

Evaluating a Simple Method for Estimating Black-White Gaps in Median Wages

by

William Johnson, Yuichi Kitamura, and Derek Neal*

Racial differences in wage rates are important measures of economic inequality among races because, for almost all individuals, labor income constitutes the most important component of lifetime income. As a consequence, the prices at which individuals may sell their time provide vital information about the distribution of welfare and economic success. However, we only observe prices when market transactions occur, and thus we only observe wage rates for individuals who are employed. Since the wages of employed workers are not randomly sampled from the distribution of potential wages, it is difficult to draw inferences concerning racial gaps in potential wages from data on observed wage rates. Richard Butler and James Heckman (1977) raised this issue in the context of assessing the impact of government policies on racial income inequality. Charles Brown (1984), James Smith and Finis Welch (1989), and Amitabh Chandra (2000) also examine the extent of racial differences in participation rates and the impact of these differences on measures of racial wage and income gaps. This topic remains salient, in part, because employment rates among working-age black males remain significantly below corresponding rates for whites.

Derek Neal and William Johnson (1996) estimated racial gaps in median wages among men by imputing wages of zero for all men in a particular cross-section who report that they have not worked at all during the survey period. Under a specific assumption concerning the distribution of missing wages, this procedure yields consistent estimates of the black-white gap in median wages conditional on observed characteristics. Below, we spell out this assumption

and use panel data to investigate the extent to which it may be violated in cross section wage analyses. Our results suggest that imputing wages of zero for unemployed individuals may provide a reasonable way to estimate median wage regressions among men.

I. A Simple Imputation Method

Consider the following linear model:

$$w_i = X_i' \beta_0 + \varepsilon_i$$

where w_i , X_i , and ε_i are the wage offer, observed characteristics, and unobserved traits for individual i . The conditional median of ε_i given X_i is assumed to be zero. We are interested in identifying the unknown parameter vector β_0 . Our problem is that w_i is not observed for those who do not work, $I_i=0$. We proceed by creating a variable y_i such that $y_i = w_i$ if $I_i = 1$ and $y_i = 0$ if $I_i = 0$, and we assume that the following condition holds: $w_i < X_i' \hat{\beta}$ if $I_i = 0$ (Condition A). Here, $\hat{\beta}$ is a hypothetical LAD estimator based on the true wage offers, w_i . Given these assumptions, LAD estimation using y_i has the following property:

$$\hat{\beta}_{\text{imputed}} = \operatorname{argmin}_{\beta} \sum_{i=1}^N |y_i - X_i' \beta| = \hat{\beta} = \operatorname{argmin}_{\beta} \sum_{i=1}^N |w_i - X_i' \beta|$$

because condition (A) implies that the LAD estimation is not affected at all by the imputations.¹

Further, since the hypothetical LAD estimator $\hat{\beta}$ is consistent, we know that $\hat{\beta}_{\text{imputed}}$ is also a consistent estimator of β_0 . (See Peter Bloomfield and William Steiger (1983) for details, pp. 44-52).

II. The Imputation Method Examined

The method we have described is easy to implement and also has important consequences

for estimates of the effect of race on wages. To see this, compare the two median wage regressions presented in the first two columns of Table 1. The dependent variable is the log of the average wage earned over the period 1990-91 and the data, from the National Longitudinal Survey of Youth, are the same observations used in Neal and Johnson (1996). We lack wage observations only for those who work neither in 1990 nor in 1991. In the first regression, we simply eliminate all individuals who did not work in either interview year. In the second regression, we replicate the Neal and Johnson results by imputing a wage of zero for all individuals who do not work.² A comparison of the two regressions reveals that the results in column one may understate the magnitude of the black-white wage gap by failing to account for the missing data problem created by individuals who are not employed. The estimated black-white gap in log wages expands by 50% from -.091 to -.134 when we add the 81 imputed wages for people who did not reporting working in either interview year.

