Tutorial 3
Economics and Genetics (FEB13089)

Erasmus School of Economics – Bachelor (Block 2)

Learning goals

- Becoming familiar with basic functionality of GCTA.
- Being able to estimate heritability using molecular genetic data.

If you know how to use PLINK (Tutorial 2), you will find it easy to use GCTA as their ways of working (e.g., inputting data and outputting results) are very similar. In this tutorial we focus on the functions that are used most often in GCTA, i.e., estimating the genetic relationship between individuals from a genome-wide scan of SNPs, and estimating the variance explained by all genotyped SNPs. We also consider bivariate analysis. At the end of this tutorial, you will get input for your Individual Assignment.

Data

In this tutorial, we continue working with the files Example.bed, Example.bim, Example.fam, and Example_height.pheno you already worked with in Tutorial 2. For Tutorial 3, you also need the Example_height_binary.pheno and Example_height_education_bivariate.pheno which are available on Canvas. For the Individual Assignment, you will need the files Females.bed, Females.bim, Females.fam, Females_income.pheno, Males.bed, Males.bim, Males.fam, and Males_income.pheno from Canvas. In this Tutorial, I'll assume you have stored all these files in the folder C:\Users\Niels\Documents\EUR\E&G\Tutorials. Whenever this path to the folder with data is mentioned in this Tutorial, please use your own path to it.

Introducing GCTA

GCTA (Genome-wide Complex Trait Analysis, https://cnsgenomics.com/software/gcta/) was originally designed to estimate the proportion of phenotypic variance explained by genome- or chromosome-wide scans of SNPs for complex traits (the GREML method), and has subsequently extended for many other analyses to better understand the genetic architecture of complex traits. Just as PLINK, GCTA is a *command line program* that does not need to be installed. Instead, you run it from the command line (CMD in Windows or Terminal in MacOS, see Tutorial 2).

GCTA can be downloaded from http://cnsgenomics.com/software/gcta/#Download. Please make sure to download the version suitable for your operating system (Windows, MacOS, or UNIX). Store the files in the downloaded folder in your data folder C:\Users\Niels\Documents\EUR\E&G\Tutorials.[1] **Important note for Windows users: In your downloaded folder, there is a subfolder "bin". Make sure to copy the contents of this folder also to your data folder C:\Users\Niels\Documents\EUR\E&G\Tutorials.**

Similar to PLINK (Tutorial 2), you need to navigate to the folder in which the GCTA executable is placed to run the program. When you are in this folder, type "gcta64" and press enter to check

[1] You are again advised to run this tutorial on your own computer, as it is somewhat complicated to run GCTA on a university PC. However, it is possible, see the document "How to run PLINK and GCTA on a university PC.pdf" on Canvas.

whether GCTA is running. ***Important note: On some MacBooks, you need to type "./gcta64" instead of "gcta64"; You need to do this every time you run GCTA in this Tutorial.***

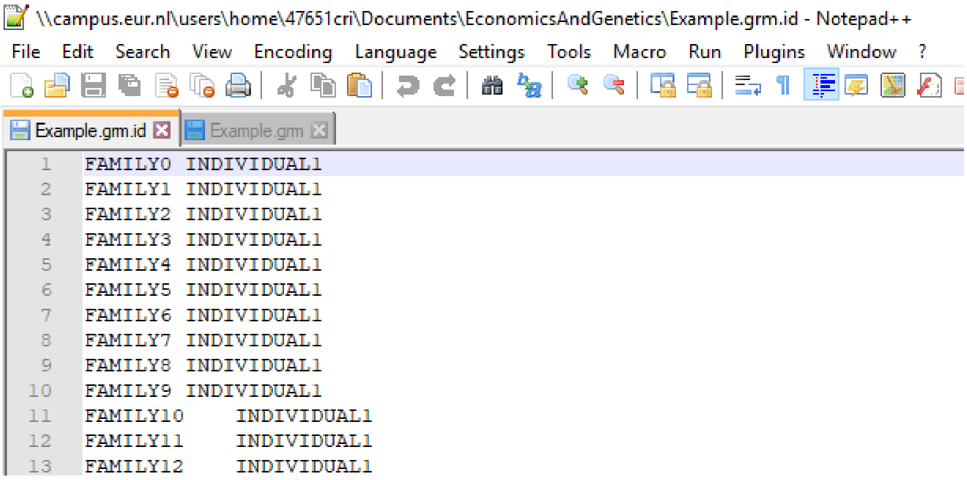
Calculating the genetic relationship matrix (GRM) from all the autosomal SNPs

