

Categorical Data Modelling

Robert Raeside

Categorical Data

In many large surveys the dependent (target) variable is often categorical in nature.

For example:

Binary: Whether has Credit Cards (like Access, Visa etc)?
No or Yes

Multinomial: e.g. type of birth assistance sought: Family only, traditional attendants, health professionals

Ordinal: e.g. how many innovations has your company produced in the last year? 0-9, 10-19, 20+

Counts: Number of child deaths experienced

Generalised linear models

- A transformation is required to allow statistical inference to be applicable.
e.g. For binary regression – use a logit transformation to linearise

$$\log it(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$

Re-transform

$$\Pr(P) = \frac{e^{\log it}}{1 + e^{\log it}}$$

Software

- SPSS – PASW18
- SAS
- STATA
- MINITAB
- MLWIN – for multilevel applications

Initial Data Preparation

- Always required
- Produce frequency tables and cross tab
- Compare means
- Consider merging categories
- Check for correlations – multicollinearity is often a major problem when building models in the social sciences
- Create new variables
- Causality - a major problematic area.

Endogeneity

Example – is a persons contraceptive practice influenced by their friends

Endogeneity – can lead to inconsistent and misleading parameter estimates.

Consequently hypothesis tests are not reliable.

One endogenous variable can seriously distort the model.

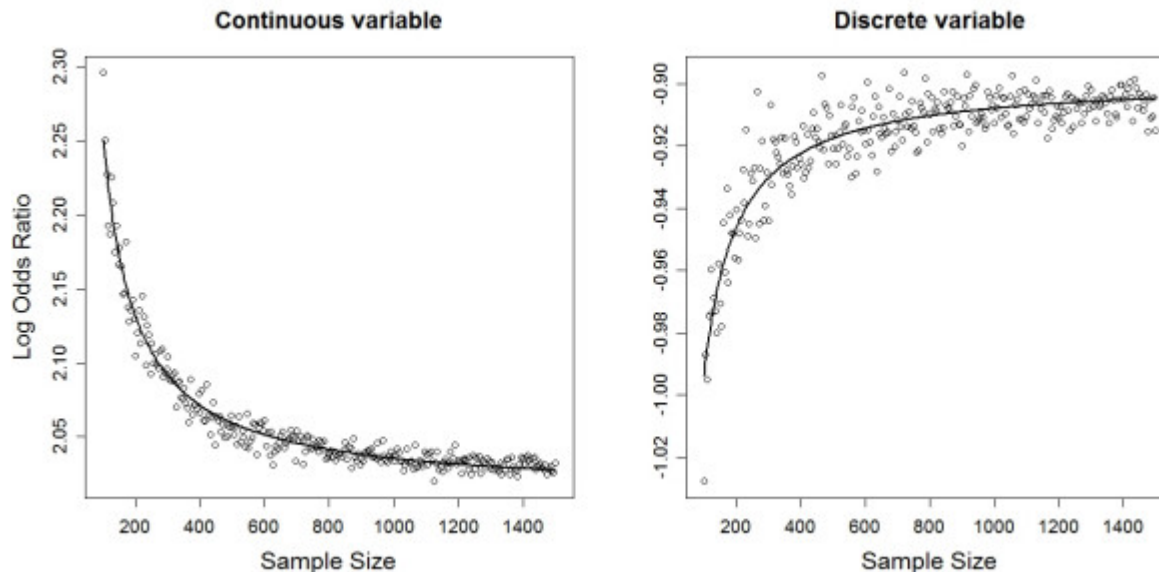
Dealing with endogeneity

- Drop endogenous variables – or substitute a proxy – but can lead to the removal of a major theoretical part of the model.
- Lag the “independent” variable so can identify cause and effect.
- In OLS can correct by using instrumental variables that are exogenous and use two stage least squares – but not available

Sample Bias

Self selecting samples – e.g. attitudes towards technology

Small samples or sub groups can distort the coefficients see Nemes et al (2009).



Coefficient estimates and its sample size dependent systematic bias in logistic regression estimates.