Joseph Altonji and Rebecca Blank (1999) question the wisdom of imputing low wages for individuals who are not working. They argue that some of the NLSY respondents who are not working may be high-wage workers who are temporarily unemployed or out of the labor force. We can never know exactly what wages these workers would have received if they had worked during 1990 or 1991. However, we can shed light on Altonji and Blank's conjecture by exploiting the panel nature of the NLSY data. We look at data from two years beyond and two years before the 1990-91 period to find wage observations for the 81 individuals who report not working in the 1990 and 1991 interview years.

Panel A of Table 2 summarizes our findings. The second column reports that in 49 of 81 cases, we are able to find a wage observation in at least one of the surveys from years 1988,

1989, 1992, or 1993. The third column breaks down the locations of these wage observations. Eight men did not report a valid wage for the 1988-89 period but did report a valid wage during the 1992-93 period. Another 23 men report the opposite, while yet another 18 report wages in both the before and after periods.

The fourth column describes the relationships between these new wage observations and the predicted median wage given each individual's characteristics. We can interpret the results from this column in the context of standard search theory. Ignoring the complication of finite life spans, simple search models predict that each worker's reservation wage will be constant over time. Regardless of the details, all models with a constant reservation wage imply that any person who reports a wage before or after 1990-91 that is lower than the predicted median wage given his characteristics must have also faced a best offer during 1990-91 that was below this predicted median. To see this, consider a worker i , who reports a wage in 1992 that is less than the predicted median given his characteristics, $W_{i92} < X_i \hat{\beta}$. We know that although $W_{i92} < X_i \hat{\beta}$, W_{i92} exceeds worker i 's reservation wage and therefore exceeds any offers that he received during the 1990-91 period. Note that this argument holds even if worker i also reports a wage for 1989 that is greater than his predicted median. If an individual's reservation wage is constant over time, the minimum of observed wages must bound his unaccepted and hence unobserved wage offers from above.

Given this framework, ten cases appear problematic. Eight individuals, who do not report wages during the 1992-93 period, do report wages during 1988-89 that exceed the predicted medians based on their characteristics. Two more individuals, who report wages in both the 1988-89 and 1992-93 periods, always report wages greater than predicted medians given their

characteristics. Because our assumption of a constant reservation wage seems less attractive in cases where persons report health problems, we are especially interested in outcomes for workers reporting no disabilities, given in parentheses for each category. Among these men, only one reports wages both before and after 1990-91 that exceed the predicted median based on his characteristics. Four others, who do not report wages in the 1992-93 period, do report wages before 1990-91 that satisfy this criterion. Thus, five of the 81 imputations, or just over six percent, seem particularly suspect.

It is difficult to draw firm conclusions based on these data. Even in the five cases noted above, it is possible that these workers lost their jobs and temporarily (during 1990-91) received wage offers that were not only below their reservation wage but also below the predicted median based on their characteristics. Thus, it is possible that all five cases involve valid imputations. On the other hand, the second column indicates that 32 individuals never report a wage during the entire 1988-1993 period. We assume that, relative to others with similar education and experience, these individuals actually face low wage offers. However, we have no direct evidence that this is true.

In sum, if those who never worked during the 1988-93 period actually faced low wage offers given their characteristics, then the vast majority of our 81 wage imputations likely involve individuals who faced wage offers during 1990-91 that were below predicted medians given their characteristics. Further, the final three columns of Table 1 show that, in this NLSY sample, estimates of racial gaps in median wages do not change much when we incorporate wage data from the 1988, 1989, 1992, and 1993 surveys. These columns report results derived from different rules for assigning 1990-91 wages based on wage observations found in other survey

years. In all cases, the original imputation procedure produces estimates that are very close to those based on the expanded data.

III. Imputations with a Mincerian Earnings Function

The specifications employed in Table 1 use scores on the Armed Forces Qualifying Test to control for skill, but these scores are not available in most data sets. Further, following Neal and Johnson, these regressions use data from only the three youngest cohorts in the NLSY. We now explore the validity of the imputation procedure described above using a more common median regression specification and data from all birth cohorts. Table 2B provides information analogous to that in 2A, but in this case, predicted medians are based on a Mincerian wage equation that includes schooling, potential experience, and potential experience squared. Further, the analysis is based on a single cross-section of the NLSY, the 1992 wave. Here, there are 294 persons with missing wage observations for 1992. Of these, 178 report wages during either the 1990-91 period or the 1993-94 period.

