Contents

COPYRIGHTED MATERIAL

Preface xv

1 Introduction 1

1.1 Real-Life Problems as Motivation for Model Building, 1

1.1.1 Many Candidate Models, 1

1.1.2 Functional Form for Continuous Predictors, 2

1.1.3 Example 1: Continuous Response, 2

1.1.4 Example 2: Multivariable Model for Survival Data, 5

1.2 Issues in Modelling Continuous Predictors, 8

1.2.1 Effects of Assumptions, 8

1.2.2 Global versus Local Influence Models, 9

1.2.3 Disadvantages of Fractional Polynomial Modelling, 9

1.2.4 Controlling Model Complexity, 10

1.3 Types of Regression Model Considered, 10

1.3.1 Normal-Errors Regression, 10

1.3.2 Logistic Regression, 12

1.3.3 Cox Regression, 12

1.3.4 Generalized Linear Models, 14

1.3.5 Linear and Additive Predictors, 14

1.4 Role of Residuals, 15

1.4.1 Uses of Residuals, 15

1.4.2 Graphical Analysis of Residuals, 15

1.5 Role of Subject-Matter Knowledge in Model Development, 16

1.6 Scope of Model Building in our Book, 17

1.7 Modelling Preferences, 18

1.7.1 General Issues, 18

1.7.2 Criteria for a Good Model, 18

1.7.3 Personal Preferences, 19

1.8 General Notation, 20

2 Selection of Variables 23

2.1 Introduction, 23
2.2 Background, 24
2.3 Preliminaries for a Multivariable Analysis, 25
2.4 Aims of Multivariable Models, 26
2.5 Prediction: Summary Statistics and Comparisons, 29
2.6 Procedures for Selecting Variables, 29
2.6.1 Strength of Predictors, 30
2.6.2 Stepwise Procedures, 31
2.6.3 All-Subsets Model Selection Using Information Criteria, 32
2.6.4 Further Considerations, 33
2.7 Comparison of Selection Strategies in Examples, 35
2.7.1 Myeloma Study, 35
2.7.2 Educational Body-Fat Data, 36
2.7.3 Glioma Study, 38
2.8 Selection and Shrinkage, 40
2.8.1 Selection Bias, 40
2.8.2 Simulation Study, 40
2.8.3 Shrinkage to Correct for Selection Bias, 42
2.8.4 Post-estimation Shrinkage, 44
2.8.5 Reducing Selection Bias, 45
2.8.6 Example, 46
2.9 Discussion, 47
2.9.1 Model Building in Small Datasets, 47
2.9.2 Full, Pre-specified or Selected Model? 47
2.9.3 Comparison of Selection Procedures, 49
2.9.4 Complexity, Stability and Interpretability, 49
2.9.5 Conclusions and Outlook, 50

3 Handling Categorical and Continuous Predictors 53

3.1 Introduction, 53
3.2 Types of Predictor, 54
3.2.1 Binary, 54
3.2.2 Nominal, 54
3.2.3 Ordinal, Counting, Continuous, 55
3.2.4 Derived, 55
3.3 Handling Ordinal Predictors, 55
3.3.1 Coding Schemes, 55
3.3.2 Effect of Coding Schemes on Variable Selection, 56
3.4 Handling Counting and Continuous Predictors: Categorization, 58
3.4.1 'Optimal' Cutpoints: A Dangerous Analysis, 58
3.4.2 Other Ways of Choosing a Cutpoint, 59
3.5 Example: Issues in Model Building with Categorized Variables, 60
3.5.1 One Ordinal Variable, 61
3.5.2 Several Ordinal Variables, 62

3.6 Handling Counting and Continuous Predictors: Functional Form, 64
3.6.1 Beyond Linearity, 64
3.6.2 Does Nonlinearity Matter? 65
3.6.3 Simple versus Complex Functions, 66
3.6.4 Interpretability and Transportability, 66
3.7 Empirical Curve Fitting, 67
3.7.1 General Approaches to Smoothing, 68
3.7.2 Critique of Local and Global Influence Models, 68
3.8 Discussion, 69
3.8.1 Sparse Categories, 69
3.8.2 Choice of Coding Scheme, 69
3.8.3 Categorizing Continuous Variables, 70
3.8.4 Handling Continuous Variables, 70

