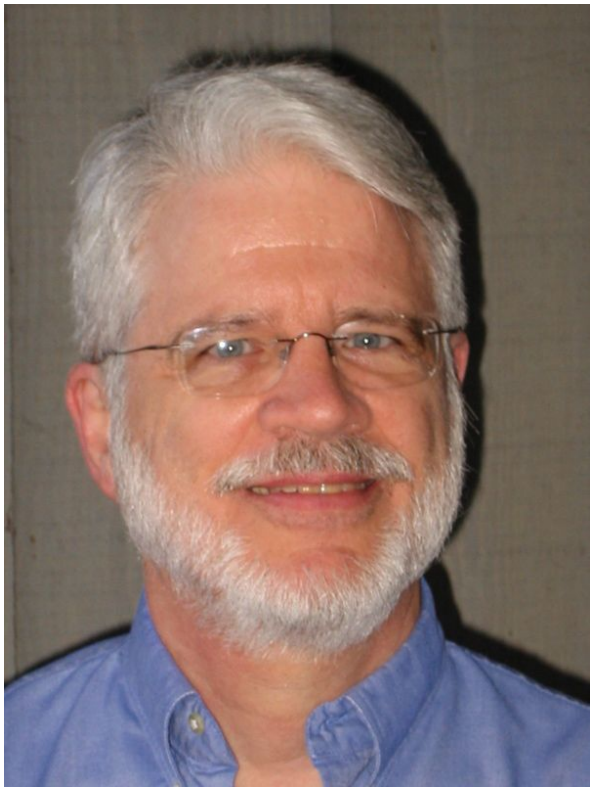
Abstract

Identification and use of patient heterogeneity
S. Stanley Young
National Institute of Statistical Sciences

Medical observational studies present numerous difficulties, not the least of which is that patients can be heterogeneous. In fact patients may be suffering from different diseases, all with the same name! I will look a several methods of clustering in an attempt to separate patients that are meaningfully different and estimate treatment differences within clusters. A large semi-synthetic data set will be used. The end result will be analysis that takes patient heterogeneity into account.

Identification and use of patient heterogeneity



S. Stanley Young

Assistant Director for Bioinformatics
National Institute of Statistical Sciences

NISS

Researcher Incentives and Empirical Methods

by

Edward L. Glaeser

The solution to this problem is not to expect a mass renunciation of data mining, selective data cleaning or opportunistic methodology selection, but rather ...in designing and using techniques that anticipate the behavior of optimizing researchers.

Put indelicately: We need methods to thwart analysis cheating.

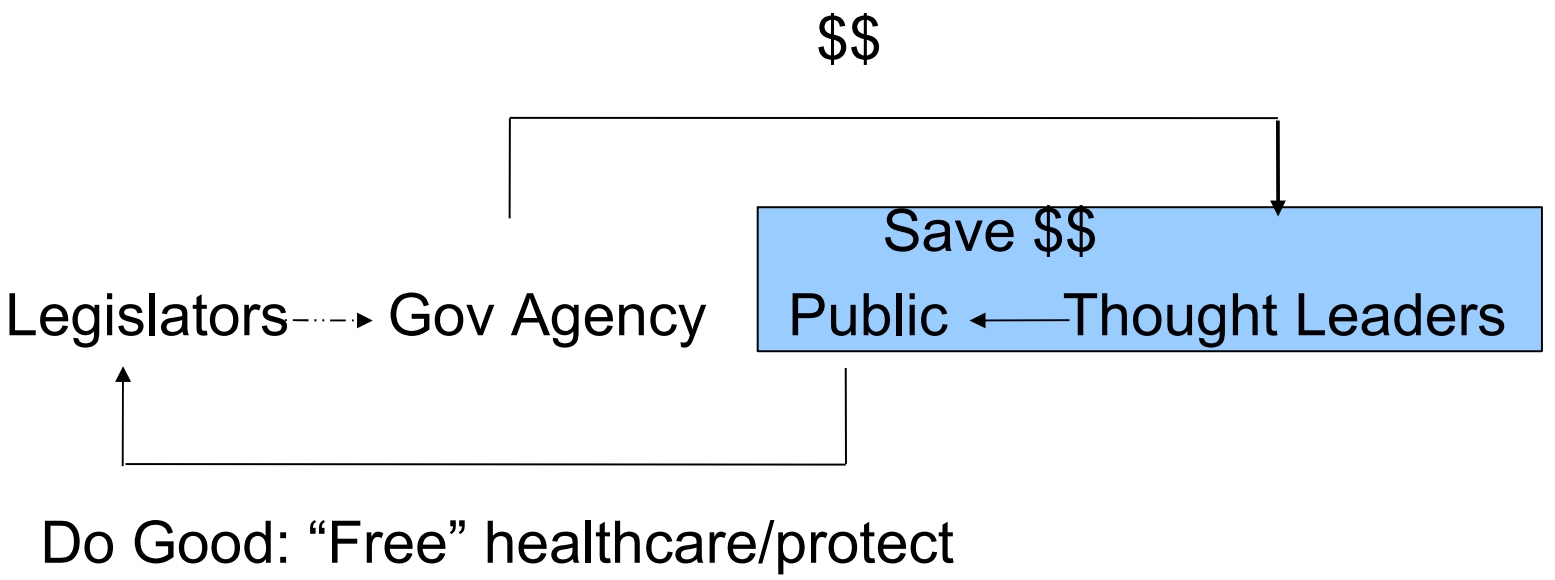
NISS

The Players

1. The workers – Analysts
2. The communicators –
   a. Gov paid analysts
   b. Industry paid analysts
   c. Etc.
3. The consumers – Gov payers, Insurance companies, HMOs, ….physicians, public.
4. The management – ? AWL, free market

