

Actuarial Statistics with R: Theory and Case Studies

Guojun Gan, PhD, FSA
Assistant Professor
Department of Mathematics
University of Connecticut
Storrs, CT, USA

Emiliano A. Valdez, PhD, FSA
Professor
Department of Mathematics
University of Connecticut
Storrs, CT, USA

April 18, 2018

Contents

Preface	xv
About the Authors	xvii
I Supervised Learning	1
1 Simple Linear Regression	3
1.1 Scatter Plots and Regression	3
1.2 Simple Linear Regression Model	5
1.3 Ordinary Least Squares Estimation	7
1.4 Model Evaluation	9
1.5 Statistical Inference	12
1.6 Residual Analysis	14
1.7 Summary	16
1.8 End-of-Chapter Exercises	16
2 Case Study: Implementing the Capital Asset Pricing Model	19
2.1 Problem Description	19
2.2 Data Description	20
2.3 Loading the Data into R	21
2.4 Data Visualization and Summarization	23
2.5 Fitting a Basic Linear Regression Model	26
2.6 Model Evaluation	29
2.7 Residual Analysis	31
2.8 Statistical Inference	35
2.9 Summary	36
2.10 End-of-Chapter Exercises	36

3	Multiple Linear Regression Models	41
3.1	Scatter Plot Matrix	41
3.2	Independent Variables and Regressors	44
3.3	Multiple Linear Regression Model	46
3.4	Ordinary Least of Squares Estimation	47
3.5	Model Evaluation	49
3.6	Statistical Inference	51
3.7	Transformations	55
3.8	Regression Diagnostics	55
3.9	Variable Selection	58
3.10	Collinearity	59
3.11	Summary	60
3.12	End-of-Chapter Exercises	60
4	Case Study: Predicting Intraday Movements	63
4.1	Problem Description	63
4.2	Data Description	64
4.3	Multiple Linear Regression	65
4.4	Fitting a Multiple Linear Regression Model	67
4.5	Model Evaluation	70
4.6	Model Selection	71
4.7	Influential Points	74
4.8	Collinearity	78
4.9	Heteroscedasticity	79
4.10	Statistical Inference	82
4.11	Summary	82
4.12	End-of-Chapter Exercises	83
5	Case Study: Estimating Fair Market Values	87
5.1	Problem Description	87
5.2	Data Description	88
5.3	Loading the Data into R	89
5.4	Selecting Variables and Preparing Training Data	90
5.5	Categorical Variables	92
5.6	Building a Multiple Linear Regression Model	93
5.7	Model Evaluation	96
5.8	Statistical Inference for Several Coefficients	98
5.9	Summary	100

6	Generalized Linear Models	101
6.1	Linear Exponential Family of Distributions	101
6.2	GLM Models	103
6.3	Maximum Likelihood Estimation	105
6.4	Model Evaluation	107
6.5	Residuals	109
6.6	Summary	110
6.7	End-of-Chapter Exercises	110
7	Case Study: Predicting Demand for Term Life Insurance	115
7.1	Problem Description	115
7.2	Data Description	116
7.3	Loading Data into R	116
7.4	Binary Dependent Variables	118
7.5	Logistic and Probit Regression Models	120
7.6	The Method of Maximum Likelihood	122
7.7	Model Evaluation	124
7.8	Summary	126
7.9	End-of-Chapter Exercises	127
8	Case Study: Modeling the Number of Auto Claims	129
8.1	Problem Description	129
8.2	Data Description	130
8.3	Poisson Regression Models	132
8.4	Negative Binomial Regression Models	135
8.5	Model Evaluation	139
8.6	Summary	141
8.7	End-of-Chapter Exercises	141
9	Case Study: Modeling the Loss Severity of Auto Claims	143
9.1	Problem Description	143
9.2	Data Description	145
9.3	The Gamma Regression Model	148
9.4	Fitting the Gamma Regression Model	151
9.5	Prediction	154
9.6	Model Evaluation	155
9.7	Summary	157

10 Decision Trees	159
10.1 Tree-Based Models	159
10.1.1 Regression Trees	159
10.1.2 Classification Trees	164
10.2 Prediction Models	165
10.2.1 Bagging	166
10.2.2 Boosting	166
10.2.3 Random Forests	167
10.3 Comparison with Linear Models	168
10.4 Summary	169
10.5 End-of-Chapter Exercises	169
11 Case Study: Decision Trees	173
11.1 Preparing Data	173
11.2 Fitting Regression Trees	174
11.3 Pruning Trees	179
11.4 Prediction with a Single Tree	181
11.5 Prediction with Many Trees	184
11.6 Summary	190
11.7 End-of-Chapter Exercises	190
II Unsupervised Learning	191
12 Data Clustering	193
12.1 The Basics of Data Clustering	193
12.2 Hierarchical Algorithms	197
12.3 Partitional Algorithms	201
12.4 Summary	203
12.5 End-of-Chapter Exercises	203
13 Case Study: Clustering Variable Annuity Policies	205
13.1 Hierarchical k-means	205
13.2 Preparing Data	206
13.3 Performing Data Clustering	210
13.4 Predictive Modeling Results	213
13.5 Summary	216

14 Principal Component Analysis	219
14.1 Principal Components	219
14.2 Empirical Principal Components	221
14.3 Computing Principal Components	222
14.4 Other Issues	224
14.5 Summary	224
14.6 End-of-Chapter Exercises	224
15 Case Study: PCA on Interest Rate Swaps	225
15.1 Interest Rate Swaps	225
15.2 Loading Swap Rates into R	226
15.3 Principal Component Analysis	229
15.4 Summary	232
III Time Series Models	233
16 Time Series Models	235
16.1 Introduction	235
16.2 Trend Models	236
16.3 Random Walk Models	236
16.4 Autoregressive Models	238
16.5 ARIMA Models	241
16.6 Smoothing Techniques	242
16.7 ARCH	243
16.8 Model Evaluation	244
16.9 Summary	246
16.10 End-of-Chapter Exercises	246
17 Case Study: Forecasting Currency Exchange Rates	249
17.1 Data Description	249
17.2 Loading Data into R	250
17.3 Fitting Trend Models	252
17.4 Fitting Random Walk Models	254
17.5 Fitting Autoregressive Models	257
17.6 ARIMA Models	259
17.7 Forecast Evaluation	260
17.8 Summary	263

IV Simulation	265
18 Case Study: Profitability Analysis of Short Term Contracts	267
18.1 Introduction to Simulation	267
18.2 Simulating Discrete Random Variables	270
18.3 Simulating Continuous Random Variables	276
18.4 Problem Description	278
18.5 Summary	289
19 Case Study: Simulating the Future Lifetime of a Person	291
19.1 Time-Until-Death Random Variables	291
19.2 Curtate Future Lifetime	295
19.3 Expectation of Life	296
19.4 Simulating Gompertz Lifetime Distributions	297
19.5 Applications to Life Insurance Pricing and Reserving	298
19.6 Simulating Makeham Lifetime Distributions	302
19.7 Simulating from a Mortality Table	305
19.8 Summary	310
A Introduction to R	313
A.1 How to Run R	314
A.2 Variables	317
A.3 Vectors	318
A.4 Matrices	325
A.5 Lists	334
A.6 Data Frames	339
A.7 Factors and Tables	342
A.8 File IO	346
A.9 Functions	350
A.10 Flow Control and Loops	352
A.11 Graphics	357
A.12 Packages	362
A.13 Summary	364
References	365
Index	370
Index of R Functions	374