Homework #3

Professor: Pat Kline

Students: Christina Brown, Sam Leone, Peter McCrory, Preston Mui

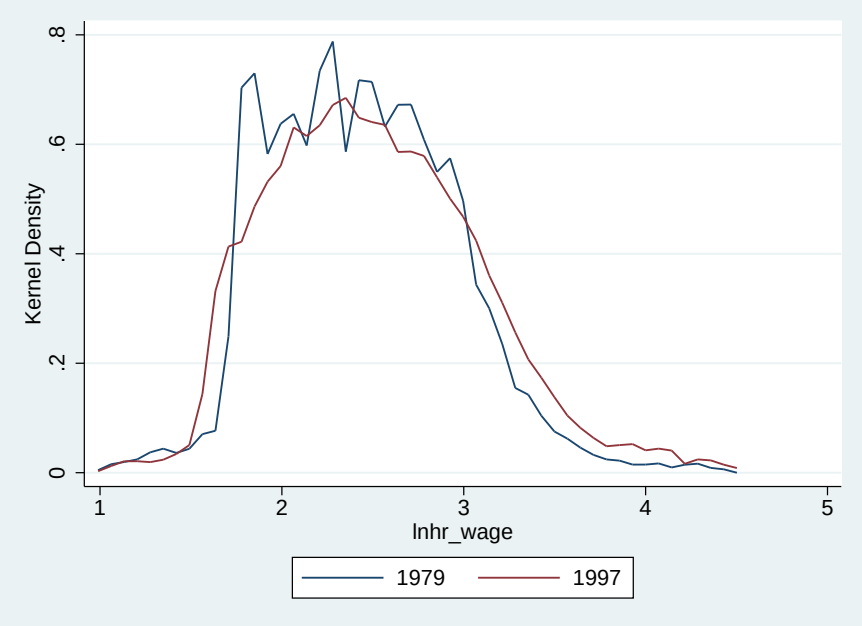
Reweighting

Subproblem A

First, the authors drop a number of demographic groups from their analysis: those self-employed, those under 18 or over 64, and those with more than 38 years of work experience.[1] To compare these groups to prime-working-age employees would be to compare apples and oranges. Second, they top code the wage. This prevents outliers from skewing the analysis. Third, they round school years to whole numbers. This makes the analysis easier in terms of both implementation and interpretation.[2]

Subproblem B

Figure 1: Kernel Density for Full Sample



Subproblem C

[1] The authors merely *note* these data-build decisions in their cleaning .do files, they do not *implement* them there.

[2] We thank Isabelle Cohen for helpful discussions on this subproblem.

The 1997 wage distribution has more mass at the tails than does the 1979 wage distribution. This is in keeping with the authors' result that wage inequality has increased over the past two decades. The 1997 wage distribution also appears to be skewed right. This is in keeping with the authors' result that increasing wage inequality is driven more by the 90/50 gap than by 50/10 gap.

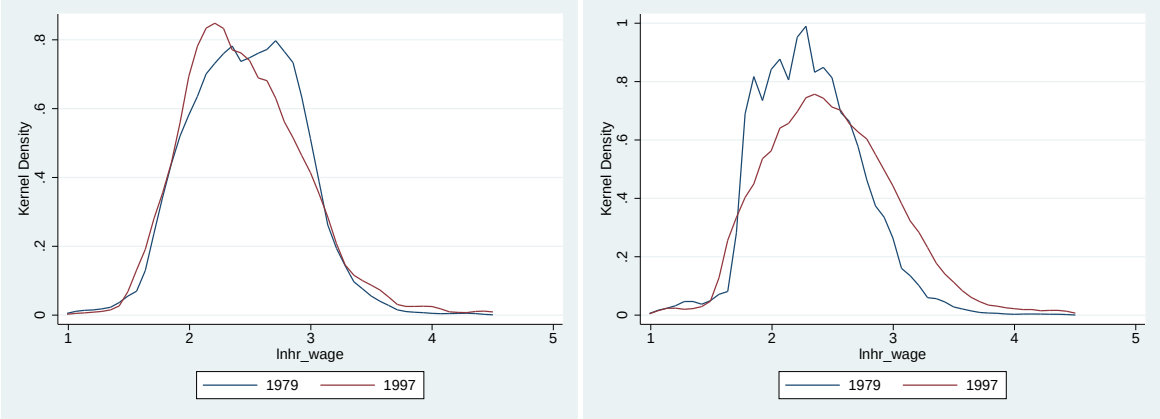
In addition, there appears to be bunching at the low-end of the 1979 distribution. This is in keeping with the hypothesis - which the authors comment on but do not argue for - that the minimum wage has been an important determinant of the U.S. wage distribution over the past decades.

