

DRAFT NOT FOR QUOTATION

**PAPER PRESENTED AT THE
MISSOURI VALLEY ECONOMIC ASSOCIATION MEETING
ST. LOUIS, MO.
FEBRUARY 24-26, 2000**

**A SIMPLE METHOD OF DETERMINING THE ECONOMIC
IMPORTANCE OF VARIABLES IN MULTIPLE REGRESSION ANALYSIS**

BY

**RALPH J. BROWN
PROFESSOR OF ECONOMICS
UNIVERSITY OF SOUTH DAKOTA**

A SIMPLE METHOD OF DETERMINING THE ECONOMIC IMPORTANCE OF VARIABLES IN MULTIPLE REGRESSION ANALYSIS

**By
Ralph J. Brown***

INTRODUCTION

In evaluating the results of multiple regression analysis the usual method is to examine the R-squared, the Durbin-Watson statistic (if time series), the sign and magnitude of the regression coefficients, and the t-statistics. All too frequently, the most important task is to determine whether a particular regression coefficient is statistically different from zero or not. If the coefficient is, that variable is considered to be an important variable. If it isn't it is considered to be unimportant. Following this approach, the scientific importance of any single explanatory variable rests on the statistical significance of the regression coefficient. Some go as far as to describe the variable with the highest level of statistical significance as the most important variable.

The elevation of statistical significance to such importance has been severely criticized by D. McCloskey.¹ Basically, she argues that the statistical significance of the individual regression coefficient estimate has come to be regarded as the scientific significance of that particular variable in explaining variation in the dependent variable. She goes on to argue that by focussing so much attention on statistical significance we ignore the economic significance of the estimate. She argues that the magnitude of the coefficient tells us the extent to which a change in X will cause a change in Y, while statistical significance is a statistical measure about the ratio of the estimate to its

* Professor of Economics of the University of South Dakota, Vermillion, SD 57069.

¹ Deirdre N. McCloskey, *The Vices of Economists* -

standard error and provides no information as to the true scientific or economic importance of an explanatory variable.

Two recent textbooks in econometrics while not necessarily agreeing with the insignificance of statistical significance do tend to agree that more attention should be paid to the economic significance of the variable than has been the case in the past where more emphasis was placed on statistical significance. For instance, Wooldridge in his text *Introductory Econometrics*² (2000) states: "Too much focus on statistical significance can lead to the false conclusion that a variable is 'important' for explaining y even though its estimated effect is modest."³ Goldberger in his text *Introductory Econometrics*⁴ (1998) states:

"We should not confuse the magnitude of a coefficient with the ratio of the coefficient to its standard error. We should not confuse economic importance with statistical significance."⁵

In a widely used undergraduate econometrics text by Studenmund⁶ he states that the t-test does not test "importance" but rather indicates the likelihood that a particular sample result could have been obtained by chance.

Each of these authors argue that we should give more attention to the economic importance of the explanatory variable than is typically the case in much empirical research. This short note is written in the spirit of that sentiment. What is presented here is a simple but useful tool for getting more economic information out of our econometric estimates. The method is embarrassingly simple but I have not seen it

² Jeffrey M. Wooldridge, *Introductory Econometrics*, USA; South-Western Publishing Co., 2000.

³ Ibid, p. 131.

⁴ Arthur S. Goldberger, *Introductory Econometrics*, Cambridge, MA: Harvard University Press, 1998.

⁵ Ibid, p. 73.

⁶ A. H. Studenmund, *Using Econometrics*, New York: Addison-Wesley, 3rd edition, 1997, p. 155.

used anywhere in the literature and it does help us to determine in a particular context which variables account for the proportional movement in the dependent variable.

DETERMINING THE ECONOMIC IMPORTANCE OF ECONOMETRIC ESTIMATES

In determining the economic importance of econometric coefficient estimates the first place to start is to check the **sign and magnitude** of the coefficient estimates. Is the sign correct and does the magnitude of the coefficient indicate a large or small change in the dependent variable for a one-unit change in the independent variable? Obviously this is important information. Another approach to extract more information about the economic significance of the variable is to calculate **elasticity** estimates. Finally, standardized coefficients or what are sometimes called **beta coefficients** can be calculated. The standardized coefficients are calculated as the ratio of the regression coefficient to the standard error of the coefficient. It measures the standard deviation changes in the dependent variable from a one standard error change in the independent variable.

