uc3m | Universidad Carlos III de Madrid

Master Degree in Statistics for Data Science
2019-2020

Master Thesis

Development of a Statistical Methodology for the Study of the Impact of Inequalities in the Genetic Predisposition of Academic and Social Success

Marta Cortés Ocaña

1st Javier González Peñas
2nd María Luz Durbán Reguera
Madrid, 2020

Abstract

The technological and medical progress has allowed to extract the information of genetic variants from the human genome and study how those are distributed in populations in order to be able to identify the risk for developing a certain disease or analyze heritage and genetic transfer of mutations and particular phenotypes.

In addition, the development of genome-wide association studies (GWAS) has facilitated the discovery of gene-environment interactions (GxE). In those, an statistical maximum likelihood method is applied to estimate the contribution of genetic variables, such as intelligence or educational attainment scores and its interactions between other features and genetic environmental factors that the subjects have been exposed to.

In 2017, the largest GWAS meta-analysis of intelligence, which included 'only' 78,000 individuals, developed a Genome-wide Polygenic Risk Scores (PRS), IQ2 (second IQ study), which finally broke the 1% barrier of previous GWAS of intelligence by predicting 3% of the variance of intelligence in independent samples. However, IQ2 still has less predictive power than the 4% of the variance explained by the Educational Attainment PRS. In addition, the predictive power of the EA PRS jumped to more than 10% of the variance in preliminary analyses with the EA3 GWAS which was derived from a sample size of more than 1 million. We can expect a similar jump in the predictive power of the IQ PRS when the sample size for GWAS meta-analyses of intelligence exceeds 1 million[1].

Other examples of previous studies applying these methods have achieved to explain 5.2% of the body mass index (BMI) variance, with an additional 1.9% after GxE interactions[2]. Or have reached explained variance of the ability (2.1%), education achievement (5.2%) and family SES (socioeconomic status) (6.6%) by the years of education PRS[3].

In this work, the monthly incomes, social class and education success are tried to be explained by the IQ PRS, cognitive performance PRS and EA PRS analysing the comparison of its performance and the additional contribution of environmental factors.

Keywords: intelligence, cognitive performance, educational attainment, academic success, social class, genetics, environment, gene-environment interaction, polygenic risk scores, childhood trauma, cannabis, explained variance, multiple regression, ordinal logistic regression.

Contents

1 Objectives

This work will be developed in three steps that involves the next objectives:

- Firstly, we will estimate the aggregated genetic scores related with educational attainment, cognitive performance and intelligence through polygenic risk models over a target population (EUGEI consortium).

  In order to do that, we need to form ourself in genetic common variation, genetic architecture of psychiatric disorders and calculation of polygenic components by RStudio and Plink.

  In addition, we need to calculate the ancestrally components and the scores of each subject in principal components form.

- Secondly, we need to analysis the magnitude in which the estimated genetic components explain indicators of social class or academic success.

  To do so, we aimed to analyse the information of the questionnaire where the data of the target subjects is collected, selecting and highlighting the relevant one, and matching it with the excel database of the project.

  After that, we will employ linear and logistic models estimating the variance metric (R2 Naggelkerke) by which the genetic components explain the next traits included in the EUGEI databse: Monthly Incomes, Social Class and Educational Level.

- Finally, we will evaluate the impact of traumatic event during the childhood, drug consumption or other personal components, and its interaction with the polygenic risk scores, in the context of academic and social success predictions.

  Then, we will compare the magnitude in which the estimated genetic scores explain the dependant variables in presence or not of traumatic variable, country of birth or gender. Then, we will apply additive regression models with gene-environmental (GxE) variables in order to study how those affect and impact in the explanation of the indicators of social class and academic level.

2 Theoretical Framework

2.1 Human Genome and Genetic Heritage

In 1860, the Swiss scientist Johann Friedrich Miescher determined for the first time the molecule DNA. Since then, many scientist have been trying to understand more about its structure and components. In 1881, Albrecht Kossel identified the 'nuclein' as a nucleic acid and he also isolated the five nucleotide bases, after that, he presented its official name: deoxyribonucleic acid (DNA).

A DNA molecule consists in two rolled chains among each other, forming an double helix structure, each of these chains has a central part called nucleotide which contains a sugar group (deoxyribose), phosphate group and a nitrogen base. There are four types of nitrogen bases: adenine (A), thymine (T), guanine (G) and cytosine (C). The order of these bases determines the genetic code. Human DNA has around 3 billion bases, and more than 99 percent of those bases are identical in all people. Due to it, DNA molecules are very long, they can't fit into cells without the right compression, in fact, DNA is coiled firmly forming the chromosomes structure. In total, humans have 23 pairs of chromosomes which are found inside the cell's nucleus.

The completed set of DNA is known as genome, it includes around 25,000 of genes and contains all of the instructions needed to build, live, maintain and reproduce the organism. It is confined in the nucleus from where the genetic information is transferred through the mechanisms of gene expression. These mechanisms set the process by which DNA is used to make ARN or proteins, responsible, to perform various important functions in the body.

Errors during the process of gene expression in the cellular machine or because external environmental factors with mutagenic capacity can lead alterations or mutations. These alterations set the genetic variety which is something that exists in all the individuals and is defined as the changes regarding to the sequence of the human referral genome. If the mutation is somatic (in not reproductive cells), the genetic variants aren't transmitted, but if the alteration is developed in the germ line as *de novo* mutations, these variants are transmitted to the offspring and can be part of the human population genetic.

Some of these variants may cause greater susceptibility to developing disorders and common traits: some, barely increase the risk, and are common in the population due to their low effect. Others are rare, sometimes only survive few generations, and can cause variety of severe pathologies, such as early neurodevelopmental disorders. The pair combination of the variants (alleles) conforms the genotypes, which make us different and shape our physical and mental observable traits (phenotypes).

In 2000, researchers completed the first full sequence of the human genome, according to a report by the National Human Genome Research Institute. However, this knowledge gave little information regarding human diseases and other projects such as *'1000 genomes project'*, *'HapMap project'* or *'Human Variome project'* were developed to describe the genetic variation across human populations and which of those increase susceptibility to physical and cognitive traits, as well as to clinically described disorders.

Thus, once we know the variants that exist in the human genome of different individuals, we can study them in at-risk populations and, through association studies, determine whether they "can" play an important role in the genesis of these disorders.

2.2 Genetic Variation and GWAS Studies

Genetic variation in the human genome can take many forms. It can take place in only one of the basic building blocks (like in the figure 1), or can involves a larger-scale variation where you might have a stretch of DNA of hundreds, or even thousands, of base pairs that is different between people. Maybe you can have three copies of that stretch and other one only has two. Or maybe it's a circumstance where you have the genes in the order 'ABC' and other has them in the order of ACB because he or she has an inversion in that. Those don't have to be pathological. In fact, most of them won't be, but it's a different kind of variation that in some instances may be playing a role in disease risk.

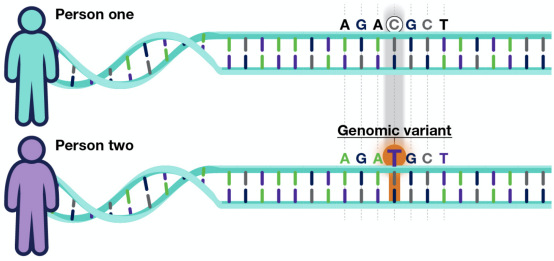

Figure 1: Single genomic variant

If the relative frequency of an allele (variant of a gene) at a particular locus (fixed position on a chromosome where the gen is located) in a population, also known as, the fraction of all chromosomes in the population that carry that allele, is higher than 1%, the variant is called polymorphism.

The most common variations in psychiatry are those that have high frequency in the population but has low associated risk and lead in complex disorders. They arise from old mutations that through some mechanism, and because they don't generate very large damage individually, have raised their allelic frequency until reaching a value to be considered polymorphisms. In most of the cases, these variants occur in a specific single nucleotide and are called SNPs (Single Nuleotide Polymorphisms), when these are tackled jointly, they determine the most part of the genetic risk of almost all known psychiatry diseases. In fact, if we analyze the source of the different disorders, we observe that with a exception of intellectual severe disability, the common variation component is the main genetic factor to predisposition.

A worldwide work known as the *HapMap Project*[4] pursue to identify and localize these genetic variants, and to learn how the variants are distributed within and among populations from different parts of the world. So far, the project has identified, more or less, 3.1 million SNPs among the human genome that are even common to individuals of different continents ancestry. The *HapMap* information has promoted a new type of research effort: the genome-wide association study (GWAS). In this kind of study, the distribution of SNPs is determined in hundreds or thousands of people with and without a particular disorder. By tagging which SNPs co-occur with disease symptoms, researchers can statistically estimate regarding the level of increased risk associated with each SNP.

Only a few years ago, a robust GWAS that genotyped thousands of individuals would have been prohibitively expensive. However, this is no longer the case thanks to the widespread use of an innovation known as the DNA microarray, which was first developed in the early 1990s[5].

This technology lets a single sample to be used simultaneously for looking for alterations at more than a million of known genetic variants.

Over the past few years, the scientific community has experienced a avalanche of knowledge derived from the use of genome-wide association studies. As SNPs are only partial contributors to an individual's risk for developing a disease, researchers must be cautious about giving too much weight to SNP profiles. Nevertheless, this has not dissuaded entrepreneurs from obtaining profit of GWAS researches. For instance, genomics companies such as *23andMe* or *Navigenics*, now offer a vary of personal genotyping and sequencing products to clients who are interested in knowing their estimated genome-based risk (aggregated score derived from the low-effect variants accumulation) for triggering of some common disorders. So, you can scanned your entire genome for markers that have been identified by GWAS and receive personalized risk calculations, for around only 1.000 usd.

But above all else, one thing is certain: Personal genetic profiles will continue to increase in their medical value as researchers increase the GWAS sample sizes, define better the case-control populations and cultivate more and more knowledge about the genetic and environmental factors that interact to contribute to the development of common disorders.

2.3 Polygenic Risk Scores

Once, we have our entire scanned genome by the GWAS, researchers can identify genomic variants associated with complex diseases by comparing the genomes of individuals with and without those diseases.
Since an enormous amount of genomic data have become available in the last decades, researchers can calculate which variants are more frequently found in groups of people with the same diseases. In fact, a disease may be affected by hundreds or even thousands of predisposing variants. With the available technology, all this genome-wide information can be used to statistically estimate individual predisposition to a particular trait or disease.

Many different methods can be used to estimate individual's predisposition to a particular trait or disease[6].
One of the most widely used methodology is calculation of Polygenic risk Scores (PRS)[7]. But, how do we now understand a polygenic risk score?

The data used for estimating a polygenic risk score comes from large scale genomic studies, so a polygenic risk score can only explain the relative risk to a disease, trait or condition. Previous studies have estimated the effect conferred by genomic variants by comparing groups with particular disease or traits to groups without that disease.
PRS indicates the individual's risk given by the aggregated contribution of all interrogated variants. These scores' absolute value is useless, but their utility comes from the comparison between different subjects with variable DNA composition. Importantly, polygenic risk scores show associations solely, not causality.

The scores may be placed on a bell curve distribution (see figure 2). In the middle would be the score of most people, denoting common risk for having a disease. Others may find themselves on the left tail end, indicating low risk. People with the score on the right tail portion, will have high-risk, they may benefit from suggestion about this risk with their doctors and genetic mentors for further health evaluations.

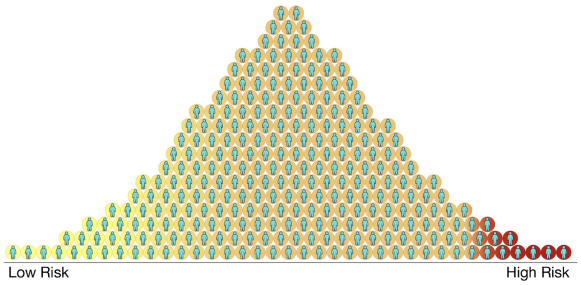

Figure 2: Bell curve distribution of the risk scores

2.4 Interactions GxE

Apart from genetic contribution, environmental factors play an important role in the etiology of mental disorders such as schizophrenia[8] or cognitive[9] and educational traits[10].

