Stastical Case Studies Project 1 part 1

Georgia Ettles, Mayez Haris and Niharika Peddinenikalva

2022-10-28

This project had equal contributions from all members. All 3 members sorted and cleaned the data together along with building the model. Mayez and Georgia refined the models to choose the most suitable one. Niharika wrote the code for box plots and other plots in the project. Georgia and Mayez wrote the analysis of the models while Niharika defined the model and formatted the document.

Introduction

The easySHARE data (https://tinyurl.com/wud2ynbs) is in the form of panel data. This report aims to analyse the data and investigate which risk factors have a significant impact on the participants' BMI, and how these risk factors may affect men and women differently. The data was collected in different waves at different points in time, leading to some individuals participating in the study multiple times, and the questions asked were slightly different for each wave. To reduce repeating entries and to simplify the data that was analysed, this report focuses on wave 5 carried out in four countries - Austria, Germany, The Netherlands, and Sweden - chosen arbitrarily.

The data has 18749 observations, with 8479 men and 10270 women, and the average participants are aged 66 (since the youngest participant was aged 30). The interviews for cities and suburbs had 28.15% of participants. 32.45% participants have graduated secondary school, 80.5% of the participants are able to make ends meet and 94.45% of the participants live in a household of 4 or fewer members. Also, 66.56% of the participants have 0-2 children. 97.37% participants have higher mobility according to the Activities of Daily living index while 57.99% frequently perform vigorous activities. The average BMI of the participants is 26.62. This is in the overweight range.

Defining Model and Checking Model Assumptions

We consider the model below to explore risk factors of obesity. Let, for n observations, $x_{i,j}$ represent the $i = 1, 2, \ldots, n$ observation of the variable x_j in the table below. The risk factors are represented by the explanatory variables x_j and described in Table 1 using the variable name on the easySHARE data guide.

$$\begin{aligned} E(Y) =& \beta_0 + \sum_{j=1}^{22} \beta_j x_j + \beta_{10,3} x_{10} x_3 + \beta_{5,1} x_5 x_1 + \beta_{2,15} x_2 x_{15} + \beta_{2,16} x_2 x_{16} \\ &+ \beta_{2,19} x_2 x_{19} + \beta_{7,18} x_7 x_{18} + \beta_{7,20} x_7 x_{20} + \beta_{7,21} x_7 x_{21} \\ &+ \beta_{12,16} x_{12} x_{16} + \beta_{14,19} x_{14} x_{19} + \beta_{16,1} x_{16} x_1 + \beta_{19,1} x_{19} x_1 \end{aligned}$$

$$\implies E(\mathbf{Y}) = X\boldsymbol{\beta}, \text{where}$$

$$X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,22} & x_{1,10}x_{1,3} & x_{1,5}x_{1,1} & \ldots & x_{1,16}x_{1,1} & x_{1,19}x_{1,1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \ldots & x_{n,22} & x_{n,10}x_{n,3} & x_{n,5}x_{n,1} & \ldots & x_{n,16}x_{n,1} & x_{n,19}x_{n,1} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{22} \\ \beta_{10,3} \\ \beta_{5,1} \\ \vdots \\ \beta_{16,1} \\ \beta_{19,1} \end{pmatrix}$$

x_j	**Variable name**	**Description of variable**	**Possible Values**
x_1	age	Age of participant	range of positive values
x_2	female	Male or female	Male=0, Female =1
x_3	dn004_mod	Born in country of interview	No=0, Yes=1
x_4	partnerinhh	Living with spouse/partner	No=0, Yes=1
x_5	mother_alive	Is mother alive?	No=0, Yes=1
x_6	father_alive	Is father alive?	No=0, Yes=1
x_7	ever_smoked	Ever smoked daily?	No=0, Yes=1
x_8	hc012_	Visited hospital in last 12 months?	No=0, Yes=1
x_9	hc029_	In a nursing home?	No=0, Yes=1
x_{10}	country_mod	Country identifier	Name of each country*
x_{11}	iv009_mod	Area of interview location	City_sub = City or suburbs; Town_rural = Small/Large towns or village
x_{12}	hhsize	Household size	Up to 4; More than 4 people
x_{13}	eurod	Depression scale of EURO-D	Not depressed = 0-3, Somewhat depressed = 4-6, Relatively depressed = 7-9, Very depressed = 10-12
x_{14}	ch001_	Number of children alive	Up to 2 or More than 2 children
x_{15}	chronic_mod	Number of chronic diseases	0; 1-3; 4-10 chronic diseases
x_{16}	adla	Activities of daily living index	High mobility = 0-2; Low mobility = 3-5
x_{17}	br015_	Vigorous activities	Often = Once a week or more; Rarely = Less than once a week
x_{18}	co007_	Is household able to make ends meet?	With difficulty = some or great difficulty; Relatively easily = fairly easily or easily
x_{19}	br010_mod	Drinking behaviour	Rarely = Less than once a month or never; Occasionally = once a month to once or twice a week; Regularly = 3 days a week to almost every day
x_{20}	maxgrip	Maximum of grip strength measure	Low = Less than or equal to 35; High = greater than 35
x_{21}	sphus	Self-perceived health	Good = Excellent/Very good/Good; Not good = Fair/Poor
x_{22}	isced1997_r	Level of Education completed (ISCED-97 coding)	Secondary and below; Post secondary; other = Still in school/other

