Body Mass Index in the world

Marine Dissler and Jonathan Lacombe

ENAC

May 10, 2012

The data base

BMI ?

Body Mass Index (BMI) is a simple index of weight-for-height that is commonly used to classify underweight, overweight and obesity in adults. It is defined as the weight in kilograms divided by the square of the height in meters (kg/m^2).

Our data base

We built our data base from different web sites proposing statistics.
We chose the year 2008 = year with the most data.
Values not always accessible = we deleted some countries and some variables.
CCI: our data base is composed of 150 countries and we kept 14 explanatory variables.

What we would expect ?

Explanatory Variables	Expected effects
Proportion of women in the country	-
Cereals consumption	+ or -
Fruits consumption	-
Proteins consumption	+
Fat consumption	+
Access to improved drinking water sources	?
Alcohol consumption	+
Undernourishment	-
Food production index	+
GDP per capita	+
Health expenditure	-
Number of Personal computers	+
Suicide rate	+
Literacy rate	-

Linear regression

Using ordinary least squares we found this equation :

$$\begin{aligned}\widehat{BMI_i} = \beta_0 &+ \beta_1 WATER_i + \beta_2 SUICIDE_i + \beta_3 LITERACY_i \\ &+ \beta_4 EXPENDITURE_i + \beta_5 ALCOHOL_i \\ &+ \beta_6 UNDERNOURISHMENT_i + \beta_7 CEREAL_i + \beta_8 FAT_i \\ &+ \beta_9 FEMALE_i + \beta_{10} FOODPROD_i + \beta_{11} FRUIT_i + \beta_{12} GDP_i \\ &+ \beta_{13} PC_i + \beta_{14} PROTEINS_i \end{aligned} \tag{1}$$

where $i = 1; 2; ...; 150$ indexes the observations.

Eviews results

Dependent Variable: AVERAGE_BMI
Method: Least Squares
Date: 05/08/12 Time: 20:26
Sample: 1 150
Included observations: 133

	Coefficient	Std. Error	t-Statistic	Prob.
C	26.42958	4.288775	6.162500	0.0000
ACCESS_WATER_SOURCE	0.017313	0.013447	1.287533	0.2004
SUICIDE_RATE	-0.056632	0.021851	-2.591755	0.0108
ADULT_LITERACY_RATE	0.035535	0.010492	3.386962	0.0010
TOTAL_EXPENDITURE_ON_HEA	0.025903	0.066866	0.387390	0.6992
ALCOHOL_CONSUMPTION	0.013261	0.039718	0.333867	0.7391
UNDERNOURISHMENT	-0.048523	0.015643	-3.102016	0.0024
CEREAL_CONSUMPTION	-0.000526	0.000539	-0.976054	0.3310
FAT_CONSUMPTION	0.009109	0.008843	1.030106	0.3051
FEMALE_PROPORTION	-0.112758	0.075554	-1.492410	0.1383
FOOD_PRODUCTION	-0.001874	0.013221	-0.141775	0.8875
FRUITS_VEG_PULSES_NUTS_C	0.002153	0.001335	1.613578	0.1093
GDP___CAPITA	-3.70E-06	1.73E-05	-0.213742	0.8311
NUMBER_OF_PC	-0.020904	0.012395	-1.686530	0.0943
PROTEINS_CONSUMPTION	0.008406	0.015013	0.559919	0.5766

R-squared	0.652439	Mean dependent var	25.04853
Adjusted R-squared	0.611203	S.D. dependent var	2.173061
S.E. of regression	1.354982	Akaike info criterion	3.551352
Sum squared resid	216.6450	Schwarz criterion	3.877331
Log likelihood	-221.1649	Hannan-Quinn criter.	3.683818
F-statistic	15.82206	Durbin-Watson stat	1.589129
Prob(F-statistic)	0.000000		

Figure : Linear regression of our model.

