

Local Linear Logistic Peters-Belson Regression and its application in employment discrimination cases

Hiro Hikawa¹, Efstathia Bura^{1,2}, and Joseph L. Gastwirth¹

¹ *Department of Statistics, George Washington University, Washington, DC*

² *Biometrics, Vertex Pharmaceuticals, Inc., Cambridge, MA*

Abstract

In cases involving possible discrimination in hiring or promotion plaintiffs allege that they were treated differently than similarly qualified majority individuals. The data are typically analyzed using logistic regression with a minority indicator variable. Alternatively, the Peters-Belson (PB) regression method, which fits a regression model to the majority data and compares the status of each minority member to its prediction obtained from the majority equation, has also been accepted by courts. The average difference estimates the disparity in treatment accounting for job-related covariates. The appropriateness of these parametric models depends on whether they reflect the process generating the data. To lessen the dependence of the ultimate inference on the assumed parametric model, the majority equation is fit by local linear logistic regression and the response of each minority is predicted from it. Large sample properties of this PB-type procedure are obtained and a simulation study shows that the method loses little power relative to parametric methods

even when the assumed parametric method is correct. Moreover, it yields more reliable estimates of the disparity when the data do not follow the assumed model. Data from the *Berger v. Iron Workers Local 201* case are used to illustrate the method.

Keywords

Covariate adjustment; Disparity studies; Employment discrimination; Legal statistics; Local likelihood estimation; Local logistic regression.

1 Introduction

In employment discrimination cases involving hiring and promotion decisions, plaintiffs often allege that they received different treatment from that of similarly qualified majority group members. Often, both the plaintiff and defendant submit statistical evidence to support or refute the claim of disparate treatment. The Supreme Court stated in *Bazemore v. Friday*¹ that a proper statistical analysis should account for the major variables or covariates influencing the response of interest. The standard approach (see Gastwirth, 1989) to account for the effects of major covariates on the binary response (e.g., 1=hired, 0=not hired) is to use an indicator or dummy variable that represents minority status in a logistic regression model.

Gastwirth and Greenhouse (1995) suggested an alternative approach based on Peters-Belson (PB) regression (Peters, 1941; Belson, 1956). This method fits a logistic regression model to the majority group data and measures the difference between each minority's response (0 or 1) and the predicted probability obtained from the majority only model. The average of the differences

¹*Bazemore v. Friday*, 478 U. S. 385, 1986

between the observed and predicted responses estimates the difference between the treatment a minority member received and the treatment s/he would have received had s/he belonged to the majority group.

While logistic regression, either in the standard or PB form, is a widely used method, its usefulness depends on whether the data follow the assumed logit link function and on how well the relationship between the response and the covariates can be approximated by a parametric model such as the linear or quadratic. When the relationship is not straightforward, nonparametric regression (see Takezawa, 2006), in which the data are used to derive the model structure, is a natural alternative. In this paper, we introduce Peters-Belson local linear logistic regression (see Loader, 1999, Chapter 4, for a discussion of local linear logistic regression) and obtain the statistical properties of the estimator of the disparity between two groups. Local linear regression was chosen as it is a nonparametric extension of linear regression and enjoys some of its optimality properties (Fan and Gijbels, 1996). The major advantage of local linear logistic regression in the context of discrimination cases is that in estimating the success probability of a minority member under a fair system, majority members with qualifications (i.e., values of covariates) that are most similar to those of the minority member receive higher weight and those with quite different qualifications receive little or no weight.

The paper is organized as follows. In section 2, we review the classical PB approach, which essentially uses regression to create a “statistical match” for each minority member. In section 3, we introduce the Local Linear Logistic PB regression method and present statistical properties of the disparity estimator. Section 4 reports the results from a simulation study comparing the statistical properties of Local Linear Logistic PB with those of other conventional methods. In section 5, the

method is applied to data from the *Berger v. Iron Workers Local 201*² case. We show that local linear PB provides a more accurate disparity estimate when the linear link logistic model does not provide a good fit to the data. We end with concluding remarks in section 6.

2 Review of the Peters-Belson method

The Peters-Belson (PB) regression method was introduced by Peters (1941) and Belson (1956) and discussed by Cochran and Rubin (1973) for continuous response variables in order to compare mean responses of two groups (e.g., majority vs. minority). It is an alternative to dummy variable regression analysis, and both have been accepted by courts (Gray, 1993). Gastwirth and Greenhouse (1995) adapted the PB method to logistic regression for a binary response. Nayak and Gastwirth (1997) extended it to a broader class of generalized linear models. In the PB method, a regression model is fitted to the majority group data, and the difference between any minority member’s actual response and its estimated response obtained from the majority only regression equation is the estimated disparity of the minority member. The average of these differences provides a summary measure of the disparity for all minority members having accounted for the relevant covariates.

Suppose that the success (e.g., hired, promoted) probabilities are determined by the following inverse logit link functions with d relevant covariates:

$$\begin{aligned} \text{Minority: } p_1(\mathbf{x}_1) &= \frac{e^{\beta_{10} + \sum_{k=1}^d \beta_{1k} x_{1k}}}{1 + e^{\beta_{10} + \sum_{k=1}^d \beta_{1k} x_{1k}}} \\ \text{Majority: } p_2(\mathbf{x}_2) &= \frac{e^{\beta_{20} + \sum_{k=1}^d \beta_{2k} x_{2k}}}{1 + e^{\beta_{20} + \sum_{k=1}^d \beta_{2k} x_{2k}}} \end{aligned}$$

where \mathbf{x}_1 and \mathbf{x}_2 are vectors of the same d covariates for minority and majority, respectively. Note

²*Berger v. Iron Workers Local 201*, 42 FEP Cases 1161 (D.D.C. 1985), 843 F.2d 1395 (D.C. Cir. 1988)

that the subscript 1 is used for minority and 2 is used for majority.

Then, the disparity for the i -th minority member is expressed by

$$\delta_i = p_1(\mathbf{x}_{1i}) - p_2(\mathbf{x}_{1i}) \quad (1)$$

and the average disparity in terms of success probability for all n_1 minority members is

$$\delta = \frac{\sum_{i=1}^{n_1} (p_1(\mathbf{x}_{1i}) - p_2(\mathbf{x}_{1i}))}{n_1} \quad (2)$$

To estimate δ_i in (1), we fit the following logistic regression model to the majority data:

$$\ln \left(\frac{p_2(\mathbf{x}_2)}{1 - p_2(\mathbf{x}_2)} \right) = \beta_{20} + \sum_{k=1}^d \beta_{2k} x_{2k} \quad (3)$$

The estimate of δ_i is the difference between the observed response value of the i -th minority and its predicted success probability from the majority logistic regression model (3):

$$D_i = Y_{1i} - \frac{e^{\hat{\beta}_{20} + \hat{\beta}_{21}x_{11i} + \hat{\beta}_{22}x_{12i} + \dots + \hat{\beta}_{2d}x_{1di}}}{1 + e^{\hat{\beta}_{20} + \hat{\beta}_{21}x_{11i} + \hat{\beta}_{22}x_{12i} + \dots + \hat{\beta}_{2d}x_{1di}}}$$

The average of D_i over all minority members

$$\bar{D} = \frac{\sum_{i=1}^{n_1} D_i}{n_1} \quad (4)$$

serves as an estimator of the average disparity δ in (2).

Using the asymptotic properties of the maximum likelihood estimators of the β_2 coefficients in (3), Gastwirth and Greenhouse (1995) obtained the asymptotic normality of \bar{D} and proposed the test statistic $\bar{D}/\sqrt{\widehat{var}(\bar{D})}$ for testing the null hypothesis of no disparity $H_0 : \delta = 0$.

In the context of legal cases, such as hiring discrimination, the PB approach is particularly attractive, for the method is intuitive and relatively easy to relate to general audiences with little or no statistical knowledge as can be the case for many judges and juries. Using the majority model

to estimate the predicted response value for a minority member provides a statistical match for the minority member’s response if s/he were majority. Thus, if all other potential factors are accounted for and the two groups received a similar treatment, one would expect the difference between the observed and the estimated response values from the majority model to be very small.

3 Local Linear Logistic PB regression method

In order to lessen the dependence of the results on the parametric model choice, we adapt local linear logistic regression to the PB method. The method uses the local likelihood approach in estimating unknown parameters, which was first proposed by Brillinger (1977) and further studied by Tibshirani (1984) and Tibshirani and Hastie (1987). Although the local likelihood approach can be applied to any response variables with a density from the exponential family, since our interest is in binary responses, here we focus on local linear logistic regression.

Suppose that the success probabilities for minority and majority group members are given by

$$\text{Minority: } p_1(\mathbf{x}_{1i}) = \frac{e^{m_1(\mathbf{x}_{1i})}}{1 + e^{m_1(\mathbf{x}_{1i})}} \quad \text{or, } \ln \left(\frac{p_1(\mathbf{x}_{1i})}{1 - p_1(\mathbf{x}_{1i})} \right) = m_1(\mathbf{x}_{1i}) \quad (5)$$

$$\text{Majority: } p_2(\mathbf{x}_{2j}) = \frac{e^{m_2(\mathbf{x}_{2j})}}{1 + e^{m_2(\mathbf{x}_{2j})}} \quad \text{or, } \ln \left(\frac{p_2(\mathbf{x}_{2j})}{1 - p_2(\mathbf{x}_{2j})} \right) = m_2(\mathbf{x}_{2j}) \quad (6)$$

where $\mathbf{x}_{1i} = (x_{11i} \ x_{12i} \ \dots \ x_{1di})^T$ and $\mathbf{x}_{2j} = (x_{21j} \ x_{22j} \ \dots \ x_{2dj})^T$ are d dimensional vectors of covariate values for the i -th minority and the j -th majority members and $m_1(\mathbf{x})$ and $m_2(\mathbf{x})$ are unknown functions of the covariates for the minority and majority groups, respectively. We assume all the second partial derivatives of $m_1(\mathbf{x})$ and $m_2(\mathbf{x})$ exist and are continuous. The true amount of disparity for the i -th minority member δ_i is as defined in (1), and the average disparity of all

minority members δ is as defined in (2).

In local linear regression, $m_2(\mathbf{x}_{2j})$ is approximated by its first order Taylor expansion around \mathbf{x}_{1i} :

$$m_2(\mathbf{x}_{2j}) \approx m_2(\mathbf{x}_{1i}) + m_2^{(1)}(x_{11i})(x_{21j} - x_{11i}) + m_2^{(2)}(x_{12i})(x_{22j} - x_{12i}) + \cdots + m_2^{(d)}(x_{1di})(x_{2dj} - x_{1di}) \quad (7)$$

where $m_2^{(c)}(x_{1ci}) = \partial m_2(\mathbf{x}_{1i}) / \partial x_{1ci}$, $c = 1, 2, \dots, d$. We call the point at which $m_2(\mathbf{x})$ is estimated, namely \mathbf{x}_{1i} in this case, the ‘‘design point.’’ Let

$$\mathbf{d}_{ij} = \begin{pmatrix} 1 & (x_{21j} - x_{11i}) & (x_{22j} - x_{12i}) & \cdots & (x_{2dj} - x_{1di}) \end{pmatrix}^T \quad (8)$$

$$\boldsymbol{\beta} = \begin{pmatrix} m_2(\mathbf{x}_{1i}) & m_2^{(1)}(x_{11i}) & m_2^{(2)}(x_{12i}) & \cdots & m_2^{(d)}(x_{1di}) \end{pmatrix}^T \quad (9)$$

To estimate m_2 amounts to maximizing the following local log-likelihood function with respect to $\boldsymbol{\beta}$,

$$l_{\mathbf{x}_{1i}}(\mathbf{d}_i; \mathbf{y}_2) = \sum_{j=1}^{n_2} K \left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h} \right) \left(y_{2j} \mathbf{d}_{ij}^T \boldsymbol{\beta} - \ln \left(1 + e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}} \right) \right) \quad (10)$$

where $\mathbf{d}_i = (\mathbf{d}_{i1}, \mathbf{d}_{i2}, \dots, \mathbf{d}_{in_2})^T$, $K(\cdot)$ is a symmetric bounded probability density function with bounded support (kernel function) and

$$\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|^2 = \sum_{k=1}^d \left(\frac{x_{2kj} - x_{1ki}}{S_k} \right)^2$$

where S_k is a sample standard deviation of $x_{2k} - x_{1k}$ for the majority members for the k -th covariate.