The results in Table 2B provide slightly less support for the imputation rule described above. In this case, 45 individuals (7+28+10) only report wages that exceed predicted medians based on their characteristics. Of these, 29 (6+16+7) do not report disabilities. These 29 individuals represent less than ten percent of the sample of imputations. The results in Table 2C mirror those in Table 2B but are based on a regression involving two-year wage averages.³ By using wage information from two years, we reduce the need for imputations. Table 2C reports a larger total sample than 2B but only 205 imputations compared to 294 in Table 2B. Only 3 of these 205 cases involve persons without a disability who report wages above predicted medians based on their characteristics both before and after the 1991-92 period. Further, only 13 cases,

or just over six percent, involve persons without disabilities who only report wages above their relevant predicted medians.

Using the data summarized in Tables 2B and 2C, we have computed regressions like those in Table 1. These results appear in Tables 3A and 3B. Once again, we find evidence that regression results based only on samples of persons who are currently working tend to understate the magnitude of the black-white wage gap. Both median regressions based on the imputation rule described above and median regressions involving imputations and additional wage data from adjacent years yield black-white wage gaps that are greater than those based on the sample of observed wages. However, the results involving imputations and additional wage data imply gaps that are slightly smaller than those implied by median regressions involving imputations alone.

IV. Conclusion

Imputing below median wages to workers with no wage observations may be a simple and fairly accurate way of handling selection problems when estimating median wage regressions among men. This procedure significantly affects estimates of racial gaps in median wages. Using data from short panels rather than single year cross-sections may mitigate the need for additional imputations and also reduce the frequency of imputation error.⁴

Table 1: Median Regression Results Using Various Wage Imputation Methods: NLSY 1990-91
(Standard errors in parentheses)

	Restricted Sample: No Imputations	Impute Zero if Missing	Use New Wage Data: 1992-93	Use New Wage Data: 1988-89	Use New Wage Data: 88,89,92&93
Black	-.091(.036)	-.134(.034)	-.141(.035)	-.138(.033)	-.139(.031)
Hispanic	.013(.039)	-.014(.038)	-.018(.038)	-.017(.037)	-.017(.035)
Age	.058(.017)	.055(.017)	.057(.017)	.061(.016)	.060(.015)
AFQT	.197(.016)	.206(.015)	.202(.015)	.200(.015)	.200(.014)
AFQT ²	.007(.014)	.010(.014)	.011(.014)	.010(.013)	.010(.013)
N	1593	1674	1674	1674	1674

For details concerning these samples and those used in tables 2A, 2B, and 2C, see <http://www.ssc.wisc.edu/~dneal>.

Table 2: New Wage Observations: Two Years After and Two Years Before Original Sample
A. Neal and Johnson (1996) - Sample Period: 1990-1991

Wage Observation in Sample Period ?	Other Wage Observation ?	Timing of other wage observation	Other Wage Observation Greater than Predicted Median given X_i ?
Yes: 1593			
No: 81(49)	No: 32(20)		
		After only: 8(6)	Yes: 0(0) No: 8(6)
		Before only: 23(11)	Yes: 8(4) No: 15(7)
		Both before and after: 18(12)	Always: 2(1) Sometimes: 5(4) Never: 11(7)

B. Mincer Regression - Sample Period: 1992

Yes: 3662			
No: 294(157)	No: 116(46)		
		After only: 31(22)	Yes: 7(6) No: 24(16)
		Before only: 78(41)	Yes: 28(16) No: 50(25)
		Both before and after: 69(48)	Always: 10(7) Sometimes: 25(16) Never: 34(25)

C. Mincer Regression with Two-Year Average Wage - Sample Period: 1991-1992

Yes: 4003			
No: 205(94)	No: 100(41)		
		After only: 27(20)	Yes: 8(7) No: 19(13)
		Before only: 49(20)	Yes: 14(3) No: 35(17)
		Both before and after: 29(13)	Always: 5(3) Sometimes: 5(1) Never: 19(9)

Note: numbers in parentheses are those not reporting disability.

