Decomposition of the Gender Wage Gap using the LASSO Estimator

by

René BÖHEIM

Philipp STÖLLINGER

Working Paper No. 2003

January 2020

Johannes Kepler University of Linz
Department of Economics
Altenberger Strasse 69
A-4040 Linz - Auhof, Austria
www.econ.jku.at

Corresponding author: rene.boeheim@jku.at

Decomposition of the Gender Wage Gap using the LASSO Estimator*

René Böheim[†] Philipp Stöllinger[‡]

Abstract

We use the LASSO estimator to select among a large number of explanatory variables in wage regressions for a decomposition of the gender wage gap. The LASSO selection with a one standard error rule removes about a quarter of the regressors. We use the LASSO-selected regressors for OLS-based gender wage decompositions. This approach results in a smaller error variance than in OLS without LASSO-selection. The explained gender wage gap is 1%-point greater than in the conventional OLS model.

Keywords: gender wage gap, LASSO, decomposition

JEL classification: J31, J71

1 Introduction

Surveys such as the PSID provide a large number of characteristics and techniques for the selection of explanatory variables have become popular in recent years

*Lawrence M. Kahn kindly provided the code for transforming the raw PSID data into the data used in Blau and Kahn (2017*a*,*b*).

†Department of Economics, Johannes Kepler University Linz, Austria. Email: Rene.Boeheim@jku.at. Böheim is also associated with CESifo, NBER, WIFO, and IZA.

‡Department of Economics, Vienna University of Economics and Business, Austria. Email: Philipp.Stoellinger@s.wu.ac.at.

(Barigozzi and Brownlees, 2013; Belloni, Chen, Chernozhukov and Hansen, 2012; Belloni, Chernozhukov and Hansen, 2014; Varian, 2014). The Least Absolute Shrinkage and Selection Operator (LASSO) estimator (Tibshirani, 1996) estimates coefficients and simultaneously selects explanatory variables, based on objective criteria. It performs better than OLS when some of many coefficients might be zero (Dormann, Elith, Bacher, Buchmann, Carl, Carré, Marquéz, Gruber, Lafourcade, Leitão, Münkemüller, McClean, Osborne, Reineking, Schröder, Skidmore, Zurell and Lautenbach, 2013; Leng, Lin and Wahba, 2006). The reduction of explanatory variables also results in specifications which are easier to interpret, however, at the potential cost of increased bias (Tibshirani, 1996).

Selection approaches are evaluated by their out-of-sample prediction accuracy and their mean-squared prediction error (Athey, 2018). An OLS regression that uses variables selected by the LASSO estimator, "OLS post-LASSO", performs at least as well as the LASSO estimator (Belloni and Chernozhukov, 2013). It has the advantage that the estimates are less biased than LASSO estimates.

We use the OLS post-LASSO approach to estimate gender wage gap decompositions (Blinder, 1973; Oaxaca, 1973) using data from the Panel Study of Income Dynamics (PSID) for 2006 and 2016. We contrast these results with results from standard decompositions. The gender wage gap decompositions based on the post-LASSO approach differ from OLS-based decompositions by the rule used for the shrinking parameter. Using a conventional rule of one standard error, the LASSO estimator removes about a quarter of the explanatory variables. This lowers the estimated error variance by about 0.001 for women and by 0.002 for men.

Our results of the OLS post-LASSO specification confirm the results obtained by the conventional approach. A comparison of the results with a conventional

OLS specification shows that the explained gender wage gap is about 1% greater than obtained by conventional OLS. We demonstrate that the OLS post-LASSO approach can improve estimates of gender wage decompositions through lower error variances.

2 Background

The standard econometric approach to study gender wage gaps are wage decompositions, based on wage regressions (e.g. Blinder, 1973; Oaxaca, 1973) or on estimating appropriate counterfactual distributions (e.g. DiNardo, Fortin and Lemieux, 1995; Firpo, Fortin and Lemieux, 2009; Machado and Mata, 2005)). Researchers aim to control for a wide range of characteristics to achieve a convincing comparison between men's and women's wages. The number of controls is typically large, potentially leading to sparsity in the estimated wage regressions.[1] In the presence of sparsity, OLS usually does not return coefficients of zero that are zero in the true underlying data generating process.

In gender wage gap studies, there is no standard set of explanatory variables. For example, Stanley and Jarrell (1998) report that in 55 analyzed studies one did not include the worker's experience and 63% did not control for a worker's industry. Weichselbaumer and Winter-Ebmer (2005) report similar results and suggest that the selection of explanatory variables is often a personal choice of the researcher.

Statistical techniques for subset-selection reduce the number of regressors from

[1] A statistical model with a coefficient vector that contains many zeros is called sparse (Hastie, Tibshirani and Friedman, 2009).

a set of explanatory variables based on some objective function.[2] The disadvantage of subset-selection techniques is potentially more bias (Tibshirani, 1996).[3] Tibshirani (1996) proposes the LASSO for subset-selection as it simultaneously performs model estimation and selects the subset of regressors. The LASSO estimator is a continuous method that shrinks some variables and drops others completely by penalizing the objective function of the OLS estimator (Hastie et al., 2009).

The OLS post-LASSO approach re-estimates the specification using OLS and the set of LASSO-selected coefficients. This removes bias caused by the LASSO-selection (Belloni and Chernozhukov, 2013).

3 Data Description

We use data from the Panel Study of Income Dynamics (PSID) (University of Michigan, 2015). The data contain the hours worked and the income earned for 1980, 1989, 1998, and every other year from 2006 to 2016 and it is the only source that includes information on actual labor-market experience for the full age range of the US population (Blau and Kahn, 2017*a*).