While all of these methods are useful and should be used in any econometric investigation there is an additional tool that can be used to evaluate the economic importance of each independent variable. For instance, it would be useful to know how important is the variation in each X in explaining the variation in Y in this particular research context. That is, what proportion of the in Y from its mean can be explained by the variation of each X from its mean. In other words, in cross-sectional analysis one might be interested in knowing how much of the variation in Y from the average person, city, state, or country can be explained by deviation in each of the Xs. Obviously, the

variation in some of the Xs will explain a large proportion of the variation in Y while others will not. This is information that can not be gleaned by looking at the magnitude of the coefficient or its statistical significance. In a time series analysis one might want to know how much of the movement in Y over the sample time period is due to movement in each X. Once again, just looking at the coefficient or its statistical significance tells us little in this regard. What we need is a methodology that both captures the sign and magnitude of the coefficient and the magnitudes of the movement in each X all in one magnitude.

Simple Tool

I am not sure what to call this tool. It might be called the deviation adjusted coefficient method. However, we will proceed without naming it. Basically this tool involves multiplying each estimated regression coefficient by the X deviation from the mean and then dividing by the total Y deviation from the mean to obtain the percentage difference in Y from the mean that can be accounted for by each X variables deviation from the mean. Assume we are interested in explaining the difference between the lowest Y observation and the mean Y. The equation for the fitted value of the lowest Y is shown below.

$$\hat{Y}_L = B_0 + B_1 X_{1L} + B_2 X_{2L} + \dots + B_k X_{kL}$$

Express in deviation from the mean form:

$$\hat{Y}_L - \bar{Y} = B_1(X_{1L} - \bar{X}_1) + B_2(X_{2L} - \bar{X}_2) + \dots + B_k(X_{kL} - \bar{X}_k).$$

Dividing by $\hat{Y}_L - \bar{Y}$ we have:

$$1.0 = \frac{\hat{B}_1(X_{1L} - \bar{X}_1) + \hat{B}_2(X_{2L} - \bar{X}_2) + \dots + \hat{B}_k(X_{kL} - \bar{X}_k)}{\hat{Y}_L - \bar{Y}}$$

Expressed in percentage terms we have:

$$100\% = \% \text{ Variation Due to } X_1 + \% \text{ Variation Due to } X_2 + \dots + \% \text{ Variation Due to } X_k$$

This method may be applied in other contexts. In this example, we are attempting to determine to what extent the Y deviation from the mean can be explained by the Xs and their coefficients. One could apply this to the high and low Y or any pair of Y observations that we are interested in.

A Cross-Section Example

Assume we have estimated a regression that is designed to explain differences in the average annual pay (wage) for all 50 states. Cross-section and time series data is pooled for the years 1993-97 which provides 250 observations. The dependent variable is the natural logarithm of average annual pay and the independent variables are variables grouped under the following categories:

1. worker characteristics,
2. job characteristics,
3. cost of living,
4. metropolitan influence,
5. fiscal environment,
6. amenities/disamenities.

The means for each of the variables are shown in Table 1.

