Alpha Traore
Stat 6620
Pr. Naranjo

Report I

Modeling of the Working Hours using Multiple Linear Regression Method

In Economic theories the basic model of labor supply, the aggregated number of working hours, is related to the repartition of each individual time between leisure and salary work. So, people choose between selling their time to institutions or spend their time in other non-remunerated activities. Therefore, the wage is the opportunity cost of the variable we call "time people don't want to work for money". In that context, the number of hours people are willing to work is positively related, ceteris paribus, to the money they can get per hours. In others words if the wage rates is getting high, and all other variables are held constants, most people will think that it reasonable to go to work than doing others non-lucrative activities. It is then known that hourly wage is not the only important factor in predicting the labor supply.

The data we study are from a national sample of 6000 households with a male head earning less than $15,000 annually in 1966. The households were classified into 39 demographic groups (e.g. factory workers, farmers, school teachers, etc.).

The objectives of this study are, firstly, the exploration of the simple linear relationship between the labor supply and the average hourly wage. Secondly, to build a model for predicting the labor supply (average hours worked) during one year based on some factors such as: average hourly wage, average yearly earnings of spouse, average yearly earnings of other family members, average yearly non-earned income, average family asset holdings (Bank account, etc.),

average age of respondent, average number of dependents, percent of white respondents, average highest grade of school completed.

The Linear Regression approach is one of the useful method for studying the relationship of the independent variables between each other and their relationships to the dependent variable which is the labor supply. The fitted model is erected based on some analyses and interpretations of SAS multiple procedures.

Let called the variables as follows:

Y --HRS: Average hours worked during the year

X1--WAGE: Average hourly wage ($)

X2--ERSP: Average yearly earnings of spouse ($)

X3--ERNO: Average yearly earnings of other family members ($)

X4--NEIN: Average yearly non-earned income

X5--ASSET: Average family asset holdings (Bank account, etc.) ($)

X6--AGE: Average age of respondent

X7--DEP: Average number of dependents

X8--RACE: Percent of white respondents

X9--SCHOOL: Average highest grade of school completed

Logy = Log(Y), transformation of the variable Y to its logarithm

The full regression model is presented in this form:

$$Yi = \beta 0 + \beta 1Xi1 + \beta 2Xi2 + \beta 3Xi3 + \beta 4Xi4 + \beta 5Xi5 + \beta 6Xi6 + \beta 7Xi7 + \beta 8Xi8 + \beta 9Xi9 + \varepsilon i$$

- Yi ; is the value of the Average hours worked during the year of one of the 39 demographics groups. i=1,2,…,39;
- β0, β1, β2, β3, β4, β5, β6, β7, β8, and β9 are parameters;
- Xi are known constants, namely, the value of the predictor variable in the itt group, i=1, 2,…,39;
- εi is the random error term of the ith group,
- We assume that the expected value of the error terms εi is zero: E(εi)=0, and its variance is constant.
- In the model building process we take $\hat{Y}i$= log(yi) (the natural logarithm) as the value of the value labor supply. We this function because the regression in logarithmic scale give a better linear relationship between the dependent the independent variables in the model.

So, our model has the following form:

$\hat{Y}i$= log(yi) =b0+b1X1i+b2X2i+b3X3i+b4X4i+b5X5i+b6X6i+b7X7i+b8X8i+b9X9i

b0, b1… b9 are the estimated value of the parameters β1… β9.

Table 1: Potential Predictor Variables and logarithmic transformation of the labor supply.

