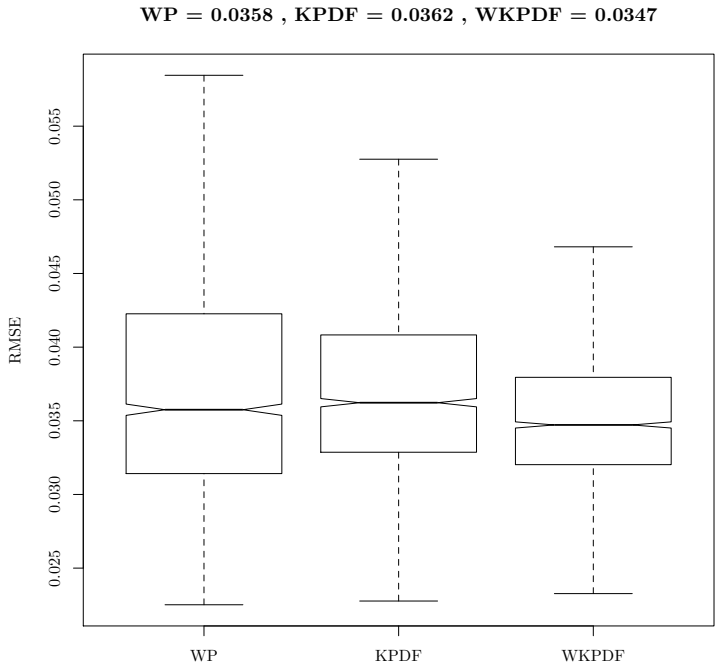
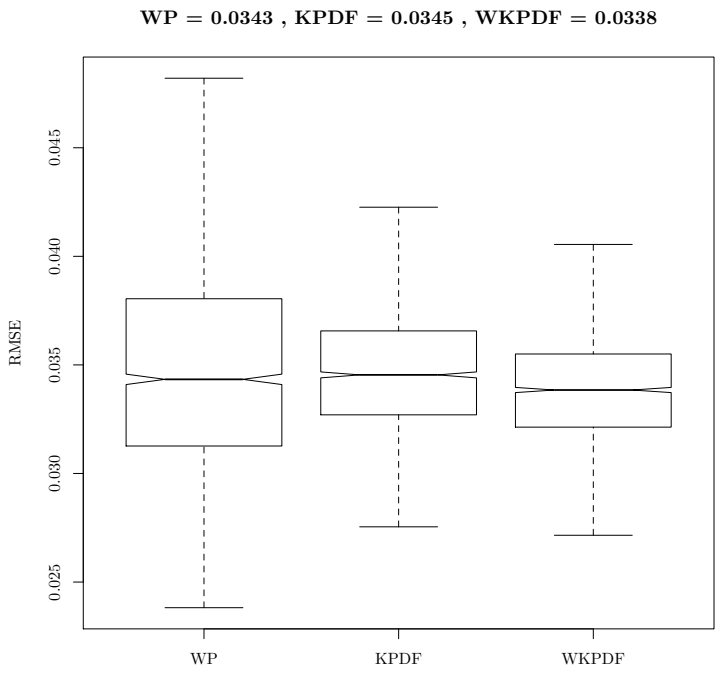

WP = 0.0471 , KPDF = 0.0477 , WKPDF = 0.0471
RMSE
0.03
0.04
0.05
0.06
WP
KPDF
WKPDF

Figure 10: Boxplot of MSE for KPDF, WKPDF, Weighted Probit under exog for $\sigma_{yx} = 0.75$

