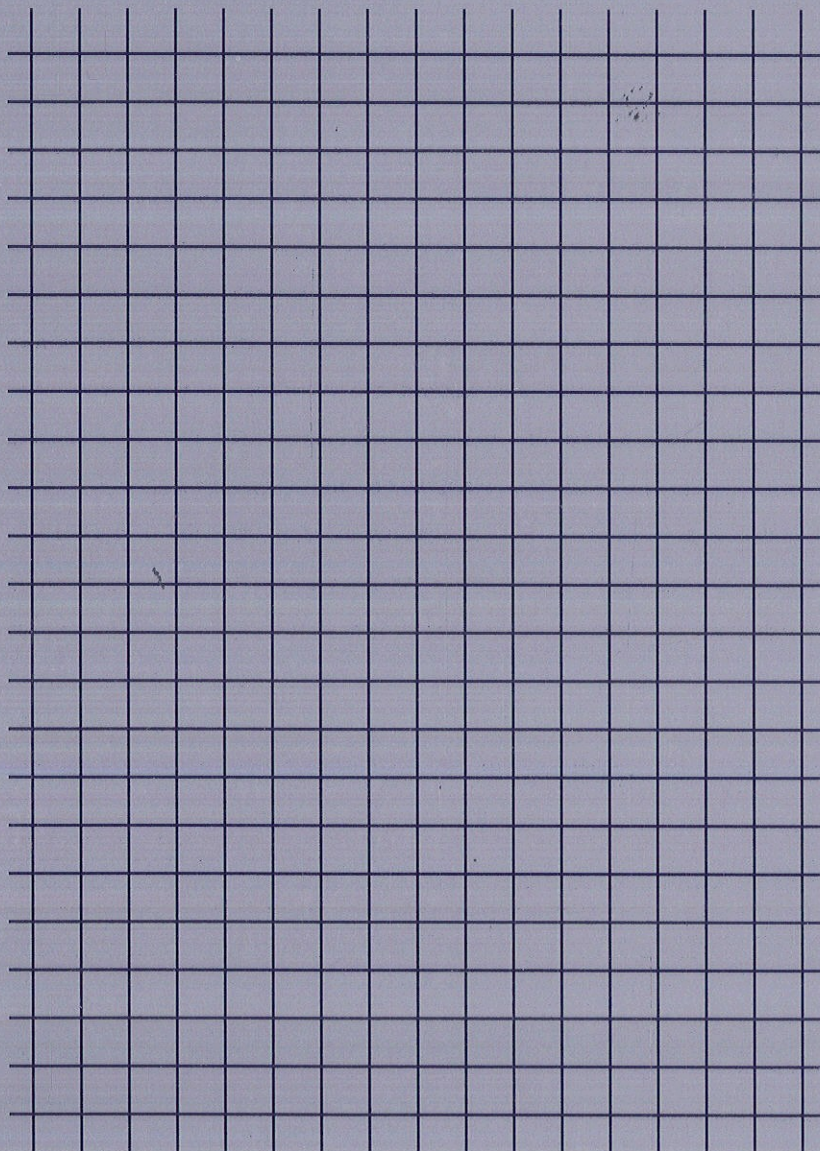


# Categorical Data Analysis

for Geographers and

Environmental Scientists



Neil Wrigley

# CONTENTS

---

Preface	xv
Acknowledgements	xvii
<b>Part 1 Some essential preliminaries</b>	<b>1</b>
<b>Chapter 1 Introduction</b>	<b>3</b>
1.1 Classifications, categorical data, and a methodological transformation	4
1.2 A framework for discussion	7
1.3 Some important tables	8
1.4 Sampling schemes	13
1.4.1 Poisson	14
1.4.2 Multinomial	14
1.4.3 Product-multinomial	14
1.5 Integration and transition	15
1.6 How to use the book and what it assumes	16
<b>Part 2 The basic family of statistical models</b>	<b>19</b>
<b>Chapter 2 Categorical response variable, continuous explanatory variables</b>	<b>21</b>
2A TWO RESPONSE CATEGORIES: THE DICHOTOMOUS CASE	21
2.1 Introduction	21
2.2 A conventional regression model approach	22

