**A Linear Answer to Global Obesity Problem**

**Modeling Adult Body Mass Index (BMI) Across the Globe**

Anna Davydova, Ryan Ebanks, Paul Jonak, William Lodge II, Eric M. Morris, Winnie Yeung

|  |
| --- |
| ***Abstract*** |

We set out to understand the factors that lead to the global obesity problem among adults with a simple linear regression model. Taking a random sample of 50 countries from a population of 152, we ran a hierarchical multiple linear regression to test for a presence of a relationship between adult Body Mass Index (BMI) and a wide spectrum of macro, behavioral, environmental, and socioeconomic factors alongside the usual culprits of diet and exercise. Our expectation was that the relationship between adult BMI, food and exercise would prove to be most significant across the globe. To our surprise, our findings showed that an individual’s background and gender have an equally if not more powerful relationship with the global adult BMI. While at the end, we might not have found the magic formula to halt the rise in the global obesity rates, we did discover several variables that have a significant relationship with the global adult BMI and could prove helpful as the world looks to fight this global concern.

|  |
| --- |
| ***Introduction*** |

The world is getting fatter. Across the globe, 39% of the adult population is overweight and 13% is obese, a rate that has doubled since 1980.[[1]](#footnote-0) While great strides have been made in addressing other major global health issues like smoking, malaria, various respiratory infections, malnutrition, HIV/AIDS and waterborne illness, no country has yet be able to halt the rise of obesity.[[2]](#footnote-1) Due to this epidemic, many will face chronic obesity-related health problems like diabetes, heart disease, high blood pressure, cancer, and asthma. In fact, global cost of obesity is estimated to be $2.1trillion annually or 2.8% of global GDP. [[3]](#footnote-2)

The charts below inspired our research. **[[4]](#footnote-3)** While the rise in global obesity is staggering, what we found particularly striking is that its rate and pace of growth differ dramatically by country and gender, even when we adjust for population growth.

|  |  |
| --- | --- |
| **Chart 1: Obese Men by Country** | **Chart 2: Obese Women by Country** |
|  | |

For example the prevalence of obesity in China was effectively 0% in 1975 and is currently hovering around 8%. During the same time the rate in the US went from 12% to 34%. At the same time as many as 40% of Saudi Arabian women are obese, well above their male counterparts at 28%.[[5]](#footnote-4) This demographic puzzle along with a lack of progress in getting the obesity rates under control and a slew of conflicting studies on diet and exercise served as the impetus for the research conducted in this paper.

|  |
| --- |
| ***Research Objective*** |

This study aims to answer the following research question: ***What is the relationship between Adult BMI and diet, exercise, environmental, macroeconomic and behavioral factors across the globe?* We will also look at gender and ethnicity as main effects and in terms of their interaction with each other and Adult BMI**.

Average adult BMI is the **dependent variable** in this study, while the following themes will be tested as the **independent variables** in this study:

**Table 1: Themes & Independent Variables\***

|  |  |
| --- | --- |
| **Theme to Be Tested** | **Corresponding Independent Variables** |
| 1. **Diet** | Protein Consumption  Carbohydrates Consumption |
| **2. Exercise** | Insufficient Activity |
| **3. Environment** | Average Annual Temperature  Air Pollution |
| **4. Ethnicity** | Continents/Regions |
| **5. Gender** | Female/Male |
| **6. Macroeconomics** | GDP, Life Expectancy, Urbanization |
| **7. Behavior** | Alcohol Consumption |

*\*A more detailed description of variables follows in the the methodology section of this paper. The list of the 11 independent variables above was narrowed down from a much broader initial list that contained 27 variables. A discussion in the methodology section of this paper will touch on the process used for the selection of the 11 variables above as well as the steps that were taken to narrow this list down even further*.

Our **conceptual model**  defines average adult BMI as a function of the independent variables in Table 1 plus an error term. The error term will account for the variability that is not explained by the model. The **Null Hypothesis** in this study states there is no statistically significant relationship between average adult BMI and any of the independent variables as defined in Table 1. In other words average adult BMI in each country is not significantly related to protein and carbohydrates consumption, activity level, average annual temperature, air pollution, Gross Domestic Product, life expectancy, urbanization and alcohol consumption rates as well as adult’s gender and region where the reading was taken. The **Alternative Hypothesis** would be that there is a statistically significant relationship between average adult BMI and at least one of the independent variables from the list above.

Prior to this analysis, a strong positive relationship was expected between average adult BMI for each country and GDP, life expectancy, urbanization rates and protein and carbohydrate consumption. A strong positive relationship was expected between adult BMI and percentage of insufficiently active adults in each country. Eating more carbohydrates and protein and not exercising enough should theoretically increase one’s BMI vs. the alternative. Likewise, countries with higher GDP levels, higher life expectancy rates and higher urbanization rates should have higher standards of living combined with lower activity levels that too should contribute to higher average BMI.

Preliminary research into this topic revealed the presence of regional and gender trends. Region and gender variables were included in this analysis as a result. A moderately strong relationship was anticipated. When this project was initially presented, several members of the public suggested ethnicity as an important factor. Ethnicity is an intriguing focus owing to regional, genetic, activity and dietary topics that it brings up. However, this would ultimately expand the project into a much larger research effort than desirable at this stage. The authors believe the included regional, activity and dietary factors will provide an accurate model, with region acting a rough proxy for ethnicity.

Lastly, a minor relationship was expected between BMI and the remaining variables included (average annual temperature, air pollution and alcohol consumption). Temperature and air pollution factors were aimed at capturing the negative endocrine effects believed to have contributed to the global rise in obesity rates. Additionally, these two factors may be capturing one’s ability to exercise outdoors throughout the year. Alcohol consumption was included as a proxy for indulgent behaviour that may contribute to overeating.

|  |
| --- |
| ***Methodology*** |

Publically-accessible data was primarily collected directly from the World Health Organization (WHO), CIA Factbook, World Bank, and agencies of the United Nations such as UNESCO and the Food and Agriculture Organization. Data was also gathered indirectly from Quandl, a data aggregator. The most recent BMI data available at time of analysis was for calendar year 2014. For each independent variable, the most recent data within the preceding 5 years (2010-2014) was taken. Custom python code was developed for data collection, management and processing. SPSS, a proprietary statistics package, was used for analysis.

