TUCSON 2019
DISCOVERY SUMMIT
EXPLORING DATA
INSPIRING INNOVATION

Finding Needles in A Haystack: Variable Selection for Models

Karen Copeland, Ph.D.
Boulder Statistics

Abstract

There are many steps to building predictive models. One key step is identifying variables to include in your model. This is particularly challenging when you have an abundance of variables to choose from, many of which are likely not important. Thus, you have needles hiding in a haystack, how can you find the needles? I explore a variable selection process that includes predictor screening followed by generalized regression with lasso fitting followed by one-click bootstrapping.

Objectives

- Robust Variable Selection
- Robust Model Construction

Tools

- Response Screening
- Predictor Screening
- Generalized Regression JMP PRO
- One-Click Bootstrapping JMP PRO

My Challenges

- *Large* number of potential variables.
- *Small* number of observations.
- Measurement reproducibility.

My Reference Frame

- Response Type = Binary (0/1)
- Industry = Medical Diagnostics
- Specific Goal = develop commercially viable diagnostic tests based on multivariate algorithms
- Small initial data sets (patients samples = $$$)

Alternative Methods

- Machine Learning
  - –Genetic Algorithms
  - –SVMs
  - –Bootstrap Forests
  - –Etc. etc.
- Train/Validate/Test Sets

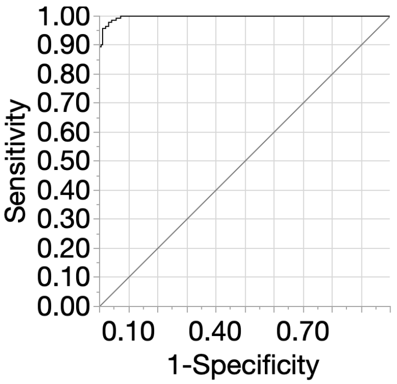
- Struggle = Reality: Small # of observations for initial development work.

Start

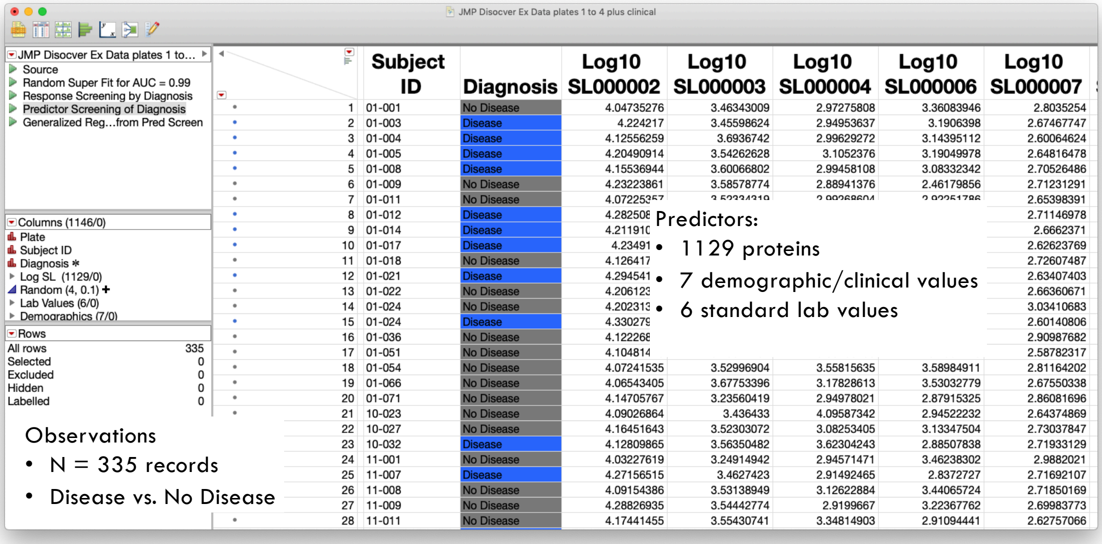
- 335 observations
- 1129+ potential predictors

Ultimate Goal

Generate a Commercially Viable Diagnostic Model



AUC
0.9976


JMP Disocver Ex Data plates 1 to 4 plus clinical
JMP Disocver Ex Data plates 1 to...
Source
Random Super Fit for AUC = 0.99
Response Screening by Diagnosis
Predictor Screening of Diagnosis
Generalized Reg...from Pred Screen
Columns (1146/0)
Plate
Subject ID
Diagnosis ✻
Log SL (1129/0)
Random (4, 0.1) ✚
Lab Values (6/0)
Demographics (7/0)
Rows