(a) $n = 200$



(b) $n = 500$



(c) $n = 1000$

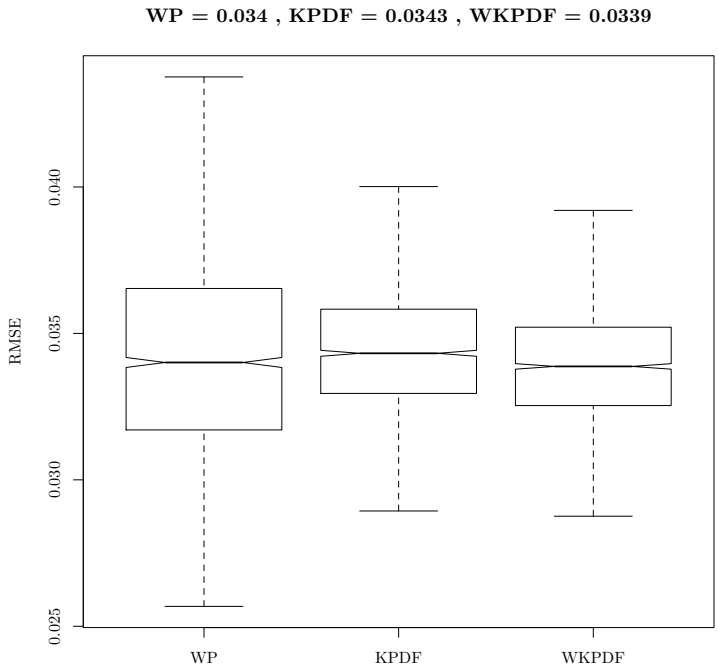


Figure 11: Predicted $Pr(Y_{any}|X)$ from WKPDF, KPDF, and Logit Versus Age Group (All Other Predictors Held Constant at Median/Mode)

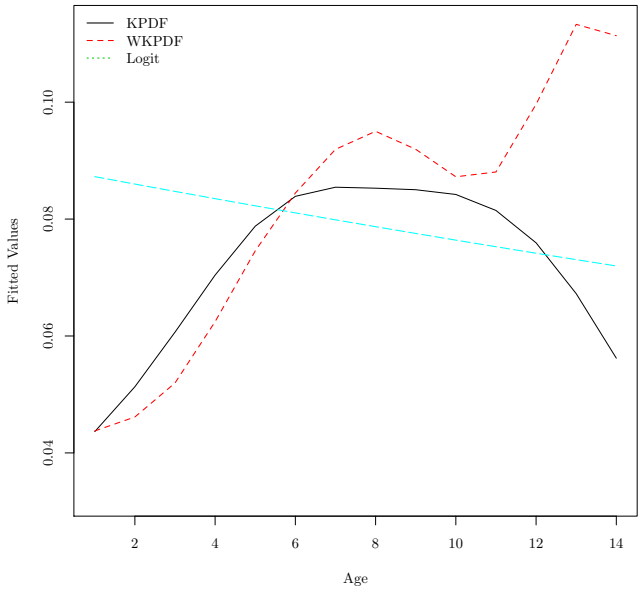


Figure 12: Model 2 Predicted $Pr(Y_{antid}|X)$ Versus Age Group (All Other Predictors Held Constant at Median/Mode)

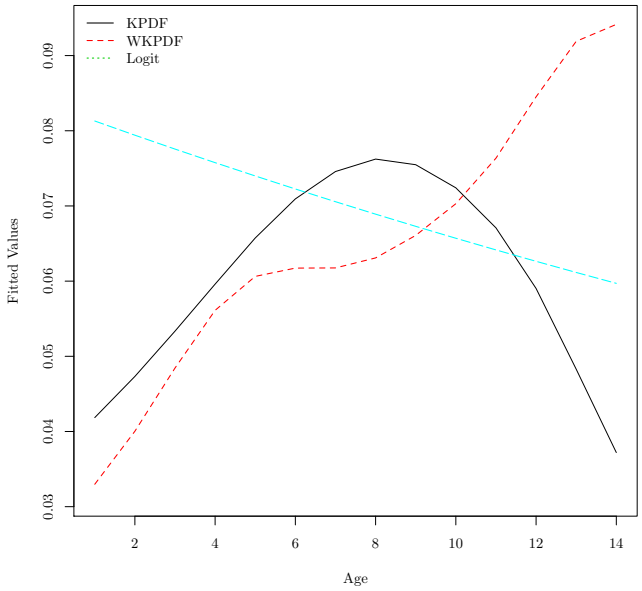


Figure 13: Predicted $Pr(Y_{antip}|X)$ Versus Age Group (All Other Predictors Held Constant at Median/Mode)