The initial dataset consisted of 173 countries and 27 independent variables. This dataset had many missing values. To address the missing data, variables were removed which had values for fewer than 75% of the countries (i.e. fewer than 130 values). Variables were also removed if they had low and insignificant correlation to the dependent variable, BMI, or high correlation (>0.7) with another independent variable deemed more robust. These steps did not fully resolve the issue of missing data. As a last resort, countries which were missing data for any of the important variables were removed. Here, an important variable was deemed to be one which had a correlation above 0.4 with the independent variable, BMI. This last step may be introducing bias into our model (see section on limitations). The resulting dataset contains 152 countries, representing 78% of all countries.

For our independent variables, we have selected and aggregated a range of macroeconomic, socioeconomic, environmental and behavioral quantitative metrics. Our extensive research leading up to this study has allowed us to group the independent variables into categories or themes that were discussed in the previous section. However, with 27 independent variables in our original list, we were concerned about overfitting our model and running into multicollinearity issues. Thus, in parallel to resolving our issues with missing data, we also worked on narrowing down our list of variables.

To begin trimming our list of independent variables, we built scatter plots of each variable against the dependent variable, with corresponding fitted lines, to assess the need for transformation.(Please attached file: VariableTransformations.pdf). Table 1, includes five variables which were determined to require transformation (natural log of GDP, air temperature squared, square root of air pollution and conversion of gender and continent variables into dummy variables). An additional 10 variables had transformations performed but were not included in the analysis owing to missing data, having very low correlation with our independent variables or highly correlated with other independent variables that were more robust. This was assessed via a two stage correlation matrix (Appendix Table 7 and 8). Please note that in the second stage of the correlation matrix, we eliminated variables (shown in Table 1). We used a correlation of 0.7 as a cut of criteria for independent variable correlation and favored those with the highest dependent variable correlation (Appendix Table #8).

Body Mass Index is the **dependent variable** in this analysis. It is the widely accepted metric for diagnosing obesity. BMI is defined as weight in kilograms divided by height in meters squared. Table 1 walks through all of our final **independent variables** in more detail as well as the necessary transformations and how we rank ordered the variables based their perceived importance. This is important for our analysis because we apply the hierarchical regression model to test the relationships between our dependent and independent variables.

In deciding the ordering of our variables, we focused on biological metrics, mechanics of weight gain, peer review, and existing research to drive our decisions. We planned to start our analysis by testing gender as the most fundamental differentiator between our observations. Next, our group chose to examine carbohydrate consumption and insufficient activity as the most direct contributors to changes in a person’s weight. During our group presentation, our peers alerted us to ethnicity as a potential driver of BMI differences between countries. We chose to incorporate dummy variables for continents as a proxy for this ethnic variation. Literature review, as well as peer review feedback, drove us to include temperature as another variable of interest. Furthermore, the literature has shown a relationship between pollution and changing metabolic rates. We included air pollution and urbanization to capture this shifting global landscape. We discuss our decision to add the interaction variable (continent\*gender) in the next section.

**Table 2: Final List of Variables, Transformations, Ranks**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dependent Variable** | **Unit** | **Type** | **Applied Transformation** |
| Body Mass Index (BMI) | kg/m^2 (year 2014) | Quantitative | None |
| **Independent Variable** | **Unit** | **Type** |  |
| 1. Male/Female | Male/Female | Categorical | 2 Gender Categories converted to 1 dummy variable (Male=0, Female =1) |
| 2. Carbohydrates Consumption | % of daily calories from carbohydrates | Quantitative | None |
| 3. Insufficient Activity | % underactive of total adults | Quantitative | None |
| 4. Continent | Regions: Africa, Asia, Caribbean, Europe, North America, Oceania, South America, Central America, Middle East | Categorical | 9 Regions Converted to 8 dummy variables (North America = baseline variable) |
| 5. Temperature | Degrees Celsius | Quantitative | Squared Transformation |
| 6. Air Pollution | Particles/Cubic Meter | Quantitative | Square Root Transformation |
| 7. Urbanization | % of Population Living in Urban Areas | Quantitative | None |
| 8. Gender & Continent Interaction | Gender\*Continent | Binary | None |

We selected a **random sample** of 50 countries from our **population** of 152 countries using the Select Cases random sampling code in SPSS. We used the first sample that we got for the regression analysis presented in this paper. We followed up with a second sample of 50 countries using the Select Cases function in SPSS to test our model.We conducted two tailed hypothesis testing and set our significance level at 5%.

To make sure that multiple linear regression is an appropriate approach in our analysis, we first ran tests to make sure our sample data is normally distributed and without highly influential outliers. Kolmogorov-Smirnov and Shapiro-Wilk tests indicated a lack of significant deviation from normality for the entire sample (D100 = 0.086, p = 0.065 and W100=0.984, p=0.280, respectively) (Appendix Table 4). However, when the data is split by gender (Appendix Table 6), there is a significant deviation from normality in the male data (D50=0.141, p=0.012). A significant deviation was not observed in the female data (D50=0.056, p=.20). The sample size (n=50) is sufficiently large enough to allow us to rely on the Central Limit Theorem to overcome our concerns regarding normality within the male set.