Linear regression

Finally, here is the equation :

$$\begin{aligned}\widehat{BMI_i} = \; & 26.43 + 0.017 WATER_i - 0.057 SUICIDE_i \\ & + 0.036 LITERACY_i + 0.026 EXPENDITURE_i \\ & + 0.013 ALCOHOL_i - 0.049 UNDERNOURISHMENT_i \\ & - 0.00053 CEREAL_i + 0.0091 FAT_i - 0.11 FEMALE_i \\ & - 0.0019 FOODPROD_i + 0.0022 FRUIT_i \\ & - 3.70.10^{-}6 GDP_i - 0.021 PC_i + 0.0084 PROTEINS_i \end{aligned} \tag{2}$$

Interpretation

Our first comments are resumed in this chart :

Explanatory Variables	Expected effect	Real effect
Proportion of women in the country	-	-
Cereals consumption	+ or -	-
Fruits consumption	-	+
Proteins consumption	+	+
Fat consumption	+	+
Access to improved drinking water sources	?	+
Alcohol consumption	+	+
Undernourishment	-	-
Food production index	+	-
GDP per capita	+	-
Health expenditure	-	+
Number of Personal computers	+	-
Suicide rate	+	-

Validity of the model

R^2 measurement

- R^2 measures how close the points are to the estimated regression line in the scatter plot.
- Closer the R^2 is to 1, better is the model.

Here R^2=0.652 so **the model is acceptable.**

t-statistic test

It tests the probability for each coefficient to be equal to 0.

For a 10 % level test, only
-Suicide rate,
-Literacy,
-Undernourishment and
-Number of computers
} are relevant !

More tests

So we performed other variable tests to check if it is necessary to drop out so many indicators.

- Correlation matrix
- F-Statistic

More tests

	ACCESS_...	ADULT_LIT...	ALCOHOL...	SUICIDE_...	TOTAL_EX...	UNDERNO...	CEREAL_...	FAT_CONS...	FEMALE_P...	FOOD_PR...	FRUITS_V...	GDP___CA...	NUMBER_...	PROTEINS...
ACCESS_W...	1.000000	0.736583	0.429078	0.039499	0.419286	-0.691817	-0.103224	0.665122	0.025462	-0.286783	0.188382	0.493879	0.499045	0.656490
ADULT_LIT...	0.736583	1.000000	0.539819	0.108979	0.405085	-0.553772	-0.245121	0.586390	0.108970	-0.313678	0.096006	0.465674	0.479654	0.584036
ALCOHOL_...	0.429078	0.539819	1.000000	0.352801	0.515274	-0.399425	-0.376479	0.512348	0.420312	-0.196145	0.145422	0.375910	0.450111	0.446516
SUICIDE_R...	0.039499	0.108979	0.352801	1.000000	0.098141	0.000791	-0.082548	0.018023	0.272374	0.072621	-0.180365	0.020421	0.073814	-0.020909
TOTAL_EXP...	0.419286	0.405085	0.515274	0.098141	1.000000	-0.331962	-0.326898	0.541272	0.260879	-0.230752	0.247479	0.513307	0.587681	0.496905
UNDERNO...	-0.691817	-0.553772	-0.399425	0.000791	-0.331962	1.000000	-0.082774	-0.707214	-0.000337	0.136696	-0.127601	-0.472648	-0.469305	-0.774950
CEREAL_C...	-0.103224	-0.245121	-0.376479	-0.082548	-0.326898	-0.082774	1.000000	-0.283510	-0.099269	0.200396	-0.334047	-0.308611	-0.343965	-0.016193
FAT_CONS...	0.665122	0.586390	0.512348	0.018023	0.541272	-0.707214	-0.283510	1.000000	0.079059	-0.311299	0.157017	0.779027	0.775240	0.859418
FEMALE_PR...	0.025462	0.108970	0.420312	0.272374	0.260879	-0.000337	-0.099269	0.079059	1.000000	0.037027	-0.251247	-0.176678	0.026485	-0.011196
FOOD_PRO...	-0.286783	-0.313678	-0.196145	0.072621	-0.230752	0.136696	0.200396	-0.311299	0.037027	1.000000	-0.258000	-0.289254	-0.252458	-0.282772
FRUITS_VE...	0.188382	0.096006	0.145422	-0.180365	0.247479	-0.127601	-0.334047	0.157017	-0.251247	-0.258000	1.000000	0.189857	0.116409	0.209302
GDP___CA...	0.493879	0.465674	0.375910	0.020421	0.513307	-0.472648	-0.308611	0.779027	-0.176678	-0.289254	0.189857	1.000000	0.882515	0.727039
NUMBER_O...	0.499045	0.479654	0.450111	0.073814	0.587681	-0.469305	-0.343965	0.775240	0.026485	-0.252458	0.116409	0.882515	1.000000	0.676687
PROTEINS_...	0.656490	0.584036	0.446516	-0.020909	0.496905	-0.774950	-0.016193	0.859418	-0.011196	-0.282772	0.209302	0.727039	0.676687	1.000000

Figure : Correlation matrix

More tests

Correlation matrix

-Correlation between GDP per capita and number of PC is high: 0.88

-Correlation between fat consumption and proteins consumption is high also: 0.86

$\Rightarrow$ we chose to delete **GDP per capita** and **proteins consumption** from the correlated pairs according to their high p-value.

