Reduced Error Logistic Regression

Daniel M. Rice, Ph.D.
Rice Analytics, St. Louis, MO
Classification Society Annual Conference
June 11, 2009

Reduced Error Logistic Regression (RELR)

- New and general form of regression where error is modeled in the Logistic Regression formulation.
- Because there are regression coefficients for error and information in the model, **estimate of error is essentially subtracted from the information portion of model**
- Results in reduced error solution in terms of regression coefficients and better goodness of fit measures for **problems with significant multicollinearity** or with significant dimensionality considerations – **problems where important interactions exist**
- Results in Parsed RELR – **entirely automated variable selection without arbitrary parameters** to give parsimonious regression models often with fewer than 10 selected variables – alternative to Stepwise Logistic Regression

Penalized Logistic Regression

- Penalized Logistic Regression (Le Cessie & Van Houwelingen, 1992) is the method that RELR most closely resembles:
- PLR maximizes the following objective:

$$LL(\beta_0,\boldsymbol{\beta},\lambda) = l(\beta_0,\boldsymbol{\beta}) - \lambda/2 \sum_{r=1}^{m} \beta_r^{\,2}$$

- PLR suffers from disadvantage that it requires **arbitrary scaling factor λ or it requires estimation of λ from validation sample** (Park and Hastie, 2008).

RELR Log Likelihood

– RELR maximizes the following LL objective:

$$LL(p,w) = \sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\ln(p_{ij}) + \sum_{l=1}^{2}\sum_{r=1}^{M}\sum_{j=1}^{C} y_{jlr}\ln(w_{jlr})$$

- summation on left is standard log likelihood that models information probability vector **p**; summation on right models error probability vector **w** like Golan et al. (1996).
- Unlike PLR, no arbitrary scaling constant in RELR.
- Fact **that error is modeled rather than smoothed is RELR advantage over PLR or Lasso**. Not an arbitrary error model function.

RELR (see 2008 JSM paper)

- Given equivalence between LL and entropy solutions in logistic regression (Golan et al., 1996), RELR also maximizes the following entropy objective:

$$(1) \qquad H(p,w) = -\sum_{i=1}^{N}\sum_{j=1}^{C} p_{ij}\ln(p_{ij}) - \sum_{l=1}^{2}\sum_{r=1}^{M}\sum_{j=1}^{C} w_{jlr}\ln(w_{jlr})$$

subject to constraints that include:

$$(2) \qquad \sum_{i=1}^{N}\sum_{j=1}^{C}(x_{ijr}\,y_{ij}) = \sum_{i=1}^{N}\sum_{j=1}^{C}(x_{ijr}\,p_{ij}) + (u_r\,w_{j1r} - u_r\,w_{j2r}) \quad \text{for } r = 1 \text{ to } M,$$

$$(3) \qquad \sum_{j=1}^{C} p_{ij} = 1 \text{ for } i = 1 \text{ to } N,$$

$$(4) \qquad \sum_{j=1}^{C} w_{jlr} = 1 \text{ for } l = 1 \text{ to } 2 \text{ and } r = 1 \text{ to } M,$$

$$(5) \qquad \sum_{i=1}^{N} y_{ij} = \sum_{i=1}^{N} p_{ij} \quad \text{for } j = 1 \text{ to } C\text{-}1,$$

RELR (see 2008 JSM paper)

Symmetrical error probability constraints – force lack of bias in positive and negative error probability ***w*** across variables:

$$(6) \qquad \sum_{j=1}^{C} \left(\sum_{r=1}^{M} s_r w_{j1r} - \sum_{r=1}^{M} s_r w_{j2r} \right) = 0$$

$$(7) \qquad \sum_{j=1}^{C} \left(\sum_{r=1}^{M} w_{j1r} - \sum_{r=1}^{M} w_{j2r} \right) = 0$$

Extreme Value Error (see JSM paper)

- **Logit for error associated with each independent variable -** error is assumed to be Extreme Value Type I error (Luce and Suppes, 1965; McFadden, 1974) that is expected to be proportional to $1/t_r$; where t_r is the t value that reflects the extent to which a correlation between the *r*th independent variable and the dependent variable is different from zero.
- Gives large expected error when t_r is small and small expected error when t_r is large.
- RELR further assumes that probability of positive and negative error is equal across all variables. **No bias in error across variables.**

t-values in RELR (see JSM paper)

