Smoothing and Semiparametric Regression For The
Social Sciences

Luke Keele
Ohio State University

April 26, 2007
Contents

1 Introduction: Global vs Local Statistics ........................................ 11
  1.1 The Consequences of Ignoring Nonlinearity ............................. 15
  1.2 Transformations ...................................................................... 17
  1.3 Nonparametric and Semiparametric Techniques ......................... 20
  1.4 Outline Of The Text .................................................................. 24

2 Smoothing and Local Regression .................................................. 27
  2.1 Simple Smoothing ..................................................................... 27
    2.1.1 Local Averaging .................................................................. 29
    2.1.2 Kernel Smoothing ................................................................ 36
  2.2 Local Polynomial Regression .................................................... 42
  2.3 Nonparametric Modeling Choices .............................................. 48
    2.3.1 The Span ........................................................................... 48
    2.3.2 Polynomial Degree and Weight Function .............................. 52
    2.3.3 A Note On Interpretation .................................................... 55
  2.4 Statistical Inference for Local Polynomial Regression ................. 56
  2.5 Multiple Nonparametric Regression ........................................... 64
  2.6 Conclusion .............................................................................. 67
  2.7 Exercises .................................................................................. 67

3 Splines ............................................................................................ 71
  3.1 Simple Regression Splines .......................................................... 72
    3.1.1 Basis Functions ................................................................... 75
  3.2 Other Spline Models and Bases ................................................... 78
    3.2.1 Quadratic and Cubic Spline Bases ........................................ 79
    3.2.2 Natural Splines .................................................................... 84
    3.2.3 B-splines ............................................................................ 84
    3.2.4 Knot Placement .................................................................... 85
    3.2.5 Comparing Spline Models .................................................... 87
  3.3 Splines and Overfitting ............................................................... 91
    3.3.1 Smoothing Splines .............................................................. 91
    3.3.2 Splines as Mixed Models ..................................................... 98
    3.3.3 Final Notes on Smoothing Splines ........................................ 103
8 Bootstrapping 233
  8.1 Classical Inference ........................................... 233
  8.2 Bootstrapping - An Overview ................................ 235
    8.2.1 Bootstrapping the Mean ................................. 236
    8.2.2 Bootstrapping Regression Models ....................... 243
    8.2.3 An Example: Presidential Elections .................... 245
  8.3 Bootstrapping Nonparametric and Semiparametric Regression Models .... 248
    8.3.1 Bootstrapping Nonparametric Fits ...................... 248
    8.3.2 Bootstrapping Nonlinearity Tests ...................... 249
  8.4 Conclusion .................................................... 253
  8.5 Exercises .................................................... 254

9 Epilogue 257

10 Appendix: Software 261
List of Tables

3.1 AIC Values for Differing Numbers of Knots ........................................... 89
4.1 Mean Squared Error For Smoother Fits .................................................. 134
4.2 Likelihood Ratio Test Between Cubic and Smoothing Splines ................. 143
5.1 Testing Spline Fits Against Linear Fits and Transformations: Challenger’s
    Vote Share ...................................................................................................... 171
5.2 A Comparison Of Parametric And Semiparametric Models of Congressional
    Voting .............................................................................................................. 172
6.1 A Comparison Of Parametric And Semiparametric Models of Supreme Court
    Overrides ......................................................................................................... 205
6.2 Grambsch and Therneau Nonproportionality Tests for Time to Onset of Racial
    Riots, 410 US Cities, 1961 to 1968 ............................................................... 210
8.1 Incumbent Vote Share: Standard and Bootstrapped Models ................. 246
# List of Figures

1.1 A linear fit and the true quadratic relationship with simulated data. . . . . 16
1.2 Residual plotted against fitted values with a nonparametric regression esti-
mate. .................................................. 22

2.1 A Scatterplot of Challengers’ Vote Share and Perot Vote, 1992 ............ 32
2.2 A Moving Average Smoother and an OLS Regression Line ................. 34
2.3 A Moving Average Smoother: Varying the Bin Width of the Smoother ... 35
2.4 An Example Bin For the Congressional Vote Data .......................... 39
2.5 The Tricube Weighted Average ............................................. 40
2.6 Kernel Smoother with 3 Different Bandwidths ............................. 41
2.7 Local Linear Regression: Fitted Y At The Focal Point ..................... 43
2.8 Loess and Lowess Fit to Challengers’ Vote Share and Support Perot, 1992 47
2.9 The Bias Variance Tradeoff in Smoothing ............................... 50
2.10 Selecting the Span for a Lowess Fit to the Data .......................... 53
2.11 Varying the Polynomial Degree for Local Polynomial Regression ....... 54
2.12 95% Confidence Bands for Loess Fit to the Challengers’ Vote Share Data 62
2.13 The Joint Effect of Overdrafts and Challenger Spending on the Challengers’  Vote Share .......................................................... 66

