

Statistics for Astrophysics Methods and Applications of the Regression

Didier Fraix-Burnet & David Valls-Gabaud (editors)

Detailed Table of Contents

REGRESSION MODELS: A BRIEF INTRODUCTION (G. Grégoire) ..	3
1 Some history	3
2 Simple Linear Regression	4
3 Multiple Linear Regression	5
4 Generalized linear regression	6
4.1 Logistic regression	7
4.2 Poissonian regression	7
5 Further in regression models	8
PRACTICAL SESSION 1. INTRODUCTION TO R (M. Clausel, G. Grégoire)	
11	
1 More about R	11
2 First steps.	12
3 Random distributions	14
4 Empirical description	15
5 Reading data in a file	17
6 A toy example	18
SIMPLE LINEAR REGRESSION (G. Grégoire)	19
1 Introduction. Statistical issues	19
2 Model, assumptions	21
3 Least Square Method and basic results	22
3.1 Least Square Method (OLS)	22
3.2 Fitted values, residuals and estimate σ^2	24
3.3 Basic statistical properties	24
3.4 Confidence Intervals and Tests	26
4 Variance decomposition. ANOVA table and R^2	27
4.1 Some comments	28
4.2 The R^2 and adjusted R^2 coefficients	28
4.3 The ANOVA table	30
5 Confidence Intervals for the mean and Forecast Intervals	31

5.1 Confidence intervals for the mean	31
5.2 Prediction intervals	31
5.3 Plots of CI and PI intervals using R	32
6 Residuals and diagnostics	32
6.1 Residuals and h-values	32
6.2 Diagnostic analysis : outliers, influent observations, observations ..	34
with high leverage effect	20
6.3 Diagnostic analysis for cars data using R	36
7 Dealing with departures from model assumptions and practical issues	38
7.1 Transformations	38
7.2 Non-linearity	39
7.3 Non-normality	39
7.4 Heteroscedasticity	40
7.5 Several data sets with different departures from model assumptions but	
same basic statistical information	41
PRACTICAL SESSION 2. SIMPLE LINEAR REGRESSION (M. Clausel,	
G. Grégoire)	41
1 Study of Galton's data on hereditary of size	41
2 Vapor tension of mercury	43
MULTIPLE LINEAR REGRESSION (G. Grégoire)	45
1 Introduction	45
2 Model and assumptions	48
3 Least Square Method and basic results	50
3.1 Least Square Method (OLS)	50
3.2 Fitted values, residuals and variance estimate	51
3.3 Basic statistical properties	52
3.4 Confidence intervals for the coefficients and individual tests	54
4 ANOVA table and R2	56
4.1 Some comments	56
4.2 The ANOVA table and the global test	56
4.3 The R2 coefficient and the adjusted R2	58
5 Confidence Intervals for the mean and Forecast Intervals	60
5.1 Confidence intervals for the mean	60
5.2 Prediction intervals	61
6 Comparing models and testing linear hypotheses	61
7 Selecting variables and model choice	64
8 Residuals and diagnostics	66
8.1 Residuals and h-values	66
8.2 Diagnostic analysis : outliers, influent observations, observations with	
high leverage effect	68
8.3 Diagnostic analysis for Wagesdata using R	69

9 Dealing with departures from model assumptions and practical issues	71
PRACTICAL SESSION 3. MULTIPLE LINEAR REGRESSION (M. Clausel, G. Grégoire)	73
1 An example of multiple linear regression	73
2 Model selection	74
3 Analysis of variance	75
SOME REGRESSION PROBLEMS IN SOLAR-TERRESTRIAL SCIENCES: LEARNING FROM MISTAKES (T. Dudok de Wit)	77
1 What not to do in regression analysis	77
2 Trend determination: is the sky falling down?	78
3 Multilinear regression: desperately searching for solar signatures	80
4 Power laws: are there lines everywhere?	83
5 Conclusion	86
LOGISTIC REGRESSION (G. Grégoire)	89
1 Introduction	89
1.1 The model	90
1.2 Odds and Odds ratio	91
1.3 Interpretation of coefficients in logistic regression	94
2 Maximum likelihood method	95
2.1 Wald, Rao, LR fundamental results in maximum likelihood theory	96
2.2 MLE in logistic regression	97
2.3 WAGES example	99
3 Models comparison. Deviance. R2 Coefficient.	103
3.1 Comparing nested models in general likelihood theory	103
3.2 Comparing nested models in logistic regression	103
3.3 Deviance, Pearson chi-square	104
3.4 Some R2 coefficients	106
4 Residuals, influence measure, h-values, classification tables and ROC curve	107
4.1 Residuals	107
4.2 h-values and leverage effect	108
4.3 Influence detection	108
4.4 Classification table and ROC curve	109
5 Some additional topics	110
5.1 Binomial data	110
5.2 Polytomous logistic regression	113
5.3 Ordinal logistic regression	114
5.4 Discriminating between Stars and Quasars	116
PRACTICAL SESSION 4. LOGISTIC REGRESSION (M. Clausel, G.	

