The Use of Peters-Belson Regression in Legal Cases

Efstathia Bura, Joseph L. Gastwirth

Department of Statistics

The George Washington University

Washington, DC 20052

Hiroyuki Hikawa

Ernst & Young LLP*

Washington, DC 20005

February 7, 2011

Abstract

In equal employment cases it is important to compare the salary, hiring or promotion status of minority employees or applicants to that of similarly qualified majority members. Standard regression methods include a binary variable indicating minority status along with major job-related variables to assess whether belonging to a minority has a negative effect. Implicitly, this method assumes that any discriminatory policy has the same effect on every minority applicant or employee. Peters-Belson regression is a form of statistical matching, akin in spirit to Bhattacharya's band-width matching. In the context of studying possible sex discrimination in pay, one fits a regression to the salaries of employees of the males and then predicts the salary of females from the male equation. This prediction estimates the salary a female employee would receive assuming her job-related characteristics were rewarded in a similar way as males, i.e. it provides a "statistical match" for the actual salary of each female. The difference, if any, between the actual and predicted salary of each female employee estimates her under-compensation. Both parametric and non-parametric approaches to Peters-Belson regression are described and their use illustrated on salary data based on the *EEOC v Shelby County* equal pay case. Implications for the analysis of the data considered by courts to decide whether a class action is appropriate, as in *Dukes v. Wal-mart*, are discussed.

*The views expressed are solely of the authors and do not necessarily reflect the views of Ernst & Young LLP.

1 Introduction

The Peters-Belson (PB) regression method was first introduced by Peters (1941) and Belson (1956) for conducting treatment-control comparisons that accounted for relevant covariates by creating statistical matches for the treatment group observations. Blinder (1973) and Oaxaca (1973) used this idea to decompose the difference between the means of the two groups into components. The PB regression method fits a regression model to the control group and estimates the response a member of the protected group would have had they been a control from the regression fitted to the control group. The difference between the value of each treatment group member's response and its prediction obtained from the control group equation estimates the effect of the treatment.

In employment discrimination cases (e.g., equal pay cases), where members of the protected group (e.g., racial minority, females) are compared to those of the unprotected group (e.g., racial majority, males), the PB regression method as well as ordinary regression with an indicator representing group status have been accepted by courts (Gray, 1993). The PB method offers some advantages compared to standard regression with a dummy or indicator variable. First, the method is intuitive and comparatively easy to understand for a general audience (e.g., judges, juries, etc.). For example, in the context of sex discrimination cases, PB regression estimates the salary equation for the majority group incorporating related covariates and then takes the difference between the minority's actual salary and the estimated salary that the minority employee would have received if s/he were paid according to the equation for majority employees. Moreover, the estimated pay differential obtained from the PB approach is individualized for each member of the protected group. In contrast, ordinary least squares linear regression with an indicator variable estimates a common overall effect of being minority, after adjusting for the relevant covariates. This approach assumes that any differential is the same across the entire range of covariate values. Another advantage of the PB method is that the females whose salaries are higher than predicted and were not discriminated against are readily identifiable.

This paper reviews the Peters-Belson regression approach in the context of legal cases and also describes a recent extension of the method, similar in spirit to Bhattacharya (1989), by the authors. In Section 2 we review parametric PB regression, based on parametric ordinary linear regression. Section 3 introduces a recent nonparametric version that increases the applicability of the PB approach. In Section 4 we apply all the methods we discuss to data from a sex discrimination case, and Section 5 shows how the PB approach is superior to the regression approaches of both parties in the *Dukes v. Wal-mart* sex discrimination case. We present concluding remarks in Section 6.

2 Parametric Peters-Belson Regression

2.1 Review of the method

Suppose the salaries, denoted by Y, for minority and majority employees are given by

$$\text{Minority:} \quad Y_{1i} = \mathbf{X}_{1i}^T \boldsymbol{\beta}_1 + \epsilon_{1i} \qquad \text{for } i = 1, \ldots, n_1 \tag{1}$$

$$\text{Majority:} \quad Y_{2j} = \mathbf{X}_{2j}^T \boldsymbol{\beta}_2 + \epsilon_{2j} \qquad \text{for } j = 1, \ldots, n_2 \tag{2}$$

$\mathbf{X}$ denotes the covariate vector and $\boldsymbol{\beta}$ the corresponding coefficient vector. The errors, ϵ, in each equation are assumed to be normally distributed with mean zero and variance σ_1^2 and σ_2^2, respectively. When both groups are paid according to the same system, the regression coefficients in the two models are equal, that is $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$.