3.1 A Simulated Nonlinear Functional Form with OLS Fit ...................... 74
3.2 A Piecewise Linear Regression Spline Fit .................................. 78
3.3 Support For The Challenger as a Function of support for Perot in the 1992  House Elections. .................................................. 83
3.4 Cubic B-Spline and Natural Spine Fit to Challengers’ Vote Share ......... 88
3.5 Differing Numbers of Knots with Natural Splines .......................... 90
3.6 Smoothing Spline Fit To Challenger Vote Share Data With Four Different  Degrees of Freedom Values. ........................................ 96
3.7 Smoothing Spline Fit To Challenger Vote Share Data With Constant Degrees of Freedom And Different Knot Numbers. ......................... 97
3.8 Confidence Bands for Natural Cubic Spline Fit to Challengers’ Vote Share 107
3.9 A Comparison of Pointwise Variability Bands and Bias Adjusted Confidence  Intervals. .................................................. 109
3.10 Plot of First Derivative for the Effect of Income on Democratic Vote Share 112
3.11 Comparison Of Smoother Fits. Dotted Line Is True Relationship, Solid Line  is Nonparametric Estimate. ........................................ 115
4.1 GCV Plot for Local Polynomial Regression Span Selection ............... 123
4.2 Lowess Fit With Span Selection Via GCV .................................. 124
4.3 Lowess Fit With Span Selection Via GCV ................................. 128
4.4 A Comparison of Estimated Smoother Fits. ................................. 133
4.5 A Simulated Nonlinear Smooth Functional Form ............................ 135
4.6 Comparing Fits Between Smoothing Splines and Lowess with Various Spans 136
4.7 Lowess Span Chosen By MSE and Smoothing Spline Fit ................. 137
4.8 Manual Selection of Degrees of Freedom Compared to Automatic Selection For Overdrafts and Challengers’ Vote Share. ......................... 140
4.9 Scatterplot of Country Level Votes For Al Gore and Overvotes with Smoothing Spline Fits, Florida 2000 ............................... 142
4.10 Standardized Residual Plots With Smoothing Spline Fits ................ 146

5.1 Additive Model For Challengers’ Vote Share as a Function of Support for Perot and Number of Overdrafts ................................. 168
5.2 Nonparametric Effects for Semiparametric Regression Model of Congressional Challenger’s Vote Share, 1992 ............................... 175
5.3 Interaction Between Support for Perot and Marginal Districts .......... 177
5.4 Interaction Between Challenger Spending and the Number of Overdrafts .. 178
5.5 The Effect of Income Proprtion, Hours Worked, and Age on Feminist Attitudes, Women (1987-2000) ........................................ 181
5.6 The Effect of Income Proprtion On Feminist Attitudes With Manual Smoothing Selection ......................................................... 182

6.1 The Nonparametric Estimates For Risk of Militarized Dispute .......... 197
6.2 The Effect of Democracy on Militarized Conflict - Probability Scale 199
6.3 The Effect of Relationship Duration on Couple Violence ................. 202
6.4 Nonparametric Estimate For Congressional Counter Variable .......... 206
6.5 The Difference in The Effect of Tenure Across Parametric and Semiparametric Models ......................................................... 208
6.6 The Effect of the Manufacturing Wages and non-White unemployment on the Hazard of Racial Riots, 1961-1968 ............................. 212

7.1 The Effect of Student SES on Math Achievement For Parochial and .... 222
7.2 Comparison of Bayesian and Frequentist Estimates of Challenger Spending on Challengers’ Vote Share ................................. 227

8.1 QQ-plot of Bootstrap Sampling Distribution .................................. 242
8.2 Histogram and QQ-plot of Bootstrap Sampling Distribution for Policy Moderation ................................................................. 247
8.3 Plug-in and Bootstrap Confidence Intervals for Effect of Perot Vote on Challenger Vote Share ......................................................... 250