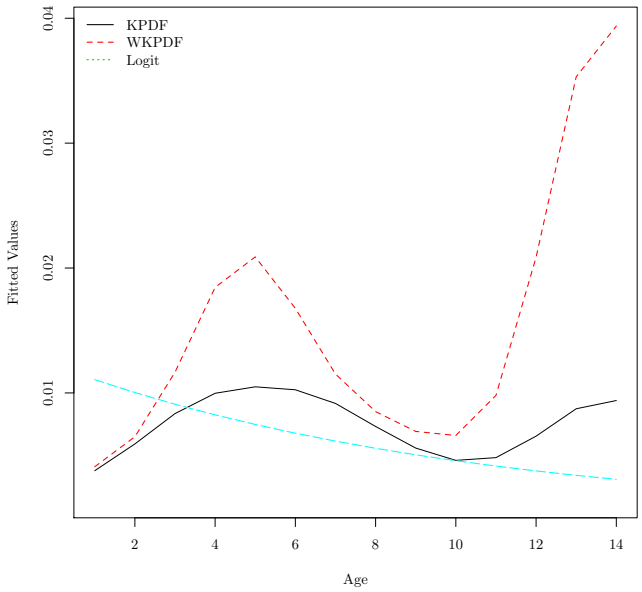


Figure 14: Predicted $Pr(Y_{benzo}|X)$ Versus Age Group (All Other Predictors Held Constant at Median/Mode)

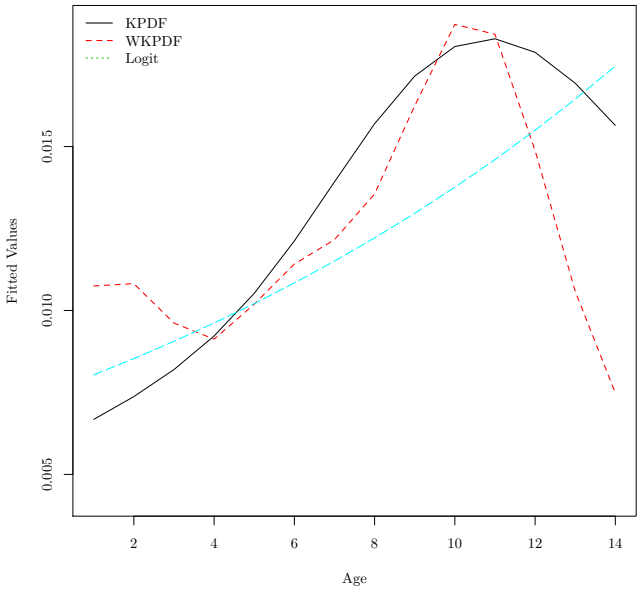


Figure 15: Predicted $Pr(Y_{any}|X)$ from WKPDF, KPDF, and Logit Versus SPS (All Other Predictors Held Constant at Median/Mode)

