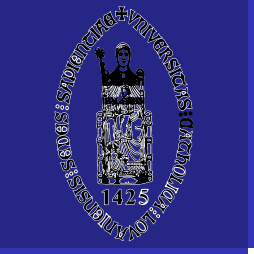


Model selection and parameter estimation with missing covariates in logistic regression models

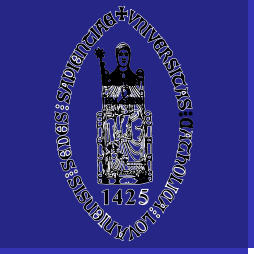
Workshop on Model Selection 2008

Fabrizio Consentino & Gerda Claeskens

ORSTAT and Leuven Statistics Research Center
Katholieke Universiteit Leuven

24 July 2008

Overview



Introduction

Problems

- Model Selection:
  Searching for the 'best' model in order to explain the phenomena of interest.
- Missing Data:
  Presence of missing observations in the data sets of interest.

Missing Data

Missingness patterns:
Structure of the missing observations

$$M = \begin{cases} 1 & X \text{ observed} \\ 0 & \text{otherwise} \end{cases}$$

Missingness mechanisms:
To describe the missing indicator M

- *Missing at random* (MAR)
  $\Rightarrow f(\mathbf{M}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}; \boldsymbol{\theta}) = f(\mathbf{M}|\mathbf{X}_{\text{obs}})$

Assumptions

- Response variable $\mathbf{Y}$ fully observed
- Design matrix $\mathbf{X}$ contains missing values
- It can be partitioned in $\mathbf{X} = (\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})$
- **MAR** assumption
- $f(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}) = f(\mathbf{Y}|\mathbf{X}; \boldsymbol{\beta}) f(\mathbf{X}; \boldsymbol{\alpha})$

Method of Weights - EM algorithm

Introduced by Ibrahim (1990), it is used in missing covariates data.

- It provides a weighted log-likelihood function in the E-step:

$$Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \int w_i \log f(y_i, x_i; \boldsymbol{\theta})\, dx_{\text{mis},i}$$
$$\text{with } w_i = f(x_{\text{mis},i}|x_{\text{obs},i}, y_i; \boldsymbol{\theta}^{(k)})$$

- $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \sum_{i=1}^{n} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ is evaluated with a Monte Carlo EM algorithm and a Gibbs sampler along with the adaptive rejection algorithm of Gilks and Wild (1992) for sampling from $(x_{\text{mis},i}|x_{\text{obs},i}, y_i; \boldsymbol{\theta}^{(k)})$
- $f(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}) = f(\mathbf{Y}|\mathbf{X}; \boldsymbol{\beta}) f(\mathbf{X}; \boldsymbol{\alpha})$

$$Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \int w_i \log f(y_i|x_i; \boldsymbol{\beta})\, dx_{\text{mis},i} + \int w_i \log f(x_i; \boldsymbol{\alpha})\, dx_{\text{mis},i}$$
$$= Q_i^{(1)}(\boldsymbol{\beta}|\boldsymbol{\theta}^{(k)}) + Q_i^{(2)}(\boldsymbol{\alpha}|\boldsymbol{\theta}^{(k)})$$

Model Selection Criteria

AIC and modifications

- One of the most popular criteria.
- AIC is twice a penalized log likelihood value,

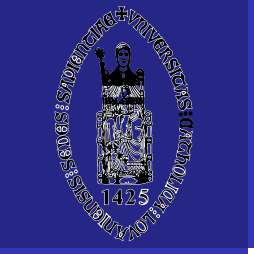
$$AIC = 2 \log L_n(\hat{\theta}) - 2 \text{ length}(\theta)$$

- Modification in the penalty term:
  Takeuchi: $\hat{p} = tr(\widehat{I}^{-1}\widehat{J})$
  Hurvich and Tsai: $\hat{p} = 2 \text{ length}(\theta) \frac{\text{n}}{\text{n}- \text{ length}(\theta)-1}$

