Jinane Amal MQE1

UNIVERSITÉ PARIS DAUPHINE – PSL COURSE M1 QE1: MEASUREMENT ISSUES FINAL PAPER 2024

1. BRIEFLY DESCRIBE THE DATASET FOR YOUR COUNTRY OF CHOICE, THE SAMPLING DESIGN AND REPRESENTATIVENESS. PROVIDE DESCRIPTIVE STATISTICS OF THE MAIN VARIABLES OF INTEREST (HOUSEHOLD INCOME, GENDER, AGE, EMPLOYMENT,…) AND NUMBER OF OBSERVATIONS. (1 PT)

2. ***DESCRIPTIVE MEASURES OF INEQUALITY AND POVERTY BASED ON HOUSEHOLD INCOME*** : GRAPH THE AVERAGE HOUSEHOLD INCOME IN YOUR COUNTRY OF REFERENCE FOR WAVE 7. CHOOSE 2 OTHER COUNTRIES IN THE DATABASE AND REPORT THE AVERAGE INCOME IN EACH OF THE TWO COUNTRIES FOR THE SAME WAVE. CAN YOU RANK THE THREE COUNTRIES IN TERMS OF INCOME LEVEL? DISCUSS THE STATISTICAL SIGNIFICANCE OF THE RANKING. (3 PTS)

3. CHOOSE NOW YOU WILL FOCUS ON YOUR COUNTRY OF REFERENCE. PROPOSE THREE DIFFERENT MEASURES OF INCOME INEQUALITY AND COMPUTE THEM IN YOUR COUNTRY OF REFERENCE. DISCUSS WHAT YOU LEARN IN TERMS OF INEQUALITY FROM THESE THREE MEASURES AND WHAT ARE THE PROS AND CONS OF EACH MEASURE. (4 PTS)

4. PROPOSE THREE DIFFERENT MEASURES OF POVERTY AND COMPUTE THEM IN YOUR COUNTRY OF REFERENCE. COMPARE AND DISCUSS THESE THREE MEASURES. (4 PTS)

5. ***MEASURES OF INEQUALITY BASED ON HEALTH OUTCOMES AT THE INDIVIDUAL LEVEL :*** PROPOSE A MEASURE OF HEALTH INEQUALITY AT THE INDIVIDUAL LEVEL BASED ON THE VARIABLES THAT ARE AVAILABLE IN THE DATA. JUSTIFY YOUR METHODOLOGY WITH A SURVEY OF THE LITERATURE. IMPLEMENT YOUR MEASURE ON YOUR COUNTRY OF REFERENCE AND COMPARE IT WITH YOUR MEASURES OF INCOME INEQUALITY. (4 PTS)

6. WHAT CAN EXPLAIN HEALTH INEQUALITIES ACROSS INDIVIDUALS? PROPOSE AN ECONOMETRIC ANALYSIS AND DISCUSS YOUR RESULTS (4PTS)

1/ Description of the DATA set :

The data set we're studying is from the generated easy SHARE data set release version 8 gathered from 2004 for wave 1 till 2020 for the last wave, in which multiple central variables have been identified to cover the diverse seven share topic's questions, from demographics , household composition, social support and network, childhood conditions, health and health behavior to functional imitation indices and money and work structure .

We have 2 variables specific to the year of interview the "int_year" and one differentiating between the two data collection methodologies: longitudinal and baseline interview "int_version".

Looking at the initial data set, we have 412110 observations of 107 variables. Throughout this whole paper though , Spain will be the reference country.

Thus, proceeding in data cleaning, we drop "int_year" as all the individuals have been interviewed the same year (2017) , filter by keeping only Spain's data in wave 7 then erase this wave column. Then we drop the "langage" and the "country_mod" to keep only one country identifier since the "country" code is the same as the "langage" code for Spain.

We currently have 4704 observations for 102 variables after dropping the named variables above.

Now providing descriptive statistics of the main variables of interest (household income, gender, age employment,…) for Spain :

We have around 3088 different households, 1834 couples living in a same household 31 of which moved from the previously lived in same household .

The median age of the participants is 74 years which explains that 54% of respondents are already retired.

Approximately 58% of respondents are female and the median income is 17107euros which we got after getting rid of the missing values captured by (-10).

As far as education is concerned, the mean of education years for this sample after getting rid of missing values is 7 years .