Rows	
All rows	335
Selected	0
Excluded	0
Hidden	0
Labelled	0

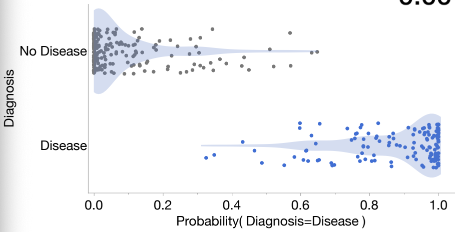
	Subject ID	Diagnosis	Log10 SL000002	Log10 SL000003	Log10 SL000004	Log10 SL000006	Log10 SL000007
1	01-001	No Disease	4.04735276	3.46343009	2.97275808	3.36083946	2.8035254
2	01-003	Disease	4.224217	3.45598624	2.94953637	3.1906398	2.67467747
3	01-004	Disease	4.12556259	3.6936742	2.99629272	3.14395112	2.60064624
4	01-005	Disease	4.20490914	3.54262628	3.1052376	3.19049978	2.64816478
5	01-008	Disease	4.15536944	3.60066802	2.99458108	3.08332342	2.70526486
6	01-009	No Disease	4.23223861	3.58578774	2.88941376	2.46179856	2.71231291
7	01-011	No Disease	4.07225357	3.52334319	2.99268604	2.92251786	2.65398391
8	01-012	Disease	4.282508				2.71146978
9	01-014	Disease	4.211910				2.6662371
10	01-017	Disease	4.23491				2.62623769
11	01-018	No Disease	4.126417				2.72607487
12	01-021	Disease	4.294541				2.63407403
13	01-022	No Disease	4.206123				2.66360671
14	01-024	No Disease	4.202313				3.03410683
15	01-024	Disease	4.330279				2.60140806
16	01-036	No Disease	4.122268				2.90987682
17	01-051	No Disease	4.104814				2.58782317
18	01-054	No Disease	4.07241535	3.52996904	3.55815635	3.58984911	2.81164202
19	01-066	No Disease	4.06543405	3.67753396	3.17828613	3.53032779	2.67550338
20	01-071	No Disease	4.14705767	3.23560419	2.94978021	2.87915325	2.86081696
21	10-023	No Disease	4.09026864	3.436433	4.09587342	2.94522232	2.64374869
22	10-027	No Disease	4.16451643	3.52303072	3.08253405	3.13347504	2.73037847
23	10-032	Disease	4.12809865	3.56350482	3.62304243	2.88507838	2.71933129
24	11-001	No Disease	4.03227619	3.24914942	2.94571471	3.46238302	2.9882021
25	11-007	Disease	4.27156515	3.4627423	2.91492465	2.8372727	2.71692107
26	11-008	No Disease	4.09154386	3.53138949	3.12622884	3.44065724	2.71850169
27	11-009	No Disease	4.28826935	3.54442774	2.9199667	3.22367762	2.69983773
28	11-011	No Disease	4.17441455	3.55430741	3.34814903	2.91094441	2.62757066

Predictors:

- 1129 proteins
- 7 demographic/clinical values
- 6 standard lab values

Observations

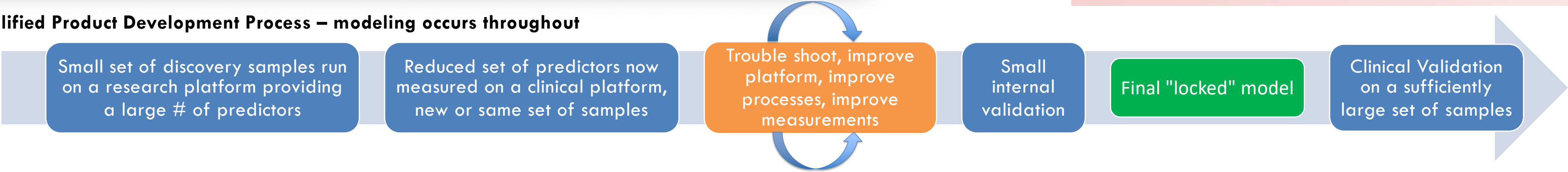
- N = 335 records
- Disease vs. No Disease



Finish

- ***Validated*** Stellar Performance
- Commercially viable product

Simplified Product Development Process – modeling occurs throughout

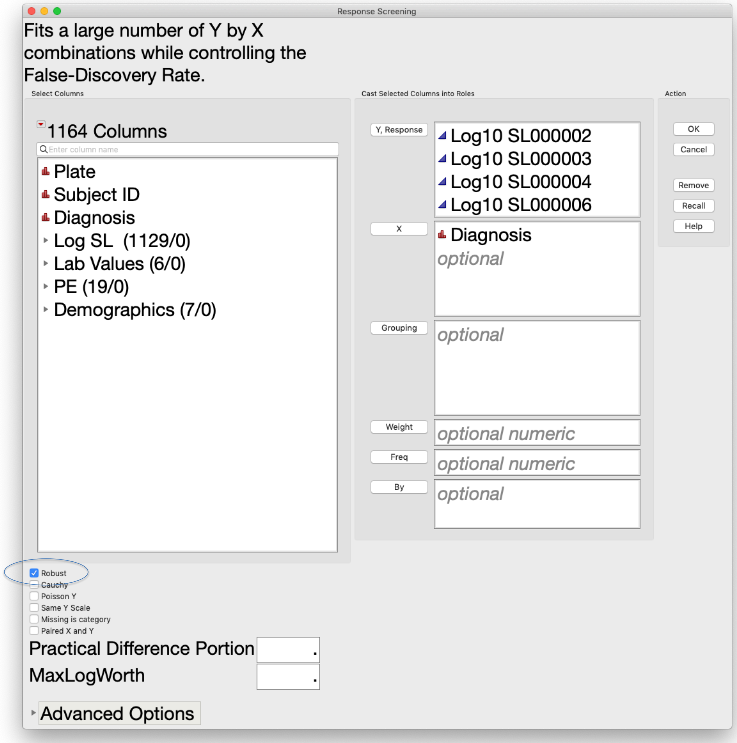

Small set of discovery samples run on a research platform providing a large # of predictors
Reduced set of predictors now measured on a clinical platform, new or same set of samples
Trouble shoot, improve platform, improve processes, improve measurements
Small internal validation
Final "locked" model
Clinical Validation on a sufficiently large set of samples

To not need to spend time and resources in the orange step is like asking for a

for

Tool #1: Response Screening

Analyze > Screening > Response Screening



Response Screening

Fits a large number of Y by X combinations while controlling the False-Discovery Rate.