Model Selection Criteria

Derivation

- Kullback-Leibler distance:
  $KL(g, f_\theta) = E_g[\log\{g(\mathbf{Y}, \mathbf{X})/f(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta})\}]$
- "Adjusted" likelihood function $\log \tilde{f}_\theta(\mathbf{y}, \mathbf{x}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta})$
- $Q(\boldsymbol{\theta}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \int \log f(y_i, x_{\text{obs},i}, x_{\text{mis},i}; \boldsymbol{\theta}) f(x_{\text{mis},i}|x_{\text{obs},i}, y_i, \boldsymbol{\theta}) dx_{\text{mis},i}$
- Kullback-Leibler distance:
  $KL(g, \tilde{f}_\theta) = [E_g\{\log g(\mathbf{Y}, \mathbf{X})\} - E_g\{\log \tilde{f}_\theta(\mathbf{Y}, \mathbf{X})\}]/n$
- $K_n = \int g(\mathbf{y}, \mathbf{x}) \int g(\tilde{\mathbf{y}}, \tilde{\mathbf{x}}) \log \tilde{f}(\tilde{\mathbf{y}}, \tilde{\mathbf{x}}; \widehat{\boldsymbol{\theta}}) d\tilde{\mathbf{y}}\, d\tilde{\mathbf{x}}\, d\mathbf{y}\, d\mathbf{x}/n$
- An estimator of K_n is $\widehat{K}_n = Q(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})/n$



Model Selection Criteria

Criteria

Following Takeuchi's information criterion (Takeuchi, 1976), we define the model robust criterion TIC for missing covariate values as

$$\text{TIC} = 2\ Q(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) - 2\ \text{tr}\{\widehat{J}(\widehat{\boldsymbol{\theta}})\widehat{I}^{-1}(\widehat{\boldsymbol{\theta}})\}$$

where

$$\widehat{I}(\widehat{\boldsymbol{\theta}}) = -\frac{1}{n}\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) \text{ and } \widehat{J}(\widehat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\dot{Q}_i(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\dot{Q}_i(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})'.$$

If the matrices I and J are equal, then the penalty in the expression of the TIC reduces to the number of parameters in the model. This simplification leads to a version of Akaike's information criterion suitable for use with missing covariate information.

$$\text{AIC} = 2\ Q(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) - 2\ \text{length}(\boldsymbol{\theta}).$$

Model Selection Criteria

Criteria

- Claeskens and Consentino (2008, Biometrics) proposed criteria using only $Q^{(1)}$

$$TIC_1 = 2\ Q^{(1)}(\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - 2\ \mathrm{tr}\{\widehat{J}(\widehat{\boldsymbol{\beta}})\widehat{I}^{-1}(\widehat{\boldsymbol{\beta}})\}$$

$$\mathrm{AIC}_1 = 2\ Q^{(1)}(\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - 2\ \mathrm{length}(\boldsymbol{\beta})$$

- 'Full' Q function: $Q^{(1)} + Q^{(2)}$
- If no missingness: $Q = \log L_n$ and $Q^{(2)} = 0$

Applications

Simulation setting

- Logistic regression model
- X_3 and X_4 generated independently from a standard normal distribution.
- X_1 and X_2 contain missing observations and are modeled using a bivariate normal regression $(X_1, X_2) \sim N_2(\boldsymbol{\mu}, \Sigma)$, $\boldsymbol{\mu}' = (\mu_{i1}, \mu_{i2})$
- The regressors are the fully observed covariates in $\mathbf{X}$
- $\mu_{it} = \alpha_{t0} + \alpha_{t1} x_{i3} + \alpha_{t2} x_{i4},\ t = 1, 2$

Simulation results

% missing x_1, x_2	Criteria	Model selection			Correctly specified	Model selection			Correctly specified
		$n = 50$				$n = 100$			
		C	O	U		C	O	U	
5% 5%	TIC_1	0.530	0.280	0.190	0.810	0.650	0.340	0.010	0.990
	AIC_1	0.563	0.223	0.214	0.786	0.653	0.333	0.014	0.986
	AIC_{orig}	0.570	0.227	0.203	0.797	0.677	0.303	0.020	0.980
	AIC_{cc}	0.527	0.253	0.220	0.780	0.680	0.300	0.020	0.980
10% 5%	TIC_1	0.503	0.297	0.200	0.800	0.630	0.360	0.010	0.990
	AIC_1	0.547	0.247	0.206	0.784	0.663	0.330	0.007	0.993
	AIC_{orig}	0.577	0.220	0.203	0.797	0.677	0.303	0.020	0.980
	AIC_{cc}	0.507	0.230	0.263	0.737	0.670	0.310	0.020	0.980
15% 15%	TIC_1	0.477	0.340	0.183	0.817	0.567	0.423	0.010	0.990
	AIC_1	0.527	0.263	0.210	0.790	0.653	0.333	0.014	0.986
	AIC_{orig}	0.577	0.220	0.203	0.797	0.677	0.303	0.020	0.980
	AIC_{cc}	0.443	0.233	0.324	0.676	0.640	0.317	0.043	0.957

