

Modern Regression Methods

Second Edition

THOMAS P. RYAN

Acworth, Georgia



WILEY

A JOHN WILEY & SONS, INC. PUBLICATION

Contents

Preface	ix
1. Introduction	1
1.1 Simple Linear Regression Model, 3	
1.2 Uses of Regression Models, 4	
1.3 Graph the Data!, 5	
1.4 Estimation of β_0 and β_1 , 6	
1.4.1 Orthogonal Regression, 11	
1.5 Inferences from Regression Equations, 11	
1.5.1 Predicting Y , 12	
1.5.2 Worth of the Regression Equation, 13	
1.5.3 Regression Assumptions, 15	
1.5.4 Inferences on β_1 , 17	
1.5.5 Inferences on β_0 , 21	
1.5.6 Inferences for Y , 22	
1.5.6.1 Prediction Interval for Y , 22	
1.5.6.2 Confidence Interval for $\mu_{Y X}$, 25	
1.5.7 ANOVA Tables, 25	
1.5.8 Lack of Fit, 26	
1.6 Regression Through the Origin, 29	
1.7 Additional Examples, 30	
1.8 Correlation, 31	
1.9 Miscellaneous Uses of Regression, 32	
1.9.1 Regression for Control, 32	
1.9.2 Inverse Regression, 33	
1.9.3 Regression Control Chart, 37	
1.9.4 Monitoring Linear Profiles, 37	

- 1.10 Fixed Versus Random Regressors, 37
- 1.11 Missing Data, 38
- 1.12 Spurious Relationships, 38
- 1.13 Software, 39
- 1.14 Summary, 40
 - Appendix, 41
 - References, 45
 - Exercises, 48

2. Diagnostics and Remedial Measures

60

- 2.1 Assumptions, 61
 - 2.1.1 Independence, 61
 - 2.1.1.1 Correlated Errors, 64
 - 2.1.1.1.1 An Example, 65
 - 2.1.1.1.2 Corrective Action, 68
 - 2.1.2 Normality, 68
 - 2.1.2.1 Supernormality Property of Residuals, 69
 - 2.1.2.2 Standardized Deletion Residuals, 70
 - 2.1.2.3 Methods of Constructing Simulation Envelopes, 70
 - 2.1.3 Constant Variance, 77
 - 2.1.3.1 Weighted Least Squares, 77
 - 2.1.3.1.1 Unknown Weights, 81
 - 2.1.3.1.2 Modeling the Variance, 83
 - 2.1.3.2 A Heteroscedastic Alternative, 87
- 2.2 Residual Plots, 88
- 2.3 Transformations, 89
 - 2.3.1 Transforming the Model, 89
 - 2.3.2 Transforming the Regressors to Improve the Fit, 91
 - 2.3.2.1 Box-Tidwell Transformation, 92
 - 2.3.3 Transform Y to Obtain a Better Fit?, 95
 - 2.3.4 Transforming to Correct Heteroscedasticity and Nonnormality, 96
 - 2.3.5 Which R^2 , 97
- 2.4 Influential Observations, 98
 - 2.4.1 An Example, 99
 - 2.4.2 Influence Statistics, 103
 - 2.4.3 Different Schools of Thought Regarding Influence, 104
 - 2.4.4 Modification of Standard Influence Measures, 105

- 2.4.5 Application of Influence Measures to Table 2.7 Data, 105
- 2.4.6 Multiple Unusual Observations, 106
- 2.4.7 Predicting Lifespan (?): An Influential Data Problem, 107
- 2.5 Outliers, 108
- 2.6 Measurement Error, 111
 - 2.6.1 Measurement Error in Y , 111
 - 2.6.2 Measurement Error in X , 112
- 2.7 Software, 112
- 2.8 Summary, 114
 - Appendix, 114
 - References, 116
 - Exercises, 120

3. Regression with Matrix Algebra 128

- 3.1 Introduction to Matrix Algebra, 128
 - 3.1.1 Eigenvalues and Eigenvectors, 130
- 3.2 Matrix Algebra Applied to Regression, 133
 - 3.2.1 Predicted Y and R^2 , 135
 - 3.2.2 Estimation of σ_ϵ^2 , 136
 - 3.2.3 Variance of Y and \widehat{Y} , 136
 - 3.2.4 Centered Data, 138
 - 3.2.5 Correlation Form, 139
 - 3.2.6 Influence Statistics in Matrix Form, 140
- 3.3 Summary, 141
 - Appendix, 141
 - References, 142
 - Exercises, 143