In this tutorial, we make use of the same example data as in Tutorial 2. The files Example.bed, Example.bim, and Example.fam are probably still in your working directory C:\Users\Niels\Documents\EUR\E&G\Tutorials. If not, please put these files there again. These data files include information about 23,825 SNPS for 2,000 individuals (have a look at the BIM and FAM file to check this).

With the --make-grm function, we can construct the Genetic Relationship Matrix (GRM). The GRM includes the pairwise genetic relationship between each pair of individuals in the dataset. For constructing the GRM, we only use autosomal SNPs (SNPs on chromosomes 1-22, with --autosomal) and we exclude SNPs with low frequency in the population (with --maf). We give the resulting GRM the name "Example".

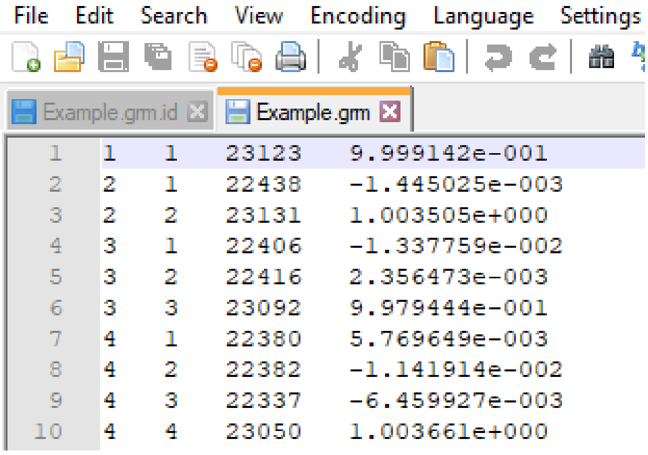
```
gcta64 --bfile Example --autosome --maf 0.01 --make-grm --out Example
```

The genetic relationship matrix is saved in two files: Example.grm.id and Example.grm.bin (the file Example.grm.N.bin is less important for now). Let's have a look at the first file first:


\\campus.eur.nl\users\home\47651cri\Documents\EconomicsAndGenetics\Example.grm.id - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
Example.grm.id
Example.grm

```
FAMILY0	INDIVIDUAL1
FAMILY1	INDIVIDUAL1
FAMILY2	INDIVIDUAL1
FAMILY3	INDIVIDUAL1
FAMILY4	INDIVIDUAL1
FAMILY5	INDIVIDUAL1
FAMILY6	INDIVIDUAL1
FAMILY7	INDIVIDUAL1
FAMILY8	INDIVIDUAL1
FAMILY9	INDIVIDUAL1
FAMILY10	INDIVIDUAL1
FAMILY11	INDIVIDUAL1
FAMILY12	INDIVIDUAL1
```

The file Example.grm.id only contains the IDs of individuals included in the GRM. The binary file Example.grm.bin contains for every pair of individuals the estimated genetic relation. Due to its binary nature, you can't directly open this file on your computer (only after some conversion). Therefore, have a careful look at the screenshot below to understand the contents of Example.grm.bin.



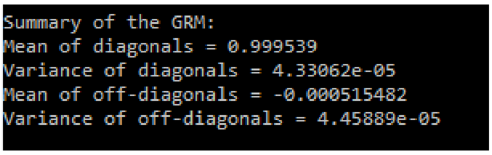
```
1	1	23123	9.999142e-001
2	1	22438	-1.445025e-003
2	2	23131	1.003505e+000
3	1	22406	-1.337759e-002
3	2	22416	2.356473e-003
3	3	23092	9.979444e-001
4	1	22380	5.769649e-003
4	2	22382	-1.141914e-002
4	3	22337	-6.459927e-003
4	4	23050	1.003661e+000
```

The first two columns in the file indicate the individuals for which the genetic relationship is estimated. The third column indicates the number of SNPs used to estimate the genetic relationship. The fourth column includes the genetic relationship. The more independent the individuals in your sample, the more these relations will be close to 0. The genetic relation of an individual to him or herself (the diagonal of the genetic relationship matrix) will be 1 if there is no inbreeding in your sample (deviations indicate that the individual has more or less homozygous SNPs than expected in the population).