Simulation results

% missing x_1, x_2	Criteria	Model selection			Correctly specified
		$n = 50$			
		C	O	U	
5% 5% **MCAR**	TIC_1	0.530	0.280	0.190	0.810
	AIC_1	0.563	0.223	0.214	0.786
	AIC_{orig}	0.570	0.227	0.203	0.797
	AIC_{cc}	0.527	0.253	0.220	0.780
5% 5% **MAR**	TIC_1	0.433	0.423	0.144	0.856
	AIC_1	0.470	0.340	0.190	0.810
	AIC_{orig}	0.583	0.280	0.137	0.863
	AIC_{cc}	0.437	0.193	0.370	0.630

Model Selection Criteria

Distribution selection

- Focussing on $f(\mathbf{X};\alpha)$.
- $Q^{(2)}$ used for deciding which distribution describes better the missing covariates.
- Criteria used

$$\text{TIC} = 2\ Q(\widehat{\theta}|\widehat{\theta}) - 2\ \text{tr}\{\widehat{J}(\widehat{\theta})\widehat{I}^{-1}(\widehat{\theta})\}$$

$$\text{AIC} = 2\ Q(\widehat{\theta}|\widehat{\theta}) - 2\ \text{length}(\theta).$$

$$\text{with } Q(\widehat{\theta}|\widehat{\theta}) = Q^{(1)}(\widehat{\beta}|\widehat{\theta}) + Q^{(2)}(\widehat{\alpha}|\widehat{\theta})$$

- Main drawback: computationally intense.

Estimation

Non-iterative method

Following Gao and Hui (1997) we propose an extension, for the estimation in logistic regression models where some of the covariates are missing, to the multivariate normal and t distributions.

- $\text{logit}\, P(Y_i = 1|\mathbf{X}_{\text{obs},i}, \mathbf{X}_{\text{mis},i}) = \log f(\mathbf{X}_{\text{mis},i}|\mathbf{X}_{\text{obs},i}, Y_i = 1) - \log f(\mathbf{X}_{\text{mis},i}|\mathbf{X}_{\text{obs},i}, Y_i = 0) + \text{logit}\, P(Y_i = 1|\mathbf{X}_{\text{obs},i})$
- $\text{logit}\, P(Y_i = 1|\mathbf{X}_{\text{obs},i}, \mathbf{X}_{\text{mis},i}) = \alpha_0 + \mathbf{X}_{\text{obs},i}\boldsymbol{\alpha}_1 + \mathbf{X}_{\text{mis},i}\boldsymbol{\alpha}_2$
- $\text{logit}\ P(Y_i = 1|\mathbf{X}_{\text{obs},i}) = \beta_0 + \mathbf{X}_{\text{obs},i}\boldsymbol{\beta}_1$
- $\mathbf{X}^t_{\text{mis},i} = \boldsymbol{\gamma}_0 + Y_i\boldsymbol{\gamma}_1 + \mathbf{X}_{\text{obs},i}\boldsymbol{\gamma}_2 + \boldsymbol{\epsilon}_i$

Estimation

Non-iterative method - error term

- If $\epsilon_i \sim N_q(\mathbf{0}, \boldsymbol{\Sigma})$

$$\alpha_0 = \beta_0 - \gamma_1^t \boldsymbol{\Sigma}^{-1}(2\gamma_0 + \gamma_1)$$
$$\boldsymbol{\alpha}_1 = \boldsymbol{\beta}_1 - \gamma_1^t \boldsymbol{\Sigma}^{-1}\gamma_2$$
$$\boldsymbol{\alpha}_2 = \gamma_1^t \boldsymbol{\Sigma}^{-1}$$