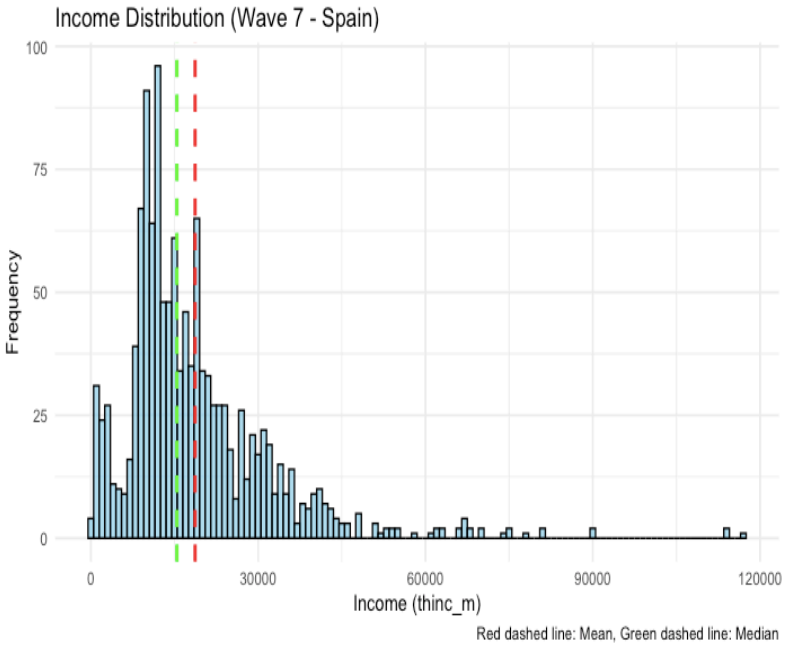
Regarding the strength, we look at the "maxgrip" variable from which we drop the missing values and we compute the mean which is 26.24 kg.

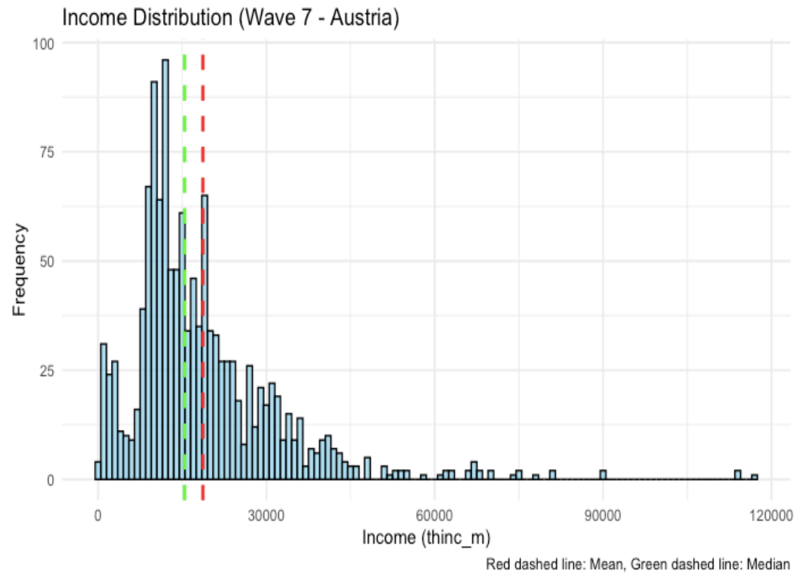
ep005_ <int>	count <int>	proportion <dbl>
-15	24	0.01875000
-12	2	0.00156250
1	686	0.53593750
2	88	0.06875000
3	19	0.01484375
4	60	0.04687500
5	356	0.27812500
97	45	0.03515625