Select Columns

1164 Columns

Enter column name

- Plate
- Subject ID
- Diagnosis
- Log SL (1129/0)
- Lab Values (6/0)
- PE (19/0)
- Demographics (7/0)

Cast Selected Columns into Roles

Y, Response: Log10 SL000002, Log10 SL000003, Log10 SL000004, Log10 SL000006

X: Diagnosis, optional

Grouping: optional

Weight: optional numeric

Freq: optional numeric

By: optional

☑ Robust
□ Cauchy
□ Poisson Y
□ Same Y Scale
□ Missing is category
□ Paired X and Y

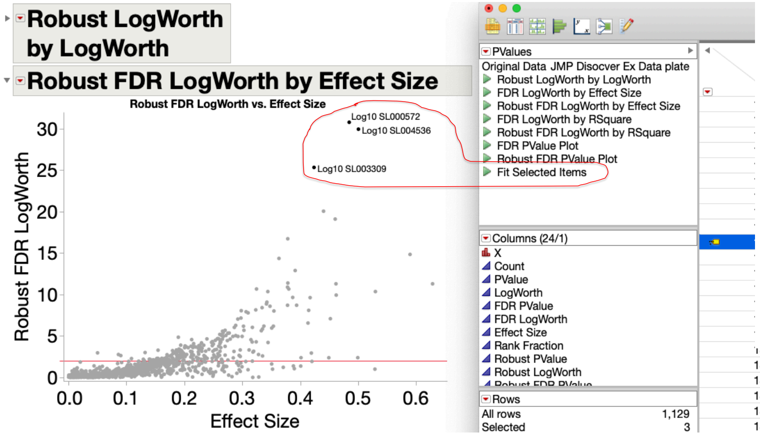
Practical Difference Portion: .

MaxLogWorth: .

Advanced Options

- Quick way to look at 1 at a time comparisons with FDR "protection"
- Use "Fit Selected Items" script in p-value table to look at relationships of interest.

- Does not capture multivariate relationships between variables



Robust LogWorth by LogWorth

Robust FDR LogWorth by Effect Size

PValues

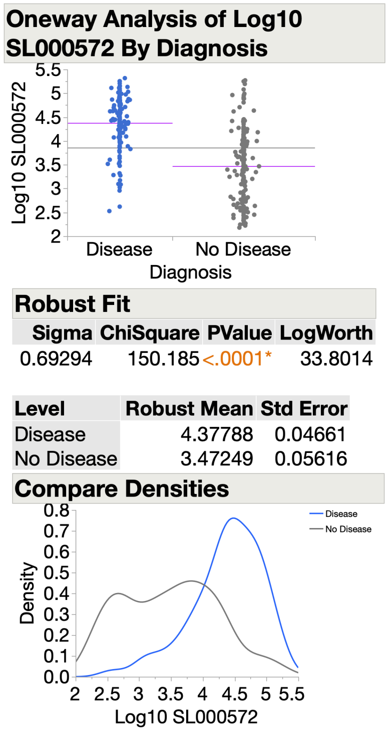
Original Data JMP Disocver Ex Data plate
- Robust LogWorth by LogWorth
- FDR LogWorth by Effect Size
- Robust FDR LogWorth by Effect Size
- FDR LogWorth by RSquare
- Robust FDR LogWorth by RSquare
- FDR PValue Plot
- Robust FDR PValue Plot
- Fit Selected Items

Columns (24/1)
- X
- Count
- PValue
- LogWorth
- FDR PValue
- FDR LogWorth
- Effect Size
- Rank Fraction
- Robust PValue
- Robust LogWorth
- Robust FDR PValue

Rows

```
All rows      1,129
Selected          3
```

```
182 Log10 SLO...  Diagnosis   335
183 Log10 SLO...  Diagnosis   335
184 Log10 SLO...  Diagnosis   335
185 Log10 SLO...  Diagnosis   335
186 Log10 SLO...  Diagnosis   335
187 Log10 SLO...  Diagnosis   335
```



Oneway Analysis of Log10 SL000572 By Diagnosis

Robust Fit

Sigma	ChiSquare	PValue	LogWorth
0.69294	150.185	<.0001*	33.8014

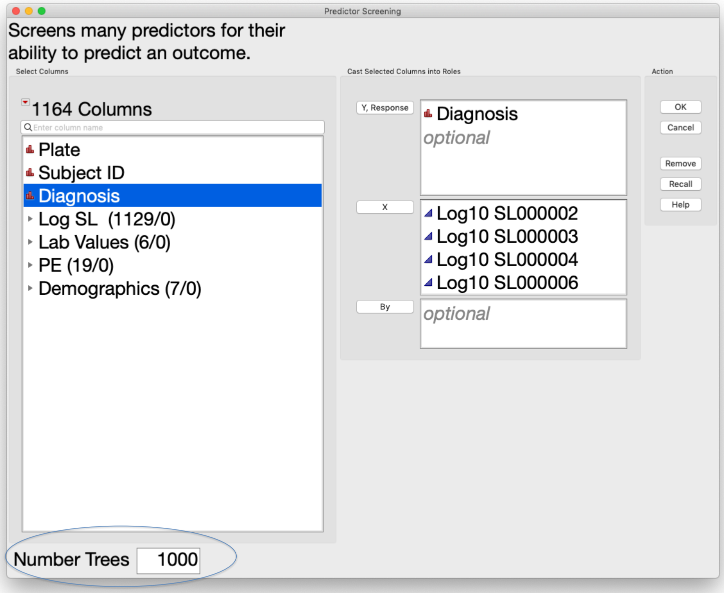
Level	Robust Mean	Std Error
Disease	4.37788	0.04661
No Disease	3.47249	0.05616

Compare Densities

- Answers the question: Do I have a simple winner?

Tool #2: Predictor Screening

Analyze > Screening > Predictor Screening



Predictor Screening

Screens many predictors for their ability to predict an outcome.

