Dependence between Models and Tests in Predictive Toxicology

Tom Aldenberg

RIVM

Bilthoven, NL

OpenTox Interaction Meeting: Műnchen, August 9-12, 2011

Dependence in Predictive Toxicology: Overview

A. Introduction: ITS Main Problem
B. Statistical Modeling
C. Conclusions

A. Introduction: ITS Main Problem

1. We have chemical categorical data, cross-tabulating (binary) predictors, e.g. model runs, tests, assays, etc., and a (binary) endpoint. ***What to do??***
2. The Bayesian Paradigm of Medical Diagnosis
3. Naïve Bayes and Independence
4. Conditional Dependence

A.1. Contingency Table of 522 chemicals for LLNA and 5 QSARS

What to do?? Zeros, (in-) dependence?

	SMARTs	DEREKfW	TIMES	TopKat	MultiCASE	LLNA+	LLNA–	
31	1	1	1	1	1	124	14	138
30	1	1	1	1	0	17	2	19
29	1	1	1	0	1	39	5	44
28	1	1	1	0	0	7	0	7
27	1	1	0	1	1	64	15	79
26	1	1	0	1	0	5	2	7
25	1	1	0	0	1	14	3	17
24	1	1	0	0	0	1	1	2
23	1	0	1	1	1	4	4	8
22	1	0	1	1	0	2	0	2
21	1	0	1	0	1	5	0	5
20	1	0	1	0	0	1	0	1
19	1	0	0	1	1	11	7	18
18	1	0	0	1	0	0	1	1
17	1	0	0	0	1	6	5	11
16	1	0	0	0	0	0	0	0
15	0	1	1	1	1	5	3	8
14	0	1	1	1	0	0	0	0
13	0	1	1	0	1	12	1	13
12	0	1	1	0	0	0	0	0
11	0	1	0	1	1	20	9	29
10	0	1	0	1	0	0	1	1
9	0	1	0	0	1	7	10	17
8	0	1	0	0	0	1	1	2
7	0	0	1	1	1	0	3	3
6	0	0	1	1	0	0	0	0
5	0	0	1	0	1	2	0	2
4	0	0	1	0	0	0	0	0
3	0	0	0	1	1	10	10	20
2	0	0	0	1	0	7	2	9
1	0	0	0	0	1	13	26	39
0	0	0	0	0	0	11	9	20
						388	134	522

Pers. comm.: Rorije, Van den Briel, Enoch, and others

A.2. The Bayesian Paradigm of Medical Diagnosis

QSAR Data: smarts

smarts		LLNA+	LLNA–	sum
1		300	59	359
0		88	75	163
		388	134	522

QSAR Data: derekfw

	derekfw	LLNA+	LLNA–	sum
	1	316	67	383
	0	72	67	139
		388	134	522

Conditional Probs: Pr (smarts| LLNA)

smarts		LLNA+	LLNA–	WoE+
1		0.773	0.440	2.45
0		0.227	0.560	-3.92
		1.000	1.000	

Conditional Probs: Pr (derekfw | LLNA)

	derekfw	LLNA+	LLNA–	WoE+
	1	0.814	0.500	2.12
	0	0.186	0.500	-4.30
		1.000	1.000	

Posterior Probs: Pr (LLNA | smarts)

smarts		LLNA+	LLNA–	WoE+
1		0.637	0.363	2.45
0		0.288	0.712	-3.92

Posterior Probs: Pr (derekfw | LLNA)

	derekfw	LLNA+	LLNA–	WoE+
	1	0.620	0.380	2.12
	0	0.271	0.729	-4.30

A.3. Naïve Bayes and Independence

MULTIPLIED FROM INDIVIDUAL SENSITIVITIES AND SPECIFICITIES:

Pr (smarts & derekfw | LLNA) = Pr (smarts | LLNA) * Pr (derekw | LLNA)

smarts	derekfw	LLNA+	LLNA–	WoE+
1	1	0.630	0.220	4.56
1	0	0.143	0.220	-1.86
0	1	0.185	0.280	-1.80
0	0	0.042	0.280	-8.23
		1.000	1.000	

Conditional Probs: Pr (smarts| LLNA)

smarts		LLNA+	LLNA–	WoE+
1		0.773	0.440	2.45
0		0.227	0.560	-3.92
		1.000	1.000	

Conditional Probs: Pr (derekfw | LLNA)

	derekfw	LLNA+	LLNA–	WoE+
	1	0.814	0.500	2.12
	0	0.186	0.500	-4.30
		1.000	1.000	

Conditional Probs multiply and WoEs add!