Table 1: List of explanatory variables

Note: x_2 to x_{22} are categorical variables while x_1 = age is a continuous variable.

Below is the summary of the data that was used for the analysis of the model.

female	dn004_mod	partnerinhh	mother_alive	father_alive	ever_smoked	hc012_	hc029_
No : 8479	No : 1808	No : 4771	No :14474	No :16714	No :8944	No :15463	No :18416
Yes:10270	Yes :16781	Yes:13978	Yes : 4236	Yes : 1883	Yes :9562	Yes : 3219	Yes : 65
	NA's: 160		NA's: 39	NA's: 152	NA's: 243	NA's: 67	NA's: 268

hhsize	sphus	ch001_	adla	br015_
Up to 4 :17709	Good :12702	Up to 2 :12480	High Mobility:18261	Often :10873
More than 4: 1040	Not Good: 6000	More than 2: 6236	Low Mobility : 435	Rarely: 7828
	NA's : 47	NA's : 33	NA's : 53	NA's : 48

br010_mod	chronic_mod	isced1997_r	country_mod	eurod
Rarely :5967	0 : 7109	Secondary and Below:12304	Austria :4279	Not Depressed :14544
Occasionally:7320	1-3 :10605	Post Secondary : 6084	Germany :5750	Somewhat Depressed : 3105
Regularly :5405	4-10: 941	Other : 3	Netherlands:4165	Relatively Depressed: 589
NA's : 57	NA's: 94	NA's : 358	Sweden :4555	Very Depressed : 58
				NA's : 453

co007_	maxgrip	iv009_mod	age	bmi
With Difficulty : 3284	Low :9797	City_sub : 5278	Min. : 30.30	Min. :12.86
Relatively easily:15093	High:7698	Town_rural:12715	1st Qu.: 58.80	1st Qu.:23.62
NA's : 372	NA's:1254	NA's : 756	Median : 65.80	Median :25.95
			Mean : 66.43	Mean :26.62
			3rd Qu.: 73.30	3rd Qu.:29.00
			Max. :100.70	Max. :81.23
			NA's :1	NA's :292

The data above contains `NA`s which were filtered out while constructing the male and female models.