More specifically, Loader (1999, p.20) suggested the following scaling

$$S_k = \left(\frac{\sum_{j=1}^{n_2} (x_{2kj} - x_{1ki} - (\sum_{j=1}^{n_2} x_{2kj} - x_{1ki}) / n_2)^2}{n_2 - 1} \right)^{1/2}$$

Alternatively, Cleveland and Devlin (1988) suggest dividing each covariate by its sample standard deviation prior to applying the distance function.

The bandwidth h defines the width of the “local neighborhood” about the design point. Since the approximation (7) is accurate only near \mathbf{x}_{1i} , the method puts greater weight on this local neighborhood. One may use a global bandwidth that is the same for all design points, or a nearest neighbor bandwidth, which uses a fixed fraction of the data to find the fitted value of y at the design point (Cleveland and Devlin, 1988; Loader, 1999). This fraction is called the smoothing parameter and determines the value of h for that design point. In this case, since the bandwidth will vary for each design point, it is more accurate to denote it by $h(\mathbf{x}_{1i})$.

In this paper, we use the Epanechnikov kernel,

$$K\left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h}\right) = \frac{3}{4} \left(1 - \left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h}\right)^2\right) I\left(\left|\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h}\right| \leq 1\right)$$

and the value of h may vary with the design points.

In the estimation process, the majority observations that lie further from \mathbf{x}_{1i} than h in the vector distance receive zero weight and those within the neighborhood receive weights that decrease as the distance from the design point grows. The fact that the link function $m(\mathbf{x})$ in (6) can be any adequately smooth function is what makes local linear regression flexible. Moreover, Loader (1999, p.61) notes that the choice of a link function is not restrictive as the local regression technique does not assume a global model.

The first element of the d dimensional vector $\hat{\boldsymbol{\beta}}$ is $\hat{m}_2(\mathbf{x}_{1i})$ and is used to compute the predicted success probability for the i -th minority member if s/he were a majority group member:

$$\hat{p}_2(\mathbf{x}_{1i}) = \frac{e^{\hat{m}_2(\mathbf{x}_{1i})}}{1 + e^{\hat{m}_2(\mathbf{x}_{1i})}}$$

The estimator of the disparity δ_i defined in (1) is

$$D_i = Y_{1i} - \frac{e^{\hat{m}_2(\mathbf{x}_{1i})}}{1 + e^{\hat{m}_2(\mathbf{x}_{1i})}}$$

with the estimator of the average disparity δ defined in (2),

$$\bar{D}_{LOC} = \frac{\sum_{i=1}^{n_1} \left(Y_{1i} - \frac{e^{\hat{m}_2(\mathbf{x}_{1i})}}{1+e^{\hat{m}_2(\mathbf{x}_{1i})}} \right)}{n_1} \quad (11)$$

In order to make inference on δ , we show that the asymptotic distribution of \bar{D}_{LOC} is normal in Theorem 1. The theorem requires the following first and second derivatives of the local log-likelihood (10):

$$l'_{\mathbf{x}_{1i}}(\mathbf{d}_i; \boldsymbol{\beta}; \mathbf{y}_2) = \frac{\partial l_{\mathbf{x}_{1i}}(\mathbf{d}_i; \boldsymbol{\beta}; \mathbf{y}_2)}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \sum_{j=1}^{n_2} K\left(\frac{\|\mathbf{x}_{2j}-\mathbf{x}_{1i}\|}{h}\right) \left(y_{2j} - \frac{e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}}{1+e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}} \right) \\ \sum_{j=1}^{n_2} (x_{21j} - x_{11i}) K\left(\frac{\|\mathbf{x}_{2j}-\mathbf{x}_{1i}\|}{h}\right) \left(y_{2j} - \frac{e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}}{1+e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}} \right) \\ \vdots \\ \sum_{j=1}^{n_2} (x_{2dj} - x_{1di}) K\left(\frac{\|\mathbf{x}_{2j}-\mathbf{x}_{1i}\|}{h}\right) \left(y_{2j} - \frac{e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}}{1+e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}} \right) \end{pmatrix} \quad (12)$$

$$l''_{\mathbf{x}_{1i}}(\mathbf{d}_i; \boldsymbol{\beta}; \mathbf{y}_2) = \frac{\partial^2 l_{\mathbf{x}_{1i}}(\mathbf{d}_i; \boldsymbol{\beta}; \mathbf{y}_2)}{\partial \boldsymbol{\beta}^2} = - \sum_{j=1}^{n_2} K\left(\frac{\|\mathbf{x}_{2j}-\mathbf{x}_{1i}\|}{h}\right) \times \begin{pmatrix} \frac{e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}}{(1+e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}})^2} & \frac{(x_{21j}-x_{11i})e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}}{(1+e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}})^2} & \dots & \frac{(x_{2dj}-x_{1di})e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}}{(1+e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}})^2} \\ \frac{(x_{21j}-x_{11i})e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}}{(1+e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}})^2} & \frac{(x_{21j}-x_{11i})^2 e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}}{(1+e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}})^2} & \dots & \frac{(x_{21j}-x_{11i})(x_{2dj}-x_{1di})e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}}{(1+e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}})^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{(x_{2dj}-x_{1di})e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}}{(1+e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}})^2} & \frac{(x_{2dj}-x_{1di})(x_{21j}-x_{11i})e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}}{(1+e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}})^2} & \dots & \frac{(x_{2dj}-x_{1di})^2 e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}}{(1+e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}})^2} \end{pmatrix} \quad (13)$$

Thus, $l'_{\mathbf{x}_{1i}}(\mathbf{d}_i; \boldsymbol{\beta}; \mathbf{y}_2)$ is a $(d+1) \times 1$ vector and $l''_{\mathbf{x}_{1i}}(\mathbf{d}_i; \boldsymbol{\beta}; \mathbf{y}_2)$ is a $(d+1) \times (d+1)$ matrix, no longer depending on the majority response \mathbf{y}_2 .

Let $H_{m_2}(\mathbf{x}_1)$ denote the $d \times d$ Hessian matrix of $m_2(\cdot)$ and ν denote any diagonal element of $\int \mathbf{u}\mathbf{u}^T K(\mathbf{u})d\mathbf{u}$. Also let $A_{i0}, A_{i1}, \dots, A_{id}$ be the 1st row elements of $(-l''_{\mathbf{x}_{1i}}(\mathbf{d}_i; \boldsymbol{\beta}; \mathbf{y}_2))^{-1}$.

Theorem 1. Suppose $h = O(1/n_2^\alpha)$ where $0 < \alpha < 1$ and $n_2 h^{3d} \rightarrow \infty$, $h \rightarrow 0$, and $n_1 = O(n_2) \rightarrow \infty$. Furthermore, suppose that $|x_{1li}|$ and $|x_{2lj}|$ are bounded by M for all $l = 1, \dots, d$ and the range of x_{1li} is contained in that of $x_{2lj} \forall l$. The joint densities of \mathbf{x}_1 and \mathbf{x}_2 are both assumed to be bounded and continuous.

Then,

$$\text{var}(\bar{D}_{LOC}) \approx \frac{1}{n_1^2} \sum_{i=1}^{n_1} p_1(\mathbf{x}_{1i}) q_1(\mathbf{x}_{1i}) + \frac{1}{n_1^2} \sum_{k=1}^{n_1} \frac{e^{m_2(\mathbf{x}_{1k})}}{(1 + e^{m_2(\mathbf{x}_{1k})})^2} \sum_{i=1}^{n_1} \sigma_{ik} \frac{e^{m_2(\mathbf{x}_{1i})}}{(1 + e^{m_2(\mathbf{x}_{1i})})^2}$$

where

$$\begin{aligned} \sigma_{ik} &= \sum_{j=1}^{n_2} \left(\sum_{l=0}^d A_{il}(x_{2lj} - x_{1li}) \right) \left(\sum_{l=0}^d A_{kl}(x_{2lj} - x_{1lk}) \right) \\ &\quad \times K \left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h} \right) K \left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1k}\|}{h} \right) p_2(\mathbf{x}_{2j}) q_2(\mathbf{x}_{2j}) \end{aligned}$$

and the asymptotic distribution of $\bar{D}_{LOC} - \delta$ is normal. Furthermore, the asymptotic bias of \bar{D}_{LOC} is

$$\frac{h^2}{2} \int \nu \text{tr}(H_{m_2}(\mathbf{x}_1)) p_2(\mathbf{x}_1) q_2(\mathbf{x}_1) f_1(\mathbf{x}_1) d\mathbf{x}_1 + o(h^2)$$

and its asymptotic variance is $O(1/(n_1^{\alpha d} n_2 h^d))$.

The proof of Theorem 1 is given in the Appendix. As one would expect, the nonparametric estimate of disparity \bar{D}_{LOC} suffers from the bias-variance trade-off. That is, we see from the asymptotic bias and bound for the asymptotic variance of \bar{D}_{LOC} that smaller bandwidths result in smaller bias for \bar{D}_{LOC} at the expense of larger variance, and vice versa.

We propose the following test statistic for $H_0 : \delta = 0$:

$$\begin{aligned} t &= \bar{D}_{LOC} / \sqrt{\hat{\text{var}}(\bar{D}_{LOC})} \\ &= \bar{D}_{LOC} / \sqrt{\frac{1}{n_1^2} \sum_{i=1}^{n_1} \hat{p}_1(\mathbf{x}_{1i}) \hat{q}_1(\mathbf{x}_{1i}) + \frac{1}{n_1^2} \sum_{k=1}^{n_1} \left(\frac{e^{\hat{m}_2(\mathbf{x}_{1k})}}{(1 + e^{\hat{m}_2(\mathbf{x}_{1k})})^2} \sum_{i=1}^{n_1} \hat{\sigma}_{ik} \frac{e^{\hat{m}_2(\mathbf{x}_{1i})}}{(1 + e^{\hat{m}_2(\mathbf{x}_{1i})})^2} \right)} \end{aligned} \tag{14}$$

which has a standard normal asymptotic distribution under H_0 based on Theorem 1.

When all d covariates are discrete, the asymptotic form of \bar{D}_{LOC} and its variance become more explicit and interpretable. Corollary 2 summarizes these results.

Corollary 2. *Let all d covariates be discretely valued. Suppose all the conditions of Theorem 1 are satisfied. Let r be the number of covariate value combinations for the minority group and n_{1k} and n_{2k} denote the numbers of minority and majority members, respectively, who have the k -th covariate value combination where $k = 1, 2, \dots, r$. Then,*

$$\bar{D}_{LOC} = \frac{1}{n_1} \sum_{k=1}^r n_{1k} (\bar{y}_{1k} - \bar{y}_{2k}) \quad (15)$$

And $\bar{D}_{LOC} - \delta$ is asymptotically distributed as normal with zero mean and variance

$$\text{var}(\bar{D}_{LOC}) = \frac{1}{n_1^2} \sum_{k=1}^r n_{1k}^2 \left(\frac{\sum_{i=1}^{n_{1k}} p_1(\mathbf{x}_{1ki}) q_1(\mathbf{x}_{1ki})}{n_{1k}^2} + \frac{\sum_{j=1}^{n_{2k}} p_2(\mathbf{x}_{2kj}) q_2(\mathbf{x}_{2kj})}{n_{2k}^2} \right)$$

where \bar{y}_{1k} and \bar{y}_{2k} are the sample mean responses of minority and majority members who have the k -th covariate value combination, and \mathbf{x}_{1ki} and \mathbf{x}_{2kj} are the minority and majority covariate vectors with the k -th covariate combination, respectively.