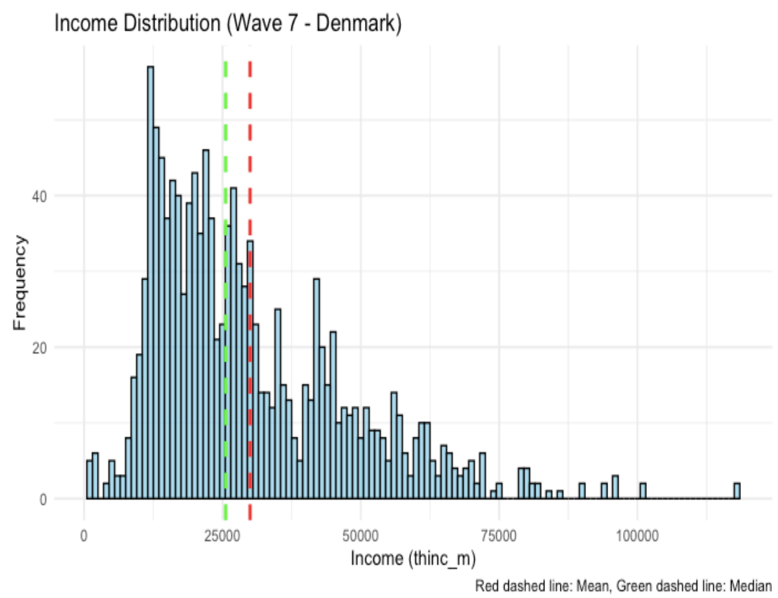
```
[1] "Proportion of females: 0.58125"
[1] "Mean of eduyears_mod after filtering: 7.25204731574158"
[1] "Mean of max_grip after filtering: 26.2420494699647"
```

Variable <chr>	Min <dbl>	Max <dbl>	Mean <dbl>	Median <dbl>	SD <dbl>
Age	46.6	98.0	73.74661	74.38984	16.91341
Income	0.0	516154.9	97300.39889	17107.71463	205353.07312

2/ Descriptive Measures of inequality and poverty based on household income :







After getting rid of the extreme values which were badly affecting our statistics table along with the income distribution (as seen in the code lines above this one), we get a more reliable table statistics describing Spain's average population income :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
23.84	10319.84	15435.66	18693.92	23799.54	116803.65

We first got rid of the zeros which could represent participants' refusal to divulgate their income. As for the highest extreme value : there is only one individual who has an income of 516154.94€, this may be true but also may just be a typing error. Either way it's not representative of Spain's population average income, so we drop this value.

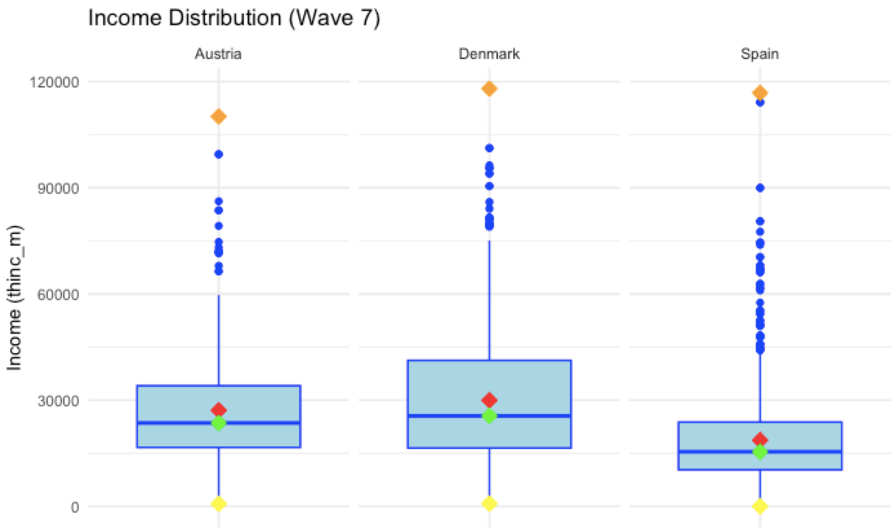
Thus, for our Country of reference "Spain", we notice a right skewed distribution with a median of 15435€, lower than its mean of 18693.92€.

In the same data cleaning process, we get rid of the out layers (the zeros and the two positive extreme values) of the Austria Data set, then we graph the income distribution. We notice that the right skewed distribution for Austria is quite similar to Spain's with a median of 23550€, lower than its mean of 27142.1€.

As for Denmark, we will only need to drop the zeros as the data doesn't contain any extremely high value. Its income distribution is also a right skewed distribution with a median of 2554€, lower than its mean of 29984.1€, but it's more widely spread : As we can see, the frequency level for this graph compared to the two others does not go over 60% for each income value in contrary to Spain and Austria in which most income values close to the median can easily go over 60%.

We can also represent the income distribution using a boxplot so as to better grasp the central and spread tendency of the data which we will use to clearly compare Spain with the two other countries: Austria and Denmark, so as to also provide a ranking at the income level .

Looking at these boxplots side by side, we notice that Denmark has a higher IQR range meaning, greater variability and spread around the median which we've already noticed in the graph above. This can also mean that there is a certain heterogeneity in the data .