4. Introduction to Multiple Linear Regression 146

- 4.1 An Example of Multiple Linear Regression, 147
 - 4.1.1 Orthogonal Regressors, 150
 - 4.1.2 Correlated Regressors, 151
 - 4.1.2.1 Partial- F Tests and t -Tests, 153
 - 4.1.2.2 Individual Regressor Effects, 155
 - 4.1.3 Confidence Intervals and Prediction Intervals, 156
- 4.2 Centering And Scaling, 158
 - 4.2.1 Centering, 158
 - 4.2.2 Scaling, 159

4.3	Interpreting Multiple Regression Coefficients, 161	
4.3.1	Multicollinearity and the “Wrong Signs” Problem, 167	
4.3.2	So Are Individual Regression Coefficients Interpretable?, 168	
4.3.3	Inflated Variances, 169	
4.3.4	Detecting Multicollinearity, 169	
4.3.5	Variance Proportions, 173	
4.3.6	What to Do About Multicollinearity?, 174	
4.4	Indicator Variables, 175	
4.5	Separation or Not?, 176	
4.6	Alternatives to Multiple Regression, 176	
4.7	Software, 176	
4.8	Summary, 177	
	References, 178	
	Exercises, 181	
5.	Plots in Multiple Regression	190
5.1	Beyond Standardized Residual Plots, 190	
5.1.1	Partial Residual Plots, 191	
5.1.2	CCPR Plot, 193	
5.1.3	Augmented Partial Residual Plots, 194	
5.1.4	CERES Plots, 194	
5.2	Some Examples, 196	
5.3	Which Plot?, 208	
5.3.1	Relationships Between Plots, 209	
5.3.2	True Model Contains Nonlinear Terms, 211	
5.4	Recommendations, 212	
5.5	Partial Regression Plots, 213	
5.5.1	Examples, 216	
5.5.2	Detrended Added Variable Plot, 217	
5.5.3	Partial Regression Plots Used to Detect Influential Observations, 218	
5.6	Other Plots For Detecting Influential Observations, 222	
5.7	Recent Contributions to Plots in Multiple Regression, 223	
5.8	Lurking Variables, 225	
5.9	Explanation of Two Data Sets Relative to R^2 , 225	
5.10	Software, 226	
5.11	Summary, 227	
	References, 228	
	Exercises, 230	

6. Transformations in Multiple Regression	234
6.1 Transforming Regressors, 234	
6.2 Transforming Y , 238	
6.2.1 Transformation Needed But Not Suggested, 238	
6.2.2 Transformation Needed and Suggested, 240	
6.2.3 Transformation Apparently Successful, 241	
6.3 Further Comments on the Normality Issue, 242	
6.4 Box–Cox Transformation, 243	
6.5 Box–Tidwell Revisited, 247	
6.6 Combined Box–Cox and Box–Tidwell Approach, 247	
6.6.1 Table 6.2 Data, 248	
6.6.2 Table 6.3 Data, 249	
6.6.3 Table 6.4 Data, 250	
6.6.4 MINITAB Tree Data, 250	
6.6.4.1 Other Analyses of the Tree Data, 253	
6.6.5 Stack Loss Data, 255	
6.6.6 Palm Beach County Data, 257	
6.7 Other Transformation Methods, 258	
6.7.1 Transform Both Sides (TBS), 259	
6.8 Transformation Diagnostics, 260	
6.8.1 Diagnostics <i>After</i> a Transformation, 261	
6.9 Software, 261	
6.10 Summary, 262	
References, 262	
Exercises, 265	
7. Selection of Regressors	269
7.1 Forward Selection, 270	
7.2 Backward Elimination, 271	
7.3 Stepwise Regression, 272	
7.3.1 Significance Levels, 272	
7.4 All Possible Regressions, 272	
7.4.1 Criteria, 273	
7.4.1.1 Mallows's C_p , 273	
7.4.1.1.1 C_p and Influential Data, 276	
7.4.1.2 Minimum $\hat{\sigma}^2$, 277	
7.4.1.3 t -Statistics, 277	
7.4.1.4 Other Criteria, 277	