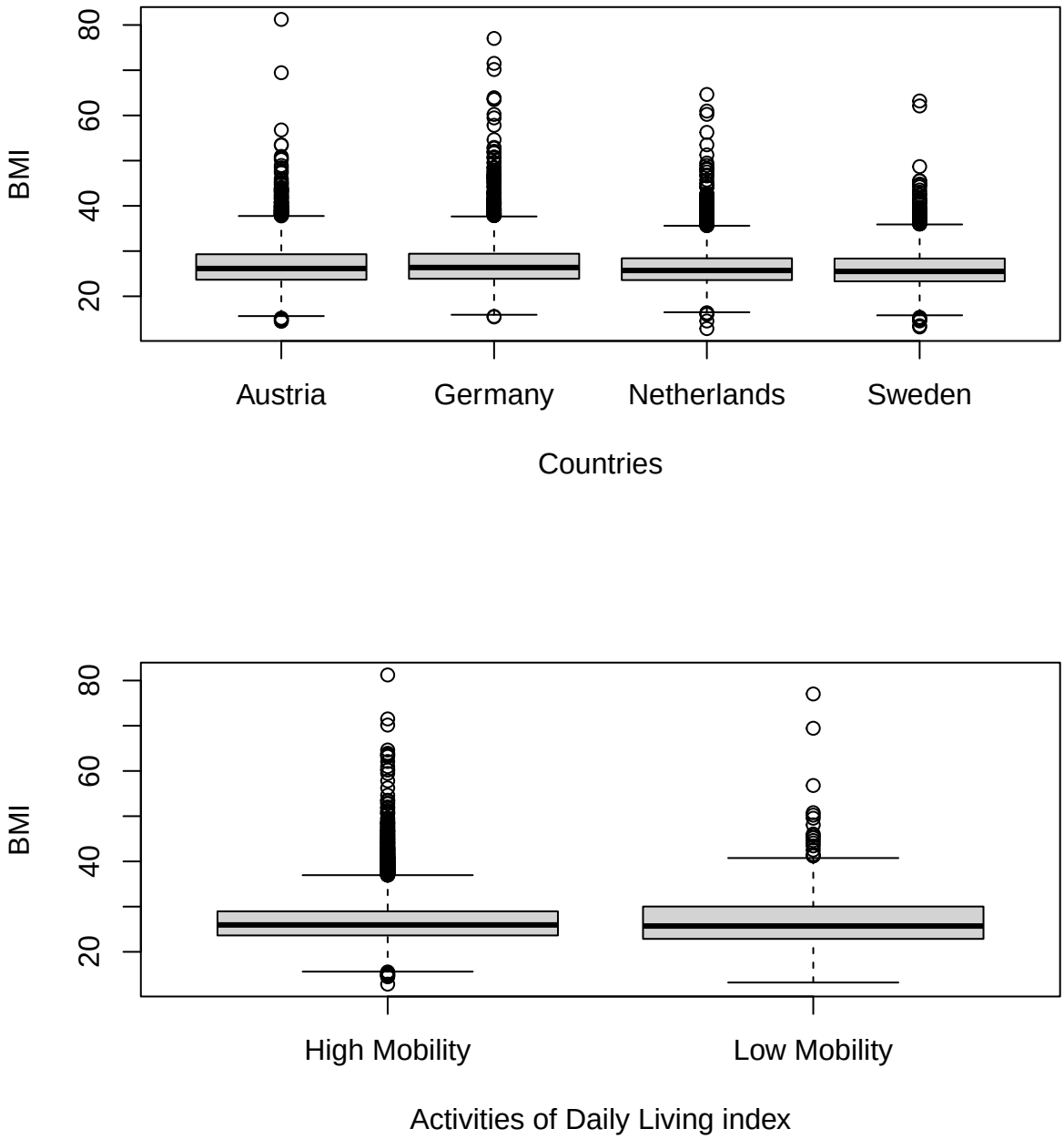

Figure 1: BMI varying across Countries and daily living activities

From the boxplots in Figure 1, it is evident that although there isn't a major difference in the value of BMI in each country, there is a difference in the range of values. Austria and Germany seem to have a larger range of values than The Netherlands and Sweden. It is also evident that there is not a large amount of variation in the BMI of those who have high mobility and those who have low mobility. However, it is worth noting that there is a larger range of values of BMI for those who have high mobility.

The model assumptions of the initial model without separating the data by sex are displayed in Figure 2.