A.4. Conditional Dependence

Joint cross table leads to Joint Cooper Statistics, i.e. generalized sensitivity and specificity. But: Conditional Probs do not multiply and WoEs do not add.

smarts	derekfw	LLNA+	LLNA–	sum
1	1	271	42	313
1	0	29	17	46
0	1	45	25	70
0	0	43	50	93
		388	134	522

JOINT COOPER STATS FROM THE JOINT TABLE

Conditional Probs: Pr (smarts, derekfw | LLNA)

smarts	derekfw	LLNA+	LLNA–	WoE+
1	1	0.698	0.313	3.48
1	0	0.075	0.127	-2.30
0	1	0.116	0.187	-2.06
0	0	0.111	0.373	-5.27
		1.000	1.000	

A.4. Conditional Dependence (2)

Naïve Bayes, or Independent, model tends to deviate more with increasing number of predictors

Modeled Joint QSAR Data: Independence

LLNA+	smarts	derekfw	times
1	*300.0*	*316.0*	*218.0*
0	*88.0*	*72.0*	*170.0*

smarts	derekfw	times	LLNA+	LLNA–	sum
1	1	1	*137.3*	*7.0*	*144.3*
1	1	0	*107.1*	*22.5*	*129.5*
1	0	1	*31.3*	*7.0*	*38.3*
1	0	0	*24.4*	*22.5*	*46.8*
0	1	1	*40.3*	*9.0*	*49.2*
0	1	0	*31.4*	*28.5*	*59.9*
0	0	1	*9.2*	*9.0*	*18.1*
0	0	0	*7.2*	*28.5*	*35.7*
			388.0	*134.0*	*522.0*

RAW DATA

LLNA+	smarts	derekfw	times
1	300	316	218
0	88	72	170

smarts	derekfw	times	LLNA+	LLNA–	sum
1	1	1	187	21	208
1	1	0	84	21	105
1	0	1	12	4	16
1	0	0	17	13	30
0	1	1	17	4	21
0	1	0	28	21	49
0	0	1	2	3	5
0	0	0	41	47	88
			388	134	522

What kind of models yield more flexibility??

B. Statistical Modeling

1. Loglinear modeling, Poisson and Multinomial GLMs
2. First order Poisson model implies independence
3. Interactions: a 2nd order Poisson regression
4. Fit diagnostics: P-values
5. Fit diagnostics: Deviance
6. Model Selection with Information Criteria

B.1. Loglinear modeling, Poisson and Multinomial GLMs

- Contingency Table modeling very well developed
  - Bishop et al. (1975) *Discrete Multivariate Analysis*
  - Fienberg (1980) *The Analysis of Cross-Classified Categorical Data*
  - Introductory textbooks on GLMs: *Generalized Linear Models*
- GLMs generalize regression procedures to non-Gaussian distributions, e.g. Poisson and Multinomial

B.1. Loglinear modeling, Poisson and Multinomial GLMs (2)

- The number of chemicals in the upper pane may be conceived as 8 (= 4 x 2) **Poisson random variables**
- The Cooper Stats in the lower panel can be thought of as 2 **Multinomial distributions**, one for each class

smarts	derekfw	LLNA+	LLNA–	sum
1	1	271	42	313
1	0	29	17	46
0	1	45	25	70
0	0	43	50	93
		388	134	522

JOINT COOPER STATS FROM THE JOINT TABLE

Conditional Probs: Pr (smarts, derekfw | LLNA)

smarts	derekfw	LLNA+	LLNA–	WoE+
1	1	0.698	0.313	3.48
1	0	0.075	0.127	-2.30
0	1	0.116	0.187	-2.06
0	0	0.111	0.373	-5.27
		1.000	1.000	

B.1. Loglinear modeling, Poisson and Multinomial GLMs (3)