In the case of educational attainment, intelligence or cognitive performance (the most widely used traits for describing brain executive and emotional functioning), some environmental factors of major importance have been particularly described. For instance, cannabis consumption[11], familiar socioeconomic status[12] or psychological trauma[13], have been found to interrupt proper neurodevelopment and consequently affect intelligence development or educational attainment.

The joint impact of genetic predisposition and described adverse environmental exposures during neurodevelopment has been widely studied[14], but the real effect of both contributions has not been clearly described, and has derived some contradictory findings[15].

The expression "GxE interaction" refers to situations where the joint effects of genetic and environmental factors are significantly greater (or significantly smaller, when there are protective factors) than would be anticipated from the aggregation of the separate effects.

In the psychiatric genetics context there is a huge problem: 'the lost heritability'. This is the proportion of the estimated genetic variance that could not be explained by the additive models of common variance. One of the possibility to explain this leap between the predicted by epidemiological models and the predicted by genome-wide additive methods such as PRS, is the gene-environment interaction (GxE). The predisposing effect of the genetic variation could be dependent on such as those previously described environmental factors.
One of the sources which explain the traits of the human behaviour is, aside of the pure variance of the genetic and environment, the variance due to the interaction between both factors.

For instance, several meta-analyses on this issue of whether cannabis use is a cause or consequence of psychosis have now been published, consistently showing that use of cannabis, increases the risk to develop later psychotic symptoms or psychotic illness. The association between cannabis and subsequent psychosis in these studies cannot be explained entirely by confounding because in these studies the effect of cannabis on psychosis outcome remained significant after adjustment for factors such as age, sex, social class, ethnicity, urbanity, and use of other drugs.[16]

But, this GxE interactions have also been described in more complex traits such as intelligence or educational attainment.[17]

3 Dataset and Pre-processing

3.1 Dataset Description

This work relies in the analysis of the Work-package 2 (WP2) and Work-package 6 (WP6) of the EUGEI (European Network of National Networks studying Gene-Environment Interactions in Schizophrenia) consortium database. This will be our target clinical data composed for all the genetic information of the subjects to study.

The aim of EUGEI consortium is to identify, over a 5-year period, the interactive genetic, clinical and environmental traits, involved in the development, severity and outcome of schizophrenia. The partners in EUGEI represent the nationally funded mental health networks of the UK, Netherlands, France, Spain, Turkey and Germany, as well as other research institutes and a number of SME's in Austria, Belgium, Ireland, Italy, Switzerland, and outside the EU in Hong Kong and Australia. Data were collected between May, 2010 and May, 2015. In order to identify the genetic, clinical and environmental features and their interactions, EUGEI will employ family-based, multidisciplinary study paradigms, which allow for efficient assessment of gene-environment interactions.

With both packages, a total of 6,960 records are collected from different kind of subjects, specifically, after 1,272 missing *Beadchip_Location* values, the dataset contains information of 1,942 patients (positive schizophrenia cases), 1,329 siblings and 2,417 healthy participants (unrelated controls) with all the genotype data available. Nevertheless, we will discard the data from the patients and siblings and we will work only with the data from healthy cases.

Regarding the environmental exposures within the limits of data availability, the project sought to examine all the environmental exposures that have previously been associated with schizophrenia spectrum disorders. In fact, thanks to this project is focused on genetics and epidemiology of schizophrenia, we have a large amount of data to analyze in the control population. In this way, a comprehensive circumstantial questionnaire is collected from all the participants. The whole questionnaire is attached in the Appendix 1.

A selection of variables is this context has been carried out in order to choose the ones that are related the most to the study goal. Thus, variables related to education, social class or incomes achieved are elected, but also some associated with vital circumstances, like *Child Trauma* and *Cannabis Consumption*.

The details of the variables are shown in the table 1:

Variable	**Description**
Beadchip_Location	ID of the individual
Package	Package of EUGEI
Type of Subject	1: Case; 2: Sibling; 3: Control
Sex	Sex of the individual (1: male; 2: female)
Age	Age of the individual
Country	Country where the individual was born. 1: Austria; 2: Belgium; 3: France; 4: Germany; 5: Ireland; 6: Italy; 7: Spain; 8: Switzerland; 9: Netherlands; 10: Turkey; 11: United Kingdom; 12: Brazil; 13: Australia; 14: Others.
Education Level	Level of studies that the individual achieved. 1: Not graduated from basic school; 2: Graduated from basic school; 3: Graduated from high school; 4: Professional training; 5: Undergraduate degree; 6: Bachelor's degree

Variable	Description
Job Type/Social Class	Type of individual's job. 1: Unemployed; 2: Inactive (housekeeper, discapacity, retired, etc); 3: Student; 4: Temporary employee; 5: Permanent employee; 6: Autonomous or entrepreneur
Monthly Income	Monthly income of the individual in Euros
Child Trauma	Childhood Trauma Questionnaire consisting of 25 questions rated on a 5-point Likert scale
Cannabis Consumption	Has the individual consumed Cannabis? (1:yes; 0:no)

Table 1: Description of variables of the dataset

On the other hand, for the fulfillment of the project we need to estimate genetic scores for the Work-packages using the discovery data of a completed genome association studies (GWAS).

Three GWAS meta-analysis are used in this project to obtain three different indicator scores: Intelligence, Cognitive Performance and Education Attainment.

For the intelligence score, a GWAS meta-analysis made by *Jeanne Savage et al. (Nature Genetics, 2018)*, in 2018 along 269,867 individuals identifies new genetic and functional links to intelligence.

GWAS was performed using various software (PLINK, SNPTEST, RAREMETALWORKER, etc.) in 14 independent cohorts, and METAL tool for the meta-analyzed. SNPs with MAC < 100, INFO < 0.6, indels, multiallelic, or N < 50,000 are excluded. The variables of this dataset are given in the table 2:

Variable	Description
SNP	Rs number
UNIQUE_ID	Unique SNP id based on chromosome, position
CHR	Chromosome number
POS	Base pair position reported on GRCh37
A1	Effect allele
A2	Non-effect allele
EAF_HRC	Effect allele frequency in the Haplotype Reference Consortium reference panel (HRC)
Zscore	Meta-analysis Z score
stdBeta_*	Standardized beta coefficient
SE	Standard error of the beta coefficient
P	P-value
N_analyzed	Sample size
minINFO	Minimum info score (SNP quality measure) across all cohorts
EffectDirection	Direction of the effect in each of the cohort

Table 2: Description of the variables of the discovery data for intelligence scores

Regarding the Cognitive Performance and Educational Attainment, a GWAS meta-analysis of all discovery cohorts (except *23andMe* for the EA) is made. The reported allele frequencies are calculated using data on European-ancestry individuals who contributed to the *1000 Genomes Project - Phase 3*. Association results are only provided for SNPs that pass standard quality-control filters described in the SI of *Lee et al. (2018)*. The sample size is 766,345 individuals for the GWAS of Education Attainment and 257,828 individuals for the Cognitive Performance.

The content of this summary statistics dataset is shown below (see table 3):

Variable	Description
MarkerName	SNP rs number
CHR	Chromosome number
POS	Base pair position
A1	Effect allele
A2	Other allele
EAF	A1 frequency in 1000 Genomes Phase 3 sample (CEU, GBR and TSI individuals)
Beta_*	Standardized regression coefficient, i.e. per-allele effect size on the phenotype that has been standardized to have unit variance
SE_*	Standard error of Beta
Pval_*	Nominal p-value of the null hypothesis that the coefficient is equal to zero.

Table 3: Description of the variables of the discovery data for cognition and educational attainment scores

3.2 Genetic Data Pre-processing

The first step needed for the development of the project is the calculation of the polygenic risk scores (PRS) from the different GWAS studies in order to estimate the probabilistic susceptibility of an individual to intelligence, cognitive performance (CP) or education attainment (EA).

Once we have the discovery sample: a case-control cohort originated from the GWAS studies with the association and risk values that we want to analyze (i.e. intelligence, CP, EA...), we select the simple variants and delete the repeated and the indels and triallelics. Also, a good practice is to filter the variants according to a minimum value of information. In this work we have selected from the GWAS of intelligence the variants with a *minINFO* higher than 80%. The other two discovery samples didn't have this parameter, so, we didn't discard data regard to that.

With all these variants filtered, we extract the association *P-values* and odd ratios (*OR*) of all the genotyping variants.

Then, with the target sample (EUGEI dataset), after eliminating the repeated variants as well, we do an overlap between the variants of both datasets. With these common variants and PLINK software we do a clumping and a *linkage disequilibrium (LD)*[18] to prune the set of variants in order to delete joined variants (which give same information) and are redundant. It is calculated using the –indep-pairwise command in `PLINK` (maximum r2=0.1 and window size=500 SNPs)). See figure 3.

A total independent samples of 134,479 SNPs for the GWAS of intelligence, 199.608 SNPs in the Cognitive Performance GWAS and 199,231 SNPs in Educational Attainment are finally used.

Finally, the software `PLINK` calculates by additive way, the relative risk of each variant in each individual. The protective alleles subtract score while the risk ones sum. The most common practice is to do different selections regarding the association *P-value*, thus, the previous step of the risk calculus is repeated for the different significance levels. In our case, we have selected the PRS obtained for a *P-value* = 1, (that is, inclusion of all SNPs), because previous studies determined it is the optimal threshold which explained most variation in the phenotype.[19]

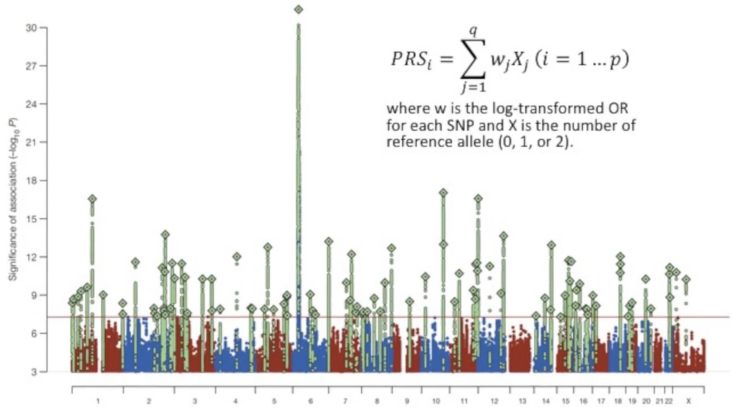

Figure 3: Polygenic Risk Scores calculation process to estimate genetic risks

The three lists of scores are joined to the rest of the environmental variables through the `vlookup` function in Excel matching the ids of both files, setting up, then, the final dataset for the next sections where regression models are executed to check if these scores can explain the academic and social success.

3.3 Environmental Data Pre-processing

In order to see the effects of the relation between the genetic and environmental factors, we need to complete the genetic dataset with the biologic features of the individuals.
Aside the *Age*, *Sex* and *Country* features, Principal Components (PCs) are calculated by the genetic data in PLINK using LD pruned variants (after combining the dataset with the Thousand Genomes reference), and the first 10 are used as ancestry covariates, in order to correct for stratification the scores with large differences in allele frequency across ancestral populations.[20]

On the other hand, the childhood adversity (*CTQ* variable) is assessed using the Childhood Trauma Questionnaire Short Form (CTQ). This consists of 25 items, rated on a 5-point Likert scale, measuring five domains of maltreatment (emotional (EA), physical (PA) and sexual abuse (SA); emotional (EN) and physical neglect (PN)). To dichotomize each childhood adversity domain (0="absent" and 1="present"), we have applied a rule based on the classification scale shown in the table 4, by which we have considered presence of abuse if the score is 'low to moderate' or more, at least in one category. These work has been done directly in *Excel*.

Category	**Score Scale**
None or minimal	$EA \leq 8$; $PA \leq 7$; $SA \leq 5$; $EN \leq 9$; $PN \leq 7$. $TOT\ CTQ \leq 36$
Low to moderate	$8 < EA \leq 12$; $7 < PA \leq 9$; $5 < SA \leq 7$; $9 < EN \leq 14$; $7 < PN \leq 9$. $36 < TOT\ CTQ \leq 51$
Moderate to severe	$12 < EA \leq 15$; $9 < PA \leq 12$; $7 < SA \leq 12$; $14 < EN \leq 17$; $9 < PN \leq 12$. $51 < TOT\ CTQ \leq 68$
Severe to extreme	$EA \geq 16$; $PA \geq 13$; $SA \geq 13$; $EN \geq 18$; $PN \geq 13$. $TOT\ CTQ \geq 69$

Table 4: Score scale of Childhood Trauma Questionnaire

Also, the use of Cannabis has also been considered as a environmental covariate, since previous studies (as mentioned in the theoretical framework) has proven that this condition can be highly related to the presence of mental disorders and therefore with the academic and social success.

