SOURCES OF VARIATION IN RESIDENTS' SALARY INCOME: MCDONOUGH, ILLINOIS AND SURROUNDING RURAL COUNTIES

Adee Athiyaman

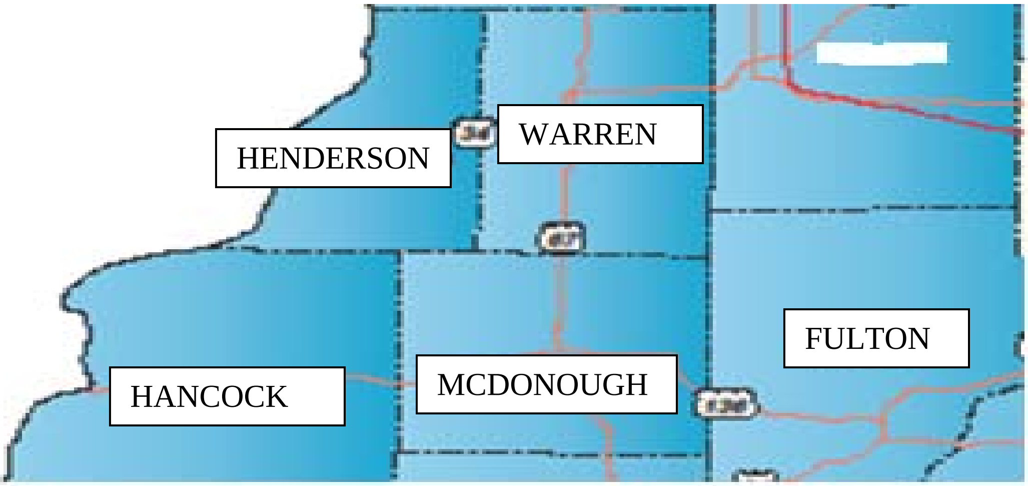
Draft, October 25, 2011

I. INTRODUCTION

> *On June 9, 2011, President Obama signed an Executive Order establishing the first White House Rural Council to accelerate the ongoing work of promoting economic growth in rural America. The Council is focused on increasing rural access to capital, spurring agricultural innovation, expanding digital and physical infrastructure in rural areas, and creating economic opportunities through conservation and outdoor recreation.*
> **Source***:* The White House: Office of the Press Secretary, *White House Rural Council Delivers Report on Rural America – Jobs and Economic Security for Rural America*, August 12, 2011.

With this in mind, we attempt to quantify the determinants of individual income[1] in McDonough, Warren, Henderson, Hancock, and Fulton counties (Figure 1). There are at least two reasons for doing this: (i) to statistically verify the relationships between demographic variables and income, and (ii) construct a predictive model of income at the individual level for use by businesses in market potential assessments. Regarding the latter, the type of automobile purchased would be a function of an individual's present and projected income. Since income depends on various combinations of demographic and occupational factors, an understanding of these factors should produce more valid predictions of income and thus market potential estimates for consumer goods such as automobiles.

Figure 1: Study Area: The Five Illinois Counties



[1] The term income refers to the total, inflation-adjusted, 2009, reported-income dollars (see the variables WAGP and ADJINC in the 2005-2009 ACS PUMS DATA DICTIONARY, January 10, 2011)

Consider the data in Table 1. It contains a sample of 10 records from the 2009 public use micro data (PUMA) provided by the US Census Bureau; *y* is income and *x* is years of schooling. For the five-county region, there were 3736 records or cases representing 45,971 individuals in the 16-years-and-above age group. From these data, we computed the variances and co-variances of *x* and *y* (Table 1). Our objective is to analyze these into their sources of variation such as race, and gender.