The key now is to disentangle the degree to which these changes come from (a) the changing demographic composition of the labor force (what you might call "selection") and (b) more fundamental labor-market changes happening within demographic groups (what you might call "treatment"). Read on!

Subproblem D

Figure 2: Kernel Density for Selected Subsamples

(a) Kernel Density for Black Men Aged 25-50 (b) Kernel Density for White Women Aged 25-50



There were more women and minorities (e.g. black and Latinos) in the labor force in 1997 than in 1979. Could this demographic shift alone explain the changes in the overall wage distribution? Probably not: When we plot the wage distributions specifically for young black men and young black women, we still changes between 1979 and 1997. In other words, even conditioning on these particular demographic characteristics, we still see that there are more fundamental economic forces at work.

So: What changes do we see exactly? For black men, the mean wage is actually lower in 1997 than it was in 1979. This could be because of increased racial discrimination, changes in government welfare programs, or changes in the skill level of the average African-American worker (e.g. if a higher portion of African Americans were working in 1997 than in 1979, the

average worker might be less skilled). For white women, the mean wage is higher in 1997 than it was in 1979, and overall the distribution is more symmetric. This is unsurprising, since women were more likely to hold high-skill/high-wage positions in the late 1990s than in the late 1970s.

Compared to the overall distribution, white woman have a similar average wage, but black men have a lower average wage.

Subproblem E

See the attached Stata code.

Subproblems F & G

Figure 3: Actual 1997 vs. Counterfactual 1997

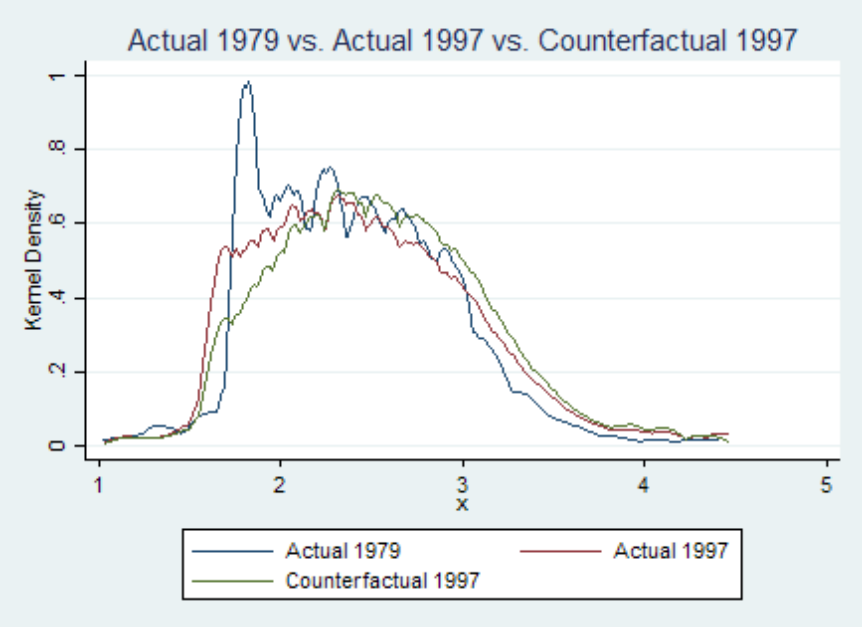
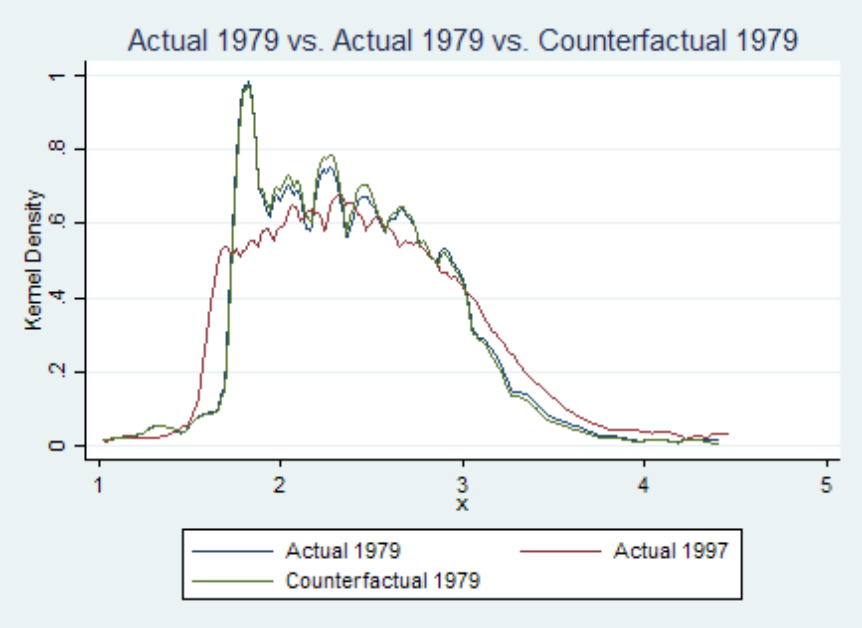