Finally, in two of the response variables (those we want to predict), particularly, *Study Level* and *Job Type/Social Class*, we have re-grouped the categories in only 3 levels more estranged in order to have best prediction performance. In this way, the groups are read as follow:

Variable	Description
Study Level	Level of studies that the individual achieved. 1: Not graduated and graduated from basic school; 2: Graduated from high school and professional training; 3: Undergraduate degree; Bachelor's degree.
Job Type/Social Class	Type of individual's job. 1: Unemployed and Inactive (housekeeper, discapacity, retired, etc); 2: Student and Temporary employee; 3: Permanent employee and Autonomous or entrepreneur.

Table 5: Classification of the categories in the levels of the categorical responses variables

3.4 Descriptive Analysis

After deleting the *Beadchip_Location* variable, filter the *type of subject* in order to select only the healthy cases and change the type of the variables (factor when categorical and numeric when continuous), we already have the data available to apply some visualization graphs and see how the data behaves. The figure 4 shows the head of the final dataset.

Sex	Age	Country	Income	ChildTrauma	Cannabis	PCA1	PCA2	PCA3	PCA4	PCA5	PCA6	PCA7	PCA8	PCA9	PCA10	ScoreInt	ScoreCog	ScoreEA
<fct>	<dbl>	<fct>	<dbl>	<fct>	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	26	7	1500	0	1	0.00480	0.0120	0.000393	0.0159	0.00794	0.00109	-0.000959	-0.00808	-0.00228	-0.00499	0.000199	-0.0000153	0.0000744
1	33	7	2200	0	0	0.00512	0.0126	0.000268	0.0127	0.00516	-0.00163	0.00121	-0.00669	-0.00443	-0.00340	0.000202	-0.0000117	0.0000700
1	50	7	3000	1	1	0.00500	0.0123	-0.000816	0.0151	0.00542	0.00338	0.00139	-0.00818	-0.000620	0.00148	0.000199	-0.0000259	0.0000637
1	29	7	3000	0	1	0.00373	0.0119	0.000414	0.0158	0.00520	0.000910	0.00121	-0.000886	0.00172	-0.000300	0.000188	-0.0000237	0.0000693
2	23	7	6000	0	0	0.00447	0.0113	0.000670	0.0136	0.00294	-0.00246	0.00480	-0.00284	0.00235	-0.00786	0.000210	-0.0000214	0.0000755
1	23	7	5000	1	1	0.00414	0.0120	-0.00152	0.0118	0.00182	-0.00517	-0.00101	0.00206	-0.00471	-0.00854	0.000203	-0.0000209	0.0000720

Figure 4: Head of the final dataset

In the figure 5 we can see the density and box plots of our numeric factors.

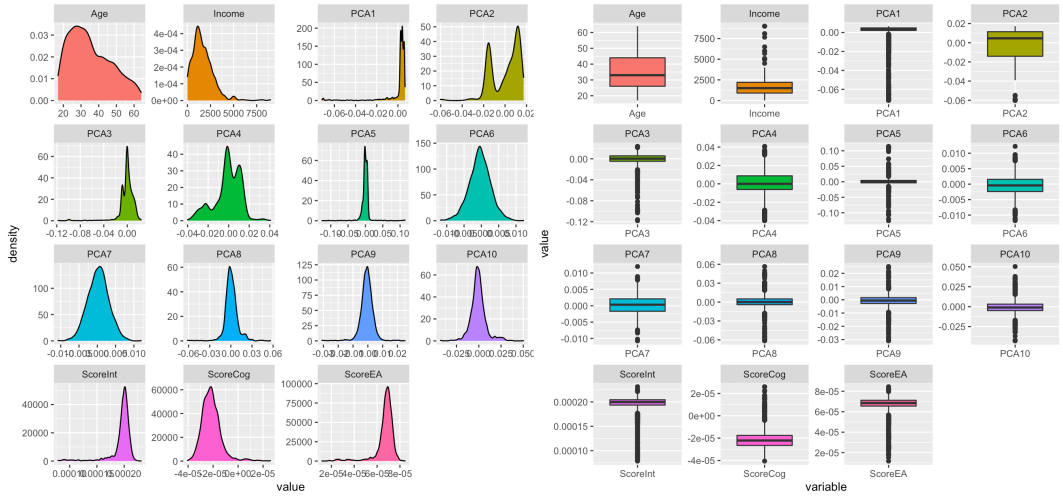

Figure 5: Density and Box plots of all numeric variables

As we can observe, more or less all variables follow a Gaussian distribution, with the exception of big tails or distant points in some cases due to the presence of some outliers.

In the case of the *Income* variable, we have applied a transformation to achieve more normal distribution since it is one of our responses. Thus, it has been change through a logarithmic and cubic operation: $(Income = log(Income)^3)$.

Also, the missing values have been treated separately with respect to the response variable that we want to study. For instance, the *Income* variable has much more outliers (35%) since it only has this information from the package WP6. On the contrary, there are only 46 missing values from the others variables. Thus, once we have split our dataset and select the response and covariates we are going to use in each of the regression models, we apply the rule to delete the missing values. We haven't decided to impute them with estimated values since it is very difficult to infer the income and it could be even worse.

Regarding the outliers, we have decided not deleting them since they don't seem to be wrong records, but, in fact, relevant observations in genetic that could give us more information to the models than even the centered ones.

On the other hand, the distribution of the categorical variables is shown in the figure 6.

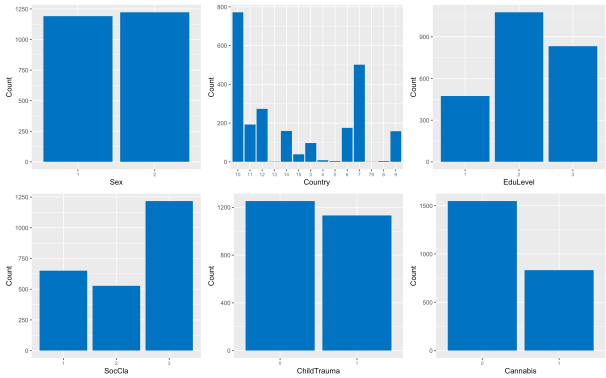

Figure 6: Distribution plots of all categorical variables

The variable *Country* has some levels with very few observations, thus, that levels has been re-grouped into the *Other Countries* category in order to have more balance between the different levels of the variable.

The other two responses variables, *SocCla* and *EduLevel* has also some imbalanced groups, but as we have enough observations en each level, we have decided to keep them in that way so we don't have to lose information.

In addition, we have checked the correlation between the variables to determine if there is multicollinearity. If so, we should delete some of them and simplify the data. In the next plot (figure 7), the Pearson correlation coefficient between each pair of variables is shown.

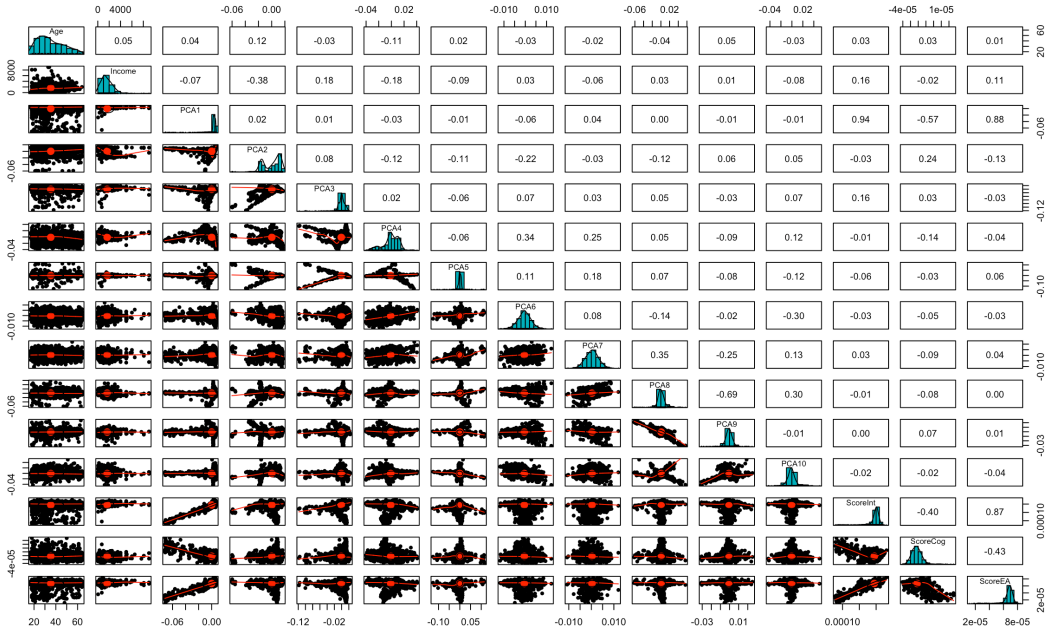

Figure 7: Pearson Correlation Coefficients between each pair of variables

The graph indicates very strong correlation between the different genetic scores (*ScoreInt*, *ScoreCog* and *ScoreEA*) and the variable *PCA1*, thus, we have deleted this last one to avoid overfitting.

Finally, we have also standardize the values of all *PCAs* and *Scores* in order to be able to have the data in the same scale and visualize it simpler.

4 Statistical Methodology

As mentioned, the aim of the project is to study the social and academic success derived from the genetic variation and environmental factors of the subjects. These traits are reflected in the dependent variables: *Income*, *SocCla* and *EduLevel*.

This section is divided into two approaches associated to multiple regressions. Section 4.1 covers Multiple Linear Regression to study the continuous numeric response *Income*, while Section 4.2 considers Multinomial Ordinal Logistic Regression to analyze the ordinal categorical response variables.

These in turn, have two parts. The first one studies the models with only genetic variables and how the different polygenic risk scores (*ScoreInt*, *ScoreCog* and *ScoreEA*) influence in the explanation of the response variables. The second part will develop the analysis of how the environmental factors (*Country*, *ChilTrauma* and *Cannabis*), and its interactions with the rest of the covariates and, in particularly, with the PRS, can increase or not the explication of the dependent variables.

4.1 Multiple Linear Regression

Normally, multiple linear regression describes the association between predictor or independent variables and one dependent or response variable. A dependent variable is modeled as a function of various independent variables with corresponding coefficients, together with the constant term.

The multiple regression equation explained above takes the following form:

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + c$$

Here, $\beta_i's$ (i=1,2. . . n) are the regression coefficients, which represent the value at which the criterion variable changes when the predictor variable changes, are used in measuring how effectively the predictor variable influences the criterion variable. It is measured in terms of standard deviation. The c term refers the errors which are suppose to be Gaussian.

In order to test the model, the metric R Adjusted Square, or Adjusted R2, is used. It is the square of the measure of association which indicates how well terms fit a line, but adjusted for the number of terms in a model. If you add more and more useless variables to a model, adjusted r-squared will decrease. If you add more useful variables, adjusted r-squared will increase. It indicates how much variance the model explain of the trait (response variable) that we are predicting. The expression of R2 and Adjusted R2 is detailed below:

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$$R^2_{adj} = 1 - \left[\frac{\left(1 - R^2\right)(n-1)}{n-k-1} \right]$$

There should be proper definition of the model and only relevant variables must be included. Thus, the insertion of the covariates has been running one by one, at the same time that, the significance and the explained variance has been pointed.

To determine whether the model is significant, we use the p-value and ANOVA test.

The p-value for each independent variable tests the null hypothesis which is that the variable has no correlation with the dependent variable. If there is no correlation, there is insufficient evidence to conclude that there is effect at the population level. Thus, the p-value for a variable would be higher than our significance level (set up at 0.05). If the p-value is lower than the significance level, your sample data provide enough evidence to reject the null hypothesis for the entire population, so, your data favor the hypothesis that there is a non-zero correlation and it is statistically significant.