NISS

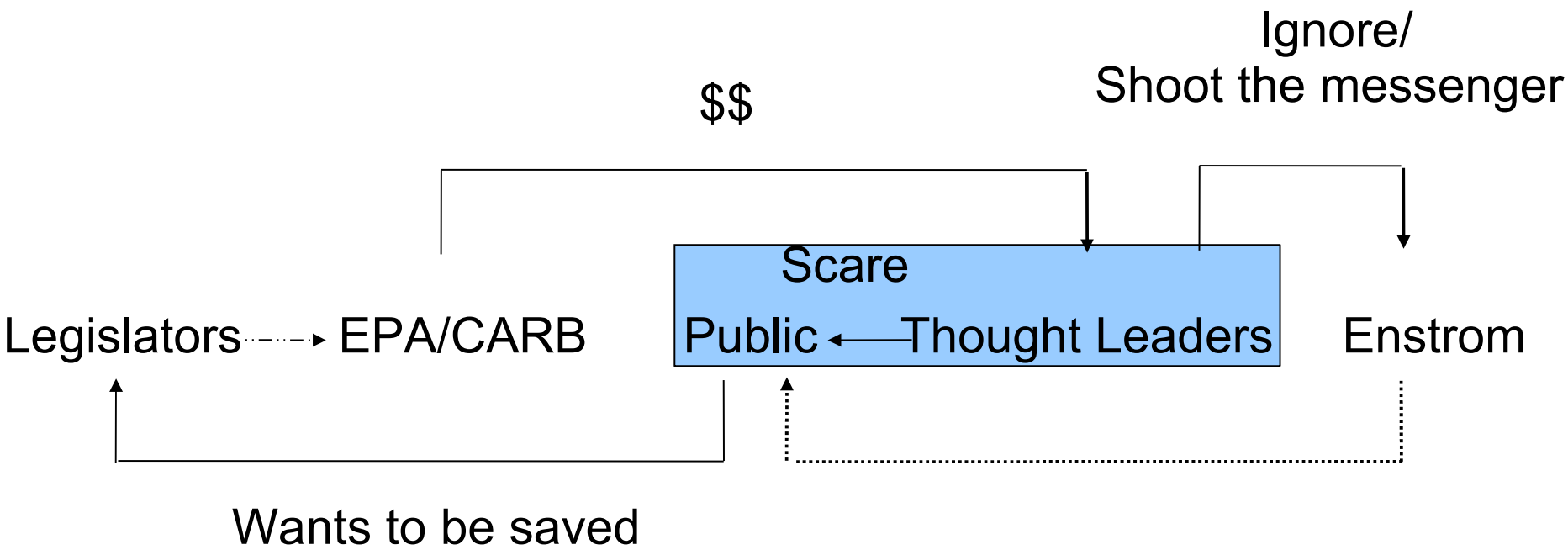
Policy diagram / Process Control



Classic “process control” diagram.
For good of public or Gov Agency?

NISS

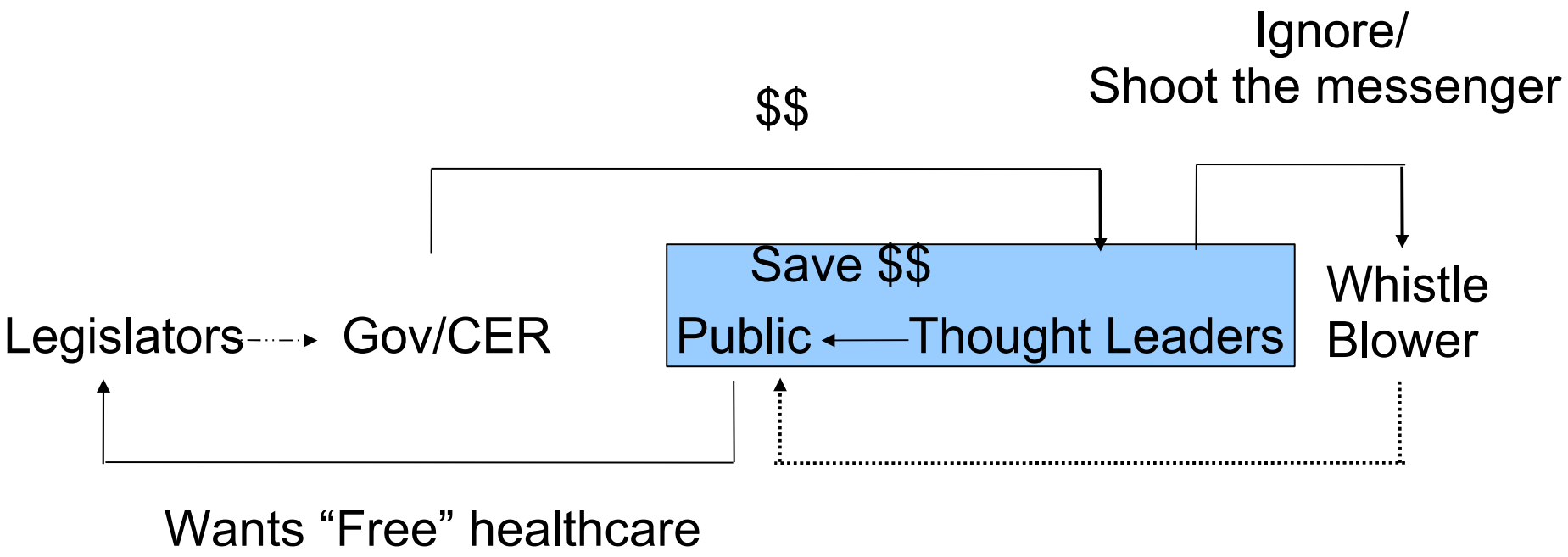
Follow the flow of $$/influence



The whole aim of practical politics is to keep the populace alarmed (and hence clamorous to be led to safety) by an endless series of hobgoblins, most of them imaginary. H. L. Mencken

NISS

Follow the policy diagram



Classic “process control” diagram.

The Strategy

1. Separate data collection and cleaning from analysis.
2. Create a test and holdout set.
3. Analysts examines test set predictors/covariates.
4. Analysts writes and files a statistical protocol.
5. Analysis of test set and paper written.
6. *Journal agrees to blind Addendum.*
7. 2nd team uses analysis method on holdout set

   and writes Addendum.

NISS

Can we do both?

Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

To Explain or to Predict?