More tests

Wald test

- F-statistic and Prob(F-statistic) are used for testing the null hypothesis :

$$H_0 : \beta_1 = 0, \beta_2 = 0, ..., \beta_k = 0$$

- We drop out the indicators with the highest Prob(F-Statistics):

- Food production 0.89
- GDP per capita 0.97
- Alcohol consumption 0.98
- Expenditure on health 0.99
- And finally,proteins consumption 0.98

New model

According to all results we found in previous section, we built a new model of our data set.
Thus, the remaining variables are :

- Access to water sources
- Suicide rate
- Adult literacy
- Undernourishment
- Cereal consumption
- Fat consumption
- Female proportion
- Fruits consumption
- Number of PC

Eviews results

Dependent Variable: AVERAGE_BMI
Method: Least Squares
Date: 05/09/12 Time: 12:05
Sample: 1 150
Included observations: 134

	Coefficient	Std. Error	t-Statistic	Prob.
C	25.36021	3.227558	7.857401	0.0000
ACCESS_WATER_SOURCE	0.018280	0.013041	1.401773	0.1635
SUICIDE_RATE	-0.055182	0.020239	-2.726581	0.0073
ADULT_LITERACY_RATE	0.036092	0.009549	3.779708	0.0002
UNDERNOURISHMENT	-0.051866	0.014155	-3.664124	0.0004
CEREAL_CONSUMPTION	-0.000526	0.000445	-1.182582	0.2392
FAT_CONSUMPTION	0.012279	0.006768	1.814239	0.0721
FEMALE_PROPORTION	-0.087954	0.057601	-1.526962	0.1293
FRUITS_VEG_PULSES_NUTS_C	0.002646	0.001181	2.240613	0.0268
NUMBER_OF_PC	-0.020631	0.008276	-2.492978	0.0140

R-squared	0.648094	Mean dependent var	25.02884
Adjusted R-squared	0.622552	S.D. dependent var	2.176843
S.E. of regression	1.337383	Akaike info criterion	3.491001
Sum squared resid	221.7854	Schwarz criterion	3.707258
Log likelihood	-223.8971	Hannan-Quinn criter.	3.578881
F-statistic	25.37404	Durbin-Watson stat	1.564890
Prob(F-statistic)	0.000000		

Figure : Linear regression of our new model.

Linear regression

Finally, here is the equation :

$$\begin{aligned}\widehat{BMI_i} &= 25.36 + 0.018 WATER_i - 0.055 SUICIDE_i \\ &\quad + 0.036 LITERACY_i - 0.052 UNDERNOURISHMENT_i \\ &\quad - 0.00053 CEREAL_i + 0.012 FAT_i - 0.088 FEMALE_i \\ &\quad + 0.0026 FRUIT_i - 0.021 PC_i \end{aligned} \tag{3}$$

Interpretation

The sign of the coefficients before each variable is the same than in the first model, and so on for the interpretation.

Validity of the model

R^2 measurement

Here R^2=0.648.
We can see that R squared is slightly smaller than the one in the first equation but it still quite high.

t-statistic test

There are still variables that have a p-value superior to 10%, nevertheless there are not extremely high p-value anymore which is **satisfying**.

Validity of the model

Heteroskedasticity Test: White

F-statistic	1.095702	Prob. F(54,79)	0.3515
Obs*R-squared	57.38308	Prob. Chi-Square(54)	0.3509
Scaled explained SS	57.63650	Prob. Chi-Square(54)	0.3423