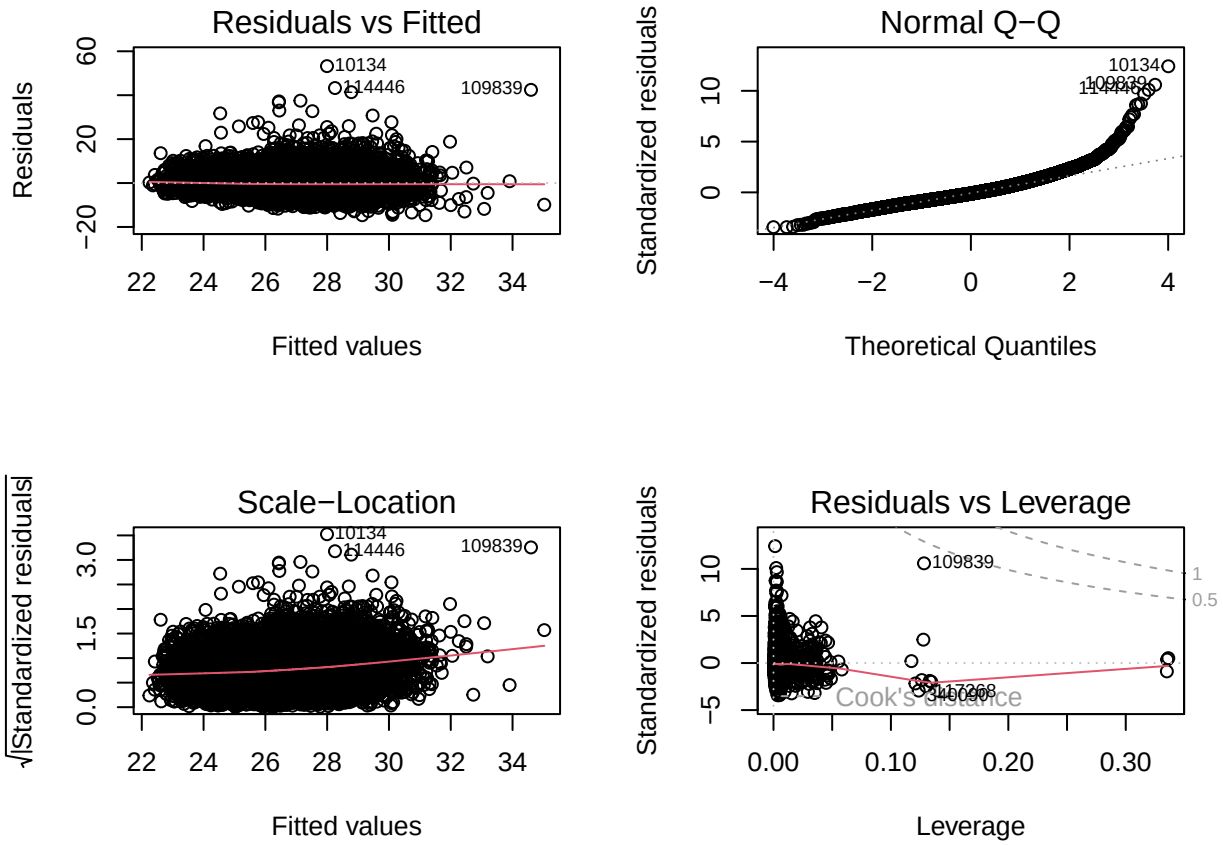

Figure 2: Model assumptions for the defined model

The residuals vs fitted plot shows a constant line at zero, justifying the assumption that the expectation is zero. The Normal Q-Q plot is positively skewed, suggesting that the model may be over-fitted. The Scale-Location plot shows only a very slight upwards trend, suggesting that the assumption of a constant variance is not far-fetched. The residuals vs leverage plot shows no data points with high residuals and high leverage, and most of the data points are very close to zero, suggesting that they all have a very similar impact on the model.

Now, the separate models for males and females may be examined. The summaries for the male and female models are show below.