Select Columns

1164 Columns

- Plate
- Subject ID
- Diagnosis
- Log SL (1129/0)
- Lab Values (6/0)
- PE (19/0)
- Demographics (7/0)

Cast Selected Columns into Roles

Y, Response: Diagnosis, optional

X: Log10 SL000002, Log10 SL000003, Log10 SL000004, Log10 SL000006

By: optional

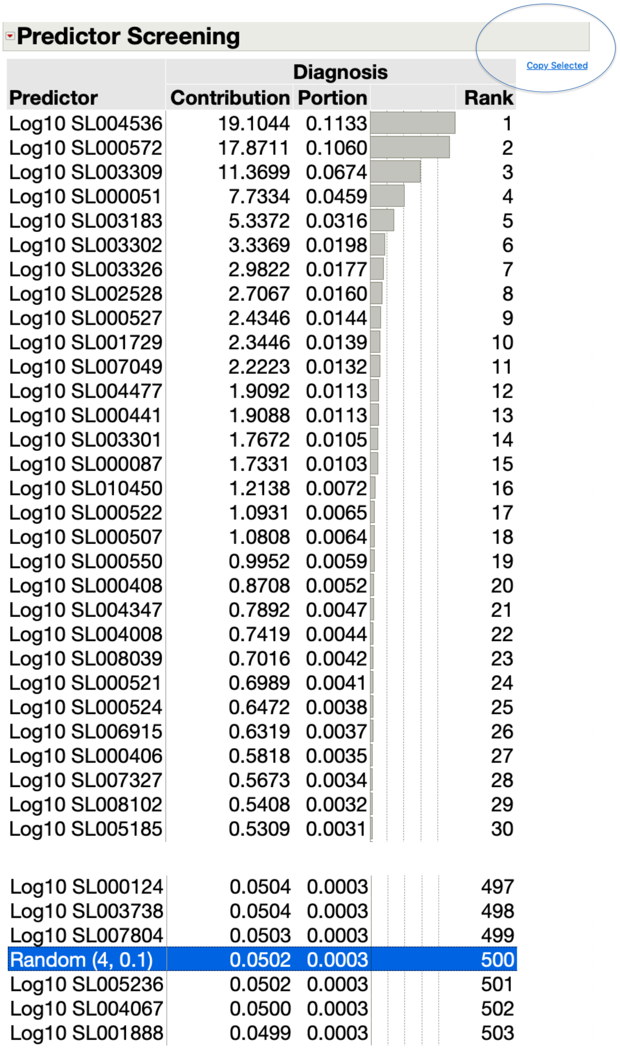
Number Trees 1000

- Uses bootstrap forests to generate a list of "interesting" variables.
- Top three align with Response Screening (this is promising!).
- Begins to capture multivariate relationships between variables.

- Contribution = G^2 (likelihood ratio chi-square)
- Portion = Contribution/∑Contributions
- Bigger is better

- Random components to bootstrap forests, so re-running can lead to a different ranking. Use a large number of trees to improve robustness ranking.
- Interpretation is "relative".

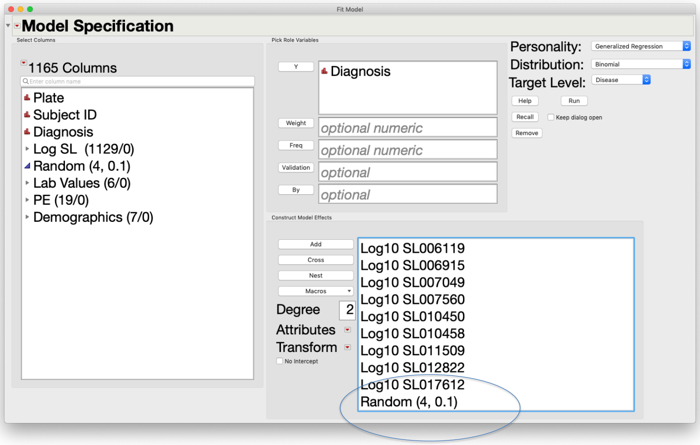
- I often use one or more "fake" variables that I know are random (as I used a random number generator to create them) as a way to judge noise vs. maybe not noise.



Predictor Screening

	Diagnosis		
Predictor	Contribution	Portion	Rank
Log10 SL004536	19.1044	0.1133	1
Log10 SL000572	17.8711	0.1060	2
Log10 SL003309	11.3699	0.0674	3
Log10 SL000051	7.7334	0.0459	4
Log10 SL003183	5.3372	0.0316	5
Log10 SL003302	3.3369	0.0198	6
Log10 SL003326	2.9822	0.0177	7
Log10 SL002528	2.7067	0.0160	8
Log10 SL000527	2.4346	0.0144	9
Log10 SL001729	2.3446	0.0139	10
Log10 SL007049	2.2223	0.0132	11
Log10 SL004477	1.9092	0.0113	12
Log10 SL000441	1.9088	0.0113	13
Log10 SL003301	1.7672	0.0105	14
Log10 SL000087	1.7331	0.0103	15
Log10 SL010450	1.2138	0.0072	16
Log10 SL000522	1.0931	0.0065	17
Log10 SL000507	1.0808	0.0064	18
Log10 SL000550	0.9952	0.0059	19
Log10 SL000408	0.8708	0.0052	20
Log10 SL004347	0.7892	0.0047	21
Log10 SL004008	0.7419	0.0044	22
Log10 SL008039	0.7016	0.0042	23
Log10 SL000521	0.6989	0.0041	24
Log10 SL000524	0.6472	0.0038	25
Log10 SL006915	0.6319	0.0037	26
Log10 SL000406	0.5818	0.0035	27
Log10 SL007327	0.5673	0.0034	28
Log10 SL008102	0.5408	0.0032	29
Log10 SL005185	0.5309	0.0031	30
Log10 SL000124	0.0504	0.0003	497
Log10 SL003738	0.0504	0.0003	498
Log10 SL007804	0.0503	0.0003	499
Random (4, 0.1)	0.0502	0.0003	500
Log10 SL005236	0.0502	0.0003	501
Log10 SL004067	0.0500	0.0003	502
Log10 SL001888	0.0499	0.0003	503

Tool #3: GenReg JMP PRO
Analyze > Fit Model



Fit Model

Model Specification

Select Columns

1165 Columns

Enter column name

- Plate
- Subject ID
- Diagnosis
- Log SL (1129/0)
- Random (4, 0.1)
- Lab Values (6/0)
- PE (19/0)
- Demographics (7/0)