Two crucial assumptions of this analysis were the lack of influential outliers and homogeneity of variance. No influential outliers were found to be present in our sample (Appendix Chart #5). Levene’s test was performed to check for homogeneity of variance for the male and female data in our sample. The results of our test did not reach significance (F1,98=0.007, p=.932), which indicated that the assumption of homogeneity of variance was not violated (Appendix Table # 5)**.** We re-checked the assumption of homogeneity, along with several other assumptions, after the completion of our regression analysis.

|  |
| --- |
| ***Regression Results*** |

We ran a 5 - step hierarchical regression, whereby one variable was added at a time in accordance with the rank order described above. Variables were removed from the model when their regression coefficient was not significantly different from 0. This constraint led to the removal of carbohydrates consumption, gender and pollution variables. We also added an interaction variable (gender\*continent). We came to this decision by observing the changes in gender coefficient significance when continent variables were added and a presence of interaction (lack of parallel slopes) in the interaction chart below (Chart #4). Please see appendix page 30 and appendix tables 1A, 2A and 3A for more details about our initial 7- step hierarchical regression, some of the issues that we ran into and resulting variable selection.

|  |  |
| --- | --- |
| **Chart 3: BMI by Gender & Continent** | **Chart 4: Gender & Continent Interaction** |

**The final model was found to predict adult BMI significantly better than no model (F19,80=12.67, p<0.001), with an adjusted R squared of 69% (Appendix Table #1, Table #2)**. The final model indicates that changes in activity levels, air temperature squared, urbanization trends as well as one’s home region and its interaction with one’s gender explain nearly 69% of the variability in adult BMI. The probability of this relationship being by chance alone is <0.1%. Thus, we were able to reject our Null Hypothesis and conclude that a statistically significant relationship exists between Adult BMI and how active adults are in that country, urbanization rates, temperature squared and the region where the reading was taken. We also noted a significant interaction between gender and continent.

The model below and our coefficient table summarize our findings from the final 5-step hierarchical regression with the final list of variables discussed above. Please see Appendix Tables 1,2,3 for more details.

|  |
| --- |
| **Estimated Adult BMI =25.45+0.04\*InsufficientActivity - 3.54\* Africa- 2.27\*Asia -1.56\*Caribbean - 1.72\*Europe-1.16\*MiddleEast -0.09\*Central\_American - 1.43\*Oceania -1.44\*South\_America - 0.002\*Annual\_Temperature\_SQRD +0.03\*UrbanPopulation+Gender\*( 1.37\*Africa+0.34\*Asia +0.79\*Caribbean -1.54\*Europe + 1.00\*MiddleEast + 1.27\*Central\_America + 1.48\*Oceania+0.09\*South\_America)** |