This corollary implies that when all the covariates are discrete and the bandwidth is sufficiently small, \bar{D}_{LOC} simply reduces to the weighted average of differences in success probabilities between the minority and the majority members having the same covariate values. Considering the 2×2 table of success probability at each covariate value combination as a stratum, the procedure is analogous to Cochran's (1954) statistic for combining the tests of significance of the differences in proportions into a summary test, which adjusts for the effect of the covariates. The difference in the form of the test statistic between ours in Corollary 2 and Cochran's is that the latter uses $n_{1k}n_{2k}/(n_{1k} + n_{2k})$ as the weight, instead of n_{1k} in our case (15), and the denominator is $\sum_{k=1}^r (n_{1k}n_{2k})/(n_{1k} + n_{2k})$, instead of n_1 .

4 Simulation results

This section compares the statistical properties of the different estimators of disparity when there is a single covariate. We consider the following methods and models:

a) *Logistic regression with a minority indicator variable:*

$$\ln \left(\frac{p(x_i)}{1 - p(x_i)} \right) = \beta_0 + \beta_1 x_i + \beta_2 \times \text{minority}_i \quad (16)$$

b) *Parametric logistic PB regression:*

$$\ln \left(\frac{p_2(x_{2i})}{1 - p_2(x_{2i})} \right) = \beta_{20} + \beta_{21} x_{2i} \quad (17)$$

c) *Local Linear Logistic PB:*

$$\ln \left(\frac{p_2(x_{2i})}{1 - p_2(x_{2i})} \right) = m_2(x_{2i}) \quad (18)$$

The estimators of the average disparity, \bar{D} , for b) parametric logistic PB and c) Local Linear Logistic PB have been defined in (4) and (11), respectively. The estimated average disparity \bar{D} using a) ordinary logit regression with an indicator variable is obtained by fitting the following logit model:

$$\ln \left(\frac{p(x_i)}{1 - p(x_i)} \right) = \beta_0 + \beta_1 x_i + \beta_2 I_i$$

where I is a 0-1 variable indicating minority status (i.e., $I = 1$ for minority). The estimated average disparity for all minority members is obtained by:

$$\bar{D} = \frac{1}{n_1} \sum_{i=1}^{n_1} \left(\frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2}} - \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i}}} \right) \quad (19)$$

We will investigate two cases: the true underlying success probabilities are given by 1) linear logistic regression models and 2) curvilinear models. For both models, the covariate values are generated from $Gamma(2, 1/2)$ for the minority and $Gamma(3, 1/2)$ for the majority group; hence,

the majority group has higher mean and variance. The Gamma distribution was chosen because Bhattacharya and Gastwirth (1999) found that prior experience for a labor union admission case, *Berger v. Iron Workers, Local 201*, was approximately Gamma distributed. After we generate the covariate values, each observation receives a success probability based on the underlying probability model of our choice (i.e., logit link models as in (20) and (21) and curvilinear model as in (22) and (23)). Then, the binary response values are randomly generated from the Bernoulli distribution with the assigned success probabilities.

Because inference with the Local Linear Logistic PB method as well as the other two parametric logistic regression based methods utilize large sample results, a study of the sample size needed for these methods is also included. In the simulations, the majority sample size is incrementally increased from 40 to 250 and the minority sample size is set to be half that of the majority. For each sample size setting, the normality of the test statistics is checked with the robust Jarque Bera (JB) test (see Gel and Gastwirth, 2008, for details) and the Shapiro-Wilk test. Also, the bias, power and alpha level computations for each method for the two sided hypothesis (i.e., $H_0 : \delta = 0$ vs. $H_1 : \delta \neq 0$) are based on 500 iterations. The bandwidth for each sample size setting is determined by automatic bandwidth selection methods using the Akaike Information Criterion (AIC) and Likelihood Cross Validation (LCV) (see Loader, 1999, pp.68-69 for details).

4.1 Logistic regression model

The success probabilities are determined by:

$$\text{Minority: } p_1(x_1) = \frac{e^{-5+3x_1}}{1 + e^{-5+3x_1}} \tag{20}$$

$$\text{Majority: } p_2(x_2) = \frac{e^{-4+3.5x_2}}{1 + e^{-4+3.5x_2}} \tag{21}$$

The values of these parameters were chosen so that the true disparity δ is near -.15. Table 1 reports bias and power calculations. The first column shows majority and minority sample sizes for each simulation setting along with the true value of δ and the bandwidth chosen by AIC and LCV (average of the two). For each set of simulations, the bias, standard deviation of the bias, and power of the test are reported. In addition, the Type I error rate is calculated assuming H_0 using $\alpha = .05$; under H_0 the minority members' success probabilities were also determined by the majority probability function (21). The p -values for the normality tests of \bar{D}_{LOC} are given in Table 2. Both the robust JB and Shapiro-Wilk tests test the null hypothesis of normality; a p -value larger than the predetermined α level indicates that \bar{D} is approximately normally distributed.

In Table 1, all the three methods show a negligible amount of bias regardless of the sample sizes. Although there is some fluctuation, Local Linear Logistic PB requires at least 75 majority observations (preferably over 150) to yield a Type I error rate near .05 and for \bar{D}_{LOC} to be approximately normally distributed (see Table 2). On the other hand, both parametric logistic PB and ordinary logistic regression with an indicator variable yielded the Type I errors that are close to .05 and achieved normality of \bar{D} even with only 40 majority members. The power of ordinary logit regression with an indicator variable and parametric logistic PB is very similar when the majority sample size is 75 or more, whereas the power of Local Linear Logistic PB, with 75 or more majority observations, is slightly lower (by less than 4%) than the other two methods.

In summary, when the underlying success probability functions are precisely modeled by the logit link function, Local Linear Logistic PB yields slightly poorer results than the other two parametric methods. The difference was rather small and its power was slightly lower. However, its use requires a larger sample size than the two parametric methods.

Table 1: Bias, power, and Type I error of the three methods under logistic regression models in (20) and (21)

$[N_{maj}, N_{min}, \delta, h]$	LLLPB*		PLPB**		OLRI***	
	Bias (Stderr)	Power (TypeI)	Bias (Stderr)	Power (TypeI)	Bias (Stderr)	Power (TypeI)
[40, 20, -.164, .110]	.017 (.101)	.554 (.100)	.014 (.084)	.460 (.064)	.011 (.084)	.354 (.040)
[50, 25, -.201, .232]	-.001 (.100)	.720 (.164)	-.002 (.082)	.698 (.068)	-.003 (.082)	.662 (.048)
[75, 37, -.153, .225]	-.003 (.087)	.590 (.058)	.006 (.069)	.600 (.054)	.000 (.068)	.606 (.042)
[100, 50, -.187, .379]	-.027 (.100)	.840 (.050)	.001 (.060)	.876 (.048)	-.001 (.060)	.878 (.042)
[150, 75, -.154, .521]	.013 (.081)	.906 (.060)	.001 (.045)	.934 (.058)	-.003 (.044)	.938 (.046)
[200, 100, -.176, .482]	-.000 (.079)	.986 (.060)	.000 (.040)	.986 (.058)	-.004 (.039)	.992 (.046)
[250, 125, -.187, .583]	-.000 (.075)	.988 (.064)	.001 (.037)	1.000 (.062)	-.000 (.036)	1.000 (.056)

* LLLPB = Local Linear Logistic PB

** PLPB = Parametric logistic PB: $logit = \beta_{20} + \beta_{21} * x$

*** OLRI = Ordinary logistic regression with indicator: $logit = \beta_0 + \beta_1 * x + \beta_2 * Indicator$

Table 2: Normality tests for \bar{D} of the three methods under logistic regression models in (20) and (21)

$[N_{maj}, N_{min}, \delta, h]$	LLLPB		PLPB		OLRI	
	P-value RJB*	P-value SW**	P-value RJB*	P-value SW**	P-value RJB*	P-value SW**
[40, 20, -.164, .110]	.000	.000	0.129	0.068	0.182	0.220
[50, 25, -.201, .232]	0.293	0.270	0.224	0.223	0.089	0.071
[75, 37, -.153, .225]	0.000	0.001	0.972	0.939	0.292	0.426
[100, 50, -.187, .379]	0.000	0.001	0.071	0.054	0.963	0.609
[150, 75, -.154, .521]	0.674	0.842	0.643	0.909	0.432	0.526
[200, 100, -.176, .482]	0.030	0.002	0.103	0.022	0.224	0.184
[250, 125, -.187, .583]	0.916	0.815	0.552	0.468	0.501	0.430

* RJB = Robust Jarque Bera test for normality

** SW = Shapiro-Wilk test for normality

4.2 Curvilinear model

To examine the potential gain in using our nonparametric method, we use the following success probability functions (curvilinear):

$$\text{Minority: } p_1(x_1) = .9 * x_1 - .5 * x_1^{1.4} \quad (22)$$

$$\text{Majority: } p_2(x_2) = 1.1 * x_2 - .5 * x_2^{1.5} \quad (23)$$

The models in (16), (17), and (18) are fitted to the data and the accuracy of the resulting inferences are examined. Again, the values of the parameters in (22) and (23) are chosen so that δ remains near -.15.

Tables 3 and 4 summarize the results. As expected, parametric logistic PB and ordinary logistic regression with an indicator variable now produce noticeably biased estimates of δ and their Type I error is not preserved at .05, especially for larger sample sizes. On the other hand, Local Linear Logistic PB yields a negligible amount of bias and a reasonable level of Type I error when the majority sample size is 75 or greater. Although there is some fluctuation in the results, the normality tests show that the distribution of \bar{D}_{LOC} seems to be normal when the majority sample size is large.

Notice that when the logistic model is not the true one, both parametric logit PB and ordinary logistic regression with an indicator variable overestimate the disparity by a practically meaningful amount (e.g., .1 in some cases). In contrast, Local Linear Logistic PB was able to fit the model reasonably well and provide a nearly unbiased estimate of δ .