We select household heads and their spouses between the ages of 25 and 64, who do not work on farms, who are not self-employed, and who do not work for the military.[4] To reduce the impact of outliers, we exclude persons who earn less

[2]For example, Bach, Chernozhukov and Spindler (2018) analyze the gender wage gap using data from the 2016 American Community Survey and use the double LASSO method to select among up to 4,382 regressors. See also Angrist and Frandsen (2019).

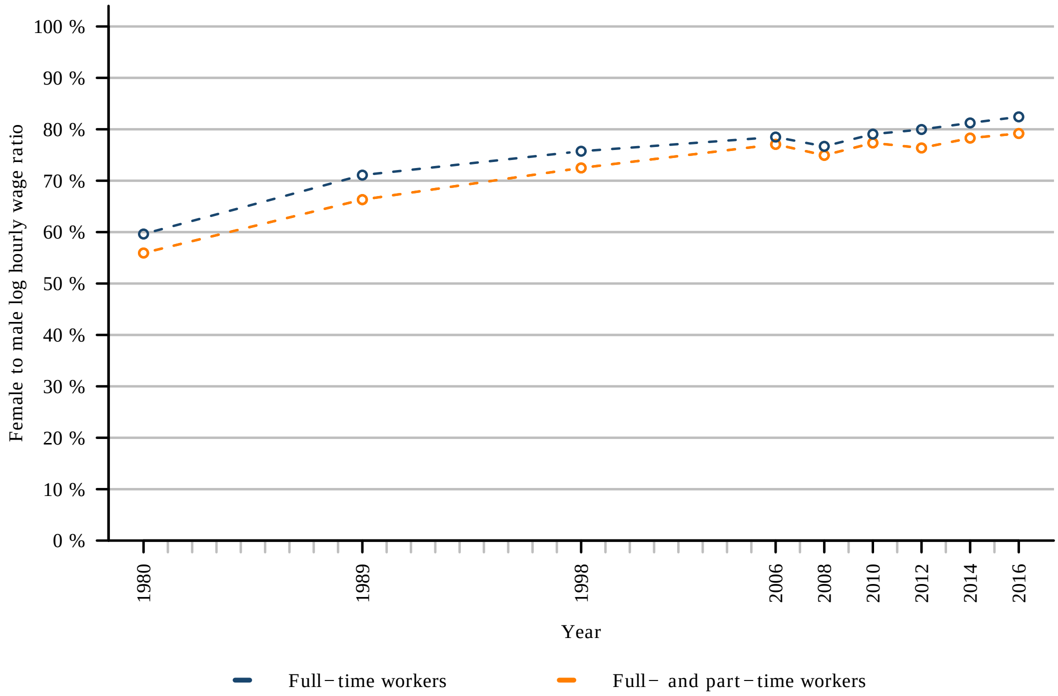
[3]Miller (1984) discusses different algorithms for the subset selection technique. The algorithms either evaluate all subsets of the set of explanatory variables or use a heuristic for which subsets to evaluate. They usually choose the subset that results in the lowest sum of squared residuals (Tibshirani, 1996).

[4]The PSID does not clearly distinguish between different sources of income for farm-workers and the self-employed.

than US$2 per hour and persons who work less than 26 weeks in a year. We drop observations with missing values for any of the explanatory variables (244 men and 235 women).

Figure 1 presents the log hourly wage ratio, women to men, unadjusted for any covariates. Between 1980 and 2016, women earned on average less per hour than men. Among full-time working women, the wage ratio rose from about 60% of full-time men's wages in 1980 up to about 82 % in 2016.

Figure 1: Women's to Men's Wages.



Note: Average of women's log hourly wage to men's wages ($e^{(\overline{log(wage_f)} - \overline{log(wage_m)})}$) using weights provided by the University of Michigan to compensate for both unequal selection probabilities and differential attrition in the PSID. Heads and spouses aged 25 and 64 who earned an hourly wage of at least US$2 (2016 prices) and who worked for at least 26 weeks during the year. Non-farming, non-military, non-self-employed wage and salary workers. 18,495 female and 19,254 male full-time workers; 22,590 female and 20,278 male workers, including part-time workers. Data from PSID, excluding observations from the Immigrant Sample added in 1997 and 1999.

Using data for 2006 and for 2016 we select 73 explanatory variables that are thought to be associated with a person's wage, such as education, experience, region, ethnicity, unionization, industry, occupation, health, family, hours housework, financial status, and job characteristics. Table 5 in the Appendix lists all variables.

Table 1 provides descriptive statistics of the explanatory variables for the years 2006 and 2016. Women were better educated than men in both 2006 and 2016.

Women's educational levels grew faster than men's from 2006 to 2016. Men had more full-time work experience than women in both years, but the gap between years spent working full-time by men and by women narrowed. All variables are standardized before estimation, but results are presented in their original scale.

4 Method

The LASSO estimator achieves subset-selection by minimization of the residual sum of squares, conditional on a penalty that depends on a tuning parameter. The objective function is given by:

$$\hat{\beta}^l = \arg\min_{\beta} \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{1}$$

where $\hat{\beta}^l$ is the vector of LASSO-estimated coefficients, and y is the vector of the dependent variables. x_j, $j = 1, ..., p$, is the vectors of the explanatory variables. p is the number of explanatory variables, and λ is a tuning parameter. The sum of the absolute values of the coefficients is less than the non-negative tuning parameter λ.

The tuning parameter controls the amount of shrinkage that is applied to the estimates. If λ is set to zero, the LASSO estimator is the OLS estimator. The larger λ, the more the LASSO estimator shrinks the coefficients towards zero. For sufficiently large λ, the LASSO estimator shrinks some coefficients to zero and the variable is eliminated from the set of explanatory variables (Tibshirani, 1996). We choose λ according to the "one standard error rule" (Breiman, Friedman, Olshen and J, 1984).[5] The one standard error rule sets λ to 0.0063.