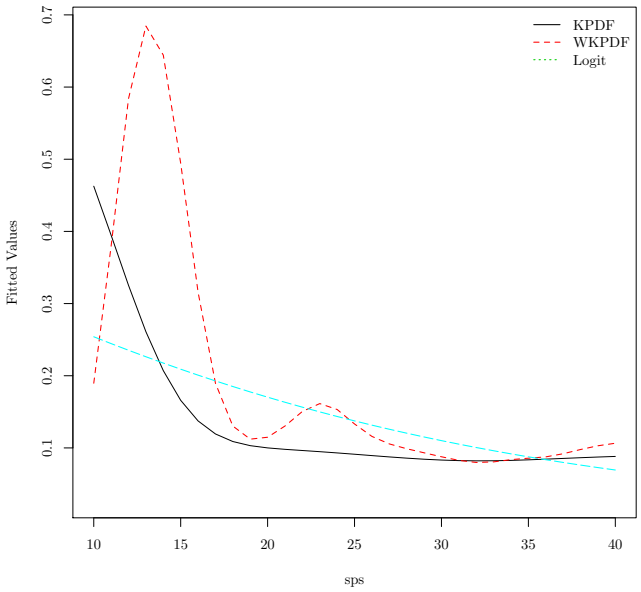


Figure 16: Predicted $Pr(Y_{antid}|X)$ from WKPDF, KPDF, and Logit Versus SPS (All Other Predictors Held Constant at Median/Mode)

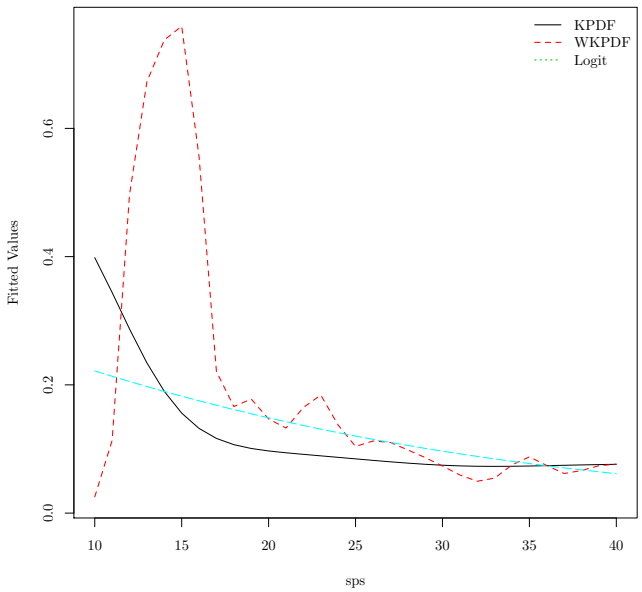


Figure 17: Model 2 Predicted $Pr(Y_{antip}|X)$ Versus SPS (All Other Predictors Held Constant at Median/Mode)

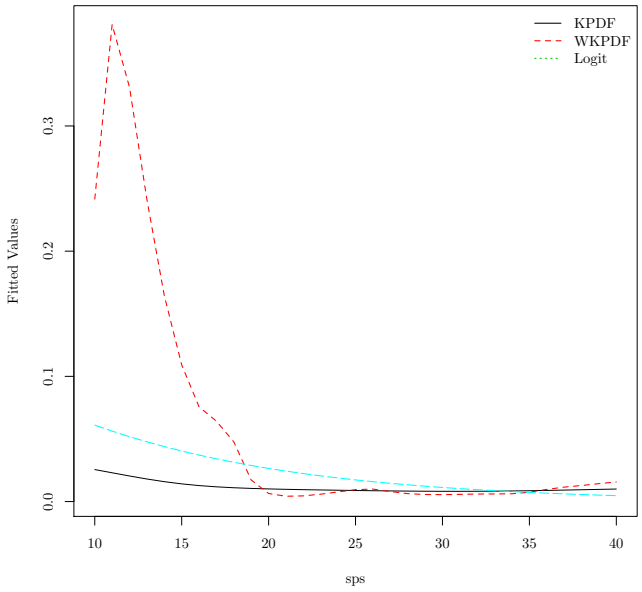


Figure 18: Predicted $Pr(Y_{benzo}|X)$ Versus SPS (All Other Predictors Held Constant at Median/Mode)