On the other hand, ANOVA (Analysis Of Variance) tests whether the more complex model is significantly better explaining the data than the simpler model. If the resulting p-value is lower than 0.05, we conclude that the more complex model is significantly better than the simpler model. If the p-value is greater, we should chose the simpler model.

In this section the response variable *Income* is represented as $Y \in \{0, +\infty\}$ and the function used to develop the regressions in `R` is `lm()`.

4.2 Ordinal Logistic Regression

An ordinal regression is an extension of a binomial logistic regression and it is used to predict responses variables that have ordered multiple categories.

The response variable Y takes values $1, ..., K$ and the cumulative probabilities (p_k) and the cutpoints (k) are represented by the next relation:

$$\tilde{p}_k = \text{Pt}(Y \leq k \mid X), \qquad \text{para } k = 1 \ldots K-1$$

The proportional odds model compares the probability of getting a response $\leq k$ with the one of getting a response $> k$, thus, the formula of the regression models take the expression below:

$$c_k(X) = \log\left(\frac{\Pr(Y \leq k \mid X)}{\Pr(Y > k \mid X)}\right) = \alpha_k - \beta_1 X_1 - \ldots - \beta_m X_m \quad k = 1 \ldots K-1$$

With expanded next equations:

$$\begin{aligned}
\text{logit}\,(p_1) &= \log\left(\frac{p_1}{1-p_1}\right) = \alpha_1 - \beta_1 X_1 + \ldots - \beta_m X_m \\
\text{logit}\,(p_2) &= \log\left(\frac{p_2}{1-p_2}\right) = \alpha_2 - \beta_1 X_1 + \ldots - \beta_m X_m \\
&\vdots \\
\log it\,(p_{K-1}) &= \log\left(\frac{p_{K-1}}{1-p_{K-1}}\right) = \alpha_{K-1} + \beta_1 X_1 + \ldots - \beta_m X_m \\
\text{logit}\,(p_K) &= 1
\end{aligned}$$

Here, $\beta_i's$ are the regression coefficients meaning the change in log-odds of the higher category versus another that is lower, associated with an increase of one unit in X_i, keeping the rest of the predictors constant.

The model assumes that the coefficients of the predictor are the same, the only difference is the intercept, α_i.

To be able to compare our estimates from the current sample with the previously reported estimates of the proportion of variance explained by PRS, a McFadden's R2[21] is calculated. It is defined as:

$$R^2_{\text{McFadden}} = 1 - \frac{log(L_c)}{log(L_{\text{null}})}$$

where L_c denotes the (maximized) likelihood value from the current fitted model, and L_{null} denotes the corresponding value but for the null model.

Also, to determine whether the model or variables are significant, we use the ANOVA test.

In this section, the response variables are represented as $Y \in \{1, 2, 3\}$ (three classes) measuring the social class and education level of control individuals, and the function used to develop the regressions in `R` is `polr()` from the package `MASS`.

In the next sections each response variable is treat separately conforming two parts.

5 Results of GxE analysis in Social and Academic Success

5.1 Genetic Models

5.1.1 Monthly Income

In order to obtain the explained variance for each of the different scores, firstly, we build a basic model with *Sex*, *Age* and all the *PCA* variables, (even if some of them are not significant at all) with the aim of correcting ancestry and gender differences[22], and secondly, we complete another model adding the score variable. The study is done with 839 observations with valid values of income and genetic information.

```
model0 - Incomes ~ Sex + Age + PCA2 + PCA3 + PCA4 + PCA5 + PCA6 + PCA7
+ PCA8 + PCA9 + PCA10
model1 - Incomes ~ Sex + Age + PCA2 + PCA3 + PCA4 + PCA5 + PCA6 + PCA7
+ PCA8 + PCA9 + PCA10 + ScoreInt
```

The difference in R2 values will give us the variance explained only by the scores. The table 6 gather this information.

Variable	R2	Coefficients	P-values	Significance
ScoreInt	**0.00725**	11.173	0.00266	Significant
ScoreCog	0.00125	5.366	0.12315	Not Significant
ScoreEA	0.00144	5.680	0.10783	Not Significant

Table 6: Statistical information of different scores in the explanation of Incomes

As we can observe in the table 6, *SoreInt* is the only sginifcant polygenic score, with a 0.73% of the total variance of monthly income explained. Therefore, from now on, we will not consider the other scores for the analysis since they are not significant.

On the other hand, the coefficient of *ScoreInt* in this first model is equal to 11.17, which means that each more unit in the (scaled) intelligence score, the estimated (transformed) monthly income increase 11.17 euros.

5.1.2 Social Class

Firstly, in order to be able to study ordinal regression correctly, we have to check if the proportional odds assumptions between the different categories are met. This means, the differences in the distance between the sets of coefficients should be similar and then the effect of the variables in the transitions between the levels of the response are homogeneous. The figure 8 shows the distances values in all the ranges or categories in each variable.

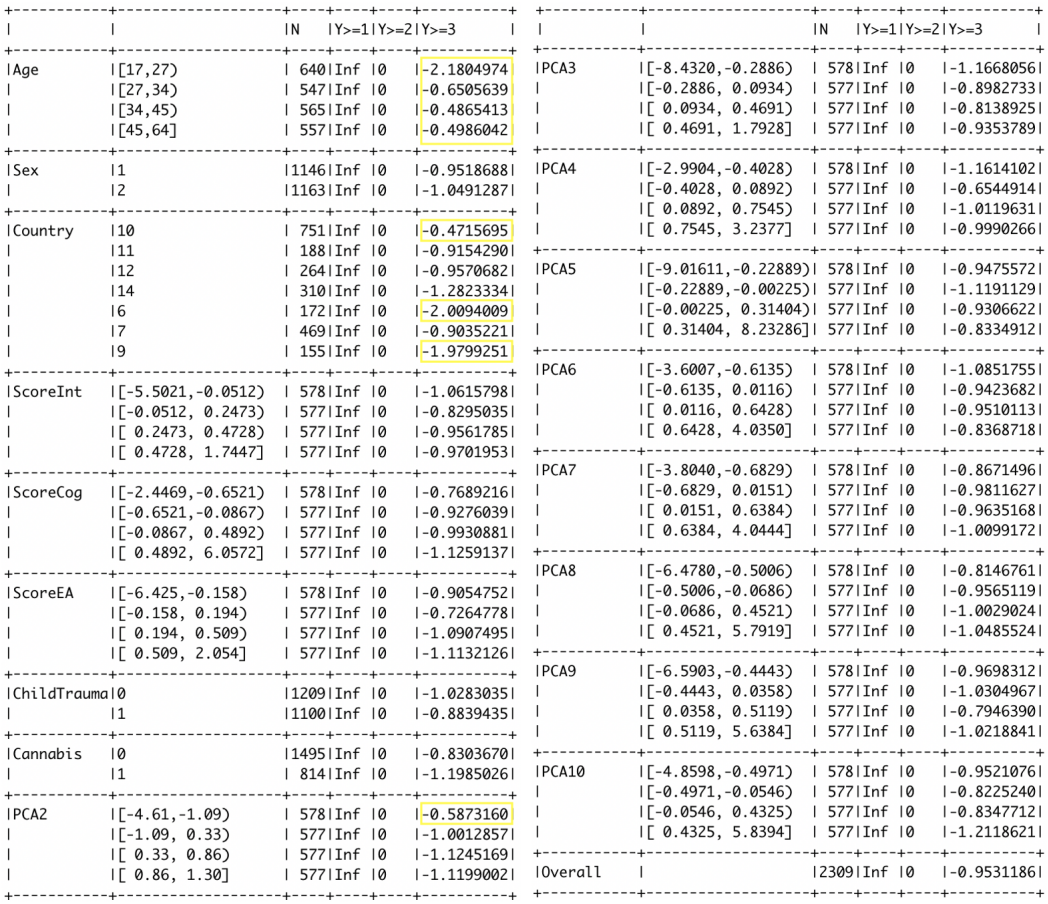

Figure 8: Proportional Odds Assumptions of Social Class variable

The differences between some sets, for instance, in *Age* or *Country*, may suggest some lack of parallel slopes for that predictors, and it could lead in worse results in terms of fit. In any case, the assumptions are generally met for the rest of variables and can be enough to do some approximation studies.

As in the previous section, we first obtain the explained variance of the scores in a basic genetic model to see how them influence the response. In this case, we have 2, 309 observations with completed information of social class and the rest of variables. The table 7 shows the detail.

Variable	R2	Coefficients	P-values	Significance
ScoreInt	**0.00136**	0.100	0.01264	Significant
ScoreCog	0.00028	-0.047	0.25799	Not Significant
ScoreEA	**0.00182**	0.115	0.00392	Significant

Table 7: Statistical information of different scores in the explanation of Social Class

The type of scores which explain the most is: *ScoreEA* with a total of 0.18% of the total variance of social class feature. Nevertheless, *ScoreInt* explain similar variance (0.14%). Both variables are significant and they will be considered in the analysis of the next sections.

As a predictors, the coefficients of *ScoreEA* and *ScoreInt* are 0.12 and 0.10 which means that the possibility of going from *class 1* to *class 2* is 0.12 or 0.10 times higher each time the *ScoreInt* and *ScoreEA* of the individual increase in one unit. Also it will be 0.12 or 0.10 times more

possible when going from *class 2* to class 3.

On the other hand, the *ScoreCog* here is not significant either. So, we can conclude that cognition performance is not a good genetic indicator to estimate the potential social class of people and it will not be considered hereafter in further analysis.

5.1.3 Educational Level

Here, we apply the previous scheme to the response *EduLEvel*. We have now, 2, 301 observations with the detail of Education Level trait.

The detail of the proportional odds assumptions and the table of the explained variance of genetic scores variables are shown below.

		N	Y>=1	Y>=2	Y>=3			N	Y>=1	Y>=2	Y>=3
Age	[17,27)	636	Inf	0	-2.703880	PCA3	[-8.4356,-0.2983)	576	Inf	0	-1.710496
	[27,34)	544	Inf	0	-1.995992		[-0.2983, 0.0947)	575	Inf	0	-2.085149
	[34,45)	560	Inf	0	-2.075571		[0.0947, 0.4701)	575	Inf	0	-2.213298
	[45,64]	561	Inf	0	-1.612827		[0.4701, 1.7950]	575	Inf	0	-2.049589
Sex	1	1140	Inf	0	-2.120887	PCA4	[-2.9843,-0.4092)	576	Inf	0	-1.841345
	2	1161	Inf	0	-1.882262		[-0.4092, 0.0904)	575	Inf	0	-1.944630
Country	10	752	Inf	0	-1.734703		[0.0904, 0.7439)	575	Inf	0	-2.024916
	11	188	Inf	0	-1.708238		[0.7439, 3.2465]	575	Inf	0	-2.272960
	12	269	Inf	0	-1.701966	PCA5	[-9.01940,-0.22853)	576	Inf	0	-1.848463
	14	308	Inf	0	-2.195675		[-0.22853,-0.00268)	575	Inf	0	-1.968522
	6	173	Inf	0	-3.281095		[-0.00268, 0.31271)	575	Inf	0	-2.085535
	7	456	Inf	0	-2.457144		[0.31271, 8.23928]	575	Inf	0	-2.128695
	9	155	Inf	0	-1.980370						
ScoreInt	[-5.5024,-0.0491)	576	Inf	0	-1.931752	PCA6	[-3.601,-0.623)	576	Inf	0	-1.922449
	[-0.0491, 0.2473)	576	Inf	0	-2.039480		[-0.623, 0.014)	575	Inf	0	-1.983263
	[0.2473, 0.4705)	574	Inf	0	-1.942716		[0.014, 0.646)	575	Inf	0	-2.058190
	[0.4705, 1.7442]	575	Inf	0	-2.234650		[0.646, 4.049]	575	Inf	0	-2.047289
ScoreCog	[-2.4446,-0.6499)	576	Inf	0	-2.074269	PCA7	[-3.8058,-0.6836)	576	Inf	0	-2.033059
	[-0.6499,-0.0879)	575	Inf	0	-2.092807		[-0.6836, 0.0164)	575	Inf	0	-1.918733
	[-0.0879, 0.4899)	575	Inf	0	-2.006816		[0.0164, 0.6375)	575	Inf	0	-2.064475
	[0.4899, 6.0494]	575	Inf	0	-1.868810		[0.6375, 4.0491]	575	Inf	0	-2.003113
ScoreEA	[-6.411,-0.158)	576	Inf	0	-1.951598	PCA8	[-6.4872,-0.5037)	576	Inf	0	-2.105277
	[-0.158, 0.191)	575	Inf	0	-2.229455		[-0.5037,-0.0709)	575	Inf	0	-1.987693
	[0.191, 0.508)	575	Inf	0	-2.109135		[-0.0709, 0.4477)	575	Inf	0	-1.746006
	[0.508, 2.049]	575	Inf	0	-1.922470		[0.4477, 5.7887]	575	Inf	0	-2.168187
ChildTrauma	0	1203	Inf	0	-2.038781	PCA9	[-6.5851,-0.4434)	576	Inf	0	-2.089951
	1	1098	Inf	0	-1.987253		[-0.4434, 0.0362)	575	Inf	0	-2.066596
Cannabis	0	1494	Inf	0	-1.956151		[0.0362, 0.5107)	575	Inf	0	-1.916221
	1	807	Inf	0	-2.094543		[0.5107, 5.6346]	575	Inf	0	-1.935617
PCA2	[-4.610,-1.086)	576	Inf	0	-1.805774	PCA10	[-4.8551,-0.4993)	576	Inf	0	-1.873322
	[-1.086, 0.329)	575	Inf	0	-2.008574		[-0.4993,-0.0576)	575	Inf	0	-1.979654
	[0.329, 0.863)	575	Inf	0	-2.250127		[-0.0576, 0.4301)	575	Inf	0	-1.924912
	[0.863, 1.304]	575	Inf	0	-1.963610		[0.4301, 5.8269]	575	Inf	0	-2.228669
						Overall		2301	Inf	0	-1.998414