Table 1: Income and Schooling: Sample Data, Variance and Covariance

White Males			White Females	
	Income in $ (y)	Years of Schooling (x)	Income in $ (y)	Years of Schooling (x)
1	14893.26	9	26539.12	9
2	64948.05	9	24859.43	9
3	25755.26	9	6595.59	9
4	11231.53	9	27994.85	10
5	78814.46	9	26875.10	9
6	106828.34	9	11253.93	9
7	10122.93	6	7838.56	9
8	19036.49	9	22395.88	9
9	2575.52	9	1343.75	10
10	6718.76	11	29114.64	11
Nonwhite Males			**Nonwhite Females**	
	Income in $ (y)	Years of Schooling (x)	Income in $ (y)	Years of Schooling (x)
1	11197.94	12	47031.35	16
2	87343.93	13	7390.64	7
3	1007.81	11	12765.65	10
4	44791.76	10	48151.14	14
5	42552.17	11	23515.67	9
6	17916.70	11	14669.30	11
7	23403.69	6	29713.89	14
8	3359.38	11	43220.20	13
9	55989.70	13	1080.50	6
10	58229.29	14	648.30	7

Note: (n=7,456)
$\sigma^2_{\text{Schooling}} = 5.236$; $\sigma^2_{\text{Income}} = 8.563\ (10^8)$; $\sigma\ (_{\text{Schooling, Income}}) = 19,\ 967.8$

To understand the factors related to income, we reviewed published literature on the topic. A Google Scholar search in the format *allintitle: "determinants of income*" produced 150 papers in business, administration, finance, and economic subjects. Of these, 47 papers were published during the last five years. While no meta-analysis on income and its relationships with other variables could be found, the general consensus among the authors includes: (i) an individual's age is a determining factor and is an approximation of years of experience in an occupation (Mincer, 1974); (ii) increases in income is associated with increases in education (Hall and Jones 1999); (iii) income of whites is higher than the nonwhites (Gruen and Klasen 2008); (iv) since labor furnishes the greater part of per capita income, the proportion of hours worked per week is an important determinant of income (Fields, Cichello, Freije, Menéndez and Newhouse, 2003), and (v) travel time to work is an indicator of motivation to earn and thus is related to income (Gottschalk and Huynh 2010).

In the following pages, we utilize these variables to explore the sources of variation in the income variable. As mentioned earlier, data were assembled from the 2009 public use micro data provided by the US Census Bureau (Appendix 1). A total of 3736 records representing 45,971 individuals in the 16+ age group are used in model estimation.

MODEL SPECIFICATION AND ESTIMATION

Consider a model of the form:

$$Log(y) = \mu + \alpha D_1 + \gamma D_2 + \beta x_1 + \beta x_2 + \beta x_3 + \beta x_4 + u \qquad (1)$$

Where, y = Wages or salary income in 2009
D_1 = 1 for females, 0 for males
D_2 = 1 for whites, 0 for others
$\beta(x_1)$ = Years of Schooling
$\beta(x_2)$ = Age
$\beta(x_3)$ = Hours worked per week
$\beta(x_4)$ = Travel-time to work.

The model calibration suggests:

$$Log(y) =$$
$$6.23(50.64) - .017(-6.65)D_1 + .057(.65)D_2 + .09(16.56)x_1 + .011(13.01)x_2 +$$
$$.47(46.17)x_3 + .003(7.10)x_4$$

(Figures in parentheses are *t* ratios)[$R^2 = .48$]

Since the dependent variable is the logarithm of income, the regression coefficients can be interpreted as the estimated percentage change in the income to a unit change in particular quality. For example, the coefficient of x_1 indicates that an additional year of schooling would increase income by 9%. Similarly, an extra minute of travel-time to work (x_4) increases earnings by one-third of a per cent (.3%). The results also indicate that females earn 1.7% less than males.

