Econometrics - Revisiting and developing (in progress)

Thiago Vinicius Gomiero

February 2024

1 Introduction

The linear regression analysis is based on the book I used, Econometric Methods by Jack Johnston, and the class notes from Econometrics 1 (and possibly 2) that I took at FEA-USP.

Several excerpts have been rewritten/adapted based on the main book (and class notes) with the aim of developing my writing skills in R and LaTeX.

Most of the statistical calculations were carried out in Rstudio with the same objective. Excel was also used to clean the database and integrate with Rstudio.

I will try to use some techniques and concepts presented in Quantitative Economics with R; A Data Science Approachlink. Its PDF can be easily found by searching online.

The statlect website, with its clear demonstrations and examples, will also be used

The main objective here will not be to present a formal approach to econometrics. Instead, I will aim to provide a concise overview of the theory necessary to understand the topic, and proceed with practical examples using R, Excel, and other tools.

2 The base of Linear Rgression

To understand Linear Regression, we need to grasp some statistical concepts. Let's start with relationships between two variables.

2.1 Bivariate Frequency Distribution

Note: The columns refer to chest circumference (inches), and the rows refer to height (inches)

	33-35	36-38	39-41	42-44	45-over	Row totals
64-65	39.00	331.00	326.00	26.00	0.00	722.00
66-67	40.00	591.00	1010.00	170.00	4.00	1815.00
68-69	19.00	312.00	1144.00	488.00	18.00	1981.00
70-71	5.00	100.00	479.00	290.00	23.00	897.00
72-73	0.00	17.00	120.00	153.00	27.00	317.00
Column totals	103.00	1351.00	3079.00	1127.00	72.00	5732.00

Note: The second table displays conditional means

	1	2	3	4	5
mean of height given chest-inches	66.31	66.84	67.89	69.16	70.53
mean of chest given height-inches	38.41	39.19	40.26	40.76	41.80

The second table shows a lot of numbers. But how were they calculated? Let's demonstrate that: The number 66.84 in the first row, second column of the second table was derived by performing the following calculation: $(64.5 \cdot 331 + 66.5 \cdot 591 + 68.5 \cdot 312 + 70.5 \cdot 100 + 72.5 \cdot 17)/1351 = 66.84$.

In a similary way, the number 38.41 in the second row, first column of the second table: $(39 \cdot 34 + 331 \cdot 37 + 326 \cdot 40 + 26 \cdot 43)/722 = 38.41$

2.2 The correlation coefficient

The correlation coefficient measures the direction and closeness of the linear association between two variables. Let´s denote the observations by (X_i, Y_i) with $i = 1, 2, 3...n$. The data can be expressed in deviation form as: $(x_i = X_i - \overline{X}), (y_i = Y_i - \overline{Y})$, where $\overline{X}, \overline{Y}$ denote sample means of X and Y. Def: $Cov(X,Y) = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})/n = \sum_{i=1}^{n} x_i y_i/n$. One problem with the sample covariance is that it is sensitive to the unit. Suppose X is measured in dollars, and so is Y. The covariance gives $dollars^2$ measure. If X and Y change to centes, it gives a coeficient that is $1 * 10^4$ the old.

The covariance of the standardized deviations is the correlation coefficient, **r** namely, measures linear association, and is dimensionless. **r** is limited,$-1 \leq$ **r** ≤ 1

$\mathbf{r} = \sum_{i=1}^{n}(x_i/s_x)(y_i/s_y)/n$, where $s_x = \sqrt{(\sum_{i=1}^{n}(x_i)^2/n)}$

