Regression Analysis of the Relationship between Income and Work Hours

Sina Mehdikarimi Samuel Norris Charles Stalzer

Georgia Institute of Technology
Econometric Analysis (ECON 3161)
Dr. Shatakshee Dhongde
Spring 2015

Abstract

This study analyzes the relationship between personal income and the usual number of hours worked per week. Labor market economics imply a positive relationship between hours worked and higher income, but also suggest a that past a specific income threshold, individuals tend to work less hours. The relationship between hours worked and personal income is modeled using data from the United States Census Bureau's American Community Survey from 2013. Simple regression analysis tested the log of adjusted personal income against hours worked, and the multiple regression expanded this analysis to include gas utility prices, number of workers in family, food stamp assistance, and number of persons in family as variables. The simple regression model demonstrates that personal income has a positive relationship with hours worked, while the multiple regression model shows that this effect diminishes as income level increases.

TABLE OF CONTENTS

1.	Introduction	2
II.	Literature Review	2
III.	Data and Methodology	6
IV.	Results	12
V.	Conclusion	16
VI.	References	18
VII.	Appendix	19

I. Introduction

The relationship between income and hours worked is a question of great practical and, for some people, perhaps even moral significance. The main question is whether or not those of high income work more or less, on average, than those of low income. Perhaps the well-paid choose to work more because each additional hour worked is highly lucrative, thus earning more simply because they work more. On the other hand, they may not work as much because they can earn enough to fulfill their needs in a shorter amount of time. Perhaps those earning low salaries must work more because that is the only way to earn an adequate income. It is also possible that those earning less work less because they do not see it as worthwhile to spend additional time working for low wages.

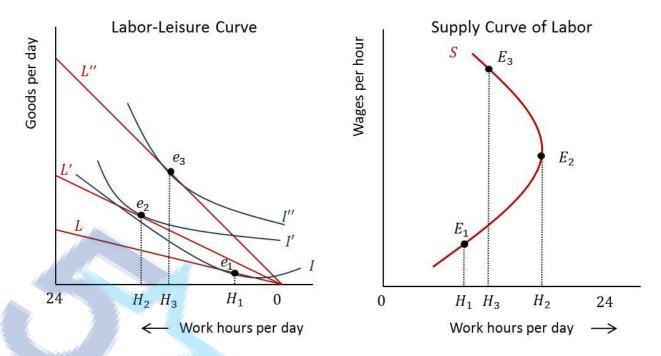
The answer to this question can be boiled down to whether or not higher wages incentivize people to work less or work more. The theoretical principles describing these incentives – the income effect and the substitution effect – will be discussed in more detail in the following section.

Before going into such details, we present the following hypothesis. The income effect suggests that higher income causes people to work fewer hours. The substitution effect suggests that higher income causes people to work more hours. Our hypothesis is that the income effect is generally stronger and, as such, higher income should lead to fewer hours worked.

II. Literature Review

Investigation of labor market dynamics has permeated mainstream economics for many years. In general, the term "labor markets" refers to the interaction between employers who seek workers and laborers looking to work for wage. To determine how much an individual will be willing to work, economists refer to labor supply curves.

Labor supply curves reflect the "labor-leisure" tradeoff, which suggests that more hours worked will result in higher wages but also will result in a reduced amount of leisure time for a worker (1). In the graphical analysis, wage rate is measured on the vertical axis while the hours worked is measured on the horizontal axis. The classic model, shown below, is said to be "backward-bending (1).



As the real wage rate increases, the opportunity cost of spending more leisure hours increases as a result, leading to laborers working more hours, illustrated by a movement along the labor supply curve from (H_1, W_1) to (H_2, W_2) . This activity is known as the substitution effect (1). However, as one can discern from the graph, hours worked past a certain real wage rate declines (also viewed as an increase in leisure time) once past a certain real wage rate. If leisure time is considered a normal good (meaning the demand for leisure increases with an increase in real wage) then an increase in real wage will increase the amount of leisure time "consumed" by workers, leading them to work less, or supply less labor (1). This phenomenon is known as the income effect. The relative sizes of the income and substitution effects determine the slope of the labor supply curve. If the substitution effect outweighs the income effect, workers will supply more labor (spend more time working), resulting in a positive slope for the labor supply curve; conversely, if past a certain real wage the income effect is stronger than the substitution effect, then workers will spend more time in leisure, and the slope of the labor supply curve is negative (1). However, if leisure time is considered an inferior good, the substitution and income effects work together, causing an unambiguous decrease in leisure (1).