Table 1: Potential Predictor Variables and logarithmic transformation of the labor supply.											
Obs	**Y**	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**X9**	**Logy**
1	2157	2.905	1121	291	380	7250	38.5	2.340	32.1	10.5	7.67647
2	2174	2.970	1128	301	398	7744	39.3	2.335	31.2	10.5	7.68432
3	2062	2.350	1214	326	185	3068	40.1	2.851	.	8.9	7.63143
4	2111	2.511	1203	49	117	1632	22.4	1.159	27.5	11.5	7.65492
5	2134	2.791	1013	594	730	12710	57.7	1.229	32.5	8.8	7.66575
6	2185	3.040	1135	287	382	7706	38.6	2.602	31.4	10.7	7.68937
7	2210	3.222	1100	295	474	9338	39.0	2.187	10.1	11.2	7.70075

Table 1: Potential Predictor Variables and logarithmic transformation of the labor supply.											
Obs	**Y**	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**X9**	**Logy**
8	2105	2.493	1180	310	255	4730	39.9	2.616	71.1	9.3	7.65207
9	2267	2.838	1298	252	431	8317	38.9	2.024	9.7	11.1	7.72621
10	2205	2.356	885	264	373	6789	38.8	2.662	25.2	9.5	7.69848
11	2121	2.922	1251	328	312	5907	39.8	2.287	51.1	10.3	7.65964
12	2109	2.499	1207	347	271	5069	39.7	3.193	.	8.9	7.65397
13	2108	2.796	1036	300	259	4614	38.2	2.040	.	9.2	7.65349
14	2047	2.453	1213	297	139	1987	40.3	2.545	.	9.1	7.62413
15	2174	3.582	1141	414	498	10239	40.0	2.064	.	11.7	7.68432
16	2067	2.909	1805	290	239	4439	39.1	2.301	.	10.5	7.63385
17	2159	2.511	1075	289	308	5621	39.3	2.486	43.6	9.5	7.67740
18	2257	2.516	1093	176	392	7293	37.9	2.042	.	10.1	7.72179
19	1985	1.423	553	381	146	1866	40.6	3.833	.	6.6	7.59337
20	2184	3.636	1091	291	560	11240	39.1	2.328	13.6	11.6	7.68891
21	2084	2.983	1327	331	296	5653	39.8	2.208	58.4	10.2	7.64204
22	2051	2.573	1194	279	172	2806	40.0	2.362	77.9	9.1	7.62608
23	2127	3.262	1226	314	408	8042	39.5	2.259	39.2	10.8	7.66247
24	2102	3.234	1188	414	352	7557	39.8	2.019	29.8	10.7	7.65064
25	2098	2.280	973	364	272	4400	40.6	2.661	53.6	8.4	7.64874
26	2042	2.304	1085	328	140	1739	41.8	2.444	83.1	8.2	7.62168
27	2181	2.912	1072	304	383	7340	39.0	2.337	30.2	10.2	7.68754
28	2186	3.015	1122	30	352	7292	37.2	2.046	29.5	10.9	7.68983
29	2108	2.786	1757	.	506	9658	43.4	.	32.6	10.2	7.65349
30	2188	3.010	990	366	374	7325	38.4	2.847	30.9	10.6	7.69074
31	2203	3.273	.	.	430	8221	38.2	2.324	22.1	11.0	7.69758
32	2077	1.901	350	209	95	1370	37.4	4.158	61.3	8.2	7.63868
33	2196	3.009	947	294	342	6888	37.5	3.047	31.8	10.6	7.69439
34	2093	1.899	342	311	120	1425	37.5	4.512	62.8	8.1	7.64635
35	2173	2.959	1116	296	387	7625	39.2	2.342	31.0	10.5	7.68386
36	2179	2.971	1128	312	397	7779	39.4	2.341	31.2	10.5	7.68662
37	2200	2.980	1126	204	393	7885	39.2	2.341	31.0	10.6	7.69621
38	2052	2.630	.	.	154	3331	40.5	.	45.8	10.3	7.62657
39	2197	3.413	1078	300	512	10450	39.1	2.297	15.5	11.3	7.69485

Model 1: With one predictor variable, simple regression

This model is the regression of the labor supply against only the potential predictor variables average hourly wage. It shows us the relationship between the two variables. To regress our model with X1 only while the others variables are holding constant in the model. The simple regression function form is:

log(yi) =b0+b1Xi

In the process of studying this relationship, the two following hypotheses are tested:

H0 (Null hypothesis): There is not a linear association between the labor supply and average hourly wage (the slope b1=0)

