Analysis Report

Xiaoyu Ouyang

March 23, 2023

Summary

This report examines the wage differential between workers employed by local governments and those in the private sector. Using data from the CPS-MORG from the year 1996, as well as data on housing supply and land availability for real estate development, we conduct exploratory data analysis on variables of interest for workers in the private, local government, state government, and federal government sectors. Our findings suggest that workers in the federal government earn significantly higher weekly wages than those in the other three categories and that men outnumber women in all four categories.

We then merge the data using the metropolitan area FIPS code column and conduct regressions with three sets of explanatory variables and the response variable as the logarithm of workers' weekly earnings. In the case where the only explanatory variable is whether or not the worker was employed by the local government, we find that local government workers earn 10.55% more per week than those in the private sector. However, after controlling for demographic variables, the wage gap decreases to 1.30%. We then introduce an additional variable for land availability and an interaction term to analyze how the wage gap varied across metropolitan areas. Our regression results show that being employed by the local government was associated with only a 0.883% increase in weekly earnings and the wage gap further decreases. These findings suggest that demographic attributes and land availability are relevant factors in explaining the wage differential between local government and private sector workers.

Furthermore, we conduct separate analyses for workers with at least a four-year college degree and those with less education. Our regression results indicate that highly educated workers earn more in the private sector than in local government, while less educated workers earn more in local government than in the private sector. However, after adding more controls, the wage difference becomes negligible in both cases. Overall, our findings suggest that being employed by the local government or not is a statistically significant factor in determining workers' earnings, especially after controlling for other factors.

Part 1 & 2

Given the CPS-MORG dataset from 1996, we concentrate on workers ages 25 to 55 who reported working at least 35 hours last week and not being self-employed. By conducting exploratory data analysis, we obtain summary statistics of variables including weekly earnings, years of schooling, gender, and age for workers in the private sector, local government, state government, and federal government separately.

Note that a direct mapping from "grade92" (representing the highest grade completed) to "yrsch" (representing years of schooling) does not exist in the data. To handle cases where a value of "grade92" corresponds to multiple possible years of schooling, a random value within the range is assigned to "yrsch". For example, if "grade92" equals 32 (indicating 1-4th grade), a random integer within the range [1, 4] is assigned to "yrsch". This approach is preferred over other methods, such as using the range average, to avoid underestimating the standard deviation of the population. The pairing of "grade92" and "yrsch" is shown below.

Grade Description	grade92	yrsch
Less than 1st grade	31	0
1st - 4th grade	32	[1,4]
5th or 6th	33	[5,6]
7th or 8th	34	[7,8]
9th	35	9
10th	36	10
11th	37	11
12th grade NO DIPLOMA	38	12
High school graduate, diploma or GED	39	12
Some college but no degree	40	[13,14]
Associate degree -- occupational/vocational	41	14
Associate degree -- academic program	42	14
Bachelor's degree (e.g. BA,AB,BS)	43	16
Master's degree (e.g. MA,MS,MEng,Med,MSW,MBA)	44	[17,18]
Professional school deg. (e.g. MD,DDS,DVM,LLB,JD)	45	19
Doctorate degree (e.g. PhD, EdD)	46	[21,23]

The tables of summary statistics are shown below. It can be easily observed that workers in the federal government receive much higher weekly earnings on average than in the other three parties, and there are more male workers than female workers in all sectors.

Summary Statistics Table for Private Sector

	mean	std	min	max
earnwke	605.71	373.42	0.00	1923.00
yrsch	13.41	2.61	0.00	23.00
sex	1.42	0.49	1.00	2.00
age	38.47	8.32	25.00	55.00

Summary Statistics Table for Local Government

	mean	std	min	max
earnwke	633.32	330.24	0.00	1923.00
yrsch	14.66	2.43	0.00	23.00
sex	1.54	0.50	1.00	2.00
age	40.85	8.30	25.00	55.00

