

Assignment for *Bioinformatics aspects of aging and rejuvenation* lecture

Zoltán Szarvas¹

lecturer: Csaba Kerepesi²

¹Faculty of Computer Science, Eötvös Loránd University

²HUN-REN Institute for Computer Science and Control

May 2024

Assignment 1

Train epigenetic clock for microarray-based methylation dataset[3]

```

CCGGGTCCGGCGGGGAAGAGCGCCTCAACGCGCAGGGGCCCATCGCGA
GAGGCCAGCGCCCGCGCGTCCAGCCCCAGGCCCGCGCCCTCGCCTTG
GGCTGCTCCTCGCGGCCCTGCACCGCCTCTGCTACTTGTGGACCGCTTC
CTCACCGCCTTCTCCACCCCGCGCGCCAGCCTCCCGCGCGCACGGTGGG
ATTCGGCCAAATAAGGAGAAAGGGCGCGGCCGTACGCGCGCCAGGTGC
GTGGGCGAGACCAGCTCACGCCCTCTCTCCAGCGCGCAAAGGCCCGGCC
ACAGCTGCTTGGCTGCAGTCGAAGCGTAGCCCGAGACAAGGAAGGGCGC
CTTGACCTCGCACTTTTGTTCGGTTCGAACGTTCTGCTCAGTGGTGGTGG
AATGACGACCGCTCTAAATCGATGCGCCTAGGAGTCCATGAATAAGCG
GTACAGGCTCTCGCGCGCGATGCGCAGCTGCGCCCTCGCGCCCT
CGGGGATGCCCCACCCCTCGTGGCGTCCCGCCCGTCCCGCGCAGCGCG
CGCTCGGGGTGCGCTGGCTCTTCGACGCGGCCATGCGCGACTCGAGC
TGACGTGGTTGAGCAGCGGATCGCAGCTTCCCGCACTTCCCAACCCCA
GGCGCTGGTATTAGGTGACCGCACAGGCGCGCCCTTGCGCGCCCGACCT
CGCGGCTTACGATGGGACCGCGTGGCCCGCGACCTTCGGGCGCGCGGG
GGGGAACCTTCCGCTTTCGGGGGCTGCTCTCTCGCGCCCTCGCGCCCT
CGTACACAGCCTGCTCTTGGGTCCAGGGACACTCGCTCGCTCTGAAGG
ACCCCGCTCTTCGCGCGCGCATCGCCTCTTGGCGCGACACTGAAG
GCGACCAACGGGGCGCATCGACTACATCGAGGCGAGTGCCAGTGCC
CGCATCTAGGGCGCTTCGCGCTTGGCGCGCGAGGGCAGCACTGGGGC
TCTGCGGCTCTGCTTGGGGGAGGGCCTTGGGGTGTCTCAGGGGGCGCG
GGACGGCGCGCTGCTTGGGTGCGCGGGAAGGTTGTGAGATTGAGCCC
CGAGGCGCGCGCGCGCTTTCGGGTGCGCGGCTTTCGGGTGCGCGG
AGCCAGGACAGCGGTGACGCGAGTTTCGCGGTCACTGGTCTCCCTGGAG
TGCCCAAGCTGAATCCAGAGGGGCCAGCTGCTTGTCTTCTTCTCTCT
CGAGCTGGATTGAGCGCCTGCCACGAGCCAGGCTTCCCTGGTGAAGA
TACGGAATGCCCAACCCAGGGAAGGGAGGCGCTGAGGCCCTCGGGAGGC