2.3 Practical examples using R

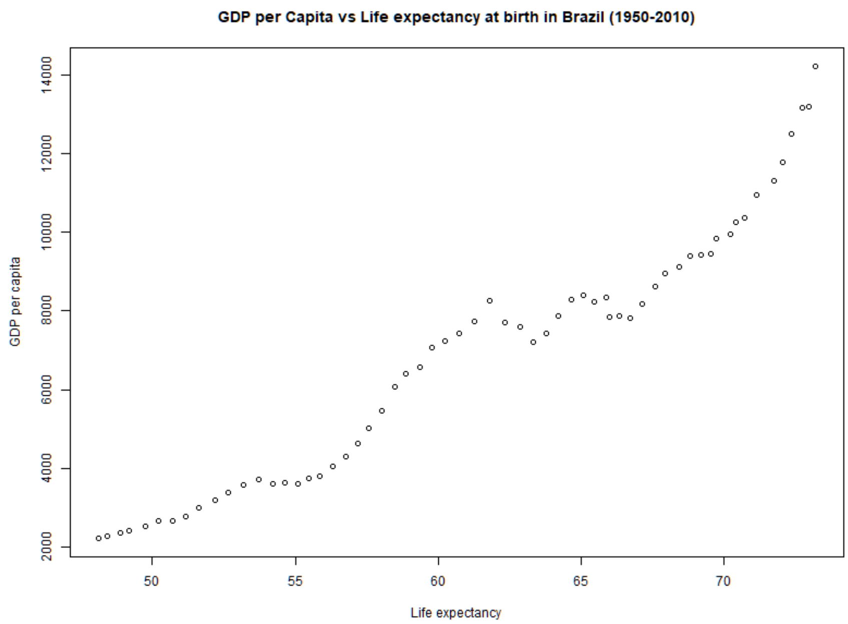
Here, we can calculate the correlation coefficient between life expectancy and GDP per capita in Brazil. We have 61 observations from 1950 to 2010. This means that for each year, we have a pair (life expectancy, GDP), and for this, we apply the formula above. The data is available here life expectancy vs GDP per-capita

	Year	Life expectancy	GDP per Capita
1	1950.00	48.12	2236.00
2	1951.00	48.43	2279.00
3	1952.00	48.87	2377.00
4	1953.00	49.20	2418.00
5	1954.00	49.75	2531.00
6	1955.00	50.22	2675.00
7	1956.00	50.71	2672.00
8	1957.00	51.17	2793.00
9	1958.00	51.64	3005.00
10	1959.00	52.19	3201.00
11	1960.00	52.66	3398.00
12	1961.00	53.18	3585.00
13	1962.00	53.71	3711.00
14	1963.00	54.21	3623.00
15	1964.00	54.65	3637.00
16	1965.00	55.08	3617.00
17	1966.00	55.47	3747.00
18	1967.00	55.87	3795.00
19	1968.00	56.31	4050.00
20	1969.00	56.75	4313.00
21	1970.00	57.17	4635.00
22	1971.00	57.59	5024.00
23	1972.00	58.03	5480.00
24	1973.00	58.47	6086.00
25	1974.00	58.88	6416.00
26	1975.00	59.35	6582.00
27	1976.00	59.79	7079.00
28	1977.00	60.24	7248.00
29	1978.00	60.72	7425.00
30	1979.00	61.25	7736.00
31	1980.00	61.78	8249.00
32	1981.00	62.33	7709.00
33	1982.00	62.86	7587.00
34	1983.00	63.33	7203.00
35	1984.00	63.77	7438.00
36	1985.00	64.20	7862.00
37	1986.00	64.64	8281.00
38	1987.00	65.08	8402.00
39	1988.00	65.45	8230.00
40	1989.00	65.87	8333.00
41	1990.00	65.98	7842.00
42	1991.00	66.31	7888.05

43	1992.00	66.71	7812.79
44	1993.00	67.11	8166.24
45	1994.00	67.57	8615.69
46	1995.00	67.92	8951.69
47	1996.00	68.41	9124.52
48	1997.00	68.81	9409.95
49	1998.00	69.19	9419.11
50	1999.00	69.52	9441.76
51	2000.00	69.74	9834.42
52	2001.00	70.19	9953.31
53	2002.00	70.41	10245.07
54	2003.00	70.72	10354.60
55	2004.00	71.13	10949.66
56	2005.00	71.75	11305.77
57	2006.00	72.04	11766.60
58	2007.00	72.37	12500.01
59	2008.00	72.72	13164.01
60	2009.00	72.95	13180.89
61	2010.00	73.18	14215.61

The correlation between GDP per capita and life expectancy at birth in Brazil (1950-2010) is 0.972979.

The graph below shows every point (Life expectancy, GDP per capita). It is reasonable to say that we can observe a positive correlation between the two variables. As one variable increases, the other also tends to increase. In fact, this relationship helps us understand the significance of the exact value of the correlation coefficient.