Ha (Alternative hypothesis): There is a linear association between the labor supply and average hourly wage (the slope b1≠0)

Table 2 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.01169	0.01169	19.10	<.0001
Error	37	0.02264	0.00061180		
Corrected Total	38	0.03432			

Table 3 Pearson Correlation Coefficients, N = 39
Prob > |r| under H0: Rho=0

	Logy	X1
Logy	**1.00000**	**0.58352** **<.0001**
X1	**0.58352** **<.0001**	**1.00000**

Table4 Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.56040	0.02469	306.24	<.0001
X1	1	0.03842	0.00879	4.37	<.0001

The simple regression model is: log(Y) = 0.03842*X1+7.56040

The results of calculations presented in these tables give some useful information. In Table 2 the F-test (F=0.01169/0.00061180) shows that the relationship is statistically significant (p-value <.0001). That means the rejection of the null hypothesis and we favor the alternative, there is a significant linear association between the labor supply and average hourly wage. In the table 4, the estimated parameter b1 is a positive number, so the association between the two variable is positive. In other word, the labor supply increases when the wage rates increase, and vice versa.

The value of the estimated coefficient of the wage rates variable is b1= 0.03842. This value would be interpreted as: The increase of one unit in the wage rate would increase of approximately (expo (0.03842)-1) = 0.039168 percentage in the labor supply.

The table 3 shows that the coefficient of correlation between the labor supply and the wage rate is 0.58352, and the correlation p-value is significant.

This low value of R-square is not very helpful for the prediction purpose. However, based on the F-test a variation in the wage rate value is positively associated with the variation of the labor supply.

Model 2: Multiple variables regression

To study the multiple association between these nine potential predictors and the labor supply, first, we observe the full model with all the variables. Secondly, we analyze the correlation between variables. Finally, we build the multi linear model that we think should be the best.

Table 5

Root MSE	0.00708	**R-Square**	0.9482
Dependent Mean	7.67268	**Adj R-Sq**	0.9223
Coeff Var	0.09222		

Table 6

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	7.52174	0.10185	73.85	<.0001
X1	1	-0.04341	0.01795	-2.42	0.0264
X2	1	0.00008995	0.00001807	4.98	<.0001
X3	1	-0.00007878	0.00002715	-2.90	0.0095
X4	1	0.00012714	0.00014115	0.90	0.3796
X5	1	-0.00000159	0.00000839	-0.19	0.8521
X6	1	0.00132	0.00135	0.98	0.3403
X7	1	0.03348	0.00594	5.64	<.0001
X8	1	-0.00115	0.00032188	-3.57	0.0022
X9	1	0.00767	0.00908	0.85	0.4090