Thanks to these histograms and boxplots, we can classify these countries in term of income level :

Indeed, we focus on the mean and the median to determine the wealth ranking of these countries.

Here Denmark, then Austria have respectively the highest mean/median values. Therefore, Denmark is the richest followed by Austria then Spain.

```
              Df    Sum Sq   Mean Sq F value Pr(>F)    
Country        2 8.419e+10 4.210e+10   168.1 <2e-16 ***
Residuals   3023 7.571e+11 2.504e+08                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = thinc_m ~ Country, data = combined_data)

$Country
                     diff         lwr       upr     p adj
Denmark-Austria  2841.980    857.6361  4826.325 0.0022893
Spain-Austria   -8448.195 -10436.1782 -6460.211 0.0000000
Spain-Denmark  -11290.175 -12761.3249 -9819.025 0.0000000
```

We tested for statistic differences between these countries using ANOVA to accept or reject this income ranking hypothesis. Indeed, the ANOVA outcome suggests that there is a statistical difference in income levels since p<0.05, and these differences are not solely due to chance. Since ANOVA is significant, we then we run the post-hoc test table: It provides pairwise comparisons between countries and their differences in means.

3/ The three different measures of inequality and their interpretations:

I suggest three different measures of income inequality :

First, the *GINI coefficient* measuring income/wealth distribution across a population, more precisely the extent to which the distribution of income within a country deviates from a perfectly equal distribution: a 0 coefficient corresponds to the “same income” perfect equality while a 1 coefficient corresponds to perfect inequality where only one individual holds all the income.

After computing the GINI coefficient for our “filtered_wave7dataSP”, we found a value of 0.32 which means that there is adequate equality with some disparities among households in Spain that could be further enhanced .

According to OECD, the *Palma ratio* is the share of all income received by the 10% people with highest disposable income divided by the share of all income received by the 40% people with the lowest disposable income. When we have a Palma ratio exceeding one, that stipulates that the top 10% holds more income than the bottom 40%. The higher the Palma ratio above one, the higher the inequality within the country’s population wealth distribution .

When computing this second measure to our “filtered_wave7dataSP”, we find a value of 1,156 ~=1.2, indicating that the top 10% holds 1.156 times more income than the bottom 40%, reflecting a somewhat moderate inequality in the country.

Last but not least, the *P90/P10 ratio* comparing the income at the 9th percentile to the 10th percentile .

The higher P90/P10 ratio, the greater disparity there is between the high earners' income and the low earners.

By computing this third measure to our "filtered_wave7dataSP", we find a value of 4.61, indicating that the 90th percentile earns 4.61 times more than those of the 10th percentile which clearly puts forth the big disparities between the two groups.

Comparing between the two percentile ratios : Palma vs P90/P10, we can notice that the disparities are strongest when comparing between the richest and poorest then it is when using palma which is less focused on extreme groups .

Overall, these three measures both have advantages as well as disadvantages :

On one hand, the GINI coefficient is the simplest measure of income inequality providing one easy to interpret summary statistic value . However, this measure can be too general in the sense that it does not capture changes in some specific groups of the sample and can be widely influenced by the middle values of the income distribution .

On the other, the Palma ratio and the P90/P10 ratio are useful to understand income disparities within certain groups (Palma: top 10% vs bottom 40% ; P90/P10: 90th percentile vs 10th percentile) highlighting by direct comparison the extreme inequality between groups, for instance: the income gap between the highest and lowest earners of the sample (P90/P10).

But since these two methods only focus on certain groups they might leave out important information and changes regarding income distribution across other groups, therefore creating a certain bias.

Different measure techniques have their strengths and weaknesses , researchers, marketers and policy makers should thus use multiple measure to better understand the topic of study, here income disparity within a country of reference (Spain) so as to find solutions to the brought up issues or design targeted interventions to resolve those issues just like we would like here to promote a greater income equality .

4/ The three different measures of poverty and their interpretations:

There are two types of poverty : Absolute and Relative.

Absolute poverty is when a household income is a state of being in which a person lacks the necessities of life, such as food, clothing, and shelter whilst relative poverty is a condition in which a person's income is insufficient to meet his or her needs when compared to others in society.

In this second sense (relative poverty) the three different measure methodologies of poverty I used are the following :

We'll keep in mind that usually due to differences in estimation methodologies and poverty lines, estimates should not be compared across countries.