CCAGAGGTTGCCAGAGGAGAACAGAGTGTTCCTGGCGCTTTCCTCTCT
CTAGGGTGTGACAGCCACTTCCCTGGACACTGCCCTGAGGAAGCGCAG
CTTCTGTGGAGCCACACACTGCGAGAGCTTCTCTACCTCTCTCCAC
GAAGCCCTCTGACTTCTGCTGCCAGCGCGGGCAGGGTTTCCCTGAGCGT
CCCCAACCATCACAGCTCAGGCCACCTCGAGAGACTCCCTTTTAGACA
GAAGCCCTGGTGCAGAGCTGCTTTGAGAGTAAGCTGAGGCTGTGAGGT
TCTTACCGCCAGTTACAGATGGGCTGCTCAGCTCAGAGAGAGGGGTGG
TGACTCCCTTAGGAACACACAGCTAGAGTGTGCTCCCTTAAAGACAGAC
CGAGGTCTGCACTTGCAGCTGCAACAGCTCGCGGTAGGTGATGGTGA
ATTCTTAAATGGTGCATGTCACTGGCCTTTACGCTGGGAGCCCAACAGG
TACCCTTGGCACGCGCCAAACCTGGCCCTGGGGATTCCCATGCTGCGC
AGTCACTCTGTGCACTTACCCTGACAGGCTAGACTCCGAGGCTTCTCT
TTTGGCCCTTCCCTGGCCAGGAGCTTGGACTGGGCTCGTGCTCATCGC
AAGCGGGGAAGCTGCCAGGCCCACTCTGTGGGCTCTCTATTCTCTTG
AGTACGGAAGTTAAGGGCTGGGTGGCAGAGGAAGGCCAGGCGCG
GGCACCTGGGCACTTCCCGCACCTTCAAGCGGTGTC
AAGTGAGCTGTGCTGGTTACAGTGGCCTTGGGAGCTCAGAGAGGTTGAG
ACATAGGCTGGGCTCACACAGCCAGGTAAACGCAAGGTGGGGTTGGAGTC
AGGGTCTAGGGTGGCAGCTGCCAAGCTGTGCAACAAAGCTGTTTCTCGC
GGAGGCTGAGGACCAACACCACTTCCCACTCAGGCTGAGCTGGAGATT
CAGAAAGAAGCCTGGAGCCAGGACAGAGGTGGTGGCGTGGATGATCT
GCTGGCAGCTGTGGTGAAGGCTTCAAGGCTTCAAGGCTGCTGGCT
AGGCTCTGGTGGAGTGGACAGGACCTCGCTGTGTGACATGGGATCGAG
CTTACTGTTGTCCAGAGGCTGCTGTGTGGCCAGGCGCACACTTCTCTCT
CCATGCTTCTCCCTCCCAACCCAGGGGCTGGCCTGGAGCACTGCTCT
CTGACGCCAGCCAACTGGGGACCTCACTTCCATCCCAAGGAACCAT
GAAAGCTGCTGTGAGCTGCTGGGCGCGCTGAGGCTGAGGTCTGGAGT
CGGTGAGGCTGTGTGAGCTGACCTCGCTTAAAGGCGAGGAGCTGGCA
CTGTACCTTCTCTCTCTCTGCTGAGTATGATGACAGGCTCC
AGCCAACTCTCCAGCTGGATCCAGGGAAATATCAGCTTGGGCACT
GCAGTGACCAAGGGGCGCGCTGCCACAGGGAACACATCTTCTGTGG
GGTTACGCGCTTCTGGGGCTGGAAGTGCCAAAGCTGGGGCAAGCT
GTGTTTACGCCACTGAACCAATTACACACAGCGGAGAAAGCAGTAA
ACAGCTTTCACC