Table 3: Bias and power of the three methods under curvilinear models in (22) and (23)

$[N_{maj}, N_{min}, \delta, h]$	LLLPB*		PLPB**		OLRI***	
	Bias (Stderr)	Power (TypeI)	Bias (Stderr)	Power (TypeI)	Bias (Stderr)	Power (TypeI)
[40, 20, -.164, .205]	0.005 (0.201)	0.156 (0.128)	-0.027 (0.138)	0.290 (0.092)	-0.024 (0.137)	0.244 (0.048)
[50, 25, -.172, .365]	-0.053 (0.164)	0.214 (0.084)	-0.103 (0.127)	0.580 (0.060)	-0.082 (0.127)	0.460 (0.052)
[75, 37, -.205, .466]	0.007 (0.101)	0.552 (0.066)	0.020 (0.098)	0.464 (0.072)	0.022 (0.098)	0.420 (0.040)
[100, 50, -.170, .501]	-0.016 (0.089)	0.574 (0.058)	-0.077 (0.086)	0.802 (0.236)	-0.066 (0.086)	0.746 (0.206)
[150, 75, -.162, .458]	-0.015 (0.074)	0.646 (0.058)	-0.103 (0.069)	0.946 (0.236)	-0.107 (0.066)	0.964 (0.228)
[200, 100, -.181, .365]	-0.002 (0.062)	0.844 (0.042)	-0.057 (0.061)	0.974 (0.364)	-0.050 (0.061)	0.972 (0.236)
[250, 125, -.168, .345]	-0.005 (0.055)	0.874 (0.068)	-0.089 (0.052)	0.996 (0.262)	-0.084 (0.051)	0.998 (0.172)

* LLLPB = Local Linear Logistic PB

** PLPB = Parametric logistic PB: $logit = \beta_{20} + \beta_{21} * x$

*** OLRI = Ordinary logistic regression with indicator: $logit = \beta_0 + \beta_1 * x + \beta_2 * Indicator$

Table 4: Normality tests for \bar{D} of the three methods under curvilinear models in (22) and (23)

$[N_{maj}, N_{min}, \delta, h]$	LLPB		PLPB		OLRI	
	P-value RJB*	P-value SW**	P-value RJB*	P-value SW**	P-value RJB*	P-value SW**
[40, 20, -.164, .205]	.000	.000	.040	.024	.935	.786
[50, 25, -.172, .365]	.231	.384	.976	.927	.505	.721
[75, 37, -.205, .466]	.916	.878	.600	.535	.719	.637
[100, 50, -.170, .501]	.011	.008	.915	.208	.140	.024
[150, 75, -.162, .458]	.001	.004	.151	.234	.107	.196
[200, 100, -.181, .365]	.179	.249	.272	.200	.341	.209
[250, 125, -.168, .345]	.952	.918	.385	.678	.736	.667

* RJB = Robust Jarque Bera test for normality

** SW = Shapiro-Wilk test for normality

5 Re-analysis of data from *Berger v. Iron Workers Local*

201

The *Berger v. Iron Workers Local 201*³ case concerned the admission exam used by Local 201 of the Iron Workers Union. The plaintiffs sued the union claiming the exam discriminatorily denied black rodmen the benefits of the union membership as black applicants had a statistically significantly lower pass rate of the exam than whites. Before taking the exam, applicants were required to serve as apprentices or work as part of an auxiliary labor pool: the auxiliary workers were called upon in periods of high demand when there was an insufficient number of union members. The average prior experience of black applicants exceeded that of whites; hence, the difference in passing rates cannot be explained by the amount of prior experience.

The data consisted of 35 black applicants and 34 white applicants with their hours of prior experience. In order to avoid the problem of extrapolation (the maximum value of hours for white members was 8,667.2, whereas that for black applicants was 10,433.6), the hours of three black applicants were truncated to 8,600; originally, they had 9,770.9, 9,905.8, and 10,433.6, respectively. All three of these black applicants failed the exam.

At trial, plaintiffs compared the fraction ($12/35=.343$) of blacks passing the exam to that of whites ($24/34=.706$). The χ^2 test for independence applied to this 2×2 table (white/black vs. pass/not pass) yielded the test statistic value of 7.71 with p -value= .0055. Since one expects that more experience should increase one's probability of passing, the fact that blacks had more hours of prior work experience suggests that the true disparity is even greater than the one obtained from the simple comparison of the proportions of passing. Table 5 presents summary statistics

³*Berger v. Iron Workers Local 201*, 42 FEP Cases 1161 (D.D.C. 1985), 843 F.2d 1395 (D.C. Cir. 1988)

Table 5: Summary statistics of “Hours” for white and black

	Mean	Median	Minimum	1st Quartile	3rd Quartile	Maximum	Std. Dev.
White (n=34)	2763	2528	1019	1883	3309	8667	1380.83
Black (n=35)	5441	5612	1408	3900	7341	8600	2144.05

showing that the typical black applicant has about twice the amount of experience of a typical white applicant. The Wilcoxon rank sum test yields the test statistic 1,022 with a p -value of almost 0, which indicates that the distribution of number of hours for blacks is shifted to the right of the hour distribution of whites.

Figure 1 shows the proportion of passing the exam by three groups created by hours. The three groups were created by dividing the data at 33rd percentile and 67th percentile values of hours:

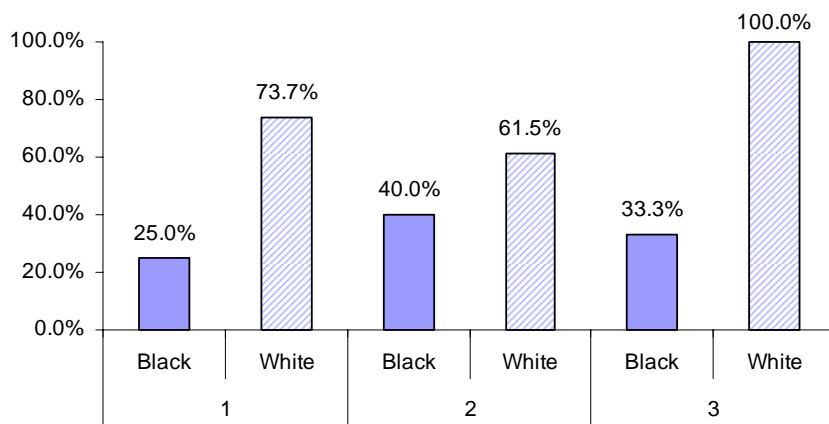
Group 1: $0 < Hours \leq 2580.27$

Group 2: $2580.27 < Hours \leq 4488.50$

Group 3: $4488.50 < Hours \leq 8667.20$

Figure 1 shows that in each group the pass rate for whites is higher than that for blacks. It also shows that the white pass rate decreases from Group 1 to Group 2 but reaches up to 100% in Group 3. On the other hand, for black applicants, having more prior experience does not seem to help their pass rate much; the pattern of pass rate change appears much flatter than that for the whites. We would like to caution, though, that since the sample sizes are small for both groups, the patterns we observe in Figure 1 could change by altering the threshold values of hours to define the groups.

Figure 1: Proportion of passing the exam by groups defined by “hours”



	1		2		3	
	Black	White	Black	White	Black	White
# Pass	1	14	4	8	7	2
# Fail	3	5	6	5	14	0
Total #	4	19	10	13	21	2
% pass	25.0%	73.7%	40.0%	61.5%	33.3%	100.0%

As before, the racial disparity will be examined using: a) ordinary logistic regression with an indicator variable, b) parametric logit PB, and c) Local Linear Logistic PB.

a) Ordinary logistic regression with an indicator variable

The ordinary logistic regression model:

$$\ln \left(\frac{p(x_i)}{1 - p(x_i)} \right) = \beta_0 + \beta_1 \text{Hours}_i + \beta_2 \text{Black}_i$$

is fitted to the data where *Black* is the indicator variable for black applicants.

b) Parametric logit PB

In applying parametric logit PB, we fit the following logistic regression model to the data on whites:

$$\ln \left(\frac{p_w(x_{wi})}{1 - p_w(x_{wi})} \right) = \beta_{w0} + \beta_{w1} \text{Hours}_{wi}$$

then the differences between the observed and the predicted response values of blacks are used to compute the average estimated disparity as defined in (4).

c) Local Linear Logistic PB

For the Local Linear Logistic PB method, we use .3 as the bandwidth (fraction of data) chosen by the automatic AIC and GCV criteria and by taking into account the relatively small sample size of whites.

Table 6 summarizes the results obtained from these three methods. Since hours of prior experience can be expected to be an important predictor, we include it in the model even though it was not statistically significant in both the ordinary logistic regression with the indicator and the parametric logit PB methods.

The estimated coefficient, $\hat{\beta} = -1.8343$, of the race indicator is highly significant ($t = -2.703$ with $p\text{-value} = .0069$). This translates to an estimated odds ratio of .067. The estimated

Table 6: Summary of results from applying the three methods to the *Berger v. Iron Workers Local*

201 data

	OLRI		PLPB*		LLLPB**	
	$\hat{\beta}$	z value	$\hat{\beta}$	z value	$\hat{\beta}$	z value
	(Std.error)	(p-value)	(Std.error)	(p-value)	(Std.error)	(p-value)
Intercept	.5760 (.5491)	1.049 (.2941)	-.4096 (1.1304)	-.362 (.717)	— —	— —
Hours	.0001 (.0001)	0.738 (.4605)	.0005 (.0004)	1.142 (.253)	— —	— —
Black	-1.8343 (.6787)	-2.703 (.0069)	— —	— —	— —	— —
\bar{D}	-.4187		-.5287		-.5712	
Test stat	-2.703		-4.7123		-7.074	
P-value	.0069		.0000		.0000	

* The results related to estimated coefficients under PLPB are for the white only model.

** The white local linear logistic regression model was fitted using bandwidth=.3.

OLRI: Ordinary logistic regression with an indicator variable

PLPB: Parametric logit PB

LLLPB: Local linear logistic PB

disparity \bar{D} as defined in (19) is:

$$\frac{1}{n_b} \sum_{i=1}^{n_b} \left(\frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \text{Hours}_{bi} + \hat{\beta}_2}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \text{Hours}_{bi} + \hat{\beta}_2}} - \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \text{Hours}_{bi}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \text{Hours}_{bi}}} \right) = -.4187$$

This means that the probability that a black applicant would pass the exam was about .42 less than that of a white with the same prior experience.

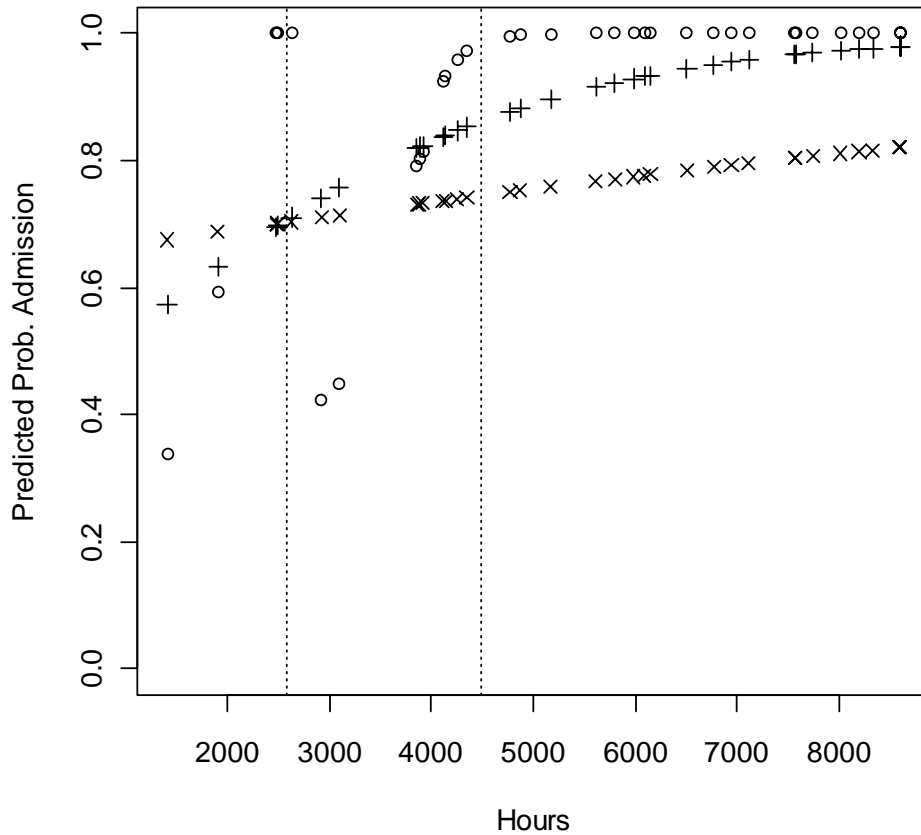
The estimated disparity obtained from the parametric logistic PB is -.5278, while that yielded by Local Linear Logistic PB is -.5712: both estimates are highly significant. All three results show a significant disparity against black applicants.

