Chapter 23
Optimal Scaling of High-Sensitivity Analysis of Health Predictors (250 Patients)

General Purpose

In linear models of health predictors (x-values) and health outcomes (y-values), better power of testing can sometimes be obtained, if continuous predictor variables are converted into the best fit discretized ones.

Specific Scientific Question

Highly expressed genes were used to predict drug efficacy. The example from chap. 22 was used once more. The gene expression levels were scored on a scale of 0–10, but some scores were rarely observed. Can the strength of prediction be improved by optimal scaling.

This chapter was previously published in "Machine learning in medicine-cookbook 1" as Chap. 8, 2013.

T.J. Cleophas, A.H. Zwinderman, *Machine Learning in Medicine - a Complete Overview*, DOI 10.1007/978-3-319-15195-3_23

G1	G2	G3	G4	G16	G17	G18	G19	G24	G25	G26	G27	O1	O2	O3	O4
8	8	9	5	7	10	5	6	9	9	6	6	6	7	6	7
9	9	10	9	8	8	7	8	8	9	8	8	8	7	8	7
9	8	8	8	8	9	7	8	9	8	9	9	9	8	8	8
8	9	8	9	6	7	6	4	6	6	5	5	7	7	7	6
10	10	8	10	9	10	10	8	8	9	9	9	8	8	8	7
7	8	8	8	8	7	6	5	7	8	8	7	7	6	6	7
5	5	5	5	5	6	4	5	5	6	6	5	6	5	6	4
9	9	9	9	8	8	8	8	9	8	3	8	8	8	8	8
9	8	9	8	9	8	7	7	7	7	5	8	8	7	6	6
10	10	10	10	10	10	10	10	10	8	8	10	10	10	9	10
2	2	8	5	7	8	8	8	9	3	9	8	7	7	7	6
7	8	8	7	8	6	6	7	8	8	8	7	8	7	8	8
8	9	9	8	10	8	8	7	8	8	9	9	7	7	8	8

Var G1-27 highly expressed genes estimated from their arrays' normalized ratios
Var O1-4 drug efficacy scores (sum of the scores is used as outcome)

Only the data from the first 13 patients are shown. The entire data file entitled "optscalingfactorplscanonical" can be downloaded from extra.springer.com.

Traditional Multiple Linear Regression

SPSS 19.0 is used for data analysis. Open the data file and command.

Command:

Analyze….Regression….Linear….Dependent: enter the 12 highly expressed genes….Independent: enter the summary scores of the 4 outcome variables (use Transform and Compute Variable command)….click OK.

Coefficients[a]

Model		Unstandardized coefficients		Standardized coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3,293	1,475		2,232	,027
	Geneone	−,122	,189	−,030	−.646	,519
	Genetwo	,287	,225	,078	1,276	,203
	Genethree	,370	,228	,097	1,625	,105
	Genefour	,063	,196	,014	,321	,748
	Genesixteen	,764	,172	,241	4,450	,000
	Geneseventeen	,835	,198	,221	4,220	,000

(continued)

Coefficients[a]

Model		Unstandardized coefficients		Standardized coefficients		
		B	Std. Error	Beta	t	Sig.
	Geneeighteen	,088	,151	,027	,580	,563
	Genenineteen	,576	,154	,188	3,751	,000
	Genetwentyfour	,403	,146	,154	2,760	,006
	Genetwentyfive	,028	,141	,008	,198	,843
	Genetwentysix	,320	,142	,125	2,250	,025
	Genetwentyseven	−,275	,133	−,092	−2,067	,040

[a] Dependent variable: summaryoutcome

The number of statistically significant p-values (indicated here with Sig.), (<0.10) was 6 out of 12. In order to improve this result the Optimal Scaling program of SPSS is used. Continuous predictor variables are converted into best fit discretized ones.

Optimal Scaling Without Regularization

Command:

Analyze….Regression….Optimal Scaling….Dependent Variable: Var 28 (Define Scale: mark spline ordinal 2.2)….Independent Variables: Var 1, 2, 3, 4, 16, 17, 18, 19, 24, 25, 26, 27 (all of them Define Scale: mark spline ordinal 2.2)….Discretize: Method Grouping….OK.

Coefficients

	Standardized coefficients				
	Beta	Bootstrap (1000) estimate of Std. Error	df	F	Sig.
Geneone	−,109	,110	2	,988	,374
Genetwo	,193	,107	3	3,250	,023
Genethree	−,092	,119	2	,591	,555
Genefour	,113	,074	3	2,318	,077
Genesixteen	,263	,087	4	9,065	,000
Geneseventeen	,301	,114	2	6,935	,001
Geneeighteen	,113	,136	1	,687	,408
Genenineteen	,145	,067	1	4,727	,031
Genetwentyfour	,220	,097	2	5,166	,007
Genetwentyfive	−,039	,094	1	,170	,681
Genetwentysix	,058	,107	2	,293	,746
Genetwentyseven	−.127	,104	2	1,490	,228

Dependent variable: summaryoutcome

There is no intercept anymore and the t-tests have been replaced with F-tests. The optimally scaled model without regularization shows similarly sized effects.

The number of p-values < 0.10 is 6 out of 12. In order to fully benefit from optimal scaling a regularization procedure for the purpose of correcting overdispersion (more spread in the data than compatible with Gaussian data) is desirable. Ridge regression minimizes the b-values such that $b_{ridge} = b\ /(1 + \text{shrinking factor})$. With shrinking factor $= 0$, $b_{ridge} = b$, with ∞, $b_{ridge} = 0$.