Pick Role Variables

Y: Diagnosis

Weight: optional numeric

Freq: optional numeric

Validation: optional

By: optional

Personality: Generalized Regression

Distribution: Binomial

Target Level: Disease

Construct Model Effects

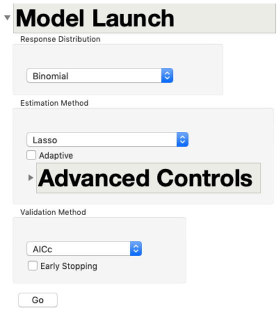
Degree 2

□ No Intercept

- Log10 SL006119
- Log10 SL006915
- Log10 SL007049
- Log10 SL007560
- Log10 SL010450
- Log10 SL010458
- Log10 SL011509
- Log10 SL012822
- Log10 SL017612
- Random (4, 0.1)

- I typically start with a sub-set of variables based on the predictor screening step. Here I used the top 50 variables as my candidate set.
- Use the copy selected from the predictor screening results window to paste variables into the model launch window.
- Run Genreg and then I start with the default model launch (Lasso with AIC validation method).
- This uses all data, no cross-validation, or other checks for model over specification.

- Include one or more "fake" variables as a way to judge noise vs. maybe not noise.



Model Launch

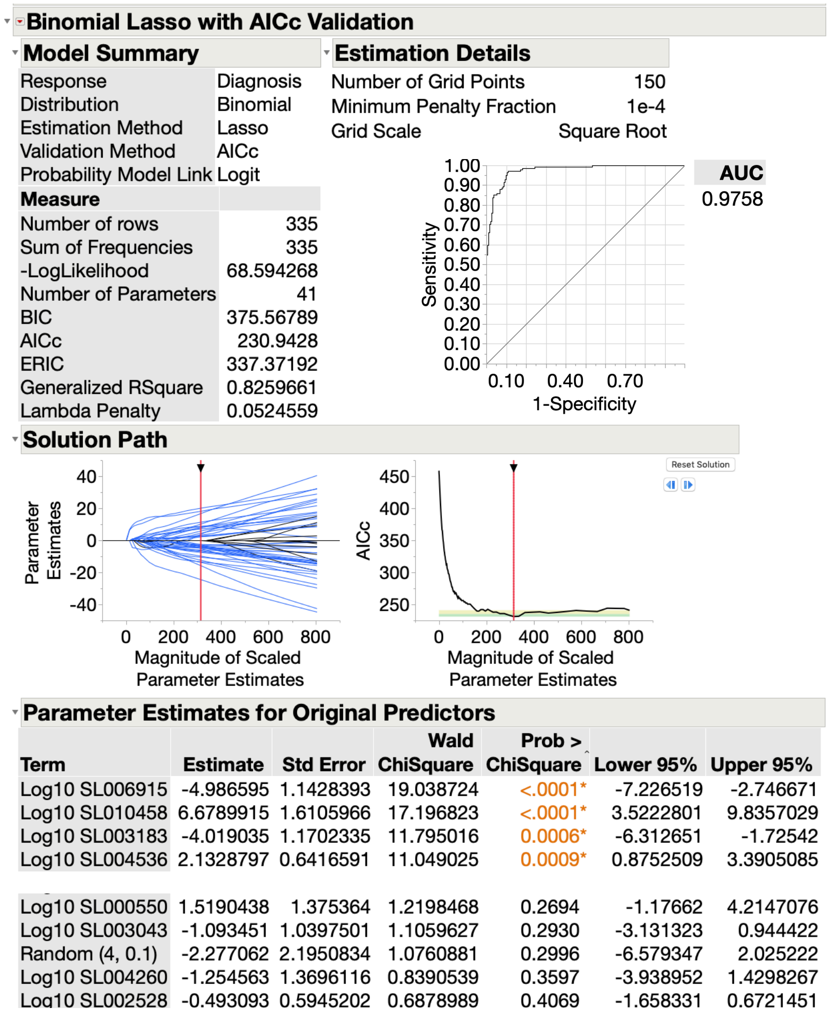
Response Distribution: Binomial

Estimation Method: Lasso

□ Adaptive

Validation Method: AICc

□ Early Stopping



Binomial Lasso with AIC Validation

Model Summary

Response	Diagnosis
Distribution	Binomial
Estimation Method	Lasso
Validation Method	AICc
Probability Model Link	Logit

Measure	
Number of rows	335
Sum of Frequencies	335
-LogLikelihood	68.594268
Number of Parameters	41
BIC	375.56789
AICc	230.9428
ERIC	337.37192
Generalized RSquare	0.8259661
Lambda Penalty	0.0524559

Estimation Details

Number of Grid Points	150
Minimum Penalty Fraction	1e-4
Grid Scale	Square Root

AUC
0.9758

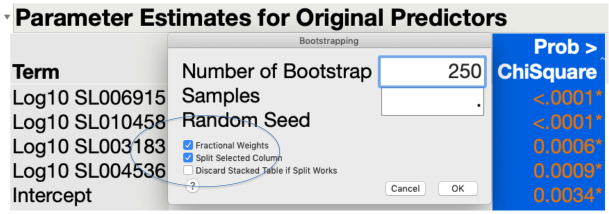
Solution Path

Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Log10 SL006915	-4.986595	1.1428393	19.038724	<.0001*	-7.226519	-2.746671
Log10 SL010458	6.6789915	1.6105966	17.196823	<.0001*	3.5222801	9.8357029
Log10 SL003183	-4.019035	1.1702335	11.795016	0.0006*	-6.312651	-1.72542
Log10 SL004536	2.1328797	0.6416591	11.049025	0.0009*	0.8752509	3.3905085
Log10 SL000550	1.5190438	1.375364	1.2198468	0.2694	-1.17662	4.2147076
Log10 SL003043	-1.093451	1.0397501	1.1059627	0.2930	-3.131323	0.944422
Random (4, 0.1)	-2.277062	2.1950834	1.0760881	0.2996	-6.579347	2.025222
Log10 SL004260	-1.254563	1.3696116	0.8390539	0.3597	-3.938952	1.4298267
Log10 SL002528	-0.493093	0.5945202	0.6878989	0.4069	-1.658331	0.6721451

