Data and Programs for "The Sorting Effect Method: Discovering Heterogeneous Effects Beyond Their Averages"

Victor Chernozhukov, Iván Fernández-Val, and Ye Luo[1]

This supplement to the paper "The Sorting Effect Method: Discovering Heterogeneous Effects Beyond Their Averages" provides added details related to the data and the programs used in the numerical and empirical examples.

1 Gender Wage Gap Application

1.1 Data

The R data file *wage2015.Rdata* contains the data used in the gender wage gap application. This data file contains 26 demographic and job-relevant variables constructed from the CPS March Supplement 2015. Table 1 includes the list of raw CPS variables extracted from the IPUMS-USA (Ruggles et al., 2017).

Sample selection: we applied the following rules:

1. Drop armed forces and children from the sample. (Drop if *popstat* equals 2 or 3)
2. Drop individuals who have worked 0 hour last year. (Drop if *wkswork1* equal to 0)
3. Keep individuals with age between 25 and 65 years. (Keep if *age* more than 24 and less than 65)
4. Drop individuals with allocated income and work variables. (Drop if *qincwage* or *qwkswork* equal to 1)
5. Keep individuals working for wages or salary. (Drop if *classwly* equals to (10,13,14,99,29,00))
6. Drop part-time workers. (Drop if *uhrsworkly* is less than 36)
7. Drop individuals who worked less than 50 weeks last year. (Drop if *wkswork1* is less than 50)
8. Drop individuals with zero annual wage. (Drop if *incwage* equals 0).
9. Keep white non-hispanic individuals. (Keep if *race* equal to 100 and *hispanic* equal to 0)
10. Drop individuals living in group quarters. (Keep if *qg* equal to 1)
11. Drop individuals with missing information on marital status, education, region, occupation, or industry). (Drop if *marst* equal to 9, *educ* equal to 999, *occly* equal to 0000, or *indly* equal to 0000)
12. Drop individuals in the military, agricultural or private household sectors. (Drop if *occly* equal to 9840, and *indly* equal to 0170, 0180, 0190, 0270, 0280, 0290, 9290, or 9890)
13. Drop individuals with hourly wage rate below $3. (Drop if *wage* less than 3)

[1]We thank Mert Demirer and Vira Semenova for capable research assistance.

Table 1: Downloaded CPS Variables

HWTSUPP	Household weight
GQ	Group Quarters status
HHINTYPE	Type of household
REGION	Region and division
STATEFIP	State (FIPS code)
ASECFLAG	Flag for ASEC
METRO	Metropolitan central city status
CBSASZ	Core-based statistical area size
HHINCOME	Total household income
MONTH	Month
PERNUM	Person number in sample unit
WTSUPP	Supplement Weight
EARNWT	Earnings weight
FAMSIZE	Number of own family members in hh
NCHILD	Number of own children in household
NCHLT5	Number of own children under age 5 in hh
RELATE	Relationship to household head
AGE	Age
SEX	Sex
RACE	Race
MARST	Marital status
POPSTAT	Adult civilian, armed forces, or child
CITIZEN	Citizenship status
NATIVITY	Foreign birthplace or parentage
HISPAN	Hispanic origin
EDUC	Educational attainment recode
EDUC99	Educational attainment, 1990
EMPSTAT	Employment status
SCHLCOLL	School or college attendance
OCC	Occupation
OCCLY	Occupation last year
INDLY	Industry last year
CLASSWLY	Class of worker last year
WKSWORK1	Weeks worked last year
WKSWORK2	Weeks worked last year, intervalled
UHRSWORK	Usual hours worked per week (last yr)
AHRSWORKT	Hours worked last week
WKSUNEM1	Weeks unemployed last year
HOURWAGE	Hourly wage
PENSION	Pension plan at work
UNION	Union membership
FIRMSIZE	Number of employees
FTOTVAL	Total family income
INCTOT	Total personal income
INCWAGE	Wage and salary income
EARNWEEK	Weekly earnings
VETSTAT	Veteran status
HEALTH	Health status
QAGE	Data quality flag for AGE
QMARST	Data quality flag for MARST
QSEX	Data quality flag for SEX
QEDUC	Data quality flag for EDUC
QCLASSWL	Data quality flag for CLASSWLY
QUHRSWORK	Data quality flag for UHRSWORKT
QWKSWORK	Data quality flag for WKSWORK1 and WKSWORK2
QEARNWEE	Data quality flag for EARNWEEK
QINCWAGE	Data quality flag for INCWAGE
MHMARNUM	Number of times married

