X-LDSC (v0.1-beta) User Manual

Huwenbo Shi

August 14, 2019

Contents

1 Introduction

X-LDSC (trans-ethnic stratified LD score regression) is a method to estimate squared trans-ethnic genetic correlation of a functional category C, $r_g^2(C)$, and its enrichment relative to squared genome-wide trans-ethnic genetic correlation, $r_g^2(C)/r_g^2$. Here, all r_g^2 refers to per-allele effect correlation.

2 Procedures

Running X-LDSC requires two steps – estimating LD scores and performing linear regression. This section provides sample scripts for running X-LDSC. For details on file format, please refer to section 3.

2.1 Step 1 – Estimating LD scores

2.1.1 Input

The following script estimates baseline annotation LD scores for chromosome 1 SNPs.

```
python <software directory>/xldsc.py \
    --score allelic \
    --ld-wind-cm 1.0 \
    --print-snps <a list of SNPs to print> \
    --bfile <EAS reference panel directory>/1000G.EAS.1 \
            <EUR reference panel directory>/1000G.EUR.1 \
    --annot <annotation directory>/baseline.1.annot.gz \
    --out <output directory>/EAS_EUR_baseline_chr1
```

Here are the meanings of the flags.

- `--score` specifies the type of LD scores to be estimated. Here, one should always put "allelic" as the score type for estimating per-allele effect correlation. Although "standardized" score type is also supported, we do not recommend estimating this quantity.
- `--ld-wind-cm` specifies the maximum window size in centimorgan for estimating LD

scores. The default and recommended value is 1.0.

- `--print-snps` specifies the file that contains a list of SNPs in plain text without header for which the LD scores are to be printed. We recommend to print all SNPs with minor allele frequency greater than 1% in both populations.
- `--bfile` takes two argument – reference panel for population 1, and reference panel for population 2. All reference panels should be in PLINK format, and have the same set of SNPs.
- `--annot` specifies the annotation file.
- `--out` specifies the prefix of output files.

2.1.2 Output

After executing this command, 4 files will be created under `<output directory>`.

1. `EAS_EUR_chr1.log` a log file that logs information of intermediate steps.
2. `EAS_EUR_chr1_pop1.gz` baseline annotation LD scores for population 1, in this case, East Asians.
3. `EAS_EUR_chr1_pop2.gz` baseline annotation LD scores for population 2, in this case, Europeans.
4. `EAS_EUR_chr1_te.gz` baseline annotation trans-ethnic LD scores.

2.2 Step 2 – Estimating $r_g^2(C)$ and its enrichment

2.2.1 Input

The following script estimates coefficient for baseline and 3 LD-related continuous annotations, and $r_g^2(C)$ and its enrichment for baseline annotations. An extra step is needed for estimating $r_g^2(C)$ and its enrichment for quintiles of continuous annotations (see sub-section 2.2.3).

```
python <software directory>/xldsc.py \
    --gcor <summary stats directory for EAS>/EAS_sumstats.gz \
           <summary stats directory for EUR>/EUR_sumstats.gz \
    --ref-ld-chr <baseline LD score directory>/EAS_EUR_baseline_chr \
```

```
                <LLD LD score directory>/EAS_EUR_avglld_chr \
                <BStat LD score directory>/EAS_EUR_bstat_chr \
                <allele age LD score directory>/EAS_EUR_alleleage_chr \
  --w-ld-chr <regression weight directory>/EAS_EUR_weight_chr \
  --frqfile <EAS MAF directory>/1000G.EAS. \
            <EUR MAF directory>/1000G.EUR. \
  --annot <baseline annotation directory>/baseline. \
          <LLD annotation directory>/avglld. \
          <BStat annotation directory>/bstat. \
          <allele age annotation directory>/alleleage. \
  --save-pseudo-coef \
  --out TRAIT_EAS_EUR.txt
```

Here are the meanings of the flags.

1. `--gcor` specifies the summary stats files. This flag takes 2 arguments – summary stats for population 1 and summary stats for population 2.
2. `--ref-ld-chr` specifies prefix of the LD score files. This flag takes one or more arguments – one may put as many LD score files as one wishes.
3. `--w-ld-chr` specifies prefix of the regression weights. These are standardized LD scores calculated from regression SNPs.
4. `--frqfile` specifies prefix of minor allele frequency files generated from PLINK.
5. `--annot` specifies prefix of the annotation files. This flags also takes one or more arguments. **Note**: the order one specifies the annotation files must be the same as the order one specifies the LD score files. The annotation files must also be the same files that one uses to obtain the LD scores.
6. `--save-pseudo-coef` If this flag is specified, jackknife pseudo values of the coefficients will be saved. This flag is optional. If one needs to estimate $r_g^2(C)$ of continuous annotations, this flag is needed to obtain the standard error or the estimates.
7. `--out` specifies the output file name.

Additionally, users may use the `--apply-shrinkage` flag to adjust the level of shrinkage. This number should be between 0 and 1, with 0.5 set as default.

2.2.2 Output

After executing the above command, 5 files will be generated.