Figure 2 shows the mean squared prediction error for different values of the natural logarithm of λ. The numbers on top of the plotted functions indicate how many coefficients are non-zero at the corresponding λ value. λ_{1se} refers to the λ-value chosen according to the one standard error rule.

We perform the following steps for the OLS post-LASSO approach: First we use the LASSO estimator on women and men combined, then we perform OLS regressions on women and men separately using only those variables selected by the LASSO estimator. We follow Belloni, Chernozhukov and Kato (2014) to perform inference for post-LASSO estimates. To compare different specifications we estimate the error variance using the estimator proposed by Fan et al. (2012) that is based on the mean squared prediction error generated by cross-validation.

5 Results

In order to evaluate the gender wage gap, we estimate wage regressions separately for men and women and use the male-based Oaxaca-Blinder decomposition (Blinder, 1973; Oaxaca, 1973).[6] We estimate wage regressions using two different specifications: An OLS specification which uses all explanatory variables, OLS^{all}; and a post-LASSO specification that is a re-estimation of the wage regressions including only the explanatory variables selected by the LASSO-estimator accord-

[5] We assess the quality of the fit using the cross-validation based, LASSO residual sum of squares estimator (Fan, Guo and Hao, 2012). Although this tends to be biased downwards, particularly for small values of λ (Fan et al., 2012), Reid, Tibshirani and Friedman (2016) show that the bias is typically not large.

[6] Our main interest is the comparison of the results arising from the OLS post-LASSO specification with results which are based on a standard OLS approach. Our specifications do not correct for selection, which could result in downward biased estimates (Albrecht, Van Vuuren and Vroman, 2009).

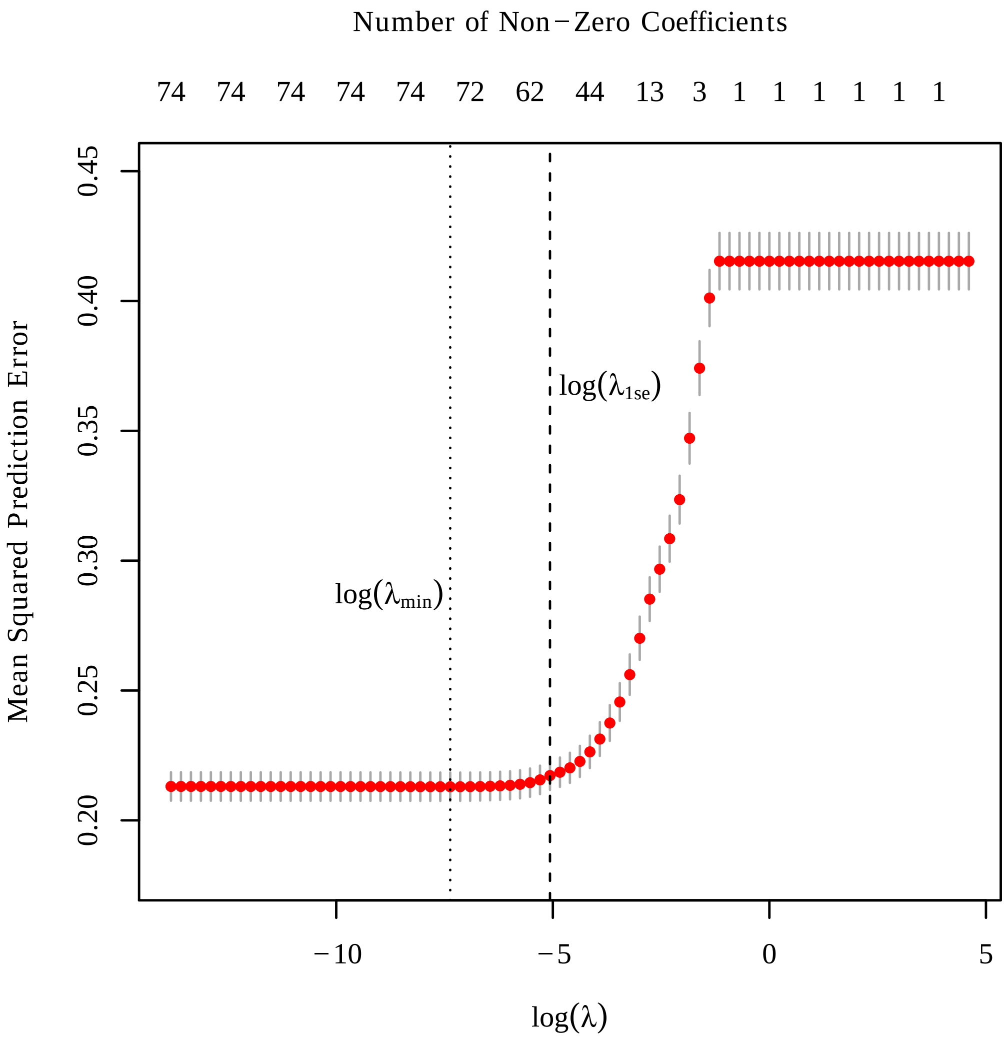
ing to the one standard error rule, POSTLASSO. Table 6 in the Appendix lists the estimated coefficients. The properties of the two different specifications are in Table 2 in the Appendix. The estimated error variance of the POSTLASSO specification is smaller than that of $\mathrm{OLS}^{\mathrm{all}}$.[7]

Figure 3 plots the gender wage gap and the explained parts of the two different specifications. The gender wage gap of 2016 was about 0.24 log points, which is 21.5 % of the average male wage of 2016. The explained gap is about 51 % of the gender wage gap according to the OLS specification. The OLS post-LASSO specification explains about 52 % of the gender wage gap. The absolute difference of the parts of the explained gender wage gap associated with the key characteristics education, experience, region, ethnicity, unionization, industry, occupation, health, family, hours of housework, financial status, and job characteristics obtained by the two different specifications is at maximum 0.01 log points. We decompose the change in the gender wage gap from 2006 to 2016 using the Smith-Welch decomposition (Smith and Welch, 1989).[8]

[7] The results of the Oaxaca-Blinder decomposition for 2016 are shown in Table 3 in the Appendix.