7.5	Newer Methods, 277	
7.5.1	Robust Variable Selection, 278	
7.6	Examples, 279	
7.7	Variable Selection for Nonlinear Terms, 280	
7.7.1	Negative C_p Values, 282	
7.8	Must We Use a Subset?, 283	
7.9	Model Validation, 284	
7.10	Software, 284	
7.11	Summary, 285	
	Appendix, 286	
	References, 287	
	Exercises, 290	
8.	Polynomial and Trigonometric Terms	296
8.1	Polynomial Terms, 296	
8.1.1	Orthogonal Polynomial Regression, 299	
8.1.1.1	When to Stop?, 299	
8.1.2	An Example, 300	
8.2	Polynomial–Trigonometric Regression, 302	
8.2.1	Orthogonality of Trigonometric Terms, 303	
8.2.2	Practical Considerations, 303	
8.2.3	Examples, 303	
8.2.4	Multiple Independent Variables, 307	
8.3	Software, 307	
8.4	Summary, 307	
	References, 308	
	Exercises, 309	
9.	Logistic Regression	312
9.1	Introduction, 312	
9.2	One Regressor, 313	
9.2.1	Estimating β_0 and β_1 , 315	
9.2.1.1	Method of Maximum Likelihood, 315	
9.2.1.2	Exact Logistic Regression, 319	
9.3	A Simulated Example, 320	
9.3.1	Complete and Quasicomplete Separation, 320	
9.3.2	Overlap: Modifying Table 9.1, 325	
9.4	Detecting Complete Separation, Quasicomplete Separation and Near Separation, 326	

- 9.5 Measuring the Worth of the Model, 326
 - 9.5.1 R^2 in Logistic Regression, 327
 - 9.5.2 Deviance, 328
 - 9.5.3 Other Measures of Model Fit, 329
- 9.6 Determining the Worth of the Individual Regressors, 330
 - 9.6.1 Wald Test, 330
 - 9.6.2 Likelihood Ratio Test, 331
 - 9.6.3 Scores Test, 331
 - 9.6.4 Exact Conditional Scores Test, 332
 - 9.6.5 Exact p -Value, 333
- 9.7 Confidence Intervals, 333
 - 9.7.1 Confidence Interval for β_i , 333
 - 9.7.2 Confidence Interval for Change in Odds Ratio, 334
 - 9.7.3 Confidence Interval for π , 334
 - 9.7.4 Exact Confidence Intervals, 335
 - 9.7.4.1 Exact Confidence Interval for β_1 , 335
 - 9.7.4.2 Exact Confidence Interval for Change in Odds Ratio, 336
- 9.8 Exact Prediction, 336
 - 9.8.1 Exact Confidence Interval for π , 337
- 9.9 An Example With Real Data, 337
 - 9.9.1 Hosmer–Lemeshow Goodness-of-Fit Tests, 340
 - 9.9.2 Which Residuals?, 343
 - 9.9.3 Application to Table 9.4 Data, 345
 - 9.9.3.1 Pearson Residuals, 346
 - 9.9.3.2 Deviance Residuals, 348
 - 9.9.4 Other Diagnostics, 349
 - 9.9.5 Partial Residual Plot, 350
 - 9.9.6 Added Variable Plot, 351
 - 9.9.7 Confidence Intervals for Table 9.3 Data, 352
- 9.10 An Example of Multiple Logistic Regression, 352
 - 9.10.1 Correct Classification Rate for Full Data Set, 355
 - 9.10.2 Influential Observations, 356
 - 9.10.3 Which Variables?, 357
 - 9.10.3.1 Algorithmic Approaches to Variable Selection, 359
 - 9.10.3.2 What About Nonlinear Terms?, 361
- 9.11 Multicollinearity in Multiple Logistic Regression, 362

- 9.12 Osteogenic Sarcoma Data Set, 366
- 9.13 Missing Data, 369
- 9.14 Sample Size Determination, 369
- 9.15 Polytomous Logistic Regression, 370
- 9.16 Logistic Regression Variations, 372
- 9.17 Alternatives to Logistic Regression, 373
- 9.18 Software for Logistic Regression, 373
- 9.19 Summary, 375
 - Appendix, 375
 - References, 376
 - Exercises, 381