Note that these results are based on the entire data set. In other words, we assumed that the subsets of coefficients in the male and female regressions are equal (*k*+1parameters): that is,

$$\mu_{Female} = \mu_{Male};\ \beta x_{1(Female)} = \beta x_{1(Male)}; \ldots.; \beta x_{4(Female)} = \beta x_{4(Male)} \qquad (2)$$

To test this assumption of (*k* + 1)-linear restrictions, we utilize the *F* test discussed by Kullback and Rosenbalt (1957):

$$F = \frac{\frac{(RRSS-URSS)}{(k+1)}}{\frac{URSS}{(n1+n2-2k-2)}}$$

The unrestricted residual sum of squares is obtained by estimating each equation separately and adding their residual sum of squares. This has degrees of freedom (n_1-k-1) + (n_2-k-1). The restricted residual sum of squares is obtained by pooling the data and estimating a single equation. This residual sum of squares has (n_1+n_2-k-1) degrees of freedom.

Table 1 shows the results of the hypothesis test of stable relationships among parameters. Put simply, the relationships specified in EQ 2 were rejected; the parameters differ for male and female.

Table 1: Tests of Linear Restrictions: H$_0$: Stable Relationship

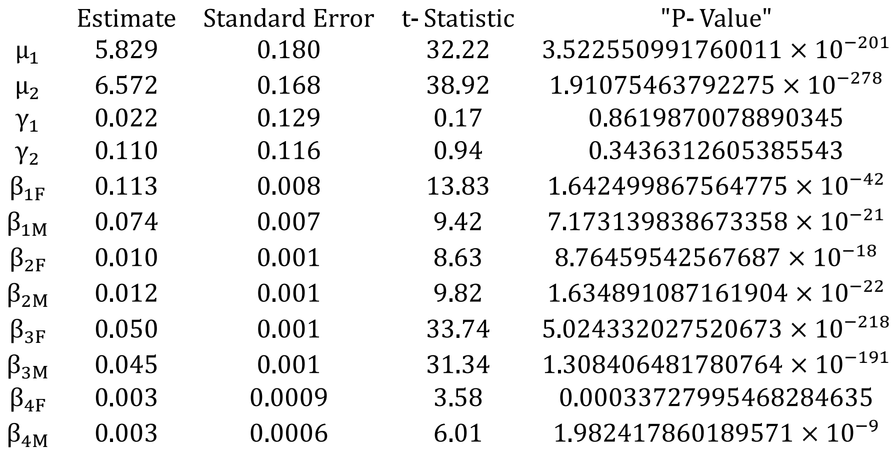
Component	**Value**	**Degrees of Freedom (df)**
Restricted Residual Sum of Squares (RRSS)	2205.12	6
Residual Sum of Squares (Female)	910.14	
Residual Sum of Squares (Male)	1281.63	
Unrestricted Residual Sum of Squares (URSS)	2191.77	3724
F	3.77932 (p= 0.002338129568043562)	6, 3724

The *F* test in Table 1 shows that the regression parameters differ between male and female. But, it doesn't tell which particular coefficients are different. To assess this, we re-specify EQ 1 as follows:

$$\binom{y_{Female}}{y_{Male}} = \mu_1\binom{1}{0} + \mu_2\binom{0}{1} + \gamma_1\binom{1}{0} + \gamma_2\binom{0}{1} + \beta_{1F}\binom{x_1}{0} + \beta_{1M}\binom{0}{x_1} + \beta_{2F}\binom{x_2}{0} + \beta_{2M}\binom{0}{x_2} + \beta_{3F}\binom{x_3}{0} + \beta_{3M}\binom{0}{x_3} + \beta_{4F}\binom{x_4}{0} + \beta_{4M}\binom{0}{x_4} + \binom{u_1}{u_2} \tag{3}$$

We calibrated this equation with 12 indicator variables without a constant term (n = 3736). The results presented in Table 2 show no race effect on income. However, the effects of schooling are more pronounced for females than males (11% increases in income for females for an additional year of schooling compared to 7% for males). Table 3 shows the 95% confidence intervals of the parameters listed in Table 2.