For example, in the first row we see that the genetic relationship between the *first* and the *first* individual in the sample equals 0.999142 (~1) and is estimated using 23,123 SNPs. In the second row we see that the genetic relationship between the *first and second* individual in the sample equals -0.00145 (~0) and is estimated using 22,438 SNPs. Later on in this tutorial, we will check if there are cryptically related individuals present in the dataset (individuals who are genetically related although they are considered to be part of different families).

When you run the following command, some additional information about the GRM is given:

```
gcta64 --bfile Example --autosome --maf 0.01 --make-grm-gz --out Example
```



```
Summary of the GRM:
Mean of diagonals = 0.999539
Variance of diagonals = 4.33062e-05
Mean of off-diagonals = -0.000515482
Variance of off-diagonals = 4.45889e-05
```

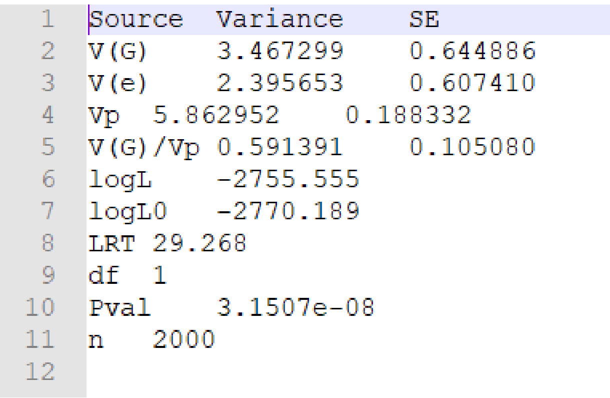
As expected, the mean of the diagonal elements of the GRM is approximately 1 (the genetic relationship of individuals to themselves). The mean of the off-diagonal elements is approximately 0, suggesting that the individuals in the sample are genetically unrelated.

GCTA-GREML analysis: estimating the variance explained by the SNPs

If you have constructed the GRM, you can use it to estimate the proportion of variance in the phenotype that can be explained by genetic factors (the SNPs in your sample):

```
gcta64 --grm Example --pheno Example_height.pheno --reml --out Example_height
```

The results will be saved in the file Example_height.hsq. You can open this file in a text editor:

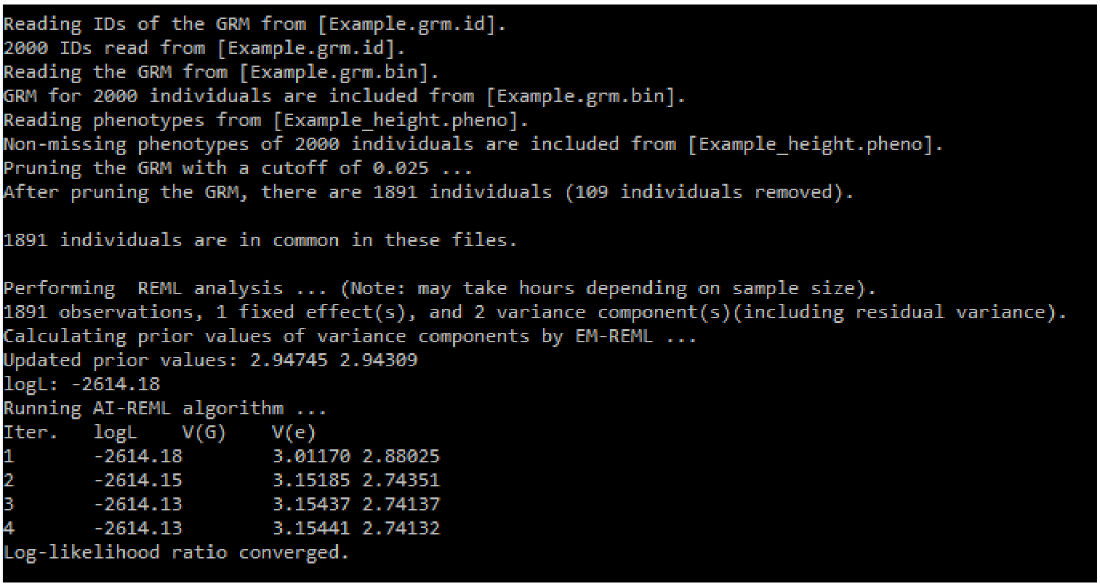


```
Source	Variance	SE
V(G)	3.467299	0.644886
V(e)	2.395653	0.607410
Vp	5.862952	0.188332
V(G)/Vp	0.591391	0.105080
logL	-2755.555
logL0	-2770.189
LRT	29.268
df	1
Pval	3.1507e-08
n	2000
```