```

Dataset

Source

- Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates (Gene Expression Omnibus) [5] [6]
- approximately 450k CpGs from **human whole blood**
- Illumina Infinium 450k (microarray)

Details

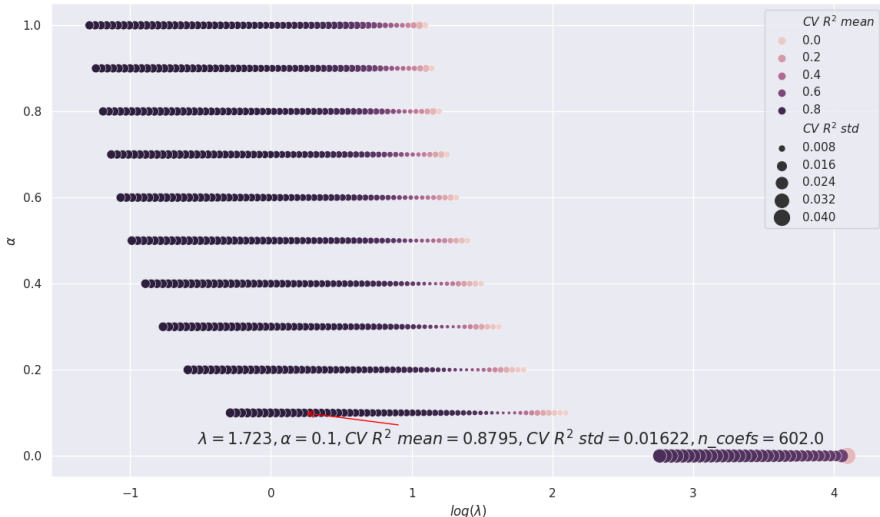
- sample N=656
- features=473 034

Model

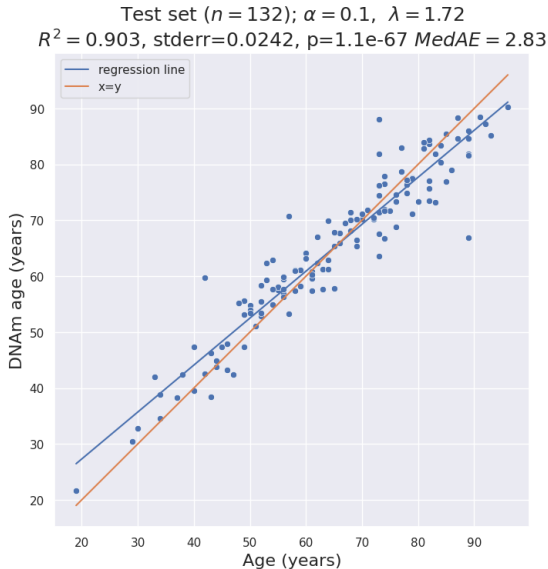
- Elasticnet using Glmnet [1]
- 80-20 train-test split
- 10-fold CV

Hyperparameter Optimization (ElasticNet)

Hyperparameter optimization result on training data
best is selected based on $(CV R^2 \text{ mean}) - 0.5 \cdot (CV R^2 \text{ std})$