To assess which method “fits” the data best, it is useful to graph their predicted probabilities. Figure 2 presents the predicted probabilities of black applicants if they were white obtained from the three methods. The vertical dotted lines represent the threshold values of hours defining the three groups in Figure 1. This graph clearly reveals the differences among the three methods. Ordinary logistic regression with an indicator variable fits an almost linear line but does not reach probability one even at the maximum number of hours of experience. Parametric logit PB starts lower than ordinary logistic regression with an indicator variable and fits a smooth quadratic type curve that almost reaches probability one for higher values of hours.

On the other hand, Local Linear Logistic PB fits a non-monotone curve around lower values of hours and better reflects the patten observed in Figure 1 where the white pass rate is lower in Group 2 (61.5%) than in Group 1 (73.7%). This demonstrates the ability of local linear logistic regression to better adjust to the data and model a non-monotone relationship.

The results from our simulation studies indicate that the Local Linear Logistic PB method requires at least 75 observations for \bar{D}_{LOC} to achieve asymptotic normality. Since the majority sample size in this Berger data is 35, the accuracy of its inferential result needs to be investigated.

Figure 2: Blacks' predicted probabilities using the three regression methods



o = Blacks' predicted prob's if they were white from **Local Linear Logistic PB**
x = Blacks' predicted prob's if they were white from **ordinary logit model with the indicator variable**
+ = Blacks' predicted prob's if they were white from **parametric logit PB**

We apply a bootstrap method and resample 35 blacks and 34 whites from the original data. Then, we fit the local linear logistic regression model to white only data and calculate the predicted probabilities for blacks from this white only model. Using these predicted probabilities, we randomly generate the 1-0 response values for blacks. Finally, we calculate \bar{D}_{LOC} by taking the average of the differences between the randomly generated 1-0 response values and their predicted probabilities. We repeat this 5,000 times and create the bootstrap null distribution of \bar{D}_{LOC} .

This null distribution ranges from -.209 to .186 and its mean and standard deviation are approximately 0 and 0.035. Since the observed value, $\bar{D}_{LOC} = -.5712$, is well outside this range, we can confidently conclude that the hypothesis testing from Local Linear Logistic PB gives a highly significant result.

Finally, Judge R. A. Posner in *Allen v. Seidman*⁴ noted that since it is easy to criticize regression studies submitted into evidence for omitting a possible predictor, the critic should demonstrate that including the additional covariate “explains” the disparity. The *Berger* case illustrates this point as the disparity between black and white pass rates increased, rather than decreased, when prior experience was included.

6 Concluding Remarks

This paper extended the Peters-Belson approach to local linear logistic regression. This method provides a more flexible way to model the relationship between a response and predictor variables. Its usefulness was illustrated on data from an actual equal employment case. Since the method does not rely on parametric assumptions, it does not rely on a particular specified model. Furthermore,

⁴*Allen v. Sideman*, 881 F.2d 375 (7th Cir. 1989)

when a reasonably large sample size is available, Local Linear Logistic PB loses little power compared to the other parametric methods even when the data follow the assumed parametric model. On the other hand, when the model generating the data differs from the assumed parametric model, Local Linear Logistic PB yields more reliable inferences. This was seen both in the simulation study and the re-analysis of the data from *Berger v. Iron Workers Local 201*. While this paper was motivated by a legal application, the method can be used in epidemiology where the exposure rate of cases is compared to that of an otherwise similar control group.

A Proof of Theorem 1

Proof. First, we asymptotically approximate the distribution of the following quantity:

$$\begin{aligned} \sum_{i=1}^{n_1} (\hat{p}_2(\mathbf{x}_{1i}) - p_2(\mathbf{x}_{1i})) &= \sum_{i=1}^{n_1} \left(\frac{e^{\hat{m}_2(\mathbf{x}_{1i})}}{1 + e^{\hat{m}_2(\mathbf{x}_{1i})}} - \frac{e^{m_2(\mathbf{x}_{1i})}}{1 + e^{m_2(\mathbf{x}_{1i})}} \right) \\ &= \sum_{i=1}^{n_1} (g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))) \end{aligned} \quad (24)$$

where $g(m_2(\mathbf{x}_{1i})) = e^{m_2(\mathbf{x}_{1i})} / (1 + e^{m_2(\mathbf{x}_{1i})})$. We will compute this distribution for two different cases:

(I) at least one of the d covariates is continuous and (II) all the d covariates are discrete.

Case I: At least one covariate is continuous

Here, $g(\hat{m}_2(\mathbf{x}_{1i}))$, for $i = 1, 2, \dots, n_1$, are locally dependent random variables. Also, $g(\hat{m}_2(\mathbf{x}_{1i}))$ and $g(\hat{m}_2(\mathbf{x}_{1j}))$ where $i \neq j$ are dependent, if the local neighborhoods of \mathbf{x}_{1i} and \mathbf{x}_{1j} overlap. If they do not overlap, $g(\hat{m}_2(\mathbf{x}_{1i}))$ and $g(\hat{m}_2(\mathbf{x}_{1j}))$ are independent.

Using similar notation as Chen (2005), let I_1 be a finite index set of cardinality n_1 . Also

let

$$\xi_i = \frac{g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))}{(\text{var}(\sum_{i=1}^{n_1} g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))))^{1/2}} \quad (25)$$

$$W = \sum_{i=1}^{n_1} \xi_i = \sum_{i=1}^{n_1} \frac{g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))}{(\text{var}(\sum_{i=1}^{n_1} g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))))^{1/2}} \quad (26)$$

Applying Theorem 4 of Fan, Heckman, and Wand (1995, p.146), we obtain $E(W) = 0$ asymptotically, and clearly $\text{var}(W) = 1$. Then, assumption (LD1) discussed by Chen (2005) on p. 17 is met; that is,

(LD1) For each $i \in I_1$, there exists $A_i \subset I_1$ such that ξ_i and $\xi_{A_i^c}$ are independent.

Here, A_i is a index set of the minority members whose neighborhoods overlap with that of the i -th minority member; hence, their $g(\hat{m}_2(\mathbf{x}_{1i}))$'s are dependent with $g(\hat{m}_2(\mathbf{x}_{1i}))$. Also, let m_i be the cardinality of A_i and $m = \max_{1 \leq i \leq n_1} m_i$. When (LD1) is met, by Stein's method of approximating the distance between two distributions, Theorems 3.1 and 3.4 of Chen (2005) give

$$\sup_z |P(W \leq z) - \phi(z)| \leq 2\psi^{1/2} \quad (27)$$

where $\phi(z)$ is the distribution function of the standard normal distribution and

$$\psi = 4E \left| \sum_{i=1}^{n_1} \left(\xi_i \sum_{j \in A_i} \xi_j - E(\xi_i \sum_{j \in A_i} \xi_j) \right) \right| + \sum_{i=1}^{n_1} E \left| \xi_i \left(\sum_{j \in A_i} \xi_j \right)^2 \right| \quad (28)$$

Thus, our goal is to show that ψ goes to zero at some rate, which proves the asymptotic normality of W . To show this, using Theorem 4 of Fan, Heckman, and Wand (1995, p.146) that states $\text{var}(g(\hat{m}_2(\mathbf{x}_{1i})))$ is of the order of magnitude $(n_2 h^d)^{-1}$, one can show that $\text{var}(\sum_{i=1}^{n_1} [g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))])$ is of the order of magnitude $m h^{-d}$. Then,

$$\xi_i = (g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))) \times O((h^d/m)^{1/2}) \quad (29)$$

Furthermore, the same theorem shows that $g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))$ is asymptotically normally distributed with the mean approaching to zero at $O(h^2)$. Thus, $(g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i})))^2 / \text{var}(g(\hat{m}_2(\mathbf{x}_{1i})))$ is approximately χ^2 distributed with one degree of freedom. Hence,

$$\text{var}((g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i})))^2) = 2[\text{var}(g(\hat{m}_2(\mathbf{x}_{1i})))^2] = O(1/n_2^2 h^{2d}) \quad (30)$$

The first expectation of ψ in (28) is

$$\begin{aligned} & 4E \left| \sum_{i=1}^{n_1} \left(\xi_i \sum_{j \in A_i} \xi_j - E(\xi_i \sum_{j \in A_i} \xi_j) \right) \right| \\ & \leq 4 \left(E \left[\sum_{i=1}^{n_1} \left(\xi_i \sum_{j \in A_i} \xi_j - E(\xi_i \sum_{j \in A_i} \xi_j) \right) \right]^2 \right)^{1/2} \\ & = 4 \left(\sum_{i=1}^{n_1} E \left[\xi_i \sum_{j \in A_i} \xi_j - E(\xi_i \sum_{j \in A_i} \xi_j) \right]^2 \right. \\ & \quad \left. + \sum_{i \neq l}^{n_1} \sum_{l=1}^{n_1} E \left[\left(\xi_i \sum_{j \in A_i} \xi_j - E(\xi_i \sum_{j \in A_i} \xi_j) \right) \left(\xi_l \sum_{j \in A_l} \xi_j - E(\xi_l \sum_{j \in A_l} \xi_j) \right) \right] \right)^{1/2} \\ & = 4 \left(\sum_{i=1}^{n_1} \text{var} \left(\xi_i \sum_{j \in A_i} \xi_j \right) + \sum_{i \neq l}^{n_1} \sum_{l=1}^{n_1} \text{cov} \left(\xi_i \sum_{j \in A_i} \xi_j, \xi_l \sum_{j \in A_l} \xi_j \right) \right)^{1/2} \end{aligned} \quad (31)$$

$$\text{var} \left(\xi_i \sum_{j \in A_i} \xi_j \right) = \sum_{j=1}^{m_i} \text{var}(\xi_i \xi_j) + \sum_{j \neq l}^{m_i} \sum_{l=1}^{m_i} \text{cov}(\xi_i \xi_j, \xi_i \xi_l) \quad (32)$$

$$\begin{aligned} \text{var}(\xi_i \xi_j) &= E(\xi_i \xi_j - E(\xi_i \xi_j))^2 \\ &= O(h^{2d}/m^2) \text{var}((g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))) (g(\hat{m}_2(\mathbf{x}_{1j})) - g(m_2(\mathbf{x}_{1j})))) \\ &= O(h^{2d}/m^2) E((g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i})))^2 (g(\hat{m}_2(\mathbf{x}_{1j})) - g(m_2(\mathbf{x}_{1j})))^2) \\ &\quad - O(h^{2d}/m^2) [E((g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))) (g(\hat{m}_2(\mathbf{x}_{1j})) - g(m_2(\mathbf{x}_{1j}))))]^2 \end{aligned} \quad (33)$$

$$\begin{aligned}
& E((g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i})))^2(g(\hat{m}_2(\mathbf{x}_{1j})) - g(m_2(\mathbf{x}_{1j}))))^2) \\
&= \text{cov}((g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i})))^2, (g(\hat{m}_2(\mathbf{x}_{1j})) - g(m_2(\mathbf{x}_{1j}))))^2) \\
&\quad + E((g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i})))^2)E((g(\hat{m}_2(\mathbf{x}_{1j})) - g(m_2(\mathbf{x}_{1j}))))^2) \\
&\leq (\text{var}((g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))))^2)\text{var}((g(\hat{m}_2(\mathbf{x}_{1j})) - g(m_2(\mathbf{x}_{1j}))))^2)^{1/2} \\
&\quad + \text{var}(g(\hat{m}_2(\mathbf{x}_{1i})))\text{var}(g(\hat{m}_2(\mathbf{x}_{1j}))) \\
&= O((n_2h^d)^{-2}) \tag{34}
\end{aligned}$$

where the last equality follows from (30) and the result $g(\hat{m}_2(\mathbf{x}_{1i}))$ is of the order of magnitude $(n_2h^d)^{-1}$.