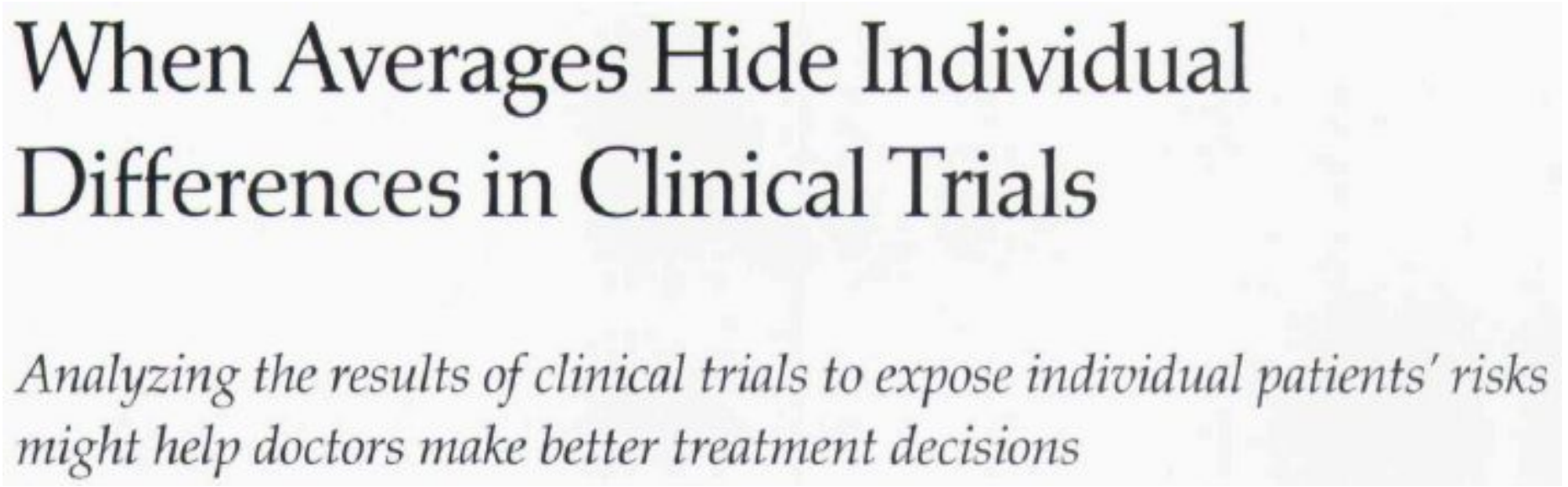
Galit Shmueli

Recursive Partitioning, single tree and multiple trees.

OptimusRP from Golden Helix.

Partition from SAS JMP.

Main Effects versus Interactions



NISS

Step-wise regression analysis

	Step	Parameter	"Sig Prob"	Seq SS	RSquare	Cp
1	1	wprevcost	0.0000	970419203765	0.4650	41505.08
2	2	offcount	0.0000	75262438994	0.5011	21856.51
3	3	age	0.0000	40780963719	0.5207	11210.84
4	4	trtm	0.0000	15739042603	0.5282	7103.47
5	5	psycpcnt	0.0000	10871488225	0.5334	4266.99
6	6	hospcount	0.0000	6364964645	0.5365	2607.14
7	7	ercount	0.0000	6870462995.5	0.5398	815.30
8	8	pain	0.0000	3057431809.7	0.5412	19.02
9	9	gender	0.0009	42221864.302	0.5412	10.00

NISS

The Big Three of Observational Studies

1. Bias - blocking

2. Multiple testing

3. Multiple modeling

NISS

Clustering

Size and guiding principles.

Unsupervised and supervised.

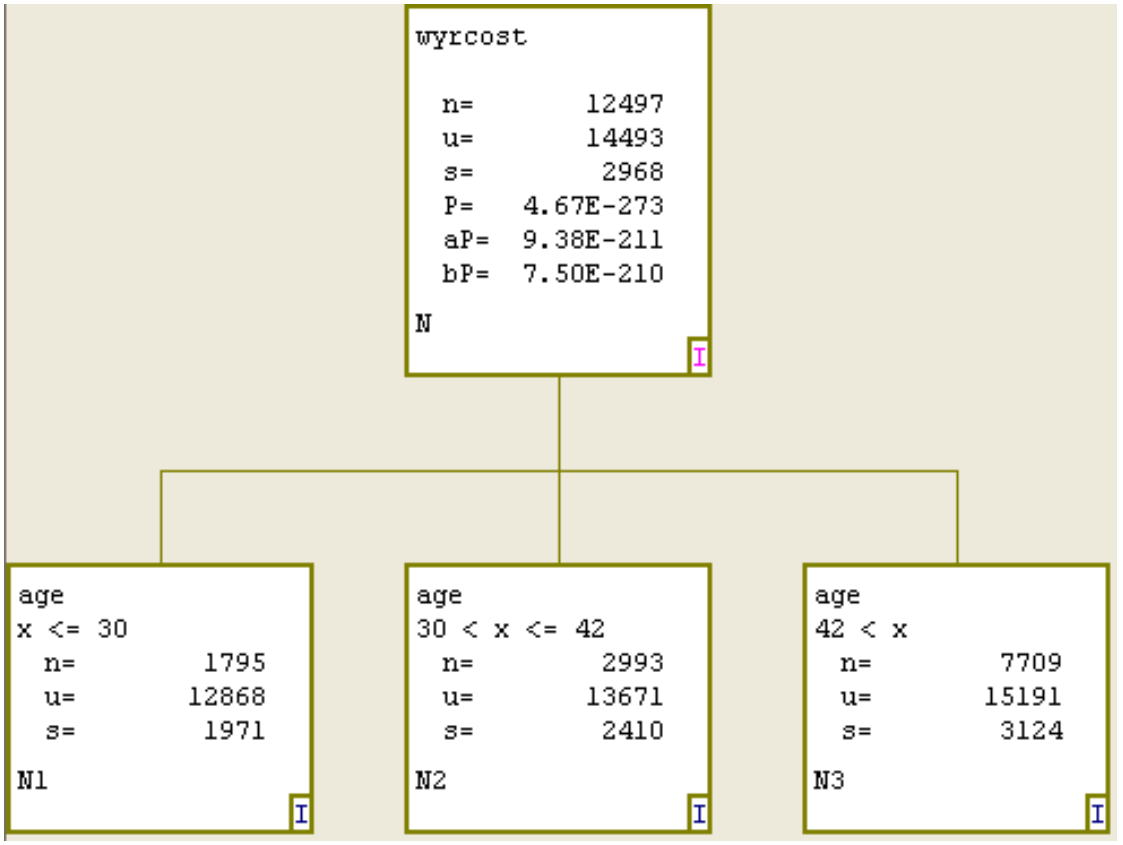
We will examine supervised.