Figure 4: Actual 1979 vs. Counterfactual 1979



Subproblem H

Above, when we estimated separate wage distributions for young black men and young white women, we were going some of the way toward disentangling demographic effects from more fundamental economic effects. But we can still go further. One strategy is to *reweight* the 1997 sample so that it has the same demographic shares as the 1997 sample. We would then have a new *counterfactual* 1997 wage distribution. And we we moved from one distribution to another, we would then be able to isolate one of our effects of interest:

$$\text{Actual 1979} \underbrace{\Rightarrow}_{Fundamentals} \text{Reweighted 1997} \underbrace{\Rightarrow}_{Demographics} \text{Actual 1997}$$

How can we produce the weights exactly?[3] Let w be the wage schedule, let z be demographic characteristics, and let t be the year. Let the conditional pdf of w be $f(w|t_w = t, t_z = t)$. Notice that t is indexed by w and by z, which indicates that we can, counterfactually, mix wage schedules and demographic characteristics from different years. For example, conditioning $t_w = 1979, t_z = 1997$ means conditioning on the wage schedule that prevailed in 1979 but the demographic characteristics that prevailed in 1997.

Consider the baseline case, where we reweight the 1997 wage distribution so that it has the 1979 demographic characteristics.

[3]We thank Christopher Campos for help with this mathematical explication.

$$
\begin{aligned}
f(w|t_w = 1997, t_z = 1979) &= \int_z f(w|z, t_w = 1997) dF(z|t_z = 1979) \\
&= \int_z f(w|z, t_w = 1997) \frac{dF(z|t_z = 1997)}{dF(z|t_z = 1997)} dF(z|t_z = 1979) \\
&= \int_z f(w|z, t_w = 1997) \Psi_z(z) dF(z|t_z = 1997)
\end{aligned}
$$

where

$$
\Psi_z(z) = \frac{dF(z|t_z = 1979)}{dF(z|t_z = 1997)}
$$

By Bayes' Rule

$$
\begin{aligned}
\Psi_z(z) &= \frac{dF(z|t_z = 1979)}{dF(z|t_z = 1997)} \\
&= \frac{P(t_z = 1979|z)F(z)}{P(t_z = 1979)} \frac{P(t_z = 1997)}{P(t_z = 1997|z)F(z)} \\
&= \frac{P(t_z = 1979|z)}{P(t_z = 1997|z)} \frac{P(t_z = 1997)}{P(t_z = 1979)}
\end{aligned}
$$

where the conditional probabilities are are the propensity scores and the unconditional probabilities are sample frequencies.

In this case, $\Psi_z(z) dF(z|t_z = 1997)$ yields the new weights. Boom! (See the attached Stata code for the actual logistic regression specification used to produce the propensity scores.)