Figure : Heteroscedasticity

White's test

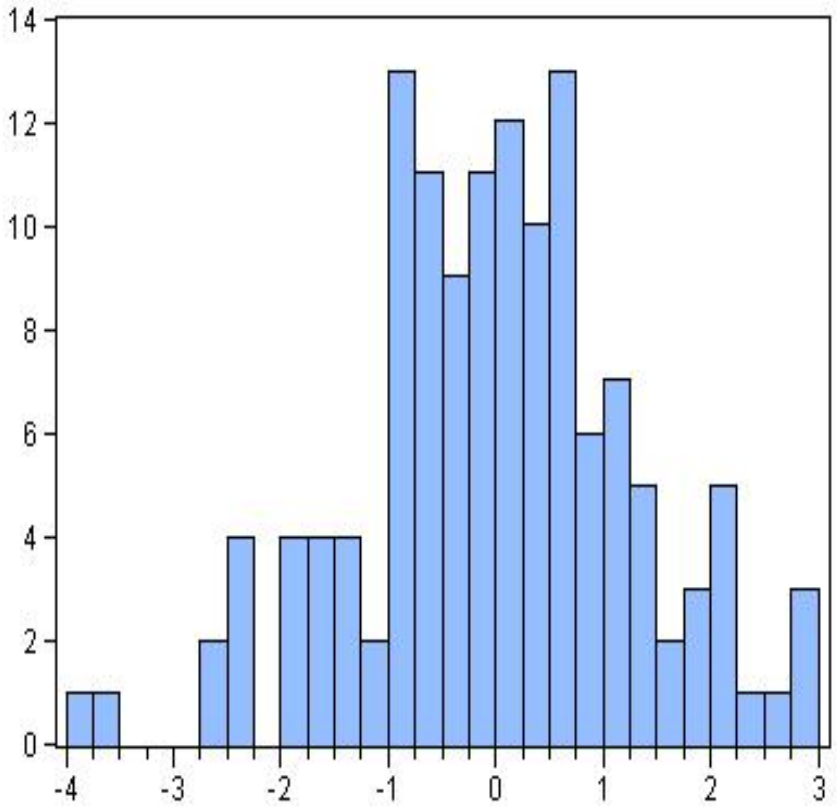
White's test is a test on the null hypothesis of no heteroscedasticity against heteroscedasticity. The probabilities are superior to 5% so we do not reject the null hypothesis.
⇒Our model is **homoscedastic**.

Residuals tests

We performed two residuals tests to check the normality of the residuals.

- Jarque Bera
- Quantile-quantile plot

Residuals tests



Series: RESID_EQ_PERFORMED	
Sample 1 150	
Observations 134	
Mean	-2.70e-15
Median	0.065654
Maximum	2.967844
Minimum	-3.895857
Std. Dev.	1.291340
Skewness	-0.195020
Kurtosis	3.345903
Jarque-Bera	1.517436
Probability	0.468266

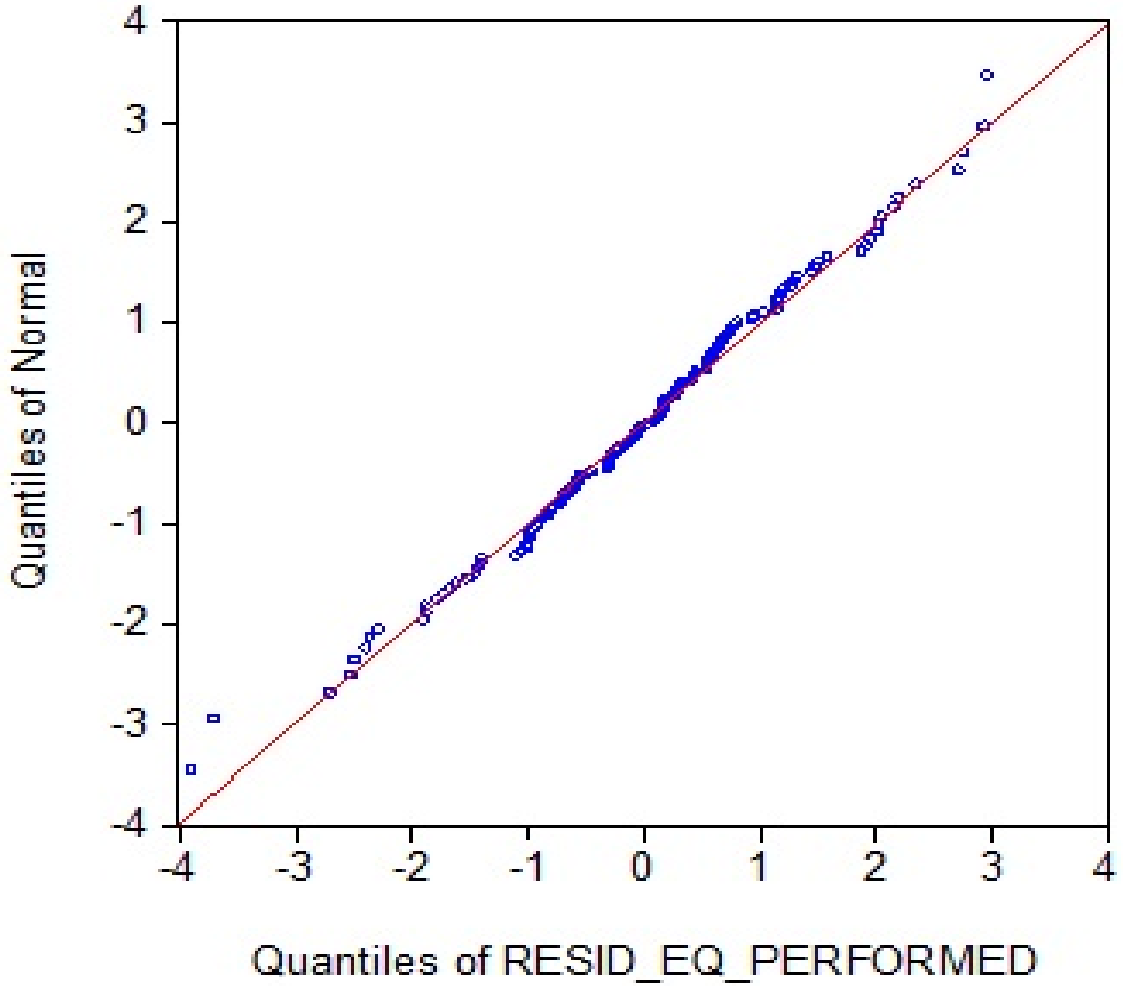
Figure : Stats and histogram.