Variable construction: after selecting the sample, we constructed new variables from the raw CPS variables. Table 2 provides the list of variables and calculation methods. The occupation and industry variables are coded as factors. The occupation factor occ2 has the levels:

1. Management, professional, and related occupations.
2. Service occupations.
3. Sales and office occupations.
4. Natural resources, construction, and maintenance occupations.
5. Production, transportation, and material moving occupations.
6. Armed forces.

The industry factor ind2 has the levels:

1. Agriculture, Forestry, Fishing, and Hunting.
2. Mining, quarrying, and oil and gas extraction.
3. Construction.
4. Manufacturing.
5. Wholesale and retail trade.
6. Transportation and utilities.
7. Information.
8. Financial activities.
9. Professional and business services.
10. Education and health services.
11. Leisure and hospitality.
12. Other services.
13. Public administration.
14. Armed Forces.

Table 2: Constructed Variables in wage2015.Rdata

Characteristics	**Variables**	**Calculation**	**Type**	**CPS Variable**
Sampling weight	weight	CPS person level weight	Continuous	wtsupp
Log Hourly Wage	wage	$\frac{annualincome}{(weeksworked * hoursworked)}$	Continuous	incwage, uhrsworklyt, wkswork1
	lwage		Continuous	
Gender	female	1 if female	Indicator	sex
Marital status	married	married	Indicator	marst
	widowed	widowed	Indicator	
	separated	separated	Indicator	
	divorced	divorced	Indicator	
	nevermarried	never married	Indicator	
Education	lhs	less than high school (years of educ $<$ 12)	Indicator	educ
	hsg	high school graduate (years of educ $=$ 12)	Indicator	
	sc	some college (13$\leq$ years of educ $\leq$ 15)	Indicator	
	cg	college graduate (16$\leq$ years of educ $\leq$ 17)	Indicator	
	ad	advanced degree (educ $\geq$ 18)	Indicator	
Region	mw	1 if living in midwest	Indicator	region
	so	1 if living in south	Indicator	
	we	1 if living in west	Indicator	
	ne	1 if living in northeast	Indicator	
Experience	exp1	max(age - years of educ -7, 0)	Continuous	age, educ
	exp2	$exp1^2/10$	Continuous	
	exp3	$exp1^3/100$	Continuous	
	exp4	$exp1^4/1000$	Continuous	
Occupation level[a]	occ	Occupation Categories in CSP (456 categories)	Categorical	occly
	occ2	Aggregated occupation (6 categories, 5 categories without armed forces)	Categorical	
Industry level[b]	ind	Industry Categories in CSP (257 categories)	Categorical	indly
	indg2	Aggregated industry (14 categories, 12 categories without agriculture and military)	Categorical	

[a] Since the original CPS occupation variable(occ) includes too many categories a second variable named occ2 is constructed by aggregating occ according to aggregation in this webpage.

[b] Similar to occupation, industry categories are aggregated based on categorization in this webpage.

1.2 Programs

We use four `R` command files in the empirical application and associated numerical simulation (R Core Team, 2018):

1. The file *gender-gap.R* contains the main commands to generate most of the results of the empirical application including Figures 1–4 and Tables 1–3.
2. The file *gender-gap-sub-ca.R* contains the commands to generate Figure 5.
3. The file *rq-nonstop.R* contains auxiliary functions used by *gender-gap.R* including a modification of the R command `rq` that does not stop when the design matrix is singular and works for sparse design matrices.
4. The file *gender-gap-mc.R* contains the commands to generate the simulation results of Table 3 in the Supplementary Material.

