# Investigating the risk factors associated with BMI: EasyShare data

Max Guy, Katrina Sanders, Kim Johnston

```
library(tidyverse)
library(gridExtra)
library(countrycode)
```

## Main Report

This investigation was carried out equally by the following people with their responsibilities: Katrina Sanders, s1901708, report write up and model building, Kim Johnston, s1936036, report write up, Max Guy, s1945362, writing the code functions.

Our report explores risk factors of obesity using the easySHARE data. The data was collected over many years, based mostly in Europe, about all aspects of life where individuals, couples, and households filled out questionnaires. The main topics of this data were surrounding health, such as social support, childhood conditions, and functional limitation indices. The data has 107 variables with 412110 observations, observed over 8 waves of surveys over 10 years. However, we found the data contained many missing values due to a range of reasons (see Appendix C). This report analyses the data to be able to make suitable conclusions on the following questions: 1. Which of the variables are associated with BMI (an indicator of obesity)? 2. Are the associations between risk factors and obesity the same for males and females?

Our report also considers the strength and nature of the associations between risk factors.

In order to begin our analysis, we chose to work with a subset of the data covering 1 wave and 4 countries. We chose wave 6 along with countries 15, 16, 23, and 25 (which represent Spain, Italy, Belgium, and Israel respectively) to use for our model. The decision process for this is detailed in Appendix A. Wave 6 was carried out in 2015 so the data is still fairly current as well as not being skewed by COVID-19.

During the data cleaning process, we found columns with a large number of "NA" values so decided that any column containing more than 15% "NA" should be removed. This resulted in removing 23 columns. Further data cleaning had us remove "year", "wave number", "mergeid", "hhid", and "coupleid".

We then started a very high level analysis on our subset od data and looked at summary statistics. The first thing we noted of importance was the vast number of variables and observations, in total there are 84 variables, and 18,769 observations.

As we wanted to focus on the factors associating to bmi, we looked at a density plot and found that the data roughly followed a normal distribution, this showed our dependent variable wasn't skewed and was unimodal (mode between 20-30).

We also explored the difference in BMI between males and females. This plot showed that males have a BMI much more concentrated around their mean, however females have a wider spread of BMI values. Something which may be of significance in our linear model.

We decided to take a generative approach to building our model (see Appendix B) where we correlated all the variables not yet in our model individually against bmi and added the most suitable variable to the model. We then evaluated this new model and evaluated the place of the new variable in the model using F-tests and by looking at p-values. If we concluded that the new variable was a good addition to the model, we would repeat this iterative process. The process terminates when adding new variables to the model no longer improves the model. This was determined by the R^2 value and when adding new variables was no longer significant.

Some interesting anomalies we encountered while building the model: When adding the "sphus" variable to the model, it failed our statistical tests (p-value = 0.66), so it was discarded from the model. We noticed high correlations between this variable and others already in the model which may explain the high p-value. Adding "sp00_3_3_mod" to the model resulted in a spike in r^2 to 0.2043 however further analysis revealed that due to too many "NA" values in the model, our number of datapoints for which the model is based upon decreased to less than 100 observations (from an initial almost 19,000 observations) and so it was also rejected as a new predictor. At iteration 13, we had to remove the "gender_partner" variable as the new "female" variable was performing better in the model and there is an interaction between the 2 variables so we can;t include both in our model. We finally decided to stop at iteration 14, as adding the next most correlated variable "sp003_2_mod" resulted in a decreased r^2 as well as increased p-values.

Therefore, our final r^2 value is 0.1401, represented by our final linear model:

From our linear model, we found the 3 most influential variables on bmi to be "female", "mobilityind", "grossmotor". Both "grossmotor" and "female" have estimates of around -1.4 meaning they reduce "bmi", in essence this means that being female results in a lower bmi than being male - which can be seen in our initial analysis. However, unexpectedly this suggests that finding motor activities easier results in a higher "bmi". Then "mobilityind" had the biggest positive effect on "bmi", meaning the less mobile someone is, the higher their "bmi". These are unexpected results as "mobilityind" and "grossmotor" are very similar variables, so we would expect them to have similar effects on bmi.

Finally, we tested the assumptions of our model by creating these plots, and making the following conclusions:

Residuals vs Fitted: The residuals "bounce randomly" around the residuals = 0 line, suggesting that the assumption that the relationship is linear is reasonable. They also roughly form a horizontal band around this line, suggesting that the variances of the error terms are equal. Finally, no residuals stand out from the basic random pattern, suggesting that there are no outliers.

Normal Q-Q: From this plot we can't be sure to assume normality since it shows a right skew. Therefore, this assumption may contribute to our low confidence in the model.

Scale-Location: The red line is approximately horizontal so the average magnitude of the standardised residuals isn't changing much across all fitted values and the spread around the red line doesn't vary much, so the variability is fairly constant.

Residuals vs Leverage: This plot helps us to find influential cases, looking out for outlying values. Since the majority of our points have low leverage they have a weak influence on the coefficients in the regression model.

The other question we wanted to answer regarding the difference in male and females was: How does the model differ if fitted on only female data vs only male data?

Looking at the difference between models fitted on only male data vs only female data. Note, this model is identical to the general model with the obvious exception; the gender predictor is no longer included. The main interesting differences are: Males have a higher base intercept which agrees with the previous coefficient conclusion. There is a notable difference in the adjusted R^2 values of the models with the females having an R^2 of 0.1565 and the males just 0.0671, implying the predictions made for females may be generally more accurate.

Overall, although we have constructed a linear model, we have little confidence in the conclusions it produces. The unexpected conclusions around "grossmotor" and "mobilityind" adds to our reduced confidence. The main reason for this is the low r^2 value, which suggests our variables in the model only explains 14% of the variability in bmi.

## Exectuive Summary