In parametric PB, a linear regression model is fitted to the data for the majority employees. Then each minority member's salary is predicted by $\mathbf{x}_{1i}^T \hat{\boldsymbol{\beta}}_2$, where $\mathbf{x}_{1i}$ is the covariate vector for the ith minority member and $\hat{\boldsymbol{\beta}}_2$ is the least squares estimate of $\boldsymbol{\beta}_2$. The difference, $D_i = y_{1i} - \mathbf{x}_{1i}^T \hat{\boldsymbol{\beta}}_2$, between the actual and predicted salaries is the estimate of the pay differential of the i-th minority employee relative to a similarly qualified majority employee. Thus, D_i is the parametric PB estimate of $\delta_i = \mathbf{x}_{1i}^T \boldsymbol{\beta}_1 - \mathbf{x}_{1i}^T \boldsymbol{\beta}_2$, the difference in pay between the i-th minority employee and a similar majority member. When the model is correct, D_i is unbiased for δ_i and the corresponding unbiased estimator for the average disparity over all minority employees

$$\delta \quad = \quad \frac{1}{n_1}\sum_{i=1}^{n_1}[\mathbf{x}_{1i}^T\boldsymbol{\beta}_1 - \mathbf{x}_{1i}^T\boldsymbol{\beta}_2] = \bar{\mathbf{x}}_1^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \tag{3}$$

equals

$$\bar{D} = \frac{\sum_{i=1}^{n_1} D_i}{n_1} = \frac{\sum_{i=1}^{n_1}[y_{1i} - \mathbf{x}_{1i}^T\hat{\boldsymbol{\beta}}_2]}{n_1} \tag{4}$$

where $\bar{\mathbf{x}}_1$ is the mean vector of the minority covariate values. The variance of $\bar{D}$ is

$$var(\bar{D}) = \frac{\sigma_1^2}{n_1} + \sigma_2^2\bar{\mathbf{x}}_1^T(\mathbf{x}_2^T\mathbf{x}_2)^{-1}\bar{\mathbf{x}}_1 \tag{5}$$

where $\mathbf{x}_2 = (\mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2})^T$ is the usual design matrix of the majority group.

When $\sigma_1^2 = \sigma_2^2 = \sigma^2$, the test statistic

$$t = \frac{\bar{D}}{\sqrt{\hat{\sigma}^2(1/n_1 + \bar{\mathbf{x}}_1^T(\mathbf{x}_2^T\mathbf{x}_2)^{-1}\bar{\mathbf{x}}_1)}} \tag{6}$$

can be used to test $\delta = 0$. Gastwirth (1989) suggested that majority observations be used to estimate the common variance σ^2 because under the hypothesis of no discrimination both majority and minority should be paid under the same system and hence the variances are the same as well. If we use $\hat{\sigma}^2 = \hat{\sigma}_2^2$, under the null hypothesis of $\delta = 0$, the test statistic in (6) is t-distributed with $n_2 - p_2$ degrees of freedom, where p_2 is the number of parameters (coefficients) in the majority model.

When the error variances are different, one can approximate the distribution of the test statistic using Welch's approximation (Welch, 1949; Scheffe, 1970; Nayak and Gastwirth, 1997). Under this approximation, the test statistic distribution under the null is approximated by a t distribution with degrees of freedom:

$$df = \frac{(\frac{1}{n_1}\sigma_1^2 + \bar{\mathbf{x}}_1^T(\mathbf{x}_2^T\mathbf{x}_2)^{-1}\bar{\mathbf{x}}_1\sigma_2^2)^2}{\sigma_1^4/(n_1^2(n_1-p_1)) + \sigma_2^4/(n_2-p_2)\bar{\mathbf{x}}_1^T(\mathbf{x}_2^T\mathbf{x}_2)^{-1}\bar{\mathbf{x}}_1\bar{\mathbf{x}}_1^T(\mathbf{x}_2^T\mathbf{x}_2)^{-1}\bar{\mathbf{x}}_1} \tag{7}$$

Since σ_1^2 and σ_2^2 are unknown, the degrees of freedom are estimated by substituting $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ in (7).

Gastwirth and Greenhouse (1995) applied the PB method to salary data as well as to logistic regression for binary responses in order to analyze the data arising from a case involving promotion decisions (*Capaci v. Katz and Besthoff* [1]). Furthermore, Nayak and Gastwirth (1997) extended the method to generalized linear models. Hikawa et al. (2010a) introduced nonparametric PB in regressions with a binary response as an alternative to logistic regression.

3 Local Linear Peters-Belson Regression

There are two problems an analyst of pay discrimination data often encounters (Hikawa et al., 2010b). The first is the difficulty in estimating the salary equation when it does not appear to follow any usual parametric forms (e.g., linear, quadratic). The second problem pertains to determining who the relevant male/majority employees are to be compared against female/minority employees of interest. Including too many irrelevant majority employees in the comparisons (e.g., male employees who are too senior compared to the target female employees) may introduce serious bias in the estimated disparity (Greiner, 2008).

To address these problems, Hikawa et al. (2010b) introduced local linear regression in PB. Local linear regression fits a linear regression in the *neighborhood* of the covariate values of each minority member. The method is well suited for equal pay cases since the estimation/prediction of the salary of a minority employee is based on majority employees whose qualifications are closest to those of the minority employee and thus should receive the greatest weight. Furthermore, the similarity of this method to matched-pairs is expected to make the results more understandable to judges and juries. Local linear regression is similar in spirit to bandwidth matching introduced by Bhattacharya (1989). However, the weight given to each majority observation decreases with the distance of the covariate values from the target minority member.