2.4 Probability Models for two variables

2.4.1 Discrete Bi-variate Probability Distribution

First, let show a table of a discrete bi-variate probability distribution.

TABLE : A bivariate probability distribution

	X_1	$\dots$	X_i	Marginal probability
Y_1	p_{11}	$\dots$	p_{i1}	$p_{.1}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Y_j	p_{1j}	$\dots$	p_{ij}	$P_.j$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Y_p	p_{1p}	$\dots$	p_{ip}	$P_.p$
Marginal probability	$p_{1.}$	$\dots$	$p_{i.}$	1

The covariance is:

$$\sigma_{X,Y} = cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)] = \sum_i \sum_j p_{ij}(X_i-\mu_x)(Y_j-\mu_y)$$

The population correlation coefficient is defined as:

$$corr(X,Y) = \rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

2.4.2 Conditional Probabilities

Consider the X_i column in the the table above. Each cell probability may be divided by the column total, $p_i.$, to give a conditional probability for Y given X_i. Thus,

$$\frac{p_{ij}}{p_{i.}} = probability\, that\, Y = Y_i\, given\, that\, X = X_i = Prob(Y_j|X_i)$$

The mean of this distribution is the conditional expectation of Y, given X_i, that is:

$$\mu_{y|x_i} = E(Y|X_i) = \sum_j (\frac{p_{ij}}{p_i.})Y_j$$

Similarly, the variance of this distribution is a conditional variance, or

$$\sigma^2_{y|x_i} = var(Y|X_i) = \sum_j (\frac{p_{ij}}{p_i})(Y_j - \mu_{y|x_i})^2$$

The conditional means and variances are both functions of X, so there is a set of "i" conditional means and variances.

Columns refer to Income (X) and rows refer to Vacation Expenditure (Y)

	20000	30000	40000
1000	0.28	0.03	0.00
2000	0.08	0.15	0.03
3000	0.04	0.06	0.06
4000	0.00	0.06	0.15
5000	0.00	0.00	0.03
6000	0.00	0.00	0.03
Marginal Probability	0.40	0.30	0.30
Mean(Y|X)	1.40	2.50	3.90
Var(Y|X)	0.44	0.85	1.09

From the table above, come the conditional probabilities:

Table: Conditional Probabilities. Columns refer to Income (X) and rows refer to Vacation Expenditure (Y).

	1000	2000	3000	4000	5000	6000
20000	0.70	0.20	0.10	0.00	0.00	0.00
30000	0.10	0.50	0.20	0.20	0.00	0.00
40000	0.00	0.10	0.20	0.50	0.10	0.10

Using the table of conditional probabilities, it's possible to calculate the last two lines of the first table: the conditional means and variances.

2.5 End of first part and considerations

Certainly, there are additional statistical concepts necessary to fully understand linear regression, which is the core of Econometrics 1. If needed for any demonstration, I will incorporate these concepts into the text. The initial part of this document is derived from the book mentioned in the introduction. For now, I will likely focus on linear regression using the class notes.

In addition to these, understanding the following concepts is also necessary: the density function of probability; knowledge of well-known distributions such as Normal, Bernoulli, and Uniform; the expected value of a continuous variable; the conditional expectation of continuous variables; and matrices, including fundamental operations like sum, multiplication, and diagonalization.

3 Linear Regression Model

The main objective here will not be to present a formal approach to econometrics. Instead, I will aim to provide a concise overview of the theory necessary to understand the topic, and proceed with practical examples using R, Excel, and other tools.

3.1 The k-variable model

For this part, I am going to use class notes and statlect by Marcos Taboga.
$y_i = B_1 x_{i1} + B_2 x_{i2} ... + B_k x_{ik} + \varepsilon_i,$
$x_{1i} = 1 \ \forall i$
Using matrix notation:

$$B_{K\times 1} = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_k \end{bmatrix}, \, x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{bmatrix} \rightarrow y_i = x_i{}'B + \varepsilon_i, \, i = 1, 2...n$$

In this form, y_i represents a scalar output variable, called dependent variable or regressand.

$x_i{}'$ is a kx1 vector of input variables, called independent variables or regressors.

N is the sample size.

B is a kx1 vector of constants, called regression coefficients.