The first three lines give the estimates for variance components in your model. V(G) is the variance component for the additive genetic part, V(e) is the variance component for the environmental part, and V(p) is the variance of the phenotype. V(G)/V(p) is the proportion of the phenotypic variance that can be explained by genetic variance. This is the parameters you are interested in, and is often denoted as h^2_{SNPs}. The *p*-value (derived from a loglikelihood test with and without the GRM) indicates that the GRM contributes significantly to the fit of the model ($p = 3.15\times10^{-8} < 0.05$).

We want to be sure that no pair of individuals that is too much related are included in the analysis sample, because strong genetic relatedness can imply environmental relatedness (This is called population stratification; In Lecture 4 we will look at another way to control for population stratification, namely using principal components). Therefore, it is common to exclude all pairs of individuals that have a genetic relationship of more than 0.025. This is equivalent to cousins two or three times removed (see the pedigree figure in Lecture 3's slides).

```
gcta64 --grm Example --pheno Example_height.pheno --reml --out Example_height_cutoff --grm-cutoff 0.025
```

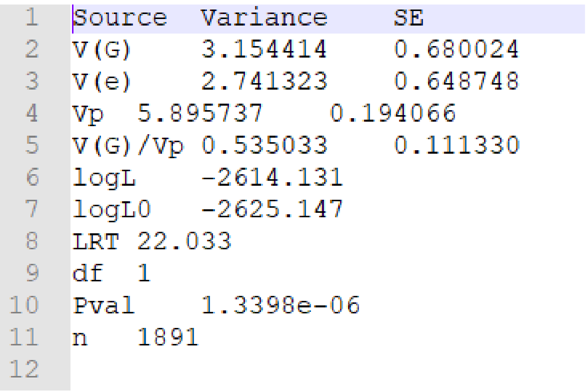


```
Reading IDs of the GRM from [Example.grm.id].
2000 IDs read from [Example.grm.id].
Reading the GRM from [Example.grm.bin].
GRM for 2000 individuals are included from [Example.grm.bin].
Reading phenotypes from [Example_height.pheno].
Non-missing phenotypes of 2000 individuals are included from [Example_height.pheno].
Pruning the GRM with a cutoff of 0.025 ...
After pruning the GRM, there are 1891 individuals (109 individuals removed).

1891 individuals are in common in these files.

Performing  REML analysis ... (Note: may take hours depending on sample size).
1891 observations, 1 fixed effect(s), and 2 variance component(s)(including residual variance).
Calculating prior values of variance components by EM-REML ...
Updated prior values: 2.94745 2.94309
logL: -2614.18
Running AI-REML algorithm ...
Iter.   logL    V(G)    V(e)
1       -2614.18        3.01170 2.88025
2       -2614.15        3.15185 2.74351
3       -2614.13        3.15437 2.74137
4       -2614.13        3.15441 2.74132
Log-likelihood ratio converged.
```