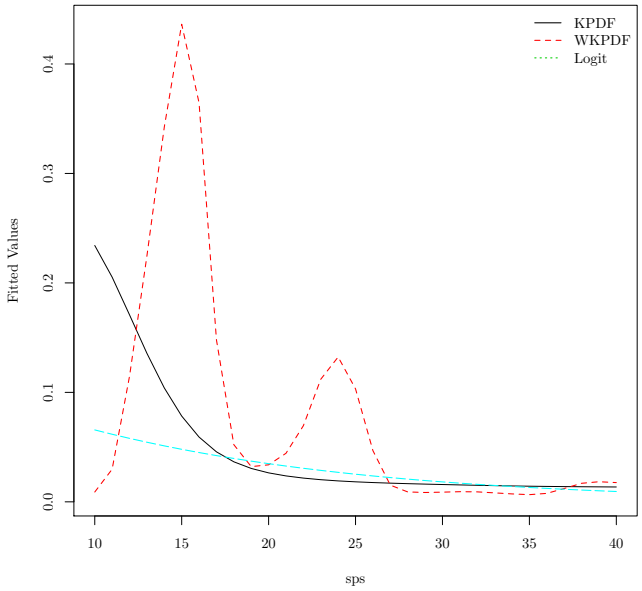


Figure 19: Predicted $Pr(Y_{any}|X)$ from WKPDF Versus Age for Insured and Unsinsured Individuals (All Other Predictors Held Constant at Median/Mode)

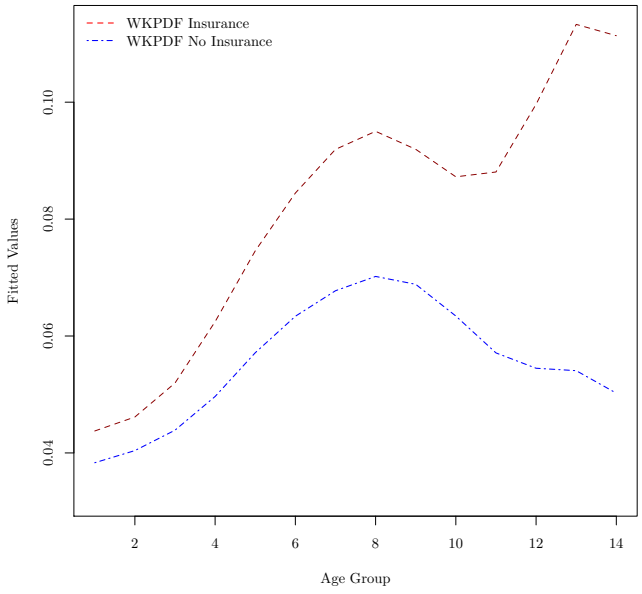


Figure 20: Predicted $Pr(Y_{antid}|X)$ from WKPDF Versus Age for Insured and Unsinsured Individuals (All Other Predictors Held Constant at Median/Mode)

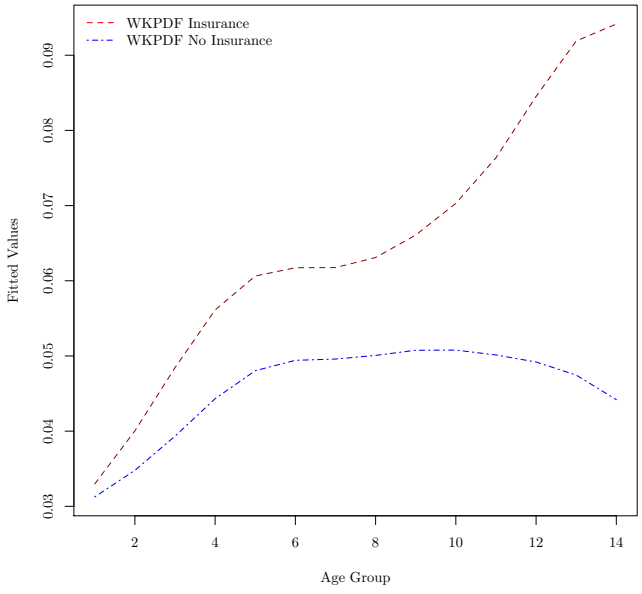


Figure 21: Predicted $Pr(Y_{antip}|X)$ from WKPDF Versus Age for Insured and Unsinsured Individuals (All Other Predictors Held Constant at Median/Mode)