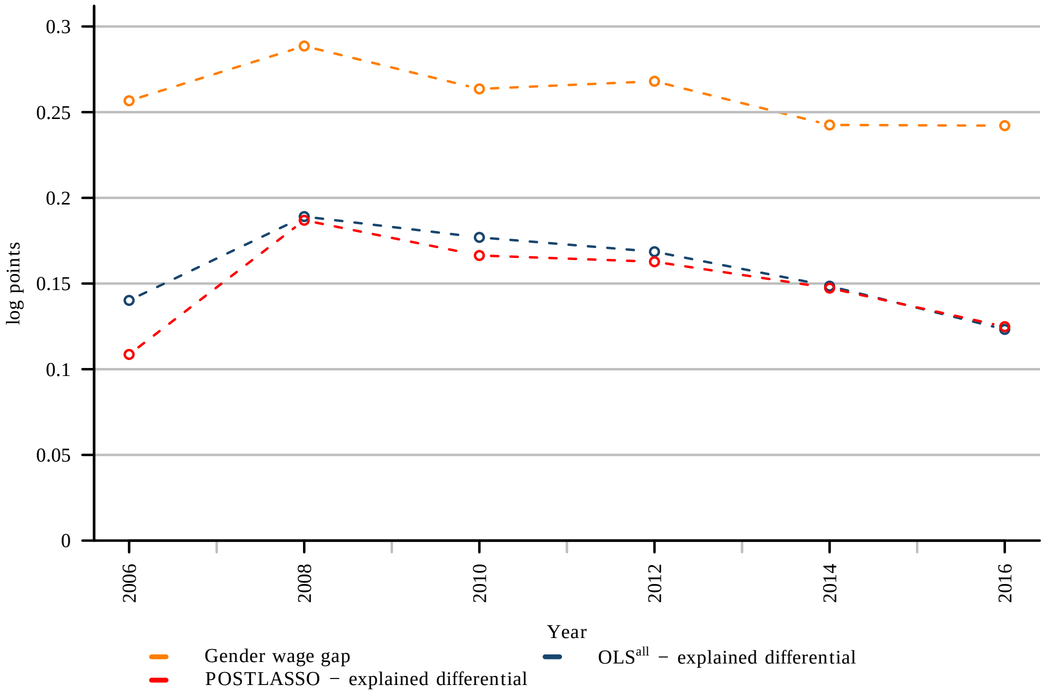
[8] The results of the Smith-Welch decomposition for the change between 2006 and 2016 are shown in Table 4 in the Appendix.

Figure 2: Cross-Validation for λ.



Source: Authors' calculations. Data from PSID. *Note:* The graph plots the mean squared prediction error, and its standard error bands, for different values of $log(\lambda)$ generated by cross-validation. λ_{min} is the λ value that minimizes the mean squared prediction error. λ_{1se} is the λ value that arises from the one standard error rule. The numbers on top of the graphs refer to the number of non-zero coefficients estimated by the LASSO estimator at the associated λ value. Weighted data for 2016 for heads and their spouses who were between 25 and 64 years of age, who earned an hourly wage of at least US$2, and who worked for at least 26 weeks. Non-farming, non-military, non-self-employed wage and salary workers. Excluding all persons with missing values for any of the explanatory variables of the wage regressions. $N = 3{,}390$ women and 2,985 men.

Figure 3: Gender Wage Gap and Explained Differential 2006 - 2016.



Source: Authors' calculations. Data from PSID.
Note: The graph plots the gender wage gap and the explained part using the male based Oaxaca-Blinder decomposition. OLS$^{\text{all}}$ is based on an OLS specification that uses all explanatory variables. POSTLASSO is based on an OLS specification that includes only the explanatory variables selected in a previous step by the LASSO estimator using the one standard error rule.
Heads and their spouses who were between 25 and 64 years of age, who earned an hourly wage of at least US$2, and who worked for at least 26 weeks in 2016. Non-farming, non-military, non-self-employed wage and salary workers. Excluding all persons with missing values for any of the explanatory variables of the wage regressions.
N = 2,756 women and 2,451 men in 2006, 2,957 women and 2,509 men in 2008, 2,945 women and 2,474 men in 2010, 3,153 women and 2,713 men in 2012, 2,635 women and 2,356 men in 2014, and 3,390 women and 2,985 men in 2016.

6 Conclusion

Our empirical analysis reveals that gender wage gap declined in the US between 2006 and 2016. The OLS post-LASSO decomposition are close to those of the conventional OLS-specification, however, it uses fewer variables and leads to more precise estimates. The OLS post-LASSO approach seems well-suited for decomposing the gender wage gap when there is a large number of explanatory variables.

References

Albrecht, James, Aico Van Vuuren and Susan Vroman (2009), 'Counterfactual distributions with sample selection adjustments: Econometric theory and an application to the Netherlands', *Labour Economics* **16**(4), 383–396.

Angrist, Joshua and Brigham Frandsen (2019), 'Machine labor', *NBER Working Paper 26584* .

Athey, Susan (2018), *The Impact of Machine Learning on Economics*, University of Chicago Press.

Bach, Philipp, Victor Chernozhukov and Martin Spindler (2018), 'Closing the US gender wage gap requires understanding its heterogeneity', *arXiv preprint arXiv:1812.04345* .