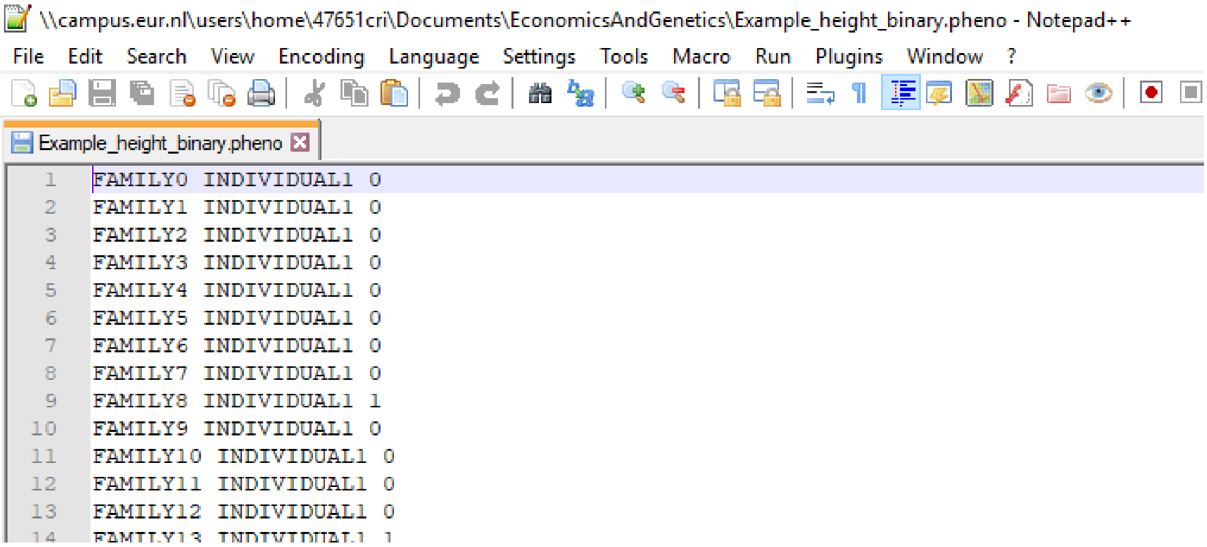
The GCTA output shows that 109 individuals are being dropped from the analysis, because they are too closely related to another individual in the dataset. From the results file "Example_height_cutoff.hsq" you can see that h^2_{SNPs} is a bit lower than in the model without relationship cutoff. Genetic relatedness is often correlated with environmental relatedness and therefore the earlier estimate was upward biased. Still, we still find a significant estimate for h^2_{SNPs} ($p = 1.34\times10^{-6} < 0.05$):



```
Source	Variance	SE
V(G)	3.154414	0.680024
V(e)	2.741323	0.648748
Vp	5.895737	0.194066
V(G)/Vp	0.535033	0.111330
logL	-2614.131
logL0	-2625.147
LRT	22.033
df	1
Pval	1.3398e-06
n	1891
```

GCTA-GREML analysis for a case-control outcome

So far, we analyzed a continuous outcome variable (i.e., body height in centimeters). As also discussed in Lecture 3, many interesting outcome variables are on a binary scale. For example, whether you have a certain disease (you are a "case") or not (you are a "control"). To analyze a case-control outcome in GCTA, the phenotypic values of cases and controls should be specified as 1 and 0, respectively. We are going to analyze such a case-control outcome now. For this purpose, download the file "Example_height_binary.pheno" from Canvas and open it in Notepad++. You'll see:



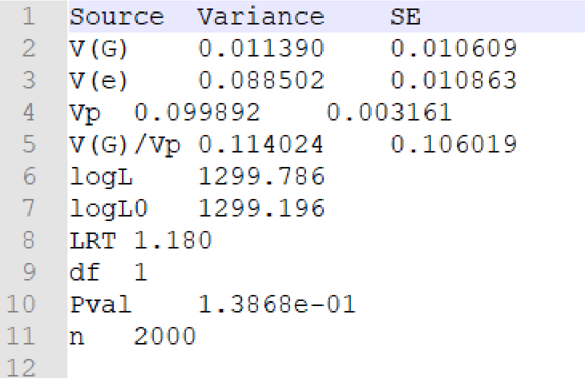
\\campus.eur.nl\users\home\47651cri\Documents\EconomicsAndGenetics\Example_height_binary.pheno - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
Example_height_binary.pheno

```
FAMILY0 INDIVIDUAL1 0
FAMILY1 INDIVIDUAL1 0
FAMILY2 INDIVIDUAL1 0
FAMILY3 INDIVIDUAL1 0
FAMILY4 INDIVIDUAL1 0
FAMILY5 INDIVIDUAL1 0
FAMILY6 INDIVIDUAL1 0
FAMILY7 INDIVIDUAL1 0
FAMILY8 INDIVIDUAL1 1
FAMILY9 INDIVIDUAL1 0
FAMILY10 INDIVIDUAL1 0
FAMILY11 INDIVIDUAL1 0
FAMILY12 INDIVIDUAL1 0
FAMILY13 INDIVIDUAL1 1
```