Model for Males

```
##
## Call:
## lm(formula = bmi ~ partnerinhh + mother_alive + ever_smoked +
##     hc029_ + iv009_mod + hhsize + eurod + chronic_mod + adla +
##     br015_ + co007_ + br010_mod + maxgrip + sphus + isced1997_r +
##     age + country_mod + mother_alive:age + ever_smoked:co007_ +
##     adla:age + br010_mod:age + ever_smoked:maxgrip + ever_smoked:sphus +
##     hhsize:adla, data = na.omit(male_data))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.422  -2.461  -0.434   1.906  52.885
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              31.51       0.82   38.54   <2e-16 ***
## partnerinhhYes                            0.21       0.12    1.73     0.08 .
## mother_aliveYes                          -5.79       1.12   -5.16   <2e-16 ***
## ever_smokedYes                           -0.65       0.35   -1.84     0.07 .
## hc029_Yes                                 2.20       1.05    2.10     0.04 *
## iv009_modTown_rural                       0.43       0.10    4.14   <2e-16 ***
## hhsizeMore than 4                        -0.04       0.20   -0.21     0.83
## eurodSomewhat Depressed                   0.07       0.15    0.45     0.65
## eurodRelatively Depressed                -0.84       0.37   -2.30     0.02 *
## eurodVery Depressed                      -3.53       0.97   -3.65   <2e-16 ***
## chronic_mod1-3                            1.25       0.10   12.18   <2e-16 ***
## chronic_mod4-10                           2.67       0.22   12.03   <2e-16 ***
## adlaLow Mobility                          6.10       3.00    2.03     0.04 *
## br015_Rarely                              0.59       0.10    5.89   <2e-16 ***
## co007_Relatively easily                  -0.56       0.23   -2.44     0.01 **
## br010_modOccasionally                    -0.12       0.85   -0.14     0.89
## br010_modRegularly                       -2.66       0.85   -3.12   <2e-16 ***
## maxgripHigh                               0.50       0.22    2.29     0.02 *
## sphusNot Good                             1.18       0.18    6.47   <2e-16 ***
## isced1997_rPost Secondary                -0.66       0.10   -6.78   <2e-16 ***
## isced1997_rOther                          3.38       3.89    0.87     0.38
## age                                      -0.08       0.01   -7.85   <2e-16 ***
## country_modGermany                       -0.18       0.13   -1.39     0.17
## country_modNetherlands                   -0.74       0.14   -5.27   <2e-16 ***
## country_modSweden                        -0.98       0.14   -6.94   <2e-16 ***
## mother_aliveYes:age                       0.09       0.02    4.91   <2e-16 ***
## ever_smokedYes:co007_Relatively easily    0.64       0.27    2.33     0.02 *
## adlaLow Mobility:age                     -0.09       0.04   -2.26     0.02 *
## br010_modOccasionally:age                 0.00       0.01    0.39     0.70
## br010_modRegularly:age                    0.04       0.01    2.84   <2e-16 ***
## ever_smokedYes:maxgripHigh                0.58       0.26    2.24     0.03 *
## ever_smokedYes:sphusNot Good             -0.74       0.22   -3.39   <2e-16 ***
## hhsizeMore than 4:adlaLow Mobility        6.85       2.03    3.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.886 on 7295 degrees of freedom
```

```
## Multiple R-squared:  0.1098, Adjusted R-squared:  0.1058
## F-statistic:  28.1 on 32 and 7295 DF,  p-value: < 2.2e-16
```

Model for Females

```
##
## Call:
## lm(formula = bmi ~ dn004_mod + mother_alive + father_alive +
##     ever_smoked + hc012_ + iv009_mod + hhsize + ch001_ + chronic_mod +
##     adla + br015_ + co007_ + br010_mod + maxgrip + sphus + isced1997_r +
##     age + country_mod + dn004_mod:country_mod + br015_:age +
##     ever_smoked:co007_ + br010_mod:age + ever_smoked:maxgrip +
##     ever_smoked:sphus + ch001_:br010_mod, data = na.omit(female_data))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.988  -3.004  -0.520   2.388  43.298
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              29.84       0.81   37.03   <2e-16 ***
## dn004_modYes                              0.52       0.36    1.44     0.15
## mother_aliveYes                          -0.39       0.14   -2.86   <2e-16 ***
## father_aliveYes                          -0.38       0.17   -2.18     0.03 *
## ever_smokedYes                           -0.44       0.26   -1.69     0.09 .
## hc012_Yes                                -0.22       0.14   -1.53     0.13
## iv009_modTown_rural                       0.19       0.11    1.70     0.09 .
## hhsizeMore than 4                        -0.39       0.23   -1.70     0.09 .
## ch001_More than 2                         0.96       0.17    5.66   <2e-16 ***
## chronic_mod1-3                            1.76       0.11   16.08   <2e-16 ***
## chronic_mod4-10                           2.55       0.27    9.30   <2e-16 ***
## adlaLow Mobility                          1.25       0.48    2.60     0.01 **
## br015_Rarely                              2.65       0.70    3.77   <2e-16 ***
## co007_Relatively easily                  -0.83       0.18   -4.55   <2e-16 ***
## br010_modOccasionally                    -3.37       0.75   -4.50   <2e-16 ***
## br010_modRegularly                       -3.38       0.95   -3.57   <2e-16 ***
## maxgripHigh                               0.40       0.22    1.79     0.07 .
## sphusNot Good                             1.60       0.16   10.14   <2e-16 ***
## isced1997_rPost Secondary                -0.62       0.11   -5.41   <2e-16 ***
## isced1997_rOther                          1.97       3.26    0.61     0.55
## age                                      -0.07       0.01   -6.19   <2e-16 ***
## country_modGermany                        0.72       0.42    1.71     0.09 .
## country_modNetherlands                   -0.09       0.55   -0.17     0.86
## country_modSweden                         0.71       0.50    1.43     0.15
## dn004_modYes:country_modGermany          -1.03       0.45   -2.32     0.02 *
## dn004_modYes:country_modNetherlands      -0.01       0.57   -0.01     0.99
## dn004_modYes:country_modSweden           -1.38       0.52   -2.65     0.01 **
## br015_Rarely:age                         -0.03       0.01   -2.99   <2e-16 ***
## ever_smokedYes:co007_Relatively easily    0.75       0.26    2.85   <2e-16 ***
## br010_modOccasionally:age                 0.04       0.01    3.64   <2e-16 ***
## br010_modRegularly:age                    0.03       0.01    2.36     0.02 *
## ever_smokedYes:maxgripHigh                0.83       0.31    2.65     0.01 **
## ever_smokedYes:sphusNot Good             -0.97       0.22   -4.30   <2e-16 ***
```

```
## ch001_More than 2:br010_modOccasionally    -0.43       0.24   -1.83     0.07 .
## ch001_More than 2:br010_modRegularly       -0.77       0.29   -2.69     0.01 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.592 on 8603 degrees of freedom
## Multiple R-squared:  0.1187, Adjusted R-squared:  0.1152
## F-statistic: 34.08 on 34 and 8603 DF,  p-value: < 2.2e-16
```