Table 2: Results of Model Calibration: The 12 Indicator Variable Model; $R^2 = 0.99$; Residual correlation = 0.05 (suggests that intercepts are not correlated with errors)



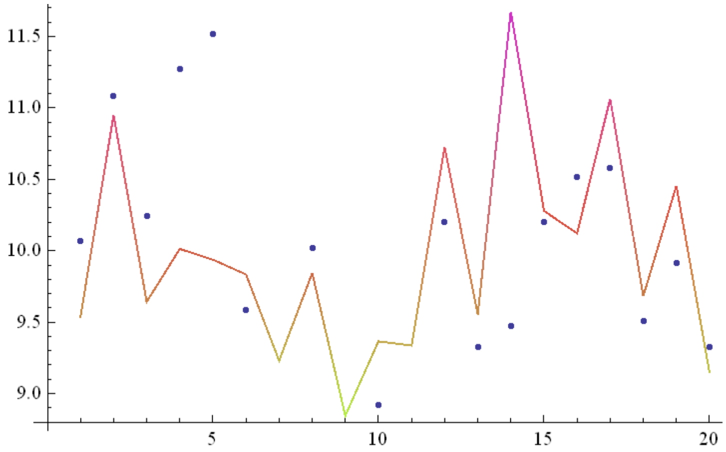
	Estimate	Standard Error	t- Statistic	"P- Value"
μ_1	5.829	0.180	32.22	$3.522550991760011 \times 10^{-201}$
μ_2	6.572	0.168	38.92	$1.91075463792275 \times 10^{-278}$
γ_1	0.022	0.129	0.17	0.8619870078890345
γ_2	0.110	0.116	0.94	0.3436312605385543
β_{1F}	0.113	0.008	13.83	$1.642499867564775 \times 10^{-42}$
β_{1M}	0.074	0.007	9.42	$7.173139838673358 \times 10^{-21}$
β_{2F}	0.010	0.001	8.63	$8.76459542567687 \times 10^{-18}$
β_{2M}	0.012	0.001	9.82	$1.634891087161904 \times 10^{-22}$
β_{3F}	0.050	0.001	33.74	$5.024332027520673 \times 10^{-218}$
β_{3M}	0.045	0.001	31.34	$1.308406481780764 \times 10^{-191}$
β_{4F}	0.003	0.0009	3.58	0.00033727995468284635
β_{4M}	0.003	0.0006	6.01	$1.982417860189571 \times 10^{-9}$

Table 3: Parameter Confidence Intervals

Parameter	**Point Estimate**	**Interval Estimates (95% Confidence Level)**	
		Minimum	**Maximum**
μ_1	5.829	5.474	6.184
μ_2	6.572	6.241	6.903
γ_1	0.022	-0.231	0.276
γ_2	0.110	-0.117	0.338
β_{1F}	0.113	0.097	0.129
β_{1M}	0.074	0.059	0.090
β_{2F}	0.010	0.008	0.013
β_{2M}	0.012	0.010	0.015
β_{3F}	0.050	0.047	0.053
β_{3M}	0.045	0.042	0.047
β_{4F}	0.003	0.001	0.005
β_{4M}	0.003	0.002	0.005

Having established the fact that it is education that necessitates differential estimation of female and male income equations, we explore the predictive capability of the model given in EQ 3 (Figure 1). A note of caution though; the accuracy of prediction depends on the stability of the coefficients between the period used for estimation and the period used for prediction. Hence care should be used in extrapolating the income function.

Figure 1: Model Fit: Sample of 20 "Hold-Out" Observations



DISCUSSION

There is a "gender" related income gap in the region[2]. On average, the male residents' wages were 1.5 times as much as the female residents' wages in 2009, *ceteris paribus* (Table 4). However, education is more rewarding for the female population; wage increases to a unit change in schooling are 4% points more for females than males.