There are 225 cases and 1,775 controls in this dataset, and the case-control status is defined based on an individual's height. An individual is a "case" if her height is above 182,5 cm. An individual is a "control" if her height is below 182,5 cm. Let's analyze the heritability of "being very tall":

```
gcta64 --grm Example --pheno Example_height_binary.pheno --reml --out Example_height_binary
```

Your output should look like:



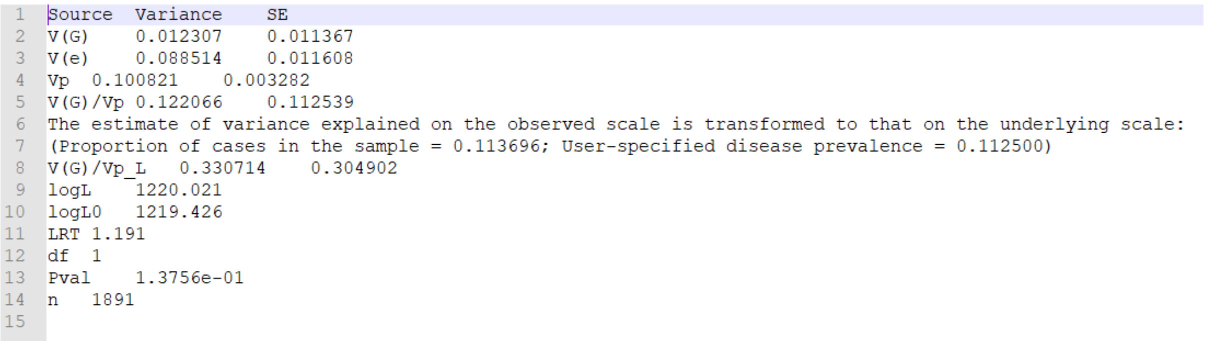
```
Source	Variance	SE
V(G)	0.011390	0.010609
V(e)	0.088502	0.010863
Vp	0.099892	0.003161
V(G)/Vp	0.114024	0.106019
logL	1299.786
logL0	1299.196
LRT	1.180
df	1
Pval	1.3868e-01
n	2000
```

- ***Exercise:** Check yourself that the heritability estimate remains insignificant (p = 0.13756) if you exclude cryptically related individuals (with the --grm-cutoff option).*

GCTA automatically detected that you were analyzing a case-control outcome variable, because it gives you the message (on the command line) ‘Note: you can specify the disease prevalence by the option --prevalence so that GCTA can transform the variance explained to the underlying liability scale’. As discussed in Lecture 3, we assume that a case-control status results from an underlying continuous liability scale. As there are 225 cases in our sample, the prevalence of our ‘trait’ is 225/2000=0.1125. We can specify this with the --prevalence flag. GCTA will now automatically transform the h^2_{SNPs} on the observed 0-1 scale to the underlying liability scale.

```
gcta64 --grm Example --pheno Example_height_binary.pheno --reml --out Example_height_binary_cutoff_liability_scale --prevalence 0.1125 --grm-cutoff 0.025
```

Your output should look like:



```
Source	Variance	SE
V(G)	0.012307	0.011367
V(e)	0.088514	0.011608
Vp	0.100821	0.003282
V(G)/Vp	0.122066	0.112539
The estimate of variance explained on the observed scale is transformed to that on the underlying scale:
(Proportion of cases in the sample = 0.113696; User-specified disease prevalence = 0.112500)
V(G)/Vp_L	0.330714	0.304902
logL	1220.021
logL0	1219.426
LRT	1.191
df	1
Pval	1.3756e-01
n	1891
```