Labor supply curves can be backwards-bending if leisure is inferior at low wages and normal at high wages or if leisure is normal at all wages while the substitution effect to work

more hours is more pronounced at low wages and the income effect to work less hours is predominant at high wages (1). Therefore, the central question of this study - whether the income or substitution effect is more influential - is related to whether leisure is an inferior or a normal good.

The class of earners who have worked the most hours has varied considerably over time in the United States. A study by Kuhn and Luzano (2005) in the National Bureau of Economic Research working paper series examines the shift of the classification of workers who worked the longest hours in the U.S. over a period of two decades (2). This study uses a large number of variables to describe the shift in working hours among American men. Variables include race, level of education attained, marital status, union membership, age, and industry, among others. From 1940 to 1970, the percentage of employed American men working 48 or more hours declined (2). However, from 1970 to 1990 the share of men working this amount increased, with the greatest increase coming in the 1980s (2). By 2000, the trend again reversed somewhat, with the percentage of men working 48 or more hours decreasing, but with the overall share of men working 48+ hours at 24.3%, a larger share than in 1970 (2). In general, this effect was strongest among highly educated, high earning, older men (2). Among the top quintile of earners, the tendency to work longer hours increased 11.7%, while earners in the bottom quintile working longer hours decreased 8.4% (2). Overall, the bottom 20% of earners were most likely to work long hours in 1979, while long hours were more common among the upper quintile in 2006 (2).

Several other earner characteristics appear to be correlated with an increase in hours worked. The number of hours worked for college graduates increased through this period, while the number of hours worked for college dropouts decreased (2). Additionally, the more educated were more likely to work longer hours (2). Earners who were married were more heavily correlated with long hours than single earners (2). Black and Hispanic workers were less likely to work long hours over the period 1970-2006 (2). Earners with union membership went from being negatively correlated with an increase in hours worked to positively correlated (2). However, each of these factors on their own do not explain the relative increase in hours worked from 1970 to 2006 (2).

The relationship between hours worked and income has been analyzed in many other countries. A study published by the New Zealand Department of Labor (2008) details the correlations between working hours and a variety of different variables. The central question of this study concerns finding out the educational attainment profile of long-hours workers. The report concludes that "while those with the highest qualifications are the most likely to work long hours, the largest group of long hours workers is in the 'no qualifications' category" (3). This study also includes many other variables such as ethnicity, age, gender, occupation, highest level of education earned, industry, location of workplace, family, and (most importantly for the purposes of this paper) income, and how they affect the number of hours worked. According to the New Zealand census, 22.68% percentage of the workforce works "long hours," or more than 50 hours per week (3). Typically, as income increased, the proportion of workers working longer hours increased, in accordance with several other studies (3).

The basis for a study done by Kimball and Shapiro (2008) is that the income and substitution effects cancel out in the overall labor market due to the fact that overall labor supply stays constant with respect to wage (4). The main variables that were taken into account by this study were marital status, optimization over time, and fixed costs of going to work. While it provides an in-depth look at elasticities in labor supply, the difference in the income and substitution effects in different income brackets was an area left unexplored. The overall results show very high labor supply elasticities in response to a wage shock, implying that the majority of people chose to reduce their hours worked when given a large and sudden increase in earnings.

The purpose of this paper is to discover the effect of individual income on the number of hours worked per week in the United States. This paper expands upon concepts from the studies already cited in a variety of ways. In both the Kimball and Shapiro and National Bureau of Economic Research studies, income was measured by household; in this paper, individual income is used instead. Instead of studying the effect on working hours of factors such as sex, age, or marital status, we measure the impact of income broken down by income bracket. Although the National Bureau of Economic Research study regresses income on hours worked via income brackets, it does this through a time series. Our study examines only one year, and it

changes the regression model by regressing work hours on income. In addition, we add several other variables not addressed by any of the aforementioned studies, including the number of people in the family, food stamp assistance workers in the family, and the monthly price of gas utilities. The effect of these variables can be used to demonstrate the impact of food stamps or rising gas prices on hours worked per week. Insights into these variables and how they affect hours worked will add to the body of knowledge in the field of labor market economics.