Three closely related forms, essentially leading to the same fits:

- **Loglinear models** focus on the logarithm of the expected counts
- **Poisson regressions** link the log-counts to the explanatory variables
- **Multinomial regressions** do the same, assuming the totals are given, so fractions add to 1.0

smarts	derekfw	LLNA+	LLNA–	sum
1	1	271	42	313
1	0	29	17	46
0	1	45	25	70
0	0	43	50	93
		388	134	522

JOINT COOPER STATS FROM THE JOINT TABLE

Conditional Probs: Pr (smarts, derekfw | LLNA)

smarts	derekfw	LLNA+	LLNA–	WoE+
1	1	0.698	0.313	3.48
1	0	0.075	0.127	-2.30
0	1	0.116	0.187	-2.06
0	0	0.111	0.373	-5.27
		1.000	1.000	

B.2. First order Poisson model implies independence

$$\begin{cases} nLLNA^{+} = e^{\beta_0+\beta_1\cdot S+\beta_2\cdot D+\beta_3\cdot T} = e^{\beta_0}\cdot e^{\beta_1\cdot S}\cdot e^{\beta_2\cdot D}\cdot e^{\beta_3\cdot T} \\ nLLNA^{-} = e^{\gamma_0+\gamma_1\cdot S+\gamma_2\cdot D+\gamma_3\cdot T} = e^{\gamma_0}\cdot e^{\gamma_1\cdot S}\cdot e^{\gamma_2\cdot D}\cdot e^{\gamma_3\cdot T} \end{cases}$$

Modeled Joint QSAR Data: Independence

LLNA+	smarts	derekfw	times
1	*300.0*	*316.0*	*218.0*
0	*88.0*	*72.0*	*170.0*

smarts	derekfw	times	LLNA+	LLNA–	sum
1	1	1	*137.3*	*7.0*	*144.3*
1	1	0	*107.1*	*22.5*	*129.5*
1	0	1	*31.3*	*7.0*	*38.3*
1	0	0	*24.4*	*22.5*	*46.8*
0	1	1	*40.3*	*9.0*	*49.2*
0	1	0	*31.4*	*28.5*	*59.9*
0	0	1	*9.2*	*9.0*	*18.1*
0	0	0	*7.2*	*28.5*	*35.7*
			388.0	*134.0*	*522.0*

RAW DATA

LLNA+	smarts	derekfw	times
1	300	316	218
0	88	72	170

smarts	derekfw	times	LLNA+	LLNA–	sum
1	1	1	187	21	208
1	1	0	84	21	105
1	0	1	12	4	16
1	0	0	17	13	30
0	1	1	17	4	21
0	1	0	28	21	49
0	0	1	2	3	5
0	0	0	41	47	88
			388	134	522

B.3. Interactions: a 2nd order Poisson regression

$$\begin{cases} nLLNA^{+} = e^{\beta_0+\beta_1\cdot S+\beta_2\cdot D+\beta_3\cdot T+\beta_{12}\cdot S\cdot D+\beta_{13}\cdot S\cdot T+\beta_{23}\cdot D\cdot T} \\ nLLNA^{-} = e^{\gamma_0+\gamma_1\cdot S+\gamma_2\cdot D+\gamma_3\cdot T+\gamma_{12}\cdot S\cdot D+\gamma_{13}\cdot S\cdot T+\gamma_{23}\cdot D\cdot T} \end{cases}$$

Modeled Joint QSAR Data: 2nd Order Dependent

LLNA+	smarts	derekfw	times
1	300.0	316.0	218.0
0	88.0	72.0	170.0

smarts	derekfw	times	LLNA+	LLNA–	sum
1	1	1	189.2	20.9	210.1
1	1	0	81.8	21.1	102.9
1	0	1	9.8	4.1	13.9
1	0	0	19.2	12.9	32.1
0	1	1	14.8	4.1	18.9
0	1	0	30.2	20.9	51.1
0	0	1	4.2	2.9	7.1
0	0	0	38.8	47.1	85.9
			388.0	134.0	522.0