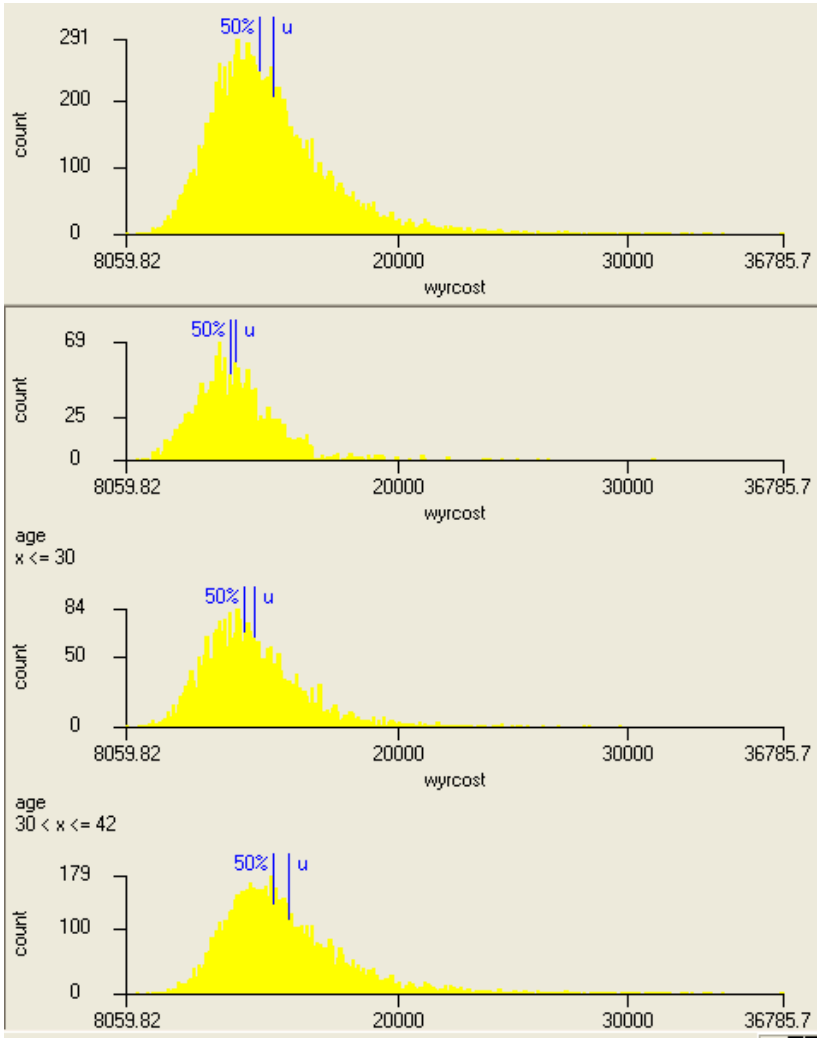
Recursive Partitioning Software

goldenhelix.com/Predictive_Analytics/optimus_rp

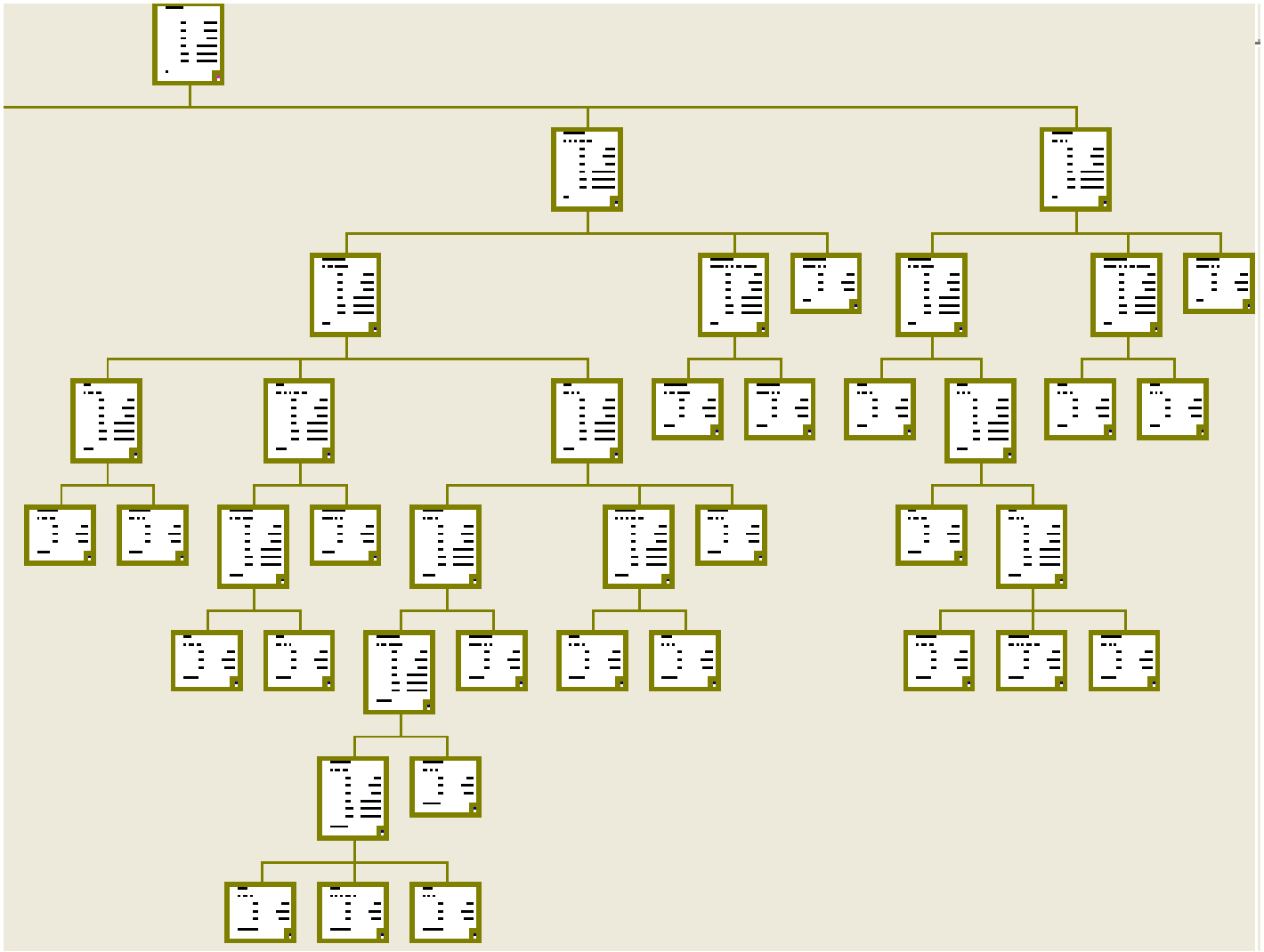
jmp.com/software/jmp9

Recursive Partitioning (1)



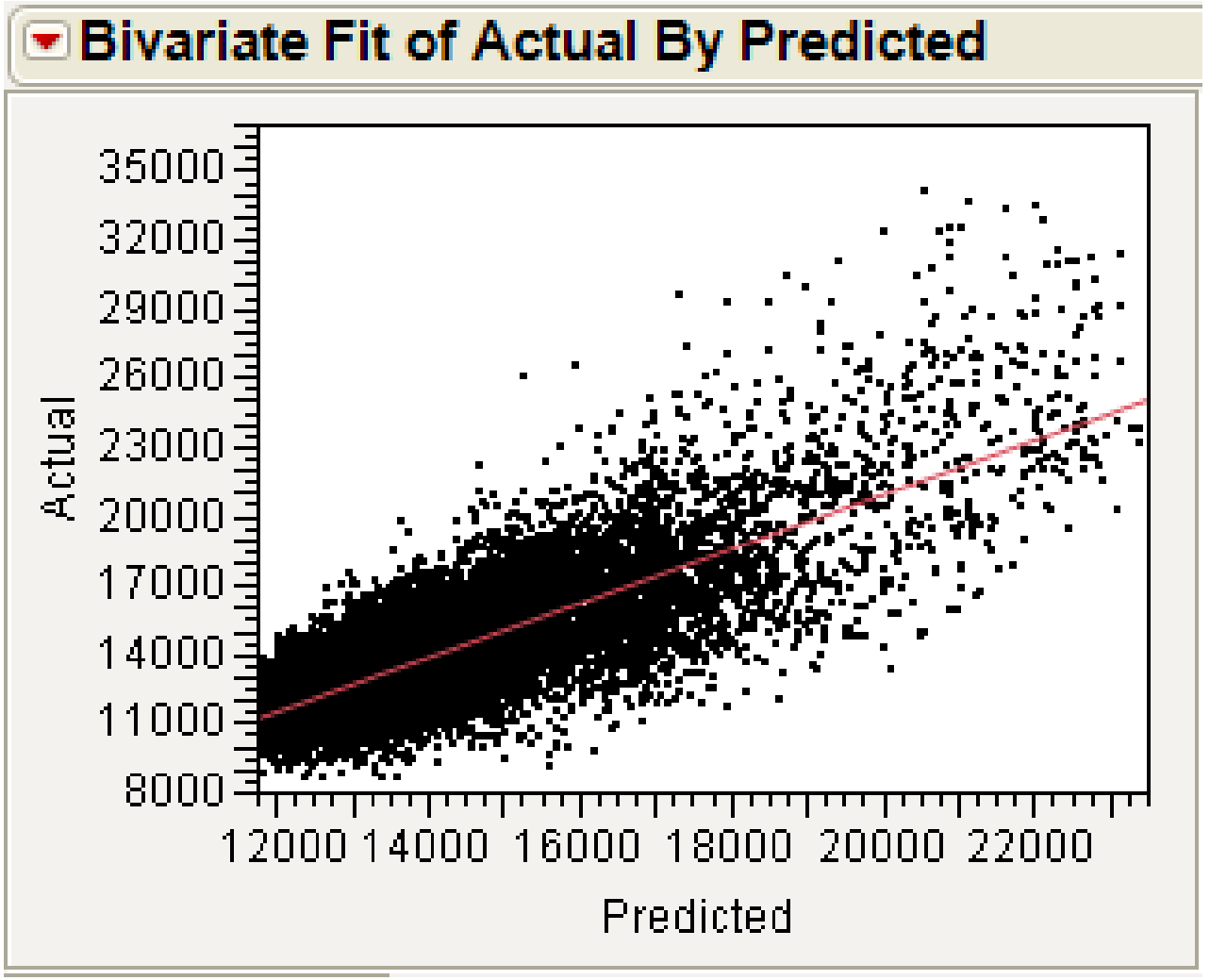


Recursive Partitioning (2)



NISS

Random Sample, 100 Trees



Summary of Fit	
RSquare	0.616161

NISS

RP, Multiple Trees

Variable Name	Column	# times used
wprevcost	10	1759
offcount	8	1492
age	3	1373
pain	5	889
psycpcnt	9	878
ercount	7	469
hospcount	6	359
gender	4	239

	wprevc	offcount	age	pain	psycpcnt	ercount	hospcour
wprevcos	1	0.93	0.81	0.6	0.71	0.45	0.43
offcount	-4.6	0.97	0.81	0.58	0.71	0.5	0.45
age	-2.5	1.3	0.83	0.5	0.65	0.44	0.37
pain	-0.1	-0.1	0.5	0.6	0.46	0.3	0.27
psycpcnt	-5.0	-0.8	4.7	2.6	0.74	0.4	0.33
ercount	-10.1	0.8	2.0	-1.8	3.2	0.51	0.24
hospcount	-6.7	0.0	-2.4	-1.8	-3.3	0.3	0.47

NISS

Local Treatment Differences

Tensions

1. Patients in group are comparable.
2. Effects of covariates can be determined.
3. Not too many “pure” nodes.
4. Stable distribution of LTDs.
5. Etc.