Figure 9: Proportional Odds Assumptions of Educational Level variable

Variable	R2	Coefficients	P-values	Significance
ScoreInt	**0.00368**	0.171	0.00003	Significant
ScoreCog	**0.00164**	0.114	0.00558	Significant
ScoreEA	**0.00657**	0.228	0.00000	Significant

Table 8: Statistical information of different scores in the explanation of Education Level

The figure 11 may suggest that there is also a bit of lack of parallel slopes assumption for the predictor *Educational Level*. Nevertheless, it happens in only one range in a couple of variables,

so, we can determine that the data for this trait behaves better than in the previous section and possibly the results of the logistic regressions will be better.

On the other hand, we can visualize in the table 8, that all the different scores are significant, although *ScoreEA* has the highest explained variance with a value of 0.66%.

Its coefficient is equal to 0.23 which means that the possibility of going from *class 1* to *class 2* is 0.23 times higher each time the *ScoreEA* of an individual increase in one unit. Also it will be 0.23 times more possible when going from *class 2* to class 3.

Therefore, education level can be predicted by any of the three scores (unique case which this happen), social class, instead, is the variable with less level of possible prediction.

5.2 Environmental Interaction Models

5.2.1 Monthly Income

In this section, we have tested, firstly, the significance of the scores whether there is a presence of environmental variables or not in the sample. In this way, we can see if some inequalities can change the genetic explanation in the responses.

	Variable	Presence	N	R2	Coeffs	P-values	Significance
ScoreInt	Sex	Men	525	0.00839	14.014	**0.00976**	**Significant**
		Women	314	0.00223	7.332	0.16354	Not Significant
	ChildTrauma	1	394	0.00173	8.969	0.18246	Not Significant
		0	445	0.01161	12.989	**0.00418**	**Significant**
	Cannabis	1	243	0.00434	11.544	0.13877	Not Significant
		0	596	0.00544	9.229	**0.02681**	**Significant**

Table 9: Statistical information of significant scores in presence of environmental variables in the explanation of Incomes

As we can see in table 9, generally, the intelligence scores are only significant explaining income trait in populations where there wasn't child trauma or don't consume *Cannabis*. On the contrary, they are only significant in men, being irrelevant to explain the salaries in women's group.

Also, the effect of having child trauma and the use of *Cannabis* is tested inside each gender in order to see if this fact influences the explanation of intelligence scores. The results of the analysis are gather in the table 10.

	Stratum	Variable	Presence	N	R2	Coeffs	P-values	Significance
ScoreInt	Men	ChildTrauma	1	244	0.00601	14.634	0.10609	Not Significant
			0	281	0.00425	11.061	0.10328	Not Significant
		Cannabis	1	181	0.02178	20.138	**0.01849**	**Significant**
			0	344	-0.00058	6.090	0.37976	Not Significant
ScoreInt	Women	ChildTrauma	1	150	-0.00136	-17.949	0.34151	Not Significant
			0	164	0.00546	8.805	0.09797	Not Significant
		Cannabis	1	62	-0.00494	-4.370	0.68532	Not Significant
			0	252	0.01218	12.551	**0.04976**	**Significant**

Table 10: Statistical information of significant scores in presence of environmental variables and sex stratum in the explanation of Incomes

The presence or not of child trauma is not related anymore with the scores, even in men. In addition, an interesting thing occurs: when the scores were only significant before when the people didn't consume *Cannabis*, now, in only male groups the consumption of this drug does boost the importance of the score in the explanation of received revenues.

In female groups, the scores weren't significant, but now, in populations of women without consumption of *Cannabis*, they became relevant in the explanation of the phenotype *Incomes*.

This could mean that the effect of this factor is more relevant in men.

To add an extra section in this further analysis, we could study the relation of the use of *Cannabis* and the presence of *ChildTrauma* in each gender in the two different countries that we have information about (Spain and Turkey), but we don't have enough observation in Turkey for these characteristics to test it. We will see this configurations in next sections for the other traits (Social Class and Education Level).

After to analyze each score and the environmental predictors separately, we have included all of them in GxE multiple linear regression models, searching through *Stepwise*[23] selection which is the better model in order to see how much the explained variance can be improved (see table 11).

	Variables	**Formula**	**R2**	**AIC**	**Coeffs**
ScoreInt	Only genetic	Income~Age+Sex+ScoreInt+PCA2 +PCA3+PCA4+PCA8+PCA10 +PCA2:PCA10+PCA4:PCA10 +Age:PCA2+Age:Sex	0.33465	7624	10.469
	With Environmental	Income~Age+Sex+ScoreInt +PCA8+PCA10+Country +Cannabis+PCA10:Country +Age:Cannabis +PCA8:Cannabis+Age:Sex	0**.37764**	7569	10.936

Table 11: Best selection models with for the prediction of Incomes

As shown in table 11, the ensemble of all genetic variables explains the 33% of the monthly incomes variance, comparing with the only 0.73% of the score variable alone. If, additionally, we add the environmental variables and its interactions to the genetic ones, the explained variance of the response increase 13% more with a total of **38%**.

On the other hand, the values of the score coefficients increase a little bit (from 10.5 to 10.9, meaning that the presence of the other variables boost the importance of the scores in the prediction of monthly incomes.

In order to test if the final model is well fitted and can be used to do predictions, we have to check some assumptions. To do so, we obtain the diagnostic plots (see figure 10) of the best model (with intelligence scores) which show the behaviour of the residuals.

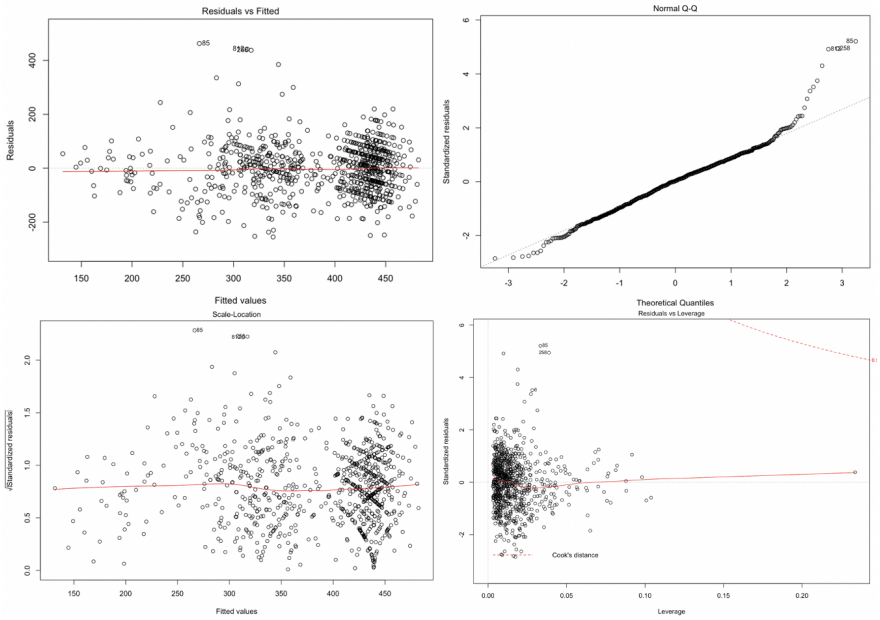

Figure 10: Diagnostic plots of best multiple linear regression

From the graphs we can indicate that the regression meet the assumptions with the exception of a couple of distant influence cases. In any case, it should be enough to be able to do obtain a primary approximation of predictions.

To end with, we can also see, the effect of the covariates in the income response (figure 11).

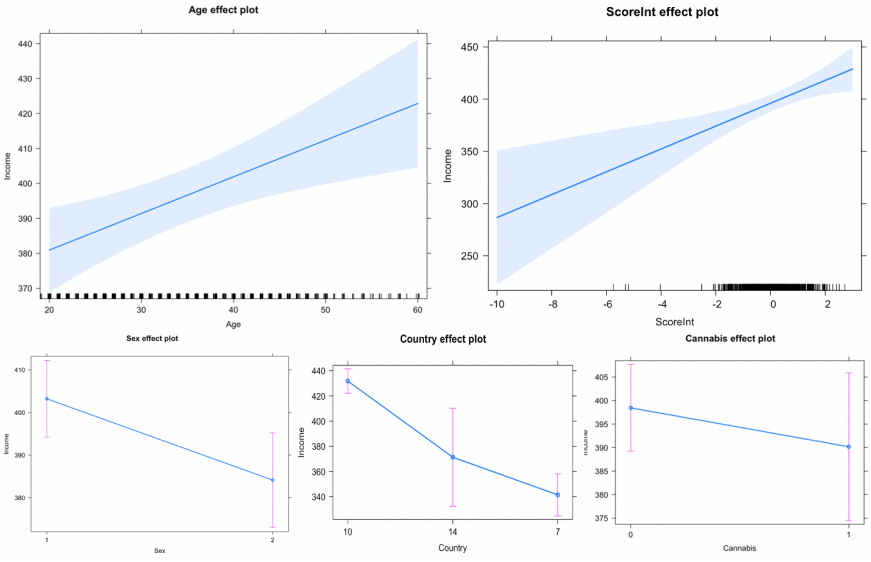

Figure 11: Effect plots of covariates in the income response variable

We can conclude saying that the only genetic scores which describe incomes are the intelligence ones, since the others haven't been significant. The fact to have a child trauma is not related with the incomes individually, although it can condition other genetic variables, over more in women, where harder child trauma could be presented.

On the contrary, the age and the intelligence scores are directly proportional of the incomes (the more, the better). By contrast, women earn much less monthly income than men, and, if you don't consume Cannabis, you have more possibility to earn more.
Regarding the countries, we can see that Turkey has much higher incomes than Spain, fact that is a bit inconsistent and make us think that maybe Turkey sample could be biased by the fact that all the participants are from upper class.
Lastly, as interpretation of the best model, 0.00001 more units in the (des-scaled) intelligence score, involve 1.58 euros more in the individual's monthly (des-transformed) income.

5.2.2 Social Class

Let's see now how the environmental variables influence the different scores describing the social class in table 12.