2.3	Alternative solutions	24
2.3.1	<i>The linear probability model</i>	25
2.3.2	<i>The logit model</i>	28
2.3.3	<i>The probit model</i>	29
2.4	Estimating the parameters of a logistic/logit model	30
2.4.1	<i>Grouped data and weighted least squares</i>	30
2.4.2	<i>Individual data and maximum likelihood</i>	35
2.5	Some simple examples	38
EXAMPLE 2.1	<i>Oil and gas exploration in south-central Kansas</i>	39
EXAMPLE 2.2	<i>Work trip mode choice in Sydney</i>	40
EXAMPLE 2.3	<i>Shopping trip mode choice in Pittsburgh</i>	42
2.6	Testing hypotheses about the logistic/logit model parameters	43
2.6.1	<i>Separate tests</i>	43
2.6.2	<i>Joint tests</i>	45
2.6.3	<i>Partial joint tests</i>	45
EXAMPLE 2.4	<i>Oil and gas exploration in south-central Kansas (continued)</i>	46
2.7	Goodness-of-fit measures, residuals, and predicted values	49
2.7.1	<i>Goodness-of-fit measures</i>	49
2.7.2	<i>Residuals</i>	52
EXAMPLE 2.5	<i>Oil and gas exploration in south-central Kansas<sup>a</sup> (continued)</i>	54
EXAMPLE 2.6	<i>Resource evaluation in Newfoundland</i>	58
2B	MULTIPLE RESPONSE CATEGORIES: THE POLYTOMOUS CASE	62
2.8	Introduction	62
2.9	The extended linear logit and logistic models	62
2.9.1	<i>The linear logit model</i>	63
2.9.2	<i>The logistic model</i>	65
2.9.3	<i>General forms</i>	66
2.10	Estimating the parameters of the extended models	67
EXAMPLE 2.7	<i>Oil and gas exploration in south-central Kansas (continued)</i>	69
2.11	Types of explanatory variables and response categories	72
2.11.1	<i>Types of explanatory variables</i>	72
2.11.2	<i>Types of response categories</i>	76

<i>EXAMPLE 2.8 Shopping trip destination choice in Pittsburgh</i>	77
<i>EXAMPLE 2.9 Shopping trip destination choice in West Yorkshire</i>	78
<i>EXAMPLE 2.10 Shopping destination and mode of travel choice in San Francisco</i>	81
<i>EXAMPLE 2.11 Shopping destination and mode of travel choice in Eindhoven</i>	82
<b>Appendix 2.1 Alternative derivations of logistic/logit models</b>	84
<b>Chapter 3 Categorical response variable, mixed explanatory variables</b>	91
3.1 Dummy variables in conventional regression models	91
3.2 Categorical explanatory variables in logistic/logit models	94
3.3 Dummy variables within the general typology of explanatory variables	96
3.4 A range of illustrative examples	97
<i>EXAMPLE 3.1 Determinants of housing tenure in Sydney</i>	97
<i>EXAMPLE 3.2 Occupational attainment in the United States</i>	99
<i>EXAMPLE 3.3 Housing choice in Pittsburgh</i>	103
<i>EXAMPLE 3.4 Work trip mode choice in Washington D.C. and the spatial transferability of models</i>	106
3.5 Summary	109
<b>Chapter 4 Categorical response variable, categorical explanatory variables: the linear logit model approach</b>	111
4.1 Linear logit models for cell (f): basic forms	112
<i>EXAMPLE 4.1 Preference for army camp location among American soldiers</i>	114
<i>EXAMPLE 4.2 Evaluation of military policemen by negro soldiers from different regions</i>	117
4.2 Weighted least squares estimation of cell (f) linear logit models	121
4.3 Goodness-of-fit and test statistics	125
4.3.1 Joint tests	126
4.3.2 Partial joint tests	127
4.3.3 Separate tests	128
<i>EXAMPLE 4.3 Automobile accidents and the accident environment in North Carolina</i>	129
4.4 Coding systems for categorical explanatory variables	132