Grégoire)	121
SURVIVAL DATA AND REGRESSION MODELS (G. Grégoire)	125
1 Introduction	125
1.1 Notation and specific statistical tools	126
1.2 Censoring	130
1.3 Statistical issues	131
2 The Kaplan-Meier estimate of S	131
3 The likelihood method	134
3.1 Some reminders on the maximum likelihood method	134
3.2 The likelihood function for censored survival data	135
4 Some parametric models	136
4.1 The exponential distribution	136
4.2 The Weibull distribution	137
4.3 The logistic and log-logistic distributions	138
5 AFT regression models	138
5.1 The Accelerated Failure Time models	138
5.2 The exponential regression model	140
5.3 The Weibull regression model	141
5.4 The log-logistic regression model	141
5.5 AFT regression models dealt with R	142
6 The Cox regression model	144
6.1 Interpretation of β_i coefficients in Cox regression	145
6.2 Maximum likelihood estimation	146
6.3 Stratified Cox model	146
6.4 Fitting a Cox regression model with R	147
LINEAR REGRESSION IN HIGH DIMENSION AND/OR FOR CORRELATED INPUTS (J. Jacques and D. Fraix-Burnet)	149
1 Introduction	149
2 Methods using Derived Input Directions	151
2.1 Regression on Principal Components	151
2.2 Partial Least Square regression	152
2.3 Application on astronomy data	152
2.3.1 The data	152
2.3.2 The R package	153
2.3.3 The analysis	153
3 Shrinkage methods	155
3.1 Ridge regression	156
3.2 Lasso regression	157
3.2.1 Least Angle Regression	158
3.3 Elastic-net regression	159
3.4 Application on astronomy data	159

3.4.1 The R packages	159
3.4.2 The analysis	160
3.5 Extensions	164

AN INTRODUCTION TO DIMENSION REDUCTION IN NONPARAMETRIC KERNEL REGRESSION (S. Girard, J. Saracco)

1 Introduction	167
2 An introduction to nonparametric kernel regression	168
2.1 Nonparametric density estimation	168
2.2 Unidimensional nonparametric regression	170
2.2.1 Moving average estimator	170
2.2.2 Kernel estimator	171
2.3 Multidimensional nonparametric regression	173
2.3.1 Kernel estimator	173
2.3.2 Bias/variance trade-off	175
2.3.3 Bandwidth selection with cross-validation	177
2.3.4 An illustration on simulated data	177
2.4 Some extensions	178
2.4.1 Structured kernels	179
2.4.2 Local polynomial regression	180
2.4.3 Local likelihood	182
3 Dimension reduction based on sliced inverse regression	183
3.1 The semi-parametric regression model	183
3.2 The basic ideas behind SIR	184
3.3 Some extensions of SIR approach	187
3.4 Closest Submodel Selection (CSS)	188
3.5 An illustration on simulated data	189
4 Application to astronomy data	191

SPATIAL PATTERNS ANALYSIS IN COSMOLOGY BASED ON MARKED POINT PROCESSES (R. Stoica)

1 Introduction	197
2 Modelling : build a marked point process	199
2.1 Poisson and interacting point process	200
2.2 Manipulating point processes	202
2.3 Models	205
3 Monte Carlo simulation	209
4 Statistical inference	215
4.1 Parameter estimation	215
4.2 Pattern detection	218
5 Morpho-statistical characterization of cosmic filaments	225
6 Conclusion and perspectives	226

Codes and Data **227**