First, "Poverty headcount ratio" at national poverty lines (% of population) which according to "The World bank" represents the percentage of population living below the national poverty line based on population-weighted subgroup estimates from household surveys (here the "easy share release 8" data set).

I decided for simplicity reasons to set the poverty line at 60% of the median income as a threshold:

Meaning that we are in this sense, we are computing the sample's percentage living with income significantly below the median who are struggling the most to access basic needs due to their below poverty line income .

After computing, we find a poverty headcount ratio of 21.5%, meaning out of 100, 21.5 people live below the poverty line which in this case is 60% of the median income.

Second, we have the poverty gap index another measure of the degree of poverty.

According to the OECDE, it is the ratio by which the mean income of the poor falls below the poverty line. We get this value by computing the mean of the gaps between the poverty line and the incomes of individuals who are below the poverty line.

The result we get through the PGI is 0.0527 which indicates that on average Spanish individuals living below the poverty line fall below the poverty line by 5.27% .

The third poverty index used is actually complementary to the previous one and it's called: squared poverty gap index.

This squared poverty gap index also known as poverty severity index according to the UNESCWA is calculated by averaging the square of the poverty gap ratio. By squaring each poverty gap data, the measure puts more weight the further a poor person's observed income falls below the poverty line.

Therefore, the SPGI indicates the average squared income shortfall of the poor relative to this chosen poverty line.

The result we get through this third method highlighting the intensity of poverty is 0.0213 which suggests that the average squared income shortfall relative to the poverty line is 2.13% which is quite low synonym to low severity of poverty in Spain.

Now let's break down the advantages and limits of these used poverty measures:

While the poverty headcount ratio is an easy direct measure of the population proportion living below the poverty line which can be used by policy makers to track the poverty progress through the years, it does not capture the depth of poverty among the poor and the differences between .

This is where the poverty gap index can be a complementing measure as it accounts for the intensity of poverty among the poor by computing the average income shortfall relative to the poverty line which could be used by the policy makers if they want to target the vulnerable groups to reduce these income shortages . But then this measure doesn't account for these income shortfall's distribution among the poorest which could be an interesting information.

Then we have squared poverty gap index, which can be useful in the case of significant income disparities among the poor; since by squaring income shortfalls we focus on the severity of poverty .

In conclusion, I believe it is always better to employ different measures to get a better overview and in-depth insights from the sample studied: PHR here will provide a more general overview on Spain's poverty prevalence while PGI & SPGI will provide more details on the depth of this poverty in Spain based on the sample's individuals studied.

5/ Measures of inequality based on health outcomes at the individual level:

Our goal here is to measure health inequalities by comparing at the individual level the health of those in the lowest socio-economic group with those in the highest group.

The concentration index is a measure of accessibility distribution evaluating the length to which accessibility inequalities are systematically associated with individuals' socioeconomic levels.

This concentration index is also defined with reference to the concentration curve (twice the area between the concentration curve and the line of equality (the 45-degree line)) (Kakwani 1977, 1980).

This index varies between -1 and +1 (when all accessibility is concentrated in the most or in the least disadvantaged person, respectively) and in the case in which there is no socioeconomic-related inequality, the concentration index is 0.

Computing the concentration index from micro-data (C) we simply use the "convenient covariance" result : $C = 2 \operatorname{cov}(y_i,R_i) / \mu$, where y is the health variable whose inequality is being measured, μ is its mean, Ri is the ith individual's fractional rank in the socioeconomic distribution (e.g. the person's rank in the income distribution), and cov(.,.) is the covariance.

The concentration index estimated using both aggregated or micro data is now becoming the standard measure to quantify income-related inequalities in health economics (Wagstaff and van Doorslaer, 2000). But this index is still better computed using the micro data just like the one we are using through the filtered "wave7dataSP".

If the health variable is a "bad" one such as ill health, a negative value of the concentration index means ill health is higher among the poor.