The income function for females in the population, based on point estimates given in Table 3, is:

$$Logy = 5.829 + .113(years\ of\ schooling) +\ .01\ (age) + .05(hrs\ worked\ per\ week) \\ +\ .003\ (travel\ time\ to\ work)$$

Thus, for example, if our interest is in predicting the annual, after-tax wages of a female high school graduate, 45 years of age, who works for 40 hours per week, and travels 10 minutes to work, it is:

$$y = Exp[5.829 + .113(12) +\ .01\ (45) + .05(40) \times 52 +\ .003\ (10)] = \$15{,}756.37$$

Similarly, the income function for males is:

$$Logy = 6.572 + .074(years\ of\ schooling) +\ .012\ (age) + .045(hrs\ worked\ per\ week) \\ +\ .003\ (travel\ time\ to\ work)$$

And the annual, average, after-tax wages for a male high school graduate, 45 years of age, who works for 40 hours per week, and travels 10 minutes to work is:

$$y = Exp[6.572 + .074(12) +\ .012\ (45) + .045(40) +\ .003\ (10)] = \$18{,}582.92$$

[2] Appendix 2 presents additional statistical evidence on this finding.

The difference between the two earnings is approximately 18%; in favor of the male worker. Note that this is the average across all occupation and industry. Future work in this area will explore industry-specific variations in income.

Table 4: Male, Female Differences in Wages & Salaries: Based on Average Values of the Predictors

Function	**Point Estimate ($)**	**Interval Estimate: Low**	**Interval Estimate: High**
$y\ (Female) = Exp[5.829 + .113(years\ of\ schooling) + .01\ (age) + .05(hrs\ worked\ per\ week) + .003\ (travel\ time\ to\ work)]$	14,955.7	6,686.03	34,929.51
$y\ (Male) = Exp[6.572 + .074(years\ of\ schooling) + .012\ (age) + .045(hrs\ worked\ per\ week) + .003\ (travel\ time\ to\ work)]$	22,766.09	10,560.26	51,144.08

SUMMARY AND CONCLUSION

The model presented in this paper can be used to measure income at the individual level. For example, it can be deployed to predict the average salary of a now 25 years old person at age 65. Such calculations would be beneficial, for example, for insurance agencies to assess the adequacy of life insurance of residents in the region, for a durable-goods manufacturer to evaluate market potential, etc.

Several market research firms offer analyses like this but at a cost. For example, a market report like this would cost several hundreds of dollars, if not thousands of dollars, if sourced from management consulting firms (see for example, county-level data prices at http://www.cement.org/market/mkt_apparent_use.asp). In contrast, this report is provided free of costs by I^2 (www.instituteintelligence.com) as a service to all interested in the study region's economic development.

REFERENCES

Conway, D. A., and Roberts, H. V. (1983). Reverse regression, fairness, and employment discrimination, *Journal of Business and Economic Statistics*, 1(January), 75-85.

Fields, G. S., Cichello, P. L., Freije, S., Menéndez, M., and Newhouse, D. (2003). For richer or for poorer? Evidence from Indonesia, South Africa, Spain, and Venezuela. *Journal of Economic Inequality*, 1(1):67–99.

Gottschalk, P. and Huynh, M. (2010). Are earnings inequality and mobility overstated? The impact of non-classical measurement error. *Review of Economics and Statistics*, 92(2):302–315.

Gruen, C. and Klasen, S. (2008). Growth, inequality, and welfare: comparisons across space and time. *Oxford Economic Papers*, 60(2):212–236.

Hall, Robert E. and Charles I. Jones (1999). Why do some countries produce so much more output per worker than others?" *Quarterly Journal of Economics* 114(1), February, 83-116.

Kullback, S and Rosenblatt, H. M. (1957). On the analysis of multiple regression in K categories, *Biometrika*, 67-83.

Mincer, Jacob (1974). *Schooling, Experience, and Earnings*, New York: Columbia University Press.

Morgan, James (1962). The anatomy of income distribution, *The Review of Economics and Statisitcs*, 277.