Barigozzi, Matteo and Christian Brownlees (2013), 'Nets: Network estimation for time series', *Journal of Applied Econometrics* .

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov and Christian Hansen (2012), 'Sparse models and methods for optimal instruments with an application to eminent domain', *Econometrica* **80**(6), 2369–2429.

Belloni, Alexandre and Victor Chernozhukov (2013), 'Least squares after model selection in high-dimensional sparse models', *Bernoulli* **19**(2), 521–547.

Belloni, Alexandre, Victor Chernozhukov and Christian Hansen (2014), 'Inference on treatment effects after selection among high-dimensional controls', *Review of Economic Studies* **81**(2), 608–650.

Belloni, Alexandre, Victor Chernozhukov and Kengo Kato (2014), 'Uniform post-selection inference for least absolute deviation regression and other z-estimation problems', *Biometrika* **102**(1), 77–94.

Blau, Francine D and Lawrence M Kahn (2017*a*), 'The gender wage gap: Extent, trends, and explanations', *Journal of Economic Literature* **55**(3), 789–865.

Blau, Francine D and Lawrence M Kahn (2017*b*), 'Online data appendix for: The gender wage gap: Extent, trends, and explanations'.
URL: *https://www.aeaweb.org/content/file?id=5300*

Blinder, Alan S (1973), 'Wage discrimination: Reduced form and structural estimates', *Journal of Human Resources* pp. 436–455.

Breiman, Leo, Jerome H Friedman, Richard Olshen and Stone Charles J (1984), *Classification and regression trees*, Chapman & Hall.

DiNardo, John, Nicole M Fortin and Thomas Lemieux (1995), 'Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach'.

Dormann, Carsten F., Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J. Leitão, Tamara Münkemüller, Colin McClean, Patrick E. Osborne, Björn Reineking, Boris Schröder, Andrew K. Skidmore, Damaris Zurell and Sven Lautenbach (2013), 'Collinearity: A review of methods to deal with it and a simulation study evaluating their performance', *Ecography* **36**(1), 27–46.

Fan, Jianqing, Shaojun Guo and Ning Hao (2012), 'Variance estimation using refitted cross-validation in ultrahigh dimensional regression', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**(1), 37–65.

Firpo, Sergio, Nicole M Fortin and Thomas Lemieux (2009), 'Unconditional quantile regressions', *Econometrica* **77**(3), 953–973.

Hastie, Trevor, Robert Tibshirani and Jerome H Friedman (2009), *The elements of statistical learning: Data mining, inference, and prediction*, New York, NY: Springer.

Leng, Chenlei, Yi Lin and Grace Wahba (2006), 'A note on the LASSO and related procedures in model selection', *Statistica Sinica* pp. 1273–1284.

Machado, José AF and José Mata (2005), 'Counterfactual decomposition of changes in wage distributions using quantile regression', *Journal of Applied Econometrics* **20**(4), 445–465.

Miller, Alan J (1984), 'Selection of subsets of regression variables', *Journal of the Royal Statistical Society. Series A (General)* pp. 389–425.

Oaxaca, Ronald (1973), 'Male-female wage differentials in urban labor markets', *International Economic Review* pp. 693–709.

Reid, Stephen, Robert Tibshirani and Jerome H Friedman (2016), 'A study of error variance estimation in LASSO regression', *Statistica Sinica* pp. 35–67.

Smith, James P. and Finis R. Welch (1989), 'Black economic progress after myrdal', *Journal of Economic Literature* **27**(2), 519–564.

Stanley, Tom D and Stephen B Jarrell (1998), 'Gender wage discrimination bias? A meta-regression analysis', *Journal of Human Resources* pp. 947–973.

Tibshirani, Robert (1996), 'Regression shrinkage and selection via the LASSO', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

University of Michigan (2015), 'Panel study of income dynamics - overview'. **URL:** *https://psidonline.isr.umich.edu/Guide/Brochures/PSID.pdf*

Varian, Hal R. (2014), 'Big data: New tricks for econometrics', *Journal of Economic Perspectives* **28**(2), 3–28.

Weichselbaumer, Doris and Rudolf Winter-Ebmer (2005), 'A meta-analysis of the international gender wage gap', *Journal of Economic Surveys* **19**(3), 479–511.

Appendix

Table 1: Descriptive Statistics by Sex, 2006 and 2016.

Year	Women	Men	Women − Men
Advanced degree			
2006	13.9%	13.3%	0.6 %-points
2016	15.5%	11.1%	4.5 %-points
Bachelor's degree			
2006	23.6%	23.4%	0.2 %-points
2016	26.4%	25.5%	0.9 %-points
Years of schooling			
2006	14.5	14.3	0.2
2016	14.7	14.3	0.4
Full-time years			
2006	15.1	18.5	−3.4
2016	14.8	16.5	−1.7
Part-time years			
2006	4.1	2.1	2.0
2016	3.7	2.2	1.5
Hours of housework			
2006	12.6	7.5	5.1
2016	12.3	7.6	4.6
Metropolitan county			
2006	67.7%	68.0%	−0.3 %-points
2016	83.6%	84.5%	−0.9 %-points
Union member			
2006	16.3%	18.1%	−1.7 %-points
2016	16.3%	15.8%	0.5 %-points
Disabled person			
2006	8.1%	7.2%	0.9 %-points
2016	7.2%	5.9%	1.2 %-points
Health status			
2006	61.1%	64.2%	−3.2 %-points
2016	56.5%	61.1%	−4.5 %-points
Mental problems			
2006	7.3%	5.1%	2.3 %-points
2016	10.7%	6.6%	4.1 %-points
Married			
2006	63.3%	71.3%	−8.0 %-points
2016	58.5%	66.4%	−7.9 %-points
Public sector job			
2006	28.0%	19.8%	8.2 %-points
2016	27.9%	17.7%	10.2 %-points
Part-time job			
2006	17.6%	3.8%	13.8 %-points
2016	16.8%	4.4%	12.4 %-points

Table 1: (continued).