III.Data

The data comes from the United States Census Bureau's American Community Survey. We have combined the individual and household data sets from 2013 to create one large data set. This was possible due to serial numbers matching individuals with their household. A few variables have been added or modified, as detailed below in the variable descriptions.

We explain each variable and its importance below. Our variable explanations are accompanied by a table containing descriptive statistics for each variable. The dependent variable is *wkhp*, the independent variable from the simple regression is *ladjpincp*, and all other variables will be used in the multiple regression.

Variable Summary Statistics					
Variables	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
wkhp	1,568,032	39.25	14.44	1	99
pincp	2,591,297	37,578	66,518	-12,200	1,272,000
ladjpincp	2,591,297	8.32	3.91	0	14.05
mid (dummy)	1,015,660	1	0	N/A	N/A
high (dummy)	847,277	1	0	N/A	N/A
npf	2,665,655	3.69	1.49	2	20
fs	548,828	1	0	N/A	N/A
wif	2,665,655	1.63	0.83	0	3
lgasp	3,087,625	2.36	2.22	0	6.30

Simple Regression Variables

wkhp

The individual's usual hours worked per week is captured with the variable *wkhp*. The purpose of this study is to determine the effect of income on hours worked, so the number of hours a person works in a typical week is a logical choice for the dependent variable. The data set provides values in the range of 1-99 hours. The mean is unsurprisingly near 40 hours per week, but there is a notable amount of variation.

ladjpincp

ladjpincp is the logarithm of the individual's adjusted personal income in the past year. This is the primary independent variable of interest, and it appears in both the simple and multiple

regression equations. It was obtained by adjusting the variable pincp (individual's personal income in the past 12 months) that appeared in the original data set, and then taking the log of that adjusted value. Of the income variables in our data set (personal, family, household), we found that personal income had the greatest effect on hours worked. We chose to take the log of personal income because a percentage change in income is economically more significant than the number of dollars that income has changed (i.e. 10% increase in income is large, but \$10 increase in income is negligible). The original variable from the data set accounts for people age 15 and older and provides data in the range -\$19999 - \$9,999,999. Because the logarithm of 0 or a negative number cannot be found, all values less than 1 were set to 1 before taking the logarithm (it should be noted that log(1) = 0). Due to the relevance of the individual personal income (before adjustment or logarithm, pincp), its descriptive statistics are also provided in the table.

Additional Multiple Regression Variables

mid

This is a binary variable representing workers with a personal income greater than \$20,599 but less than or equal to \$104,096. Since economic theory on the competing substitution and income effects suggests that the effect of income on hours worked may change as income increases, we have included dummy variables to account for low, middle, and high income. This variable, if it has a value of 1, denotes an individual of middle personal income. If zero, the individual is either in the top or bottom income quintile.

high

The binary variable *high* indicates whether personal income in the past 12 months is greater than \$104,096. As stated above, we seek to account for how the effect of income on hours worked varies with income level. This variable, if it has a value of 1, denotes an individual of high personal income. The choice of these specific income values (for both the *high* and *mid* variables) is based on the distribution of income in the United States. The lower fifth quintile lies

below \$20,600, the second through fourth quintiles range from \$20,600 to \$104,096, and the fifth quintile consists of all incomes larger than \$104,096 (5).

npf

The number of persons in the worker's family is represented by the variable npf. The number of people in a person's family may affect the amount of time he or she chooses to work. Perhaps people with large families must work more to provide for a large group of people. On the other hand, someone with a large family could place an unusually high value on leisure and time with family, which could lead them to work less. This variable only accounts for family households, and it provides data in the range 2-20.

fs

The variable *fs* is a binary value indicating whether or not a person receives food stamps. A person receiving food stamps may work longer in order to avoid requiring government support. Alternatively, a person receiving food stamps may work less because some of their needs are provided by the government. The data set originally provided values of 1 for those receiving assistance and 2 for those not receiving food stamps. In order to be consistent with the commonly used values of 0 and 1 for binary variables, *fs* was recoded so that those not receiving food stamps have a value of 0.

wif

The number of workers in the family during the last 12 months is represented by the variable *wif*. In families with a large number of workers, each individual worker may not have to work as many hours as someone who is the only worker in their family. This variable only accounts for family households and provides data in the range 0 - 3 (with households with more than 3 workers being capped at 3).

lgasp

The log of the monthly cost of gas utilities is represented by lgasp. People may choose to work more if the cost of utilities is high. We obtained this variable by taking the log of the variable gasp (monthly cost of gas utilities) provided in the original data set. The original data set provides values for the range of \$4 – \$999 per month. It assigned values of 1-3 for qualitative outcomes for circumstances like including gas utilities in rent, included them electricity, or not using gas utilities. We have assigned those values to 1 (it should be noted that log(1) = 0) because the person is not paying an independent gas bill. We chose the log of the variable because a percentage change will be more meaningful than an absolute change (i.e. 10% increase is noteworthy, while a \$10 increase may not be large).