Figure 3 plots the three distributions on top of each other. We can apply this logic to our problem at hand: Observe Figure 3 above. Start with the actual 1979 distribution (blue line) and move to the reweighted 1997 distribution (green line). There is a big change. Now move to the actual 1997 distribution (red line). There is a small change. This suggests that the evolution of the wage distribution is not merely due to demographic characteristics, but rather results from other economic forces affecting the wage schedule.

The same logic applies going backward in time instead of forward in time. Observe Figure 4, which instead plots the counterfactual 1979 distribution using 1997 weights. There is a big difference between the actual 1997 distribution (red line) and the reweighted 1979 distribution (green line), but not between the reweighted 1979 distribution and the actual 1979 distribution (blue line). This serves as a robustness check, and provides further evidence that more fundamental economic forces are at play.

Subproblem I

Table 1: Actual 1979

variable	mean	sd	p10	p25	p50	p75	p90
age	34.99355	12.77136	20	24	32	45	55

Table 2: Counterfactual 1997 Reweighted with 1979 Demographic Characteristics

variable	mean	sd	p10	p25	p50	p75	p90
age	38.94351	9.939499	26	32	39	46	52

Subproblem J

Table 3: Actual 1979 (African Americans)

variable	mean	sd	p10	p25	p50	p75	p90
age	35.52812	12.33177	21	25	33	45	54

Table 4: Counterfactual 1997 Reweighted with 1979 Demographic Characteristics (African Americans)

variable	mean	sd	p10	p25	p50	p75	p90
age	38.19553	9.680713	25	31	38	45	51

Subproblem K & L

Observe from Tables 1-4 that, at least for age, the 1979 and the reweighted 1997 distributions do not have the same moments/quantiles. This is problematic, since the weights are supposed to have been calculated in order to achieve similar demographic characteristics between the two distributions. This suggests that our propensity scores might be misspecified. Below we try a few different specifications, namely (a) adding a cubic term in age, and (b) eliminating the quadratic term in age. Neither makes a difference.

Table 5: Actual 1979: Logit with Age, Age^2, Age^3

variable	mean	sd	p10	p25	p50	p75	p90
age	34.99355	12.77136	20	24	32	45	55

Table 6: Counterfactual 1997: Logit with Age, Age^2, Age^3

variable	mean	sd	p10	p25	p50	p75	p90
age	39.18865	9.905478	26	32	39	46	52

Table 7: Actual 1979: Logit with Age, Age^2, Age^3 (African Americans)

variable	mean	sd	p10	p25	p50	p75	p90
age	35.52812	12.33177	21	25	33	45	54

Table 8: Counterfactual 1997: Logit with Age, Age^2, Age^3 (African Americans)

variable	mean	sd	p10	p25	p50	p75	p90
age	38.46548	9.677537	25	31	38	46	51

Table 9: Actual 1979: Logit with Age Only

variable	mean	sd	p10	p25	p50	p75	p90
age	34.99355	12.77136	20	24	32	45	55

Table 10: Counterfactual 1997: Logit with Age Only

variable	mean	sd	p10	p25	p50	p75	p90
age	39.35282	11.33751	24	31	39	48	55

Table 11: Actual 1979: Logit with Age Only (African Americans)

variable	mean	sd	p10	p25	p50	p75	p90
age	35.52812	12.33177	21	25	33	45	54

Table 12: Counterfactual 1997: Logit with Age Only (African Americans)

variable	mean	sd	p10	p25	p50	p75	p90
age	38.41906	11.01582	24	30	38	47	54

Matching

(a) To run OLS on $Y_{i2} - Y_{i1} = \gamma_2 - \gamma_1 + \beta D_{i2} + \epsilon_{i2} - \epsilon_{i1}$, the identifying restriction is that $E\left[D_{i2}(\epsilon_{i2} - \epsilon_{i1})\right] = 0$; that is, program participation is uncorrelated with the change in error terms $\epsilon_{i2} - \epsilon_{i1}$.