[1] *Capaci v. Katz and Besthoff* 711 F. 2d 647, 5th Cir., 1983

3.1 Description of Local Linear PB

Suppose we have d covariates and the data consist of n_1 minority observations, $(\mathbf{x}_{11}, y_{11}), \ldots, (\mathbf{x}_{1n_1}, y_{1n_1})$, and n_2 majority observations, $(\mathbf{x}_{21}, y_{21}), \ldots, (\mathbf{x}_{2n_2}, y_{2n_2})$, where $\mathbf{x}$ is a vector of d fixed covariate values. The response in the two groups can always be modeled as follows

$$\text{Minority:} \quad Y_{1i} = m_1(\mathbf{x}_{1i}) + \epsilon_{1i}, \qquad \epsilon_{1i} \sim N(0,\, \sigma_1^2) \tag{8}$$

$$\text{Majority:} \quad Y_{2j} = m_2(\mathbf{x}_{2j}) + \epsilon_{2j}, \qquad \epsilon_{2j} \sim N(0,\, \sigma_2^2) \tag{9}$$

with $\mathrm{E}(\epsilon_1) = \mathrm{E}(\epsilon_2) = 0$. The only assumption we make on the unknown functions modeling the mean response in the two groups, $m_1(\mathbf{x})$ and $m_2(\mathbf{x})$, is that they are twice differentiable.

As in the parametric PB definition of disparity, the pay disparity for the i-th minority member is

$$\delta_i = m_1(\mathbf{x}_{1i}) - m_2(\mathbf{x}_{1i}), \tag{10}$$

and the average disparity of all minority members is

$$\delta = \frac{\sum_{i=1}^{n_1} \left(m_1(\mathbf{x}_{1i}) - m_2(\mathbf{x}_{1i})\right)}{n_1}. \tag{11}$$

Let

$$\mathbf{Z}_i = \begin{pmatrix} 1 & (x_{211} - x_{11i}) & \cdots & (x_{2d1} - x_{1di}) \\ 1 & (x_{212} - x_{11i}) & \cdots & (x_{2d2} - x_{1di}) \\ \vdots & \vdots & & \vdots \\ 1 & (x_{21n_2} - x_{11i}) & \cdots & (x_{2dn_2} - x_{1di}) \end{pmatrix} \tag{12}$$

and $\mathbf{W}_i = diag\big(W(||\mathbf{x}_{2j} - \mathbf{x}_{1i}||/h)\big)$, where W is a kernel weight function and $||\cdot||$ is a norm. Denoting the elements of the first row of $(\mathbf{Z}_i^T\mathbf{W}_i\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{W}_i$ by $S_{i1}, S_{i2}, \ldots, S_{in_2}$, the fitted value for the design point $\mathbf{x}_{1i}$ is given by

$$\hat{m}_2(\mathbf{x}_{1i}) = \sum_{j=1}^{n_2} S_{ij} y_{2j} \tag{13}$$

The estimated disparity of the i-th minority member is

$$D_i = y_{1i} - \hat{m}_2(\mathbf{x}_{1i}) \tag{14}$$

The estimators of the overall average disparity and its variance are

$$\bar{D}_{LOC} = \frac{\sum_{i=1}^{n_1} D_i}{n_1} = \frac{\sum_{i=1}^{n_1}(y_{1i} - \sum_{j=1}^{n_2} S_{ij} y_{2j})}{n_1} \tag{15}$$

$$\text{var}(\bar{D}_{LOC}) = var(\bar{D}) = \frac{1}{n_1^2}\Big(n_1\sigma_1^2 + \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} S_{ij}^2\sigma_2^2 + \sum_{i\neq l}^{n_1}\sum^{n_1}\sum_{j=1}^{n_2} S_{ij}S_{lj}\sigma_2^2\Big) \tag{16}$$

A nearest neighbor bandwidth that fixes the fraction of data that contribute to the estimation (Cleveland and Devlin, 1988; Loader, 1999) is used in fitting the local linear regression model. The Epanechnikov kernel, defined by

$$W(u) = \frac{3}{4}(1 - u^2) I_{(|u|\leq 1)}, \tag{17}$$

where $u = ||\mathbf{x}_{2j} - \mathbf{x}_{1i}||/h$ and h is the bandwidth, is used as the weight function. It is the optimal kernel in the sense that roughly both the asymptotic mean squared error and the mean integrated squared error are minimized (see Thm. 3.4 in Fan and Gijbels, 1996, p.75). Loader (1999, p.20) suggests scaling the vector norm. We use

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

where $s_k = \sqrt{(\sum_{j=1}^{n_2}(x_{2kj} - x_{1ki} - \sum_{j=1}^{n_2}(x_{2kj} - x_{1ki})/n_2)^2)/(n_2-1)}$. Since x_{1ki} is fixed, s_k is simply the sample standard deviation of x_{2k} about x_{1ki} for the majority members for the k-th covariate.