**Table 3**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Step 1** | **B** | **SE B** | **Beta** | **P** |
| **Constant** | **23.36 [ 22.46 , 24.26 ]** | **0.45** |  |  |
| **InsufficientActivity** | **0.095 [ 0.062 , 0.13 ]** | **0.017** | **0.50** | **<0.001** |
| **Note: R2adj = 0.24 Change Statistics: F1,98 = 32.88, p < 0.001** | | | | |
| **Step 2** | **B** | **SE B** | **Beta** | **P** |
| **Constant** | **25.56 [ 23.96 , 27.56 ]** | **0.80** |  |  |
| **InsufficientActivity** | **0.066 [ 0.036 , 0.096 ]** | **0.015** | **0.35** | **<0.001** |
| **Africa** | **-3.44 [-4.96 ,-1.91 ]** | **0.77** | **-0.70** | **<0.001** |
| **Asia** | **-1.74 [-3.36 ,-0.12 ]** | **0.82** | **-0.29** | **0.035** |
| **Carribean** | **-0.79 [-3.24 , 1.66 ]** | **1.23** | **-0.053** | **0.52** |
| **Europe** | **-1.07 [-2.59 , 0.45 ]** | **0.76** | **-0.22** | **0.16** |
| **MiddleEast** | **-0.17 [-1.81 , 1.47 ]** | **0.83** | **-0.028** | **0.84** |
| **Central America** | **0.36 [-1.46 , 2.18 ]** | **0.92** | **0.41** | **0.70** |
| **Oceania** | **-2.43 [-4.89 , 0.037 ]** | **1.24** | **-0.16** | **0.054** |
| **South America** | **-0.88 [-2.61 , 0.86 ]** | **0.87** | **-0.11** | **0.32** |
| **Note: R2adj = 0.54 Change Statistics: F8,90 = 9.03, p < 0.001** | | | | |
| **Step 3** | **B** | **SE B** | **Beta** | **P** |
| **Constant** | **27.38 [ 25.54 , 29.23 ]** | **0.93** |  |  |
| **InsufficientActivity** | **0.062 [ 0.033 , 0.090 ]** | **0.014** | **0.33** | **<0.001** |
| **Africa** | **-3.26 [-4.71 ,-1.81 ]** | **0.73** | **-0.67** | **<0.001** |
| **Asia** | **-2.72 [-4.36 ,-1.09 ]** | **0.82** | **-0.45** | **0.001** |
| **Carribean** | **-0.55 [-2.87 , 1.77 ]** | **1.17** | **-0.037** | **0.54** |
| **Europe** | **-2.57 [-4.25 ,-0.89 ]** | **0.85** | **-0.54** | **0.003** |
| **MiddleEast** | **-0.64 [-2.22 , 0.93 ]** | **0.79** | **-0.11** | **0.42** |
| **Central America** | **0.52 [-1.21 , 2.25 ]** | **0.87** | **0.059** | **0.55** |
| **Oceania** | **-2.04 [-4.38 , 0.30 ]** | **1.18** | **-0.14** | **0.087** |
| **South America** | **-1.15 [-2.80 , 0.50 ]** | **0.83** | **-0.15** | **0.17** |
| **SQRT (Annual Temp)** | **-0.003 [-0.005 ,-0.001]** | **0.001** | **-0.39** | **0.001** |
| **Note: R2adj = 0.59 Change Statistics: F1,89 = 11.61, p = 0.001** | | | | |
| **Step 4** | **B** | **SE B** | **Beta** | **P** |
| **Constant** | **25.35 [ 23.21 , 27.49 ]** | **1.08** |  |  |
| **InsufficientActivity** | **0.050 [ 0.022 , 0.078 ]** | **0.014** | **0.26** | **0.001** |
| **Africa** | **-2.85 [-4.25 ,-1.81 ]** | **0.70** | **-0.58** | **<0.001** |
| **Asia** | **-2.06 [-3.66 ,-0.46 ]** | **0.81** | **-0.34** | **0.012** |
| **Carribean** | **-1.12 [-3.35 , 1.11 ]** | **1.12** | **-0.075** | **0.32** |
| **Europe** | **-2.47 [-4.07 ,-0.88 ]** | **0.80** | **-0.52** | **0.003** |
| **MiddleEast** | **-0.73 [-2.23 , 0.77 ]** | **0.75** | **-0.12** | **0.34** |
| **Central America** | **0.55 [-1.09 , 2.19 ]** | **0.83** | **0.062** | **0.51** |
| **Oceania** | **-0.80 [-3.15 , 1.54 ]** | **1.18** | **-0.054** | **0.50** |
| **South America** | **-1.41 [-2.98 , 0.16 ]** | **0.79** | **-0.18** | **0.078** |
| **SQRT (Annual Temp)** | **-0.002 [-0.004 , 0.00 ]** | **0.001** | **-0.29** | **0.014** |
| **UrbanPopulation** | **0.030 [ 0.012 , 0.04]** | **0.009** | **0.30** | **0.001** |
| **Note: R2adj = 0.63 Change Statistics: F1,88 = 10.85, p = 0.001** | | | | |
| **Step 5** | **B** | **SE B** | **Beta** | **P** |
| **Constant** | **25.45 [ 23.48 , 27.42 ]** | **0.99** |  |  |
| **InsufficientActivity** | **0.043 [ 0.015 , 0.071 ]** | **0.014** | **0.23** | **0.003** |
| **Africa** | **-3.54 [-4.91 ,-2.17 ]** | **0.69** | **-0.72** | **<0.001** |
| **Asia** | **-2.27 [-3.87 ,-0.66 ]** | **0.81** | **-0.38** | **0.006** |
| **Carribean** | **-1.56 [-4.19 , 1.07 ]** | **1.32** | **-0.10** | **0.24** |
| **Europe** | **-1.72 [-3.26 ,-0.18 ]** | **0.77** | **-0.36** | **0.029** |
| **MiddleEast** | **-1.16 [-2.66 , 0.34 ]** | **0.75** | **-0.19** | **0.13** |
| **Central America** | **-0.086[-1.87 , 1.70 ]** | **0.90** | **-0.010** | **0.92** |
| **Oceania** | **-1.43 [-4.12 , 1.26 ]** | **1.35** | **-0.096** | **0.29** |
| **South America** | **-1.44 [-3.10 , 0.22 ]** | **0.83** | **-0.19** | **0.088** |
| **SQRT (Annual Temp)** | **-0.002[-0.004 ,-0.001]** | **0.001** | **-0.29** | **0.007** |
| **UrbanPopulation** | **0.031 [ 0.014 , 0.048 ]** | **0.008** | **0.31** | **<0.001** |
| **Gender\_Africa** | **1.37 [ 0.39 , 2.34 ]** | **0.49** | **0.21** | **0.007** |
| **Gender\_Asia** | **0.34 [-0.91, 1.59 ]** | **0.63** | **0.042** | **0.59** |
| **Gender\_Carribbean** | **0.079[-2.50 , 4.09 ]** | **1.66** | **0.038** | **0.63** |
| **Gender\_Europe** | **-1.54 [-2.47 ,-0.61 ]** | **0.47** | **-0.25** | **0.001** |
| **Gender\_CentralAmerica** | **1.27 [-0.61 , 3.18 ]** | **0.96** | **0.10** | **0.19** |
| **Gender\_Oceania** | **1.48 [-1.82 , 4.78 ]** | **1.66** | **0.070** | **0.37** |
| **Gender\_SouthAmerica** | **0.086 [-1.57 , 1.74 ]** | **0.83** | **0.008** | **0.92** |
| **Gender\_MiddleEast** | **1.00 [-0.25 , 2.25 ]** | **0.63** | **0.12** | **0.12** |
| **Note: R2adj = 0.69 Change Statistics: F8,80 = 3.11, p = 0.004** | | | | |
| **Overall Statistics: F19,80 = 12.67, p < 0.001** | | | | |

Tying together all the work leading up to to the regression analysis, we ran tests to screen for any violations of the key linear regression model assumptions (linearity, normality, homoscedasticity/homogeneity of variance, independence). We tested for Linearity and Homoscedasticity using standardized residuals vs. predicted value plots. We did not find any violations of linearity, normality or homoscedasticity. This conclusion was based on the scatter plot lacking any notable curving or fanning out patterns and the points in the P-P plot followed the ideal diagonal line reasonably well (Appendix Chart #1 and Chart #2). We will also applied Levene’s test to check for homogeneity of variances by gender(Appendix Table 5). We found no significant deviation from homoscedasticity (W1,98=0.007, p=.932). We applied Durbin-Watson test to check for any autocorrelation issues (Appendix Table #1) and ended up with a value of 1.87, which was close enough to the 2.0 threshold indicating that the assumption of independence of the residual terms was met. As discussed earlier we used box plots, histograms plots and applied Kolmogorov-Smirnov and Shapiro-Wilk tests to check for normal distribution (Appendix Table #4,6; Charts #4,5,6,7). Finally, we also checked for multicollinearity issues with our final list of variables by looking at the Variance Inflation Factors (VIF) and noting that none of our final variables were above the 10 threshold (Appendix Table #3).

|  |
| --- |
| ***Implications*** |

The final model has many variables which makes unpacking the implications difficult. If we look at each variable separately, the results appear quite logical, albeit quite different from our initial expectations. While not all of the dummy variables are significant (see the table above), we will discuss the variables that are significant at the 5% level, the initial threshold that was set. It's important to note that this study was not experimental and therefore we cannot prove causation between our dependent variables and the independent variable. Nonetheless, we are capturing the magnitude, the direction and the significance of the relationships between them. When discussing each variable's individual effect below, we assume that all other factors are being held constant.