10. Nonparametric Regression

385

- 10.1 Relaxing Regression Assumptions, 386
 - 10.1.1 Bootstrapping, 386
- 10.2 Monotone Regression, 387
- 10.3 Smoothers, 390
 - 10.3.1 Running Line, 393
 - 10.3.1.1 Inferences for Running Line, 397
 - 10.3.2 Kernel Regression, 398
 - 10.3.2.1 Inferences in Kernel Regression, 399
 - 10.3.3 Local Regression, 400
 - 10.3.3.1 Inferences and Diagnostics, 403
 - 10.3.4 Splines, 403
 - 10.3.4.1 Piecewise Linear Regression (Linear Splines), 403
 - 10.3.4.1.1 Model Representation, 405
 - 10.3.4.2 Splines with Polynomial Terms, 405
 - 10.3.4.3 Smoothing Splines, 407
 - 10.3.4.4 Splines Compared to Local Regression, 408
 - 10.3.5 Other Smoothers, 409
 - 10.3.6 Which Smoother?, 409
 - 10.3.7 Smoothers for Multiple Regressors, 409
- 10.4 Variable Selection, 410
- 10.5 Important Considerations in Smoothing, 410
- 10.6 Sliced Inverse Regression, 410
- 10.7 Projection Pursuit Regression, 411
- 10.8 Software, 411

- 10.9 Summary, 412
- Appendix, 413
- References, 414
- Exercises, 418

11. Robust Regression 421

- 11.1 The Need for Robust Regression, 421
- 11.2 Types of Outliers, 423
- 11.3 Historical Development of Robust Regression, 426
 - 11.3.1 Breakdown Point, 427
 - 11.3.2 Efficiency, 427
 - 11.3.3 Classes of Estimators, 428
 - 11.3.3.1 M -Estimators, 428
 - 11.3.3.2 Bounded Influence Estimators, 428
 - 11.3.3.3 High Breakdown Point Estimators, 429
 - 11.3.3.4 Two-Stage Procedures, 429
 - 11.3.3.5 MM -Estimator (Three Stages), 429
- 11.4 Goals of Robust Regression, 430
- 11.5 Proposed High Breakdown Point Estimators, 430
 - 11.5.1 Least Median of Squares, 430
 - 11.5.2 Least Trimmed Squares, 432
 - 11.5.2.1 LTS Applications, 434
 - 11.5.3 S -Estimators, 434
- 11.6 Approximating HBP Estimator Solutions, 435
 - 11.6.1 Application to Hawkins–Bradu–Kass Data Set, 436
 - 11.6.2 Another Application: One Regressor, 440
 - 11.6.3 A Proposed Sequential Procedure, 441
 - 11.6.4 Application to Multiple Regression, 442
- 11.7 Other Methods for Detecting Multiple Outliers, 446
- 11.8 Bounded Influence Estimators, 446
 - 11.8.1 Shortcomings of Bounded Influence Estimators, 449
 - 11.8.2 Application of Welsh Estimator, 450
- 11.9 Multistage Procedures, 452
- 11.10 Other Robust Regression Estimators, 454
- 11.11 Applications, 456
- 11.12 Software for Robust Regression, 456
- 11.13 Summary, 457
 - References, 458
 - Exercises, 462

12. Ridge Regression	466
12.1 Introduction, 466	
12.2 How Do We Determine k ?, 470	
12.3 An Example, 471	
12.4 Ridge Regression for Prediction?, 476	
12.5 Generalized Ridge Regression, 477	
12.6 Inferences in Ridge Regression, 477	
12.7 Some Practical Considerations, 477	
12.8 Robust Ridge Regression?, 478	
12.9 Recent Developments in Ridge Regression, 478	
12.10 Other Biased Estimators, 479	
12.11 Software, 480	
12.12 Summary, 480	
Appendix, 481	
References, 482	
Exercises, 485	
13. Nonlinear Regression	488
13.1 Introduction, 488	
13.2 Linear Versus Nonlinear Regression, 489	
13.3 A Simple Nonlinear Example, 489	
13.3.1 Iterative Estimation, 491	
13.4 Relative Offset Convergence Criterion, 493	
13.5 Adequacy of the Estimation Approach, 494	
13.6 Computational Considerations, 495	
13.7 Determining Model Adequacy, 496	
13.7.1 Lack-of-Fit Test, 496	
13.7.2 Residual Plots, 497	
13.7.3 Multicollinearity Diagnostics, 498	
13.7.4 Influence and Unusual Data Diagnostics, 499	
13.7.4.1 Leverage, 499	
13.7.4.2 Influence, 499	
13.8 Inferences, 501	
13.8.1 Confidence Intervals, 501	
13.8.2 Prediction Interval, 502	
13.8.3 Hypothesis Tests, 502	
13.9 An Application, 502	
13.9.1 When Is a Linear Fit Not Good Enough?, 507	