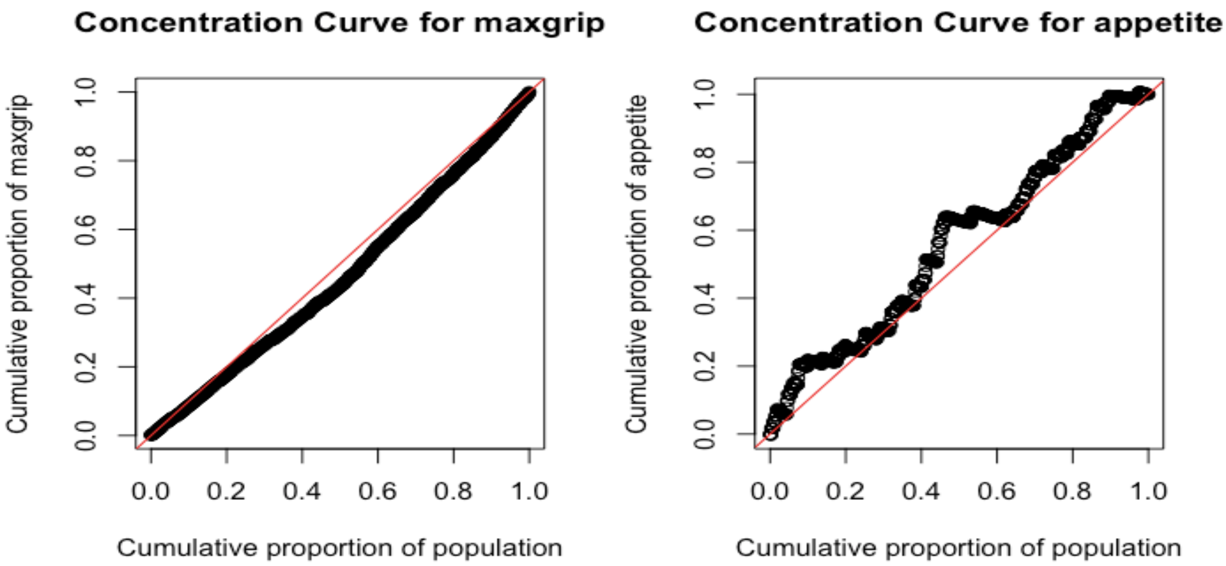
From our health variables present in the data share wave 7, we have chosen the following to compute their respective concentration index in relation to the ranked income: Doctor visits annually 'hc002_mod' ; Self-perceived health 'sphus'; the number of chronic diseases 'chronic_mod'; the appetite variable 'euro8'; the strength variable 'maxgrip' ; the hospital stays number variable 'hc012_'

On one hand, we find that concentration Index for number of doctor visits (–0.025), for self-perceived health status (-0.031), for number of chronic diseases (-0.029), for hospital stays (0.001) are all four practically null denoting no systematic pattern of inequality regarding these studied types of health outcomes based on the individuals' income. So regardless of income, we have the same weighted : doctor visits , number of hospital stays, number of chronic diseases and self-perceived health status outcomes

On the other, we find that the health outcome 'maxgrip' evaluating strength has a positive concentration index of 0.1: Thus, individuals with higher income are somewhat stronger .

And for the appetite variable 'euro8' we notice a negative CI of -0.2 suggesting tat individuals with lower income are hungrier which makes sense.

Here are the concentration index curves linked to these two variables :



We can't really compare these CI with the three measures of income inequality used in question 3 because concentration index measures income disparities based on health access/ outcomes whilst the other 3 measures only focus on the income, how it is distributed and the depth of the income inequality .

However, these different measures can nevertheless be intertwined and influence each other .

We could tell by looking at the gini coefficient 0.32 that most health outcome concentration indexes will be close to 0 since this coefficient is quite low and synonym to adequate equality within Spain's households therefore these individuals would have on average equitable health access and health outcomes: Afterall , the gini coefficient and the concentration index are computed in a same way but only vary in their range interval (Drs. Carlos Castillo-Salgado).

<u>6/ Explaining inequalities across individuals through a regression</u>

To better grasp the cause of these inequalities across these individuals, we will compute a linear regression which is a data analysis technique that predicts the value of a dependent variable based on an independent variable. The greater the linear relationship between the independent variable and the dependent variable, the more accurate the prediction is.

We chose as a dependent variable: 'casp' which measures the quality of life and well-being index of individuals to assess how the quality of life of the studied individuals could be influenced by explanatory variables such as income (wealth) and more health focused independent variables: 'mobilityind', 'maxgrip' etc . After computing different regressions with different sets of explanatory variables, we finally found one which could be interesting enough .