Guided clustering using RP

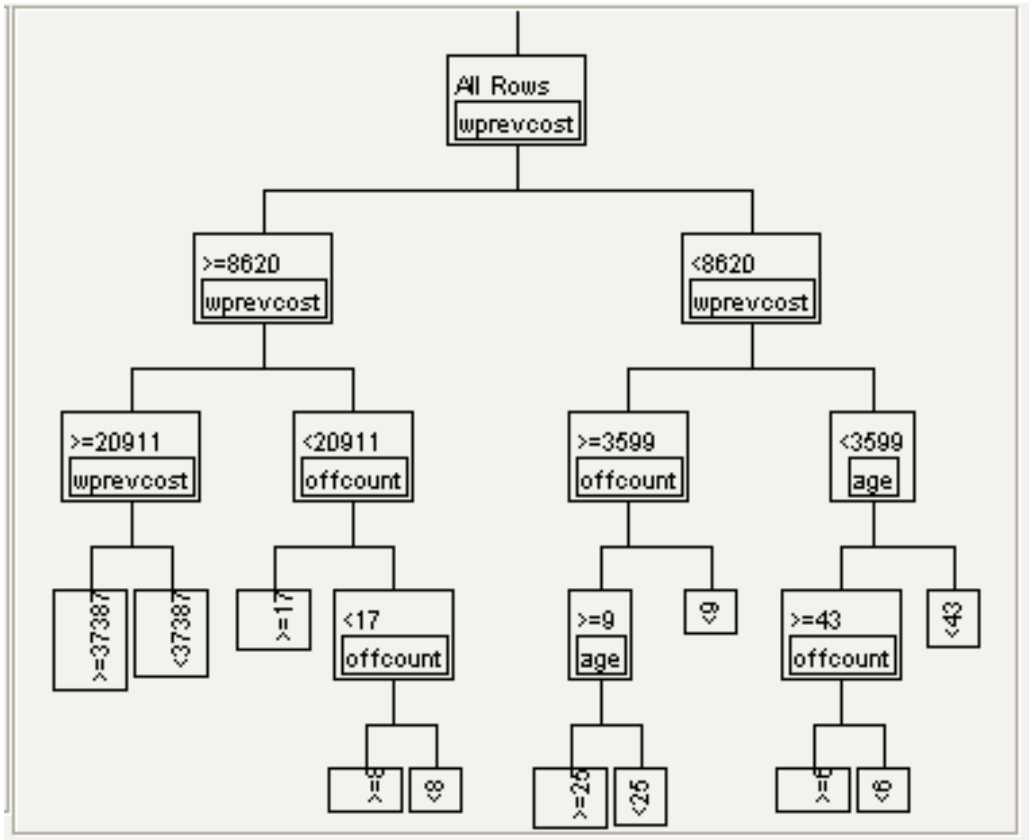
Find combinations of Xs

that cluster Treatments.

Find combinations of Xs

that cluster Costs.

JMP RP 10 y=trt

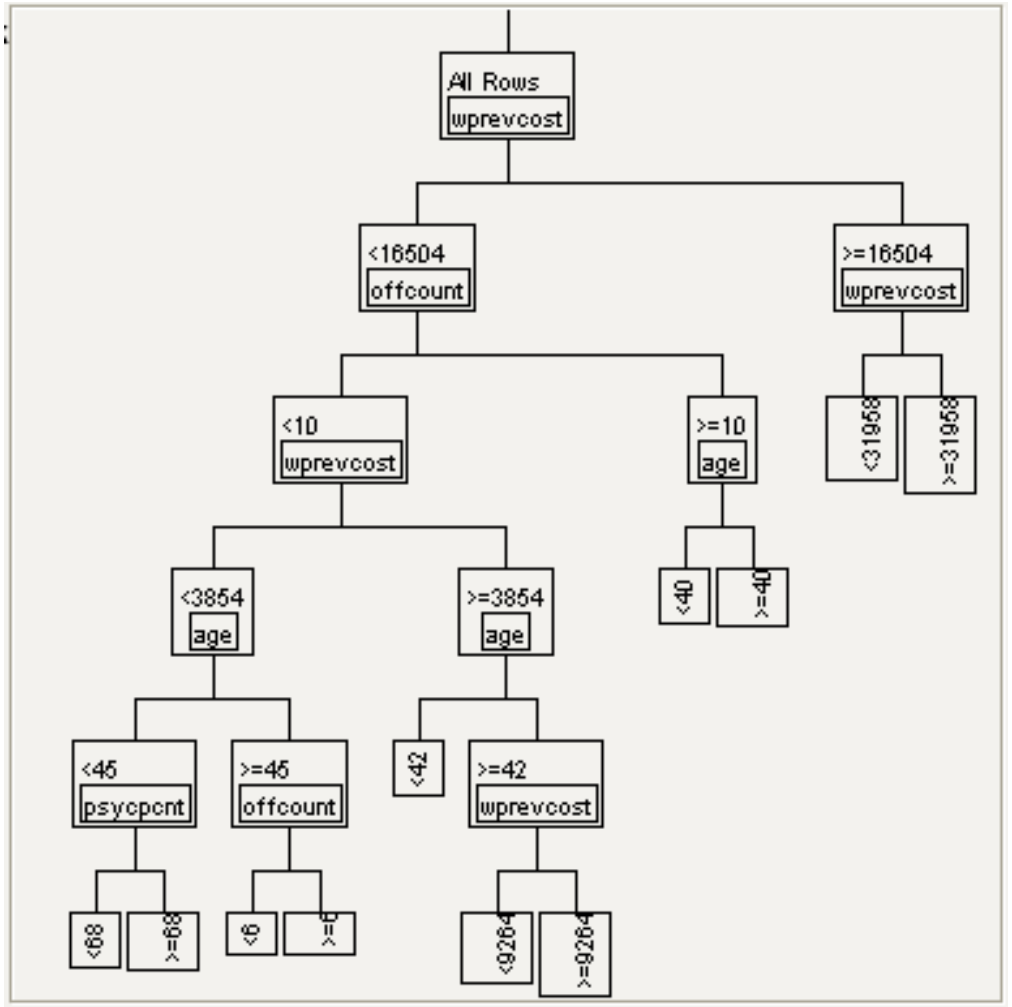


	Leaf #Splits...	N Rows	Dif
1	1	5657	-4706.46
2	2	14937	-2269.35
3	3	7332	-1995.36
4	4	27424	-1058.20
5	5	23481	-470.39
6	6	20505	-778.97
7	7	1482	299.44
8	8	52254	-96.39
9	9	10518	-203.59
10	10	36694	94.53
11	11	49674	317.59

High cost patients appear to benefit most.