Optimal Scaling With Ridge Regression

Command:

Analyze….Regression….Optimal Scaling….Dependent Variable: Var 28 (Define Scale: mark spline ordinal 2.2)….Independent Variables: Var 1, 2, 3, 4, 16, 17, 18, 19, 24, 25, 26, 27 (all of them Define Scale: mark spline ordinal 2.2)….Discretize: Method Grouping, Number categories 7….click Regularization….mark Ridge…. OK.

Coefficients

	Standardized coefficients		df	F	Sig.
	Beta	Bootstrap (1000) estimate of Std. Error			
Geneone	,032	,033	2	,946	,390
Genetwo	,068	,021	3	10,842	,000
Genethree	,051	,030	1	2,963	,087
Genefour	,064	,020	3	10,098	,000
Genesixteen	,139	,024	4	34,114	,000
Geneseventeen	,142	,025	2	31,468	,000
Geneeighteen	,108	,040	2	7,236	,001
Genenineteen	,109	,020	2	30,181	,000
Genetwentyfour	,109	,021	2	27,855	,000
Genetwentyfive	,041	,038	3	1,178	,319
Genetwentysix	,098	,023	2	17,515	,000
Genetwentyseven	–,017	,047	1	,132	,716

Dependent variable: 20–23

The sensitivity of this model is better than the above two methods with 7 p-values < 0.0001, and 9 p-values < 0.10, while the traditional and unregularized Optimal Scaling only produced 6 and 6 p-values < 0.10. Also the lasso regularization model is possible (Var = variable). It shrinks the small b values to 0.

Optimal Scaling With Lasso Regression

Command:

Analyze….Regression….Optimal Scaling….Dependent Variable: Var 28 (Define Scale: mark spline ordinal 2.2)….Independent Variables: Var 1, 2, 3, 4, 16, 17, 18, 19, 24, 25, 26, 27 (all of them Define Scale: mark spline ordinal 2.2)…. Discretize: Method Grouping, Number categories 7….click Regularization…. mark Lasso…. OK.

Coefficients

	Standardized coefficients				
	Beta	Bootstrap (1000) estimate of Std. Error	df	F	Sid.
Geneone	,000	,020	0	,000	
Genetwo	,054	,046	3	1,390	,247
Genethree	,000	,026	0	,000	
Genefour	,011	,036	3	,099	,960
Genesixteen	,182	,084	4	4,684	,001
Geneseventeen	,219	,095	3	5,334	,001
Geneeighteen	,086	,079	2	1,159	,316
Genenineteen	,105	,063	2	2,803	,063
Genetwentyfour	,124	,078	2	2,532	,082
Genetwentyfive	,000	,023	0	,000	
Genetwentysix	,048	,060	2	,647	,525
Genetwentyseven	,000	,022	0	,000	

Dependent variable: 20–23

The b-values of the genes 1, 3, 25 and 27 are now shrunk to zero, and eliminated from the analysis. Lasso is particularly suitable if you are looking for a limited number of predictors and improves prediction accuracy by leaving out weak predictors. Finally, the elastic net method is applied. Like lasso it shrinks the small b-values to 0, but it performs better with many predictor variables.

Optimal Scaling With Elastic Net Regression

Command:

Analyze….Regression….Optimal Scaling….Dependent Variable: Var 28 (Define Scale: mark spline ordinal 2.2)….Independent Variables: Var 1, 2, 3, 4, 16, 17, 18, 19, 24, 25, 26, 27 (all of them Define Scale: mark spline ordinal 2.2)….Discretize: Method Grouping, Number categories 7….click Regularization….mark Elastic Net….OK.

Coefficients

	Standardized coefficients				
	Beta	Bootstrap (1000) estimate of Std. Error	df	F	Sig.
Geneone	,000	,016	0	,000	
Genetwo	,029	,039	3	,553	,647
Genethree	,000	,032	3	,000	1,000
Genefour	,000	,015	0	,000	
Genesixteen	,167	,048	4	12,265	,000
Geneseventeen	,174	,051	3	11,429	,000
Geneeighteen	,105	,055	2	3,598	,029
Genenineteen	,089	,048	3	3,420	,018
Genetwentyfour	,113	,053	2	4,630	,011
Genetwentyfive	,000	,012	0	,000	
Genetwentysix	,062	,046	2	1,786	,170
Genetwentyseven	,000	,018	0	,000	

Dependent variable: 20–23

The results are pretty much the same, as it is with lasso. Elastic net does not provide additional benefit in this example but works better than lasso if the number of predictors is larger than the number of observations.

Conclusion

Optimal scaling of linear regression data provides little benefit due to overdispersion. Regularized optimal scaling using ridge regression provides excellent results. Lasso optimal scaling is suitable if you are looking for a limited number of strong predictors. Elastic net optimal scaling works better than lasso if the number of predictors is large.

Note

More background, theoretical and mathematical information of optimal scaling with or without regularization is available in Machine learning in medicine part one, Chaps. 3 and 4, entitled “Optimal scaling: discretization”, and “Optimal scaling: regularization including ridge, lasso, and elastic net regression”, pp 25–37, and pp 39–53, Springer Heidelberg Germany, 2013, from the same authors.