ε_i is an unobservable error term.This includes the sources of variability in y_i that are not accounted for in the input vector x_i,such as measurement errors and input variables that are not observed by the statistician.

3.1.1 Example:

Suppose there is a sample of countries for which GDP, life expectancy, and crime rates are observed. We aim to establish a linear regression model to predict GDP based on life expectancy and crime rates. $g_i = B_1 + B_2 l_i + B_3 c_i + \varepsilon_i$

g_i, l_i, c_i denote GDP, life expectancy and crime rates.
B_1, B_2, B_3 are regression coefficients.
ε_i is an error term.
The equation can be written in a vector notation as:

$$y_i = {x_i}' B + \varepsilon_i$$

by defining

$$y_i = g_i$$

$$x_i = \begin{bmatrix} 1 \\ l_i \\ c_i \end{bmatrix}$$

$$B = \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix}$$

3.1.2 Matrix Notation

Denote y by the Nx1 vector of outputs

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

X by the NxK matrix of inputs

$$X = \begin{bmatrix} {x_1}' \\ {x_2}' \\ \vdots \\ {x_N}' \end{bmatrix}$$

and ε, the Nx1 error terms by:

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

we have, then:

$$y_{Nx1} = X_{NxK} B_{Kx1} + \varepsilon_{Nx1}$$

3.1.3 Design Matrix

Definition: A design matrix contains information/data about multiple characteristics of many objects of interest. Each row corresponds to an individual and each column to a characteristic.

In the context of Linear Regression, it is often represented by the X.

Example: Consider the linear regression $y_i = x_i{}'B + \varepsilon_i$, where y_i is the dependent variable, $x_i{}'$ is a Kx1 vector containing the K explanatory variables, B is a Kx1 vector of regression coefficients, ε_i is the error term and there are N observations

All the observations can be collected in the design matrix:

$$X = \begin{bmatrix} x_1{}' \\ x_2{}' \\ \vdots \\ x_N{}' \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & & & \vdots \\ 1 & x_{N2} & \cdots & x_{Nk} \end{bmatrix}$$

where it sets the first column, $x_{i1} = 1$, because in the matrix form $Y = XB + \varepsilon$, this 1 in each row for each individual forms the intercept of the regression line.

3.1.4 Initiating a practical example using R

Not considering the context of biased estimator and similar concepts, we can begin the visualization of the regression I mentioned earlier: Suppose there is a sample of countries for which GDP per capita, life expectancy, and crime rates are observed. We aim to establish a linear regression model to predict GDP based on life expectancy and crime rates. $g_i = B_1 + B_2 l_i + B_3 c_i + \varepsilon_i$, which is equivalent to $y_i = x_i{}'B + \varepsilon_i$ and $Y = XB + \varepsilon$

Database: For the crime rate, the variable "homicide rate" is set, while for GDP and life expectancy, it´s possible to use this data separately, here.

All variables are available for many years, from 1990 to 2010, for example. For simplification, I will choose one year that contains the largest number of countries for which we have data for all variables.

Did that, this is the available data, as follows, to perform the regression.

Country and relevant data

Entity	Code	Year	Life expectancy	Homicide rate	GDP per capita
Afghanistan	AFG	2010	60.8508	3.4870927	1627.6716
Albania	ALB	2010	77.9359	4.0845757	9222.9730
Algeria	DZA	2010	73.8081	0.7083823	12587.7440
Argentina	ARG	2010	75.7208	5.8004694	18979.9920
Armenia	ARM	2010	73.1597	1.9006935	8330.8120
Australia	AUS	2010	82.0550	1.0490860	45400.2230
Austria	AUT	2010	80.4643	0.7294182	40288.3480
Azerbaijan	AZE	2010	69.5291	2.2301123	16153.8370
Bahrain	BHR	2010	78.7480	0.1647929	34057.7150
Bangladesh	BGD	2010	68.6384	2.6874920	2599.2085
Barbados	BRB	2010	75.7115	11.2845860	11954.3220
Belarus	BLR	2010	71.1219	4.1206700	17172.4600