4 Fractional Polynomials for One Variable 71
4.1 Introduction, 72
4.2 Background, 72
4.2.1 Genesis, 72
4.2.2 Types of Model, 73
4.2.3 Relation to Box–Tidwell and Exponential Functions, 73
4.3 Definition and Notation, 74
4.3.1 Fractional Polynomials, 74
4.3.2 First Derivative, 74
4.4 Characteristics, 75
4.4.1 FP1 and FP2 Functions, 75
4.4.2 Maximum or Minimum of a FP2 Function, 75
4.5 Examples of Curve Shapes with FP1 and FP2 Functions, 76
4.6 Choice of Powers, 78
4.7 Choice of Origin, 79
4.8 Model Fitting and Estimation, 79
4.9 Inference, 79
4.9.1 Hypothesis Testing, 79
4.9.2 Interval Estimation, 80
4.10 Function Selection Procedure, 82
4.10.1 Choice of Default Function, 82
4.10.2 Closed Test Procedure for Function Selection, 82
4.10.3 Example, 83
4.10.4 Sequential Procedure, 83
4.10.5 Type I Error and Power of the Function Selection Procedure, 84
4.11 Scaling and Centering, 84
4.11.1 Computational Aspects, 84
4.11.2 Examples, 85
4.12 FP Powers as Approximations to Continuous Powers, 85
4.12.1 Box–Tidwell and Fractional Polynomial Models, 85
4.12.2 Example, 85

4.13 Presentation of Fractional Polynomial Functions, 86
4.13.1 Graphical, 86
4.13.2 Tabular, 87
4.14 Worked Example, 89
4.14.1 Details of all Fractional Polynomial Models, 89
4.14.2 Function Selection, 90
4.14.3 Details of the Fitted Model, 90
4.14.4 Standard Error of a Fitted Value, 91
4.14.5 Fitted Odds Ratio and its Confidence Interval, 91
4.15 Modelling Covariates with a Spike at Zero, 92
4.16 Power of Fractional Polynomial Analysis, 94
4.16.1 Underlying Function Linear, 95
4.16.2 Underlying Function FP1 or FP2, 95
4.16.3 Comment, 96
4.17 Discussion, 97

5 Some Issues with Univariate Fractional Polynomial Models 99
5.1 Introduction, 99
5.2 Susceptibility to Influential Covariate Observations, 100
5.3 A Diagnostic Plot for Influential Points in FP Models, 100
5.3.1 Example 1: Educational Body-Fat Data, 101
5.3.2 Example 2: Primary Biliary Cirrhosis Data, 101
5.4 Dependence on Choice of Origin, 103
5.5 Improving Robustness by Preliminary Transformation, 105
5.5.1 Example 1: Educational Body-Fat Data, 106
5.5.2 Example 2: PBC Data, 107
5.5.3 Practical Use of the Pre-transformation $g_\delta(x)$, 107
5.6 Improving Fit by Preliminary Transformation, 108
5.6.1 Lack of Fit of Fractional Polynomial Models, 108
5.6.2 Negative Exponential Pre-transformation, 108
5.7 Higher Order Fractional Polynomials, 109
5.7.1 Example 1: Nerve Conduction Data, 109
5.7.2 Example 2: Triceps Skinfold Thickness, 110
5.8 When Fractional Polynomial Models are Unsuitable, 111
5.8.1 Not all Curves are Fractional Polynomials, 111
5.8.2 Example: Kidney Cancer, 112
5.9 Discussion, 113

6 MFP: Multivariable Model-Building with Fractional Polynomials 115
6.1 Introduction, 115
6.2 Motivation, 116
6.3 The MFP Algorithm, 117
6.3.1 Remarks, 118
6.3.2 Example, 118