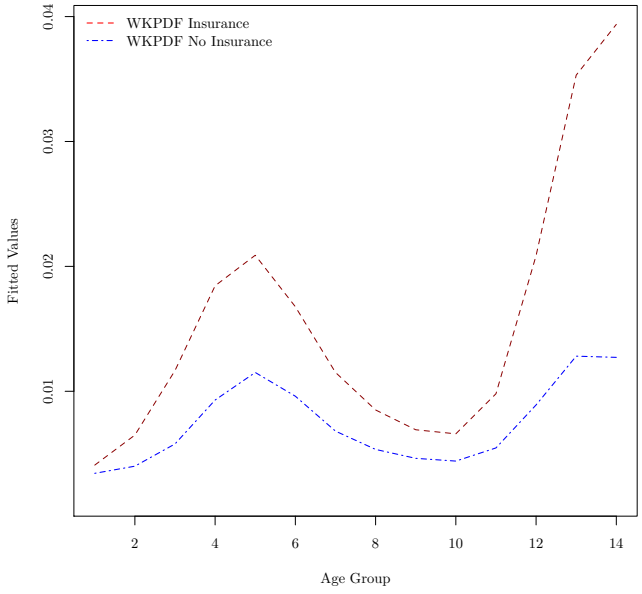


Figure 22: Predicted $Pr(Y_{benzo}|X)$ from WKPDF Versus Age for Insured and Unsinsured Individuals (All Other Predictors Held Constant at Median/Mode)

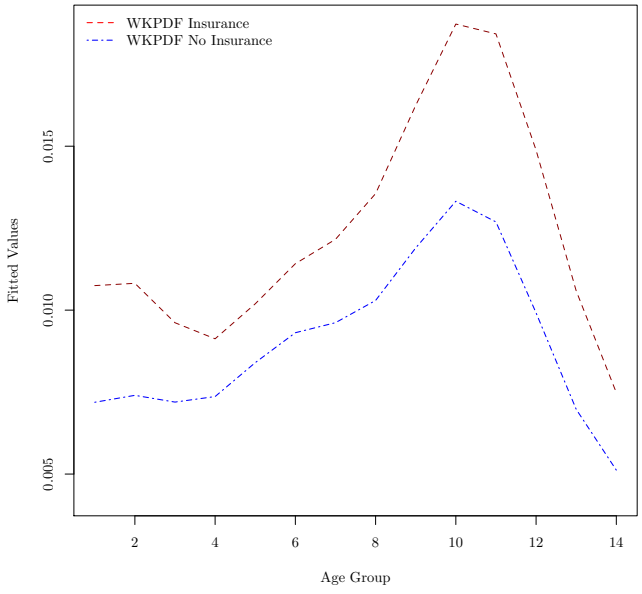


Figure 23: Predicted $Pr(Y_{any}|X)$ from WKPDF Versus SPS for Insured and Unsinsured Individuals (All Other Predictors Held Constant at Median/Mode)