Country and relevant data *(continued)*

Entity	Code	Year	Life expectancy	Homicide rate	GDP per capita
Bosnia and Herzegovina	BIH	2010	77.0729	1.4693967	8982.4580
Botswana	BWA	2010	60.0133	14.4860710	12544.7280
Brazil	BRA	2010	73.1821	26.6300330	14215.6120
Bulgaria	BGR	2010	73.8348	1.9493505	14686.4610
Cambodia	KHM	2010	67.7123	2.3044472	2426.6870
Cameroon	CMR	2010	56.5820	5.0256470	2420.1328
Canada	CAN	2010	81.3469	1.6400001	41209.4260
Chile	CHL	2010	78.5015	3.1815740	18909.8480
China	CHN	2010	75.5991	0.9946659	9658.4190
Colombia	COL	2010	75.0327	34.4942930	11191.6870
Costa Rica	CRI	2010	78.6704	11.4013700	12031.5710
Croatia	HRV	2010	76.8049	1.4191923	19511.3500
Cuba	CUB	2010	77.6610	4.4728193	6578.1904
Cyprus	CYP	2010	79.6650	0.7081617	27630.1040
Czechia	CZE	2010	77.5703	0.9842567	25922.3950
Denmark	DNK	2010	79.2459	0.7566411	42932.4000
Dominica	DMA	2010	71.5499	21.8165950	8768.4780
Dominican Republic	DOM	2010	72.0933	25.3075080	11276.4610
Ecuador	ECU	2010	75.4301	17.5054870	9327.2130
Egypt	EGY	2010	69.6638	2.1076782	10719.2910
El Salvador	SLV	2010	71.8478	65.2106250	7351.1147
Estonia	EST	2010	75.7485	5.2570906	20713.4260
Eswatini	SWZ	2010	46.6191	18.9104670	7192.7520
Finland	FIN	2010	80.0117	2.2001498	37615.1130
France	FRA	2010	81.4275	1.2747306	36086.7270
Georgia	GEO	2010	72.1277	4.8738136	8443.4370
Germany	DEU	2010	80.0864	0.9750988	41109.5820
Ghana	GHA	2010	61.1571	1.6500670	2946.0393
Greece	GRC	2010	80.5102	1.5951011	26517.4650
Guatemala	GTM	2010	70.8783	40.9815750	6526.4575
Haiti	HTI	2010	46.0185	6.8780680	1556.5504
Honduras	HND	2010	71.0884	73.7906700	4305.8920
Hong Kong	HKG	2010	82.9930	0.4907158	43170.5500
Hungary	HUN	2010	74.5192	1.3818206	20036.3340
Iceland	ISL	2010	81.7003	0.6282729	36272.6880
India	IND	2010	66.9086	3.7449210	4525.7456
Iraq	IRQ	2010	67.0622	8.6071030	10274.3300
Ireland	IRL	2010	80.5226	1.2155810	48623.8100
Israel	ISR	2010	81.7443	2.0468190	28575.5310
Italy	ITA	2010	82.1371	0.8842834	34765.9380
Jamaica	JAM	2010	72.6219	52.9281270	7025.9300
Japan	JPN	2010	82.9193	0.3629823	35011.4020
Jordan	JOR	2010	73.9996	1.6735778	11601.0880
Kazakhstan	KAZ	2010	68.1157	8.3414340	19964.5300
Kenya	KEN	2010	60.6490	4.6389630	2579.5310
Kuwait	KWT	2010	77.9817	2.0384890	68865.3100
Lebanon	LBN	2010	78.1559	3.2627410	15622.3670
Liberia	LBR	2010	59.4331	3.2089903	854.2350
Lithuania	LTU	2010	73.4236	6.9767030	18663.7620
Luxembourg	LUX	2010	80.5731	1.9721162	54086.3360
Malawi	MWI	2010	56.3806	3.5329876	1091.5876
Malaysia	MYS	2010	74.4423	1.8803715	18574.2990
Malta	MLT	2010	81.8134	0.9552125	23632.6250
Mauritius	MUS	2010	73.9987	2.5714352	16269.3200
Mexico	MEX	2010	74.1901	22.8885200	14697.2980
Moldova	MDA	2010	69.3569	7.2046385	4615.3240
Mongolia	MNG	2010	67.1822	8.8435970	7830.0640
Montenegro	MNE	2010	76.1387	2.3770134	14509.7660
Morocco	MAR	2010	70.8288	1.3891941	6931.5580
Mozambique	MOZ	2010	54.1984	3.7921925	956.5305
Myanmar	MMR	2010	63.3294	1.6217533	3772.9985
Namibia	NAM	2010	56.0153	14.9575770	7905.5410
Nepal	NPL	2010	66.8138	2.9932000	2152.2747
Netherlands	NLD	2010	80.8186	0.8665763	43812.3480