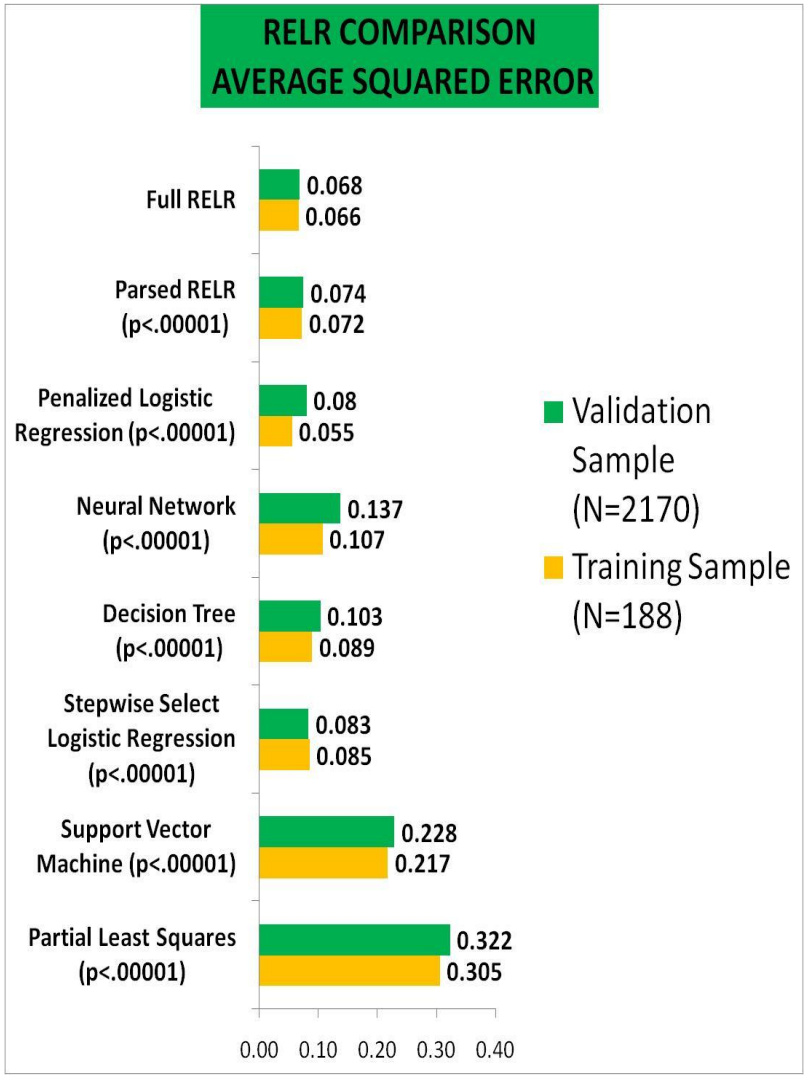
Method	Relationship Between β and t	Interactions
Naïve Bayes (Mitchell, 2005)	$\beta_r \propto t_r$	Does not model
RELR/Many IVs with large magnitude t-values (Often Full RELR)	$\beta_r \propto t_r$ (approximately)	Does model
RELR/Few Important IVs (Always Parsed RELR; Sometimes Full RELR)	β_r also depends upon covariance with other IVs	Does model