	Variable	Presence	N	R2	Coeffs	P-values	Significance
ScoreInt	Sex	Men	1146	0.00088	0.082	0.18208	Not Significant
		Women	1163	0.00234	0.129	**0.01652**	**Significant**
	ChildTrauma	1	532	0.00011	0.027	0.62951	Not Significant
		0	614	0.00457	0.190	**0.00100**	**Significant**
	Cannabis	1	814	0.00045	0.055	**0.39077**	Not Significant
		0	1495	0.00249	0.146	**0.00771**	**Significant**
ScoreEA	Sex	Men	1146	0.00127	0.097	0.10811	Not Significant
		Women	1163	0.00335	0.156	**0.00415**	**Significant**
	ChildTrauma	1	532	0.00024	0.040	0.47228	Not Significant
		0	614	0.00572	0.214	**0.00023**	**Significant**
	Cannabis	1	814	0.00001	0.026	0.68687	Not Significant
		0	1495	0.00425	0.189	**0.00051**	**Significant**

Table 12: Statistical information of significant scores in presence of environmental variables in the explanation of Social Class

Both types of scores are only significant in the explanation of the trait *Social Class* when the population are all women, there hasn't been child trauma and there is not consumption of *Cannabis*.

Nevertheless, the explained variances of educational attainment scores are higher in all the cases than the intelligence ones as we already saw in the first section.

Since, only the genetic scores are significant in women, let's study further how the gender can work as stratification parameter to see how the environmental variables behave in this group and into country divisions (only in Turkey, Brazil, Spain and Other countries, since there is not enough observations for testing it in the rest of the geographies). See Table 13 for the details.

	Stratum	Variable	Presence	N	R2	Coeffs	P-values	Significance
ScoreInt	Men	ChildTrauma	1	532	0.00062	0.066	0.44185	Not Significant
			0	614	0.00159	0.115	0.19552	Not Significant
		Cannabis	1	500	0.00069	0.072	0.41089	Not Significant
			0	646	0.00035	0.054	0.54653	Not Significant
	Women	ChildTrauma	1	568	0.00002	0.012	0.87372	Not Significant
			0	595	0.00982	0.286	**0.00045**	**Significant**
		Cannabis	1	314	0.00027	0.040	0.67292	Not Significant
			0	849	0.00461	0.202	**0.00450**	**Significant**

	Stratum	Variable	Presence	N	R2	Coeffs	P-values	Significance
ScoreEA	Men	ChildTrauma	1	532	0.00030	0.045	0.59195	Not Significant
			0	614	0.00356	0.169	0.05268	Not Significant
		Cannabis	1	500	0.00008	0.025	0.78338	Not Significant
			0	646	0.00223	0.128	0.12954	Not Significant
	Women	ChildTrauma	1	568	0.00060	0.063	0.39762	Not Significant
			0	595	0.00984	0.287	**0.00044**	**Significant**
		Cannabis	1	314	0.00040	0.048	0.60861	Not Significant
			0	849	0.00619	0.237	**0.00099**	**Significant**

Table 13: Statistical information of significant scores in presence of environmental variables into sex stratum in the explanation of Social Class

In all the cases, the results meet the expected criteria, being the genetic scores significant in populations where the subjects hadn't child trauma or didn't consume *Cannabis*.

We also add an extra section with a study of the relation of the presence of environmental variables inside each gender in the countries where we have enough information (Turkey, Brazil, Spain and Others). In any case, this is a primary analysis and we would need much more observations and data from more countries to conclude something determinant. The details are collected in the table 14.

	Stratum	Variable	Presence	N	R2	Coeffs	P-values	Significance
ScoreInt	Men & ChildTrauma=0	Country	Turkey	146	0.00542	-0.741	0.35110	Not Significant
			Brazil	78	0.00011	0.044	0.90731	Not Significant
			Spain	203	0.00009	0.087	0.85217	Not Significant
	Men & Cannabis=0		Turkey	323	0.00620	-0.774	0.11395	Not Significant
			Brazil	89	0.00278	-0.204	0.53021	Not Significant
			Spain	68	0.01815	-1.586	0.14478	Not Significant
	Women & ChildTrauma=0	Country	Turkey	166	0.01346	1.195	0.05424	Not Significant
			Brazil	80	0.07919	1.116	**0.00044**	**Significant**
			Spain	122	0.00001	0.028	0.96103	Not Significant
	Women & Cannabis=0		Turkey	404	0.00709	0.807	**0.02537**	**Significant**
			Brazil	120	0.02695	0.688	**0.01168**	**Significant**
			Spain	89	0.03800	1.782	**0.01166**	**Significant**
ScoreEA	Men & ChildTrauma=0	Country	Turkey	146	0.00010	-0.074	0.84075	Not Significant
			Brazil	78	0.00012	0.040	0.89572	Not Significant
			Spain	203	0.01813	0.904	0.14507	Not Significant
	Men & Cannabis=0		Turkey	323	0.00053	-0.192	0.77134	Not Significant
			Brazil	89	0.00664	0.312	0.36015	Not Significant
			Spain	68	0.00135	0.209	0.47909	Not Significant
	Women & ChildTrauma=0	Country	Turkey	166	0.00041	0.162	0.73692	Not Significant
			Brazil	80	0.02446	0.615	**0.01629**	**Significant**
			Spain	122	0.04890	1.541	**0.00422**	**Significant**
	Women & Cannabis=0		Turkey	404	0.00711	0.652	**0.02519**	**Significant**
			Brazil	120	0.02446	0.615	**0.01629**	**Significant**
			Spain	89	0.04890	1.541	**0.00422**	**Significant**

Table 14: Statistical information of significant scores in different countries into sex plus environmental feature stratum in the explanation of Social Class

As saw before, the scores aren't significant in any country when men populations. Regarding women and no presence of child trauma or consumption of *Cannabis*, generally, in all countries, the scores are significant as it should be regard previous studies, but this doesn't meet for Turkey or Spain when there isn't child trauma.

In addition, a configuration of additive models has been developed with the purpose to study the additional amount of explained variance of the trait social class in the presence of more variables and interactions. Lastly, a best model has been found with *Stepwise* criterion. See table 15 for the details.

	Variables	Formula	R2	AIC	Coeffs
ScoreInt	Only genetic	SocCla~Age+Sex+ScoreInt+PCA2 +PCA3+PCA4+PCA8+PCA9 +Sex:PCA2+Age:Sex+Age:PCA4 +Age:PCA2+Sex:PCA3+PCA8:PCA9 +ScoreInt:PCA4+Sex:PCA9	0.06700	4472	0.189
	With Environmental	SocCla~Age+Sex+ScoreInt+PCA5 +PCA8+PCA9+Country+ChildTrauma +Sex:Country+Age:Country +Age:Sex+PCA8:PCA9+Age:PCA5 +ScoreInt:ChildTrauma+Age:ScoreInt	0.08641	4408	0.502
ScoreEA	Only genetic	SocCla~Age+Sex+ScoreEA+PCA2 +PCA3+PCA4+PCA5+PCA6+PCA8 +PCA9+PCA10+Sex:PCA2+Age:Sex +Age:PCA4+Age:PCA2+Sex:PCA3 +PCA8:PCA9+ScoreEA:PCA6 +Sex:PCA5+PCA4:PCA10+Sex:PCA9	0.07123	4462	0.132
	With Environmental	SocCla~Age+Sex+ScoreEA+PCA2 +PCA4+PCA5+PCA6+PCA8+PCA9 +Country+ChildTrauma+Cannabis +Sex:Country+Age:Country+Age:Sex +PCA8:PCA9+Age:PCA5 +ScoreEA:PCA6+Age:Cannabis +ScoreEA:ChildTrauma+PCA2:PCA8 +PCA6:Cannabis	**0.08810**	4402	0.253

Table 15: Best selection models with for the prediction of Social Class

The scores that explain the most the response variable are the educational attainment ones. If we use these in a ensemble model of genetic and environmental variables plus its GxE interactions, the explanation of the response social class increase an additional 1.7%, getting a total explained variance of **8.8%**.

To end with, we've tested the best model in terms of prediction, dividing our data in training and testing and calculating the confusion matrix[24] with true and false positives/negatives values. Also, we've applied the sensibility test obtaining the AUC value from the ROC curve[25].

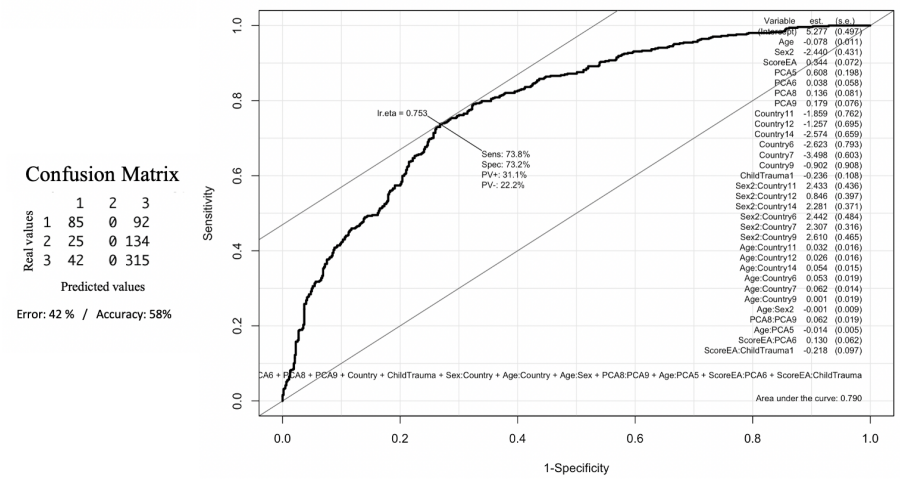

Figure 12: Confusion Matrix and ROC Curve of Social Class best model predictions

As we can observe in figure 12, the performance is not very good, as the model doesn't classify any observation in the *class 2*, and it has only a **58**% of accuracy. Conversely, the AUC of **0.79** means not bad performance, but is not surprisingly that with such few variables, the lack in the proportional odds assumptions and the vague definition and differentiation of the response levels, the prediction model still has much to improve.

To end with, we show the effect of the covariates derived from the best regression model (with educational attainment scores) in the impact of social class (figure 13).

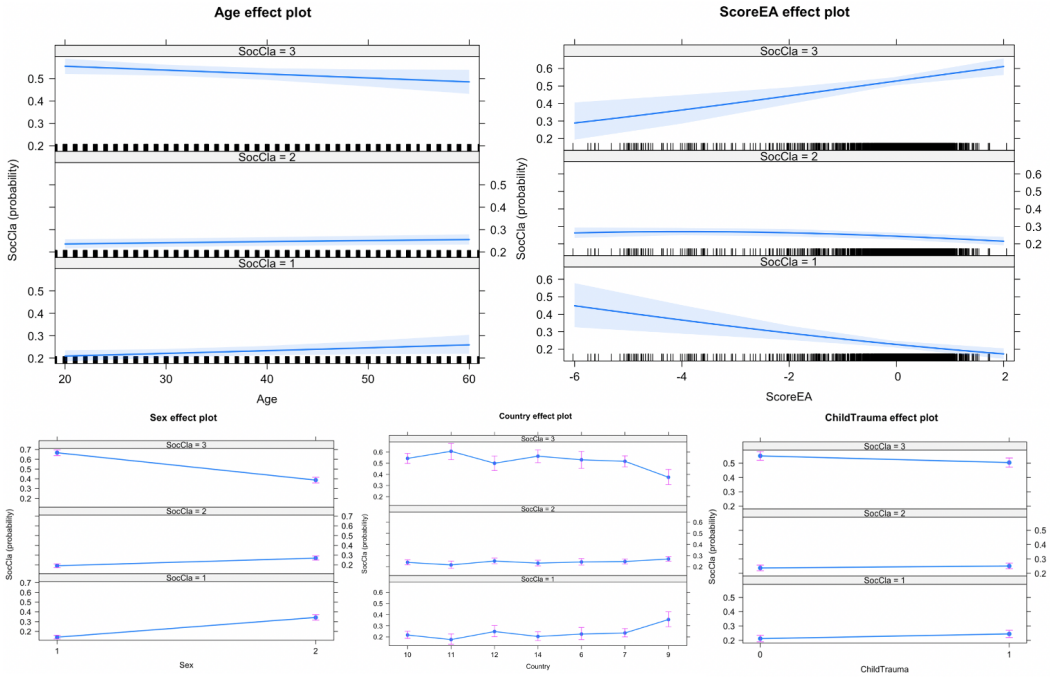

Figure 13: Effect plots of covariates in the Social Class response variable

In this case, the age doesn't influence positively in the belonging to upper class. Neither the female sex or have had a child trauma. On the contrary, the higher the educational attainment

score, the greater the probability to have bigger 'status'.
Finally, regarding the countries we can see again a rare behave when Netherlands have the highest probability to belong to the first social class, this make us think that the sample can be also biased because maybe all the individuals went to do the survey because they are retired or unemployed.