TABLE 1: MEAN AND SOUTH DAKOTA VALUE FOR VARIABLES

VARIABLE LABEL	BRIEF DESCRIPTION	Source	1997 MEAN	SD VALUE
WAGE \$	Average Annual Pay	BLS	\$28,135	\$21,645
PARTTIME %	Part-time worker %	BLS	18.57	20.50
HIGHSCHOOL %	High School Graduate %	Census	31.08	33.70
ASSOCIATEDEG %	Associate Degree %	Census	7.40	6.30
BACHELORS %	Bachelor's Degree %	Census	12.96	12.30
GRADUATE %	Graduate Degree %	Census	6.79	4.90
WOMENWORKER %	Women Worker %	BLS	46.66	48.00
BLACKWORKER %	Black Worker %	BLS	8.74	1.00
MEDAGE	Median Age	BLS	35.06	35.20
INDUSTRYMIX \$	Industry Mix Wage	Computed	\$30,241	\$30,144
OCCUPATIONMIX \$	Occupation Mix Wage	Computed	\$28,363	\$27,316
FIRMSIZE	Employees Per Establishment	CBP	14.63	12.00
FRINGEBENEFIT %	Fringe Benefit %	Computed	17.12	16.06
UNION %	Union Membership %	BNA	12.82	6.90
UNRATE %	Unemployment Rate %	BLS	4.71	3.10
COLI %	Cost of Living Index	Harvard	0.99	0.93
METRO %	Metropolitan Population %	Census	67.25	33.30
TAXBURDEN %	Tax Burden %	BEA	13.94	10.10
EDQUALINDEX	Education Quality Index	Computed	100.00	93.04
HWYEXPCAPITA \$	Highway Expenditures Per Capita	Census	\$361	\$526
WELFEXPCAPITA \$	Welfare Expenditures Per Capita	Census	\$693	\$559
SUNNYDAYS	Sunny Days	USNOAA	56.84	57.00
HEATDEGREE	Heating Degree Days	USNOAA	5090.90	7809.00
COOLDEGREE	Cooling Degree Days	USNOAA	1215.28	744.00
LIVABILITY	Livability Index	Morgan	25.66	29.62
COAST	Coast Location	Computed	0.58	0.00
VIOLENTCRIME	Violent Crime Rate	Census	493.10	197.40

The general form of the regression equation was:

$$(1) \ln W_{it} = B_0 + B_1 WC_{it} + B_2 JC_{it} + B_3 C_{it} + B_4 M_{it} + B_5 F_{it} + B_7 A_{it} + u_{it}$$

where:

W_{it} = average annual pay variable for i th state in year t ,

WC_{it} = vector of employee (worker) characteristics,

JC_{it} = vector of job characteristics,

C_{it} = cost of living,

M_{it} = metropolitan influence,

F_{it} = vector of fiscal environment variables,

A_{it} = vector of amenity/disamenity variables,

u_{it} = random error term.

i = state where $i = 1, \dots, 50$,

t = Year where $t = 1993, \dots, 1997$.

The regression results are shown in Table 2. Since we are mainly interested in the techniques developed in this paper we will not give the regression itself much attention other than to say it did a good job of explaining wage differences between states and the coefficients all had the expected signs, though not all were statistically significant. Assume that we want to determine what factors explain the difference in wages between the lowest-wage state (South Dakota) and the average of all states. Obviously, this is just one of many possibilities. We could examine the difference between the low and high-wage state (Connecticut) or between any two states or groups of states.

TABLE 2: REGRESSION RESULTS

Dependent Variable: LOG(WAGE?)

Method: Pooled Least Squares

Date: 02/17/00 Time: 15:55

Sample: 1993 1997

Included observations: 5

Total panel observations 250

White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	8.021075	0.288576	27.79536	0.0000
PARTTIME?	-0.009248	0.001674	-5.524164	0.0000
HIGHSCHOOL?	-0.001978	0.001254	-1.576710	0.1163
ASSOCIATEDEG?	0.003658	0.003033	1.206305	0.2290
BACHELORS?	0.008068	0.002555	3.156991	0.0018
GRADUATE?	0.015869	0.003353	4.732766	0.0000
WOMENWORKER?	-0.008572	0.002869	-2.987669	0.0031
BLACKWORKER?	-0.000467	0.000602	-0.775794	0.4387
MEDAGE?	0.000476	0.002301	0.206836	0.8363

INDUSTRYMIX?	5.03E-05	7.79E-06	6.455793	0.0000
OCCUPATIONMIX?	1.00E-05	6.25E-06	1.602509	0.1105
FIRMSIZE?	0.010097	0.002188	4.613677	0.0000
FRINGEBENEFIT?	-0.002807	0.002016	-1.392780	0.1651
UNION?	0.004498	0.000856	5.257089	0.0000
UNRATE?	0.012768	0.003165	4.033739	0.0001
COLI?	0.248246	0.073980	3.355595	0.0009
METRO?	0.001437	0.000354	4.059214	0.0001
TAXBURDEN?	0.010093	0.002496	4.043982	0.0001
EDQUALINDEX?	0.001032	0.000396	2.606723	0.0098
HWYEXPCAPITA?	1.00E-05	3.71E-05	0.270320	0.7872
WELFEXPCAPITA?	1.72E-05	2.38E-05	0.720245	0.4721
SUNNYDAYS?	-0.001613	0.000421	-3.834462	0.0002
HEATDEGREE?	8.41E-06	4.28E-06	1.965361	0.0506
COOLDEGREE?	1.22E-05	6.23E-06	1.958717	0.0514
LIVABILITY?	-0.000426	0.000519	-0.821574	0.4122
COAST?	-0.015818	0.007907	-2.000587	0.0467
VIOLENTCRIME?	5.86E-05	1.82E-05	3.221768	0.0015
D93?	0.055254	0.031324	1.763953	0.0791
D94?	0.067950	0.025643	2.649864	0.0086
D95?	0.059624	0.019528	3.053177	0.0025
D96?	0.037343	0.011433	3.266289	0.0013
R-squared	0.956938	Mean dependent var	10.15937	
Adjusted R-squared	0.951039	S.D. dependent var	0.147225	
S.E. of regression	0.032577	Sum squared resid	0.232411	
F-statistic	162.2232	Durbin-Watson stat	0.462629	
Prob(F-statistic)	0.000000			