- If $\epsilon_t \sim t_q(\nu)$

$$\alpha_0 = \beta_0 - \left(\tfrac{\nu+q}{\nu}\right) \gamma_1^t \boldsymbol{\Sigma}^{-1}(2\gamma_0 + \gamma_1)$$
$$\boldsymbol{\alpha}_1 = \boldsymbol{\beta}_1 - \left(\tfrac{\nu+q}{\nu}\right) \gamma_1^t \boldsymbol{\Sigma}^{-1}\gamma_2$$
$$\boldsymbol{\alpha}_2 = \left(\tfrac{\nu+q}{\nu}\right) \gamma_1^t \boldsymbol{\Sigma}^{-1}$$

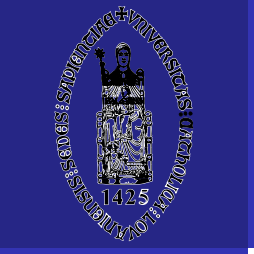
Model Selection Criteria

Distribution selection

- Based on $\mathbf{X}^t_{\text{mis},i} = \gamma_0 + Y_i\gamma_1 + \mathbf{X}_{\text{obs},i}\gamma_2 + \epsilon_i$
- For selecting the distribution we restrict attention to the use of the model for $\mathbf{X}_{\text{mis}}$ given Y and $\mathbf{X}_{\text{obs}}$. The corresponding AIC is

$$\text{AIC}_{distr} = -2\ \log\{f(\mathbf{X}_{\text{mis}}, \gamma|\mathbf{X}_{\text{obs}}, Y)\} + 2\ p_\gamma,$$

  with p_γ the number of parameters in the model.
- The smallest obtained value of this AIC indicates the best distribution for modeling the data.



Applications

Simulation setting

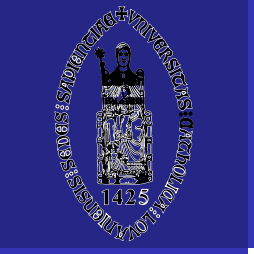
- Logistic regression model
- $X_3, \ldots, X_6$ generated independently from a standard normal distribution.
- X_1 and X_2 contain missing observations and are modeled using a bivariate normal regression and a bivariate t-distribution, with one of four different degrees of freedom df$= (5, 7, 15, 50)$
- Four different sample sizes $n = 50, 100, 200$ and 500; three different choices of percentages of missingness $(5\%, 5\%)$, $(15\%, 5\%)$ and $(30\%, 5\%)$
- For each setting we run $N = 2000$ simulations.

Applications

Simulation results

Fitted data	α_0	α_1	α_2	α_3	α_4	α_5	α_6
Normal	1.353	0.821	−0.003	−0.264	−0.560	−1.340	0.539
	(0.721)	(0.210)	(0.177)	(0.498)	(1.128)	(0.739)	(1.495)
t_{50}	1.219	0.966	−0.003	−0.114	−0.874	−1.202	0.237
	(0.700)	(0.248)	(0.183)	(0.512)	(1.239)	(0.722)	(1.528)
t_{15}	1.071	1.094	−0.004	0.053	−1.222	−1.050	−0.097
	(0.729)	(0.327)	(0.192)	(0.596)	(1.655)	(0.759)	(1.833)
t_7	0.799	1.313	−0.004	0.359	−1.860	−0.772	−0.709
	(0.925)	(0.560)	(0.210)	(0.936)	(3.220)	(0.977)	(3.124)
t_5	0.568	1.487	−0.005	0.620	−2.403	−0.534	−1.231
	(1.239)	(0.834)	(0.229)	(1.415)	(5.373)	(1.318)	(4.974)
CC	6.373	7.120	0.405	0.543	−8.138	−8.234	0.994
	(5917.983)	(4010.07)	(626.35)	(1824.39)	(11000.55)	(6845.76)	(6825.57)

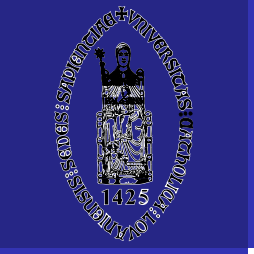
- True values: $\boldsymbol{\alpha}^t = (1, 1, 0, 0, -1, -1, 0)$