5.2.3 Educational Level

In the case of *Educational Level* response, we operate the same way than with *Social Class* and the details of the summary tables and figures are presented below.

	Variable	Presence	N	R2	Coeffs	P-values	Significance
ScoreInt	Sex	Men	1140	0.00189	0.129	**0.03704**	**Significant**
		Women	1161	0.00606	0.210	**0.00016**	**Significant**
	ChildTrauma	1	531	0.00127	0.095	0.09182	Not Significant
		0	609	0.00694	0.249	**0.00004**	**Significant**
	Cannabis	1	807	0.00387	0.164	**0.01258**	**Significant**
		0	1494	0.00492	0.212	**0.00012**	**Significant**
ScoreCog	Sex	Men	1140	0.00194	0.123	**0.03451**	**Significant**
		Women	1161	0.00131	0.104	0.07848	Not Significant
	ChildTrauma	1	531	0.00155	0.106	0.06321	Not Significant
		0	609	0.00242	0.146	**0.01545**	**Significant**
	Cannabis	1	807	0.00410	0.177	**0.01020**	**Significant**
		0	1494	0.00047	0.063	0.23489	Not Significant
ScoreEA	Sex	Men	1140	0.00474	0.198	**0.00096**	**Significant**
		Women	1161	0.00944	0.270	**0.00000**	**Significant**
	ChildTrauma	1	531	0.00401	0.171	**0.00276**	**Significant**
		0	609	0.00932	0.286	**0.00000**	**Significant**
	Cannabis	1	807	0.00876	0.249	**0.00017**	**Significant**
		0	1494	0.00717	0.255	**0.00000**	**Significant**

Table 16: Statistical information of significant scores in presence of environmental variables in the explanation of Social Class

In this case, the *intelligence scores* are significant in the explanation of the educational level in all cases, except when there is child trauma. On the contrary, the *cognition scores* aren't significant in women, when there is child trauma or the subjects don't consume *Cannabis*.

The educational *attainment scores* are always significant describing educational level.

	Stratatum	Variable	Presence	N	R2	Coeffs	P-values	Significance
ScoreInt	Men	ChildTrauma	1	531	0.00336	0.165	0.05675	Not Significant
			0	609	0.00092	0.095	0.29063	Not Significant
		Cannabis	1	495	0.00213	0.132	0.14974	Not Significant
			0	645	0.00251	0.160	0.07282	Not Significant
	Women	ChildTrauma	1	567	0.00049	0.056	0.45729	Not Significant
			0	594	0.01721	0.376	**0.00001**	**Significant**
		Cannabis	1	312	0.00708	0.201	**0.03601**	**Significant**
			0	849	0.00670	0.241	**0.00074**	**Significant**
ScoreCog	Men	ChildTrauma	1	531	0.00121	0.091	0.25247	Not Significant
			0	609	0.00351	0.176	**0.03918**	**Significant**
		Cannabis	1	495	0.00718	0.231	**0.00813**	**Significant**
			0	645	0.00019	0.040	0.62422	Not Significant

	Stratatum	Variable	Presence	N	R2	Coeffs	P-values	Significance
	Women	ChildTrauma	1	567	0.00162	0.112	0.17503	Not Significant
			0	594	0.00168	0.122	0.15712	Not Significant
		Cannabis	1	312	0.00089	0.086	0.45673	Not Significant
			0	849	0.00077	0.082	0.25153	Not Significant
ScoreEA	Men	ChildTrauma	1	531	0.00670	0.225	**0.00717**	**Significant**
			0	609	0.00303	0.166	0.05524	Not Significant
		Cannabis	1	495	0.00632	0.225	**0.01304**	**Significant**
			0	645	0.00529	0.219	**0.00921**	**Significant**
	Women	ChildTrauma	1	567	0.00288	0.142	0.07020	Not Significant
			0	594	0.01956	0.413	**0.00000**	**Significant**
		Cannabis	1	312	0.01370	0.287	**0.00353**	**Significant**
			0	849	0.00866	0.283	**0.00012**	**Significant**

Table 17: Statistical information of significant scores in presence of environmental variables into sex stratum in the explanation of Education Level

The *intelligence scores* aren't significant anymore in men, while they are still significant in women (except in presence of child trauma as before). Regarding, *cognition scores*, generally, they aren't related now with the educational level.

Finally, *educational attainment scores* are relevant in the explanation of educational levels, particularly, in men in presence of child trauma (not being so for men without traumas), something that don't happen in women, in any other score and in any other response.

	Stratum	Variable	Presence	N	R2	Coeffs	P-values	Significance
ScoreInt	Men & ChildTrauma=1	Country	Turkey	196	0.01213	1.100	**0.03950**	**Significant**
			Brazil	46	0.00098	-0.107	0.78936	Not Significant
			Spain	87	0.00216	0.457	0.56209	Not Significant
	Men & ChildTrauma=0		Turkey	146	0.00228	0.432	0.42731	Not Significant
			Brazil	81	0.01573	0.460	0.10871	Not Significant
			Spain	197	0.00955	0.837	0.06069	Not Significant
	Men & Cannabis=1		NA	NA	NA	NA	NA	NA
	Men & Cannabis=0		Turkey	322	0.00866	0.880	**0.02281**	**Significant**
			Brazil	92	0.00122	0.117	0.63924	Not Significant
			Spain	67	0.00687	0.953	0.37703	Not Significant
	Women & ChildTrauma=1	Country	Turkey	243	0.01054	0.962	**0.03357**	**Significant**
			Brazil	60	0.01487	0.582	0.20713	Not Significant
			Spain	54	0.04825	2.067	**0.02248**	**Significant**
	Women & ChildTrauma=0		Turkey	167	0.02307	1.516	**0.00786**	**Significant**
			Brazil	82	0.09760	1.143	**0.00006**	**Significant**
			Spain	118	0.04569	2.067	**0.00156**	**Significant**
	Women & Cannabis=1		NA	NA	NA	NA	NA	NA
	Women & Cannabis=0		Turkey	406	0.01089	0.985	**0.00420**	**Significant**
			Brazil	122	0.05567	0.946	**0.00026**	**Significant**
			Spain	86	0.03728	1.827	**0.01507**	**Significant**
ScoreCog	Men & ChildTrauma=1	Country	Turkey	196	0.00996	0.408	0.06205	Not Significant
			Brazil	46	0.00104	-0.070	0.78323	Not Significant
			Spain	87	0.00432	0.249	0.41190	Not Significant
	Men & ChildTrauma=0		Turkey	146	0.00118	-0.144	0.56835	Not Significant
			Brazil	81	0.00022	-0.041	0.84988	Not Significant
			Spain	197	0.00947	0.373	0.06174	Not Significant

	Stratum	Variable	Presence	N	R2	Coeffs	P-values	Significance
	Men & Cannabis=1		NA	NA	NA	NA	NA	NA
	Men & Cannabis=0		Turkey	322	0.00287	0.218	0.19025	Not Significant
			Brazil	92	0.00099	-0.070	0.67422	Not Significant
			Spain	67	0.02545	0.672	0.08903	Not Significant
	Women & ChildTrauma=1		Turkey	243	0.00187	0.199	0.37089	Not Significant
			Brazil	60	0.00102	0.084	0.74137	Not Significant
			Spain	54	0.01548	0.568	0.19619	Not Significant
	Women & ChildTrauma=0		Turkey	167	0.00641	0.360	0.16127	Not Significant
			Brazil	82	0.02607	-0.460	**0.03803**	**Significant**
			Spain	118	0.00727	0.313	0.20712	Not Significant
	Women & Cannabis=1		NA	NA	NA	NA	NA	NA
	Women & Cannabis=0		Turkey	406	0.00289	0.244	0.14004	Not Significant
			Brazil	122	0.00010	-0.027	0.87618	Not Significant
			Spain	86	0.01927	0.520	0.08053	Not Significant
ScoreEA	Men & ChildTrauma=1	Country	Turkey	196	0.00044	0.144	0.69456	Not Significant
			Brazil	46	0.01179	0.375	0.35379	Not Significant
			Spain	87	0.01314	0.803	0.15234	Not Significant
	Men & ChildTrauma=0		Turkey	146	0.00862	0.741	0.12244	Not Significant
			Brazil	81	0.00243	0.160	0.52884	Not Significant
			Spain	197	0.01617	0.680	**0.01466**	**Significant**
	Men & Cannabis=1		NA	NA	NA	NA	NA	NA
	Men & Cannabis=0		Turkey	322	0.00492	0.494	0.08624	Not Significant
			Brazil	92	0.00100	0.095	0.67145	Not Significant
			Spain	67	0.03483	1.351	**0.04666**	**Significant**
	Women & ChildTrauma=1		Turkey	243	0.00288	0.427	0.26637	Not Significant
			Brazil	60	0.04148	0.928	**0.03515**	**Significant**
			Spain	54	0.08365	2.358	**0.00266**	**Significant**
	Women & ChildTrauma=0		Turkey	167	0.00019	0.110	0.80953	Not Significant
			Brazil	82	0.07483	0.976	**0.00044**	**Significant**
			Spain	118	0.06330	1.818	**0.00020**	**Significant**
	Women & Cannabis=1		NA	NA	NA	NA	NA	NA
	Women & Cannabis=0		Turkey	406	0.00109	0.251	0.36452	Not Significant
			Brazil	122	0.05142	0.846	**0.00044**	**Significant**
			Spain	86	0.08877	2.006	**0.00018**	**Significant**

Table 18: Statistical information of significant scores in different countries into sex plusenvironmental feature stratum in the explanation of Education Level

In this case, generally all the scores aren't significant anymore in any country in men populations, this could happen because the reduction of the sample sizes.In fact, we weren't be able to analyze the cases with *Cannabis* consumption because there wasn't enough observations in Turkey. But there is a couple of exceptions: the *intelligence scores* are relevant in Turkey when there has been child trauma and there isn't *Cannabis* consumption, and *educational attainment scores* are in Spain when there hasn't been child trauma and there isn't *Cannabis* consumption.

Regarding women, since the scores explained more before, they still significant in almost all the cases, with the exception of *cognition scores* and *educational attainment* ones in Turkey.

	Variables	Formula	R2	AIC	Coefficients
ScoreInt	Only genetic	EduLevel~Age+Sex+ScoreInt+PCA2 +PCA4+PCA5+PCA9+Age:PCA2 +ScoreInt:PCA5+ScoreInt:PCA4 +Sex:PCA2+Age:Sex+Age:ScoreInt +PCA2:PCA4+PCA2:PCA9	0.06012	4574	0.584
	With Environmental	EduLevel~Age+ScoreInt+PCA5+PCA7 +PCA8+PCA10+Country+ChildTrauma +Cannabis+Age:Country +ScoreInt:Country+Age:Cannabis +PCA7:PCA8+PCA5:Cannabis +Age:PCA7+Age:PCA10	0.08705	4476	1.108
ScoreCog	Only genetic	EduLevel~Age+ScoreCog+PCA2+PCA4 +PCA5+PCA9+Age:PCA2+PCA2:PCA4 +PCA2:PCA9+Age:ScoreCog	0.04778	4623	-0.218
	With Environmental	EduLevel~Age+Sex+ScoreCog+PCA2 +PCA3+PCA5+PCA7+PCA8+PCA9 +PCA10+Country+ChildTrauma +Cannabis+Age:Country+PCA5:Country +Age:Cannabis+Sex:PCA2+Age:Sex +Age:PCA7+Age:PCA3+PCA2:PCA9 +ScoreCog:PCA8+PCA7:PCA8 +PCA8:PCA10+PCA2:Cannabis +Age:PCA10+PCA5:Cannabis+Age:PCA2	0.08731	4487	0.108
ScoreEA	Only genetic	EduLevel~Age+Sex+ScoreEA+PCA2 +PCA3+PCA4+PCA5+PCA8 +Age:PCA2+ScoreEA:PCA3 +Sex:PCA2+Age:Sex +PCA2:PCA3+Sex:PCA5	0.05714	4586	0.370
	With Environmental	EduLevel~Age+Sex+ScoreEA+PCA2 +PCA3+PCA7+PCA8+PCA10 +Country+ChildTrauma+Cannabis +Age:Country+ScoreEA:Country +Age:Cannabis+PCA2:PCA3 +Sex:PCA2+Age:Sex+Age:PCA7 +PCA7:PCA8+Age:PCA10	**0.09421**	4449	0.416

Table 19: Best selection models with for the prediction of Education Level

The scores which are more related to the trait *Education Level* are the educational attainment scores, this make a lot of sense since they are very connected because if you have high education capacities is normal that you achieved good education level. In total, this genetic variable along with the rest of biological and environmental factors (and its interactions) reach to explain a total of **9.4%** of the variance of the education level trait.