Summary Statistics Table for State Government

	mean	std	min	max
earnwke	619.84	321.86	0.00	1923.00
yrsch	15.06	2.85	1.00	23.00
sex	1.52	0.50	1.00	2.00
age	40.79	8.16	25.00	55.00

Summary Statistics Table for Federal Government

	mean	std	min	max
earnwke	725.59	349.45	0.00	1923.00
yrsch	14.28	2.27	3.00	23.00
sex	1.43	0.50	1.00	2.00
age	41.55	7.80	25.00	55.00

Part 3

To merge data on the amount of land unavailable for real-estate development to CPS-MORG data, we use the column "msafips" (representing metropolitan FIPS code) as the key to merge two data frames at the intersection.

Part 4

To analyze the wage gap between local government workers and private sector workers, we run a regression with the formula:

$$log(earnwke) = \beta_0 + \beta_1(I_{lc}) + \varepsilon$$

where *earnwke* represents the weekly earnings and I_{lc} is the binary variable of whether the worker is employed by the local government or not. The regression result is as follows.

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.2756	0.003	2339.262	0.000	6.270	6.281
localgov	0.1003	0.009	11.166	0.000	0.083	0.118

The regression table indicates that "localgov" is a statistically significant variable with a negligible p-value. The estimated coefficient of 0.1003 suggests that being employed by the local government is associated with a 10.55% increase in weekly earnings, as compared to those working in the private sector. Therefore, it can be concluded that public sector workers earn more.

Part 5

To add controls for workers' demographics, we include binary variables including whether the worker is male, whether the worker is black, whether the worker is Hispanic as well as numeric variables including the worker's age, the square of the worker's age, and years of education of the worker as explanatory variables in the regression with the formula:

$$log(earnwke) = \beta_0 + \beta_1 (I_{lc}) + \beta_2 (I_{Male}) + \beta_3 (I_{Black}) + \beta_4 (I_{Hispanic}) + \beta_5 (age) + \beta_6 (age^2) + \beta_7 (yrsch) + \varepsilon$$

where I_{Male}, I_{Black}, $I_{Hispanic}$ are binary variables defined similarly as I_{lc} and $yrsch$ is years of schooling defined above. The regression result is as follows.

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.4779	0.049	70.728	0.000	3.381	3.574
localgov	0.0129	0.008	1.684	0.092	-0.002	0.028
male	0.2877	0.004	66.042	0.000	0.279	0.296
black	-0.2085	0.007	-30.679	0.000	-0.222	-0.195
hispanic	-0.1325	0.008	-17.422	0.000	-0.147	-0.118
age	0.0655	0.003	26.197	0.000	0.061	0.070
age_squared	-0.0007	3.17e-05	-22.185	0.000	-0.001	-0.001
yrsch	0.0914	0.001	109.101	0.000	0.090	0.093

Based on the regression results table above, the p-value of "localgov" is not statistically significant at the 5% level, with a value of 0.092. This indicates that there is insufficient evidence to reject the null hypothesis that the coefficient of "localgov" equals zero. However, when additional demographic variables are included in the model, the coefficient of "localgov" decreases from 0.1003 to 0.0129, suggesting that being employed by the local government is associated

with a minimal 1.30% increase in weekly earnings. Therefore, it can be concluded that the wage gap between local government and private sector workers decreases massively after controlling for demographic factors.

Part 6

Construct variable for land availability

Note that the variable for the amount of land available is not given in the MSA data and the definition of the variable "unaval" is ambiguous. Therefore, we make certain assumptions about the variable "unaval" and "aval" to conduct further analysis. Assume the "unavailability" of a metropolitan region is defined as the area of unavailable land for real-estate development divided by the entire area of that region, which is a percentage from 0 to 1:

$$unavailability = \frac{area\ of\ unavailable\ land}{entire\ area}$$

Similarly, "availability" is defined below and equals to 1-unavailability:

$$availability = \frac{area\ of\ available\ land}{entire\ area} = 1 - unavailability$$