Table 7 Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations										
	Logy	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**X9**
Logy	1.00000 39	0.58352 <.0001 39	0.08188 0.6299 37	-0.23959 0.1593 36	0.67907 <.0001 39	0.71280 <.0001 39	-0.10171 0.5378 39	-0.34648 0.0357 37	-0.83331 <.0001 31	0.67004 <.0001 39
X1	0.58352 <.0001 39	1.00000 39	0.51009 0.0013 37	0.05804 0.7367 36	0.70240 <.0001 39	0.77760 <.0001 39	0.03149 0.8491 39	-0.59195 0.0001 37	-0.66363 <.0001 31	0.88397 <.0001 39
X2	0.08188 0.6299 37	0.51009 0.0013 37	1.00000 37	-0.04101 0.8123 36	0.28799 0.0839 37	0.31578 0.0569 37	0.05029 0.7675 37	-0.69188 <.0001 36	-0.23307 0.2237 29	0.50526 0.0014 37
X3	-0.23959 0.1593 36	0.05804 0.7367 36	-0.04101 0.8123 36	1.00000 36	0.35044 0.0361 36	0.28378 0.0935 36	0.77546 <.0001 36	0.05000 0.7721 36	0.11040 0.5760 28	-0.29212 0.0838 36
X4	0.67907 <.0001 39	0.70240 <.0001 39	0.28799 0.0839 37	0.35044 0.0361 36	1.00000 39	0.98788 <.0001 39	0.47490 0.0023 39	-0.51831 0.0010 37	-0.68746 <.0001 31	0.53902 0.0004 39
X5	0.71280 <.0001 39	0.77760 <.0001 39	0.31578 0.0569 37	0.28378 0.0935 36	0.98788 <.0001 39	1.00000 39	0.39950 0.0117 39	-0.50967 0.0013 37	-0.73912 <.0001 31	0.63048 <.0001 39
X6	-0.10171 0.5378 39	0.03149 0.8491 39	0.05029 0.7675 37	0.77546 <.0001 36	0.47490 0.0023 39	0.39950 0.0117 39	1.00000 39	-0.04635 0.7853 37	0.10079 0.5895 31	-0.31605 0.0500 39
X7	-0.34648 0.0357 37	-0.59195 0.0001 37	-0.69188 <.0001 36	0.05000 0.7721 36	-0.51831 0.0010 37	-0.50967 0.0013 37	-0.04635 0.7853 37	1.00000 37	0.39128 0.0358 29	-0.59861 <.0001 37
X8	-0.83331 <.0001 31	-0.66363 <.0001 31	-0.23307 0.2237 29	0.11040 0.5760 28	-0.68746 <.0001 31	-0.73912 <.0001 31	0.10079 0.5895 31	0.39128 0.0358 29	1.00000 31	-0.78387 <.0001 31
X9	0.67004 <.0001 39	0.88397 <.0001 39	0.50526 0.0014 37	-0.29212 0.0838 36	0.53902 0.0004 39	0.63048 <.0001 39	-0.31605 0.0500 39	-0.59861 <.0001 37	-0.78387 <.0001 31	1.00000 39

The table 5 (R-square=0.9482) give a strong linear association between all the predictor variables in the full model and the labor supply. That mean 94.82 % of the labor supply variation is explained by the multi linear model below:

$$\log(Yi) = 7.52174 - 0.04341X1i + 0.00008995X2i - 0.00007878X3i + 0.00012714X4i - 0.00000159X5i + 0.00132X6i + 0.03348X7i - 0.00115X8i + 0.00767X9i$$

This model is very good for predicting the labor supply based on the potential predictors in this study. However, this model is a "good model" and we should drop some predictor variables for the reasons we are going to give by interpreting the analyses of the table 6 and 7.

The analysis of the table 6 show some interesting result of the test concerning the parameters (β1…., β9). The coefficients of X4, X5, X6, and X9 are not significant. Also, some of the parameter values are negative (the coefficients of X1, X3, X5, and X8 negative). This negative slopes of some parameters shown in this table are explained by the effect of mutual relationships between X4, X5, X6, and X9 and the other predictor variables, and also to their correlation with the response variable (Y). In addition, this table shows that the p-value of the estimated parameters, b4, b5, b6 and b9 are >0.05. We cannot reject the null hypothesis for any those regression coefficients base on this test.

The table 7 shows that there are some mutual correlations between some predictor variables. For example, X1 has a pairwise linear association with X2, X4, X5, and X9. X9 has pairwise association with X2, X4, X5 and X8. Also the linear association between X4 and X5 is very strong. The effect of these inter correlations between predictor variables would be the creation of wrong coefficients in the multivariate model.

Table 7

Table 8 Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Type I SS**	**Type II SS**
Intercept	1	7.52174	0.10185	73.85	<.0001	1648.36078	0.27306
X1	1	-0.04341	0.01795	-2.42	0.0264	0.00477	0.00029275
X2	1	0.00008995	0.00001807	4.98	<.0001	0.00048094	0.00124
X3	1	-0.00007878	0.00002715	-2.90	0.0095	0.00085811	0.00042138
X4	1	0.00012714	0.00014115	0.90	0.3796	0.00534	0.00004062
X5	1	-0.00000159	0.00000839	-0.19	0.8521	0.00179	0.00000179
X6	1	0.00132	0.00135	0.98	0.3403	0.00129	0.00004803
X7	1	0.03348	0.00594	5.64	<.0001	0.00093383	0.00159
X8	1	-0.00115	0.00032188	-3.57	0.0022	0.00099451	0.00063682
X9	1	0.00767	0.00908	0.85	0.4090	0.00003577	0.00003577

