LogXact® 8

with Cytel Studio™

The Fastest and the Most Powerful Exact Logistic Regression Package Available

New Milestone in Predictive Modeling

Cytel helped pioneer exact methods for binary logistic regression and multinomial regression. Culminating over 20 years developing exact statistics and logistic regression analysis tools, LogXact® 8 represents a new milestone in predictive modeling

The eighth major release of LogXact's adds important new methods with the functionalities most requested by users.

Firth's PMLE procedure and profile penalized likelihood confidence intervals

LogXact 8 is first to offer this major improvement in logistic regression modeling: Firth's Penalized Maximum Likelihood Estimate (Biometrika [1993], 80, 1, pp27 – 38). This remarkable procedure effectively corrects for separation bias, while exhibiting lower mean square error than usual Maximum Likelihood (ML) estimations.

The PMLE procedure in LogXact 8 more accurately estimates coefficients by finding the maximum of a penalized likelihood function and adding a penalty term to the log-likelihood. The penalty term equals the logarithm of the square root of the determinant of the information matrix. Heinze and Schemper's extensive simulation shows PMLE's reliably provide better logistic regression estimate of coefficients in data sets exhibiting separation (see example within).

New LogXact 8 also computes the profile likelihood based confidence intervals with or without penalizing the likelihood by Jeffreys' invariant prior.

Exact prediction with logistic regression

The inverse logit is known as the probability of success, which can be evaluated using the ML estimates of the parameters.

The ML estimates are biased to the order n-1 where n is the sample size; so smaller sample sizes will produce estimates with more bias. In this situation exact methods are the better alternative and LogXact's Exact Prediction utilizes exact estimates for the most accurate predictions possible.

Exact Poisson regression with or without stratification

The asymptotic estimates of unstratified Poisson model is computed in many commercial software packages. But only in LogXact is the estimates of a stratified Poisson model computed, and the only package that produces exact estimates of this model - with or without stratification.

Adjacent-Categories Logits using Exact Methods

In fitting a polytomous regression where the response variable is categorical and ordered, models using adjacent-categories logit are often preferred. The only software that produces these exact estimates is LogXact 8. See Hirji (JASA, 87: 487-492, 1992)

For multi-category responses, LogXact offers fitting baseline-category logit models, proportional odds model, and continuation ratio models.

GLM in Presence of Missing Categorical Covariates

Accurate missing value analysis is another reason why so many statisticians rely on LogXact. When presented with missing discrete covariates, LogXact 8 employs Chen and Ibrahim's EM algorithm to produce GLM parameter estimates for the covariates. Logit, Probit, CLogLog, Poisson, and Normal links are all supported.

LogXact® 8 Features

Improved syntax of command language and viewing tools in Cytel Studio 8

Penalized Maximum Likelihood Estimate (PMLE) method to handle both complete separation and quasi Separation

The profile likelihood (penalized likelihood) based confidence intervals as proposed by Venzon and Moolgavkar (Applied Statistics 1988; 37: 87-94)

Exact inference for logistic regression, exact prediction, Post-fit diagnostics, such as classification table and ROC curve.

Conditional likelihood inference for matched case-control data under general M;N matching

Exact inference for Poisson regression and stratified Poisson regression.

Exact estimates for unordered category model, and adjacent category model

Proportional odds model

Generalized linear models in presence of missing categorical covariates

Time-saving Monte Carlo procedures for fast exact inference in large data sets

Cytel
STATISTICAL SOFTWARE & SERVICES

LogXact 8
with Cytel Studio™

New tests answer difficult statistical questions

Why Firth's Penalized Maximum Likelihood Method (PMLE) and the CI using the penalized profile likelihood method?

"…These studies and others, which have not yet been published, and the analysis of the examples provided by the present paper confirm the penalized maximum likelihood approach to be an easy-to-use method in the analysis of logistic regression problems when conventional asymptotic methods are in doubt and exact results are unavailable or unnecessarily conservative." - Heinze G (2006)

When visualizing data as points in covariate space, "separation" occurs if it's possible to find a hyper-plane separating responses from non-responses. Separation phenomenon is especially common in small and medium-sized datasets. See [1] Paul Allison's explanation of separation and its occurrence in practice vML estimates for coefficients are infinite for datasets exhibiting separation. This is undesirable — it's very unlikely that a covariate is a perfect response predictor.