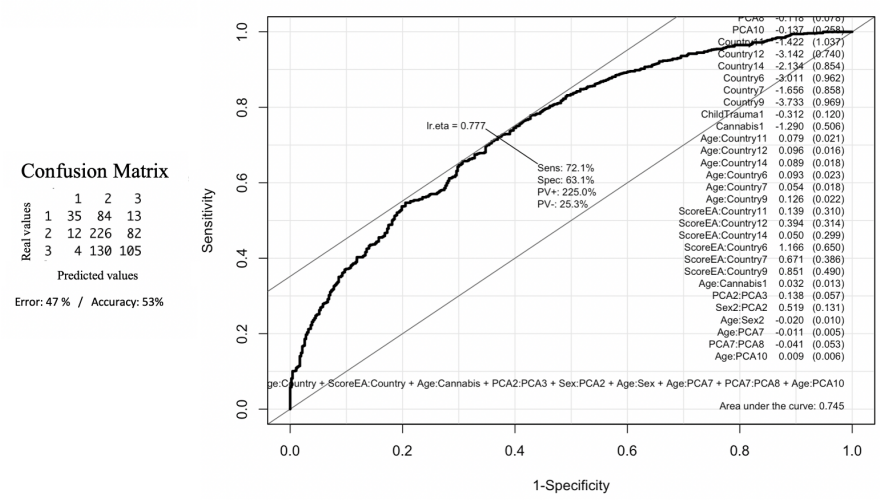

Figure 14: Confusion Matrix and ROC Curve of Educational Level best model predictions

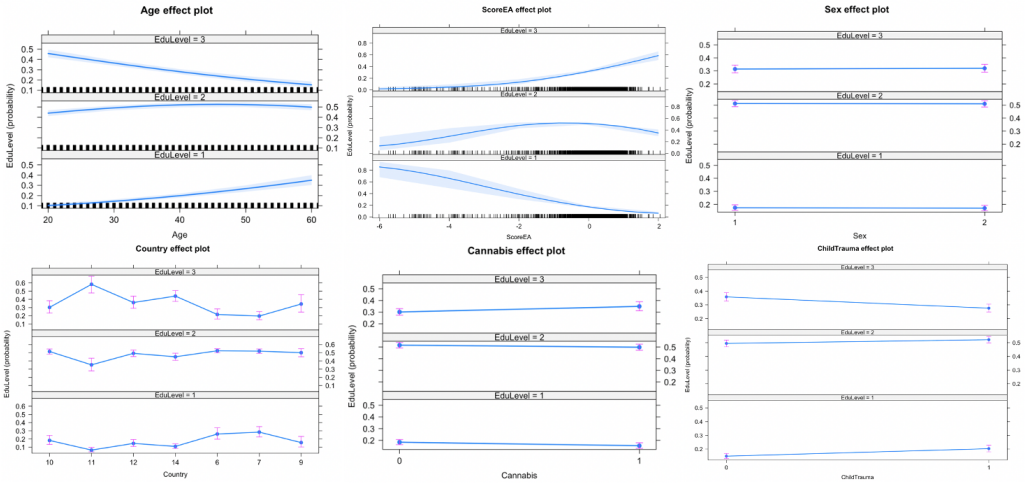

Figure 15: Effect plots of covariates in the Educational Level response variable

To end with, the best model (with *ScoreEA*) and environmental variables is used to estimate predictions of the education level. The result is not so much good, as the accuracy only reach **53%**, but it can be a good primary vision.

Regarding the effects of the variables in the response, we can conclude that the age and the presence of child trauma is not a good indicator to rise higher classes, on contrary, as more age, less probability to reach the third class. Conversely, greater educational attainment scores and the consumption of Cannabis boost the probability to be in upper classes. The gender is not relevant in the definition of education level, and, UK and Netherlands have much more probability top be in the third education level than Turkey, Brazil, Italy or Spain. In fact, these two last countries have the highest probability to be in the lowest level.

6 Conclusions

For concluding, we summarize the most important results that we have obtained and the important facts we have realised during the studies.

Regarding the different scores, we can conclude that the cognition ones aren't too good indicators of the studied traits. By contrast, *educational attainment scores* are usually the best in the explanations, as long as they are significant. For example, for *Income* trait, the intelligence ones are the only feasible, but in the categorical responses, the *educational attainment scores* always explain the highest variance in the best model. This make sense because the GWAS sample, with which the scores are built, is the greatest one (around three more times). Therefore, for further analysis, is important to keep in mind that we should do genetic studies with the highest possible discovery data sample size.

The best predicted trait is the *Income* with a 39% of explanation of its total variance, while Social Class is much less explained, this could be due to the nature of the categorical responses against a continuous numerical variable. To improve this reality, we could define better the levels of the categorical responses and balance the amount of observations in the categories.

Generally, the effect of the covariates in the responses are the greater the scores, the better. Females usually have less probabilities to reach big salaries or upper classes and the presence of harmful factors decrease the wages and the probability to have better education/social levels.

On the other hand, regarding the interaction of environmental factors, the scores explain better when other variables don't interfere. For instance, if the individual consume Cannabis or have had child traumas, those factors condition the predictions in such a way that finally, the scores are not longer relevant in the explanation of the traits.

In addition, the effect of have had traumas is dependent of the gender. In women, the traumas affect more, maybe because those traumas are worst cases or because the society haven't implemented good helps to resolve this.

As well, the use of *Cannabis* seems related to the gender. Usually, the fact to consume this drug is a bad indicator and it make the scores to explain less, but in women, this circumstance doesn't affect as much as in men, where the impact is really clear.

Regarding the countries, we would need much more observations to conclude something but at first vision, looks like where there is less 'biological meritocracy' is in Turkey, where the scores are less significant and there should be other factors that disrupt the genetic expression reaching a particular social or education level.

Finally, the efficiency of the prediction models is not too much good. This is not surprisingly due to the absence of proportional odds assumptions, the short amount of used variables and the lack in the definition of the categories.

If we use GWAS with more sample size, more biological and environmental variables, response classes with better definition and other advanced machine learning models, we could, for sure, improve much more the accuracy and the power of the prediction models and venture really important social and academic features using our genetic component.

7 Bibliography

[1] Robert Plomin and Sophie von Stumm. The new genetics of intelligence. 2018.

[2] J. Sulc, N. Monier. Quantification of the overall contribution of gene-environment interaction for obesity-related traits. 2020. Retrieved from: https://www.researchgate.net/publication/339918607 _Quantification_of_the_overall_contribution_of_gene-environment _interaction_for_obesity- related_traits

[3] E. Smith-Woolley, J.Pingault. Differences in exam performance between pupils attending selective and non-selective schools mirror the genetic differences between them. 2018. Retrieved from: https://www.nature.com/articles/s41539-018-0019-8

[4] L. D. Brooks. International HapMap Project. 2012. Retrieved from: https://www.genome.gov /10001688/international-hapmap-projectal-7

[5] K. Norrgard. Genetic Variation and Disease: GWAS. 2008. Retrieved from: https://www.nature. com/scitable/topicpage/genetic-variation-and-disease-gwas-682/

[6] B. Pasaniuc, A. L. Price. Dissecting the genetics of complex traits using summary association statistics. 2007. Retrieved from: https://www.nature.com/articles/nrg.2016.142).

[7] A. Torkamani, N.E. Wineinger. The personal and clinical utility of polygenic risk scores. 2018. Retrieved from: https://www.nature.com/articles/s41576-018-0018-x? spm=smpc. content.content.1.15473376001127D5jliP

[8] J. Van Os, G. Kenis. The environment and schizophrenia. 2010. Retrieved from: https://www.nature.com/articles/nature09563?page=22

[9] A. Ardila, S. Lundberg. Gender gaps in the effects of childhood family environment: Do they persist into adulthood?. 2017. Retrieved from: https://www.sciencedirect.com/science/article /abs/pii/S0014292117300740

[10] M. Crous-Bou, M. Gascon. Impact of urban environmental exposures on cognitive performance and brain structure of healthy individuals at risk for Alzheimer's dementia. 2020. Retrieved from: https://www.sciencedirect.com/science/article/pii/S0160412019321130

[11] M. Melchior, C. Bolze. Early cannabis initiation and educational attainment: is the association causal? Data from the French TEMPO study. 2017. Retrieved from: https://academic.oup.com /ije/article/46/5/1641/3830700 C. L. Millsaps, R. L. Azrin MS. Neuropsychological Effects of Chronic Cannabis Use on the Memory and Intelligence of Adolescents. 2008. Retrieved from: https://www.tandfonline.com /doi/abs/10.1300/J029v03n01_05

[12] M. Trzaskowski, N. Harlaar. Genetic influence on family socioeconomic status and children's intelligence. 2014. Retrieved from: https://www.sciencedirect.com/science/article/pii/ S0160289613001682; https://journals.sagepub.com /doi/abs/10.1177/1043463109348987)

[13] S. Rudenstine, A. Espinosa. Examining the role of trait emotional intelligence on psychiatric symptom clusters in the context of lifetime trauma. 2018. Retrieved from: https://www.sciencedirect.com/science/article/abs/pii/S0191886918300953

[14] F. Ullén, D. Z. Hambrick. Rethinking expertise: A multifactorial gene–environment interaction model of expert performance. 2016. Retrieved from: https://psycnet.apa.org/record/2015-56715-001

[15] D. N. Figlio, J. Freese. Socioeconomic status and genetic influences on cognitive development. 2017. Retrieved from: https://www.pnas.org/content/114/51/13441.short

[16] C. Henquet, M. Di Forti. Gene-Environment Interplay Between Cannabis and Psychosis. 2008. Retrieved form: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2632498/

[17] W. Johnson, I. J. Deary. Genetic and environmental transactions underlying educational attainment. 2009. https://www.sciencedirect.com/science/article/abs/pii/S0160289609000877

[18] Linkage disequilibrium. 2020. Retrieved from: https://en.wikipedia.org/wiki/ Linkage_disequilibrium

[19] R. E. Marioni, S. J. Ritchie. Genetic variants linked to education predict longevity. 2016. Retrieved from: https://www.pnas.org/content/pnas/113/47/13366.full.pdf

[20] E.Vassos, M. Di Forti. An Examination of Polygenic Score Risk Prediction in Individuals With First-Episode Psychosis. 2016. Retrieved from: https://sci-hub.ee/10.1016/j.biopsych.2016.06.028

[21] P. Allison. What's the Best R-Squared for Logistic Regression?. 2013. Retrieved from: https://statisticalhorizons.com/r2logistic

[22] L. Duncan, H. Shen. Analysis of polygenic risk score usage and performance in diverse human populations. 2019. Retrieved from: https://www.nature.com/articles/s41467-019-11112-0.pdf

[23] Stepwise regression. 2020. Retrieved from: https://en.wikipedia.org/wiki/Stepwise_regression

[24] Confusion matrix. 2020. Retrieved from: https://en.wikipedia.org/wiki/Confusion_matrix

[25] Receiver operating characteristic (ROC). 2020. Retrieved from: https://en.wikipedia.org/wiki /Receiver_operating_characteristic

8 Appendix

8.1 Genetic Data Discovery

Savage Jansen Intelligence Meta-analysis 2018.txt

8.2 Genetic Data Target

EUGEI WP2 WP6 PassingQC All Population.bim

8.3 Circumstantial Questionnaire: CRD

CRD EUGEI.pdf

8.4 Excel of Data Target WP2 Package

EUGEI WP2.xlsx

8.5 Excel of Data Target WP6 Package

EUGEI WP6.xlsx

8.6 Final Genetic and Environmental Dataset for the project

Final Regression Dataset.xlsx

8.7 Code of the project

Coding.R