From the summaries above, the adjusted R-squared values for both models are low, but this is due to variability in the data. After performing model selection procedures, some explanatory variables for both the male and female models were removed, with there now being 18 common variables to both models, 6 and 7 variables being exclusive to males and females respectively. The relationship between BMI and each of the variables for both males and females is shown in the tables below, where a positive or negative relationship shows that the value of the variable is proportional or inversely proportional respectively to the value of BMI. "Varying for different levels" suggests that each level of the variable may have a different relationship with BMI.

Variable Name	**Relationship (Male)**	**Relationship (Female)**
mother_alive	Negative	Negative
ever_smoked	Negative	Negative
iv009_mod	Positive	Positive
hhsize	Negative	Negative
chronic_mod	Positive	Positive
adla	Positive	Positive
br015_	Positive	Positive
co007_	Negative	Negative
br010_mod	Negative	Negative
maxgrip	Positive	Positive
sphus	Positive	Positive
isced1997_r	Varying for different levels	Varying for different levels
age	Negative	Negative
country_mod	Negative	Varying for different levels
ever_smoked:co007_	Positive	Positive
br010_mod:age	Positive	Positive
ever_smoked:maxgrip	Positive	Positive
ever_smoked:sphus	Negative	Negative

Table 6: The relationship between BMI and variables common between both sexes.

Variable Name	**Relationship**
partnerinhh	Positive
hc029_	Positive
eurod	Varying for different levels
mother_alive:age	Positive
adla:age	Negative
hhsize:adla	Positive

Table 7: The relationship between BMI and the variables exclusive to males.

Some variables did not show a very significant relationship with BMI, for example `country_mod`. However, when considering this variable in interactions with other variables, a significant relationship was present. Hence, the country where the interview was conducted may not have made a difference on its own, but it did have a significant impact on other variables such as `dn004_mod`. This was true for other variables as well,

Variable Name	Relationship
dn004_mod	Positive
father_alive	Negative
hc012_	Negative
ch001_	Positive
dn004_mod:country_mod	Negative
br015_:age	Negative
ch001_:br010_mod	Negative

Table 8: The relationship between BMI and the variables exclusive to females.

including `hhsize`. These variables could be considered as confounding variables.

Corresponding to Tables 6,7 and 8, the boxplots below (in Figure 3) help visualise the relationship between BMI and some of the explanatory variables.