Results on Test Set



Assignment 2

Train epigenetic clock for RRBS-based methylation dataset (with missing data)[4]

```
CCCGGGTCCGGCGGGGAAGAGCGGCTCAACGGCAGAGGGGCCATCGCGA
GAGGCCACCGCCCCCGCGCTCCAGCCCCAGGCCCGCGCCCTCGCCCTG
GGCTGCTCCTCGCGGCCCTGCAACGGCTCTGCTGCTACTTGTGGAACGCTTC
CTCAACGCTTCTCCACCCCGCGCGCCGAGCCTCCCGCGCGCACGGTGGG
ATTCGGCCAAATAAGGAGAAAGGGCGCGGCCGTACGCGCGCCAGGTGC
GTGGGCGAGACAGCTCACGCCCTCTCCAGCGCGCAAAGGCCCGGCC
ACAGCTGCTTGGCTGCAGTCAAGAAGCGTAGCCCGAGACAAGGAAGGCGCG
CTTGACTCGCACTTTTGTCCGGTTTCGAACGTTTCTGCTCAGTGGTGCCTGG
AATGACGACGGCTCTAAATCGATGGCGCTAGAGGTCCATGAATAACG
GTACAGGCTCTCGCGCGCGATCGACCGCTGCGCCCTCGCGCCCT
CGGGGATGCCCCACCCCTCGTGGCGTCCCGCGCTCCCGCGAGCGCG
CGCTCGGGTGCCTGGCTCTTCGCAACGGCGCCATGCGCGACTCGAGC
TGAGCTGGTTGAGCAGCGGATCGCGAGCTTCCCGCACTTCCCAACCCCA
GGCGCTGGTATTAGGTGCAACGACAGGCGCGCCCTCGGGCGCCCGACCT
CGCGGCTTACGATGGGACGGCGTGGCCCGCGACCTCGCGGCGCGCGGG
GGGAACCCCTCGCTTTCGGGGGCGCGGCTGCTGCTCTCGCGCCCT
CGTACACAGCCTGCTTCTGGGTCCAGGACACTCGCTCGCTCTGAAGG
ACCCCGCTCTCTTCGCGCGCGCATCGGCTCTGCGCGACACTGAAG
GCGACCCAAGGGGCGCATCGACTACATCGAGGCGAGTGCCAGTGCC
CGCATCTAGGGCGCTTCGCGCTTGGCGCGCGAGAGGCGAGCACTGGGG
TCTGCGGCTCTGCTTGGGGGAGGGCTTTGGGGTGCTTCAGGGGGCGCG
GGACGGCGCGCTGCTTGGGTGGCGCGGGAAGGTTGTGAGATTGAGCCC
CGAGGGCGCGCGAGCTTCGCGGCTCGCGGCTCGCGGCTCGCGGCTCG
AGCCAGGACAGCGTGACGAGTTTCGCGGTCAAGTTGCTTCTCTGGAG
TGCCAAAGCTGAATCCACAGGGGCCAGCTGCTTGTCTTCTTCTCTCT
CGAGCTGGATTAGGCGGCTGCCACGAGCCAGGCGCTTCCCTGGTGAAGA
TCAACGGAATGCCCAAGGGAAGGAGGCGCTGAGGCGCTCGGGAGAGC
CCAAAGAGTGGCCAGGGGAGAACAGAGTGTCTTGGCGCTTCTGCTCTC
CTAGGGTGTGACAGCCACTTCCCTGGACACTGCCCTGAGGAAGCGAG
CTTGTGGAGCACAGCACTGCGAGAGCTCTCTTCTACCTCTCTGCA
GAAGCCCTCTGACCTTCTGCGAGGCGGGGACAGGTTTCCCTGAGCG
CCCCAACCATACAGCTCAGGCCACTCGAGAGACTCCCTTTTAGACA
GAAGCCCTGGTGCAGAGCTGCTTGAAGTAAAGTGAAGGCGCTGACGTT
TCTACAGCCAGTTACAGATGGGCTGCTCAGCTCAGAGAGAGGGGTGG
TGACTCCCTTAGGAACACAGCACTAGAGTGTGCTCCCTTAAAGACAGAC
CGAGTCTGTGACTGTGACTGTGAGACAGCTCGGTAAGTGTGGTGGTGG
ATTCTTAAATGTCATGTCACTGGCTTTTACGCTGGGAGCCCAACAGG
TACCCTTTGCCACGGCCAAACCTTGGCCCTGGGGATTCCCATGCTGCG
AGTCACTCTGTGCTTACCTGACAGGCGTAGACTCGAGGCTTCTCT
TTTGGCCCTTCCCTGGCCAGGAGCTTGGACTGGGCTCGTGCTCATCG
AAGCGGGGAAGCTGCGAGGCCCACTCTGTGGGCTCTCTATTCTCTGG
AGTACGGAAGTTAGAGGCTGGGTTGGCAGAGAGGGCAGAGGCG
GGCACTGTGGCACTTCCCGACCTTCAAGCGGTGTC
AAGTGAGCTGTGCTGGTTACAGTGGCTTGGAGCTCAGAGAGGTTGAG
ACATAGGCTGGGCTCACACAGCCAGGTAAACAGCAAGTGGGGTTGGAGTC
AGGGTCTAGGGTGGCAGCTGCCAAGCTGTGCAACAAAGCTGTTTCTCG
GGAGGCTGAGGACCAACACCACTTCCCACTCAGGCTGAGCTGGAGATT
CAGAAAGAAGCCTGGAGCCAGGACAGAGGTGGTGGCTGGATGATCT
GCTGGCACTGTGAGTGAAGGCTTCAAGGCTGCTGCTGCTGCTGCTGCT
AGGCTCTGTGGAGTGGGACAGGACTCGCTGTGTGACATGGGATCGAG
CTTACTGTTGTCCAGAGGCTGCTGTGTGGCAGGCGACACTTCTCTCTC
CCATGCTTCTCCCTCCCAACCCAGGGGCTGGGCTGGAGCACTGCTCT
CTGACGCCAGGCAACTGGGGACCTCACTTCCATCCCAAGGAACCAT
GAACTGCTGCTGTGAGCTGCTGGGCGCTGCTGAGGCTGAGGTCTGGAGT
CGGTGAGGCTGTGTGAGCTGAGCTCGCTTAAAGGAGGAGGAGCTGGCA
CTGTACCTTCTCTCTCTCTCTGCTGAGTATGATGACAGAGGCTCC
AGCCAACTCTCCAGCTGGATCCAGGGAAATATCAGCTTGGGCACT
GCAGTGACAGGGGCGCGGCTGCCACAGGGAACACATCTTCTGTGG
GGTTACGCGCTTCTGGGGCTGGAAGTGCCAAAGCTGGGGCAAGCT
GTGTTTACGCCACTGAACCAATTACACACAGCGGAGAAACGAGTAA
ACAGCTTTCCAC
```

Dataset

Source

- Data from "Depression and suicide risk prediction models using blood-derived multi-omics data" [2]
- Reduced Representation Bisulfite Sequencing (RRBS)