Tool #4: One-Click Bootstrap
Right Click on Statistic of Interest

- Column > Column Viewer
- Select *Show Quantiles*
- Order by Median *p*-value



Parameter Estimates for Original Predictors

Term	Prob > ChiSquare
Log10 SL006915	<.0001*
Log10 SL010458	<.0001*
Log10 SL003183	0.0006*
Log10 SL004536	0.0009*
Intercept	0.0034*

Bootstrapping

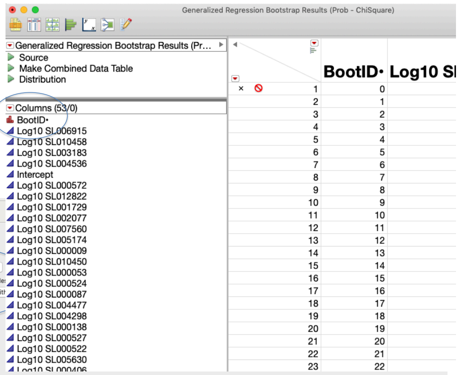
Number of Bootstrap Samples: 250

Random Seed: .

☑ Fractional Weights
☑ Split Selected Column
□ Discard Stacked Table if Split Works

- Run 250 models using bootstrap samples from your data set.
- Evaluate the *p*-values on the estimates to try to separate true signal from lucky signals (i.e., noise).
- The more models an estimate appears in with a small *p*-value the higher your confidence that you may have a true signal.

- Right click on the *p*-value column for bootstrap dialog.
- I use fractional weights.
- Uncheck "Discard Stacked Table if Split Works" if you want to build the visual.
- I use 250 to 500 bootstrap samples.

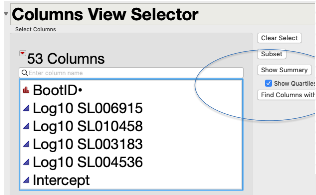


Generalized Regression Bootstrap Results (Prob - ChiSquare)

Generalized Regression Bootstrap Results (Pr…
- Source
- Make Combined Data Table
- Distribution

Columns (53/0)
BootID•
Log10 SL006915
Log10 SL010458
Log10 SL003183
Log10 SL004536
Intercept
Log10 SL000572
Log10 SL012822
Log10 SL001729
Log10 SL002077
Log10 SL007560
Log10 SL005174
Log10 SL000009
Log10 SL010450
Log10 SL000053
Log10 SL000524
Log10 SL000087
Log10 SL004477
Log10 SL004298
Log10 SL000138
Log10 SL000527
Log10 SL000522
Log10 SL005630
Log10 SL000406

	BootID•	Log10 Sl
1	0	
2	1	
3	2	
4	3	
5	4	
6	5	
7	6	
8	7	
9	8	
10	9	
11	10	
12	11	
13	12	
14	13	
15	14	
16	15	
17	16	
18	17	
19	18	
20	19	
21	20	
22	21	
23	22	