$$\begin{aligned}
& [E((g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))) (g(\hat{m}_2(\mathbf{x}_{1j})) - g(m_2(\mathbf{x}_{1j}))))]^2 \\
&= (\text{cov}((g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))), (g(\hat{m}_2(\mathbf{x}_{1j})) - g(m_2(\mathbf{x}_{1j}))))))^2 \\
&\leq ((\text{var}(g(\hat{m}_2(\mathbf{x}_{1i})))\text{var}(g(\hat{m}_2(\mathbf{x}_{1j}))))^{1/2})^2 \\
&= O((n_2h^d)^{-2}) \tag{35}
\end{aligned}$$

Using (34) and (35) in (33),

$$E(\xi_i \xi_j - E(\xi_i \xi_j))^2 \leq O\left(\frac{1}{m^2 n_2^2}\right) \tag{36}$$

Similarly, one can show that

$$\begin{aligned}
\text{cov}(\xi_i \xi_j, \xi_i \xi_l) &= E(\xi_i^2 \xi_j \xi_l) - E(\xi_i \xi_j)E(\xi_i \xi_l) \\
&\leq O\left(\frac{h^{2d}}{m^2}\right) \times O\left(\frac{1}{n_2^2 h^{2d}}\right) + O\left(\frac{h^{2d}}{m^2}\right) \times O\left(\frac{1}{n_2^2 h^{2d}}\right) \\
&= O\left(\frac{1}{m^2 n_2^2}\right) \tag{37}
\end{aligned}$$

Using (36) and (37) in (32) yields

$$\text{var}\left(\xi_i \sum_{j \in A_i} \xi_j\right) \leq O\left(\frac{1}{n_2^2}\right) \tag{38}$$

Next, the second piece of (31) is

$$\sum_{i \neq l}^{n_1} \sum_{j \in A_i}^{n_1} \text{cov} \left(\xi_i \sum_{j \in A_i} \xi_j, \xi_l \sum_{j \in A_l} \xi_j \right) = \sum_{i=1}^{n_1} \sum_{\substack{l \in A_i^* \\ i \neq l}} \text{cov} \left(\xi_i \sum_{j \in A_i} \xi_j, \xi_l \sum_{j \in A_l} \xi_j \right) \quad (39)$$

where A_i^* is an index set of the minority members whose $\xi_l \sum_{j \in A_l} \xi_j$ is dependent with $\xi_i \sum_{j \in A_i} \xi_j$.

In other words, these minority members are within $4h$ in the vector distance $\|\mathbf{x}_{1i} - \mathbf{x}_{1l}\|$ from \mathbf{x}_{1i} .

Let the cardinality of A_i^* be m_i^* and $m^* = \max_{1 \leq i \leq n_1} m_i^*$. Then, (39) becomes

$$\sum_{i=1}^{n_1} \sum_{\substack{l \in A_i^* \\ i \neq l}} O(m^2 \text{cov}(\xi_i \xi_j, \xi_l \xi_k)) \leq O\left(\frac{n_1 m^* m^2}{m^2 n_2^2}\right) = O\left(\frac{m^*}{n_2}\right) \quad (40)$$

Finally, using (38) and (40) in (31) yields

$$4E \left| \sum_{i=1}^{n_1} \left(\xi_i \sum_{j \in A_i} \xi_j - E(\xi_i \sum_{j \in A_i} \xi_j) \right) \right| \leq O\left(\frac{1}{n_2}\right) + O\left(\frac{m^*}{n_2}\right) = O\left(\frac{m^*}{n_2}\right) \quad (41)$$

Recall that m_i^* is the number of minority observations that lie within $4h$ in the vector distance $\|\mathbf{x}_{1i} - \mathbf{x}_{1l}\|$ from \mathbf{x}_{1i} . In other words, the space which contains the m_i^* minority members can be defined as a d dimensional ball with \mathbf{x}_{1i} as its center and $4h$ as its radius. The volume of this ball can be obtained by $(4h)^d \pi^{d/2} / \Gamma(d/2 + 1)$ (see Wade, 2004, p 444), where d is the number of covariates and is finite. Let the joint density of \mathbf{x}_1 be bounded by B . Then, with $h = 1/n_1^\alpha$ where $0 < \alpha < 1$,

$$m_i^* \leq (4h)^d \pi^{d/2} / \Gamma(d/2 + 1) \times B \times n_1 = (4n_1^{-\alpha})^d \pi^{d/2} / \Gamma(d/2 + 1) \times B \times n_1 = O(n_1^{1-\alpha d}) \quad \forall i \quad (42)$$

Thus, (41) is $O(n_1^{-\alpha d})$ which goes to zero.

Now, the second component of ψ in (28) is

$$\begin{aligned} & \sum_{i=1}^{n_1} E \left| \xi_i \left(\sum_{j \in A_i} \xi_j \right)^2 \right| \\ &= \sum_{i=1}^{n_1} E \left| O(h^{3d/2}/m^{3/2}) \times (g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))) \left(\sum_{j \in A_i} g(\hat{m}_2(\mathbf{x}_{1j})) - g(m_2(\mathbf{x}_{1j})) \right)^2 \right| \quad (43) \end{aligned}$$

Factoring out the squared term in (43), we have four types of components:

- (a) one of $(g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i})))^3 \times O(h^{3d/2}/m^{3/2})$
- (b) $m_i - 1$ of $(g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))) (g(\hat{m}_2(\mathbf{x}_{1j})) - g(m_2(\mathbf{x}_{1j})))^2 \times O(h^{3d/2}/m^{3/2})$
- (c) $2(m_i - 1)$ of $(g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i})))^2 (g(\hat{m}_2(\mathbf{x}_{1j})) - g(m_2(\mathbf{x}_{1j}))) \times O(h^{3d/2}/m^{3/2})$
- (d) $m_i^2 - 3m_i + 2$ of $(g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))) (g(\hat{m}_2(\mathbf{x}_{1j})) - g(m_2(\mathbf{x}_{1j}))) (g(\hat{m}_2(\mathbf{x}_{1l})) - g(m_2(\mathbf{x}_{1l}))) \times O(h^{3d/2}/m^{3/2})$

Each of the four forms has the same rate, so we will show it using only (a).

$$\begin{aligned} & E \left| (g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i})))^3 \times O(h^{3d/2}/m^{3/2}) \right| \\ & \leq O\left(\frac{h^{3d/2}}{m^{3/2}}\right) \sqrt{E(g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i})))^4 E(g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i})))^2} \end{aligned}$$

by the Cauchy-Schwarz inequality. Using (30),

$$\begin{aligned} & E \left| (g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i})))^3 \times O(h^{3d/2}/m^{3/2}) \right| \\ & \leq O\left(\frac{h^{3d/2}}{m^{3/2}}\right) \sqrt{O(1/(n_2^2 h^{2d})) O(1/(n_2 h^d))} = O\left(\frac{1}{m^{3/2} n_2^{3/2}}\right) \end{aligned} \tag{44}$$

Using (44) for all the components in (a) through (d) above, we get

$$\sum_{i=1}^{n_1} E \left| \xi_i \left(\sum_{j \in A_i} \xi_j \right)^2 \right| \leq n_1 \left(O\left(\frac{1}{m^{3/2} n_2^{3/2}}\right) + O\left(\frac{m}{m^{3/2} n_2^{3/2}}\right) + O\left(\frac{m}{m^{3/2} n_2^{3/2}}\right) + O\left(\frac{m^2}{m^{3/2} n_2^{3/2}}\right) \right) \tag{45}$$

Recall $m = \max_{1 \leq i \leq n_1} m_i$ where m_i is the number of minority members (including the i -th minority) whose estimated $m_2(\cdot)$ depends on some common majority members that are used to find the estimated $m_2(\cdot)$ for the i -th minority member. Those m_i observations are within $2h$ in the vector

distance $\|\mathbf{x}_{1k} - \mathbf{x}_{1i}\|$ from \mathbf{x}_{1i} . In other words, the space which contains the m_i minority members can be defined as a d dimensional ball with \mathbf{x}_{1i} as its center and $2h$ as its radius. The volume of this ball equals $(2h)^d \pi^{d/2} / \Gamma(d/2 + 1)$ (see Wade, 2004, p 444), where d is the number of covariates and is finite. We let the joint density of \mathbf{x}_1 be bounded by B . Then, with $h = 1/n_1^\alpha$, where $0 < \alpha < 1$,

$$m_i \leq (2h)^d \pi^{d/2} / \Gamma(d/2 + 1) \times B \times n_1 = (2n_1^{-\alpha})^d \pi^{d/2} / \Gamma(d/2 + 1) \times B \times n_1 = O(n_1^{1-\alpha d}) \quad \forall i \quad (46)$$

Using (46) in (45),

$$\sum_{i=1}^{n_1} E \left| \xi_i \left(\sum_{j \in A_i} \xi_j \right)^2 \right| \leq O \left(\frac{n_1 m^2}{m^{3/2} n_2^{3/2}} \right) = O \left(\frac{m^{1/2}}{n_1^{1/2}} \right) = O \left(n_1^{-1/2\alpha d} \right) \quad (47)$$

Combining (41) and (47) in (27), we get

$$\sup_z |P(W \leq z) - \phi(z)| \leq O \left(n_1^{-1/4\alpha d} \right) \quad (48)$$

Therefore, the distribution of W can be approximated by the standard normal distribution and, as a result, we can claim that

$$\sum_{i=1}^{n_1} (\hat{p}_2(\mathbf{x}_{1i}) - p_2(\mathbf{x}_{1i})) = \sum_{i=1}^{n_1} (g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))) \stackrel{D}{\sim} Normal \quad (49)$$

Case II: All the covariates are discrete

When all covariates are bounded and discrete, the potential values of the vector distance $\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|$ will be discrete and finite as well. As $h \rightarrow 0$, h becomes smaller than the minimum distance between any two discrete values of $\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|$. Thus, only those majority observations \mathbf{x}_{2j} having the same covariate values as the design point \mathbf{x}_{1i} will receive non-zero weight in computing $\hat{m}_2(\mathbf{x}_{1i})$. When this occurs, the local log-likelihood (10) becomes equivalent to

$$l_{\mathbf{x}_{1i}}(\mathbf{d}_i \boldsymbol{\beta}; \mathbf{y}_2) = \sum_{j=1}^{n_2} K \left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h} \right) (y_{2j} m_2(\mathbf{x}_{1i}) - \ln(1 + e^{m_2(\mathbf{x}_{1i})})) \quad (50)$$

This is the local log-likelihood for local constant regression. Then, the estimator of $p_2(\mathbf{x}_{1i})$ can be explicitly solved as

$$\hat{p}_2(\mathbf{x}_{1i}) = \frac{\sum_{j=1}^{n_2} K\left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h}\right) y_{2j}}{\sum_{j=1}^{n_2} K\left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h}\right)} \quad (51)$$

which is the local constant estimator.