4.5 Interactions, saturation, hierarchies and parsimony	136
<i>EXAMPLE 4.4 Determinants of homeownership in Boston and Baltimore</i>	140
4.6 Model selection	143
<i>EXAMPLE 4.5 Byssinosis amongst U.S. cotton textile workers</i>	147
4.7 A more general matrix formulation	152
 <b>Chapter 5 All variables categorical but no division into response and explanatory</b>	155
5.1 The hypothesis of independence and the chi-square test	156
5.2 Towards a log-linear model of independence	157
5.3 A hierarchical set of log-linear models for two-dimensional contingency tables	161
<i>EXAMPLE 5.1 Some simple two-dimensional contingency tables</i>	163
(a) <i>Pebbles in glacial till</i>	164
(b) <i>Lifetime residential mobility and retirement migration</i>	166
(c) <i>Farm acreage under woodland</i>	168
5.4 Log-linear models for multidimensional contingency tables	169
<i>EXAMPLE 5.2 Age, decay and use of buildings in north-east London</i>	175
<i>EXAMPLE 5.3 Industrial location in Hull</i>	176
5.5 Abbreviated notation systems for log-linear models	179
5.6 Estimation of the parameters and the expected cell frequencies	182
5.6.1 <i>The iterative proportional fitting procedure</i>	184
5.6.2 <i>The iterative weighted least squares procedure</i>	188
5.6.3 <i>The Newton–Raphson procedure</i>	190
5.7 Model selection	190
5.7.1 <i>Strategy 1. Stepwise selection</i>	191
5.7.2 <i>Strategy 2. Abbreviated stepwise selection</i>	194
5.7.3 <i>Strategy 3. Screening</i>	196
5.7.4 <i>Strategy 4. Aitkin’s simultaneous test procedure</i>	200
<i>EXAMPLE 5.4 Shopping behaviour in Manchester</i>	204
<i>EXAMPLE 5.5 Non-fatal deliberate self-harm in Bristol</i>	207
5.8 The analysis of residuals	211
<i>EXAMPLE 5.6 Opinions about a television series in urban and rural areas</i>	213

<b>Chapter 6 Categorical response variable, categorical explanatory variables: the log-linear model approach</b>	<b>215</b>
6.1 Fixed marginal totals	215
6.2 Log-linear models for mixed explanatory/response variable situations	216
<i>EXAMPLE 6.1 A hypothetical three-dimensional table</i>	217
<i>EXAMPLE 6.2 Shopping behaviour in Manchester (continued)</i>	219
<i>EXAMPLE 6.3 Relationships between ethnic origin, birthplace, age, and occupation in Canada in 1871</i>	221
6.3 Log-linear models as logit models	223
6.3.1 The dichotomous response variable case	223
6.3.2 The multiple-category response variable case	227
6.3.3 Discussion	230
 <b>Chapter 7 Computer programs for categorical data analysis</b>	 <b>233</b>
7.1 A classification of available programs	233
7.2 Programs which use function maximization algorithms	233
7.3 Programs based upon weighted least squares algorithms	236
7.3.1 Iterative weighted least squares	236
7.3.2 Non-iterative weighted least squares	237
7.4 Programs which use iterative proportional fitting algorithms	237
 <b>Part 3 Extensions of the basic statistical models</b>	 <b>239</b>
 <b>Chapter 8 Special topics in logistic/logit modelling</b>	 <b>241</b>
8.1 Logistic regression diagnostics and resistant fitting	241
8.2 Some logistic/logit model analogues to classical spatial analysis models	246
8.2.1 Trend surface models	246
8.2.2 Space-time models	249
<i>EXAMPLE 8.1 Aircraft noise disturbance around Manchester Airport</i>	250
<i>EXAMPLE 8.2 The space-time pattern of housing deterioration in Indianapolis</i>	252