The estimated regression model was used to account for the South Dakota/U.S. average wage gap. To determine the causes of the wage gap between South Dakota and the nation we apply the regression coefficient estimates to the South Dakota/all-state differential ($X_{SDj} - X_{AVGj}$) for each of the variables in the regression. Since the dependent variable in this case is in logs we can interpret the coefficient as the percent change in Y from a one-unit change in X. Likewise the difference between Ys are percent differences. The percentage of the pay gap explained by the variable X_j South Dakota/All-State Average differential is calculated as:

$$\begin{aligned} &\% \text{ Pay Gap} \\ &\text{Explained by } X_j = b_j \times ((X_{SDj} - X_{AVGj}))/\%Y \text{ Gap} \\ &\text{Differential} \end{aligned}$$

where:

b_j = estimated regression coefficient for variable j ,

X_{SDj} = South Dakota value for variable j ,

X_{AVGj} = All-state average for variable j .

We perform the same analysis for each X to determine its contribution to explaining the Y deviation. The decomposition of the pay gap is presented in Table 3. The regression coefficient estimates are presented in column (1). The regression coefficient estimates are taken from regression equation based on the equation (1) shown above. Columns (2) and (3) show the South Dakota value and all-state average value for each of the explanatory variables. The percent of the wage gap explained by each explanatory variable is shown in column (4). Columns (5) and (6) show the percent of the gap explained by the aggregated characteristics.

TABLE 3: ACCOUNTING FOR THE PAY GAP
Dependent Variable = Average Annual Pay for 50 States

	(1)	(2)	(3)	(4)	(5)	(6)
Independent Variable	Regression Coefficient	SD Value	All-State Average	% Gap Explained	% Gap By Category	Category
PARTTIME?	-0.009248	20.5	18.6	7%		
HIGHSCHOOL?	-0.001978	33.7	31.1	2%		
ASSOCIATEDEG?	0.003658	7.4	6.3	-2%		
BACHELORS?	0.008068	12.3	13.0	2%		
GRADUATE?	0.015869	4.9	6.8	12%		
WOMENWORKER?	-0.008572	48.0	46.7	5%		
BLACKWORKER?	-0.000467	1.0	8.7	-1%		
MEDAGE?	0.000476	35.2	35.1	0%	25%	worker characteristics
INDUSTRYMIX?	5.03E-05	\$30,144	\$30,241	2%		
OCCUPATIONMIX?	1.00E-05	\$27,316	\$28,363	4%		
FIRMSIZE?	1.01E-02	12.00	14.63	11%		
FRINGEBENEFIT?	-0.002807	16.06	17.12	-1%		
UNION?	0.004498	6.90	12.82	11%		
UNRATE?	0.012768	3.10	4.71	8%	34%	job characteristics
COLI?	0.248246	0.93	0.99	6%	6%	cost of living

METRO?	0.001437	33.30	67.25	19%	19%	metropolitan influence
TAXBURDEN?	0.010093	10.10	13.94	15%		
EDQUALINDEX?	0.001032	93.04	100.00	3%		
HWYEXPCAPITA?	1.00E-05	\$526	\$361	-1%		
WELFEXPCAPITA?	1.72E-05	\$559	\$693	1%	18%	fiscal setting
SUNNYDAYS?	-1.61E-03	57	56.84	0%		
HEATDEGREE?	8.41E-06	7,809	5,091	-9%		
COOLDEGREE?	1.22E-05	744	1,215	2%		
LIVABILITY?	-4.26E-04	30	26	1%		
COAST?	-0.015818	0.00	0.58	-4%		
VIOLENTCRIME?	5.86E-05	197	493	7%	-3%	amenity/disamenity
VIOLENTCRIME?	5.86E-05			100.0%	100.0%	