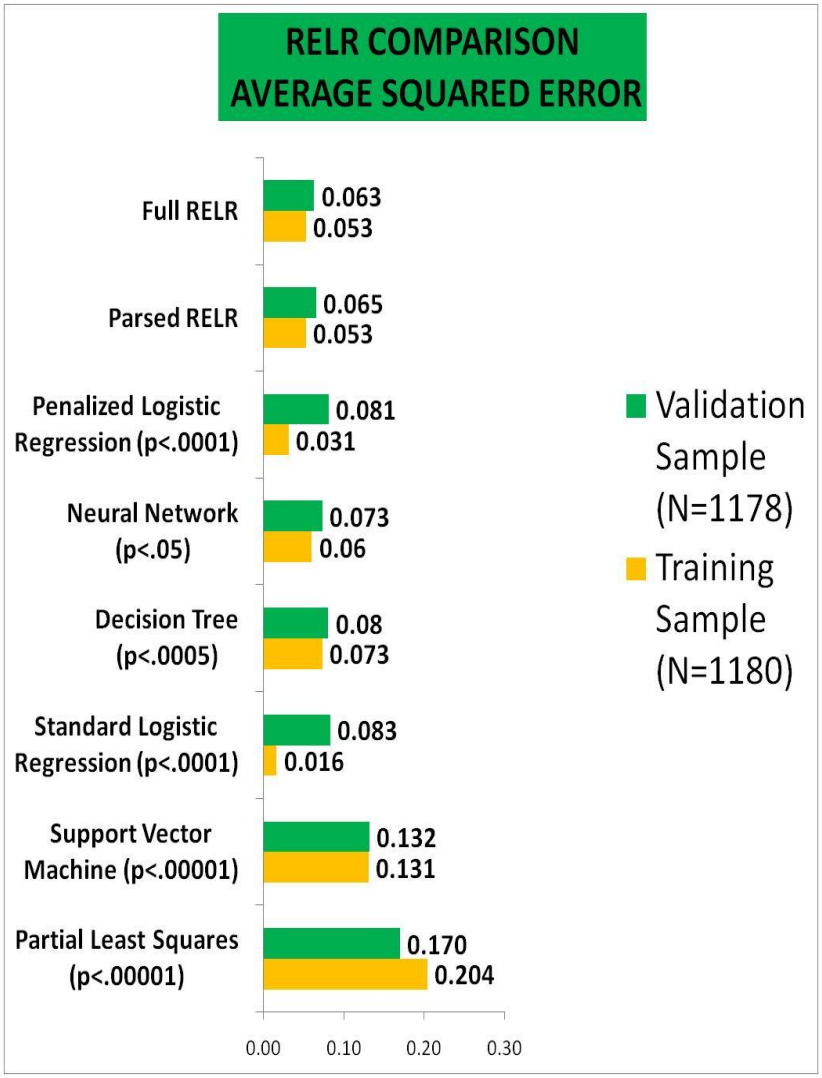
- Because β_r is roughly proportional to t_r across many important IVs, **RELR screens IVs based upon magnitude of t_r and avoids curse of dimensionality.** Model most important 400 IVs instead of all 50,000 IVs and get β_r's with almost identical magnitude ordering.
- Because approximate proportionality between β_r and t_r breaks down with few important IVs, **Parsed RELR** solutions with few IVs have β_r's that also depend upon covariance between IVs where **variables with largest magnitude t_r's are not always selected.**