Appendix 1: Data Source

This report is based on the PUMA boundaries that include the five-county study region (PUMA code 200 for Illinois). A complete description of the 2009 data used in the report can be found at:

> U.S. Census Bureau, *A Compass for Understanding and Using American Community Survey Data: What PUMS Data Users Need to Know* U.S. Government Printing Office, Washington, DC, 2009.

The variables used in the report include:

- Age (AGEP): 1 to 99 years;
- School (SCHL): Educational attainment. Coded from "01" to indicate "no schooling" to "16" to represent "doctoral degree";
- Gender (Sex). Coded 1 = Male, 2 = Female. We recoded them as 1 = Female, and 0 = Male.
- Race: Coded 1 = White; 0 = other;
- Wages (WAGP): Wage or salary income for the last 12 months;
- Hours worked per week for the past 12 months (WKHPP);
- Travel time to work (JWMNP): Coded from 1 to 200 minutes.

Appendix 2: Additional Evidence on Salary Discrimination

To further explore gender differences in salary, we specify a model of the form (Conway and Roberts, 1983):

$$Log(y) = \beta_1 x_1 + \beta_2 x_2 + u \qquad \text{(A1)}$$

Where, *y* = salary
x_1 = true qualifications, and
x_2 = gender ; 1 = Male, and 0 = Female
u = error

Our focus is on β_2: the effect of gender on income. The estimated equation shows gender discrimination (figures in parentheses are *t* ratios):

$$Log(y) = 0.1290342781834275(\mathrm{x}_1) + 0.5370669977617308\ (\mathrm{x}_2) + 8.48605197726124$$
(17.645) (16.327) (85.293)

However, to highlight any gender bias in the region, it is essential to demonstrate that among men and women receiving equal salaries, the men possess lower qualifications. This requires calibrating a model of the form:

$$X_1 = \gamma_1 y + \gamma_2 x_2 + w \qquad \text{(A2)}$$

The estimates for A2 are (figures in parentheses are *t* ratios):

$$X_1 = 0.000027685416934587885(\mathrm{y}) - 0.7920364186482719(\mathrm{x}_2) + 12.169942586328965$$
(20.01) (10.92) (168.12)

Again, the results indicate gender discrimination - among men and women receiving equal salaries, men possess lower qualifications.

These analyses raise a methodological issue about bias in the estimator β_1 in EQ A1. For instance, assume that x_1 is measured with error – years of schooling are impure measures of true qualifications. Let:

X_1 = measured qualifications = $x_1 + v$; where v = error.

The parameter β_1 is estimated as:

$$\hat{\beta}_1 = \frac{\sum(y\ (x1+v))}{\sum(\llbracket x1+v)\rrbracket^2} \qquad = \frac{\sigma^2_{x_1 y}}{\sigma^2_{x_1} + \sigma^2_v}$$

Since population parameter = $\beta_1 = \frac{\sigma^2_{x_1y}}{\sigma^2_{x_1}}$, we have

plim $\hat{\beta}_1 = \frac{\beta}{1+\frac{\sigma^2_v}{\sigma^2_x}}$

Thus, in our model, $\hat{\beta}_1$ is an underestimate of β_1. Having established the lower bound for β_1, we explore its upper bound.

Consider the reverse regression $\beta_{x,y}$. Where *x* is years of schooling, and y is salary. It can be shown that $\hat{\beta}_{xy}$, under the errors-in-variables conceptualization is:

$$\text{plim}\,\frac{1}{\hat{\beta}_{xy}} = \beta\left[1+\frac{\sigma^2_e+\sigma^2_v}{\beta\sigma^2_x}\right] \qquad \text{(A3)}$$

EQ A3 is the upper bound for $\hat{\beta}_1$. Computations suggest that β_1 would lie between:

$0.44821 < \beta_1 < 0.61002.$

Since we are not proposing policy changes to address gender discrimination, we do not pursue this type of modeling: that is, measurement-error analysis further. However, readers should note that the analysis presented do suggest gender inequality in wages in the study area.