1. `TRAIT_EAS_EUR.txt` output file containing the estimates. See sub-section 3.4 for details on output file format.
2. `TRAIT_EAS_EUR.txt.log` log file containing information of intermediate steps.
3. `TRAIT_EAS_EUR.txt.pseudo_tau1.gz` jackknife pseudo values for τ coefficients for population 1.
4. `TRAIT_EAS_EUR.txt.pseudo_tau2.gz` jackknife pseudo values for τ heritability coefficients for population 2.
5. `TRAIT_EAS_EUR.txt.pseudo_theta.gz` jackknife pseudo values for θ genetic covariance coefficients.

The `TRAIT_EAS_EUR.txt.pseudo_*.gz` files will be used for estimating $r_g^2(C)$ of quintiles of continuous annotations.

2.2.3 Continuous annotations

For estimating $r_g^2(C)$ and its enrichment, one needs to execute an extra command.

```
python <software directory>/cont_annot_gcor.py \
    --coef TRAIT_EAS_EUR.txt \
    --frqfile <EAS MAF directory>/1000G.EAS. \
              <EUR MAF directory>/1000G.EUR. \
    --annot <baseline annotation directory>/baseline. \
            <LLD annotation directory>/avglld. \
            <BStat annotation directory>/bstat. \
            <allele age annotation directory>/alleleage. \
    --names AVGLLD BSTAT ALLELEAGE \
    --nbins 5 \
    --out TRAIT_EAS_EUR_contannot.txt
```

Here are the meaning of the flags.

1. `--coef` specifies the output from the previous step (see sub-section 2.2.2).

2. `--frqfile` specifies prefix of minor allele frequency files generated from PLINK.
3. `--annot` specifies prefix of the annotation files. This flags also takes one or more arguments. **Note**: the order one specifies the annotation files must be the same as the order of annotations in `TRAIT_EAS_EUR.txt`.
4. `--names` specifies the names of the continuous annotations for which one wishes to compute $r_g^2(C)$ and its enrichment at their quintiles.
5. `--nbins` specifies the number of bins to bin the SNPs based on the values of their continuous annotation. The default is 5 (i.e. quintiles).
6. `--out` specifies the output file name.

Additionally, users may use the `--apply-shrinkage` flag to adjust the level of shrinkage. This number should be between 0 and 1, with 0.5 set as default.

After executing the above command, 2 files will be created.

1. `TRAIT_EAS_EUR_contannot.txt` contains estimates of $r_g^2(C)$ and its enrichment. See sub-section 3.4 for details on output format.
2. `TRAIT_EAS_EUR_contannot.txt.log` is the log file that logs information of intermediate steps.

3 File format

Input files to X-LDSC are almost identical as those for S-LDSC, with only minor differences.

3.1 Summary stats file format

The format of summary stats file is identical to that of S-LDSC. Necessary columns are "SNP", "BP", "A1" effect allele, "A2" non-effect allele, "Z", "N" sample size. The file can be either gzip compressed or uncompressed.

3.2 Annotation file format

The format of annotation file is identical to that of S-LDSC with a minor constraint. The first 4 columns are "CHR", "BP", "SNP", "CM", and can be in any order. The first annotation must start from the 5-th column.

3.3 LD score file format

The format of LD score file is also identical to that of S-LDSC with a minor constraint. The first 3 columns are "CHR", "SNP", "BP", and can be in any order. LD scores for the first annotation must start from the 4-th column.

3.4 Output file format

The output files of X-LDSC contain the following columns.

1. `ANNOT` name of the annotations
2. `NSNP` sum of annotation values
3. `STD` standard deviation of the annotation across SNPs
4. `TAU1` heritability coefficient of population 1
5. `TAU1_SE` standard error heritability coefficient of population 1
6. `TAU2` heritability coefficient of population 2
7. `TAU2_SE` standard error heritability coefficient of population 2
8. `THETA` trans-ethnic genetic covariance coefficient
9. `THETA_SE` standard error of trans-ethnic genetic covariance coefficient
10. `HSQ1` heritability in population 1
11. `HSQ1_SE` standard error of heritability in population 1
12. `HSQ2` heritability in population 2
13. `HSQ2_SE` standard error of heritability in population 2
14. `GCOV` trans-ethnic genetic covariance
15. `GCOV_SE` standard error of trans-ethnic genetic covariance
16. `GCOR` trans-ethnic genetic correlation ($r_g(C)$)
17. `GCOR_SE` standard error of trans-ethnic genetic correlation ($r_g(C)$)

18. `GCORSQ` squared trans-ethnic genetic correlation ($r_g^2(C)$)
19. `GCORSQ_SE` standard error of squared trans-ethnic genetic correlation ($r_g^2(C)$)
20. `HSQ1_ENRICHMENT` heritability enrichment in population 1
21. `HSQ1_ENRICHMENT_SE` standard error heritability enrichment in population 1
22. `HSQ2_ENRICHMENT` heritability enrichment in population 2
23. `HSQ2_ENRICHMENT_SE` standard error heritability enrichment in population 2
24. `GCOV_ENRICHMENT` genetic covariance enrichment
25. `GCOV_ENRICHMENT_SE` standard error of genetic covariance enrichment
26. `GCORSQ_ENRICHMENT` squared trans-ethnic genetic correlation enrichment ($r_g^2(C)/r_g^2$)
27. `GCORSQ_ENRICHMENT_SE` standard error of squared trans-ethnic genetic correlation enrichment ($r_g^2(C)/r_g^2$)