Previously with separation LogXact's EX procedure computed the Median Unbiased Estimator (MUE) for the coefficient. While the estimates were finite, practitioners remarked they were sometimes unreliable.

When the logistic regression model is used on small or sparse data, then the inference based on confidence intervals is relied upon more than the inference based on the point estimates. Heinze and Schemper (Statistics in Medicine, 2002; 21: 2409-2419), Heinze (Statistics in Medicine 2006; 25: 4216-4226) have shown that the confidence interval using the penalized profile likelihood method based on Venzon and Moolgavkar (Applied Statistics 1988; 37: 87-94) shows better coverage probability than the exact method.

Only LogXact 8 provides this exceptional method for estimating coefficients in unstratified logistic regression: Firth's Penalized Maximum Likelihood Estimates (PMLE) procedure.

Example 1: Erythrocyte Sedimentation Rate (ESR) Study

Consider a 46-patient study of non-metastatic osteogenic sarcoma by Goorin, Perz-Atayde, Gebhardt, and Andersen. They sought the predictors for a three-year disease-free interval (DFI3), with covariates of interest: GENDER, osteoid pathology (AOP), and lymphocytic infiltration (LYINF).

This dataset clearly exhibits separation, therefore the ML estimates for a logistic regression model with these covariates does not exist (see LogXact 8 output, below). Notice that EX gave conditional MLE estimates for GENDER and AOP, but the conditional MLE estimate for LYINF doesn't exist, so a MUE was computed.

Parameter Estimates

Model Term	Point Estimate			Confidence Interval and P-Value for Beta			
					95 %CI		2*1-sided
	Type	Beta	SE(Beta)	Type	Lower	Upper	P-Value
%Const	MLE	?	?	Asymptotic	?	?	?
LYINF	MLE	?	?	Asymptotic	?	?	?
	MUE	-1.886	NA	Exact	-INF	0.1615	0.07418
GENDER	MLE	?	?	Asymptotic	?	?	?
	CMLE	-1.548	0.8884	Exact	-4.024	0.3627	0.1392
AOP	MLE	?	?	Asymptotic	?	?	?
	CMLE	-1.156	0.7506	Exact	-2.997	0.5114	0.2182

Selecting LogXact 8's Firth's PMLE procedure provides the following vastly improved output:

Parameter Estimates

Model Term	Point Estimate			Confidence Interval and P-Value for Beta			
					95 %CI		2*1-sided
	Type	Beta	SE(Beta)	Type	Lower	Upper	P-Value
%Const	PMLE	4.29	1.663	Asymptotic	1.031	7.55	0.009893
				Asymptotic(Profile)	1.814	9.324	
LYINF	PMLE	-2.461	1.553	Asymptotic	-5.504	0.5818	0.1129
				Asymptotic(Profile)	-7.363	-0.1887	
	MUE	-1.886	NA	Exact	-INF	0.1615	0.07418
GENDER	PMLE	-1.415	0.8441	Asymptotic	-3.07	0.2392	0.09361
				Asymptotic(Profile)	-3.251	0.1151	
	CMLE	-1.548	0.8884	Exact	-4.024	0.3627	0.1392
AOP	PMLE	-1.104	0.7316	Asymptotic	-2.538	0.3299	0.1313
				Asymptotic(Profile)	-2.604	0.2807	
	CMLE	-1.156	0.7506	Exact	-2.997	0.5114	0.2182

PMLE: Penalized MLE for bias correction (Firth's method).

Notice the penalized ML estimates using Firth's method produces all finite coefficients values. The point estimates are quite close to CMLE estimates computed by EX for GENDER and AOP. There is an appreciable difference for the LYINF coefficient. Simulation experiments suggest the Firth estimate is a more reliable point estimate than the MUE.

Additionally the output shows the Profile Likelihood Confidence Intervals which are in general more reliable than the confidence intervals produced by other methods.

$$P_1(\bar{R}_j^{opt}) \geq P_1(\bar{R}_j) \text{ for all } j = 1, \ldots, K$$

EXAMPLE 2: Missing Categorical Covariates Method – Logistic Regression Model

This dataset is from a longitudinal study of air pollution's health effects. The binary response is the mother's wheezing status,and a binary covariate of main interest is the city of residence. Two more binary covariates indicate socioeconomic status and the child's previous medical condition. We considered a data subset comprising 2106 baseline subjects, where baseline wheezing status is non-missing.