RAW DATA

LLNA+	smarts	derekfw	times
1	300	316	218
0	88	72	170

smarts	derekfw	times	LLNA+	LLNA–	sum
1	1	1	187	21	208
1	1	0	84	21	105
1	0	1	12	4	16
1	0	0	17	13	30
0	1	1	17	4	21
0	1	0	28	21	49
0	0	1	2	3	5
0	0	0	41	47	88
			388	134	522

B.4. Fit diagnostics: P-values (**Linear**)

- We have two **Linear** models, one for **LLNA+** and one for **LLNA-**, so we have two sets of coefficient estimates: intercept and 3 linear terms
- P-values are the classic indicators of significance of each coefficient, separately. Interpret with caution!
- For **LLNA+** the linear model coefficients seem OK
- For **LLNA-** both Smarts and DerekfW are "insignificant"

	LLNA+					LLNA–			
Term	Estimate	Std Error	P-Value	Signif	Term	Estimate	Std Error	P-Value	Signif
1	1.97	0.16			1	3.35	0.15		
S	1.23	0.12	4.7E-24	***	S	-0.24	0.17	1.7E-01	---
D	1.48	0.13	9.7E-30	***	D	0.00	0.17	1.0E+00	---
T	0.25	0.10	1.5E-02	*	T	-1.16	0.20	1.1E-08	***

B.4. Fit diagnostics: P-values (2nd Order)

- Again, we have two, **both 2nd Order** models, one for **LLNA+** and one for **LLNA-**, so we have two sets of 7 coefficients: intercept, 3 linear terms, and 3 2nd order terms
- The coefficients of the linear terms are *negative*; the 2nd order interaction terms have *positive* coefficients

	LLNA+					LLNA–			
Term	Estimate	Std Error	P-Value	Signif	Term	Estimate	Std Error	P-Value	Signif
1	3.66	0.16			1	3.85	0.14		
S	-0.70	0.26	6.4E-03	**	S	-1.29	0.30	1.6E-05	***
D	-0.25	0.23	2.7E-01	---	D	-0.81	0.25	1.4E-03	**
T	-2.22	0.37	1.2E-09	***	T	-2.78	0.49	1.4E-08	***
S.D	1.70	0.31	3.4E-08	***	S.D	1.30	0.40	1.1E-03	**
S.T	1.55	0.30	3.2E-07	***	S.T	1.63	0.50	1.1E-03	**
D.T	1.51	0.34	9.8E-06	***	D.T	1.15	0.51	2.3E-02	*

B.5. Fit diagnostics: Deviance (1)

- *Deviance* is the 'Sum of Squares' of discrete GLMs

$$D = 2 \cdot \sum observed \cdot \log_e \left(\frac{observed}{fitted} \right)$$

- If the fit is exact (*Saturated model*), the *Deviance* is zero
- Simpler models have larger *Deviances*
- *Null model*, only a constant, has the largest *Deviance*
- If an extended model has one more parameter, the Likelihood Ratio Test dictates that the extra term is 'significant', if the *Deviance Drop* exceeds 5.0, i.e. ChiSquare value for 1 df at 97.5%

B.4. Fit diagnostics: Deviance (**Linear**)

- Deviance Table displays Deviance Drops for each term added to the model. If below 5.0, then this term is not significant
- The Residual Deviances, 140.9 for **LLNA+**, and 45.3 for **LLNA-**, indicate that both linear models may benefit from extension

	LLNA+		
Term	Deviance	Resid DF	Resid Deviance
1		7	434.9
S	122.4	6	312.5
D	165.6	5	146.9
T	6.0	4	140.9

	LLNA–		
Term	Deviance	Resid DF	Resid Deviance
1		7	85.7
S	1.9	6	83.7
D	0.0	5	83.7
T	38.4	4	45.3

B.4. Fit diagnostics: Deviance (2nd Order)

- Deviance Table displays Deviance Drops for each term added to the model. If below 5.0, then this term is not significant
- The Residual Deviances, 2.8 for **LLNA+**, and 0.0 for **LLNA-**, indicate that both 2nd order models fit very well
- In fact, the insignificant terms (red) could be removed, to yield a nonhierarchical model with some lower order terms lacking