Without loss of generality, assume $E(\hat{p}_2(\mathbf{x}_{1i})) = 0$ for all i .

$$\sum_{i=1}^{n_1} \hat{p}_2(\mathbf{x}_{1i}) = \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} \frac{1}{\sum_{j=1}^{n_2} K\left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h}\right)} K\left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h}\right) y_{2j} \quad (52)$$

Thus,

$$\text{var}\left(\sum_{i=1}^{n_1} \hat{p}_2(\mathbf{x}_{1i})\right) = \sum_{j=1}^{n_2} \left(\sum_{i=1}^{n_1} \frac{K\left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h}\right)}{\sum_{j=1}^{n_2} K\left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h}\right)}\right)^2 p_2(\mathbf{x}_{2j}) q_2(\mathbf{x}_{2j}) \quad (53)$$

Since $K(\cdot)$ is bounded,

$$\text{var}\left(\sum_{i=1}^{n_1} \hat{p}_2(\mathbf{x}_{1i})\right) = O(n_2) \quad (54)$$

Then,

$$\begin{aligned} & \frac{1}{[\text{var}(\sum_{i=1}^{n_1} \hat{p}_2(\mathbf{x}_{1i}))]^{3/2}} \sum_{i=1}^{n_1} E(|\hat{p}_2(\mathbf{x}_{1i})|^3) \\ &= O\left(\frac{1}{n_2^{3/2}}\right) \sum_{i=1}^{n_1} E\left(\left|\frac{\sum_{j=1}^{n_2} K\left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h}\right) y_{2j}}{\sum_{j=1}^{n_2} K\left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h}\right)}\right|^3\right) \\ &\leq O\left(\frac{1}{n_2^{3/2}}\right) \sum_{i=1}^{n_1} \left|\frac{\sum_{j=1}^{n_2} K\left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h}\right)}{\sum_{j=1}^{n_2} K\left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h}\right)}\right|^3 = O\left(\frac{1}{n_2^{1/2}}\right) \end{aligned} \quad (55)$$

Thus, it meets Lyapunov's condition and (49) holds.

To obtain the form of $\text{var}(\sum_{i=1}^{n_1} (g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))))$, let $h(\hat{m}_2(\mathbf{x}_1)) = \sum_{i=1}^{n_1} g(\hat{m}_2(\mathbf{x}_{1i}))$

where $\hat{m}_2(\mathbf{x}_1) = (\hat{m}_2(\mathbf{x}_{11}), \hat{m}_2(\mathbf{x}_{12}), \dots, \hat{m}_2(\mathbf{x}_{1n_1}))$. The first order Taylor expansion of $h(\hat{m}_2(\mathbf{x}_1))$

around $m_2(\mathbf{x}_1)$ is

$$h(\hat{m}_2(\mathbf{x}_1)) \approx h(m_2(\mathbf{x}_1)) + h'(m_2(\mathbf{x}_1))(\hat{m}_2(\mathbf{x}_1) - m_2(\mathbf{x}_1)) + o(h^2) \quad (56)$$

where $h'(m_2(\mathbf{x}_1)) = \left(\frac{e^{m_2(\mathbf{x}_{11})}}{(1+e^{m_2(\mathbf{x}_{11})})^2} \quad \dots \quad \frac{e^{m_2(\mathbf{x}_{1n_1})}}{(1+e^{m_2(\mathbf{x}_{1n_1})})^2} \right)$. The remainder of the expansion above is $o(h^2)$ due to Theorem 3 of Fan, Heckman, and Wand (1995, p.146). Using (56)

$$\text{var}(h(\hat{m}_2(\mathbf{x}_1))) \approx h'(m_2(\mathbf{x}_1))\Sigma h'(m_2(\mathbf{x}_1))^T \quad (57)$$

where Σ is the variance covariance matrix of $\hat{m}_2(\mathbf{x}_1)$ whose (i, k) -th element is:

$$\begin{aligned} \sigma_{ik} = & \sum_{j=1}^{n_2} \left(\sum_{l=0}^d A_{il}(x_{2lj} - x_{1li}) \right) \left(\sum_{l=0}^d A_{kl}(x_{2lj} - x_{1lk}) \right) \\ & \times K \left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1k}\|}{h} \right) p_{2j}(\mathbf{x}_{2j}) q_{2j}(\mathbf{x}_{2j}) \end{aligned} \quad (58)$$

$A_{i0}, A_{i1}, \dots, A_{id}$ are the first row elements of $(-l''_{\mathbf{x}_{1i}}(\mathbf{d}_i\boldsymbol{\beta}; \mathbf{y}_2))^{-1}$. To obtain this, first, we approximate $l'_{\mathbf{x}_{1i}}(\mathbf{d}_i\hat{\boldsymbol{\beta}}; \mathbf{y}_2)$ using a first order Taylor expansion around $\boldsymbol{\beta}$:

$$\begin{aligned} l'_{\mathbf{x}_{1i}}(\mathbf{d}_i\hat{\boldsymbol{\beta}}; \mathbf{y}_2) &= 0 \approx l'_{\mathbf{x}_{1i}}(\mathbf{d}_i\boldsymbol{\beta}; \mathbf{y}_2) + l''_{\mathbf{x}_{1i}}(\mathbf{d}_i\boldsymbol{\beta}; \mathbf{y}_2)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + o(h^2) \\ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= -(l''_{\mathbf{x}_{1i}}(\mathbf{d}_i\boldsymbol{\beta}; \mathbf{y}_2))^{-1} l'_{\mathbf{x}_{1i}}(\mathbf{d}_i\boldsymbol{\beta}; \mathbf{y}_2) + o(h^2) \end{aligned} \quad (59)$$

Again, the remainder term of the above expansion is $o(h^2)$ due to Theorem 3 of Fan, Heckman, and Wand (1995, p.146). The first element of the $(d+1) \times 1$ vector $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ is $\hat{m}_2(\mathbf{x}_{1i}) - m_2(\mathbf{x}_{1i})$ and can be expressed as:

$$\begin{aligned} (\hat{m}_2(\mathbf{x}_{1i}) - m_2(\mathbf{x}_{1i})) &= \sum_{l=0}^d A_{il} \left(\sum_{j=1}^{n_2} (x_{2lj} - x_{1li}) K \left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h} \right) \left(Y_{2j} - \frac{e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}}{1 + e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}} \right) \right) + o(h^2) \\ &\text{where } x_{20j} - x_{10i} = 1 \\ &= \sum_{j=1}^{n_2} \left(\sum_{l=0}^d A_{il}(x_{2lj} - x_{1li}) \right) K \left(\frac{\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|}{h} \right) \left(Y_{2j} - \frac{e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}}{1 + e^{\mathbf{d}_{ij}^T \boldsymbol{\beta}}} \right) + o(h^2) \end{aligned}$$

From this, one can obtain (58). Thus,

$$\begin{aligned}
\text{var} \left(\sum_{i=1}^{n_1} (g(\hat{m}_2(\mathbf{x}_{1i})) - g(m_2(\mathbf{x}_{1i}))) \right) &\approx h'(m_2(\mathbf{x}_1))^T \Sigma h'(m_2(\mathbf{x}_1)) \\
&= \frac{e^{m_2(\mathbf{x}_{11})}}{(1 + e^{m_2(\mathbf{x}_{11})})^2} \sum_{i=1}^{n_1} \sigma_{i1} \frac{e^{m_2(\mathbf{x}_{1i})}}{(1 + e^{m_2(\mathbf{x}_{1i})})^2} + \dots \\
&\quad + \frac{e^{m_2(\mathbf{x}_{1n_1})}}{(1 + e^{m_2(\mathbf{x}_{1n_1})})^2} \sum_{i=1}^{n_1} \sigma_{in_1} \frac{e^{m_2(\mathbf{x}_{1i})}}{(1 + e^{m_2(\mathbf{x}_{1i})})^2} \\
&= \sum_{k=1}^{n_1} \left(\frac{e^{m_2(\mathbf{x}_{1k})}}{(1 + e^{m_2(\mathbf{x}_{1k})})^2} \sum_{i=1}^{n_1} \sigma_{ik} \frac{e^{m_2(\mathbf{x}_{1i})}}{(1 + e^{m_2(\mathbf{x}_{1i})})^2} \right) \quad (60)
\end{aligned}$$

Next, we will show that $\sum_{i=1}^{n_1} (Y_{1i} - p_1(\mathbf{x}_{1i}))$ is asymptotically normally distributed because it satisfies the Lyapunov condition:

$$\frac{1}{(\text{var}(\sum_{i=1}^{n_1} (Y_{1i} - p_1(\mathbf{x}_{1i})))^{3/2}} \sum_{i=1}^{n_1} E(|Y_{1i} - p_1(\mathbf{x}_{1i})|^3) \leq O\left(\frac{1}{n_1^{3/2}}\right) O(n_1) = O\left(\frac{1}{n_1^{1/2}}\right)$$

Therefore,

$$\sum_{i=1}^{n_1} Y_{1i} - \sum_{i=1}^{n_1} p_1(\mathbf{x}_{1i}) \stackrel{D}{\sim} \text{Normal} \quad (61)$$

and $\text{var}(\sum_{i=1}^{n_1} Y_{1i} - \sum_{i=1}^{n_1} p_1(\mathbf{x}_{1i})) = \sum_{i=1}^{n_1} p_1(\mathbf{x}_{1i}) q_1(\mathbf{x}_{1i})$. Finally, combining (49) and (61) and because of their independence:

$$\begin{aligned}
\frac{1}{n_1} \left(\sum_{i=1}^{n_1} Y_{1i} - \sum_{i=1}^{n_1} p_1(\mathbf{x}_{1i}) - \sum_{i=1}^{n_1} \hat{p}_2(\mathbf{x}_{1i}) + \sum_{i=1}^{n_1} p_2(\mathbf{x}_{1i}) \right) \\
= \frac{1}{n_1} \left(\sum_{i=1}^{n_1} Y_{1i} - \sum_{i=1}^{n_1} \hat{p}_2(\mathbf{x}_{1i}) \right) \\
\quad - \frac{1}{n_1} \left(\sum_{i=1}^{n_1} p_1(\mathbf{x}_{1i}) - \sum_{i=1}^{n_1} p_2(\mathbf{x}_{1i}) \right) \\
= \bar{D}_{LOC} - \delta \stackrel{D}{\sim} \text{Normal}
\end{aligned}$$

Furthermore, using (60),

$$\text{var}(\bar{D}_{LOC}) \approx \frac{1}{n_1^2} \sum_{i=1}^{n_1} p_1(\mathbf{x}_{1i}) q_1(\mathbf{x}_{1i}) + \frac{1}{n_1^2} \sum_{k=1}^{n_1} \left(\frac{e^{m_2(\mathbf{x}_{1k})}}{(1 + e^{m_2(\mathbf{x}_{1k})})^2} \sum_{i=1}^{n_1} \sigma_{ik} \frac{e^{m_2(\mathbf{x}_{1i})}}{(1 + e^{m_2(\mathbf{x}_{1i})})^2} \right)$$

In order to obtain the form of the asymptotic bias, notice

$$E(\bar{D}_{LOC} - \delta) = E\left(\frac{\sum_{i=1}^{n_1}(Y_{1i} - p_1(\mathbf{x}_{1i}))}{n_1} - \frac{\sum_{i=1}^{n_1}(\hat{p}_2(\mathbf{x}_{1i}) - p_2(\mathbf{x}_{1i}))}{n_1}\right) \quad (62)$$

$$\begin{aligned} &= 0 - E\left(\frac{\sum_{i=1}^{n_1}(\hat{p}_2(\mathbf{x}_{1i}) - p_2(\mathbf{x}_{1i}))}{n_1}\right) \\ &= \sum_{i=1}^{n_1} \frac{1}{n_1} E(\hat{p}_2(\mathbf{x}_{1i}) - p_2(\mathbf{x}_{1i})) \end{aligned} \quad (63)$$

Using Theorem 4 of Fan, Heckman, and Wand (1995) and applying the strong law of large numbers

$$E(\bar{D}_{LOC} - \delta) = \frac{h^2}{2} \int \nu \text{tr}(H_{m_2}(\mathbf{x}_1)) p_2(\mathbf{x}_1) q_2(\mathbf{x}_1) f_1(\mathbf{x}_1) d\mathbf{x}_1 + o(h^2)$$

The order of the asymptotic variance is obtained as follows.