Pew 2004 Election Survey

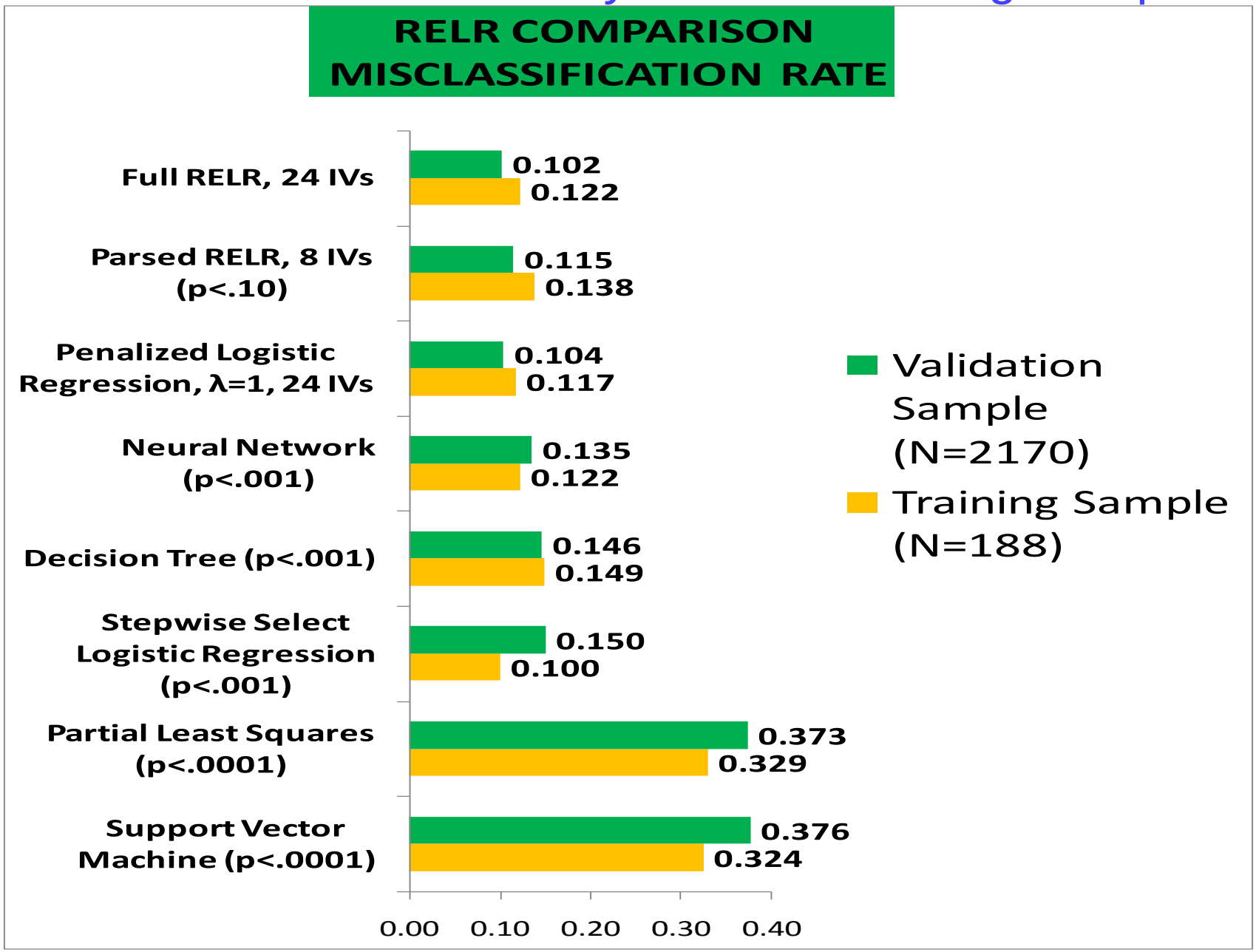
- Weekend before election; Bush vs. Kerry; Kerry as the target condition;
- Balanced data **(election was almost 50/50 outcome); no separate error measures for target and non-target conditions**
- Highly multicollinear dataset (correlations between variables as high as .9); important variables like Democratic and Republican were highly negatively correlated
- Telephone survey data with many missing values in independent variables – **impacts accuracy of model around .5 probability level**
- **PLR results used validation sample data to get *λ***