We found a significant relationship between average adult BMI and Insufficient Activity, Annual Temperature SQRD and Urban Population. A single percentage point increase in under-active adults led to a 0.04 increase in average adult BMI. This relationship is significant (b=0.04 [0.015,0.071], SE=0.014, beta=0.23, t=3.09, p=0.003). Adding insufficient activity to our model had a significant positive effect with adjusted R squared going up 0.25 or 25% (F change 1,98=33, p<0.001). This was in line with our expectations discussed earlier. Similarly, a single percentage point increase in urbanization led to a 0.03 increase in average adult BMI (B=0.03 [0.01,0.05], SE=0.008, beta=0.31, t=3.71, p<0.001). Adding this variable to our model had a significant positive effect with adjusted R squared increasing 0.04 or 4% (F change 1,88=11, p=0.001). This result is also in line with our expectations.

Average air temperature squared proved to have a significant negative relationship with BMI (b=-0.002 [-0.004,0.0001], s.d.=0.001, beta=-0.26, t=-2.75, p=0.007) . This result tells us that a 1 degree celsius increase in average annual temperature squared is associated with BMI going down by 0.002points. This is somewhat confusing as a change of 1 degree celsius squared is equivalent to both an increase and decrease of 1 degree celsius. It follows that people tend to have lower BMI when living in more extreme climates which are either very cold or very hot. Adding this variable to our model had a significant positive effect on our R squared with adjusted R squared increase of 0.05 or 5% (F1,89=12, p=0.001).

Regional and ethnic differences were assessed by comparing against North America. It was found that average adult BMI is 3.54 points lower for men from the African continent than men in North America, all else equal (B=-3.54[-4.91,-2.17], SE=0.69, beta=-0.72, t=-5.13, p<0.001). Similarly, Asian male BMI is 2.27pts lower than North American male average (B=-2.27 [-3.87,-0.66], SE=0.81, beta=-0.38, t=-2.89, p=0.006) and European males have a BMI 1.72 points lower than North American men (B=-1.72 [-3.26,-0.18], SE=0.77, beta=-0.36, t=-2.22, p=0.029). Other continents did not show a significant difference (i.e not significantly different from 0) versus North America. However, the general trend remained negative.

The difference narrowed for some and expanded for others when we looked at the female BMI as can be seen from our gender\*continent interaction variable. We saw two significant relationships emerge. The difference for females in Africa and North America narrowed by 1.37 points from the 3.54 point gap that we saw for their male counterparts. In other words being a woman and being in Africa is associated with BMI 2.17 pts lower than that for a woman from North America, all else equal. This difference in the two gender gaps is significant (B=1.37[ 0.39,2.34], SE=0.49, beta=0.21, t=2.79, p=0.007). This trend reversed when we looked at the European data. The BMI gap between women in Europe and North America was 1.54 wider than that for their male counterparts. In other words, an average European female BMI was 1.72+1.54=3.26 lower than the average BMI for women in North America, this difference between the female and male BMI gaps is significant (B=-1.54[-2.47,-0.61], SE=0.47, beta=-0.25, t=-3.29, p=0.001). These results highlight the importance of gender on comparing regional BMI differences.

While the moderately significant impact from urbanization and air temperature were not too far from our expectations as they each added 0.04 or 4% and 0.05 or 5% to adjusted R squared respectively, we were surprised by the magnitude and significance of continent main effects and their interaction with gender. Recall that our initial expectations were focused mostly on diet and exercise. As we can see from the coefficient table in the previous section, adding the continent variable increased the model’s R squared by 0.33 or 33% (F change 8,90 = 9.03, p<0.001), while the interaction variables increased adjusted R squared by another 0.08 or 8% (F change 8,80=3.11, p=0.004). At the same time, adding the insufficient activity variable increased R squared by 0.24 or 24% (F change 1,98=33, p<0.001), while carbohydrates consumption ended up not being significant in combination with the other factors. More broadly, in our final model gender and region explain more than half of the variability in the global adult BMI.

|  |
| --- |
| ***Application*** |

To test the fit of our model on another sample in our population, we randomly selected another group of 50 countries and applied the formula from our first sample to calculate the **predicted BMI** value below. The results were somewhat disappointing. As the scatter plot between our predicted and actual **Mean BMI by Gender** below demonstrates, the points don’t fall perfectly on a 45% angle line with a 0 intercept, which means that our predicted values deviate from the actual BMI readings. On the positive side, the correlation between our predicted values and the actual readings is fairly high at 0.80 (albeit not 1.0) and variability in the predicted BMI explains 64% of the variability in the actual reading. Hence, while we have not come up with the perfect model, it certainly points us in the right direction and with more time and more robust data, we believe that a more accurate model is possible.