Columns View Selector

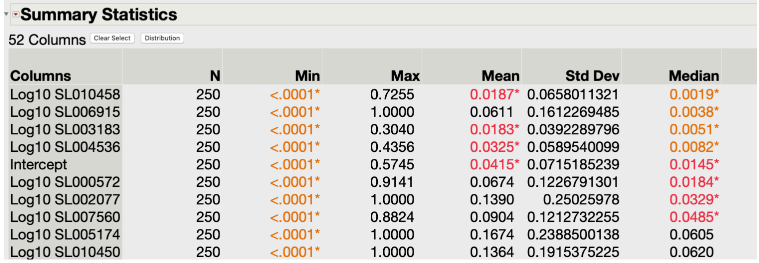
Select Columns

53 Columns

Enter column name

BootID•
Log10 SL006915
Log10 SL010458
Log10 SL003183
Log10 SL004536
Intercept

☑ Show Quartiles

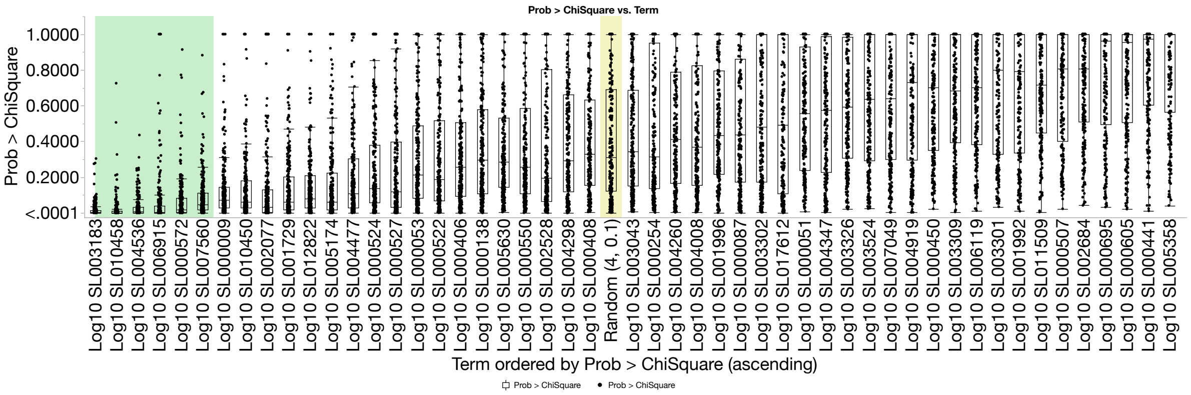


Summary Statistics

52 Columns

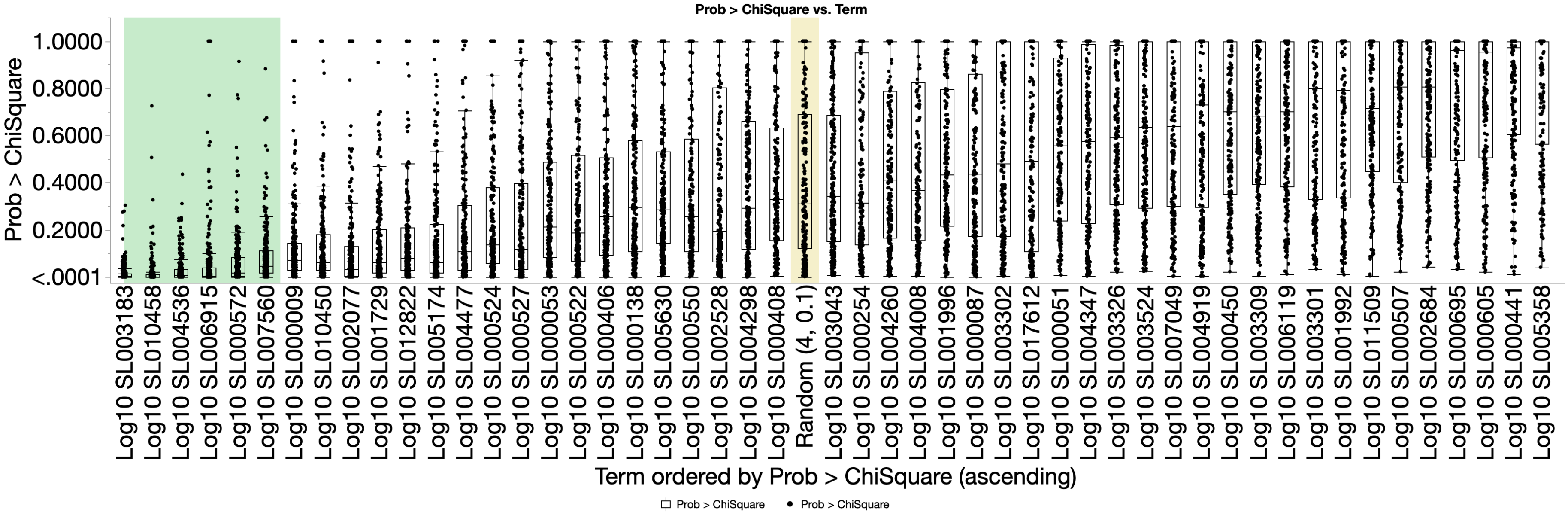
Clear Select
Distribution

Columns	N	Min	Max	Mean	Std Dev	Median
Log10 SL010458	250	<.0001*	0.7255	0.0187*	0.0658011321	0.0019*
Log10 SL006915	250	<.0001*	1.0000	0.0611	0.1612269485	0.0038*
Log10 SL003183	250	<.0001*	0.3040	0.0183*	0.0392289796	0.0051*
Log10 SL004536	250	<.0001*	0.4356	0.0325*	0.0589540099	0.0082*
Intercept	250	<.0001*	0.5745	0.0415*	0.0715185239	0.0145*
Log10 SL000572	250	<.0001*	0.9141	0.0674	0.1226791301	0.0184*
Log10 SL002077	250	<.0001*	1.0000	0.1390	0.25025978	0.0329*
Log10 SL007560	250	<.0001*	0.8824	0.0904	0.1212732255	0.0485*
Log10 SL005174	250	<.0001*	1.0000	0.1674	0.2388500138	0.0605
Log10 SL010450	250	<.0001*	1.0000	0.1364	0.1915375225	0.0620



- Bootstrapping the *p*-values (i.e., building multiple models from variations of the dataset at hand) helps to identify robust variables, variables that appear in your model regardless of the variation of the data used, however, they do not guarantee truth.

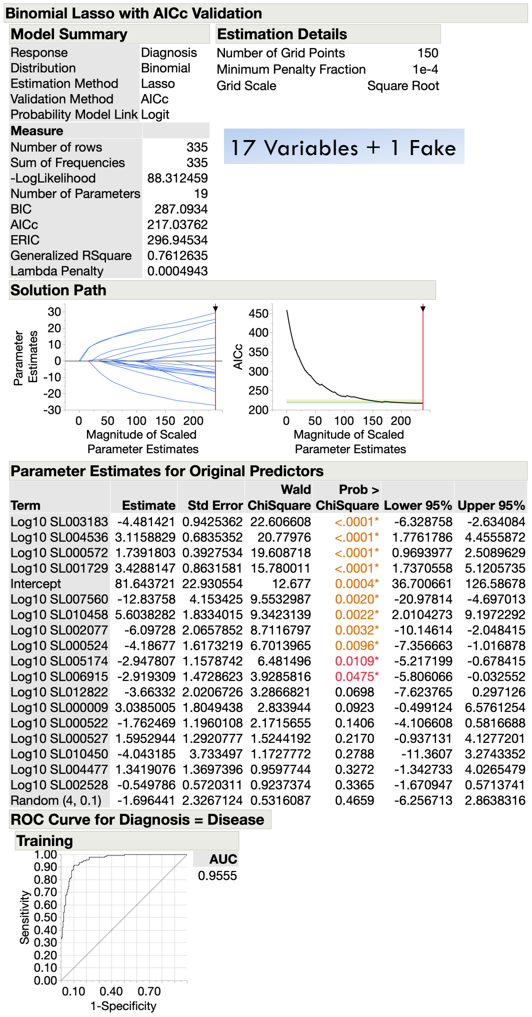
p-values by predictor for 250 models from bootstrap samples



- For each variable there are 250 data points, one p-value for each model.
- Each model is built on a bootstrap sample, that is, each model uses a varied data set.
- Even apparently strong variables have models that they either don't appear in (p-value = 1.0) or are a weak contributor to (0.5 < p-value < 1.0).
- Even apparently weak variables have models in which they are a strong predictor (small p-values).
- The goal is to use variables that are strong in most data set variants so that they generalize to other data sets.