Table 3A: Median Regression Results Using Various Wage Imputation Methods: NLSY 1992
(Standard errors in parentheses)

	Restricted Sample: No Imputations	Impute Zero if Missing	Use New Wage Data: 1993-94	Use New Wage Data: 1990-91	Use New Wage Data: 90,91,93&94
Black	-.300(.021)	-.362(.022)	-.351(.023)	-.343(.023)	-.338(.025)
Hispanic	-.079(.024)	-.091(.025)	-.089(.027)	-.081(.027)	-.084(.029)
Highest Grade Completed	.090(.005)	.101(.005)	.099(.005)	.101(.005)	.098(.006)
N	3662	3956	3956	3956	3956

Each regression also includes controls for potential experience and its square.

Table 3B: Median Regression Results Using Various Imputation Methods: NLSY 1991-92
(Standard errors in parentheses)

	Restricted Sample: No Imputations	Impute Zero if Missing	Use New Wage Data: 1993-94	Use New Wage Data: 1989-90	Use New Wage Data: 89,90,93&94
Black	-.302(.017)	-.335(.017)	-.325(.018)	-.329(.017)	-.322(.017)
Hispanic	-.097(.020)	-.102(.020)	-.102(.020)	-.098(.020)	-.099(.019)
Highest Grade Completed	.076(.004)	.081(.004)	.081(.004)	.081(.004)	.081(.004)
N	4003	4208	4208	4208	4208

Each regression also includes controls for potential experience and its square.

References:

- Altonji, Joseph and Blank, Rebecca. "Race and Gender in the Labor Market" in Orley Ashenfelter and David Card, eds., Handbook of Labor Economics. Vol. 3, Amsterdam: North Holland, 1999.
- Bloomfield, Peter and Steiger, William L. Least Absolute Deviations: Theory, Applications, and Algorithms. Boston: Birkhauser, 1983.
- Brown, Charles. "Black-White Earnings Ratios since the Civil Rights Acts of 1964: The Importance of Labor Market Dropouts." Quarterly Journal of Economics , February 1984, 91(1), pp. 31-44.
- Butler, Richard and Heckman, James J. "The Government's Impact on the Labor Market Status of Black Americans: A Critical Review," in L. Hausman, et al, eds. Equal Rights and Industrial Relations. Madison: Industrial Relations Research Association, 1977.
- Chandra, Amitabh. "Is the Convergence in the Racial Wage Gap Illusory ?", mimeo, University of Kentucky, 2000.
- Heckman, James J. "The Impact of Government." in S. Shulman and W. Darity, eds., The Question of Discrimination. Middletown, CT: Wesleyan University Press, 1989, pp. 50- 80.
- Neal, Derek and Johnson, William."The Role of Premarket Factors in Black-White Wage Differences." Journal of Political Economy , October 1996, 104(5), pp. 869-895.
- Smith, James and Welch, Finis. "Black Economic Progress after Myrdal." Journal of Economic Literature. June 1989, 27(2), pp. 519-564.

Endnotes:

*University of Virginia; University of Wisconsin; and University of Wisconsin and NBER

1. Here, we assume that the LAD estimator is unique.

2. The coefficients in these columns do not match Neal and Johnson (1996) exactly because the NLSY data are edited over time as coding errors are found.

3. Here, the total sample is larger than in Table 2B. Some persons who report a valid wage in 1991 and did not report a valid wage in 1992 were actually working in both years, but coding problems contaminated their 1992 wage records. In our 1992 cross-section analyses, we eliminate these workers from the sample. We do not impute wages of zero unless individuals report that they did not work during the sample period in question.

4. Note that computing average wages over short panels involves implicit imputations since only years with valid wage data contribute to the average calculations.