Since σ_1^2 and σ_2^2 are usually unknown, the estimated variance of $\bar{D}_{LOC}$ in (16) can be obtained by using estimates $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ which can be obtained from the residuals of the separate local linear regression models within each group. Details of the estimation approach are given in Hikawa et al. (2010b). The estimated variance of $\bar{D}_{LOC}$ is given by

$$\widehat{var}(\bar{D}_{LOC}) = \frac{1}{n_1^2}\Big(n_1\hat{\sigma}_1^2 + \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} S_{ij}^2\hat{\sigma}_2^2 + \sum_{i\neq l}^{n_1}\sum_{j=1}^{n_2} S_{ij}S_{lj}\hat{\sigma}_2^2\Big) \tag{18}$$

Cleveland and Devlin (1988) showed that, under the assumption of normal errors and negligible bias of $\hat{m}_j(\mathbf{x})^2$, the distribution of $(\eta_j^2\hat{\sigma}_j^2)/(\kappa_j\sigma_j^2)$ can be approximated by a χ^2 distribution with degrees of freedom η_j^2/κ_j, where

$$\eta_j = tr(\mathbf{I}-\mathbf{S}_j)(\mathbf{I}-\mathbf{S}_j)^T \tag{19}$$

$$\kappa_j = tr[(\mathbf{I}-\mathbf{S}_j)(\mathbf{I}-\mathbf{S}_j)^T]^2 \tag{20}$$

for $j = 1, 2$. Let $\mathbf{S}_j$ be the $n_j \times n_j$ matrix whose $(i,k)^{th}$ element is S_{ik} obtained from fitting a separate smooth curve for the minority and majority groups (i.e., separately estimating $m_1(x_i)$ for $i = 1, \ldots n_1$ and $m_2(x_i)$ for $i = 1, \ldots n_2$). The test statistic for testing for lack of disparity $(H_0 : \delta = 0)$ is defined to be

$$t = \frac{\bar{D}_{LOC}}{\sqrt{v\hat{a}r(\bar{D}_{LOC})}} \tag{21}$$

When the two variances are equal (i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$), we estimate the common variance by the majority variance estimate $\hat{\sigma}^2 = \hat{\sigma}_2^2$, and the estimated variance of $\bar{D}_{LOC}$ in (18) becomes

$$\widehat{var}(\bar{D}_{LOC}) = \frac{\hat{\sigma}^2}{n_1^2}\Big(n_1 + \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} S_{ij}^2 + \sum_{i\neq l}^{n_1}\sum_{j=1}^{n_2} S_{ij}S_{lj}\Big)$$

Therefore, under the assumption of equal variances, the test statistic for group disparity is t distributed with η^2/κ degrees of freedom, where $\eta = \eta_2$ and $\kappa = \kappa_2$ as defined in (19) and (20). The

[2] Cleveland and Devlin (1988) state that $\hat{m}(\mathbf{x})$ in (13) is unbiased only when the underlying mean function $m(\cdot)$ is linear. In general, the bias goes to zero asymptotically at the rate of h^2 as $h \to 0$ and $nh \to \infty$ (Fan and Gijbels, 1996, p.62).

distribution of the test statistic when the variances are unequal is given in the Appendix .

4 Application: Re-analysis of Data from *EEOC v. Shelby County Government*

In this section, we illustrate the methods on a data set from the pay discrimination case, *EEOC v. Shelby County Government*. In 1983, a female clerical employee of Shelby County Criminal Court brought a charge of salary discrimination. The data, presented in Table 1, consist of salaries per pay period as the response and seniority measured in months since hired as the covariate with 16 male and 15 female clerical employees as of February 1, 1984. Fitting a simple linear regression to each gender separately yields:

$$\hat{Y}_f = 885.1 + 4.3X_f \quad (22)$$

$$\hat{Y}_m = 913.9 + 6.0X_m \quad (23)$$

where the subscripts m and f stand for male and female. The estimated variances of the error terms are $\hat{\sigma}_m^2 = 7494.5$ and $\hat{\sigma}_f^2 = 8757.1$, and the sample means and variances of seniority are: $\bar{X}_m = 81$, $S_m^2 = 160.3.3$, $\bar{X}_f = 69.3$, and $S_f^2 = 2195.5$. Table 2 summarizes the ANOVA tables for both the male and female models.

Since the sample size of 16 males is too small to fit a local linear regression model, we augmented the data set following the method used by Bhattacharya (1989) and Bhattacharya and Gastwirth (1999), where they analyze data from *Berger v. Iron Workers Local 201*[3]. Fitting a Gamma distribution to the seniority data yielded $Gamma(4, 20)$ for males and $Gamma(2, 33)$ for females. Then, we generated additional salary data for 84 males and 85 females according to the fitted models in (22). The error variances were set equal to $\hat{\sigma}_m^2 = 7494.5$ and $\hat{\sigma}_f^2 = 8757.1$ in order to match the estimated variances of the error terms from the fitted regression models.

The data were simulated based on unequal variances for males and females. In consequence,

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

Table 1: Shelby County pay discrimination case data

Male		Female	
Salary (Y)	Seniority (X)	Salary (Y)	Seniority (X)
1666	125	1474	128
1666	124	1474	96
1666	124	1403	121
1666	120	1403	117
1548	117	1403	114
1548	116	1336	89
1548	113	1336	97
1548	95	1157	105
1548	73	1000	64
1306	69	1000	51
1157	51	1000	13
1157	47	1000	17
1157	46	929	13
1157	18	929	8
1000	41	929	6
1000	17		

Table 2: ANOVA tables for male and female models from the Shelby County data

	Male				Female			
Sources	DF	SS	MS	F	DF	SS	MS	F
Seniority	1	852412	852412	113.74*	1	575666	575666	65.74*
Error	14	104924	7495		13	113842	8757	
Total	15	957336			14	689508		
$R^2 = .89$					$R^2 = .83$			