Solution

- The most effective solution is good research design

Different Models

- Logistic model: this is used most frequently – relies on the concept of the odds ratio.
- Probit models: – more like OLS for interpretation as predictions are z scores – easier – can compute marginal means
- Tobit models: when all of the dependent variable is not observable –allows censoring

Example – Binary Logistic

PhD Kaberi Gayen

Context: A survey of 724 women in rural Bangladesh

Data: Socio-economic, Reproductive, Family Planning and Sociometric data were collected

Method: Interviewing – with questionnaire

Target variable: currently using contraception

Independent variables: women's education level, women's age, number of children who died, husbands education, female autonomy, socioeconomic status, religion (1 if Hindu), influence of TV & radio

Data Collection



Two Women in Brhammin-Shasan and Their Children

'Ja'-Network in Kamarpara



Null model

Classification Table^{a, b}

Observed			Predicted		
			FP practice		Percentage Correct
			No	Yes	
Step 0	FP practice	No	0	220	.0
		Yes	0	466	100.0
Overall Percentage			67.9		

a. Constant is included in the model.

b. The cut value is .500

Final Predictions

Classification Table^a

Observed			Predicted		
			FP practice		Percentage Correct
			No	Yes	
Step 1	FP practice	No	33	187	15.0
		Yes	27	439	94.2
Overall Percentage			68.8		

a. The cut value is .500

Model output

Omnibus Tests of Model Coefficients

	Chi-square	df	Sig.
Step 1	44.360	9	.000
Block	44.360	9	.000
Model	44.360	9	.000

Justifies the inclusion of new variables

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	816.427 ^a	.063	.088

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

The Cox-Snell R² and Nagelkerke R² are attempts to provide a logistic analogy to R² in OLS regression.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5.615	8	.690

Compares observed probabilities to expected don't want to be significant

Model

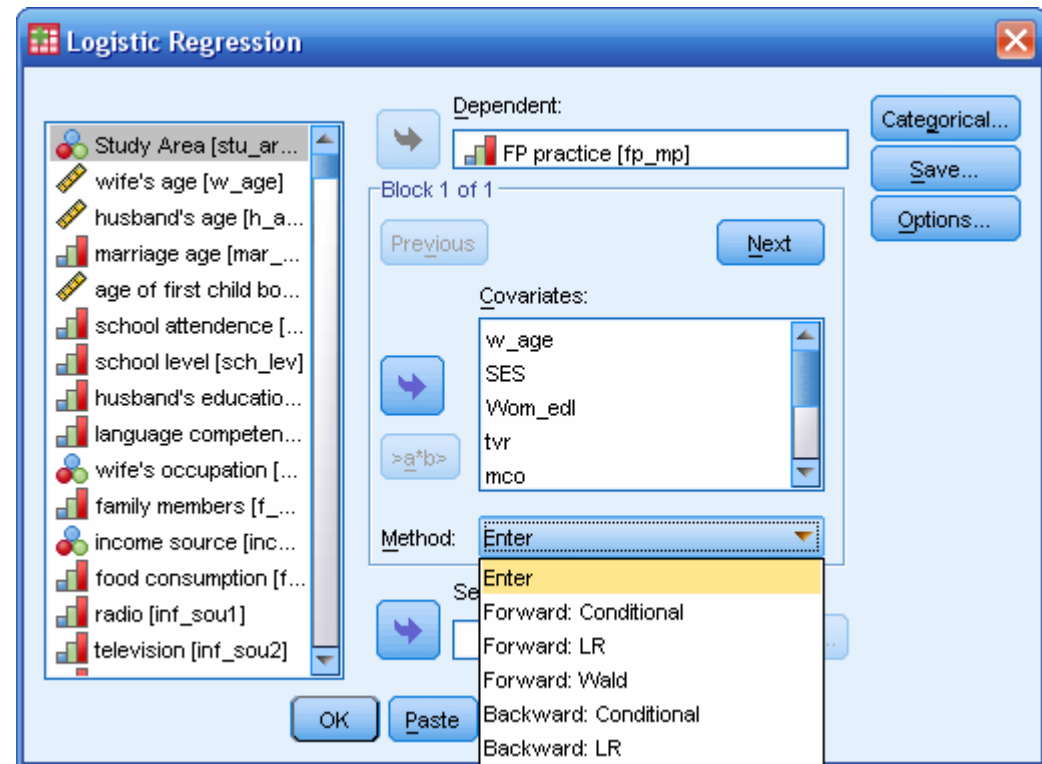
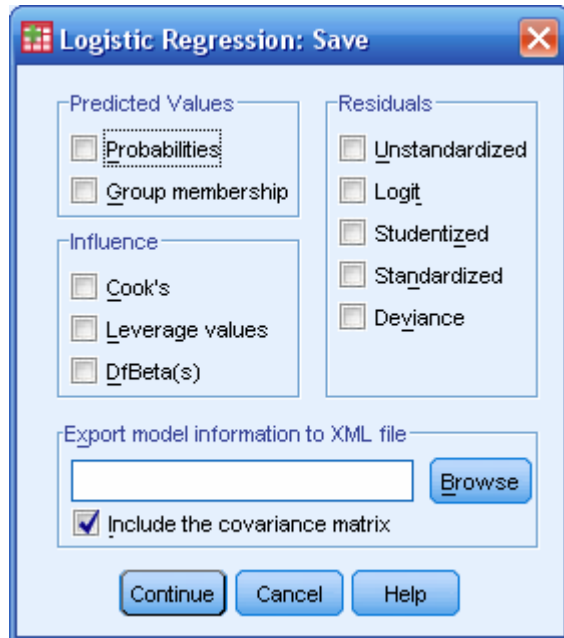
Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a						
w_age	.008	.014	.326	1	.568	1.008
SES	.306	.150	4.179	1	.041	1.358
Wom_edl	.007	.128	.003	1	.958	1.007
lvr	.110	.093	1.421	1	.233	1.117
mco	.567	.182	9.705	1	.002	1.763
religion	-.373	.192	3.779	1	.052	.689
totcd	-.622	.219	8.052	1	.005	.537
fauto_sc	.205	.087	5.613	1	.018	1.228
wjob	-.119	.329	.131	1	.717	.888
Constant	.491	.441	1.237	1	.266	1.634