For replication purposes, note that the programs can take several hours running and employ parallel computing with multiple processors. All the command files assume that the data sets are located in the current folder and generate output files in this folder. We use the packages `boot` (Canty and Ripley, 2017), `ggplot2` (Wickham, 2009), `Hmisc` (Harrell, 2018), `quantreg` (Koenker, 2018), and `xtable` (Dahl, 2016), which need to be installed.

2 Mortgage Application

2.1 Data

The Stata data file *mortgage.dta* contains the data on mortgage applications in Boston from 1990 (Munnell et al., 1996). We obtained the data from the companion website of Stock and Watson (2011). The file contains the following variables:

- *deny*: indicator for mortgage application denied.
- *p_irat*: monthly debt to income ratio.
- *black*: indicator for black applicant.
- *hse_inc*: monthly housing expenses to income ratio.
- *loan_val*: loan to assessed property value ratio (not used in analysis).
- *ccred*: consumer credit score with 6 categories (1 if no "slow" payments or delinquencies, 2 if one or two "slow" payments or delinquencies, 3 if more than two "slow" payments or delinquencies, 4 if insufficient credit history for determination, 5 if delinquent credit history with payments 60 days overdue, and 6 if delinquent credit history with payments 90 days overdue).
- *mcred*: mortgage credit score with 4 categories (1 if no late mortgage payments, 2 if no mortgage payment history, 3 if one or two late mortgage payments, and 4 if more than two late mortgage payments).
- *pubrec*: indicator for any public record of credit problems (bankruptcy, charge-offs, collection actions).
- *denpmi*: indicator for applicant applied for mortgage insurance and was denied.
- *selfemp*: indicator for self-employed applicant.

- *single*: indicator for single applicant.
- *hischl* indicator for high school graduated applicant.
- *probunmp*: 1989 Massachusetts unemployment rate in the applicant's industry (not used in analysis).
- *condo*: indicator for unit is a condominium (not used in analysis).
- *ltv_med*: indicator for medium loan to property value ratio $[.80, .95]$.
- *ltv_high*: indicator for high loan to property value ratio $> .95$.

2.2 Programs

We use the `R` command file *mortgage.R* to generate all the results including Figures 4 and Tables 4–5 of the Supplementary Material (R Core Team, 2018). For replication purposes, note that this program employes parallel computing with multiple processors, assumes that the data set is located in the current folder, and generates output files in this folder. We use the packages `boot` (Canty and Ripley, 2017), `foreign` (R Core Team, 2017), and `xtable` (Dahl, 2016), which need to be installed.

References

[1] Canty, A., and B. Ripley (2017). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-20.

[2] Dahl, D. B. (2016): *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-2. URL `https://CRAN.R-project.org/package=xtable`.

[3] Harrell Jr, F. E., with contributions from C. Dupont and many others, (2018). *Hmisc: Harrell Miscellaneous.* R package version 4.1-1. URL `https://CRAN.R-project.org/package=Hmisc`.

[4] Koenker, R. (2018). *quantreg: Quantile Regression*. R package version 5.35. URL `https://CRAN.R-project.org/package=quantreg`.

[5] Munnell, A. H., Tootell, G. M. B., Browne, L. E., and J. McEneaney (1996). "Mortgage Lending in Boston: Interpreting HMDA Data," *The American Economic Review*, Vol. 86, No. 1, pp. 25-53.

[6] R Core Team (2017). *foreign: Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ...* R package version 0.8-69. URL `https://CRAN.R-project.org/package=foreign`.

[7] R Core Team (2018): *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`.

[8] Ruggles, S., Genadek, K., Goeken, R., Grover, J., and M. Sobek (2017). *Integrated Public Use Microdata Series: Version 7.0 [dataset]*. Minneapolis: University of Minnesota. URL `https://doi.org/10.18128/D010.V7.0`.

[9] Stock, J., and M. Watson (2011). *Introduction to Econometrics* (3rd edition). Addison Wesley Longman.

[10] Wickham, H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.