	LLNA+				LLNA–		
Term	Deviance	Resid DF	Resid Deviance	Term	Deviance	Resid DF	Resid Deviance
1		7	434.9	1		7	85.7
S	122.4	6	312.5	S	1.9	6	83.7
D	165.6	5	146.9	D	0.0	5	83.7
T	6.0	4	140.9	T	38.4	4	45.3
S.D	59.7	3	81.2	S.D	19.4	3	25.9
S.T	56.8	2	24.4	S.T	20.4	2	5.5
D.T	21.5	1	2.8	D.T	5.5	1	0.0

B.6. Model Selection with Information Criteria

- AIC, BIC, modified AIC, etc. all variations of the Akaike Information Criterion, adapted by Bozdogan (1987)

$$-2 \cdot LogLikelihood + nPar \cdot [1 + \log_e(nCells)]$$

- The second term is proportional to the number of parameters (terms) and protects from overfitting
- 3 QSARs: *nCells* = 2^3 = 8, hence the 'penalty' is approximately *nPar*.3

B.6. Model Selection with Information Criteria (2)

- There are 128 possible models with 3 QSARs. One chooses the model with the best (lowest) modified AIC.
- This leaves the **LLNA-** model unchanged (7 parameter full 2nd order), but the procedure drops the linear Derekfw term from the model for **LLNA+** reducing the number of parameters to 6 (cf. its P-value on slide #16)

	LLNA+		
Term	Deviance	Resid DF	Resid Deviance
1		7	434.9
S	122.4	6	312.5
D	165.6	5	146.9
T	6.0	4	140.9
S.D	59.7	3	81.2
S.T	56.8	2	24.4
D.T	21.5	1	2.8

	LLNA+		
Term	Deviance	Resid DF	Resid Deviance
1		7	434.9
S	122.4	6	312.5
T	6.0	5	306.5
S.D	225.3	4	81.3
S.T	56.8	3	24.4
D.T	20.4	2	4.0

B.6. Model Selection with Information Criteria (3): Ames from 4 Rules

				Ames (MODEL)				Ames (DATA)		
Kazius	BenigniBossa	UFZ NN	CAESAR	1	0	Total		1	0	Total
1	1	1	1	***1245.4***	***16.0***	1261.4		**1247**	**16**	1263
1	1	1	0	***1.1***	***14.0***	15.1		**1**	**14**	15
1	1	0	1	***53.6***	***44.0***	97.6		**52**	**44**	96
1	1	0	0	***0.0***	***21.0***	21.0		**0**	**21**	21
1	0	1	1	***73.2***	***5.2***	78.4		**77**	**6**	83
1	0	1	0	***0.3***	***2.1***	2.3		**0**	**2**	2
1	0	0	1	***12.3***	***5.8***	18.1		**10**	**5**	15
1	0	0	0	***0.0***	***29.9***	30.0		**0**	**30**	30
0	1	1	1	***37.4***	***0.0***	37.4		**35**	**0**	35
0	1	1	0	***1.1***	***3.2***	4.3		**2**	**2**	4
0	1	0	1	***1.6***	***0.2***	1.8		**4**	**0**	4
0	1	0	0	***0.0***	***46.6***	46.6		**0**	**48**	48
0	0	1	1	***73.2***	***1.8***	75.0		**70**	**1**	71
0	0	1	0	***0.3***	***27.7***	28.0		**0**	**29**	29
0	0	0	1	***12.3***	***2.0***	14.3		**14**	**3**	17
0	0	0	0	***0.0***	***404.5***	404.5		**0**	**403**	403
			Total	***1512.0***	***624.0***	2136.0		1512	624	2136

Pers. comm.: Buist, Benfenati, Benigni, Schüürmann, Kühne, and others

C. Conclusions

1. The one-test Bayesian paradigm for ‘toxicological diagnosis’ can be generalized to a battery setup through joint Cooper statistics (sensitivity, specificity) for cross-tabulated model and test results
2. However, direct use of the raw counts are hampered by sparsely filled cells and missing data
3. The Naïve Bayes approach is based on the independence of the one-predictor Cooper statistics, but predictors, when developed for the same endpoint, cannot be assumed to be independent
4. Model selection techniques of discrete multivariate analysis allow the modeling of the dependence structure of multiple predictors on the basis of Poisson and Multinomial regression.