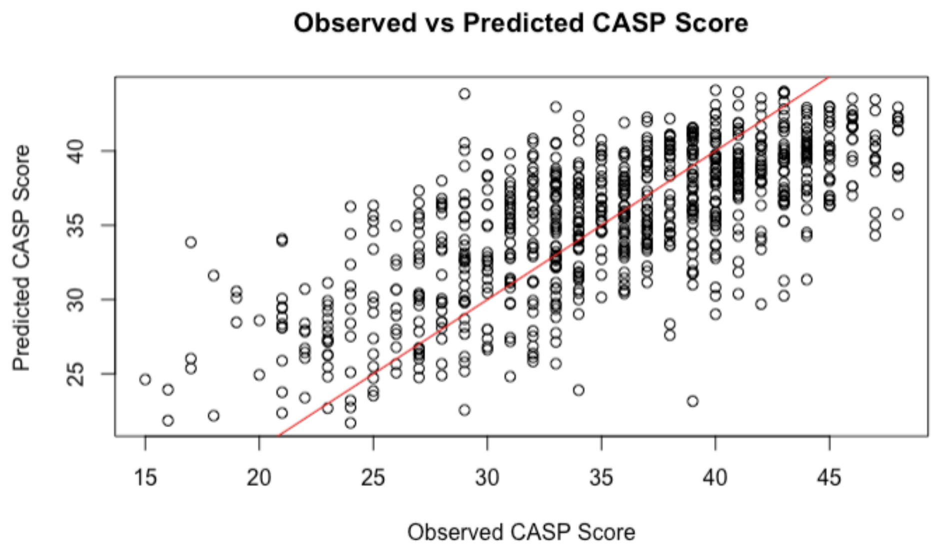
Of course, prior to computing the final regression tests results, we want to make sure that we've filtered any data value underneath 0 which alludes to missing values that could create major bias in our regression.

```
lm(formula = casp ~ mobilityind + lgmuscle + numeracy_2 + smoking +
    maxgrip + euro2 + euro5 + euro6 + euro10 + chronic_mod +
    sphus + thinc_m, data = filtdata)

Residuals:
     Min       1Q   Median       3Q      Max
-16.8557  -3.1503   0.2228   3.3525  15.8342

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.038e+01  1.217e+00  33.192  < 2e-16 ***
mobilityind -5.211e-01  2.367e-01  -2.202  0.02796 *
lgmuscle    -2.238e-01  2.129e-01  -1.051  0.29358
numeracy_2   3.234e-01  1.035e-01   3.124  0.00185 **
smoking      2.452e-01  1.449e-01   1.692  0.09103 .
maxgrip      3.938e-02  2.088e-02   1.886  0.05972 .
euro2       -3.298e+00  4.206e-01  -7.840 1.45e-14 ***
euro5       -9.811e-01  3.758e-01  -2.611  0.00920 **
euro6       -2.797e+00  4.919e-01  -5.686 1.82e-08 ***
euro10      -2.084e+00  4.210e-01  -4.950 9.08e-07 ***
chronic_mod -8.800e-02  1.515e-01  -0.581  0.56154
sphus       -1.585e+00  2.179e-01  -7.277 8.19e-13 ***
thinc_m      4.500e-05  1.496e-05   3.008  0.00271 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.883 on 793 degrees of freedom
Multiple R-squared:  0.4961,    Adjusted R-squared:  0.4885
F-statistic: 65.07 on 12 and 793 DF,  p-value: < 2.2e-16
```



Regression results interpretation:

When all explanatory variables are null, the 'CASP' score is at 40.38 which is initially quite high for everyone . Now taking into account the independent variables, for each unit increase in 'mobilityind' (for which the higher the index, the more difficulties with activities), the 'CASP' score decreases by approximately 0.521 , for each unit increase in numeracy_2 (test on Mathematical performance), the ‘CASP’ score increases by approximately 0.323.

The “euro” s variables have significant negative coefficients, indicating that higher values of these variables are associated with lower ‘CASP’ scores. ‘Sphus’ the self-perceived health goes from 1 being excellent to 5 poor therefore, for each unit increase in ‘sphus’, the ‘CASP’ score decreases by approximately 1.585 and for each unit increase in ‘thinc_m’ (income), the ‘CASP’ score increases by approximately 0.000045. As for ‘lgmuscle’, ‘smoking’, ‘chronic_mod’ ; there is no significant relationship between them and ‘CASP’ score (p-value > 0.05) so we may neglect them.

The coefficient of determination (R-squared) is 0.4961, indicating that approximately 49.61% of the variance in the CASP score is explained by the independent variables employed in the model.