8.3 Logistic/logit models for ordered categories and systems of equations	254
8.3.1 <i>Ordered response categories</i>	254
8.3.2 <i>Systems of logistic/logit models</i>	255
8.4 Some extensions of the general matrix formulations of cell (f) linear logit models	259
8.4.1 <i>A multiple response category model</i>	260
8.4.2 <i>A repeated-measurement research design example</i>	261
8.4.3 <i>Paired comparison experiment examples</i>	264
EXAMPLE 8.3 <i>Residential preferences of schoolchildren in Southend-on-Sea</i>	269
8.5 An alternative treatment of error structures in cell (f) linear logit models	271
<b>Chapter 9 Special topics in log-linear modelling</b>	275
9.1 Combining categories and collapsing tables	275
EXAMPLE 9.1 <i>Oak, hickory and maple distributions in Lansing Woods, Michigan</i>	277
9.2 Sampling zeros	279
9.2.1 <i>Sampling zeros and saturated log-linear models</i>	279
9.2.2 <i>Sampling zeros and unsaturated log-linear models</i>	281
EXAMPLE 9.2 <i>Non-fatal deliberate self-harm in Bristol (continued)</i>	282
EXAMPLE 9.3 <i>Relationships between tree species and tree height in the forests of South Island, New Zealand</i>	284
9.3 Structural zeros and incomplete contingency tables	285
EXAMPLE 9.4 <i>Filtering in the housing market of Kingston, Ontario</i>	289
EXAMPLE 9.5 <i>Plant type, soil type and slope aspect</i>	292
9.4 Outliers or rogue cells	293
EXAMPLE 9.6 <i>Opinions about a television series in urban and rural areas (continued)</i>	293
9.5 Square tables, symmetry, and marginal homogeneity	295
9.5.1 <i>Symmetry</i>	295
9.5.2 <i>Marginal homogeneity</i>	296
9.5.3 <i>Quasi-symmetry</i>	297
9.5.4 <i>Symmetry and marginal homogeneity in multidimensional tables</i>	299
9.5.5 <i>Alternative log-linear models for square tables</i>	300

<i>EXAMPLE 9.7 Filtering in the housing market of Kingston, Ontario (continued)</i>	301
<b>9.6 Some remaining issues</b>	303
9.6.1 <i>The multiplicative form of the log-linear model</i>	303
9.6.2 <i>Log-linear models for tables with ordered categories</i>	304
9.6.3 <i>Causal analysis with log-linear models</i>	305
9.6.4 <i>Log-linear models and spatially dependent data</i>	307
 <b>Part 4 Discrete choice modelling</b>	 311
<b>Chapter 10 Statistical models for discrete choice analysis</b>	313
10.1 Random utility maximization, discrete choice theory and multinomial logit models	313
<i>EXAMPLE 10.1 The collapse and re-opening of the Tasman Bridge</i>	318
10.2 The IIA property and its implications	324
10.3 The search for less restrictive discrete choice models	326
10.3.1 <i>The multinomial probit model</i>	327
10.3.2 <i>The dogit model</i>	328
10.3.3 <i>The nested logit model</i>	329
10.3.4 <i>Elimination-by-aspects models</i>	332
10.3.5 <i>Weight shifting models</i>	334
<i>EXAMPLE 10.2 Location decisions of clothing retailers in Boston</i>	337
<i>EXAMPLE 10.3 Travel mode choice in Montreal</i>	340
<i>EXAMPLE 10.4 Travel mode choice in the Rotterdam/Hague Metropolitan area.</i>	341
10.4 Assessing and comparing the performance of alternative discrete choice models	343
10.4.1 <i>Tests of the IIA property of the MNL</i>	344
10.4.2 <i>Tests of the MNL against specific alternative discrete choice models</i>	347
10.4.3 <i>A generalized test procedure for comparing the performance of any pair of discrete choice models</i>	349
10.5 A brief guide to some remaining statistical issues	350
10.5.1 <i>Statistical transformations and the search for appropriate functional form</i>	350
10.5.2 <i>Sample design and parameter estimation</i>	351
10.5.3 <i>Panel data and dynamic modelling</i>	353



10.5.4 <i>Specification analysis: improper exclusion or inclusion of explanatory variables</i>	355
10.5.5 <i>Wider themes of empirical application</i>	357
 <b>Part 5 Towards integration</b>	 359
<b>Chapter 11 An alternative framework</b>	361
11.1 The central classification scheme reconsidered	361
11.2 The GLM framework	362
11.2.1 <i>The linear predictor</i>	363
11.2.2 <i>The link function</i>	364
11.2.3 <i>The error distribution</i>	364
11.2.4 <i>Some examples of GLMs</i>	365
11.3 Conclusion	366
 References	 367
 Author index	 383
 Example index	 387
 Subject index	 389