Suppose we have a data set of size n of unavailability $u_1, u_2, \ldots, u_n$, then the "unaval" variable $\widetilde{u_i}$ as a z-scored measure of land unavailability for $i = 1, 2, \ldots,\ n$ is computed as

$$\widetilde{u_i} = \frac{u_i - \bar{u}}{S_u}$$

where $\bar{u}$ is the mean and S_u is the standard deviation as

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$$

$$S_u = \sqrt{\frac{\sum_i\ (u_i - \bar{u})^2}{n-1}}$$

We then construct a variable "aval" as the z-scored measure of availability with symbols $\widetilde{v_i}$ for $i = 1, 2, \ldots n$, which is defined as

$$\widetilde{v_i} = \frac{v_i - \bar{v}}{S_v} = \frac{(1-u_i) - (1-\bar{u})}{S_u} = -\frac{u_i - \bar{u}}{S_u} = \ -\widetilde{u_i}$$

We use the following relationships in the equation above:

$$\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i = \frac{1}{n}\sum_{i=1}^{n}(1-u_i) = 1 - \bar{u}$$

$$S_v = \sqrt{\frac{\sum_i\ (v_i - \bar{v})^2}{n-1}} = \ \sqrt{\frac{\sum_i\ (u_i - \bar{u})^2}{n-1}} = \ S_u$$

Therefore, we obtain the z-scored measure of land availability ("aval") as the negative of "unaval" and add this variable into the data frame.

Run regressions

To analyze how the local government worker – private sector wage gap differs across areas based on land availability, we run the following regression with two additional variables "aval" and "interaction" (representing the interaction between land availability and whether employed by local government, calculated as the product of "aval" and "localgov"):

$$log(earnwke) = \beta_0 + \beta_1 (I_{lc}) + \beta_2 (I_{Male}) + \beta_3 (I_{Black}) + \beta_4 (I_{Hispanic}) + \beta_5 (age) + \beta_6 (age^2) + \beta_7 (yrsch) + \beta_8 (aval) + \beta_9 (interaction) + \varepsilon$$

Note that using clustered standard errors is necessary because data points from the same metropolitan area are not independent. This correlation violates the regression assumption and brings underestimated standard errors. And the regression result with clustered standard errors is as follows.

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.4798	0.057	61.582	0.000	3.368	3.591
localgov	0.0088	0.011	0.803	0.423	-0.013	0.030
male	0.2878	0.009	32.741	0.000	0.270	0.305
black	-0.2085	0.009	-22.106	0.000	-0.227	-0.190
hispanic	-0.1362	0.015	-9.173	0.000	-0.166	-0.107
age	0.0655	0.003	24.026	0.000	0.060	0.071
age_squared	-0.0007	3.43e-05	-20.473	0.000	-0.001	-0.001
yrsch	0.0914	0.002	46.649	0.000	0.088	0.095
aval	-0.0067	0.008	-0.857	0.393	-0.022	0.009
interaction	-0.0294	0.011	-2.600	0.010	-0.052	-0.007

Based on the regression table above, the p-value of "localgov" is not statistically significant with a value of 0.423. This indicates that there is insufficient evidence to reject the null hypothesis. However, when additional variables "aval" and "interaction" are included, the coefficient of "localgov" decreases further to 0.0088. This suggests that being employed by the local government is associated with a minimal 0.883% increase in weekly earnings. Therefore, the wage gap between local government and private sector workers further decreases after controlling for availability and interaction between "aval" and "localgov".