Additionally, thanks to the F-statistic p-value (< 2.2e-16), we can conclude that the model as a whole is statistically significant with several variables such as ‘numeracy_2’, self-perceived health (sphus), and certain mental health indicators (e.g., euro2, euro5, euro6, euro10) have significant associations with the ‘CASP’ score, while others like ‘lgmuscle’, ‘smoking’, and ‘chronic_mod’ do not appear to significantly impact the ‘CASP’ score.

In conclusion, there are many factors that could explain inequalities across individuals.

Some of them might be individual mental health status, some important physical health criteria, behavioral risks, cognitive function and many more that we might have disregarded in this regression.

We should also know that the factors affecting the inequality, impact it at a different scale from one another . We should also be careful with data cleaning and analysis as we could easily misuse a database creating false results and therefore failing to address the underlying determinants of the issue studied or planning interventions aiming to reduce health outcome disparities in our case for instance.

Important note :

To respect the number of 10 maximum pages for this final paper, I decided to add some notes to better understand my view and thinking process on some questions which you can find written above the line codes dedicated to each of those questions: You can thus, read them on the R quarto document or the pdf file generated from this quarto .

Additionally, not all descriptive statistics tables or graphics are shown in this final synthesis so please refer to the quarto document and generated pdf file to get a better overall overview of my work .

Thank you for taking the time to review my work and sharing your expertise as a professor while teaching us in Measurements' Issues Class.

Sources and references :

Relative or absolute inequality : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8992138/

R use for CI : https://search.r-project.org/CRAN/refmans/accessibility/html/concentration_index.html

The importance of CI : https://countdown2030.org/documents/Country_workshops/concentration_index.pdf

Measuring health inequalities : https://www3.paho.org/english/sha/be_v22n1-Gini.htm

Quantitative Techniques for Health Equity Analysis : https://countdown2030.org/documents/Country_workshops/concentration_index.pdf

Analyzing health equity : https://www.worldbank.org/content/dam/Worldbank/document/HDN/Health/HealthEquityCh8.pdf

Calculating the concentration index when income is grouped : https://pure.eur.nl/ws/portalfiles/portal/46815521/2012-10-11+16_26_51.pdf

Palma ratio: https://data.oecd.org/inequality/income-inequality.htm#:~:text=The%20Palma%20ratio%20is%20the,with%20the%20lowest%20disposable%20income.

Poverty headcount ratio at national poverty lines (% of population) : https://data.worldbank.org/indicator/SI.POV.NAHC

Handbook on poverty statistics (concept & methodologies & policy use) : https://unstats.un.org/unsd/methods/poverty/pdf/un_book%20final%2030%20dec%2005.pdf

Unescwa Economic and social commission for Western Asia : https://archive.unescwa.org/squared-poverty-gap-index

Word Bank : Analyzing Health equity with concentration curves : https://www.worldbank.org/content/dam/Worldbank/document/HDN/Health/HealthEquityCh7.pdf

Kakwani, Nanak, and Hyun H. Son, 'Concentration Curves', *Economic Inequality and Poverty: Facts, Methods, and Policies* (Oxford, 2022; onlineedn, OxfordAcademic,22Sept.2022):https://doi.org/10.1093/oso/9780198852841.003.0009

Van Doorslaer E, Wagstaff A, van der Burg H, Christiansen T, De Graeve D, Duchesne I, Gerdtham UG, Gerfin M, Geurts J, Gross L, Häkkinen U, John J, Klavus J, Leu RE, Nolan B, O'Donnell O, Propper C, Puffer F, Schellhorn M, Sundberg G, Winkelhake O. Equity in the delivery of health care in Europe and the US. J Health Econ. 2000 Sep;19(5):553-83. doi: 10.1016/s0167-6296(00)00050-3. PMID: 11184794.

Henao, S. & Arbelaez, M . 2001. Epidemiological Surveillance System for Acute Pesticide Poisoning. Epidemiological Bulletin, Vol. 22 No. 4, 4-11 December 2001.

Measuring Health Inequalities:Gini Coefficient and Concentration Index: Prepared by Drs. Carlos Castillo-Salgado, Cristina Schneider, Enrique Loyola, Oscar Mujica, Ms. Anne Roca and Mr. Tom Yerg of PAHO's Special Program for Health Analysis (SHA).-from *Epidemiological Bulletin, Vol. 22 No. 1, March 2001*-