Pew 2004 Election Survey

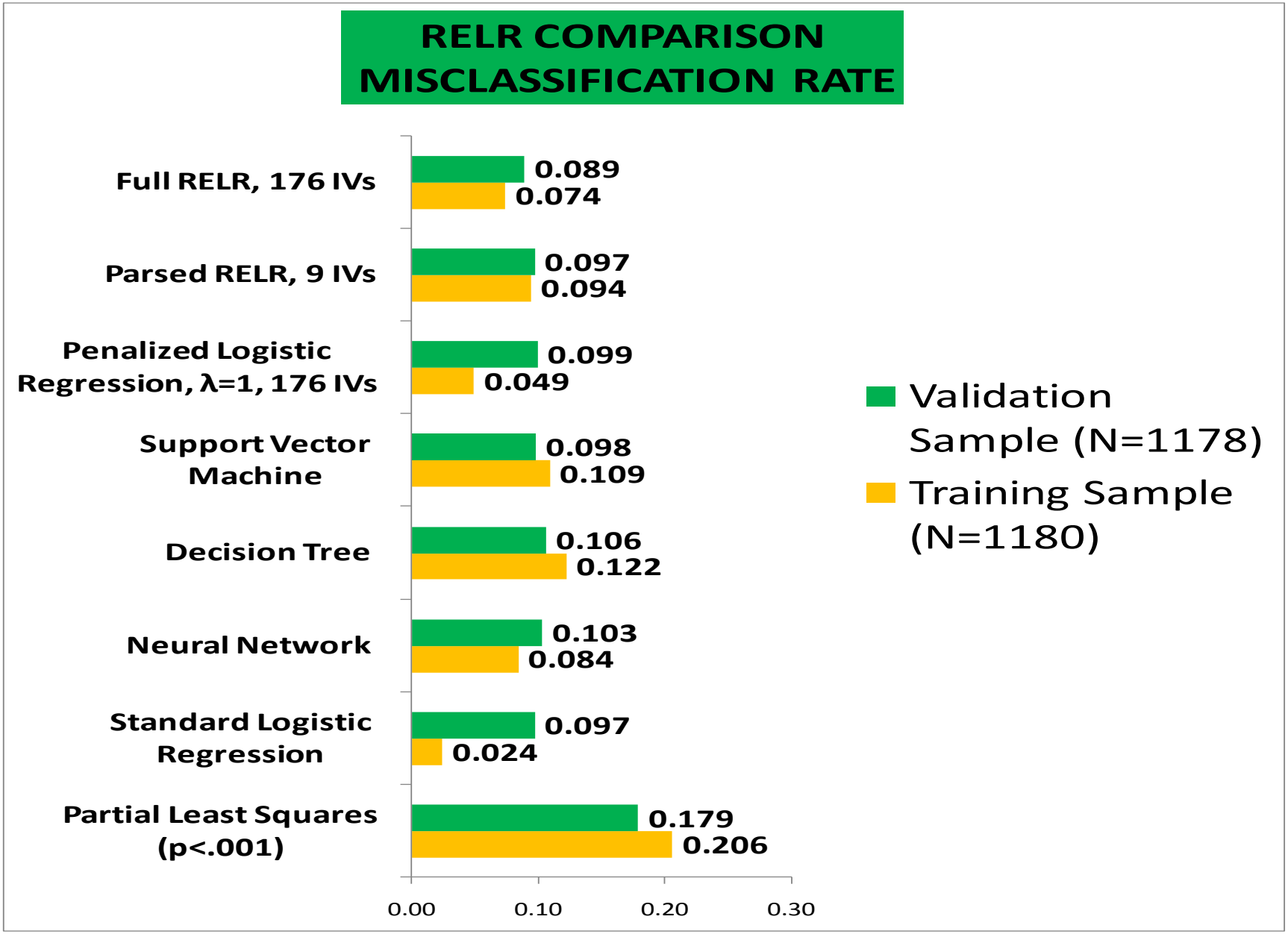




2004 Pew Election Survey Smaller Training Sample



2004 Pew Election Survey Larger Training Sample



Consistency of Standardized Logit Coefficients from Larger Sample

Table 1. Full RELR's Coefficients (r=.976 between Training and Validation)

PARAMETER	ESTIMATE	STDERR	PROB
IRAQWASWRONG	0.364	0.047	<.0001
USLOSINGWARONTERROR	0.266	0.035	<.0001
NORISKINCHANGEDURINGWAR	0.263	0.034	<.0001
REPUBLICAN	-0.247	0.032	<.0001
BUSHWILLLOSEELECTION	0.244	0.032	<.0001
REGISTEREDxIRAQWASRIGHT	-0.236	0.031	<.0001
DEMOCRAT	0.217	0.028	<.0001

Table 2. Full PLR's Coefficients, λ=1 (r=.238 between Training and Validation)

PARAMETER	ESTIMATE	STDERR	PROB
REPUBLICANxREGISTERED^3	-7.405	4.442	0.096
DEMOCRATxPARENT^3	-6.132	2.868	0.033
REPLUBICANxBORNAGAIN^2	3.823	2.011	0.057
REGISTEREDxIRAQWASRIGHT^3	3.632	1.930	0.060
PARTYLINEVOTERxDEMOCRAT	-3.211	4.016	0.424
PARTYLINEVOTER	-3.208	4.014	0.424
RISKINCHANGExATTENDCHURCH^3	-3.171	1.154	0.006