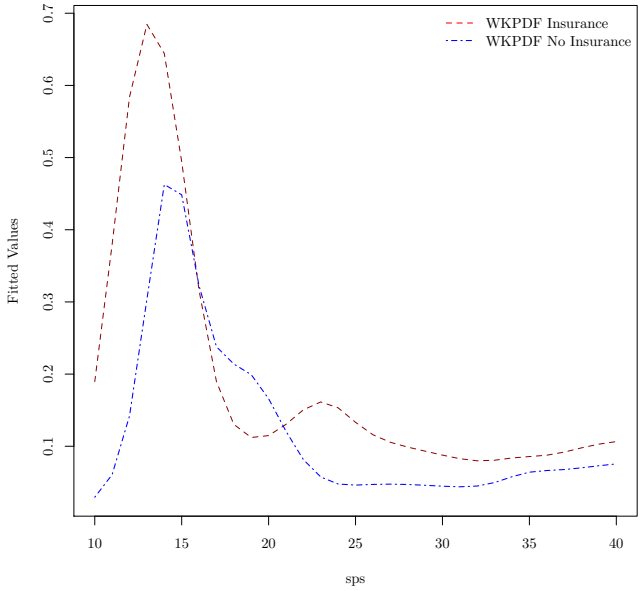


Figure 24: Predicted $Pr(Y_{antid}|X)$ from WKPDF Versus SPS for Insured and Unsinsured Individuals (All Other Predictors Held Constant at Median/Mode)

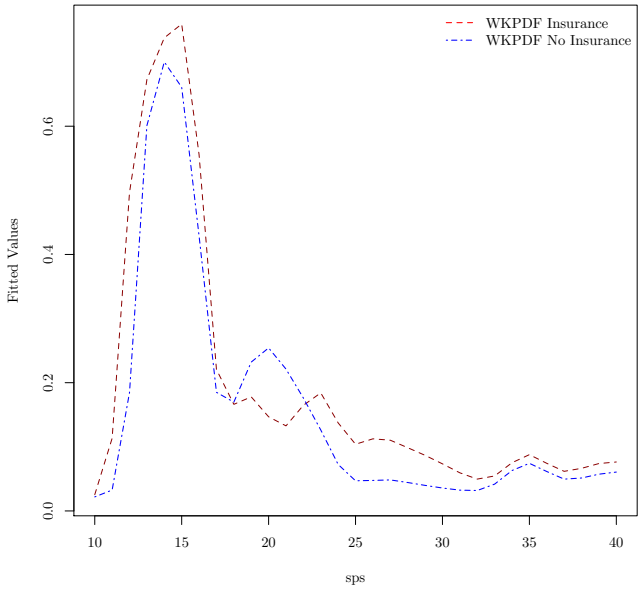


Figure 25: Predicted $Pr(Y_{antid}|X)$ from WKPDF Versus SPS for Insured and Unsinsured Individuals (All Other Predictors Held Constant at Median/Mode)

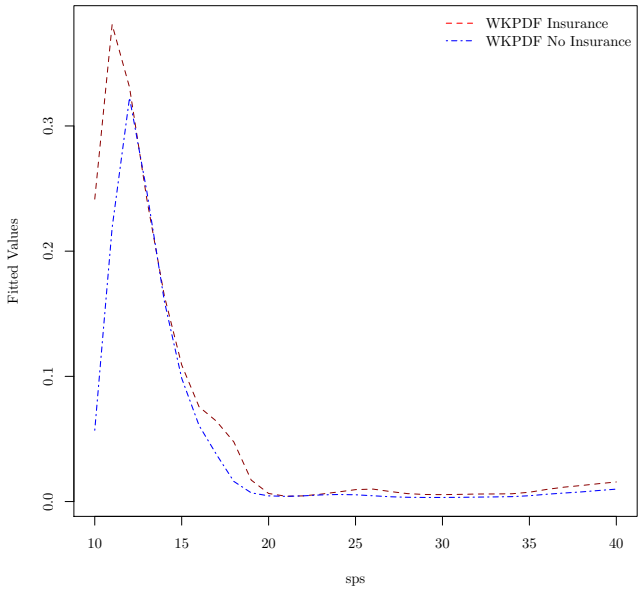


Figure 26: Predicted $Pr(Y_{benzo}|X)$ from WKPDF Versus SPS for Insured and Unsinsured Individuals (All Other Predictors Held Constant at Median/Mode)