|  |  |
| --- | --- |
|  | Table: 4 |

|  |
| --- |
| ***Limitations*** |

* BMI as a measure of one's weight status has some limitations and varies in accuracy from person to person. (e.g. it does not differentiate between muscle mass and fat).
* Our analysis is static and only applicable to the 2010-14 timeframe. With more time, we would look to continue to test and refine our model as new data becomes available.
* We don’t expect our model to accurately capture metabolic changes worldwide. This is an important variable as many studies have suggested the rise of endocrine issues and the changes in one’s stomach bacteria have had a meaningful impact on the global weight gain.
* There are also cultural aspects that drive the perception of beauty/healthy weight that may not be captured in this model.
* Ethnicity is another variable that could be important. The data that we have come across that splits ethnicity by region is inconsistent and very difficult to disentangle. So instead, we have included the continent variable as a rough proxy for ethnicity. We have also done a lot of reading on the topic of genetics and its impact on weight. The evidence thus far is inconclusive on whether or not genes are to blame for one’s weight issues. The research agrees that the genetic factors have only a small impact on whether or not one becomes obese[[6]](#footnote-5). However, our regression results point otherwise.
* We also understand that despite the credible sources (WHO, CIA, UN, etc.), our dataset may contain inaccuracies especially for smaller countries.
* We understand that our initial population data set may be biased by our decision to remove 21 countries. We removed the following countries from our original data base due to very sparse data: Antigua, Bahrain, Bhutan, Burundi, Comoro, Czech Republic, Equatorial Guinea, Eritrea, Grenada, Ivory Coast, Libya, Papua New Guinea, Qatar, St.Lucia, Truveo, Tonga, Somalia, Singapore, Seychelles, Syria, St.Vincent Grenadine. The lack of robust data could be associated with the less developed status of the country’s economy which might mean our population was more biased towards developed economies. That being said, our resulting population of 152 countries still accounted for 78% of all the countries in the world.

|  |
| --- |
| ***Team Members’ Responsibilities*** |

While we have assigned people to areas that fit their strengths, the expectation was that team members will contribute in all areas.

* Data Aggregation: Eric, Paul, Ryan, Winnie
* Data Analysis: Everyone
* Data Visualization: Winnie, Paul, Ryan, Anna
* Final Paper: Anna, William
* Proposal & Poster: Anna, William, Eric

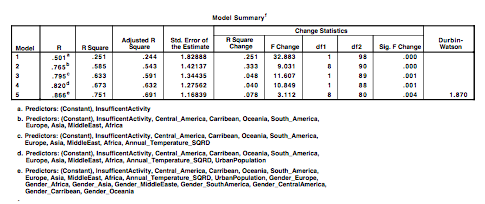
|  |
| --- |
| ***Data Sources*** |

We are using publically available data within 2010-2014 timeframe. We have aggregated data primarily from the World Health Organization, both on their website directly and from Quandl (data aggregator). We have also collected variables from the United Nations, CIA Factbook, the World Bank, UNESCO and Food and Agriculture Organization. A custom Python code was developed to aggregate, process and organize the data. The data was then imported into SPSS for analysis

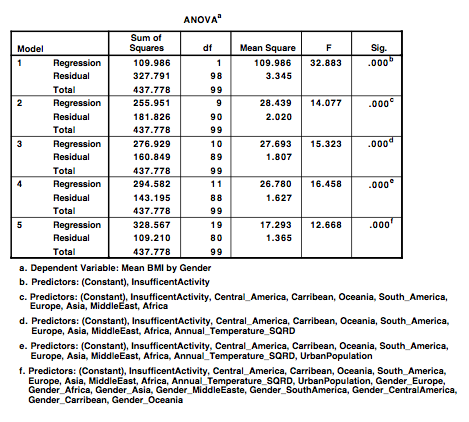
1. WHO:<http://apps.who.int/gho/data/node.imr>
2. Qunadl:<https://www.quandl.com/data/WHO-World-Health-Organization/documentation/documentation>
3. UN: <http://data.un.org>
4. CIA: <https://www.cia.gov/library/publications/the-world-factbook/>
5. World Bank:<http://data.worldbank.org>
6. UNESCO: <http://www.uis.unesco.org/Pages/default.aspx>
7. FAO: <http://www.fao.org/faostat/en/#home>

|  |
| --- |
| ***Appendix 1 : Tables*** |

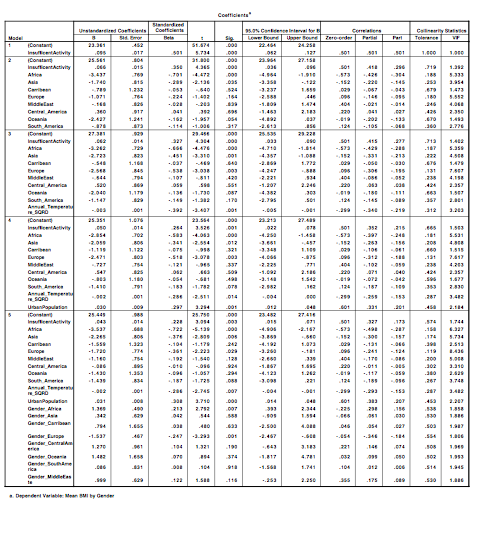
Appendix Table 1: Model Summary



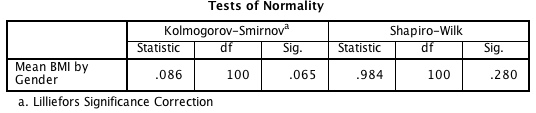
Appendix Table 2: ANOVA



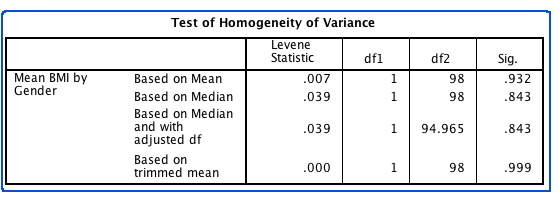
Appendix Table 3: Regression Coefficients