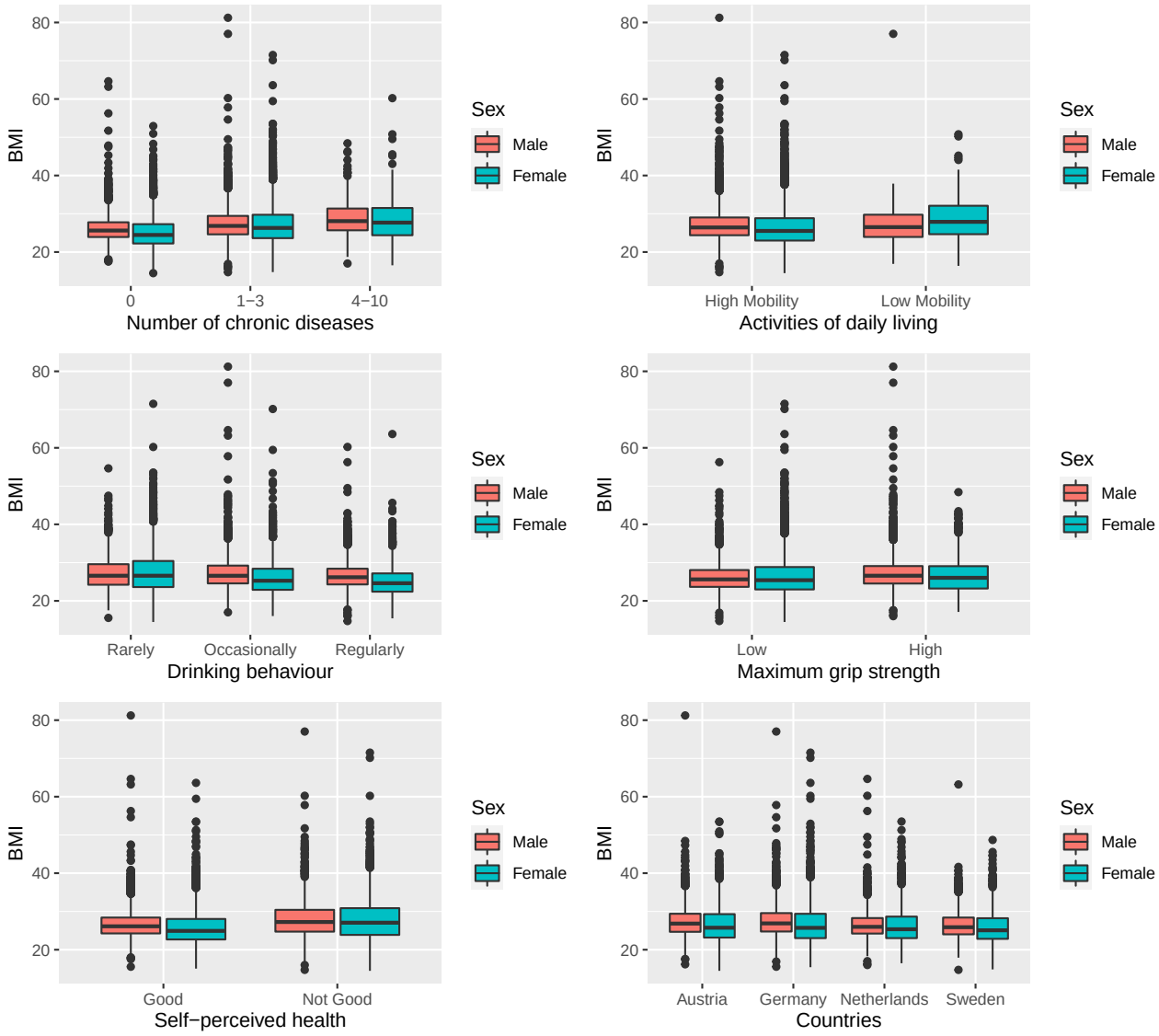

Figure 3: Potential risk factors of obsesity varying with BMI

Conclusion

From the given data, there are some risk factors that affect both males and females, for example smoking habits, level of daily activity, number of chronic diseases, etc. However, some factors exclusively affect males, e.g. depression (EURO-D score) or their age in interaction with their daily activities. On the other hand, some factors exclusively affect females, e.g. number of children, or the amount of vigorous activities in interaction with their age.

Executive Summary

This report illustrates some of the risk factors associated with obesity with BMI as a measure. The data includes participants from four countries - Austria, The Netherlands, Germany and Sweden. When comparing the effect of different risk factors for males and females, it was evident that there is generally a difference in how each factor affects each sex. For example, males with a EURO-D score of 10 or higher were much more likely to have a smaller BMI than those who were less depressed, suggesting that depression may not have much of an impact on their BMI until it becomes significantly severe. On the other hand, there was no apparent relationship between BMI and severity of depression for females at all. The number of chronic illnesses that participants had from each sex seemed to increase their risk of obesity. In addition to this, participants who were generally more active and had a healthier lifestyle, and perceived themselves to be healthier may have a lower risk of being obese. Surprisingly, those who drank regularly had a lower chance of being obese. Participants from Austria and Germany generally had higher values of BMI on average than The Netherlands and Sweden. Being able to make ends meet financially seemed to increase the risk of being obese for both males and females, with the risk being slightly higher for females. Some of this information is visually represented by Figure 4.