Final Thoughts

- When the signal is big and clear modeling is easy.
- When the signal is multivariate and subtle, modeling can be hard.
- Understanding your measurement systems is important.



Binomial Lasso with AICc Validation

Model Summary

Response	Diagnosis
Distribution	Binomial
Estimation Method	Lasso
Validation Method	AICc
Probability Model Link	Logit

Measure	
Number of rows	335
Sum of Frequencies	335
-LogLikelihood	88.312459
Number of Parameters	19
BIC	287.0934
AICc	217.03762
ERIC	296.94534
Generalized RSquare	0.7612635
Lambda Penalty	0.0004943

Estimation Details

Number of Grid Points	150
Minimum Penalty Fraction	1e-4
Grid Scale	Square Root

17 Variables + 1 Fake

Solution Path

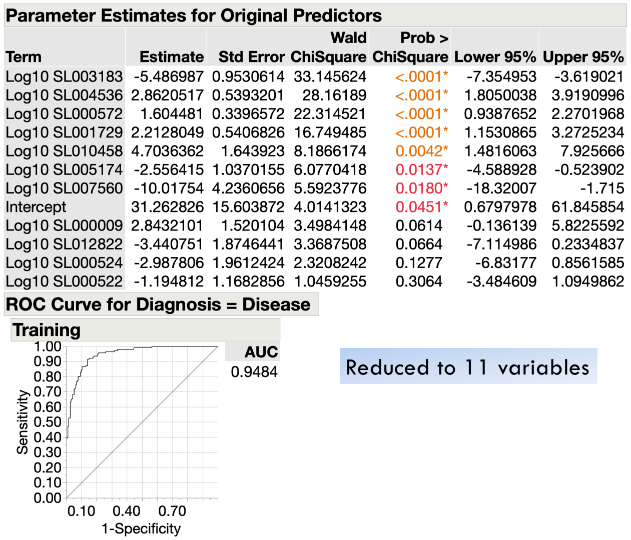
Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Log10 SL003183	-4.481421	0.9425362	22.606608	<.0001*	-6.328758	-2.634084
Log10 SL004536	3.1158829	0.6835352	20.77976	<.0001*	1.7761786	4.4555872
Log10 SL000572	1.7391803	0.3927534	19.608718	<.0001*	0.9693977	2.5089629
Log10 SL001729	3.4288147	0.8631581	15.780011	<.0001*	1.7370558	5.1205735
Intercept	81.643721	22.930554	12.677	0.0004*	36.700661	126.58678
Log10 SL007560	-12.83758	4.153425	9.5532987	0.0020*	-20.97814	-4.697013
Log10 SL010458	5.6038282	1.8334015	9.3423139	0.0022*	2.0104273	9.1972292
Log10 SL002077	-6.09728	2.0657852	8.7116797	0.0032*	-10.14614	-2.048415
Log10 SL000524	-4.18677	1.6173219	6.7013965	0.0096*	-7.356663	-1.016878
Log10 SL005174	-2.947807	1.1578742	6.481496	0.0109*	-5.217199	-0.678415
Log10 SL006915	-2.919309	1.4728623	3.9285816	0.0475*	-5.806066	-0.032552
Log10 SL012822	-3.66332	2.0206726	3.2866821	0.0698	-7.623765	0.297126
Log10 SL000009	3.0385005	1.8049438	2.833944	0.0923	-0.499124	6.5761254
Log10 SL000522	-1.762469	1.1960108	2.1715655	0.1406	-4.106608	0.5816688
Log10 SL000527	1.5952944	1.2920777	1.5244192	0.2170	-0.937131	4.1277201
Log10 SL010450	-4.043185	3.733497	1.1727772	0.2788	-11.3607	3.2743352
Log10 SL004477	1.3419076	1.3697396	0.9597744	0.3272	-1.342733	4.0265479
Log10 SL002528	-0.549786	0.5720311	0.9237374	0.3365	-1.670947	0.5713741
Random (4, 0.1)	-1.696441	2.3267124	0.5316087	0.4659	-6.256713	2.8638316

ROC Curve for Diagnosis = Disease

Training

AUC
0.9555



Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Log10 SL003183	-5.486987	0.9530614	33.145624	<.0001*	-7.354953	-3.619021
Log10 SL004536	2.8620517	0.5393201	28.16189	<.0001*	1.8050038	3.9190996
Log10 SL000572	1.604481	0.3396572	22.314521	<.0001*	0.9387652	2.2701968
Log10 SL001729	2.2128049	0.5406826	16.749485	<.0001*	1.1530865	3.2725234
Log10 SL010458	4.7036362	1.643923	8.1866174	0.0042*	1.4816063	7.925666
Log10 SL005174	-2.556415	1.0370155	6.0770418	0.0137*	-4.588928	-0.523902
Log10 SL007560	-10.01754	4.2360656	5.5923776	0.0180*	-18.32007	-1.715
Intercept	31.262826	15.603872	4.0141323	0.0451*	0.6797978	61.845854
Log10 SL000009	2.8432101	1.520104	3.4984148	0.0614	-0.136139	5.8225592
Log10 SL012822	-3.440751	1.8746441	3.3687508	0.0664	-7.114986	0.2334837
Log10 SL000524	-2.987806	1.9612424	2.3208242	0.1277	-6.83177	0.8561585
Log10 SL000522	-1.194812	1.1682856	1.0459255	0.3064	-3.484609	1.0949862

ROC Curve for Diagnosis = Disease

Training

AUC
0.9484

Reduced to 11 variables

Objectives

- Robust Variable Selection
- Robust Model Construction

Measure Of Success?

- Does the model predict outcomes on a set of data *not* used in the model construction?
- Does the model predict across diverse data sets?
  - –Different patient groups
  - –Different lots of materials
  - –Different labs, instruments