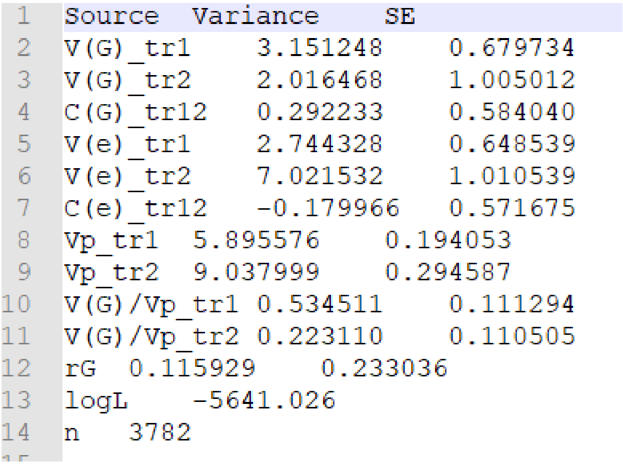
You can see that the heritability estimate has been adjusted upwards (to 0.331, SE=0.305), but has remained insignificant (*p*=0.138). These results are exemplary for the notion that (given your sample size) the statistical power to detect a significant heritability estimate is larger when analyzing a continuous outcome variable than when analyzing a case-control outcome variable.

Bivariate REML

In Lecture 3, we also discussed the estimation of bivariate genetic correlations with GREML. The --reml-bivar option of GCTA produces the SNP heritability of two traits and the genetic correlation between them. It can be performed with two quantitative traits, two binary traits, or a continuous and binary trait Here, we will perform the analysis with two continuous traits: Height in cm (which we analyzed before) and educational attainment (in years). For this purpose, you have downloaded the file “Example_height_education_bivariate.pheno” from Canvas. Instead of --reml we use --reml-bivar 1 2 (indicating that we use the first and second phenotype in the file test_bivariate.phen.

```
gcta64                --grm               Example               --pheno
Example_height_education_bivariate.pheno  --reml-bivar  1  2  --out
Example_height_education_bivariate --grm-cutoff 0.025
```

The results are as follows:



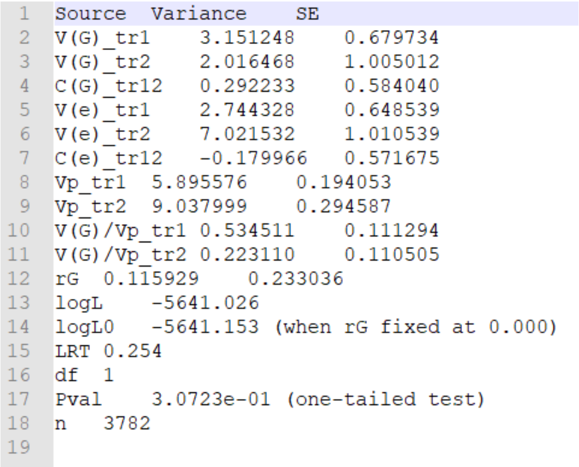
```
 1  Source  Variance    SE
 2  V(G)_tr1    3.151248    0.679734
 3  V(G)_tr2    2.016468    1.005012
 4  C(G)_tr12   0.292233    0.584040
 5  V(e)_tr1    2.744328    0.648539
 6  V(e)_tr2    7.021532    1.010539
 7  C(e)_tr12   -0.179966   0.571675
 8  Vp_tr1  5.895576    0.194053
 9  Vp_tr2  9.037999    0.294587
10  V(G)/Vp_tr1 0.534511    0.111294
11  V(G)/Vp_tr2 0.223110    0.110505
12  rG  0.115929    0.233036
13  logL    -5641.026
14  n   3782
```

In line 10 and 11 we see the SNP heritability estimates for phenotype 1 (height) and 2 (educational attainment) (53% and 22%, respectively). In row 12, we see the estimate for the genetic correlation (r_g, the proportion of variance that two traits share due to shared genetic origins). In our case, it equals 0.116 (SE=0.233). Maybe you are surprised to see the sample size 3782. This equals the number of individuals in the sample after the genetic relatedness cutoff (1981) multiplied by the number of phenotypes (2).