* P-value was smaller than .0001.

we compute the variance of $\bar{D}$ based on the assumption of unequal variances and approximate the degrees of freedom of the test statistics using the Welch's approximation approach for Local Linear PB by (26) and parametric PB by (7). Table 3 summarizes the results from applying the four methods: (1) Local Linear PB, (2) parametric PB, (3) ordinary regression with and without interaction between seniority and gender indicator variables, and (4) propensity score (Rosenbaum and Rubin, 1983; Zanutto, 2006). This last method creates subgroups in which males and females have similar covariate values. This is accomplished by creating a logistic regression predicting gender from the covariates. The quintiles of the females' estimated probabilities or propensity scores are used to form the comparable subgroups. Within each subgroup, the average salaries of males and females are compared and the subgroup differences are aggregated to form the estimate of the overall disparity.

Table 4 presents the predicted salaries of female employees that were in the original data before augmentation.

Table 3: Analysis of the augmented Shelby County Pay Discrimination data.

	Local PB	Parametric PB	Ordinary Reg. w\o Int.	Ordinary Reg. w\ Int.	Propensity Score
$\bar{D}$	-128.5	-126.6	-140.3	-126.6	-123.6
$\sqrt{\widehat{Var}(\bar{D})}$	15.4	13.8	14.8	13.7	23.1
$\bar{D}/\sqrt{\widehat{Var}(\bar{D})}$	-8.3	-9.2	-9.5	-9.2	-5.4
P-value	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$

Table 4: Actual and predicted salary of female employees.

Actual	Local PB	Parametric PB	Ordinary Reg. w\o Int.	Ordinary Reg. w\ Int.
1474	1681	1686	1629	1686
1474	1496	1493	1475	1493
1403	1639	1643	1595	1643
1403	1616	1619	1576	1619
1403	1598	1601	1561	1601
1336	1460	1450	1441	1450
1336	1501	1499	1479	1499
1157	1544	1547	1518	1547
1000	1297	1300	1320	1300
1000	1205	1221	1258	1221
1000	1026	992	1075	992
1000	1046	1016	1094	1016
929	1026	992	1075	992
929	1003	962	1051	962
929	993	950	1041	950

The negative values of $\bar{D}$ indicate that female employees were underpaid on average compared to their similarly qualified male counterparts. Local Linear PB (-128.5), parametric PB (-126.6), and ordinary first order regression with interaction (-126.6) give similar estimated average pay differential. From Table 4 we can see that OLS without interaction yields a higher average estimate than the others as it predicts higher salaries for the less senior female employees and underpredicts the salaries more senior females should receive. The propensity score method yielded a slightly smaller estimate (-123.6) and its standard error (23.1) is much larger than the other methods.

Since the estimated amount of pay differential from all the methods is quite large, the p-values of all test statistics are very small. Hence, all methods would reject the null hypothesis of no pay differential and confirm the court's conclusion that the female employees were discriminated in their pay with an average differential about \$126-128 per pay period.

5 The Suitability of Peters-Belson Regression for Studying the Class Certification Issue

In order for the claims of many individual plaintiffs, who are represented by a relatively small number of named plaintiffs, to be combined into a single class action, four criteria need to be met (see Fed. R. Civ. P. 23). Statistical studies are often submitted to satisfy the criterion (23(a)) that there are questions of law or fact *common* to the members of the class. In *Caridad v. Metro-North Commuter Railroad* plaintiffs used a multiple regression controlling for job related factors, to demonstrate that African American employees received fewer promotions but were disciplined more often than similar whites and that this occurred in many locations of the defendant's operation. Perhaps the best known controversy over the evidence plaintiffs need to submit to support a class action is a lower court's certification of a class of female employees in *Dukes et al. v. Wal-mart*, which will be considered by the U.S. Supreme Court in 2011.

A statistical issue in Wal-mart, noted in the 6-5 split decision in the 9th Circuit upholding the District court's certification of the class, concerned the appropriate degree of aggregation of the employee data. The plaintiffs analyzed the data on all employees in each of the 41 regions. Each region contains about 80 to 85 stores and the number of employees in a store ranges from 200 to 500. The defendant argued that the data should be analyzed store by store and before the store-wide data are combined into a single sample the plaintiffs' expert should have applied a Chow (1960) or Rao (1952)[4] test of the equality of the regression equations before pooling the data. The defendants' expert actually analyzed the data at the sub-store level, i.e. similar departments within each store were combined. This led to about three regression analyses being conducted per store. While the Wal-mart case is used as a familiar back-drop, the salary data are not available for independent reanalysis. Thus, only general statistical implications are discussed here.

As noted in Gastwirth, Miao and Zheng (2003) in the context of pooling 2×2 tables, requiring that the odds ratios in all the tables be equal is too strong a requirement. The Cochran-

[4]While Chow was the first economist to consider testing whether two regressions were identical, the problem had also been considered in the statistical literature.