a. Variable(s) entered on step 1: w_age, SES, Wom_edl, lvr, mco, religion, totcd, fauto_sc, wjob.

Further Options

- Can save predicted probabilities
- Diagnostics



- Model selection

Forward Selection Ward's method

Classification Table^a

Observed			Predicted		
			FP practice		Percentage Correct
			No	Yes	
Step 1	FP practice	No	0	220	.0
		Yes	0	466	100.0
	Overall Percentage				
Step 2	FP practice	No	0	220	.0
		Yes	0	466	100.0
	Overall Percentage				
Step 3	FP practice	No	29	191	13.2
		Yes	24	442	94.8
	Overall Percentage				
Step 4	FP practice	No	30	190	13.6
		Yes	22	444	95.3
	Overall Percentage				
Step 5	FP practice	No	29	191	13.2
		Yes	26	440	94.4
	Overall Percentage				

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	mco	.606	.174	12.090	1	.001	1.833
	Constant	.528	.102	26.957	1	.000	1.695
Step 2 ^b	SES	.389	.114	11.577	1	.001	1.475
	mco	.654	.176	13.816	1	.000	1.924
Step 3 ^c	Constant	.533	.103	26.821	1	.000	1.705
	SES	.368	.115	10.198	1	.001	1.445
	mco	.661	.177	13.904	1	.000	1.937
Step 4 ^d	totcd	-.637	.202	9.924	1	.002	.529
	Constant	.663	.112	34.946	1	.000	1.941
	SES	.361	.116	9.661	1	.002	1.435
	mco	.584	.181	10.399	1	.001	1.792
Step 5 ^e	totcd	-.562	.206	7.462	1	.006	.570
	faulto_sc	.185	.085	4.737	1	.030	1.203
	Constant	.685	.113	36.482	1	.000	1.984
	SES	.387	.117	10.933	1	.001	1.473
	mco	.584	.181	10.351	1	.001	1.793
	religion	-.342	.189	3.267	1	.071	.710
	totcd	-.603	.207	8.437	1	.004	.547
	faulto_sc	.208	.086	5.815	1	.016	1.231
	Constant	.800	.131	37.193	1	.000	2.225

a. Variable(s) entered on step 1: mco.

b. Variable(s) entered on step 2: SES.

c. Variable(s) entered on step 3: totcd.

d. Variable(s) entered on step 4: faulto_sc.

e. Variable(s) entered on step 5: religion.