Gauss-Markov Assumptions

In order for our regression to produce unbiased estimates and for our statistical significance test to be valid, we must make sure our sample meets the Gauss-Markov assumptions.

1. Linear in parameters

This is satisfied through a population model written as:

$$wkhp = \beta_0 + \beta_1 ladjpincp + \delta_0 mid + \delta_1 high + \beta_2 npf + \delta_2 fs + \beta_3 wif + \beta_4 lgasp + u$$

2. Random sampling

We assume the United States Census Bureau uses proper sampling techniques to obtain a sample representative of the American population. The American Community Survey results display high coverage rates for many segments of the American people (6).

3. No perfect collinearity

None of the independent variables had a constant value. Additionally, we calculated the variance inflation factor (VIF) for each independent variable, and all were less than 3.5. This is low enough to conclude that there are no problems of multicollinearity in our regression. We used the dummy variables *mid* and *high*, with our base value being individual income at or below \$20,599.

Variable	VIF
high	3.44
ladjpincp	3.39
mid	2.66
npf	1.42
fs	1.30
wif	1.24
lgasp	1.03
Mean VIF	2.07

4. Zero Conditional Mean

The model does not necessarily avoid omitted variable bias due to the complex nature of the dataset and the large number of possible variables. However, if we assume normality (as discussed below), then the zero conditional mean condition is implied. In addition, the variables that are included in the multiple regression are statistically significant, and we have justified the reasoning behind the inclusion of each. Therefore, we can assume a zero conditional mean.

5. Homoskedasticity

Although the data may violate homoskedasticity, the assumption of normality (discussed below) allows us to assume homoskedasticity for the purposes of this study.

6. Normality

Since the sample size for the multiple regression is well over 1.2 million, we can safely assume normality for the dataset.

Note: Information about these assumptions and the methods for testing them was obtained from a site maintained by UCLA (7).

IV. Results

The simple regression tests the effect of the log of adjusted personal income on the usual hours an individual works per week. Below is the simple regression population equation.

$$wkhp = \beta_0 + \beta_1 ladjpincp + u$$

After running the simple regression in STATA, we obtained the following simple regression equation. Under each constant and variable coefficient in the equation is its standard error. The STATA output may be viewed in the appendix.

$$wkhp = -13.320 + 5.157ladjpincp$$

$$(0.082) \quad (0.008)$$

$$n = 1568032 \qquad R^2 = 0.209$$

While the simple regression yields information relevant to the hypothesis, it could be made more accurate with the inclusion of additional relevant variables. The multiple regression tests how usual hours worked per week is affected by the log of adjusted income, income level, number of people in the family, food stamp status, workers in the family, and the log of gas utility prices. The multiple regression population equation is printed below:

$$wkhp = \beta_0 + \beta_1 ladjpincp + \delta_0 mid + \delta_1 high + \beta_2 npf + \delta_2 fs + \beta_3 wif + \beta_4 lgasp + u$$

Running the multiple regression in STATA provided the following equation. Under the constant and each variable coefficient in the equation is its standard error. The STATA output can be viewed in the Appendix.

$$wkhp = -26.770 + 6.804ladjpincp - 2.077mid - 7.574high - 0.293npf - 0.402fs$$

$$(0.139) \quad (0.014) \quad (0.032) \quad (0.054) \quad (0.008) \quad (0.035)$$

The results comparison for the simple and multiple regression are shown in the following table.