Residuals tests

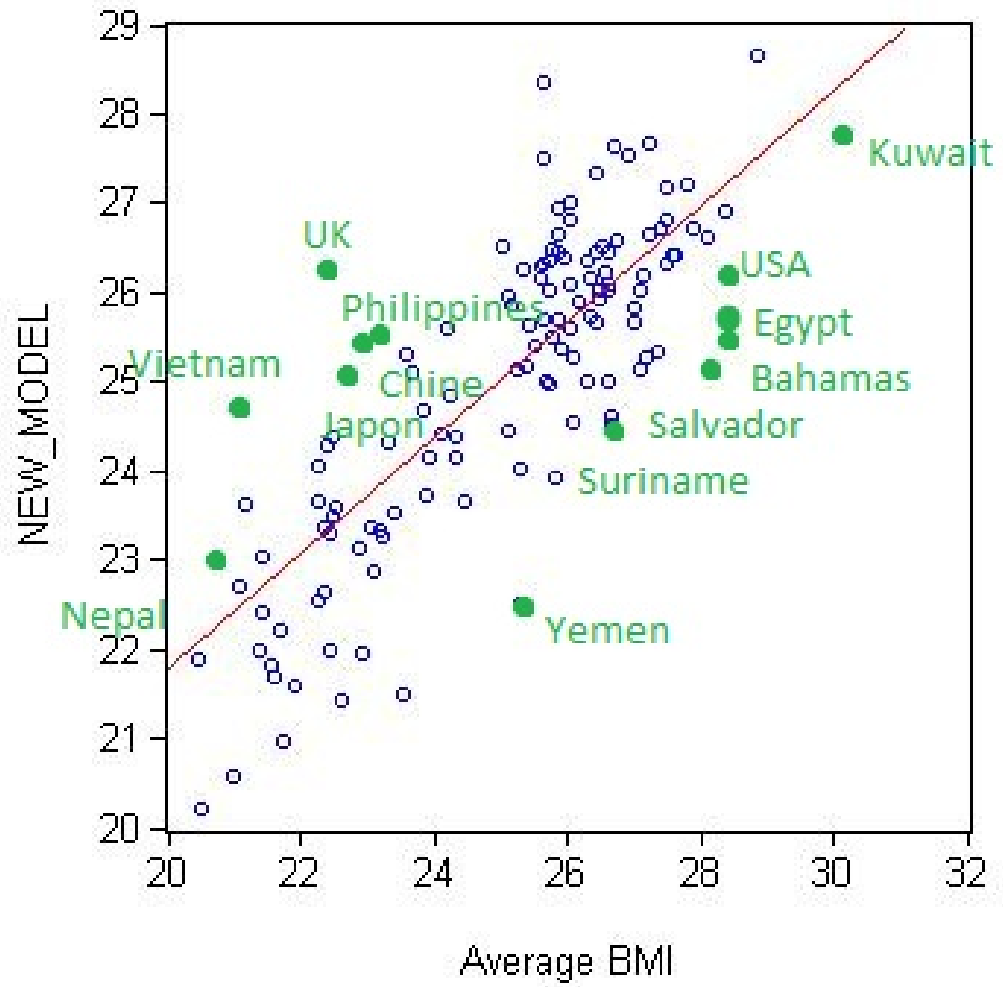
Jarque Bera

The Jarque-Bera statistic has a χ^2 distribution with two degrees of freedom under the null hypothesis of normally distributed errors. Here the associated probability is 0.47 and it is superior to 5% so **we do not reject** H_0.

Quantile-quantile



Conclusions of the model optimized



Conclusions of the model optimized

Possible explanations

1. The underestimation of Yemen :
   - high undernourishment rate,
   - low fat consumption, fruits and vegetables consumption and the literacy rate.
2. The overestimation of the UK :