Final model

Variable	B	S.E.	P level	Odds ratio
Wife's Age	0.007	0.024	0.760	1.007
Number of children	-0.034	0.111	0.760	0.967
Socio-economic status	0.282	0.203	0.165	1.326
Female Autonomy Score	0.223	0.131	0.088	1.249
Micro credit affiliation	0.583	0.274	0.033	1.792
in-degree Centrality	0.006	0.002	0.009	1.006
Out degree Centrality	0.042	0.004	0.000	1.043
Women's Education level	0.073	0.190	0.701	1.076
Exposure to Radio and TV	0.427	0.149	0.004	1.532
Exposure to Friends & Relatives	0.368	0.126	0.004	1.445
Exposure to Family Welfare Assistants	0.456	0.079	0.000	1.577
Husbands approval	1.333	0.284	0.000	3.791
Constant	-3.840	0.729	0.000	0.021

Multinomial Logistic Regression

PhD Michael Lewrick

- Study of Innovation in 170 German companies
- Total Innovation taken as response

(3 categories 0-9, 10 - 30 and >30)

Use 12 management capabilities formed from a questionnaire and used factor analysis

Capability
Knowledge
Competitor Orientation
Interorganisational
Organisational
Knowledge Acquisition

Market Orientation
Performance
Informal Networks
Outcomes
Formal Networks
Key performance

Multinomial Logistic Regression

Parameter Estimates

Total_innovativeness		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence	
								Lower Bound	Upper Bound
2 = 10 - 30 innovations	Intercept	1.307	.302	18.767	1	.000			
	Capability	.732	.253	8.392	1	.004	2.079	1.267	3.412
	Knowledge Enhancement	.138	.226	.373	1	.541	1.148	.738	1.786
	Competitor Orientation	.269	.271	.987	1	.320	1.309	.770	2.225
	Interorganisational Network	-.213	.241	.783	1	.376	.808	.503	1.296
	Organisational Learning	.885	.287	9.477	1	.002	2.423	1.379	4.255
	Knowledge Acquisition	1.120	.295	14.420	1	.000	3.066	1.720	5.466
	Market Orientation	.033	.256	.017	1	.897	1.034	.626	1.706
	Performance Measurement	.801	.255	9.874	1	.002	2.228	1.352	3.673
	Informal Networks	-.879	.288	9.344	1	.002	.415	.236	.730
	Outcomes	1.128	.306	13.578	1	.000	3.089	1.695	5.627
	Formal Networks	.206	.240	.738	1	.390	1.229	.768	1.965
	Key performance indicators	-.405	.261	2.410	1	.121	.667	.400	1.112
	3 = > 30 innovations	Intercept	-.573	.439	1.706	1	.192		
Capability		.952	.364	6.837	1	.009	2.592	1.269	5.293
Knowledge Enhancement		1.273	.405	9.869	1	.002	3.573	1.614	7.908
Competitor Orientation		.323	.331	.951	1	.329	1.382	.722	2.645
Interorganisational Network		.237	.313	.575	1	.448	1.268	.687	2.341
Organisational Learning		1.336	.406	10.847	1	.001	3.805	1.718	8.430
Knowledge Acquisition		1.905	.390	23.844	1	.000	6.722	3.129	14.443
Market Orientation		.309	.321	.929	1	.335	1.362	.726	2.555
Performance Measurement		1.058	.359	8.710	1	.003	2.882	1.427	5.820
Informal Networks		-.777	.390	3.984	1	.046	.460	.214	.986
Outcomes		1.300	.366	12.594	1	.000	3.670	1.790	7.524
Formal Networks		.214	.328	.424	1	.515	1.238	.651	2.354
Key performance indicators		-.279	.346	.654	1	.419	.756	.384	1.489

Multinomial Regression

Classification

Observed	Predicted			
	1 = < 10 innovations	2 = 10 - 30 innovations	3 = > 30 innovations	Percent Correct
1 = < 10 innovations	35	14	2	68.6%
2 = 10 - 30 innovations	15	70	5	77.8%
3 = > 30 innovations	2	15	13	43.3%
Overall Percentage	30.4%	57.9%	11.7%	69.0%

Modelling counts

- Poisson or Negative Binomial
- Can model rates – by using an offset
- Should be rare events and for Poisson the expected value should be similar to the variance
- Use Generalized linear models in SPSS