Country and relevant data *(continued)*

Entity	Code	Year	Life expectancy	Homicide rate	GDP per capita
New Zealand	NZL	2010	81.1019	0.9893385	31586.0900
Nicaragua	NIC	2010	72.0090	13.4056640	3916.9658
North Macedonia	MKD	2010	75.1187	2.0536550	11038.2900
Norway	NOR	2010	81.0401	0.5930784	78476.1500
Oman	OMN	2010	76.2695	1.6655595	46182.1600
Pakistan	PAK	2010	64.4361	6.7830777	4354.2670
Palestine	PSE	2010	73.0045	0.9768859	4548.5312
Panama	PAN	2010	76.4479	12.6944960	15169.1250
Paraguay	PRY	2010	71.8872	12.8453770	7139.0270
Philippines	PHL	2010	70.7542	9.1349340	5694.0470
Poland	POL	2010	76.3175	0.9948869	20608.6930
Portugal	PRT	2010	80.0350	1.1710927	25463.1640
Puerto Rico	PRI	2010	78.0100	27.7036480	33195.4400
Qatar	QAT	2010	78.4251	0.2334398	134802.7800
Romania	ROU	2010	73.9959	1.3080759	16377.3280
Russia	RUS	2010	69.3860	11.6515620	21737.3850
Rwanda	RWA	2010	62.5397	2.8227677	1416.3923
Saint Lucia	LCA	2010	72.7196	25.7407780	10058.6230
Sao Tome and Principe	STP	2010	65.2486	3.2941964	3093.3452
Serbia	SRB	2010	74.3753	1.7246451	11963.5550
Seychelles	SYC	2010	72.9703	9.7393640	20147.1780
Sierra Leone	SLE	2010	53.6932	2.5323544	1421.1521
Singapore	SGP	2010	81.6910	0.3679611	58612.7300
Slovakia	SVK	2010	75.4950	1.6492403	21941.2130
Slovenia	SVN	2010	79.7039	0.7291154	26000.9200
South Africa	ZAF	2010	58.8987	30.6904010	11318.6045
South Korea	KOR	2010	80.7630	0.9928375	31537.7730
Spain	ESP	2010	82.0214	0.8610181	31786.4040
Sri Lanka	LKA	2010	73.2150	3.6915977	8319.7030
Sweden	SWE	2010	81.5187	0.9699704	42634.7540
Switzerland	CHE	2010	82.2831	0.6647546	57219.5000
Syria	SYR	2010	73.8808	2.0727417	6520.6123
Tajikistan	TJK	2010	67.7420	2.3747735	2975.9268
Tanzania	TZA	2010	60.1048	8.6188310	2068.8638
Thailand	THA	2010	76.1313	5.3522390	13343.5470
Trinidad and Tobago	TTO	2010	72.7254	33.5390600	30235.9940
Tunisia	TUN	2010	75.4206	2.6250422	10721.8120
Turkey	TUR	2010	75.0693	4.1860585	16528.1200
Turkmenistan	TKM	2010	68.2933	1.9931778	14397.1490
Uganda	UGA	2010	57.0547	9.7706590	1831.2906
Ukraine	UKR	2010	70.6207	4.3517265	9601.2090
United Arab Emirates	ARE	2010	78.3335	0.7899294	60112.4020
United Kingdom	GBR	2010	80.4008	1.2045882	34754.4730
United States	USA	2010	78.7724	4.7309804	49266.9140
Uruguay	URY	2010	76.8580	6.1145644	16415.6290
Venezuela	VEN	2010	72.8901	45.5510700	17105.5800
Vietnam	VNM	2010	73.5126	1.5284115	4571.8486
Yemen	YEM	2010	67.2800	4.4414907	4812.3450
Zambia	ZMB	2010	56.7991	5.9019350	3032.0679
Zimbabwe	ZWE	2010	50.6523	5.5374823	1401.8564

Let´s see a simplified version of the design matrix and linear regression.