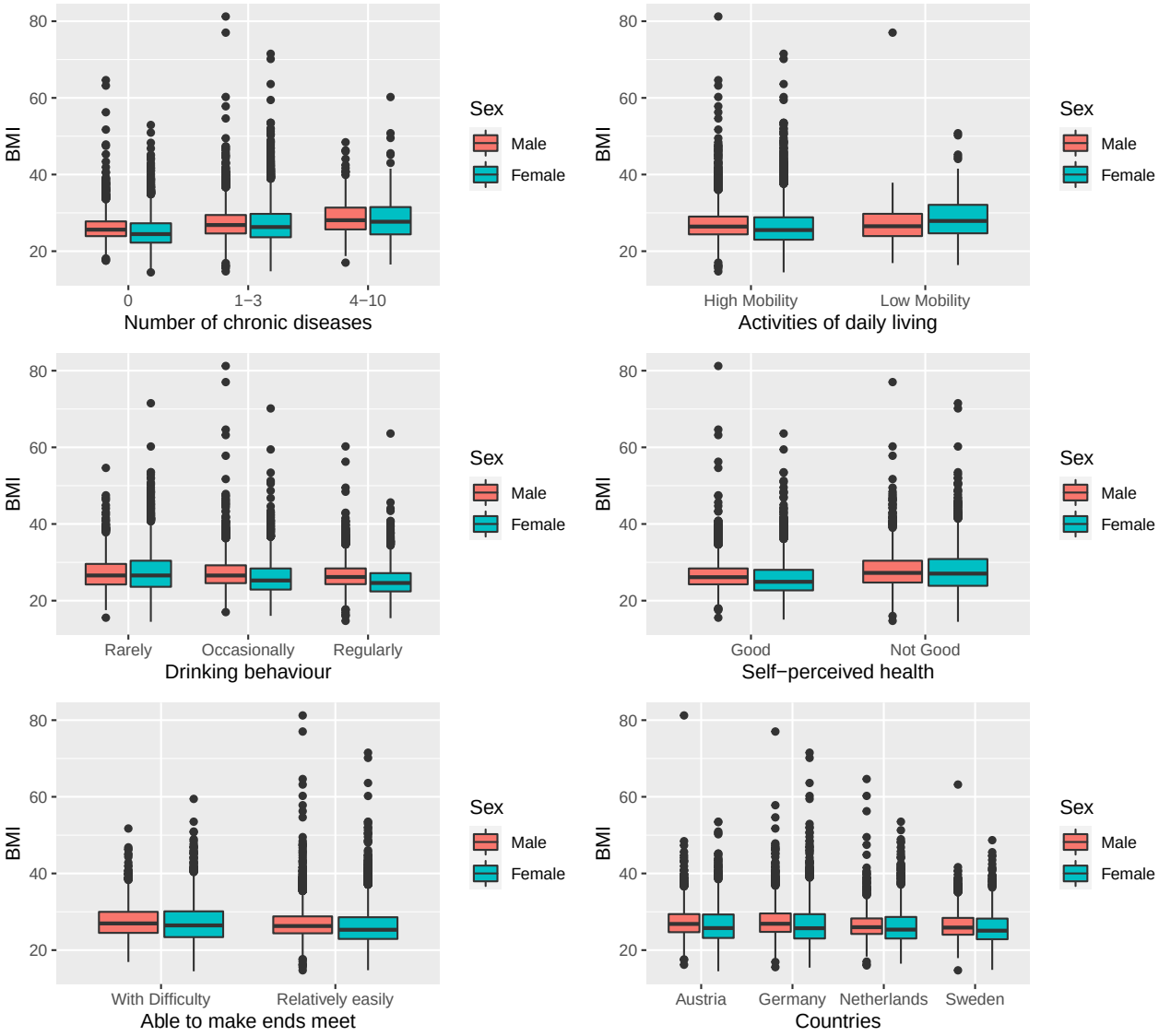

Figure 4: Possible risk factors of obsesity