Part 7

Workers with at least a 4-year college degree

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.6322	0.005	1409.775	0.000	6.623	6.641
localgov	-0.0411	0.013	-3.268	0.001	-0.066	-0.016

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.4581	0.101	34.139	0.000	3.260	3.657
localgov	-0.0226	0.012	-1.882	0.060	-0.046	0.001
male	0.2204	0.008	26.460	0.000	0.204	0.237
black	-0.2058	0.016	-13.212	0.000	-0.236	-0.175
hispanic	-0.1682	0.021	-8.088	0.000	-0.209	-0.127
age	0.0905	0.005	18.873	0.000	0.081	0.100
age_squared	-0.0010	6.11e-05	-16.503	0.000	-0.001	-0.001
yrsch	0.0690	0.003	23.062	0.000	0.063	0.075

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.4623	0.113	30.774	0.000	3.240	3.684
localgov	-0.0260	0.013	-2.041	0.043	-0.051	-0.001
male	0.2204	0.011	19.376	0.000	0.198	0.243
black	-0.2056	0.018	-11.239	0.000	-0.242	-0.170
hispanic	-0.1730	0.036	-4.778	0.000	-0.244	-0.102
age	0.0903	0.005	17.656	0.000	0.080	0.100
age_squared	-0.0010	6.51e-05	-15.470	0.000	-0.001	-0.001
yrsch	0.0689	0.004	17.187	0.000	0.061	0.077
aval	-0.0079	0.010	-0.799	0.425	-0.028	0.012
interaction	-0.0300	0.012	-2.483	0.014	-0.054	-0.006

Regression results above for workers with at least a 4-year college degree show that "localgov" is statistically significant and brings a decrease of 4.02% of weekly earnings when "localgov" is the only explanatory variable. After controlling for demographic variables, "localgov" is statistically significant and brings a smaller decrease of 0.022% to weekly earnings. After further controlling for the availability of land, there is still no evidence supporting that "localgov" is significant and brings a minimal decrease of 0.026% to weekly earnings.

In conclusion, these regressions together show that being employed by the lo-

cal government or private sector becomes less significant to weekly earnings of highly educated workers when considering their demographic attributes and land availability. Highly educated workers earn more in the private sector than in the local government but the wage difference becomes ignorable by adding more controls.

Workers with less education

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.1251	0.003	2108.610	0.000	6.119	6.131
localgov	0.0433	0.011	3.831	0.000	0.021	0.066

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.8802	0.059	66.307	0.000	3.766	3.995
localgov	0.0164	0.010	1.615	0.106	-0.003	0.036
male	0.3127	0.005	61.548	0.000	0.303	0.323
black	-0.1992	0.007	-26.893	0.000	-0.214	-0.185
hispanic	-0.1592	0.008	-19.383	0.000	-0.175	-0.143
age	0.0564	0.003	19.460	0.000	0.051	0.062
age_squared	-0.0006	3.66e-05	-16.105	0.000	-0.001	-0.001
yrsch	0.0701	0.001	49.985	0.000	0.067	0.073

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.8817	0.061	64.019	0.000	3.762	4.001
localgov	0.0112	0.015	0.764	0.446	-0.018	0.040
male	0.3128	0.009	33.526	0.000	0.294	0.331
black	-0.1991	0.011	-17.892	0.000	-0.221	-0.177
hispanic	-0.1624	0.014	-11.775	0.000	-0.190	-0.135
age	0.0563	0.003	20.177	0.000	0.051	0.062
age_squared	-0.0006	3.46e-05	-17.026	0.000	-0.001	-0.001
yrsch	0.0701	0.002	31.735	0.000	0.066	0.074
aval	-0.0056	0.007	-0.786	0.433	-0.020	0.008
interaction	-0.0315	0.016	-1.962	0.051	-0.063	0.000

Regression results above for workers with less education show that "localgov" is statistically significant and brings an increase of 4.43% of weekly earnings when "localgov" is the only explanatory variable. After controlling for demographic variables, "localgov" is statistically significant and brings a smaller decrease of

0.017% to weekly earnings. After further controlling for the availability of land, there is still no evidence supporting that "localgov" is significant and brings a minimal decrease of 0.011% to weekly earnings.

In conclusion, the regressions reveal that the influence of being employed by either the local government or the private sector on the weekly earnings of highly educated workers also diminishes as demographic characteristics and land availability are included. In contrast, for less educated workers, working for the local government yields higher wages than the private sector; however, this wage differential becomes negligible upon the inclusion of additional controls.