Details

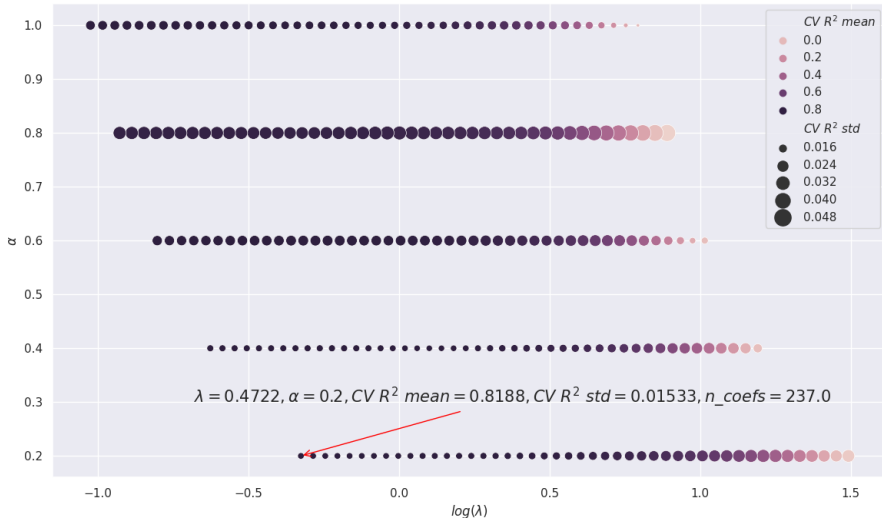
- sample N=182 (Healthy/Control=87, Major Depressive Disorder (MDD)=39, Suicide Attempters (SA)=56)
- features=8 722 096

Model

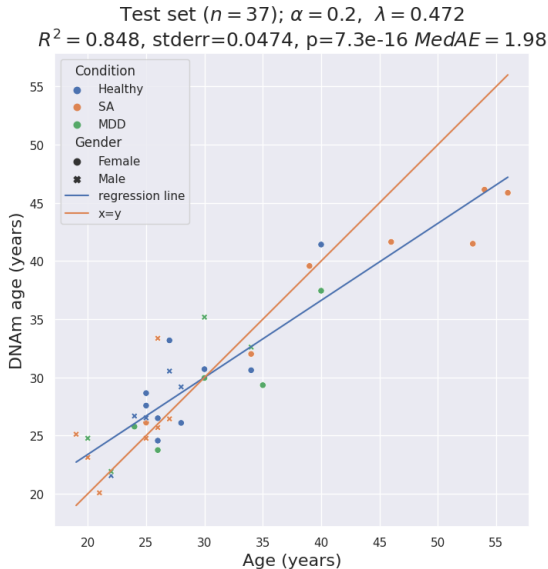
- Elasticnet using Glmnet [1]
- 80-20 train-test split, 10-fold CV
- data for missing CpG sites was filled with **mean**

Hyperparameter Optimization (ElasticNet)

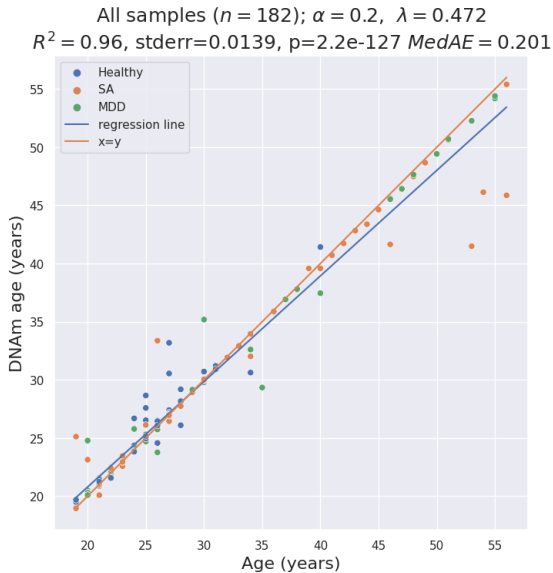
Hyperparameter optimization result on training data
best is selected based on $(CV R^2 \text{ mean}) - 0.5 \cdot (CV R^2 \text{ std})$