   - high values for access to water sources, literacy rate and fat consumption,
   - low value for suicide rate, undernourishment and cereal consumption
3. Underestimation of the USA :
   - number of fast food, TV time, video games ... may be significant ?

Conclusion

Limits

The model should still be improved :

- Some variables do not seem to be significant enough,
- We might have missed some important ones.

Positive point

- Quite good model according to R^2,
- Validation of the linear regression.

Questions

Thank you for your attention.
Any questions ?

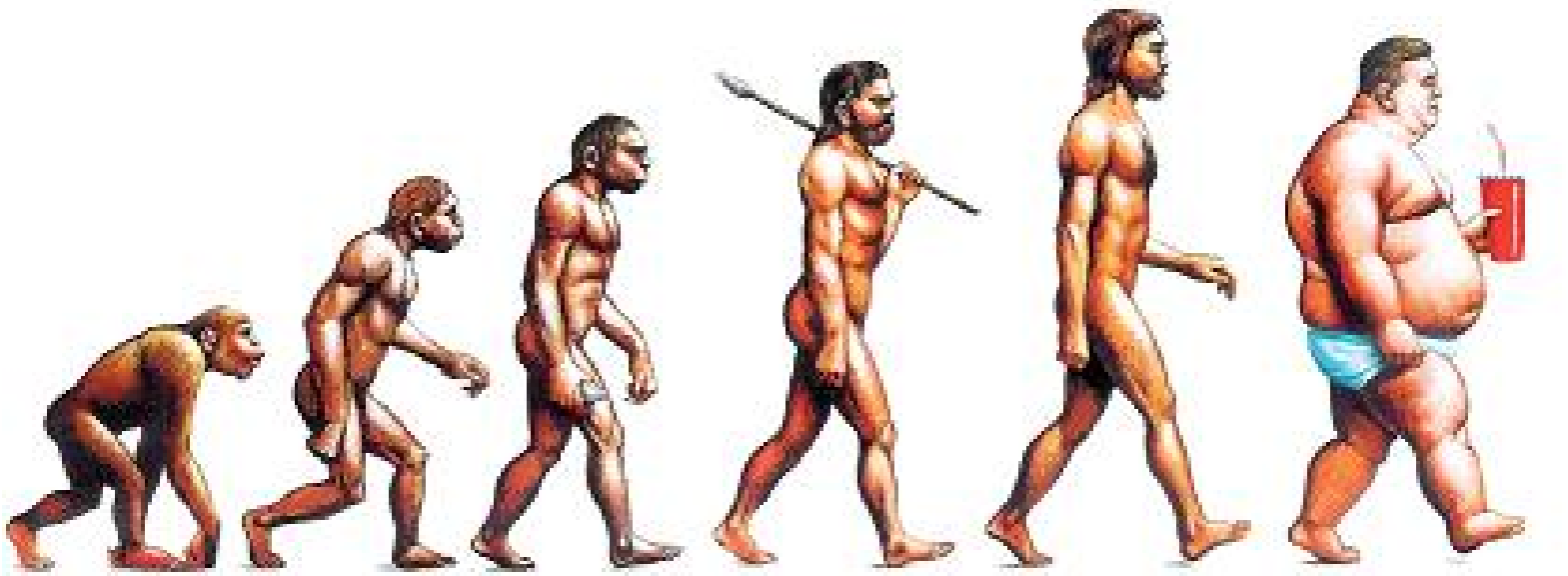


Figure : Our BMI destiny !