In many cases, we are interested in the question whether the genetic correlation between two phenotypes is significantly different from 0. If you want to test whether the genetic correlation is different from a particular value (values used most often are -1, 0, and 1), you can use the --reml-bivar-lrt-rg command, e.g. --reml-bivar-lrt-rg 0 performs a loglikelihood ratio test of r_g being different from 0.

```
gcta64  --grm  Example  --pheno  Example_height_education_bivariate.pheno  --
reml-bivar 1 2 --out Example_height_education_bivariate2 --grm-cutoff 0.025
--reml-bivar-lrt-rg 0
```

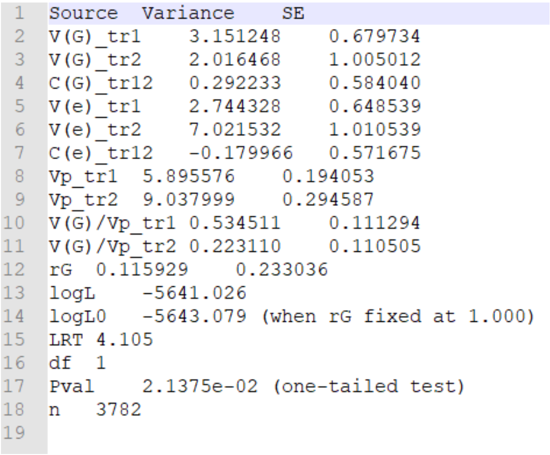
The results show (see the screenshot below) that r_g is not significantly different from 0 (p = 0.307 > 0.05). This result is exemplary for the notion that large samples are needed to estimate genetic correlations which are significantly different from zero.



```
Source  Variance    SE
V(G)_tr1    3.151248    0.679734
V(G)_tr2    2.016468    1.005012
C(G)_tr12   0.292233    0.584040
V(e)_tr1    2.744328    0.648539
V(e)_tr2    7.021532    1.010539
C(e)_tr12   -0.179966   0.571675
Vp_tr1  5.895576    0.194053
Vp_tr2  9.037999    0.294587
V(G)/Vp_tr1 0.534511    0.111294
V(G)/Vp_tr2 0.223110    0.110505
rG  0.115929    0.233036
logL    -5641.026
logL0   -5641.153 (when rG fixed at 0.000)
LRT 0.254
df  1
Pval    3.0723e-01 (one-tailed test)
n   3782
```

However, in some cases we are interested in testing whether two phenotypes are genetically the same. Hence, whether r_g is equal to 1. Let's also test this with our data:

```
gcta64 --grm Example --pheno Example_height_education_bivariate.pheno --reml-bivar 1 2 --out Example_height_education_bivariate3 --grm-cutoff 0.025 --reml-bivar-lrt-rg 1
```



```
Source  Variance     SE
V(G)_tr1     3.151248     0.679734
V(G)_tr2     2.016468     1.005012
C(G)_tr12    0.292233     0.584040
V(e)_tr1     2.744328     0.648539
V(e)_tr2     7.021532     1.010539
C(e)_tr12    -0.179966    0.571675
Vp_tr1  5.895576     0.194053
Vp_tr2  9.037999     0.294587
V(G)/Vp_tr1 0.534511     0.111294
V(G)/Vp_tr2 0.223110     0.110505
rG  0.115929     0.233036
logL    -5641.026
logL0   -5643.079 (when rG fixed at 1.000)
LRT 4.105
df  1
Pval    2.1375e-02 (one-tailed test)
n   3782
```

The results show (see the screenshot above) that r_g is significantly different from 1 ($p = 0.021 < 0.05$).

Individual assignment

Download the files "Females.bed", "Females.bim", "Females.fam", "Females_income.pheno", "Males.bed", "Males.bim", "Males.fam", and "Males_income.pheno" from Canvas. These files contain income data from 3,500 females and 3,500 males. The genetic files contain a representative genome-wide scans of SNPs (38,560 SNPs) for both samples.

Your goal is to provide new evidence for the heritability of personal income (among females *and* among males; a separate model for each sex) using GREML.[2] In your report, describe the main intuition behind GREML and its main differences with the classical twin study. Discuss your findings and use at least 1 scientifically formatted table to report the estimation results in this section. [+/- 500 words]

Hints:

- The individuals in the sample are not closely related, but do make sure that cryptically related pairs of individuals (which may nevertheless be present) are excluded from the analysis (use the --grm-cutoff function with an appropriate threshold).
- For your report, it is not needed to perform a case-control GREML analysis. It is also not needed to perform a bivariate GREML analysis.

[2] Don't forget to include your analysis code in the Appendix of your report.