6.4 Presenting the Model, 120
6.4.1 Parameter Estimates, 120
6.4.2 Function Plots, 121
6.4.3 Effect Estimates, 121
6.5 Model Criticism, 123
6.5.1 Function Plots, 123
6.5.2 Graphical Analysis of Residuals, 124
6.5.3 Assessing Fit by Adding More Complex Functions, 125
6.5.4 Consistency with Subject-Matter Knowledge, 129
6.6 Further Topics, 129
6.6.1 Interval Estimation, 129
6.6.2 Importance of the Nominal Significance Level, 130
6.6.3 The Full MFP Model, 131
6.6.4 A Single Predictor of Interest, 132
6.6.5 Contribution of Individual Variables to the Model Fit, 134
6.6.6 Predictive Value of Additional Variables, 136
6.7 Further Examples, 138
6.7.1 Example 1: Oral Cancer, 138
6.7.2 Example 2: Diabetes, 139
6.7.3 Example 3: Whitehall I, 140
6.8 Simple Versus Complex Fractional Polynomial Models, 144
6.8.1 Complexity and Modelling Aims, 144
6.8.2 Example: GBSG Breast Cancer Data, 144
6.9 Discussion, 146
6.9.1 Philosophy of MFP, 147
6.9.2 Function Complexity, Sample Size and Subject-Matter Knowledge, 148
6.9.3 Improving Robustness by Preliminary Covariate Transformation, 148
6.9.4 Conclusion and Future, 149

7 Interactions **151**
7.1 Introduction, 151
7.2 Background, 152
7.3 General Considerations, 152
7.3.1 Effect of Type of Predictor, 152
7.3.2 Power, 153
7.3.3 Randomized Trials and Observational Studies, 153
7.3.4 Predefined Hypothesis or Hypothesis Generation, 153
7.3.5 Interactions Caused by Mismodelling Main Effects, 154
7.3.6 The ‘Treatment–Effect’ Plot, 154
7.3.7 Graphical Checks, Sensitivity and Stability Analyses, 154
7.3.8 Cautious Interpretation is Essential, 155
7.4 The MFPI Procedure, 155
7.4.1 Model Simplification, 156
7.4.2 Check of the Results and Sensitivity Analysis, 156

7.5 Example 1: Advanced Prostate Cancer, 157
7.5.1 The Fitted Model, 158
7.5.2 Check of the Interactions, 160
7.5.3 Final Model, 161
7.5.4 Further Comments and Interpretation, 162
7.5.5 FP Model Simplification, 163
7.6 Example 2: GBSG Breast Cancer Study, 163
7.6.1 Oestrogen Receptor Positivity as a Predictive Factor, 163
7.6.2 A Predefined Hypothesis: Tamoxifen–Oestrogen Receptor Interaction, 163
7.7 Categorization, 165
7.7.1 Interaction with Categorized Variables, 165
7.7.2 Example: GBSG Study, 166
7.8 STEPP, 167
7.9 Example 3: Comparison of STEPP with MFPI, 168
7.9.1 Interaction in the Kidney Cancer Data, 168
7.9.2 Stability Investigation, 168
7.10 Comment on Type I Error of MFPI, 171
7.11 Continuous-by-Continuous Interactions, 172
7.11.1 Mismodelling May Induce Interaction, 173
7.11.2 MFPIgen: An FP Procedure to Investigate Interactions, 174
7.11.3 Examples of MFPIgen, 175
7.11.4 Graphical Presentation of Continuous-by-Continuous Interactions, 179
7.11.5 Summary, 180
7.12 Multi-Category Variables, 181
7.13 Discussion, 181

8 Model Stability **183**
8.1 Introduction, 183
8.2 Background, 184
8.3 Using the Bootstrap to Explore Model Stability, 185
8.3.1 Selection of Variables within a Bootstrap Sample, 185
8.3.2 The Bootstrap Inclusion Frequency and the Importance of a Variable, 186
8.4 Example 1: Glioma Data, 186
8.5 Example 2: Educational Body-Fat Data, 188
8.5.1 Effect of Influential Observations on Model Selection, 189
8.6 Example 3: Breast Cancer Diagnosis, 190
8.7 Model Stability for Functions, 191
8.7.1 Summarizing Variation between Curves, 191
8.7.2 Measures of Curve Instability, 192
8.8 Example 4: GBSG Breast Cancer Data, 193
8.8.1 Interdependencies among Selected Variables and Functions in Subsets, 193

8.8.2 Plots of Functions, 193
8.8.3 Instability Measures, 195
8.8.4 Stability of Functions Depending on Other Variables Included, 196
8.9 Discussion, 197
8.9.1 Relationship between Inclusion Fractions, 198
8.9.2 Stability of Functions, 198

