2021 International Statistical Genetics Workshop

DAY 3. TUTORIAL – PART 2. CHEATSHEET.

Here you will find the main information that can be found in the PLINK websites (as in June 2021) and that will help you navigate the practical.

Extracted from:
https://www.cog-genomics.org/plink/1.9/assoc#linear

Regression with multiple covariates

--linear ['hide-covar'] ['standard-beta']
--logistic ['hide-covar'] ['beta']

Given a quantitative phenotype and possibly some covariates (in a --covar file), **--linear** writes a linear regression report to plink.assoc.linear. Similarly, **--logistic** performs logistic regression given a case/control phenotype and some covariates. If either flag is used with --all-pheno, the type of regression will automatically adapt based on whether the current phenotype is case/control or not.

- '**hide-covar**' removes covariate-specific lines from the main report.
- For logistic regressions, the '**beta**' modifier causes regression coefficients instead of odds ratios to be reported.
- With --linear, the '**standard-beta**' modifier standardizes the phenotype and all predictors to zero mean and unit variance before regression. (This happens separately for each variant, since different samples can have missing genotypes for different variants.)

Extracted from:
https://www.cog-genomics.org/plink/1.9/formats#assoc_linear

.assoc.linear, .assoc.logistic (multi-covariate association analysis report)

Produced by --linear/--logistic.

A text file with a header line, and **T** lines per variant typically with the following nine fields (where **T** is normally the number of terms, but the 'genotypic' and 'hethom' modifiers and the --tests flag can change this):

CHR	Chromosome code. *Not present with 'no-snp' modifier.*
SNP	Variant identifier. *Not present with 'no-snp'.*
BP	Base-pair coordinate. *Not present with 'no-snp'.*
A1	Allele 1 (usually minor). *Not present with 'no-snp'.*

TEST	Test identifier
NMISS	Number of observations (nonmissing genotype, phenotype, and covariates)
'BETA'/'OR'	Regression coefficient (--linear, "--logistic beta") or odds ratio (--logistic without 'beta')
STAT	T-statistic
P	Asymptotic p-value for t-statistic

If --ci 0.*xy* has also been specified, the following three fields are inserted before 'STAT':

SE	Standard error of beta (log-odds) estimate
L*xy*	Bottom of *xy*% symmetric approx. confidence interval
H*xy*	Top of *xy*% approx. confidence interval

Refer to the PLINK 1.07 documentation for more details.

Extracted from: https://www.cog-genomics.org/plink/1.9/basic_stats#freq

Allele frequency

--freq [{counts | case-control}] ['gz']

By itself, **--freq** writes a minor allele frequency report to plink.frq. If you add the '**counts**' modifier, an allele count report is written to plink.frq.count instead. Alternatively, you can use --freq with --within/--family to write a cluster-stratified frequency report to plink.frq.strat, or use the '**case-control**' modifier to write a case/control phenotype-stratified report to plink.frq.cc.

Extracted from: https://www.cog-genomics.org/plink/1.9/formats#frq

.frq (basic allele frequency report)

Produced by --freq. Valid input for --read-freq.

A text file with a header line, and then one line per variant with the following six fields:

CHR	Chromosome code
SNP	Variant identifier
A1	Allele 1 (usually minor)
A2	Allele 2 (usually major)
MAF	Allele 1 frequency
NCHROBS	Number of allele observations

Extracted from: https://www.cog-genomics.org/plink/1.9/formats#frq_cc

.frq.cc (case/control phenotype-stratified allele frequency report)

Produced by "--freq case-control". *Not* valid input for --read-freq.

A text file with a header line, and then one line per variant with the following eight fields:

CHR	Chromosome code
SNP	Variant identifier
A1	Allele 1 (usually minor)
A2	Allele 2 (usually major)
MAF_A	Allele 1 frequency in cases
MAF_U	Allele 1 frequency in controls
NCHROBS_A	Number of case allele observations
NCHROBS_U	Number of control allele observations

Extracted from: https://www.cog-genomics.org/plink/1.9/input#pheno

Phenotypes

Loading from an alternate phenotype file

--pheno <filename>

--mpheno <n>
--pheno-name <column name>
--all-pheno

--pheno causes phenotype values to be read from the 3rd column of the specified space- or tab-delimited file, instead of the .fam or .ped file. The first and second columns of that file must contain family and within-family IDs, respectively.

In combination with --pheno, **--mpheno** lets you use the (**n**+2)th column instead of the 3rd column, while **--pheno-name** lets you select a column by title. (In order to use --pheno-name, there must be a header row with first two entries 'FID' and 'IID'.)

Phenotype encoding

--1

Case/control phenotypes are expected to be encoded as 1=unaffected (control), 2=affected (case); 0 is accepted as an alternate missing value encoding. If you use the **--1** flag, 0 is interpreted as unaffected status instead, while 1 maps to affected. *This also forces phenotypes to be interpreted as case/control.*

Extracted from: https://www.cog-genomics.org/plink/1.9/input#covar

Covariates

--covar <filename> ['keep-pheno-on-missing-cov']

--covar-name <column ID(s)/range(s)...>
--covar-number <column number(s)/range(s)...>

--covar designates the file to load covariates from. The file format is the same as for --pheno (optional header line, FID and IID in first two columns, covariates in remaining columns). By default, the main phenotype is set to missing if any covariate is missing; you can disable this with the '**keep-pheno-on-missing-cov**' modifier.

--covar-name lets you specify a subset of covariates to load, by column name; separate multiple column names with spaces or commas, and use dashes to designate ranges. (Spaces are not permitted immediately before or after a range-denoting dash.) **--covar-number** lets you use column numbers instead.

For example, if the first row of the covariate file is

FID IID SITE AGE DOB BMI ETH SMOKE STATUS ALC

then the following two expressions have the same effect:

--covar-name AGE, BMI-SMOKE, ALC
--covar-number 2, 4-6, 8

Extracted from: https://www.cog-genomics.org/plink/1.9/assoc#misc

Miscellaneous options

--ci <confidence interval size>

For --model and case/control --assoc, '**--ci** *X*' causes size-*X* centered confidence intervals to be reported for odds ratios. (E.g. "--ci 0.95" corresponds to a 95% confidence interval.)