Year	Women	Men	Women − Men
# of observations			
2006	2,756	2,451	305
2016	3,390	2,985	405

Source: Authors' calculations. Data from PSID.
Note: Weighted data for 2016 for heads and their spouses who were between 25 and 64 years of age, who earned an hourly wage of at least US$2, and who worked for at least 26 weeks. Non-farming, non-military, non-self-employed wage and salary workers. Excluding all persons with missing values for any of the explanatory variables of the wage regressions.

Table 2: Comparison of Different Regression Models.

	Women		Men	
	OLS^{all}	POSTLASSO	OLS^{all}	POSTLASSO
# observations	3,390	3,390	2,985	2,985
# coefficients	73	57	73	57
$\hat{\sigma}^2_{MPE}$	0.2013	0.2003	0.2321	0.2302
adj. R^2	0.5014	0.4983	0.5291	0.5262

Note: The table shows number of non-zero coefficients generated by different models, the error variance estimated based on the mean squared prediction error generated by cross-validation, and the adjusted coefficient of determination for different models by gender.

OLS^{all} is based on an OLS specification that uses all explanatory variables. POSTLASSO is a re-estimation by OLS-regression of the wage regressions including only the explanatory variables selected by the LASSO-estimator according to the one standard error rule.

Weighted data for 2016 for heads and their spouses who were between 25 and 64 years of age, who earned an hourly wage of at least US$2, and who worked for at least 26 weeks. Non-farming, non-military, non-self-employed wage and salary workers. Excluding all persons with missing values for any of the explanatory variables of the wage regressions.

Table 3: Oaxaca-Blinder Decomposition for 2016 - Grouped Variables.

	OLS$^{\text{all}}$		POSTLASSO	
Variable group	log points	% of gap	log points	% of gap
Education	-0.0319	-13.2	-0.0319	-13.2
Experience	0.0208	8.6	0.0226	9.4
Region	0.0013	0.6	0.0009	0.4
Ethnicity	0.0051	2.1	0.0052	2.2
Unionization	-0.0007	-0.3	-0.0007	-0.3
Industry	0.0376	15.5	0.0276	11.4
Occupation	0.0582	24.0	0.0632	26.1
Health	0.0257	10.6	0.0264	10.9
Family	0.0092	3.8	0.0096	4.0
Hours housework	0.0053	2.2	0.0065	2.7
Financial Status	0.0020	0.8	0.0021	0.9
Job characteristics	-0.0095	-3.9	-0.0066	-2.7
Explained differential	0.1232	50.9	0.1248	51.6
Unexplained differential	0.1189	49.1	0.1173	48.4
Gender wage gap	0.2421	100.0	0.2421	100.0

Note: The table shows the gender wage gap, the explained differential, and the unexplained differential calculated using the male based Oaxaca-Blinder decomposition. The dependent variable is the logarithm of the hourly wage. The presented gender wage gap is the result of the mean male log hourly wage minus the female counterpart. For each variable group, the table shows the part of the gender wage gap that is explained by the variable group.

OLS$^{\text{all}}$ is based on an OLS specification that uses all explanatory variables. POSTLASSO is a re-estimation by OLS-regression of the wage regressions including only the explanatory variables selected by the LASSO-estimator according to the one standard error rule.

Weighted data for 2016 for heads and their spouses who were between 25 and 64 years of age, who earned an hourly wage of at least US$2, and who worked for at least 26 weeks. Non-farming, non-military, non-self-employed wage and salary workers. Excluding all persons with missing values for any of the explanatory variables of the wage regressions. N = 3,390 women and 2,985 men.

Table 4: Smith-Welch Decomposition of the Change in the Gender Wage Gap between 2006 and 2016.

	OLS$^{\text{all}}$	POSTLASSO
Main effect		
Education	−0.0167	−0.0168
Experience	−0.0182	−0.0212
Region	−0.0027	−0.0019
Ethnicity	−0.0008	−0.0008
Unionization	−0.0030	−0.0031
Industry	0.0053	0.0019
Occupation	−0.0065	−0.0088
Health	0.0037	0.0042
Family	−0.0008	−0.0001
Hours housework	−0.0001	−0.0001
Financial Status	0.0061	0.0066
Job characteristics	0.0011	0.0025
Sum main effect	−0.0325	−0.0377
Year interaction effect		
Education	−0.0070	−0.0066
Experience	0.0000	0.0011
Region	0.0003	−0.0007
Ethnicity	−0.0001	0.0001
Unionization	0.0000	0.0000
Industry	−0.0326	0.0195
Occupation	0.0693	0.0431
Health	0.0094	0.0109
Family	0.0011	0.0019
Hours housework	0.0048	0.0058
Financial Status	−0.0011	−0.0006
Job characteristics	−0.0284	−0.0206
Sum year interaction effect	0.0155	0.0538
Gender interaction effect	−0.0184	−0.0114
Gender-year interaction effect	0.0208	−0.0193
Change in gender wage gap	−0.0145	−0.0145