9 Some Comparisons of MFP with Splines 201
9.1 Introduction, 201
9.2 Background, 202
9.3 MVRS: A Procedure for Model Building with Regression Splines, 203
9.3.1 Restricted Cubic Spline Functions, 203
9.3.2 Function Selection Procedure for Restricted Cubic Splines, 205
9.3.3 The MVRS Algorithm, 205
9.4 MVSS: A Procedure for Model Building with Cubic Smoothing Splines, 205
9.4.1 Cubic Smoothing Splines, 205
9.4.2 Function Selection Procedure for Cubic Smoothing Splines, 206
9.4.3 The MVSS Algorithm, 206
9.5 Example 1: Boston Housing Data, 207
9.5.1 Effect of Reducing the Sample Size, 208
9.5.2 Comparing Predictors, 212
9.6 Example 2: GBSG Breast Cancer Study, 214
9.7 Example 3: Pima Indians, 215
9.8 Example 4: PBC, 217
9.9 Discussion, 219
9.9.1 Splines in General, 220
9.9.2 Complexity of Functions, 221
9.9.3 Optimal Fit or Transferability? 221
9.9.4 Reporting of Selected Models, 221
9.9.5 Conclusion, 222

10 How To Work with MFP 223
10.1 Introduction, 223
10.2 The Dataset, 223
10.3 Univariate Analyses, 226
10.4 MFP Analysis, 227
10.5 Model Criticism, 228
10.5.1 Function Plots, 228
10.5.2 Residuals and Lack of Fit, 228
10.5.3 Robustness Transformation and Subject-Matter Knowledge, 229
10.5.4 Diagnostic Plot for Influential Observations, 230
10.5.5 Refined Model, 231
10.5.6 Interactions, 231
10.6 Stability Analysis, 232

10.7 Final Model, 235
10.8 Issues to be Aware of, 235
10.8.1 Selecting the Main-Effects Model, 235
10.8.2 Further Comments on Stability, 236
10.8.3 Searching for Interactions, 238
10.9 Discussion, 238

11 Special Topics Involving Fractional Polynomials 241
11.1 Time-Varying Hazard Ratios in the Cox Model, 241
11.1.1 The Fractional Polynomial Time Procedure, 242
11.1.2 The MFP Time Procedure, 243
11.1.3 Prognostic Model with Time-Varying Effects for Patients with Breast Cancer, 243
11.1.4 Categorization of Survival Time, 245
11.1.5 Discussion, 246
11.2 Age-specific Reference Intervals, 247
11.2.1 Example: Fetal growth, 247
11.2.2 Using FP Functions as Smoothers, 248
11.2.3 More Sophisticated Distributional Assumptions, 249
11.2.4 Discussion, 249
11.3 Other Topics, 250
11.3.1 Quantitative Risk Assessment in Developmental Toxicity Studies, 250
11.3.2 Model Uncertainty for Functions, 251
11.3.3 Relative Survival, 252
11.3.4 Approximating Smooth Functions, 253
11.3.5 Miscellaneous Applications, 254

12 Epilogue 255
12.1 Introduction, 255
12.2 Towards Recommendations for Practice, 255
12.2.1 Variable Selection Procedure, 255
12.2.2 Functional Form for Continuous Covariates, 257
12.2.3 Extreme Values or Influential Points, 257
12.2.4 Sensitivity Analysis, 257
12.2.5 Check for Model Stability, 258
12.2.6 Complexity of a Predictor, 258
12.2.7 Check for Interactions, 258
12.3 Omitted Topics and Future Directions, 258
12.3.1 Measurement Error in Covariates, 258
12.3.2 Meta-analysis, 258
12.3.3 Multi-level (Hierarchical) Models, 259
12.3.4 Missing Covariate Data, 259
12.3.5 Other Types of Model, 259
12.4 Conclusion, 259

Appendix A: Data and Software Resources 261

A.1 Summaries of Datasets, 261

A.2 Datasets used more than once, 262

A.2.1 Research Body Fat, 262

A.2.2 GBSG Breast Cancer, 262

A.2.3 Educational Body Fat, 263

A.2.4 Glioma, 264

A.2.5 Prostate Cancer, 264

A.2.6 Whitehall I, 265

A.2.7 PBC, 265

A.2.8 Oral Cancer, 266

A.2.9 Kidney Cancer, 266

A.3 Software, 267

Appendix B: Glossary of Abbreviations 269

References 271

Index 285