NISS

JMP RP 10 nodes, y=cost, (no trt)



RSquare	RMSE	N	Number of Splits
0.493	2056.5633	249958	10

	Node	Trt0	Trt1	Dif
1	1	12285.96	12606.77	320.81
2	2	13220.72	13317.44	96.72
3	3	13216.89	13297.07	80.18
4	4	14227.32	14008.04	-219.28
5	5	13529.09	13477.43	-51.66
6	6	14214.35	13967.06	-247.28
7	7	15456.05	14707.81	-748.24
8	8	14332.86	14233.11	-99.75
9	9	16629.60	15446.33	-1183.27
10	10	18422.33	16837.44	-1584.89
11	11	24706.07	20673.56	-4032.51

NISS

Stats for Two Methods

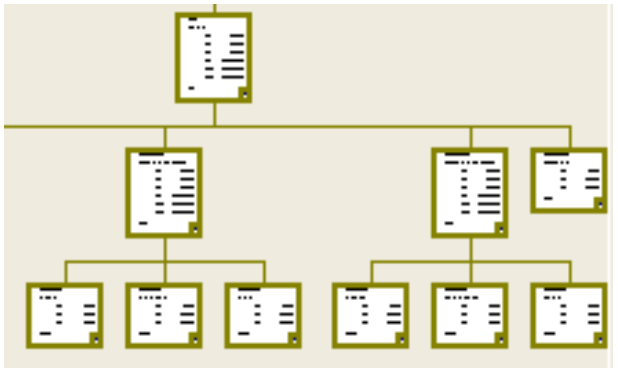
RP by Cost

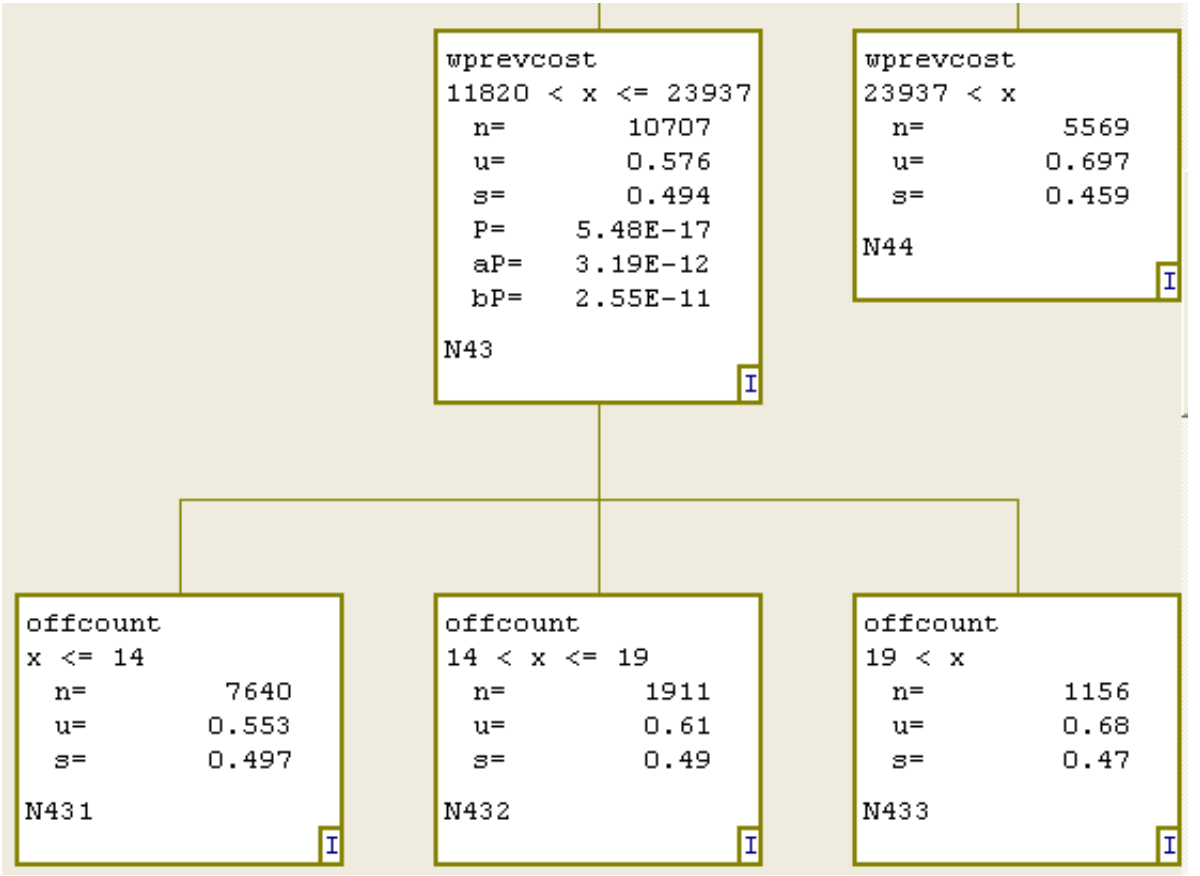
	Leaf#10Splits Cost	N Rows	Prob(trtm)
1	1	44953	0.3564
2	2	10311	0.3895
3	3	33804	0.3849
4	4	8631	0.4278
5	5	25740	0.4164
6	6	37700	0.4350
7	7	15836	0.4976
8	8	10112	0.4697
9	9	31211	0.5281
10	10	23401	0.5976
11	11	8259	0.7071

