MRS Model Experimentation

MRS model construction and parameter optimization

Basic first models

Basic MRS modeling – each model is trained on a single cohort using elastic net regression on the top 50% most variable CpGs.

Table 1: Trained in WHI

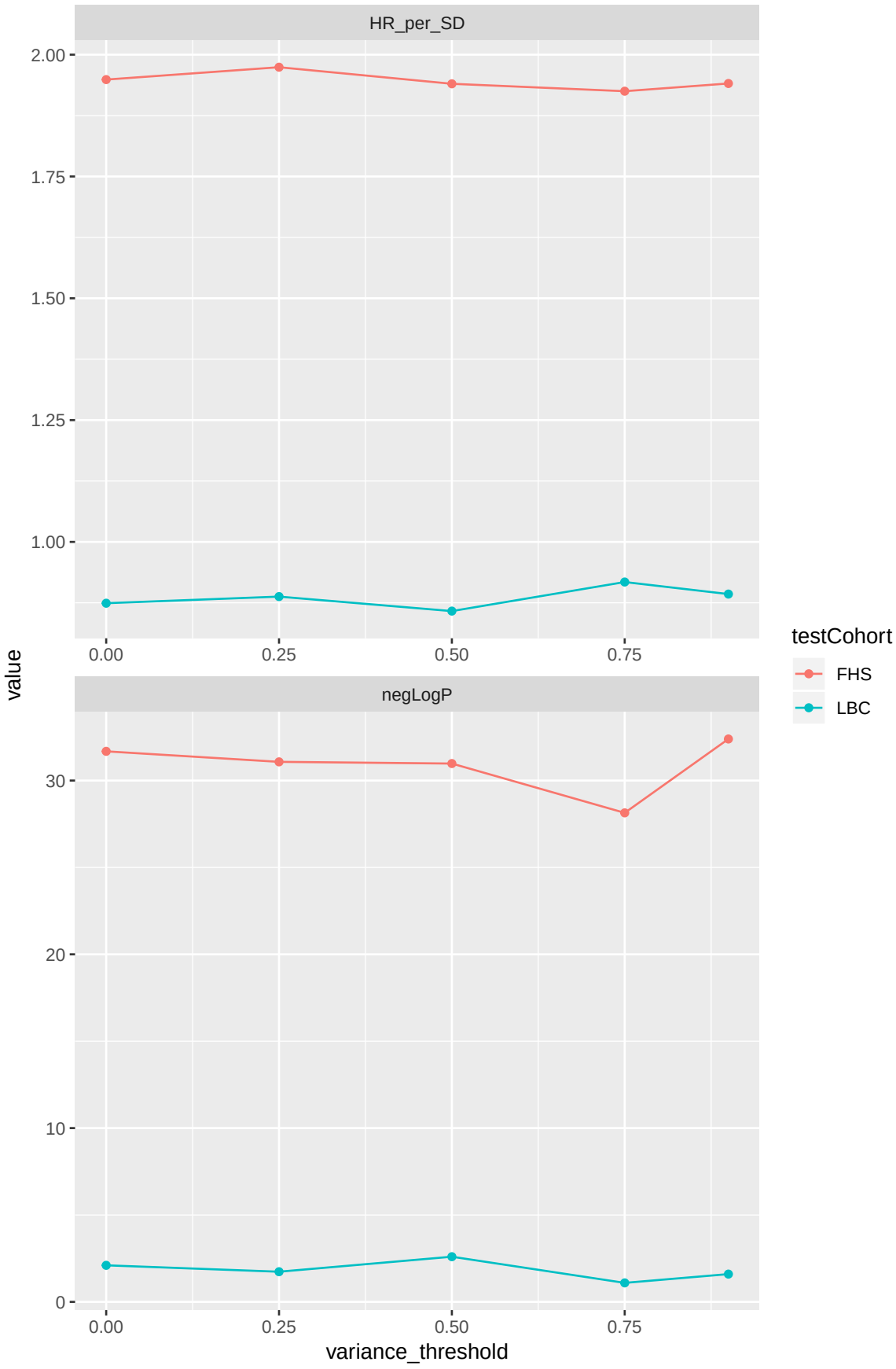
testCohort	HR_per_SD	p
FHS	1.86	0

Table 2: Trained in FHS

testCohort	HR_per_SD	p
WHI	1.2	0

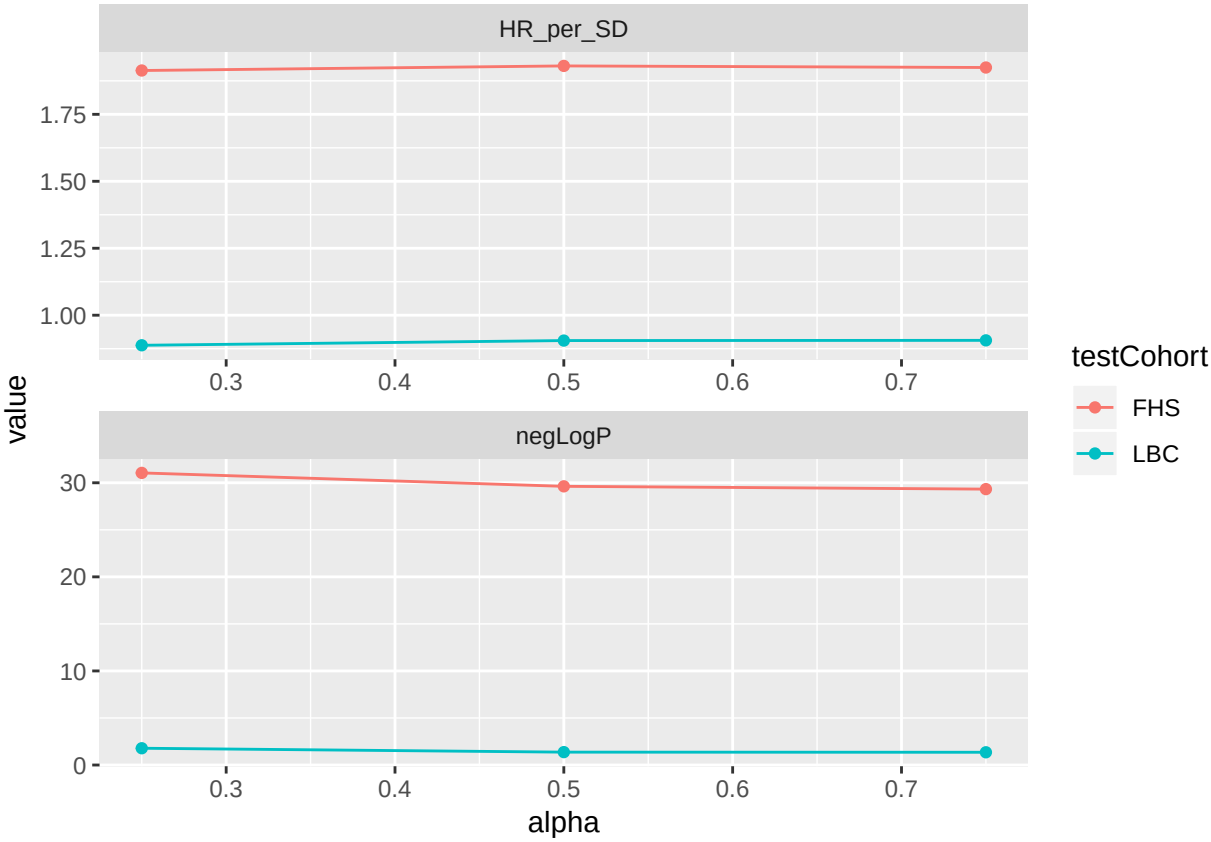
Variance thresholds

Training in WHI, testing in FHS & LBC.


HR_per_SD
negLogP
value
variance_threshold
2.00
1.75
1.50
1.25
1.00
30
20
10
0
0.00
0.25
0.50
0.75
testCohort
FHS
LBC

Manipulation of alpha

Does the elastic net alpha parameter have any notable effect on the results? Try it out using 75th quantile variance threshold.



Covariate exploration

When training in WHI, does adjustment for age or cell counts impact the eventual generalization to other cohorts?

Table 3: Trained in WHI with age adjustment

testCohort	HR_per_SD	p
FHS	1.447	0.000
LBC	0.955	0.354

Table 4: Trained in WHI with cell count adjustment

testCohort	HR_per_SD	p
FHS	1.426	0
LBC	0.827	0

Final multi-cohort methylation model

Trained in WHI & FHS together.