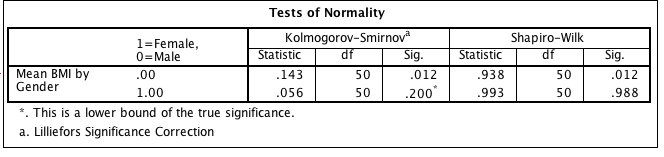
Appendix Table 4: Test for Normality



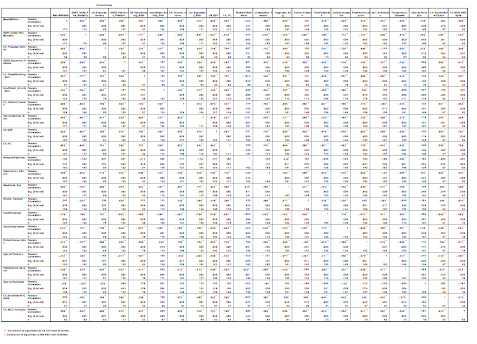
Appendix Table 5: Levene’s Test



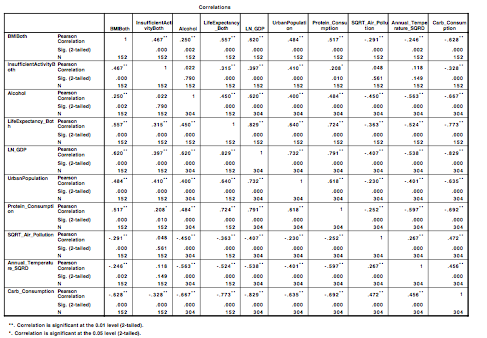
Appendix Table 6: Normality by Gender



Appendix Table 7: First Correlation



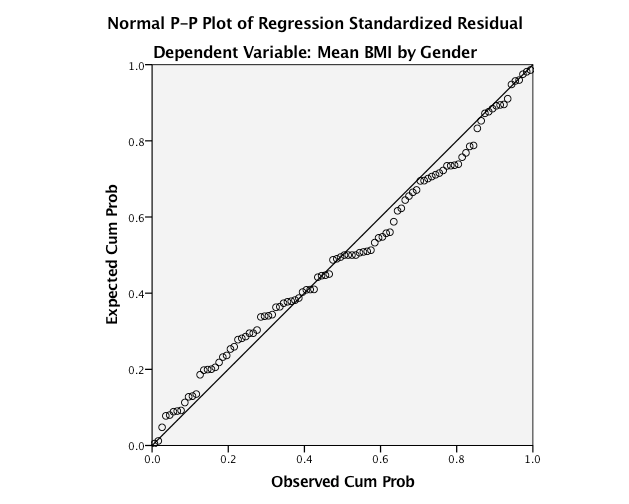
Appendix Table 8: Second Correlation

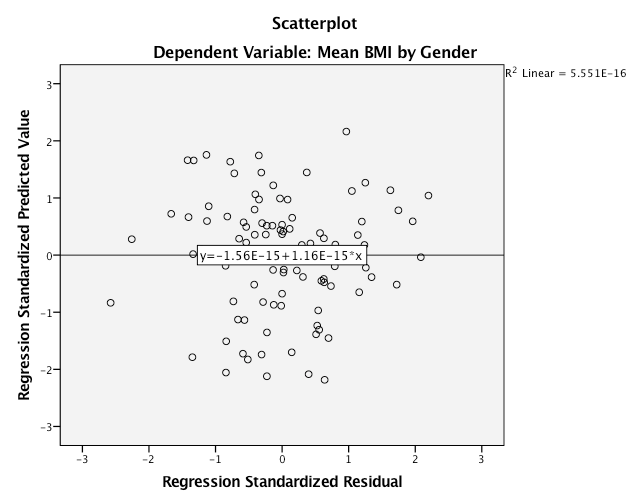


*\*Please let us know if you would like for us to send you the pdf version of these tables as we did with the transformations .*

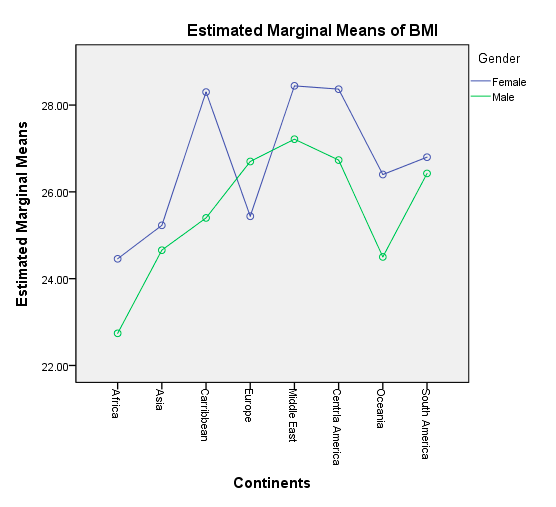
|  |
| --- |
| ***Appendix 2 : Charts*** |

Appendix Chart 1: P-P Plot



Appendix Chart 2: Residual Scatter Plot

Appendix Chart 3: Interaction Between Gender & Continent



|  |  |
| --- | --- |
| Appendix Chart 4: Sample Histogram | Appendix Chart 5: Box Plot by Gender |
| Appendix Chart 6: Histogram Female BMI | Appendix Chart 7: Histogram Male BMI |

|  |
| --- |
| ***Appendix 3 : Regression*** |

**Troubleshooting Our Regression Model**: We ran into several issues before getting to our final model. In the first step of our initial hierarchical regression, the model containing just gender did not predict adult BMI significantly better than no model (F1,98=2.63, p=0.11), and had a very low adjusted R squared of 1.6% (Appendix Table 1.A, Step #1). Normally, these findings would be cause for removing gender from the analysis. The background research strongly suggested there is a link between BMI and gender, therefore we did not remove it. In the last step of the initial hierarchical regression that we attempted, gender does have a significant impact on adult BMI (Gender Coefficient = 3.43[1.17,5.69], s.d=1.13, beta=0.82, t=3.03, p=0.003) (Appendix Table 3A Step#8). This model was able to predict adult BMI significantly better than no model (F22,77=12, p<0.001), with an adjusted R squared of 71%. (Appendix Table 1A, 2A, Step#8). However, due to multicollinearity concerns for gender (Gender’s VIF = 26) along with several continent and interaction variables (Appendix Table 3A, Step #8), a decision was made to remove the gender variable, which while reducing our adjusted R squared to 69% remedied all of our multicollinearity issues. At the same time, the inclusion of the interaction variable gender\*continent still ensured that the gender effect was being captured.