Applications

Simulation results

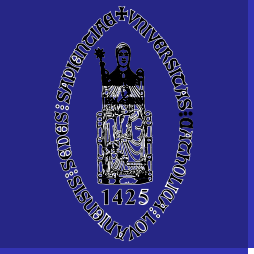
Sample Size	Simulated data	Distribution selection				
		missingness= (30%, 5%)				
		Norm	t_{50}	t_{15}	t_7	t_5
50	Norm	0.000	0.846	0.088	0.050	0.016
	t_{50}	0.000	0.805	0.110	0.062	0.023
	t_{15}	0.000	0.715	0.136	0.096	0.053
	t_7	0.000	0.518	0.185	0.156	0.141
	t_5	0.000	0.399	0.185	0.197	0.220
100	Norm	0.295	0.547	0.127	0.024	0.005
	t_{50}	0.232	0.544	0.166	0.056	0.003
	t_{15}	0.118	0.472	0.254	0.130	0.025
	t_7	0.032	0.258	0.278	0.272	0.162
	t_5	0.009	0.123	0.222	0.309	0.337
200	Norm	0.564	0.316	0.112	0.007	0.000
	t_{50}	0.440	0.328	0.214	0.019	0.000
	t_{15}	0.165	0.307	0.389	0.132	0.008
	t_7	0.022	0.088	0.335	0.406	0.148
	t_5	0.005	0.025	0.134	0.421	0.415
500	Norm	0.668	0.281	0.051	0.000	0.000
	t_{50}	0.416	0.419	0.165	0.000	0.000
	t_{15}	0.066	0.279	0.588	0.067	0.000
	t_7	0.000	0.009	0.251	0.648	0.092
	t_5	0.000	0.000	0.025	0.418	0.557



Applications

Dataset

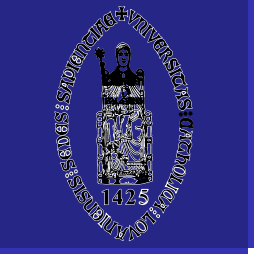
- The European Values Study (EVS) is a large-scale, cross-national and longitudinal survey research program.
- Data related to Belgium.
- 1603 observations and 6 variables.
- Binary outcome variable that indicates if the workers are satisfied with their job hours
- Variables x_1, age when education was completed, contains missing values



Applications

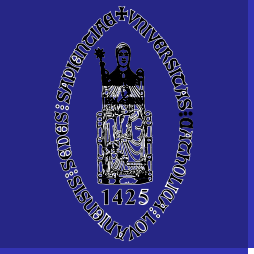
Dataset

Method	Missing Covariate Models	AIC	Goodness of fit	penalty term	Timing
			$Q^{(2)}$		
Q-function	Normal	7658.384	3820.192	9	21'42"
	t_{50}	7580.776	3781.388	9	13h59'00"
	t_{15}	7471.422	3726.711	9	17h55'45"
	t_5	7403.142	3692.571	9	21h39'55"
			LogLik		
Non iterative	Normal	7389.142	3685.571	9	2"
	t_{50}	7317.908	3649.954	9	2"
	t_{15}	7220.962	3601.481	9	2"
	t_5	7125.912	3553.956	9	2"



Conclusions

- Directly comparable with criteria with fully observed variables
- Including the missingness process provides better results for the estimation
- Ignoring the missing cases provides biased results
- The proposed criteria include the significant variables for the phenomenon under investigation
- The distribution selection criterion chooses the most suitable parametric family for fitting the missing covariates



References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, B. Petrov and F. Csáki (editors), 267–281, Akadémiai Kiadó, Budapest.
- Claeskens, G., and Consentino, F. (2008). Variable Selection with Incomplete Covariate Data. *Biometrics*, To appear.
- Gao, S., and Hui, S. L. (1997). Logistic Regression Models with Missing Covariate Values for Complex Survey Data. *Statistics in Medicine*, **16**, 2419-2428.
- Ibrahim, J.G., Chen, M.H. and Lipsitz, S.R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics*, **55**, 591-596.
- Kotz, S. and Nadarajah, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge University Press, Cambridge.