PR by Trt

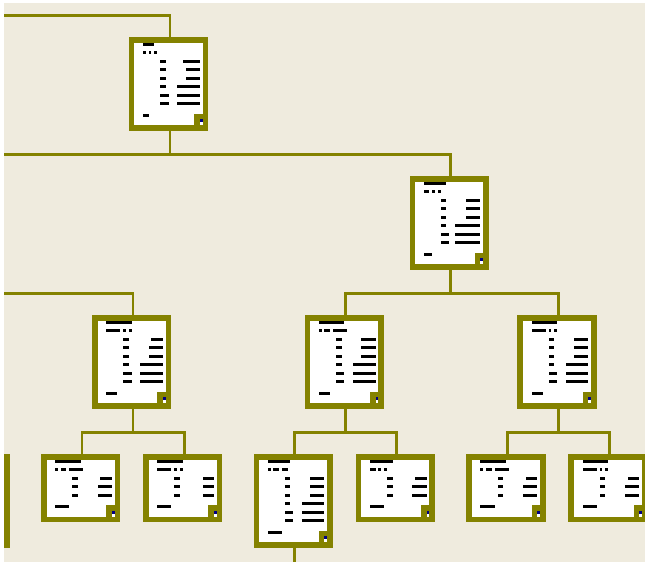
	Node	Trt0	Trt1	Dif
1	1	12285.96	12606.77	320.81
2	2	13220.72	13317.44	96.72
3	3	13216.89	13297.07	80.18
4	4	14227.32	14008.04	-219.28
5	5	13529.09	13477.43	-51.66
6	6	14214.35	13967.06	-247.28
7	7	15456.05	14707.81	-748.24
8	8	14332.86	14233.11	-99.75
9	9	16629.60	15446.33	-1183.27
10	10	18422.33	16837.44	-1584.89
11	11	24706.07	20673.56	-4032.51

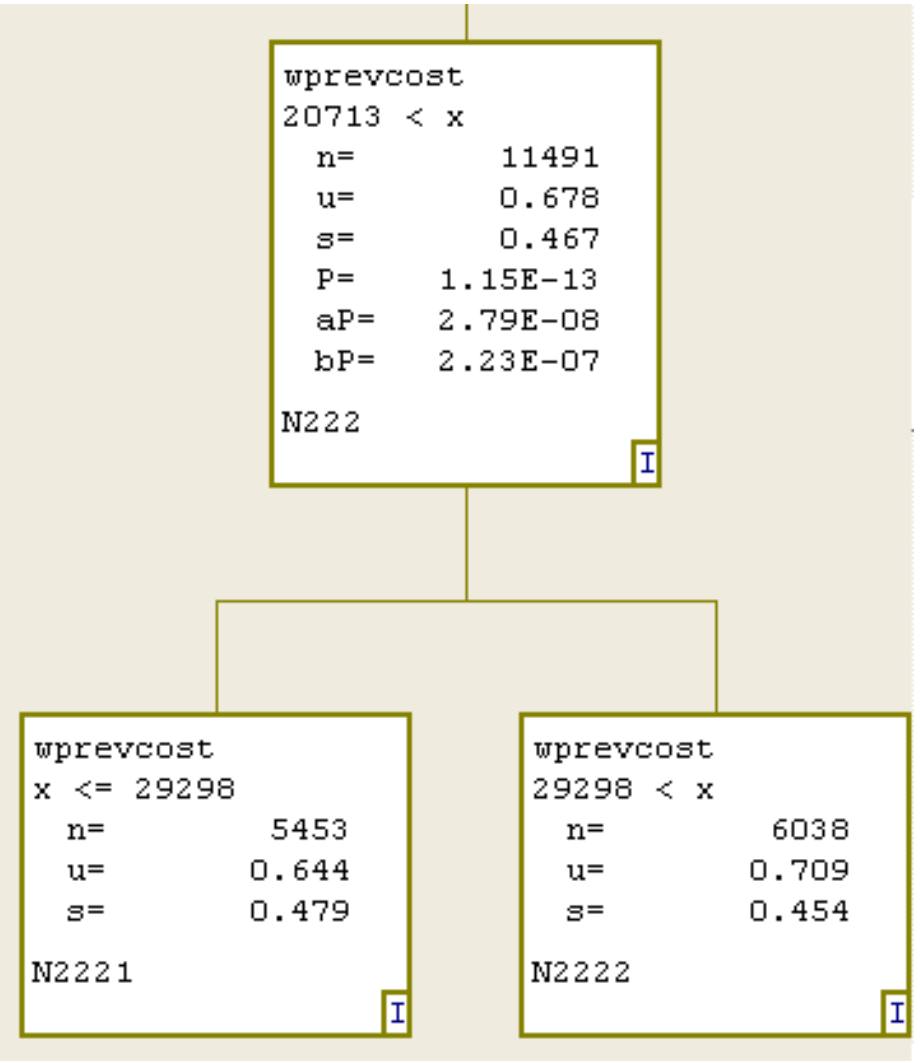
Splits 4 node size 250 (1)





Splits 2 node size 500 (1)





NISS

Design of Experiments

Evaluation

1. rMSE.

2. # missing values.

Factors

1. Trt versus Cost for y.

2. Number in daughter nodes.

3. Significance for a split.

4. Etc.

Want terminal nodes small, but not “pure”.

Initial Design, 2^(3-1)

	Treatment		Cost	
	250	500	250	500
4	x			x
2		x	x	

NISS

Initial Design, 2^(3-1)

rMSE, # terminal nodes

	Treatment (0/1)		Cost ($$)	
	250	500	250	500
4	926, 61			949,335
2		896,55	**857,529**	

2 daughter nodes better
250 Min in node better
Cost better than treatment.

NISS

Follow up experiments

	Treatment (0,1)		Cost ($$)	
	250	500	250	500
4	926, 61			949,335
2		896,55	**857,529**	

Cost, 2, 125	**822**, 780

NISS

Summary

Clustering on Cost vs Trt is effective.

Recursive partitioning is competitive for LTDs.

DoE looks to be useful to tune LTD process.

Hint that 2-way splitting is better than multi-way.

Contact Information

Stan Young

www.niss.org

young@niss.org

919 685 9328

NISS

Reproducible Research

"A piece of reproducible research is an article that provides readers with all the materials that are needed to produce the same results as described in the publication."

1. Protocol
2. Electronic data
3. Analysis code, e.g. SAS code.

Biometrical Journal 51 (2009), 553–555

NISS