Note: The table shows the components of the Smith-Welch decomposition. The dependent variable is the logarithm of the hourly wage. The components are defined as follows:
Main endowments effect $= ((\bar{X}_{m,2016} - \bar{X}_{f,2016}) - (\bar{X}_{m,2006} - \bar{X}_{f,2006}))\hat{\beta}_{m,2006}$,
year interaction effect $= (\bar{X}_{m,2016} - \bar{X}_{f,2016})(\hat{\beta}_{m,2016} - \hat{\beta}_{m,2006})$,
gender interaction effect $= (\bar{X}_{f,2016} - \bar{X}_{f,2006})(\hat{\beta}_{m,2006} - \hat{\beta}_{f,2006})$,
gender-year interaction effect $= \bar{X}_{f,2016}((\hat{\beta}_{m,2016} - \hat{\beta}_{f,2016}) - (\hat{\beta}_{m,2006} - \hat{\beta}_{f,2006}))$,
change in gender wage gap $= (\bar{y}_{m,2016} - \bar{y}_{f,2016}) - (\bar{y}_{m,2006} - \bar{y}_{f,2006})$,
where $\bar{X}_{g,y}$ is the vector of mean explanatory variables of gender g in year y, $\bar{y}_{g,y}$ is the mean of the dependent variable and $\hat{\beta}_{g,y}$ is the vector of estimated coefficients. The table shows the main endowments effect and the year interaction effect for each variable group.
OLS$^{\text{all}}$ is based on an OLS specification that uses all explanatory variables. POSTLASSO is a re-estimation by OLS-regression of the wage regressions including only the explanatory variables selected by the LASSO-estimator according to the one standard error rule
Weighted data for 2006 and for 2016 for heads and their spouses who were between 25 and 64 years of age, who earned an hourly wage of at least US$2, and who worked for at least 26 weeks. Non-farming, non-military, non-self-employed wage and salary workers. Excluding all persons with missing values for any of the explanatory variables of the wage regressions. N = 2,756 women and 2,451 men in 2006, and 3,390 women and 2,985 men in 2016.

Table 5: Explanatory Variables.

Name	Description
Education	
Advanced degree	1 if the participant holds any degree higher than a bachelor's
Bachelor's degree	1 if the participant has only a bachelor's degree
Foreign education	1 if the participant was educated abroad
No US education	1 if the participant was not educated in the US
Years of schooling	Number of years the participant was schooled
Experience	
Full-time years	Number of years the participant worked full-time
Full-time years squared	Square of full-time years
Part-time years	Number of years the participant worked part-time
Part-time years squared	Square of part-time years
Tenure	Number of weeks the participant has been with their current employer
Tenure squared	Tenure squared
Region	
Metropolitan county	1 if participant lives in metropolitan area as defined by USDA
North-central	1 if participant lives in the north-central US
North-east	1 if participant lives in the north-eastern US
South	1 if participant lives in the southern US
Ethnicity	
Black	1 if participant is Afro-American
Hispanic	1 if participant is Hispanic
Other ethnicity	1 if participant is non-Afro-American, non-Hispanic and non-white
Unionization	
Union member	1 if participant's job is covered by a union contract
Industry	
Communications	
Durables	Durable manufacturing
Finance, real estate	Includes insurance industry
Hotels, restaurants	
Medical	
Mining, construction	
Non-durables	Non-durable manufacturing
Professional services	
Public administration	
Retail, trade	
Social work, recreation	Includes arts
Transportation sector	
Utilities	
Wholesale	
Occupation	
Administration	
Architect, engineer	
Artist, athlete	Includes designers, entertainers and media-area jobs
Builder, cleaner	
Business specialist	
Computer specialist	Includes mathematics specialists

(continues)

Table 5: (continued).

Name	Description
Construction job	Includes extraction and installation jobs
Financial specialist	
Food, personal care	
Health-care support	
Higher education	
Lawyer, physician	Includes judges and dentists
Life, social science	Includes physical science jobs
Nurse, health-care	
Production	
Protective services	
Sales	
Social worker	
Training	Includes non-post-secondary education, legal and library jobs
Transportation	
Health	
Disabled person	1 if participant has a disability that negatively affects their work
Drinks alcohol often	1 if participant drinks alcohol at least several times a week
Health status	1 if participant reports their health status to be at least "very good"
Heavy exerciser	1 if participant does heavy exercise for at least 10 min a week
Light exerciser	1 if participant does light exercise for at least 10 min a week
Mental problems	1 if participant has any diagnosed mental problems
Smoker	1 if participant smokes cigarettes
Family	
Child between 5 and 18	1 if there is a 5 to 18 year old in the family unit
Child born last year	1 if participant or his/her spouse gave birth to a child last year
Child in care center	1 if any of participant's children are enrolled in a childcare-center
Child younger than 5	1 if there is somebody younger than 5 in the family unit
Married	1 if participant is currently married
Number of children	Number of children in the household
Widowed or divorced	1 if participant has ever been widowed, divorced or separated
Hours housework	
Hours of housework	On average per week
Financial Status	
Inheritances and gifts	Value of large gifts or inheritances during the last 2 years
Insured by employer	1 if participant's employer provides health insurance
Job characteristics	
Part-time job	1 if participant works part-time only
Public sector job	1 if participant works for federal, state or local government
Size of employer's firm	Number of people employed by the participant's employer

Table 6: Wage Regressions for 2016, by Gender.