(b) $Y_{i2} = Y_{i1} + (\gamma_2 - \gamma_1) + \beta D_{i2} + \epsilon_{i2} - \epsilon_{i1}$

(c) Taking expectations,

$$E\left[Y_{i2}|Y_{i1}, D_{i2} = 1\right] = Y_{i1} + (\gamma_2 - \gamma_1) + \beta + E[\epsilon_{i2} - \epsilon_{i1}|Y_{i1}, D_{i1} = 1]$$
$$E\left[Y_{i2}|Y_{i1}, D_{i2} = 0\right] = Y_{i1} + (\gamma_2 - \gamma_1) + E[\epsilon_{i2} - \epsilon_{i1}|Y_{i1}, D_{i1} = 0]$$

the difference of these two identifies β if

$$E[\epsilon_{i2} - \epsilon_{i1}|Y_{i1}, D_{i2} = 1] - E[\epsilon_{i2} - \epsilon_{i1}|Y_{i1}, D_{i2} = 0] = 0$$
$$E\left[E[\epsilon_{i2} - \epsilon_{i1}|D_{i2} = 1] - E[\epsilon_{i2} - \epsilon_{i1}|D_{i2} = 0]\middle| Y_{i1}\right] = 0$$
$$E\left[D_{i2}(\epsilon_{i2} - \epsilon_{i1})\middle| Y_{i1}\right] = 0$$

(d) This assumption is stricter than that of the assumption in part (a); it is the same restriction in part (a), but conditional on Y_{i1}.

(e) The Smith and Todd (2005) matched differences-in-differences estimator requires that selection is independent on the differences in untreated outcomes between time periods, conditional on the propensity. Formally,

$$E\left[D_{i2}(\gamma_2 + \epsilon_{i2} - \gamma_1 - \epsilon_{i1})\middle| P_i\right] = E[D_{i2}]E\left[(\gamma_2 + \epsilon_{i2} - \gamma_1 - \epsilon_{i1})\middle| P_i\right]$$
$$\implies E\left[D_{i2}(\epsilon_{i2} - \epsilon_{i1})\middle| P_i\right] = 0$$

which is the same condition from above, except instead of conditioning on Y_{i1} we are conditioning on P_i, the conditions are not really "weaker" than the ones above.

2SLS

(a) The moment conditions defining the 2SLS estimator are

$$E\left[Z_i\epsilon_i\right] = 0$$
$$E\left[Z_i^2\epsilon_i\right] = 0$$

they imply each other, since functions of uncorrelated variables are uncorrelated with each other.

(b) The moment conditions defining the control function estimator are

$$\begin{aligned} E\left[Z_i \epsilon_i\right] &= 0 \\ E\left[Z_i v_i\right] &= 0 \\ E\left[\epsilon_i | v_i\right] &= \rho \frac{\sigma_\epsilon}{\sigma_v} v_i \end{aligned}$$

$$E\left[(Y_i - \beta_0 - \beta_1 X_i - \beta_2 X_i^2 - \rho \frac{\sigma_\epsilon}{\sigma_v} v_i) X_i\right] = 0$$

$$E\left[(Y_i - \beta_0 - \beta_1 X_i - \beta_2 X_i^2 - \rho \frac{\sigma_\epsilon}{\sigma_v} v_i) X_i^2\right] = 0$$

$$E\left[(Y_i - \beta_0 - \beta_1 X_i - \beta_2 X_i^2 - \rho \frac{\sigma_\epsilon}{\sigma_v} v_i) v_i\right] = 0$$

(c) The control function estimator relies on stronger conditions than the 2SLS estimator. The only condition in the 2SLS estimator is that the instrument, Z, is uncorrelated with the error term ϵ. However, the control function estimator relies on specifying the exogenous relationship between X and Z, in particular that the conditional expection of X given Z is linear.

(d) The control function would work better when Z is weak, because it brings to bear more moment restrictions than the 2SLS estimator, whereas the 2SLS estimator only relies on the exogeneity of Z.

(e) The table below contains the estimates for β_1 and β_2 under the 2SLS and control function methods.

(i) Overall, for all values of γ, the control function estimate appears to be more efficient (the spread of the estimates is much lower), because the control function's "first-stage" is correctly specified, and the estimation involves more moment restrictions.