Models

$$P(x \text{ dead children}) = \frac{e^{-\lambda} \lambda^x}{x!} \text{ where } \lambda = \alpha + \sum_{i=1}^6 \beta_i x_i + \sum_{i=1}^4 \delta_i y_i + \sum_{i=1}^5 \phi_i z_i + \varepsilon$$

X = socio-cultural and economic variables

Y = communication variables

Z = area dummy variables

Log (No. of child deaths) – Log (total number of children ever born)

$$= \alpha + \sum_{i=1}^6 \beta_i x_i + \sum_{i=1}^4 \delta_i y_i + \sum_{i=1}^5 \phi_i z_i + \varepsilon$$

Modelling counts

The screenshot shows the 'Generalized Linear Models' dialog box with the 'Statistics' tab selected. The 'Analysis Type' is set to 'Type III' and the 'Confidence Interval Level (%)' is 95. Under 'Chi-square Statistics', 'Wald' is selected. Under 'Confidence Interval Type', 'Wald' is selected and the 'Tolerance level' is .0001. The 'Log-Likelihood Function' is set to 'Full'. In the 'Print' section, several options are checked, including 'Case processing summary', 'Descriptive statistics', 'Model information', 'Goodness of fit statistics', 'Model summary statistics', and 'Parameter estimates'. Other options like 'Contrast coefficient (L) matrices' and 'Iteration history' are unchecked. The 'Print Interval' is set to 1.

Generalized Linear Models

Type of Model | Response | Predictors | Model | Estimation | **Statistics** | EM Means | Save | Export

Model Effects

Analysis Type: Confidence Interval Level (%):

Chi-square Statistics

Wald
 Likelihood ratio

Confidence Interval Type

Wald
 Profile likelihood

Tolerance level:

Log-Likelihood Function:

Print

Case processing summary
 Descriptive statistics
 Model information
 Goodness of fit statistics
 Model summary statistics
 Parameter estimates
 Include exponential parameter estimates
 Covariance matrix for parameter estimates
 Correlation matrix for parameter estimates

Contrast coefficient (L) matrices
 General estimable functions
 Iteration history
Print Interval:
 Lagrange multiplier test of scale parameter or negative binomial ancillary parameter

OK | Paste | Reset | Cancel | Help

Poisson model of child death in rural Bangladesh

Variable	Total number of dead children			Proportion of Children Ever born who died		
	Coefficient	S.E.	P value	Coefficient	S.E.	P value
Constant	-2.424	0.639	<0.001	-3.954	0.69	0.496
<i>Demographic variables</i>						
Number of children	0.272	0.042	<0.001			
Wife's age				0.048	0.012	<0.001
Wife's Age at marriage	-0.047	0.029	0.102	-0.045	0.0281	0.111
<i>Socio-Economic & Cultural variables</i>						
Wife's education level	-0.373	0.139	0.007	-0.241	0.14	0.086
Husbands education level	-0.046	0.104	0.655	-0.051	0.103	0.618
Micro Credit Affiliation	0.476	0.171	0.006	0.419	0.173	0.015
Female autonomy score	-0.001	0.081	0.987	-0.046	0.081	0.572
SES	0.156	0.12	0.194	0.083	0.122	0.5
Religion	-0.085	0.231	0.712	-0.207	0.236	0.379
<i>Model Performance Measures</i>						
Deviance/DF	0841			0.794		
Pearson Chisq/DF	1.277			1.263		
Log Likelihood	-341.8			-327.6		

References

- Gayen, K. and Raeside, R.(2007). Social networks, normative influence and health delivery in rural Bangladesh, *social Science and Medicine*, 65(5):900-14 .
- Hosmer, D. W. and Lemeshow, S., (2000). *Applied Logistic Regression* 2nd ed, John Wiley and Sons, Chichester.
- Lawal, B., (2003). *Categorical Data analysis with SAS and SPSS Applications*, Lawrence Erlbaum Associates, London.
- Nemes, S.,Miao, J., Genell, A. and Steineck, G., (2009), Bias in odds ratios by logistic regression modelling and sample size, *BMC Med Res Methodol*,
doi: 10.1186/1471-2288-9-56.PMCID
- Simonoff, J.S., (2003). *Analyzing Categorical Data*, Springer, London.