Mantel-Haenszel test is consistent for detecting an overall shortfall across the strata provided there is no interaction, i.e. individual odds ratios are consistent with all being less (greater) than one. Thus, that summary test is applicable in many situations when the odds ratios in each table are not equal. The Gail-Simon (1985) test for interaction can be used in place of the test of equality of the odds ratios. Similarly, the null hypothesis tested in the Chow-Rao test is whether the coefficients for every predictor, e.g. seniority, gender, are identical in all the strata (stores in Wal-mart). Again, this is too strong a requirement. For example, if the data for each store fit a model relating salary to gender and seniority, the coefficient reflecting the added salary increment of each additional year of employment, could vary among the stores but the coefficient for being female could be negative in every store.

Because the PB approach provides an estimate of the underpayment for each female employee compared to a statistically matched male, it is appropriate for examining the issue of whether female employees are generally underpaid relative to similarly qualified males. Whether one fits the regression to males in each store or to stores in an entire region or to stores in a sub-region, e.g. those located in the same or an economically similar labor market, one has an estimate D_i of the difference between each female employee's salary and her "statistical match." Under the hypothesis of equal treatment approximately half of the D_i will be positive and half negative and the average of them will be near zero. A pattern of negative D_i suggests that the issue of underpayment of female employees, relative to comparable males, is a common situation. Of course, in Wal-mart and similar cases, the regression equation should include information on the major predictors that were considered in the salary setting process, e.g. seniority, job type, special education or experience, etc.

In the Wal-mart case both parties could present their estimates of the D_i for their preferred models, i.e. at the level of aggregation they believe is appropriate. When looking at highly disaggregated data, such as examining each store by itself, the power of a test of significance of the average difference is likely to be low. Thus, a finding of non-significance in each of many small strata or sub-groups of a large data set (stores in Wal-mart) may have little meaning unless a power analysis shows that one could detect a legally meaningful difference (see Bura and Gastwirth, 2009 for a discussion of the power of regression analyses in a securities law case and Gastwirth and Pan, 2011 give an example where a sample of over 900 venire members was insufficient to reliably detect a legally meaningful under-representation of minority in a jury pool). By examining the pattern of negative D_i obtained from the sub-store analyses advocated by Wal-mart or 41 regional regressions based on more aggregated data, advocated by the plaintiffs, one can assess whether there is a general pattern of underpayment of women. Thus, it is possible that when the estimated salary obtained from a PB regression analysis falls short for each female employee it could support the plaintiffs' position. On the other hand, if a pattern of predominantly negative differences is not seen or only occurs in a small subset of stores, this would be inconsistent with a nation-wide practice of underpaying females.

The issue of homogeneity of the regression models in the individual stratum (the plaintiffs' regions or the defendant's sub-store units) becomes less important when the PB method is used because one is working with the individualized estimates, D_i of underpayment of each female employee obtained from a regression based on comparable males. If the differences observed from the plaintiffs' analysis of the much larger regional data sets are "explained" or substantially decreased, when the defendant incorporates local factors, the latter would explain the pattern of differences observed in the more highly aggregated analysis.

So far in the Wal-mart proceedings, it appears that neither side has submitted a PB regression. Both submitted regressions in which the gender effect is reflected by an "indicator" variable, along with other predictors. As noted previously, this approach assumes that the expected underpayment of a female employee is the same, regardless of the values of the other predictors and often it is less powerful than PB when the data does not follow this assumption. A pattern of discrimination in pay may exist even if the coefficients of the regression models are not identical in each store or stratum. Thus, the Chow test should not be required before pooling the data. For example, if some stores started men and women at the same salary but gave men a higher percentage raise or other stores gave similar pay raises but started women at a lower salary the coefficients of the fitted equations would not be equal. Yet, there is a common pattern of underpayment of females.

To illustrate these issues we conducted a small simulation study in a simple situation. While the concepts discussed are general and applicable to class certification cases, they need to be interpreted in the context of a particular case.

5.1 Statistical Properties of the two Regression Approaches when the Data are Stratified

The issue of how to analyze data when a firm has many locations has arisen in a number of major discrimination cases. In particular, the brief to the appeals court filed by Wal-mart essentially claims that the Chow-Rao test must be used before the data are pooled. Here we present a simple example showing that this is too stringent. Combining the estimated differentials obtained by the Peters-Belson approach in the strata is more appropriate than ordinary least squares applied to the pooled data. Nonetheless, it will be seen that one can reach a correct inference when analyzing the pooled data with either OLS or PB.

Consider a company that has two stores with employees of both sexes. Women are underpaid relative to comparable men in both stores but the system leading to the disparity differ in the two stores. In the first, men and women start at the same salary but men receive better raises over time while in the second store men start at a slightly higher salary and also receive higher raises over time. Formally the salary equations are

$$y_i = \begin{cases} 20000 + 200x_i + \epsilon_{1i} & \text{when } i \text{ is female} \\ 20000 + 250x_i + \epsilon_{2i} & \text{when } i \text{ is male} \end{cases} \tag{24}$$

for store 1 and

$$y_i = \begin{cases} 20000 + 200x_i + \epsilon_{1i} & \text{when } i \text{ is female} \\ 20500 + 250x_i + \epsilon_{2i} & \text{when } i \text{ is male} \end{cases} \tag{25}$$

for store 2. We fitted PB regression and ordinary least squares with an indicator variable for gender.