Table 8

Number in Model	R-Square	C(p)	AIC	BIC	Variables in Model
1	0.7569	60.4591	-242.3484	-243.9161	X8
1	0.4639	162.2326	-220.2080	-223.0504	X5
2	0.7828	53.4395	-243.5106	-246.0746	X7 X8
2	0.7605	61.1890	-240.7727	-243.6334	X3 X8
3	0.8300	39.0431	-248.3724	-251.1315	X2 X7 X8
3	0.7935	51.7349	-242.9205	-246.5413	X5 X7 X8
4	0.8660	28.5366	-253.0372	-255.3378	X2 X7 X8 X9
4	0.8600	30.6338	-251.8030	-254.4057	X1 X2 X7 X8
5	0.9107	15.0302	-262.3850	-261.5759	X1 X2 X5 X7 X8
5	0.9078	16.0359	-261.4919	-261.0354	X1 X2 X4 X7 X8
6	0.9450	5.1138	-273.9524	-265.8760	X1 X2 X3 X4 X7 X8
6	0.9429	5.8510	-272.8928	-265.4827	X1 X2 X3 X5 X7 X8
7	0.9460	6.7505	-272.4897	-263.0837	X1 X2 X3 X4 X6 X7 X8
7	0.9453	6.9966	-272.1246	-262.9907	X1 X2 X3 X4 X7 X8 X9
8	0.9481	8.0358	-271.5779	-260.2439	X1 X2 X3 X4 X6 X7 X8 X9
8	0.9461	8.7146	-270.5434	-260.1049	X1 X2 X3 X4 X5 X6 X7 X8
9	0.9482	10.0000	-269.6335	-257.1397	X1 X2 X3 X4 X5 X6 X7 X8 X9

The analysis of the table 7 shows that the four smaller Regression Sum of Squares Type II are: X4, X5, X6, and X9. That means putting each of these variables in the model given that all the others variables are already in this model doesn't diminish significantly the Error Sum of Square. So, these variables should not have a big impact on the reduced model we want to build.

The table 8 shows that the best two models, containing six predictor variables, regarding the bigger R-Square value, and the smaller AIC, and BIC are:

1) Log(Y) versus X1 X2 X3 X4 X7 X8
2) Log (Y) versus X1 X2 X3 X5 X7 X8

Table 9

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	0.01643	0.00274	60.11	<.0001
Error	21	0.00095690	0.00004557		
Corrected Total	27	0.01739			

Table 10 Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.61620	0.02086	365.06	<.0001
X1	1	-0.03754	0.00678	-5.54	<.0001
X2	1	0.00009442	0.00001422	6.64	<.0001
X3	1	-0.00007776	0.00002064	-3.77	0.0011
X4	1	0.00013460	0.00002376	5.67	<.0001
X7	1	0.03329	0.00432	7.71	<.0001
X8	1	-0.00113	0.00013326	-8.47	<.0001

Table 11			
Root MSE	0.00675	**R-Square**	0.9450
Dependent Mean	7.67268	**Adj R-Sq**	0.9293
Coeff Var	0.08798		

The first model should be the good model in our study of the relationship between the average hours worked during the year and the potential predictors average hourly wage, average yearly earnings of spouse, average yearly earnings of other family members, average yearly non-earned income, average number of dependents, and percent of white respondents.

Final Model:

Log (Yi) = 7.61620-0.03754X1i+0.00009442X2i-0.00007776X3i+0.00013460X4i

+0.03329X5i-0.00113X6i

The F test is significant p-value<0.0001

R-Square=0.9450: 94.50% of the labor supply variation is explained by our model.

All the parameters are significant at a level 0.001.