Parsed RELR Standardized Logit Coefficients

Training Sample – Large Sample

PARAMETER	ESTIMATE	STDERR	PROB
INTERCEPT	-0.040		
DEMOCRAT	1.049	0.112	<.0001
NOTPARTYLINEVOTER	-0.947	0.105	<.0001
NOTPARTYLINEVOTERxDEMOCRAT	-0.929	0.103	<.0001
IRAQWASWRONG	1.653	0.176	<.0001
USLOSINGWARONTERROR	1.223	0.139	<.0001
LOWRISKINCHANGE	1.281	0.136	<.0001
HIGHRISKINCHANGE	-0.978	0.114	<.0001
BUSHWILLLOSE	1.153	0.126	<.0001
REPUBLICAN	-1.207	0.128	<.0001

Validation Sample

PARAMETER	ESTIMATE	STDERR	PROB
INTERCEPT	-0.040		
DEMOCRAT	0.937	0.093	<.0001
NOTPARTYLINEVOTER	-0.937	0.102	<.0001
NOTPARTYLINEVOTERxDEMOCRAT	-0.908	0.100	<.0001
IRAQWASWRONG	1.255	0.121	<.0001
USLOSINGWARONTERROR	0.759	0.087	<.0001
LOWRISKINCHANGE	0.856	0.091	<.0001
HIGHRISKINCHANGE	-0.748	0.083	<.0001
BUSHWILLLOSE	0.788	0.084	<.0001
REPUBLICAN	-0.951	0.094	<.0001

Parsed RELR Variable Selection

- Backward Selection – If M original variables, then M steps; avoid Stepwise Data Torturing and Overfitting
- Mechanical process without arbitrary criteria
- Smallest set of most important variables should give largest RELR Log Likelihood. We need to drop all unimportant variables that are either redundant or have small effects.

1. Choose shortlist of M original independent variables based upon magnitude of t values (univariate importance). Choose shortlist size as large as possible until larger makes no difference in maximal step 5 solution (very often 100 or less).
2. Maximize RELR Log Likelihood function:

$$LL(p,w) = \sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\ln(p_{ij}) + \sum_{l=1}^{2}\sum_{r=1}^{M}\sum_{j=1}^{C} y_{jlr}\ln(w_{jlr})$$

3. Drop the least important variable (multivariate importance), so *M* now equals *M*-1.
4. Store value of *LL(p,w);* if M>=1, then go back to step 2.
5. Parsed RELR maximum likelihood solution has maximal *LL(p,w)* of all solutions in the backward selection path.

OTHER INFORMATION

We thank the Pew Research Center for making their data available for this type of research.

We thank SAS Institute for their continued partnership.

Generalized RELR Method is currently pending patent.

Communications to:

Dan Rice
info@RiceAnalytics.com

REFERENCES

Golan, A., Judge, G., and Miller, D. (1996), A maximum entropy approach to recovering information from multinomial response data. *Journal of the American Statistical Association*, 91: 841-853.

Lecessie S. Van Houwelingen, JC, Ridge estimators in logistic regression, Appl Stat-J Roy STC 41 (1): 191-201, 1992.

Luce, R.D. and Suppes, P. (1965). Preference, utility and subjective probability, in R.D. Luce, R.R. Bush and E. Galanter (eds), Handbook of Mathematical Psychology, Vol. 3, Wiley and Sons, New York, NY, pp. 249-410.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (ed) Frontiers in

Econometrics, New York, Academic Press, pp. 105-142.

Mitchell, T. (2005), Generative and discriminative classifiers: Naïve Bayes and Logistic Regression. Online draft.

Park, M.Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. Online manuscript: plr_interactions.pdf/penalized-logistic-regression-for.pdf

Rice, D.M. (2008), Generalized Reduced Error Logistic Regression Machine, *Section on Statistical Computing - JSM Proceedings 2008*, pp. 3855-3862.