As shown in Table 3, the percent of the pay gap accounted for by each characteristic is as follows:

<u>Characteristic</u>	<u>% Gap Explained</u>
Worker Characteristics	25%
Job Characteristics	34%
Cost of Living	6%
Metropolitan Influence	19%
Fiscal Environment	18%
Amenities/Disamenities	-3%
Total	100%

A negative percentage indicates that this characteristic actually raises South Dakota pay relative to the all-state average. This occurs because South Dakota's cold weather (heating degree-days) is a disamenity that reduces the supply workers thereby raising wages and causing a negative gap.

The most important individual variables that explained the pay gap were as follows:

<u>Variable</u>	<u>% Gap Explained</u>
Metropolitan Influence	19%
Education (High School to Graduate)	15%
Tax Burden	15%
Firm Size	11%
Union Membership	11%
Unemployment Rate	8%
Cost of Living	6%

Part-Time	7%
Violent Crime (amenity)	<u>7%</u>
Total of These Variables	99%

Once again, this approach can be applied with many variations. For instance, in another application, not presented here, the differences in wages between South Dakota and each of the regional states of Iowa, Minnesota, Montana, Nebraska, North Dakota, and Wyoming was calculated. The point of all this, is that the information content of econometric analysis using this approach is much richer than would be available just by examining signs and magnitudes, elasticities, and standardized regression coefficients.

Time Series Example

The time series model presented herein is the estimation of demand curve for pork. Pork consumption per capita is shown in Figure 1. As shown in Figure 1, pork consumption is volatile with a slight downward trend since 1970. The basic model uses annual data for 1970-1997. A demand curve is estimated using a double log specification. No great claims are made for this regression other than it provides an example to show how the technique developed herein can be applied to time series analysis.

$$(2) \quad \ln(qpork_t) = B_0 + B_1 \ln(ppork_t/CPI_t) + B_2 \ln(dpi_t/CPI_t) + B_3 \ln(pbeef_t/CPI_t) + B_4 \ln(pchicken_t/CPI_t) + u_{it}$$

where:

$qpork_t$ = pork consumption per capita,

$ppork_t$ = price of pork from CPI index,

dpi_t = disposable income per capita,
 $pbeef_t$ = price of beef from CPI index,
 $pchicken_t$ = price of chicken from CPI index,
 t = Year where $t = 1990, \dots, 1997$.

The regression results are shown in Table 4. They indicate an own-price elasticity of -0.83, income elasticity of -0.20 (inferior good), cross-price elasticity of beef of 0.36 and for chicken of 0.17. The Adj R-squared is 0.856 and the Durbin-Watson statistic indicates some positive serial correlation. Once again this equation is used for illustrative purposes only.