The values of the seniority predictor variable for females and males were generated from the Gamma distribution with scale parameters 3 and 2 for females and males, respectively, and shape 2 for both sexes. In both stores, the number of males is 40 and the number of females is 30. The error variances were the same. One thousand replicates were used in the simulation.

First, we analyzed the data for each store using PB and OLS with an indicator for Gender regression. In store 1 the PB method detected a significant disparity at the .05 level 82.15% of the time while OLS found a statistically significant disparity 75.5% of the time. In store two both methods had 100% power.

Combining both stores into a single sample, we followed the approach of the plaintiffs' expert in Wal-mart who used OLS with indicators for store and gender. The Gujarati (1970a,b) implementation of the Chow-Rao test for the equality of the coefficients in two regressions was carried out by fitting a model that included gender by store and gender by seniority interactions. Then the partial F test for both coefficients is used to assess whether they are needed. Although the average p-value of the test was .0164, indicating that most of the time the Chow-Rao test would find that the coefficients are statistically significantly different at the 0.05 level, both the PB and OLS applied to the pooled sample detected an overall negative effect on female employees every time.

In Table 5 the estimated disparities obtained by the different methods are reported. Because each store employed the same number of males and females, the PB estimate of the overall disparity is the average of the individual store disparities. Notice that using OLS with a gender indicator, yielded a smaller disparity than PB in each store. When the pooled data were analyzed, the estimates obtained from both regression approaches were close to the average of those obtained from the analysis of the stores individually. While this may not be a general phenomenon, the example shows that fitting PB or OLS regression to data pooling over strata or stores where the regressions are not identical but reflect a similar type of disparity in the stores, still detected the general pattern that females were disadvantaged. Analogous to checking for a cross-over effect when the results from several two by two contingency tables are combined (Gail and Simon, 1985; Gastwirth, Miao and Zheng, 2003) one should check that the regressions in a substantial majority of stores follow similar pattern of disparity against one group. In stores where the sample sizes are small, especially when there is an imbalance between the number of employees of the two sexes, the power of any regression type test to find a statistically significant disparity in a single store is likely to be small.

Table 5: Estimated Disparities

	Average $\bar{D}$	Average $\hat{\beta}_2$	St. Dev.	Average $\bar{D} - \delta$	Average $\hat{\beta}_2 - \delta$
Store 1 - PB	-303.561		99.914	-1.198	
Store 1 - OLS		-271.098	98.34		31.265
Store 2 - PB	-803.570		98.962	-1.207	
Store 2 - OLS		-771.184	97.059		31.179
Pooled - PB	-553.566		71.856	-1.202	
Pooled - OLS		-521.141	70.613		31.223
OLS with Store Effect		-521.134	70.612		31.222
OLS with Store by x and Store by Gender Effect		-271.098	98.34		281.265

6 Concluding Remarks

Even though the PB approach is related to estimating the "counterfactual" in modern treatments of causality (Pearl, 2000, p. 308; Heckman 2005; Rubin, 2005) it is not widely used outside the area of measuring minority disparities (Sinclair and Pan, 2009; Graubard, 2009). Depending on one's prior knowledge concerning the relationship between the response of interest and the predictor variables one can estimate the regression from a parametric model or a non-parametric one.

Bhattacharya's bandwidth matching approach computes two estimates of a differential based on the Mann-Whitney-Wilcoxon measure. The first matches each female to a set of male "near-neighbors" and the second matches each male to a set of "nearby" females. This approach may well have greater statistical power in some situations than the PB method. In the context of the Equal Pay Act, however, the law is asymmetric in the sense that the salaries of the underpaid group need to be increased to those of the majority group, but the employer cannot decrease the salaries of the employees of the majority group.

Regression and matching procedures lose statistical power and efficiency when the distribution of predictors varies noticeably in the two groups. The article by Hall, Leng and Muller (2008) provides a useful approach to this issue, which deserves further study.

Using the PB estimates of the pay differential in each stratum and combining them to obtain a summary test of equality avoids the problem of aggregation that occurs in many class action cases. While we have seen that the Chow-Rao test is too stringent, further research is needed to develop a test for "no interaction" between the strata and the response of interest, i.e., to ensure that the pattern of signs and values of the stratum specific estimated D_i is consistent with a common general policy.

7 Appendix

The distribution of (21) when the variances are unequal: As in parametric PB, Welch's approximation can be used to find the approximate distribution of the test statistic. The expression for the variance of $\bar{D}_{LOC}$ in (16) can be simplified as follows:

$$\begin{aligned} var(\bar{D}_{LOC}) &= \frac{1}{n_1^2}\Big(n_1\sigma_1^2 + \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} S_{ij}^2\sigma_2^2 + \sum_{i\neq l}^{n_1}\sum_{j=1}^{n_2} S_{ij}S_{lj}\sigma_2^2\Big) \\ &= \sigma_1^2\frac{1}{n_1} + \sigma_2^2\frac{1}{n_1^2}\Big(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} S_{ij}^2 + \sum_{i\neq l}^{n_1}\sum_{j=1}^{n_2} S_{ij}S_{lj}\Big) \\ &= \sigma_1^2 C_1 + \sigma_2^2 C_2 \end{aligned}$$

When $\delta = 0$ is true, the test statistic in (21) is approximately t distributed with degrees of freedom

$$df_{LOC} = \frac{(\sigma_1^2 C_1 + \sigma_2^2 C_2)^2}{\sigma_1^4\kappa_1/\eta_1^2 C_1^2 + \sigma_2^4\kappa_2/\eta_2^2 C_2^2} \tag{26}$$

In most practical situations σ_1^2 and σ_2^2 are unknown and the degrees of freedom are approximated by plugging $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ in (26).