Dependent Variable log(adjpincp)			
Independent Variables	Wkhp (simple)	Wkhp (multiple)(2)	
Log adjpincp standard error:	5.157 (0.008)	6.804 (0.014)	
mid (dummy)	N/A	-2.077 (0.032)	
High (dummy)	N/A	-7.574 (0.054)	
npf	N/A	-0.293 (0.008)	
fs	N/A	-0.402 (0.035)	
wif	N/A	-0.657 (0.016)	
Log (gasp)	N/A	0.189 (0.004)	
No. of obs.	1568032	1291298	
Intercept	-13.20	-26.77	
R-square	0.209	0.298	

Before interpreting the results and discussing their economic and practical significance, it is important to test the statistical significance of each variable coefficient. If a variable's coefficient is not statistically significant, we cannot be sure that the variable has any effect at all on usual hours worked per week. Statistical significance is tested using a t distribution, and all tests are two-tailed tests. Since the critical value for the 1% significance level is 2.576 and |t| = 642.95 for *ladjpincp* in the simple regression with 1,568,030 degrees of freedom, *ladjpincp* is statistically significant at the 1% level.

The same test is done for all independent variables in the multiple regression. All of them are significant at the 1% level, for which the critical value is 2.576. The t values for each independent variable are reported below. The degrees of freedom in multiple regression are 1,291,290.

ladjpincp:
$$|t| = 476.27 > 2.576$$

 $mid: |t| = 65.76 > 2.576$
 $high: |t| = 139.87.27 > 2.576$
 $npf: |t| = 35.01 > 2.576$
 $fs: |t| = 11.42 > 2.576$
 $wif: |t| = 41.24 > 2.576$

Additionally, we must test whether the independent variables added for the multiple regression are jointly significant. An F test can be used to test the joint statistical significance of mid, high, npf, fs, wif, and lgasp. The unrestricted model will be the multiple regression (with n-k-1 degrees of freedom), and the restricted model will be the simple regression (with

lgasp: |t| = 44.07 > 2.576

 $F = \frac{(R_{unrestricted}^2 - R_{restricted}^2)/q}{(1 - R_{unrestricted}^2)/(n - k - 1)}$ $F = \frac{(0.298 - 0.209)/6}{(1 - 0.298)/(1291290)}$

q restrictions). The F statistic for our regressions is calculated below.

F = 27285.093

Since the critical value of the $F_{6,\ 1291290}$ distribution at a 1% significance level is 2.80 and F=27285.093>2.80, we can say that mid, high, npf, fs, wif, and lgasp are jointly significant at the 1% level.

Although every independent variable in the model is statistically significant, it is important to discuss the practical and economic significance of the variable coefficients. All of the independent variables in the multiple regression had a statistically significant effect on usual

hours worked per week, and the variables added in after the simple regression were jointly significant.

For each 1% increase in adjusted personal income, hours worked per week goes up by 0.06804, using the simple regression. This means that a 10% increase leads to a predicted 40.8 more minutes of work per week, and we would expect someone whose salary is double that of their neighbor to work nearly 7 hours more. This preliminary observation suggests that the substitution effect is stronger than the income effect in the overall population.

However, the effect of income on work hours is not the same across all income levels. Someone of middle income (\$20,600 - \$104,096) is predicted to work over 2 hours less per week than a low income (< \$20,600) individual who is otherwise the same. The effect is even greater for high income (> (\$104,096) individuals, who are predicted to work 7.574 hours less than an otherwise equal low earner. This observation supports our hypothesis that as income increases, hours worked decreases.

Each additional person in a worker's family results in about 17.6 fewer predicted minutes of work per week. A possible explanation for this is that as the number of children in a family increases, the amount of care and attention necessary for kids also increases, which would lower the amount of time spent working per week by the family's primary earner. Regardless of the explanation, the effect of just over a quarter of an hour per person is rather small.

Someone receiving food stamp assistance is predicted to work a little over 24 minutes less per week than someone not receiving food stamps. This could be explained by the fact that a worker receiving food stamps will aim to keep their income below the maximum threshold for food stamp assistance (if the worker's income gets too high, they may lose food stamp benefits, resulting in greater loss of net income). This effect is quite small.

Each additional worker in a family leads to about 40 fewer predicted minutes of work per week per worker. A possible reason for this result is that with more workers in a family, the overall family income increases. When the total family income increases, each worker has less

of an obligation to work more hours, as the burden of earning additional income has been split among multiple earners. The effect of additional workers is somewhat small.

Finally, a 1% increase in the price of gas utilities leads to a 0.00189 increase in predicted hours worked per week. Thus, a 10% increase in gas price increases the predicted work time by just over a minute, and doubling gas prices raises predicted work time by less than 12 minutes. This effect is practically negligible.