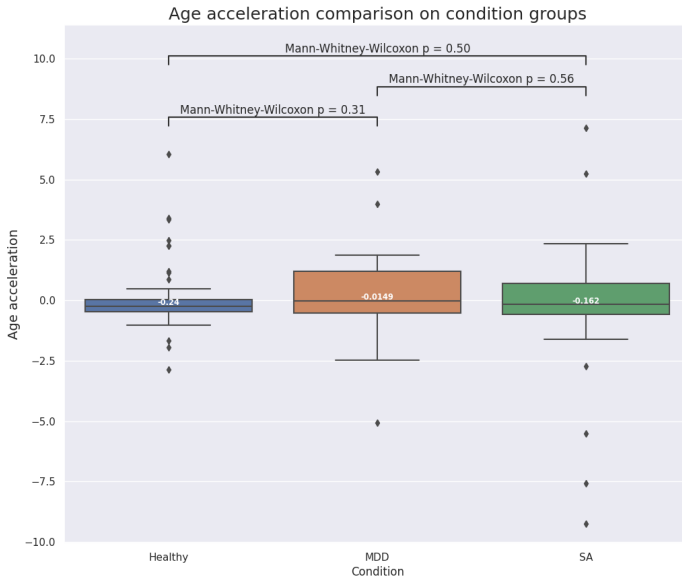
Results on Test Set



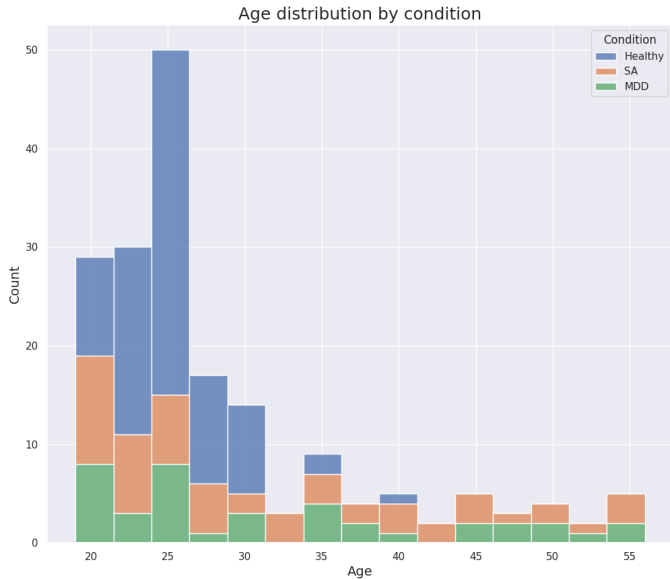
Results for All Instances



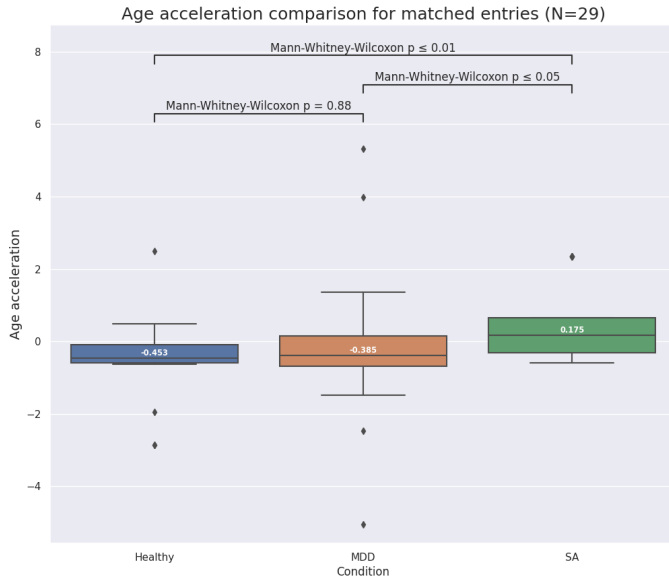
Age acceleration comparison by condition



Age distribution by condition



Age acceleration comparison by condition (controlled)



References I

- [1] *An Introduction to 'glmnet' — glmnet.stanford.edu.*
<https://glmnet.stanford.edu/articles/glmnet.html>.
[Accessed 16-05-2024].
- [2] Youngjune Bhak et al. “Depression and suicide risk prediction models using blood-derived multi-omics data”. *en. In: Transl. Psychiatry* 9.1 (Oct. 2019), p. 262.
- [3] *Epigenetic clock assignment 1.*
https://github.com/szazo/epigenetic-clock/blob/main/assignment1_microarray.ipynb.
2024.
- [4] *Epigenetic clock assignment 2.*
https://github.com/szazo/epigenetic-clock/blob/main/assignment2_rrbs.ipynb. 2024.

References II

- [5] *Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates*. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40279>. [Accessed 13-05-2024]. 2012.
- [6] Gregory Hannum et al. “Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates”. In: *Molecular Cell* 49.2 (Jan. 2013), pp. 359–367. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2012.10.016. URL: <http://dx.doi.org/10.1016/j.molcel.2012.10.016>.