Six Cities data (partially shown)

smoke	soc	cond	Count
.	.	.	17
.	0	0	23
.	0	1	2
.	1	0	7
0	.	.	454
0	0	0	829
0	0	1	128
0	1	0	141
1	.	.	160
1	0	0	233
1	0	1	54
1	1	0	58

	Wheeze	city	smoke	soc	cond	var
1	0	1	0	0	0	
2	0	0	0	.	.	
3	1	1	0	1	0	
4	1	1	1	0	0	
5	0	0	0	.	.	
6	1	0	0	.	.	
7	0	1	.	1	0	
8	0	0	0	0	0	
9	1	1	0	0	0	
10	0	0	0	0	0	

In the results, the complete case analysis (omitting all the missing observations) shows:

- the covariate "smoke" is not a significant factor (p=0.2788)

- there is no significant interaction between covariates "smoke" and "soc" (p=0.0584)

However, the result from LogXact 8, under missing data analysis, shows quite differently:

- the covariate "smoke" is significant at 90% confidence level (p=0.087)

- the interaction between variables "smoke" and "soc" is significant (p=0.0494)

Regression with Categorical Missing Covariates				
Variables	smoke	soc	cond	smoke*soc
Missing %	2.30%	30%	30%	30%

	LogXact 7 with Cytel Studio						SAS's PROC MI and PROC MIANALYZE					
	Complete Case			Missing Data			Output-1+			Output-1++		
	Beta	Std_Err	P-value	Beta	Std_Err	P-value	Beta	Std_Err	P-value	Beta	Std_Err	P-value
% Const	-2.2971	0.1290	0.0000	-2.2028	0.1167	0.0000	-2.0979	0.1953	<0.0001	-2.0815	0.1911	<0.0001
city	0.5533	0.1449	0.0001	0.3369	0.1249	0.0070	0.2897	0.1178	0.0139	0.2895	0.1169	0.0133
smoke	0.2048	0.1891	0.2788	0.2932	0.1713	0.0870	0.3240	0.1523	0.0335	0.2829	0.1599	0.0770
soc	2.4311	0.2033	0.0000	2.4011	0.2009	0.0000	2.2504	0.2478	<0.0001	2.2443	0.2531	<0.0001
cond	2.9690	0.1933	0.0000	2.9657	0.1915	0.0000	2.2479	0.6676	0.0010	2.2207	0.6425	0.0008
smoke*soc	-0.6982	0.3689	0.0584	-0.7228	0.3679	0.0494	-0.5327	0.3356	0.1126	-0.4814	0.3378	0.1543

\+ Imputation sequence is soc-cond-smoke, # of imputation=100
++ Imputation sequence is smoke-soc-cond, # of imputation=100

For imputing binary covariate, we used "Logistic" option in PROC MI. The "Logistic" option requires monotonic missing pattern in the data. In our data, the missing pattern is non-monotone. We employed a common sequential imputation method. Notice the covariates "soc" and "cond" are always missing together, so we impute "cond" and "soc" together.

We used two different sequences. In both, the SAS output shows the interaction term "smoke*soc" is highly non-significant. In SAS Output-1, the covariate "smoke" is significant, but it is not significant in Output-2. The covariate "city" is not significant at the 99% confidence level. However, it is significant in complete case analysis — and in LogXact's missing data analysis.

Referenced Papers

1.Allison,P. (2004) Convergence Problems in Logistic Regression Chapter 10 (pp 238-252) in Numerical Issues in Statistical Computing for the Social Scientist, Altman, Gill, and McDonald(ed.) Wiley-Interscience (2004).

2.Firth,D. (1993) Bias Reduction of Maximum Likelihood Estimates. Biometrika,80,27-38.

3.Heinze,G.and Schemper,M. (2002) A solution to the problem of separation in logistic regression. Statistics inMedicine,21,2409-2419.

4. Heinze, G (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. Statistics in Medicine, 25, 4216-4226.

5.Ibrahim,J.G.(1990),"Incomplete Data in Generalized Linear Models", JASA,85,765-769.

6. Lipsitz,S.R.and Ibrahim,J.G. (1996a),``A Conditional Model For Incomplete Covariates in Parametric Regression Models", Biometrika,83,916-922.