Air pollution and carbohydrates consumption were removed owing to lack of model improvement or having regression coefficients not significantly different from 0. The addition of air pollution did not significantly improve our model (F1,86= 0.01, p=0.91), with an adjusted R square change of 0.0% (Appendix Table 1A, Step #6). Carbohydrates consumption did improve the model when initially added (Adjusted R square change = 26%; F1,97=34, p<0.001) (Appendix Table 1A, step#2), but later in the analysis its regression coefficient became insignificant (B=-0.009 [-0.046,0.029], SE=0.019, beta=-0.053, t=-0.45, p=0.65) (Appendix Table 3A, Step 8). Urbanization does remain in our final model and is highly correlated with carbohydrates consumption (correlation=0.64) (Appendix Table 8, pg 25), therefore it may be capturing some effect by carbohydrates consumption.

**Our Initial Model:**

Table: 1A

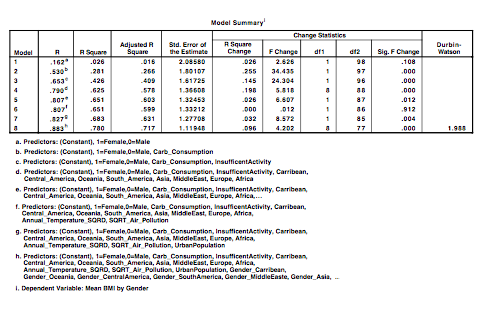


Table: 2A

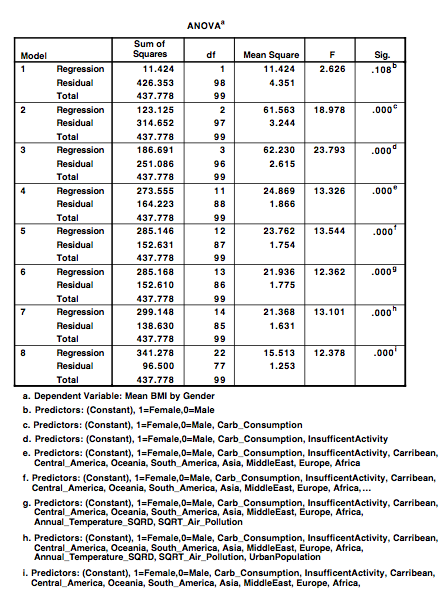
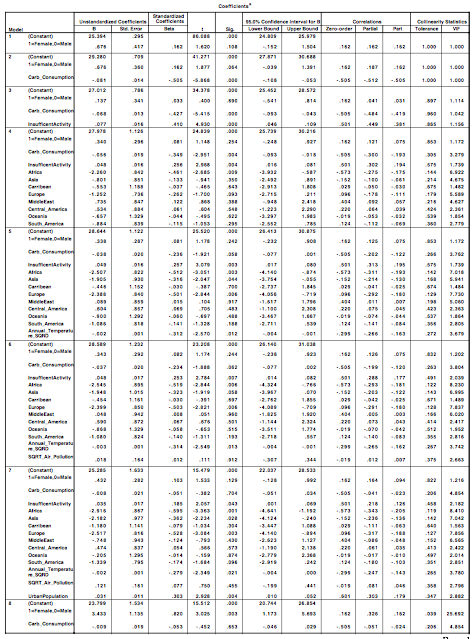
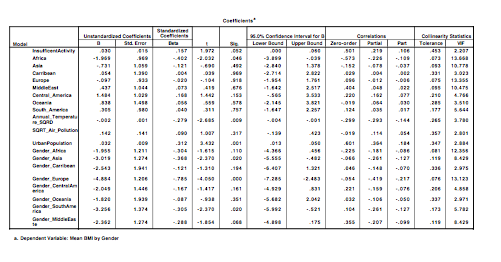


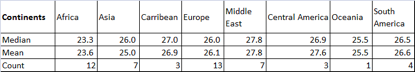
Table: 3A

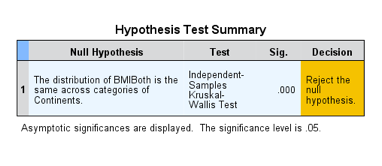


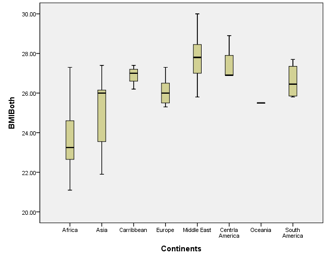
Continued on the next page

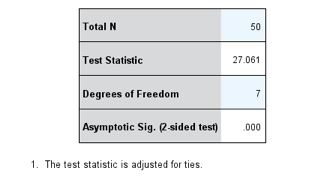
Table 3A continued

 **Result Output for Kruskal-Wallis Test on Continents:**









**Results**: There is a significant difference in BMI of mixed sex as a function of the continents of the country (H7=27.061, p<0.001)

1. “Obesity and Overweight Factsheet”, World Health Organization, <http://www.who.int/mediacentre/factsheets/fs311/en/>, (June, 2016) [↑](#footnote-ref-0)
2. “Patchy progress on obesity prevention: emerging examples, entrenched barriers, and new thinking,”

   The Lancet, <http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(14)61744-X/fulltext>, ( June, 2015) [↑](#footnote-ref-1)
3. “How the world could better fight obesity”, McKinsey Global Institute, <http://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/how-the-world-could-better-fight-obesity>, (November, 2014) [↑](#footnote-ref-2)
4. Stacked Plots, NCD Risk Factor Collarboration, http://www.ncdrisc.org/obesity-prevalence-plot.html (2014) [↑](#footnote-ref-3)
5. Country Profiles, NCD Risk Factor Collarboration, <http://www.ncdrisc.org/obesity-prevalence-plot.html> (2014) [↑](#footnote-ref-4)
6. “Genes Are Not Destiny”, Harvard T.H. Chan School of Public Health, <https://www.hsph.harvard.edu/obesity-prevention-source/obesity-causes/genes-and-obesity/> (accessed November 28, 2016) [↑](#footnote-ref-5)