- 13.10 Rational Functions, 510
- 13.11 Robust Nonlinear Regression, 510
- 13.12 Applications, 510
- 13.13 Teaching Tools, 511
- 13.14 Recent Developments, 511
- 13.15 Software, 511
 - 13.15.1 SAS Software, 511
 - 13.15.1.1 Cautions, 512
 - 13.15.2 SPSS, 512
 - 13.15.3 BMDP, 512
 - 13.15.4 S-Plus and *R*, 512
 - 13.15.5 MINITAB, 513
- 13.16 Summary, 513
 - Appendix, 513
 - References, 516
 - Exercises, 520

14. Experimental Designs for Regression

525

- 14.1 Objectives for Experimental Designs, 525
- 14.2 Equal Leverage Points, 525
 - 14.2.1 Simple Linear Regression, 526
 - 14.2.2 Multiple Linear Regression, 526
 - 14.2.2.1 Construction of Equileverage Designs—
Two Regressors, 527
 - 14.2.2.1.1 Inverse Projection
Approach, 531
- 14.3 Other Desirable Properties of Experimental Designs, 537
 - 14.3.1 *D*-Optimality, 537
 - 14.3.2 *G*-Optimality, 538
 - 14.3.3 Other Optimality Criteria, 539
- 14.4 Model Misspecification, 540
- 14.5 Range of Regressors, 541
- 14.6 Algorithms for Design Construction, 541
- 14.7 Designs for Polynomial Regression, 541
- 14.8 Designs for Logistic Regression, 542
- 14.9 Designs for Nonlinear Regression, 542
- 14.10 Software, 543
- 14.11 Summary, 543
 - References, 544
 - Exercises, 547

- 15. Miscellaneous Topics in Regression** **550**
- 15.1 Piecewise Regression and Alternatives, 550
 - 15.2 Semiparametric Regression, 550
 - 15.3 Quantile Regression, 552
 - 15.4 Poisson Regression, 556
 - 15.4.1 Exact Poisson Regression, 560
 - 15.4.2 Zero-Inflated Poisson Regression, 561
 - 15.4.3 Zero-Truncated Poisson Regression, 562
 - 15.5 Negative Binomial Regression, 562
 - 15.5.1 Zero-Inflated Negative Binomial Regression, 563
 - 15.5.2 Zero-Truncated Negative Binomial Regression, 563
 - 15.6 Cox Regression, 564
 - 15.7 Probit Regression, 564
 - 15.8 Censored Regression and Truncated Regression, 565
 - 15.8.1 Tobit Regression, 566
 - 15.9 Constrained Regression, 566
 - 15.10 Interval Regression, 567
 - 15.11 Random Coefficient Regression, 567
 - 15.12 Partial Least Squares Regression, 568
 - 15.13 Errors-in-Variables Regression, 568
 - 15.14 Regression with Life Data, 568
 - 15.15 Use of Regression in Survey Sampling, 569
 - 15.16 Bayesian Regression, 569
 - 15.17 Instrumental Variables Regression, 570
 - 15.18 Shrinkage Estimators, 571
 - 15.19 Meta-Regression, 571
 - 15.20 Classification and Regression Trees (CART), 571
 - 15.21 Multivariate Regression, 572
 - References, 572
 - Exercises, 576
- 16. Analysis of Real Data Sets** **577**
- 16.1 Analyzing Buchanan's Presidential Vote in Palm Beach County in 2000, 577
 - 16.2 Water Quality Data, 578
 - 16.3 Predicting Lifespan?, 588
 - 16.4 Scottish Hill Races Data, 591
 - 16.5 Leukemia Data, 593
 - 16.5.1 Y Binary, 593
 - 16.5.2 Y Continuous, 598

16.6	Dosage Response Data, 599	
16.7	A Strategy for Analyzing Regression Data, 602	
16.8	Summary, 604	
	References, 604	
	Exercises, 606	
Answers to Selected Exercises		609
Statistical Tables		617
Author Index		625
Subject Index		637