References

[1] Bhattacharya, P. K. (1989). Estimation of treatment main effect and treatment-covariate interaction in observational studies using bandwidth-matching. Technical Report 188, Division of Statistics, University of California, Davis.

[2] Bhattacharya, P. K. and Gastwirth, J. L. (1999). Estimation of the Odds-Ratio in an Observational Study Using Bandwidth-Matching. *Journal of Nonparametric Statistics*, **11**, 1-12.

[3] Belson, W. A. (1956). A technique for studying the effects of a television broadcast. *Applied Statistics*, **5**, 195-202.

[4] Blinder, A. S. (1973). Wage discrimination: reduced form and structural estimates. *Journal of Human Resources*, **8**, 436-455.

[5] Bura, E. and Gastwirth, J. L. (2009). How accurate are the power calculations relied upon the SEC? *Law, Probability and Risk*, **8**, 277-288.

[6] Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, **28**, 591-603.

[7] 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.

[8] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* London: Chapman and Hall.

[9] Gail, M. H. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subjects. *Biometrics*, **41**, 361-372.

[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] Gastwirth, J. L., Miao, W. and Zheng, G. (2003). Statistical issues arising in disparate impact cases and the use of the expectancy curve in assessing the validity of pre-employment tests. *International Statistical Review*, **71**, 565-580.

[13] Gastwirth, J. L. and Pan, Q. Statistical Measures and Methods for Assessing the Representativeness of Juries: A Reanalysis of the data in Berghuis v. Smith. To appear in *Law, Probability and Risk*.

[14] Graubard, B. I. (2009). Comment on "Using the Peters-Belson method in equal employment opportunity personnel evaluations" by Sinclair and Pan. *Law, Probability and Risk*, **8**, 119-122.

[15] 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.

[16] Greiner, D. J. (2008). Causal inference in civil rights litigation. *Harvard Law Review*, **122**, 533-598.

[17] Gujarati, D. (1970a). Use of dummy variables in testing for equality between sets of coefficients in two linear regressions: A note. *The American Statistician*, **24**, 50-52.

[18] Gujarati, D. (1970b). Use of dummy variables for equality between sets of coefficients in linear regressions: A generalization. *The American Statistician*, **24**, 18-22.

[19] Hall, P., Leng, X. Y. and Muller H. G. (2009). Weighted-bootstrap alignment of explanatory variables. *Journal of Statistical Planning and Inference*, **138**, 1817-1827.

[20] Heckman, J. J. (2005). The scientific model of causality. *Sociological Methodology*, **35**, 1-97.

[21] Hikawa, H., Bura, E. and Gastwirth, J.L. (2010a). Local Linear Logistic Peters-Belson Regression and its application in employment discrimination cases. *Statistics and its Interface*, **3**, 125-144.

[22] Hikawa, H., Bura, E. and Gastwirth, J. L. (2010b). Robust Peters-Belson Type Estimators of Measures of Disparity and their Applications in Employment Discrimination Cases. *Technical Report*, Department of Statistics, George Washington University.

[23] Loader, C. (1999). *Local Regression and Likelihood*. New York: Springer-Verlag.

[24] Nayak, T. K. and Gastwirth, J. L. (1997). The Peters-Belson approach to measures of economic and legal discrimination. In: Johnson N.L., Balakrishnan N. eds. *Advances in the Theory and Practice of Statistics: a Volume in Honor of Samuel Kotz*. Hoboken, NJ: Wiley.

[25] Oaxaca, R. (1973). Male-Female differentials in urban labor markets. *International Economic Review*, **14**, 693-709.

[26] Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

[27] Peters, C. C. (1941). A method of matching groups for experiments with no loss of populations. *Journal of Educational Research*, **34**, 606-612.

[28] Rao, C. R. (1952). *Advanced Statistical Methods in Biometric Research*. New York: John Wiley & Sons.

[29] Rosenbaum, P. and Rubin, D. B. (1983). Central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41-55.

[30] Rubin, D. B. (2005). Causal inference using potential outcomes: design, modeling, decisions. *Journal of the American Statistical Association*, **100**, 322-331.

[31] Scheffe, H. (1970). Practical solutions of the Behrens-Fisher problem. *J. Amer. Statist. Assoc.*, **65**, 1501-1508.

[32] Sinclair, M. D. and Pan, Q. (2009). Using the Peters-Belson Method in EEO Personnel Evaluations. *Law, Probability & Risk*, **8(2)**, 95-117 with Comments from Dr. Barry Graubard: 119-122.

[33] Welch, B. L. (1949). Further note on Mrs. Aspin's tables and on certain approximations to the tabled function. *Biometrika*, **36**, 293-296.

[34] Zanutto, E. L. (2006). A comparison of propensity score and linear regression: analysis of complex survey data. *Journal of Data Science*, **4**, 67-91.