FIGURE 1: PORK CONSUMPTION PER CAPITA

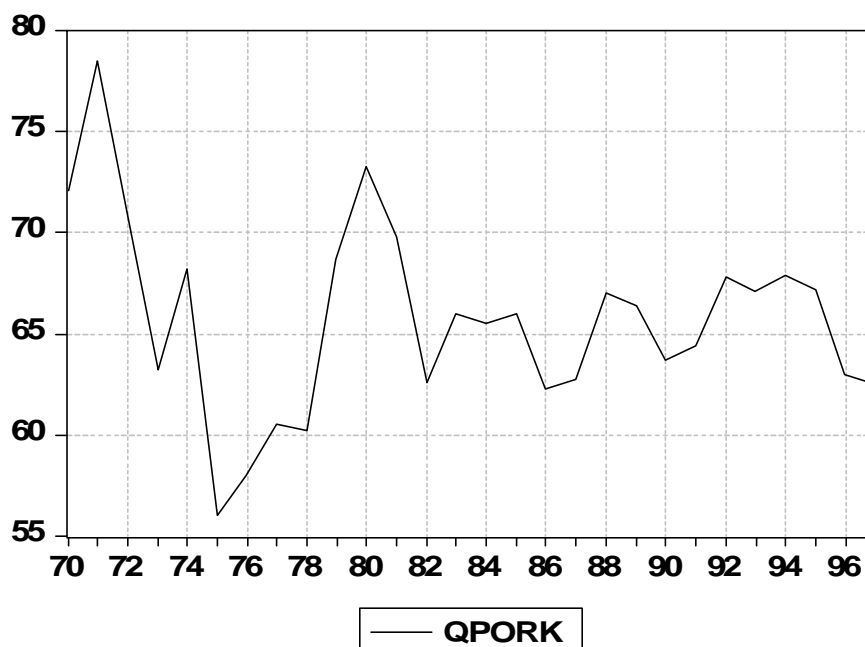


TABLE 4: REGRESSION RESULTS

Dependent Variable: LOG(QPORK)
 Method: Least Squares

Sample: 1970 1997
 Included observations: 28

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.147593	0.256491	20.06928	0.0000
LOG(PPORK/CPIURS)	-0.834210	0.082090	-10.16220	0.0000
LOG(DPI/CPIURS)	-0.201165	0.053214	-3.780296	0.0010
LOG(PBEEF/CPIURS)	0.364976	0.076612	4.763937	0.0001
LOG(PCHICKEN/CPIURS)	0.174669	0.085890	2.033645	0.0537
R-squared	0.876910	Mean dependent var		4.183597
Adjusted R-squared	0.855503	S.D. dependent var		0.071601
S.E. of regression	0.027217	Akaike info criterion		-4.209483
Sum squared resid	0.017038	Schwarz criterion		-3.971590
Log likelihood	63.93277	F-statistic		40.96370
Durbin-Watson stat	1.514370	Prob(F-statistic)		0.000000

The deviation model is applied using the difference between the 1997 and 1970 values for Y and Xs. In other words, we are trying to explain what factors account for the change in pork consumption over the 1970-97 period. Over the 1970-97 period, pork consumption fell by 14.2 percent. The results of this analysis are shown in Table 5.

As shown in Table 5, the price of pork variable, $\ln(\text{ppork}/\text{CPI})$, explained 6.4 percentage points or -45% of the change in pork consumption. The negative number indicates that if the only change had been the pork price change we would have expected pork consumption to rise since real pork prices fell over this period. The rise in disposable income combined with the negative (pork is an inferior good) coefficient shows that *ceteris paribus* pork consumption would have fallen 11 percentage points. The interpretation of the other coefficients is the same.

TABLE 5: CHANGE IN PORK CONSUMPTION ACCOUNTED FOR BY PRICE AND INCOME VARIABLES

		1970	1997		% Change
Variable	Coefficient (b)	Mean(M_{70})	Mean(M_{97})	$b \times (M_{97} - M_{70})$	Explained
$\ln(\text{ppork}/\text{CPI})$	-0.83421	4.690	4.614	0.064	-45.0%
$\ln(\text{dpi}/\text{CPI})$	-0.201165	9.448	9.997	-0.110	77.7%
$\ln(\text{pbeef}/\text{CPI})$	0.364976	4.647	4.483	-0.060	42.3%
$\ln(\text{pchicken}/\text{CPI})$	0.174669	4.834	4.630	-0.036	25.0%
			Total	-0.142	100.0%

SUMMARY AND CONCLUSIONS

This note has provided a methodology for extracting additional information from our regression results. The technique is easy to calculate and to understand and provides useful information. For instance, in the cross-section analysis we were able to determine what is different about a low-wage state and the average state that explains wage differences. The time series analysis allowed us to determine what factors account for changes in pork consumption over the 1970-97 time period. While this simple technique is no theoretical breakthrough it provides an additional tool for better understanding the meaning of our econometric results.

% Difference
Explained by $X_j = b_j \times ((X_{ij} - X_{AVGj})) / (Y_i - Y(\text{mean}))$
Differential

And the $\sum b_j \times ((X_{ij} - X_{AVGj})) = (Y_i - Y(\text{mean})) = 100\%$

where:

b_j = estimated regression coefficient for variable j ,

X_{ij} = value for i th observation on variable j ,

X_{AVGj} = average for variable j ,

Y_i = value for i th observation of the dependent variable.