V. Conclusions

The regression analysis is largely in agreement with economic theory and partially in line with our hypothesis. The substitution effect appears to be relevant because of the positive and significant effect of the log of adjusted income on usual hours worked per week. This suggests that as workers earn more, they will tend to work more. This lines up with the substitution effect making leisure time more expensive to those earning high incomes. Nevertheless, this effect appears to lessen in higher income brackets. According to our model, middle income workers work a predicted two hours less than low income workers, ceteris paribus. High income workers work a predicted seven and a half hours less than low income workers, ceteris paribus. This suggests the relevance of the income effect in causing workers to consume more leisure time as their income increases. While our hypothesis did not account for the importance of the substitution effect, the results of our regression analysis generally support economic theory.

Possible shortcomings of the multiple regression model are violations of the Gauss-Markov assumptions, namely the zero conditional mean and homoscedasticity; however, these violations are not a serious concern, as the large number of observations allows us to assume normality. Although the personal income variable only counts income for people age 15 and older, age was not a variable accounted for in the multiple regression. It is possible that age would have an effect on whether the income effect or substitution effect is more significant. In order to further the research in this area of economics, one could add independent variables describing the age and sex of the primary earner to this model.

Although the results of this study do not yield obvious policy prescriptions, the conclusions from the model confirm established labor market economic theory. The independent variables of the greatest practical and economic significance were the log of adjusted income and the binary variables for income level. This suggests that, even when other variables are included in the analysis, income is still an important predictor of weekly work hours.



VI. References

- Perloff, Jeffrey. *Microeconomics*. 6th ed. Boston, Massachusetts: Pearson Addison Wesley, 2012.
- Kuhn, Peter, and Fernando Lozano. "The Expanding Workweek? Understanding Trends
 In Long Work Hours Among U.S. Men, 1979-2004." National Bureau of
 Economic Research, 2005. Accessed March 3, 2015.
- 3. "Working Long Hours in New Zealand." *New Zealand Department of Labor*, 2008.

 Accessed March 10, 2015. http://www.dol.govt.nz/PDFs/long-hours.pdf.
- 4. Kimball, Miles, and Matthew Shapiro. "Labor Supply: Are the Income and Substitution Effects Both Large or Both Small?" *National Bureau of Economic Research*, 2008. Accessed April 9, 2015. http://www.nber.org/papers/w14208.pdf
- 5. "Household Income Quintiles 2000-2012." *Census Bureau Historical Income Tables*, October 23, 2015. Accessed April 10, 2015.
- 6. "Coverage Rates." *American Community Survey*, 2013. Accessed April 10, 2015. http://www.census.gov/acs/www/methodology/coverage_rates_data/index.php
- 7. "Stata Web Books Regression with Stata Chapter 2 Regression Diagnostics." *UCLA Institute for Digital Research and Education*, 2015. Accessed April 7, 2015. http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm

VII. Appendix

Simple Regression STATA Output

Source	SS	df	MS		Number of obs		
Model Residual	68276510.2 2589836991		76510.2 5.165015	Prob > F = R-squared		= 0.0000 = 0.2086	
Total	3272602091	1568031 20	8.707742		Adj R-squared Root MSE	= 12.852	
wkhp	Coef.	Std. Err.	t	P> t	[95% Conf.	Interval]	
ladjpincp _cons	5.156952 -13.31999	.0080208 .0824114	642.95 -161.63	0.000	5.141232 -13.48151	5.172673 -13.15847	

Multiple Regression STATA Output

Source Model Residual	SS df MS 66363187.9 7 9480455.41 1562375501291290 120.993387	5 4
Total	2226007381291297 172.385391	
wkhp	Coef. Std. Err. t	P> t [95% Conf. Interval]
ladjpincp mid high npf fs wif lgasp _cons	6.803827 .0142856 476.27 -2.076793 .0315813 -65.76 -7.574277 .0541522 -139.87 2930213 .0083686 -35.01 4016238 .0351826 -11.42 6566068 .0159225 -41.24 .1893719 .0042967 44.07 -26.76996 .1388313 -192.82	0.000 -2.138691 -2.014895 0.000 -7.680414 -7.468141 0.0003094235276619 0.00047058053326671 0.00068781436253993 0.000 .1809505 .1977934

PAPER DUE? WHY SUFFER? THAT'S OUR JOB!