EXAMPLE 3: Missing Categorical Covariates Method – Quality of Life Study – Multiple Linear Regression Model

The data from a "quality of life" study were analyzed using Missing Categorical Covariates method introduced in LogXact 7. The results:

Multiple Linear Regression (with missing categorical covariates)

regression (type=mlr, model(ltime = age trt_e trt_f trt_g lang phy1 phy2), estimate(age trt_e trt_f trt_g lang phy1 phy2), method=asym

Basic Information

Data file	QOLm.cyd
Model	ltime=%Const+age+trt_e+trt_f+trt_g+lang+PHY1+PHY2
Analysis type	Estimate :: Asymptotic
Number of terms in model	8
Number of term(s) dropped	0
Number of observations in analysis	402
Number of records rejected	0
Number of groups	402
Terms with missing values	2
EM iterations used	33

Summary of Covariates with Missing Values

Covariate	Count	%Missing
PHY1	51	12.69
PHY2	124	30.85
One of the term	155	38.56

Parameter Estimates

	Point Estimate			Confidence Interval and P-Value for Beta			
					95 %CI		2*1-sided
Model Term	Type	Beta	SE(Beta)	Type	Lower	Upper	P-Value
%Const	MLE	6.381	0.4931	Asymptotic	5.415	7.348	0.0000
age	MLE	0.003473	0.0078	Asymptotic	-0.01182	0.01876	0.6561
trt_e	MLE	-0.2823	0.1309	Asymptotic	-0.5389	-0.02565	0.03109
trt_f	MLE	-0.2882	0.1335	Asymptotic	-0.5499	-0.02656	0.03086
trt_g	MLE	-0.1409	0.1393	Asymptotic	-0.4139	0.132	0.3116
lang	MLE	-0.2516	0.1099	Asymptotic	-0.467	-0.03619	0.02206
PHY1	MLE	0.2531	0.2171	Asymptotic	-0.1725	0.6787	0.2438
PHY2	MLE	-0.2645	0.2272	Asymptotic	-0.7099	0.1808	0.2444
Tau	MLE	1.145	0.08102	Asymptotic	0.9859	1.304	0.0000

A comparison of analyses with "complete cases", and with "all cases" (including cases with missing covariates):

	Using Complete Cases		Using All Cases (missing covariates method)	
Parameter	**Beta**	**P-value**	**Beta**	**P-Value**
% C	5 9	0	6 3813	0
age	0.009339	0.2227	0.0035	0.6561
trt_e	-0.04467	0.7378	-0.2823	0.0311
trt_f	-0.03466	0.796	-0.2882	0.0309
trt_g	-0.09303	0.4979	-0.1409	0.3116
lang	-0.1327	0.2224	-0.2516	0.0221
PHY1	0.2063	0.2888	0.2531	0.2438
PHY2	-0.1861	0.3422	-0.2645	0.2444
a			1 1447	0

About Cytel

Founded in 1987, Cytel Inc. is the leading provider of specialized statistical applications, and clinical trial design methodologies and services for the bio-pharmaceutical industry. Founders Drs. Mehta and Patel are Fellows of the American Statistical Association and have published more than 100 papers in refereed journals.

Are You a SAS User?

Access LogXact® PROCs Directly Within SAS

For all the features of SAS, LogXact PROCs solves problems that SAS alone cannot. Examine for yourself at: www.cytel.com/Products/LogXact/examples.asp

LogXact PROCs and SAS – A PowerfulCombination, An Unbeatable Pairing

Seamless integration between LogXact PROCs and SAS's latest version isn't just a claim, it's a fact thanks to Cytel's renowned developers and technical support.

No more application shuttling, no more troublesome data transfer between programs. Immediate access to Cytel's world-recognized exact methods within their SAS environment is why so many statisticians choose LogXact 8 PROCs.

Not Found in SAS, LogXact PROCs Delivers:

- Exact inference for other GLM regression, i. e., Poisson model and Adjacent category model
- Exact prediction for binary logistic regression
- The new PROC XMISS allows fitting GLM in presence of missing categorical covariates

Cytel STATISTICAL SOFTWARE & SERVICES

Cytel Inc.
675 Massachusetts Avenue
Cambridge, MA 02139-3309

T: +1 617.661.2011
F: +1 617.661.4405
sales@cytel.com
www.cytel.com