(ii) As γ_1 shrinks, the mean 2sls estimate falls (away from the true value of 1.0), although the median appears to remain unbiased. The interquartile range of both estimators increases, although the IQR for the 2sls estimate increases much faster than that of the control function. The control function estimates appear to remain both median and mean-unbiased.

(iii) If one wanted to test the hypothesis that $\rho \frac{\sigma_\epsilon}{\sigma_v} = 0$, one would need an additional term in the variance of the coefficient on v_i. In particular, by using the two-stage estimate formulas for the asymptotic variance of the estimate of $\rho \frac{\sigma_\epsilon}{\sigma_v}$, one would see that in the "sandwich" estimator, one would replace the term

$$E[s_i(\theta_0, \gamma^*) s_i(\theta_0, \gamma^*)']$$

where $\theta_0 = (\beta_0, \beta_1, \beta_2, \rho \frac{\sigma_\epsilon}{\sigma_v})$ and $\gamma^* = (\gamma_0, \gamma_1)$, and s_i being the score function (gradient of the squared residual with respect to θ_0), with this term

$$\begin{gathered} E[g_i(\theta_0, \gamma^*) g_i(\theta_0, \gamma^*)'] \\ \text{s. t. } g_i(\theta_0, \gamma^*) \equiv s_i(\theta_0, \gamma^*) + E\left[\nabla_\gamma s_i(\theta_0, \gamma^*)\right] r_i(\theta_0, \gamma^*) \end{gathered}$$

where r_i is the asymptotic variance of $\sqrt{N}(\hat{\gamma} - \gamma^*)$ and can be obtained from the "first-stage" regression. To construct an estimate of the standard error, one can simply plug in the population analogues into this sandwich estimator and run a t-test with the modified standard error.

Statistic	β_1^{2sls}	β_2^{2sls}	β_1^{cf}	β_2^{cf}
$\gamma = 0.3$				
mean	1.003234	1.076318	.9938653	1.000297
sd	.3412886	3.060352	.1081107	.0204986
p50	.9992083	.9904965	.9961268	1.000156
iqr	.1585197	.3704991	.1449331	.0277689
$\gamma = 0.2$				
mean	.7418318	2.253272	.9936412	.999546
sd	8.550667	47.84852	.1688314	.0207697
p50	.9829823	1.005342	.9991214	.9994298
iqr	.315193	.7726767	.2297587	.0277564
$\gamma = 0.1$				
mean	.8100506	.839673	.9693889	.9988507
sd	7.576605	17.05286	.5007757	.0224328
p50	1.029302	1.044092	1.005126	.9992241
iqr	.7045021	1.448808	.4600121	.0299107

Bootstrap OLS

(a) See the regression results in table below. The clustered standard error on D is 0.0925.

(b) The p value for the null hypothesis that the coefficient on D equals zero is 0.008.

(c) See column 2 of the table below. Using a block bootstrap our new standard error on D is 0.0965 and the p value is 0.003.

(d) The p value of the symmetric percentile t-test is 0.005.

(e) Using the Wild bootstrap with Rademacher weights, the p value of the null hypothesis is 0.009.

(f) Since we have only have 20 clusters, we need to use a bootstrap to correct the standard errors. When the limiting distribution is normal a regular bootstrap will give us correct standard errors (our answer in b). The refinement we did in part c, will give us a slightly better estimate. However, the most correct standard errors when we do not have a normal limiting distribution or are concerned about heteroskedasticity is the wild bootstrap imposing the null used in CGM.

Bootstrap Probit See attached Stata code.

Table 13: Bootstrap

	(1) Standard Reg	(2) Bootstrap
D	0.276***	0.276***
	(0.09248)	(0.09651)
X	0.147	0.147
	(0.11505)	(0.12219)
X2	0.216***	0.216***
	(0.02678)	(0.02602)
Constant	0.392***	0.392***
	(0.09661)	(0.10193)
R^2	0.381	0.381
Observations	200	200

Standard errors in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Table 14: Comparison of p-Values on Null Hypothesis of $D = 0$ from Various Tests

	Cluster Robust Wald Test	**Symmetric Bootstarp Percentile t-test**	**Cluster Robust Score Test**	**Clustered Score Bootstrap**
p-value:	0.005	0.003	0.005	0.018