	Women		Men	
	OLS^{all}	POSTLASSO	OLS^{all}	POSTLASSO
(Intercept)	1.7737***	1.864***	1.9669***	2.0182***
Education				
Advanced degree	0.3193***	0.3165***	0.3385***	0.3350***
Bachelor's degree	0.1877***	0.1861***	0.1749***	0.1793***
Foreign education	0.0304	0.0313	0.0466	0.0504
No US education	0.0066	0.0176	0.1016	0.1099
Years of schooling	0.0439***	0.0435***	0.0392***	0.0394***
Experience				
Full-time years	0.0231***	0.0244***	0.0238***	0.0277***
Full-time years squared	−0.0005***	−0.0005***	−0.0005***	−0.0006***
Part-time years	−0.0122**	−0.0051	−0.0010	−0.0031
Part-time years squared	0.0004°	–	−0.0003	–
Tenure	0.0004***	0.0003***	0.0004***	0.0002***
Tenure squared	−0.0001*	–	−0.0001***	–
Region				
Metropolitan county	0.1584***	0.1590***	0.1147***	0.1115***
North-central	−0.0914***	−0.0892***	−0.1304***	−0.1121***
North-east	−0.0016	–	−0.0358	–
South	−0.0740**	−0.0703*	−0.0563*	−0.0366
Ethnicity				
Black	−0.0793**	−0.0817*	−0.1630***	−0.1649***
Hispanic	−0.0016	0.0021	−0.1541***	−0.1484***
Other ethnicity	0.1076*	0.1084	−0.0090	−0.0059
Unionization				
Union member	0.0843**	0.0742	0.1441***	0.1382***
Industry				
Communications	0.1642*	0.0706	0.2125**	0.1410
Durables	0.3475***	0.2251***	0.2094**	0.1200**
Finance, real estate	0.1484***	0.0273	0.3214***	0.2438***
Hotels, restaurants	−0.0370	−0.1374*	0.0341	−0.0432
Medical	0.1205**	0.0348	0.0822	−0.0063
Mining, construction	0.0942	−0.0203	0.1830**	0.1032°
Non-durables	0.1278*	–	0.1569*	–
Professional services	0.1622***	–	0.0673	–
Public administration	0.1119**	–	0.0964	–
Retail, trade	−0.0595	−0.1806***	−0.1672*	−0.2502***
Social work, recreation	−0.0126	−0.1020°	−0.0084	−0.0806
Transportation sector	0.1345*	0.0336	0.2315**	0.1509*
Utilities	0.1801°	0.0871	0.3354***	0.2383*
Wholesale	0.1464*	–	0.0642	–

(continues)

Table 6: (continued).

	Women		Men	
	OLS^{all}	POSTLASSO	OLS^{all}	POSTLASSO
Occupation				
Administration	−0.2900***	−0.2592***	−0.3180***	−0.2925***
Architect, engineer	0.1207	0.1803	0.0038	0.0389
Artist, athlete	−0.2172**	−0.1839	−0.0980	−0.0697
Builder, cleaner	−0.4436***	−0.4263***	−0.6048***	−0.5913***
Business specialist	−0.0381	–	−0.0145	–
Computer specialist	0.1326	0.1509	0.1011*	0.1215
Construction job	−0.1205	−0.0773	−0.2498***	−0.2235***
Financial specialist	−0.1031*	–	−0.0583	–
Food, personal care	−0.2767***	−0.2607***	−0.4178***	−0.3945***
Health-care support	−0.2878***	−0.2850***	−0.7937***	−0.7384***
Higher education	−0.0800	−0.1376	−0.1889	−0.2232
Lawyers, physicians	0.2553***	0.2984***	0.2607***	0.2877**
Life, social science	−0.0356	–	−0.0634	–
Nurses, health-care	0.0294	–	−0.1006	–
Production	−0.4738***	−0.4361***	−0.3499***	−0.3111***
Protective services	−0.1876°	–	−0.1416*	–
Sales	−0.0776*	–	−0.0114	–
Social worker	−0.2307***	−0.2220***	−0.5936***	−0.5698***
Training	−0.3561***	−0.3936***	−0.4393***	−0.4575***
Transportation	−0.2917***	−0.2556***	−0.3026***	−0.2773***
Health				
Disabled person	−0.0826**	−0.0796	−0.0403	−0.0424
Drinks alcohol often	0.0639**	0.0650	0.1259***	0.1269***
Health status	0.0613***	0.0615*	0.0593**	0.0591*
Heavy exerciser	0.0080	0.0135	0.03220	0.0364
Light exerciser	0.0468°	0.0407	0.0200	0.0167
Mental problems	0.0028	0.0039	−0.1688***	−0.1731***
Smoker	−0.0697**	−0.0734°	−0.0747**	−0.0800°
Family				
Child between 5 and 18	−0.0424	−0.0476	0.0579°	0.0701
Child born last year	0.1182*	0.1356	0.0747	0.0607
Child in care center	0.0024	–	0.0259	–
Child younger than 5	0.0185	–	−0.0370	–
Married	0.0195	0.0134	0.1393***	0.1387***
Number of children	0.0395**	0.0454***	−0.0008	−0.0054
Widowed or divorced	0.0046	–	0.0001	–
Hours housework				
Hours of housework	−0.0028**	−0.0030*	−0.0011	−0.0014
Financial status				
Inheritances and gifts	0.0001	0.0001	0.0001	0.0001
Insured by employer	0.2162***	0.2183***	0.2078***	0.2146***

(continues)

Table 6: (continued).

	Women		Men	
	OLS^{all}	POSTLASSO	OLS^{all}	POSTLASSO
Job characteristics				
Part-time job	−0.0382	−0.0373	0.1260**	0.1222
Public sector job	−0.0689**	−0.0956***	−0.0477	−0.0716
Size of employer's firm	0.0001**	0.0001°	0.0001**	0.0001°

Note: The table shows the estimated coefficients for different models for men and women. The dependent variable is the logarithm of the hourly wage.

OLS^{all} is based on an OLS specification that uses all explanatory variables. POSTLASSO is a re-estimation by OLS-regression of the wage regressions including only the explanatory variables selected by the LASSO-estimator according to the one standard error rule.

Weighted data for 2016 for heads and their spouses who were between 25 and 64 years of age, who earned an hourly wage of at least US$2, and who worked for at least 26 weeks. Non-farming, non-military, non-self-employed wage and salary workers. Excluding all persons with missing values for any of the explanatory variables of the wage regressions. $N = 3{,}390$ women and 2,985 men.

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '°' 0.1, significance codes for POSTLASSO estimates calculated by the method proposed by Belloni, Chernozhukov and Kato (2014).