Table 5: Performance in 30% held-out test set

HR_per_SD	p
1.31	0

Table 6: Testing full model in LBC

Test cohort	HR_per_SD	p
LBC (full)	0.801	0.000
LBC21	0.859	0.141
LBC36	1.057	0.337

Tentative conclusions

- A variance threshold of 50% reaches an approximate asymptote in performance while keeping the computational burden under control.
- There is not strong evidence to suggest that a change in alpha from 0.5 will provide notably better performance.
- The LBC datasets are complicated: they are somewhat older, and have non-intuitive relationships between traditional CVD risk factors and CVD prevalence/incidence. At the very least, the two cohorts (1921 and 1936) should be modeled separately.

Genetic risk in the same cohorts

A genetic risk score (GRS) was calculated based on the model used by Khera et al. 2016 – currently, ~48 of the 50 constituent genotypes are available for entry into the model. This GRS was first tested to confirm its associations in these datasets (genotypes available for WHI and FHS). While all CVD cases are incident in WHI, past and incident events were merged into a single binary variable for FHS in order to test the GRS.

Table 7: GRS results

cohort	OR_per_SD	p
WHI	1.14	0.005
FHS	1.19	0.001

MRS performance across demographic and risk-based strata

Also worth checking this in REGICOR.

Across sexes

Table 8: WHI -> FHS sex-stratified

Sex	HR_per_SD	p
Male	1.92	0
Female	1.78	0

Table 9: WHI + FHS -> WHI + FHS sex-stratified (Note: doesn't use an external validation set)

Sex	HR_per_SD	p
Male	2.15	0
Female	1.64	0

Based on race

Table 10: FHS -> WHI race-stratified

Race	HR_per_SD	p
White	1.11	0.016
Black	1.24	0.000
Hispanic	1.31	0.000

Based on history of CVD

Table 11: WHI -> FHS stratified by history of CVD

Past event	HR_per_SD	p
Yes	1.29	0.012
No	1.77	0.000

Across tertiles of genetic risk

Table 12: WHI -> FHS stratified by genetic risk score

Tertile	HR_per_SD	p
1	1.56	0
2	2.33	0
3	2.08	0
4	1.88	0

Table 13: FHS -> WHI stratified by genetic risk score

Tertile	HR_per_SD	p
1	1.10	0.064
2	1.25	0.000

Tertile	HR_per_SD	p
3	1.31	0.000
4	1.08	0.484

MRS stability

```
## [1] "ICC: 0.663188642850793"
```

Validation of MRS with respect to prevalent CVD (MI) in REGICOR

Quite predictive alone and with adjustment for cell counts/technical factors/smoking, marginally significant in the presence of traditional risk factors.

GWIS with MRS

EWIS with GRS

Table 14: EWIS results in WHI

CpG	coef	z	p	fdr
cg03307581	51.48	4.56	0	1
cg06470943	11.21	4.54	0	1
cg10759651	10.01	4.48	0	1
cg07554357	7.99	4.34	0	1
cg20374595	7.21	4.32	0	1
cg05983061	-7.03	-4.27	0	1
cg17258816	8.27	4.24	0	1
cg20800956	27.34	4.21	0	1
cg20230308	118.25	4.21	0	1
cg24050474	-6.92	-4.20	0	1

Table 15: EWIS results in FHS

CpG	coef	z	p	fdr
cg11895489	11.25	4.46	0	0.999
cg04208750	-92.53	-4.41	0	0.999
cg16930811	-6.89	-4.11	0	0.999
cg20361001	-119.23	-4.02	0	0.999
cg01974180	47.63	3.97	0	0.999
cg07922411	5.20	3.97	0	0.999
cg00481227	-5.33	-3.96	0	0.999
cg12424468	-9.16	-3.96	0	0.999
cg17699374	-51.02	-3.94	0	0.999
cg00576886	-12.71	-3.85	0	0.999

Full modeling of genetic plus epigenetic risk

Start with a standard ML model (random forest for example?) and try:

- Methylation only
- Genotypes only
- Risk factors only
- All three together
- Different groups of two depending on the results