$$\begin{aligned} \text{var}(\bar{D}_{LOC} - \delta) &= \text{var}\left(\frac{\sum_{i=1}^{n_1}(Y_{1i} - p_1(\mathbf{x}_{1i}))}{n_1} - \frac{\sum_{i=1}^{n_1}(\hat{p}_2(\mathbf{x}_{1i}) - p_2(\mathbf{x}_{1i}))}{n_1}\right) \\ &= \text{var}\left(\frac{\sum_{i=1}^{n_1}(Y_{1i} - p_1(\mathbf{x}_{1i}))}{n_1}\right) - \text{var}\left(\frac{\sum_{i=1}^{n_1}(\hat{p}_2(\mathbf{x}_{1i}) - p_2(\mathbf{x}_{1i}))}{n_1}\right) \end{aligned}$$

$$\begin{aligned} \text{var}\left(\frac{\sum_{i=1}^{n_1}(Y_{1i} - p_1(\mathbf{x}_{1i}))}{n_1}\right) &= \sum_{i=1}^{n_1} \frac{1}{n_1^2} \text{var}(Y_{1i}) \\ &= \sum_{i=1}^{n_1} \frac{1}{n_1^2} p_1(\mathbf{x}_{1i}) q_1(\mathbf{x}_{1i}) \\ &= O(1/n_1) \end{aligned} \quad (64)$$

since $p_1(\mathbf{x}_{1i})q_1(\mathbf{x}_{1i})$ is bounded.

$$\begin{aligned}
\text{var} \left(\frac{\sum_{i=1}^{n_1} (\hat{p}_2(\mathbf{x}_{1i}) - p_2(\mathbf{x}_{1i}))}{n_1} \right) &= \frac{1}{n_1^2} \text{var} \left(\sum_{i=1}^{n_1} \hat{p}_2(\mathbf{x}_{1i}) \right) \tag{65} \\
&= \frac{1}{n_1^2} \left[\sum_{i=1}^{n_1} \text{var}(\hat{p}_2(\mathbf{x}_{1i})) + \sum_{i \neq l}^{n_1} \sum_{l=1}^{n_1} \text{cov}(\hat{p}_2(\mathbf{x}_{1i}), \hat{p}_2(\mathbf{x}_{1l})) \right] \\
&= \frac{1}{n_1^2} \left[\sum_{i=1}^{n_1} \text{var}(\hat{p}_2(\mathbf{x}_{1i})) + \sum_{i=1}^{n_1} \sum_{\substack{l=1 \\ i \neq l}}^{m_i} \text{cov}(\hat{p}_2(\mathbf{x}_{1i}), \hat{p}_2(\mathbf{x}_{1l})) \right] \\
&\leq \frac{1}{n_1^2} \sum_{i=1}^{n_1} \text{var}(\hat{p}_2(\mathbf{x}_{1i})) + \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{\substack{l=1 \\ i \neq l}}^{m_i} \sqrt{\text{var}(\hat{p}_2(\mathbf{x}_{1i})) \text{var}(\hat{p}_2(\mathbf{x}_{1l}))} \tag{66}
\end{aligned}$$

Theorem 4 of Fan, Heckman, and Wand (1995) states that the asymptotic variance of $\hat{p}_2(\mathbf{x}_{1i})$ is of order $O((n_2 h^d)^{-1})$. Using this in (66) and the fact that $m \leq O(n_1^{1-\alpha d})$, the asymptotic variance of (65) is $O\left(\frac{1}{n_1^{\alpha d} n_2 h^d}\right)$. Hence, the asymptotic variance of $\bar{D}_{LOC} - \delta$ is $O\left(\frac{1}{n_1^{\alpha d} n_2 h^d}\right)$.

This completes the proof of this theorem. □

B Proof of Corollary 2

Proof. The distribution of \bar{D}_{LOC} is normal due to Theorem 1.

As discussed in the proof of Theorem 1, when all the covariates are bounded and discrete, the potential values of the vector distance $\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|$ will be discrete and finite as well. As $h \rightarrow 0$, h becomes smaller than the minimum distance between any two discrete values of $\|\mathbf{x}_{2j} - \mathbf{x}_{1i}\|$. Thus, only those majority members \mathbf{x}_{2j} having the same covariate values as the design point \mathbf{x}_{1i} will receive a non-zero weight in finding $\hat{m}_2(\mathbf{x}_{1i})$; all these majority members receive the same weight,

and we denote it by $K\left(\frac{\|\mathbf{x}_{2j}-\mathbf{x}_{1i}\|}{h}\right) = w$. Then, using (51), \bar{D}_{LOC} becomes

$$\begin{aligned}
\bar{D}_{LOC} &= \frac{\sum_{i=1}^{n_1} y_{1i}}{n_1} - \frac{\sum_{i=1}^{n_1} \hat{p}_2(\mathbf{x}_{1i})}{n_1} \\
&= \frac{\sum_{i=1}^{n_1} y_{1i}}{n_1} - \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{\sum_{j=1}^{n_2} K\left(\frac{\|\mathbf{x}_{2j}-\mathbf{x}_{1i}\|}{h}\right) y_{2j}}{\sum_{j=1}^{n_2} K\left(\frac{\|\mathbf{x}_{2j}-\mathbf{x}_{1i}\|}{h}\right)} \\
&= \frac{\sum_{k=1}^r \sum_{i=1}^{n_{1k}} y_{1ki}}{n_1} - \frac{\sum_{k=1}^r \sum_{i=1}^{n_{1k}} \frac{\sum_{j=1}^{n_{2k}} w y_{2kj}}{n_{2k} w}}{n_1} \\
&= \frac{\sum_{k=1}^r \sum_{i=1}^{n_{1k}} y_{1ki}}{n_1} - \frac{\sum_{k=1}^r \frac{n_{1k}}{n_{2k}} \sum_{j=1}^{n_{2k}} y_{2kj}}{n_1} \\
&= \frac{1}{n_1} \sum_{k=1}^r \left(\sum_{i=1}^{n_{1k}} y_{1ki} - \frac{n_{1k}}{n_{2k}} \sum_{j=1}^{n_{2k}} y_{2kj} \right) \tag{67}
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{n_1} \sum_{k=1}^r n_{1k} \left(\frac{\sum_{i=1}^{n_{1k}} y_{1ki}}{n_{1k}} - \frac{\sum_{j=1}^{n_{2k}} y_{2kj}}{n_{2k}} \right) \\
&= \frac{1}{n_1} \sum_{k=1}^r n_{1k} (\bar{y}_{1k} - \bar{y}_{2k}) \tag{68}
\end{aligned}$$

Using the form of \bar{D}_{LOC} in (68),

$$E\left(\frac{1}{n_1} \sum_{k=1}^r n_{1k} (\bar{y}_{1k} - \bar{y}_{2k})\right) = \frac{1}{n_1} \sum_{k=1}^r n_{1k} (p_1(\mathbf{x}_{1k}) - p_2(\mathbf{x}_{1k})) \tag{69}$$

where \mathbf{x}_{1k} indicates the minority covariate vector with the k -th covariate value combination. Now,

$$\begin{aligned}
\delta &= \frac{1}{n_1} \sum_{i=1}^{n_1} (p_1(\mathbf{x}_{1i}) - p_2(\mathbf{x}_{1i})) \\
&= \frac{1}{n_1} \sum_{k=1}^r \sum_{i=1}^{n_{1k}} (p_1(\mathbf{x}_{1ki}) - p_2(\mathbf{x}_{1ki})) \\
&= \frac{1}{n_1} \sum_{k=1}^r n_{1k} (p_1(\mathbf{x}_{1k}) - p_2(\mathbf{x}_{1k})) \tag{70}
\end{aligned}$$

Since $E(\bar{D}_{LOC})$ in (69) and δ in (70) are the same, $E(\bar{D}_{LOC} - \delta) = 0$.

Using the form of \bar{D}_{LOC} in (67),

$$\begin{aligned}
\text{var}(\bar{D}_{LOC}) &= \text{var}\left(\frac{1}{n_1} \sum_{k=1}^r \left(\sum_{i=1}^{n_{1k}} y_{1ki} - \frac{n_{1k}}{n_{2k}} \sum_{j=1}^{n_{2k}} y_{2kj}\right)\right) \\
&= \frac{1}{n_1^2} \sum_{k=1}^r \left(\sum_{i=1}^{n_{1k}} \text{var}(y_{1ki}) + \sum_{j=1}^{n_{2k}} \frac{n_{1k}^2}{n_{2k}^2} \text{var}(y_{2kj})\right) \\
&= \frac{1}{n_1^2} \sum_{k=1}^r \left(\sum_{i=1}^{n_{1k}} p_1(\mathbf{x}_{1ki}) q_1(\mathbf{x}_{1ki}) + \sum_{j=1}^{n_{2k}} \frac{n_{1k}^2}{n_{2k}^2} p_2(\mathbf{x}_{2kj}) q_2(\mathbf{x}_{2kj})\right)
\end{aligned}$$

This completes the proof of this corollary.

□

References

- [1] BELSON, W. A. (1956). A technique for studying the effects of a television broadcast. *Applied Statistics*, **5**, 195-202.
- [2] BHATTACHARYA, P. K. and GASTWIRTH, J. L. (1999). Estimation of the odds-ratio in an observational study using bandwidth-matching. *Nonparametric Statistics*, **11**, 1-12.
- [3] BRILLINGER, D. R. (1977). Discussion of Stone (1977). *The Annals of Statistics*, **5**, 622-623.
- [4] CHEN, L. (2005). *Introduction to Stein's Method*. River Edge, NJ: World Scientific Publishing Company, Inc.
- [5] CLEVELAND, W. S. and DEVLIN, S. J. (1988). Locally-weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.*, **83**, 597-610.
- [6] COCHRAN, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, **10**, 417-451.
- [7] COCHRAN, W. G. and RUBIN, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya, Series A*, **35**, 417-446.
- [8] FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.
- [9] FAN, J., HECKMAN, N.E., and WAND, M.P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.*, **90**, 141-150.
- [10] GASTWIRTH, J. L. (1989). A clarification of some statistical issues in Watson vs. Fort Worth Bank and Trust. *Jurimetrics Journal*, **29**, 267-284.

- [11] GASTWIRTH, J.L. and GREENHOUSE, S.W. (1995). Biostatistical Concepts and Methods in the Legal Setting. *Statistics in Medicine*, **14**, 1641-1653.
- [12] GEL, Y. R. and GASTWIRTH, J. L. (2008). A robust modification of the JarqueBera test of normality. *Economics Letters*, **99**, 30-32.
- [13] GRAY, M. W. (1993). Can statistics tell us what we do not want to hear? The case of complex salary structures. *Statistical Science*, **8**, 144-179.
- [14] KARR, A. F. (1993). *Probability*. New York: Springer-Verlag.
- [15] LOADER, C. (1999). *Local Regression and Likelihood*. New York: Springer-Verlag.
- [16] NAYAK, T. K. and GASTWIRTH, J. L. (1997). The Peters-Belson approach to measures of economic and legal discrimination. *Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz*, ed. by Johnson, N.L. and Balakrishnan, N., 587-601.
- [17] PETERS, C. C. (1941). A method of matching groups for experiments with no loss of populations. *Journal of Educational Research*, **34** 606-612.
- [18] TAKEZAWA, K. (2006). *Introduction to Nonparametric Regression*. Hoboken, New Jersey: John Wiley and Sons.
- [19] TIBSHIRANI, R. J. (1984). Local Likelihood Estimation. *Ph.D. thesis*, Department of Statistics, Stanford University.
- [20] TIBSHIRANI, R. J. and HASTIE T. (1987). Local Likelihood Estimation. *J. Amer. Statist. Assoc.*, **82**, 559-567

[21] WADE, W. R. (2004). *An Introduction to Analysis*. Upper Saddle River, NJ: Pearson Prentice Hall.