Lecture 11
Advanced Topics I

STAT 8020 Statistical Methods II
September 24, 2020

Whitney Huang
Clemson University

Notes

Agenda

1. Nonlinear Regression
2. Non-parametric Regression
3. Ridge Regression

Notes

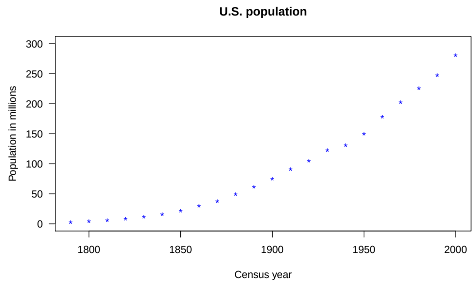
Moving Away From Linear Regression

- We have mainly focused on linear regression so far
- The class of polynomial regression can be thought as a starting point for relaxing the linear assumption
- In this lecture we are going to discuss non-linear and non-parametric regression modeling

Notes

Population of the United States

Let's look at the `USPop` data set, a bulit-in data set in R. This is a decennial time-series from 1790 to 2000.

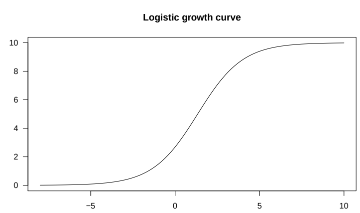


Notes

Logistic Growth Curve

A simple model for population growth is the logistic growth model,

$$
\begin{aligned}
Y &= m(X, \boldsymbol{\phi}) + \varepsilon \\
&= \frac{\phi_1}{1 + \exp\left[-(x - \phi_2)/\phi_3\right]} + \varepsilon
\end{aligned}
$$

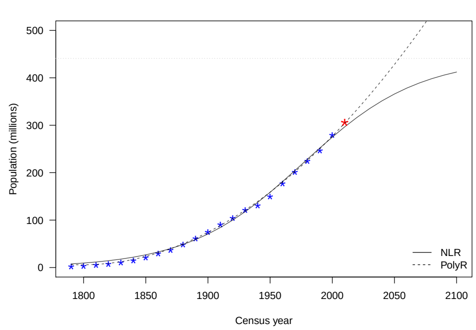


We are going to fit a logistic growth curve to the U.S. population data set

Notes

Fitting logistic growth curve to the U.S. population

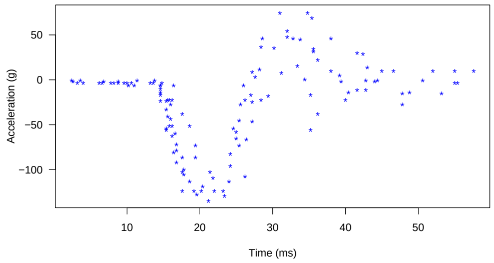
$$\hat{\phi}_1 = 440.83,\ \hat{\phi}_2 = 1976.63,\ \hat{\phi}_3 = 46.29$$



Notes

Non-parametric Regression

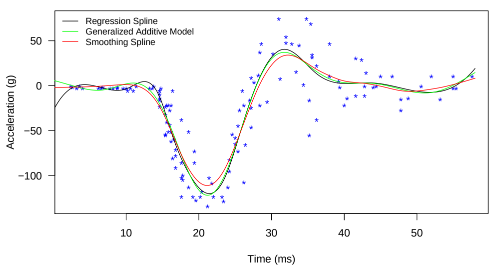
Let's use the motor-cycle impact data as an illustrative example. This data set is taken from a simulated motor-cycle crash experiment in order to study the efficacy of crash helmets.



Notes

Non-parametric Regression Fits

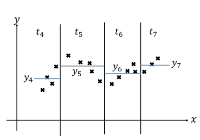
The main idea "non-parametric" regression modeling is to fit the data "locally". Therefore, no global structure assumption made when fitting the data.

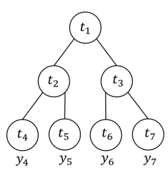


Notes

Regression Tree

- Partitioning X-space into sub-regions and fit simple model to each sub-region
- The partitioning pattern is encoded in a tree structure

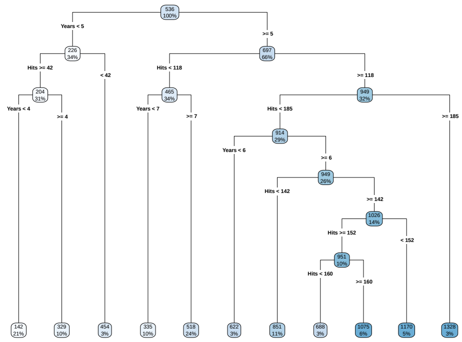




We will use Major League Baseball Hitters Data from the 1986–1987 season to give you a quick idea of what a regression tree might look like

Notes

Regression Tree



Notes

Longley's Economic Regression Data

We are going to use Longley's data set, which provides a well-known example of multicollinearity, to illustrate Ridge regression.

```
     GNP.deflator     GNP Unemployed Armed.Forces Population Year Employed
1947         83.0 234.289      235.6        159.0    107.608 1947   60.323
1948         88.5 259.426      232.5        145.6    108.632 1948   61.122
1949         88.2 258.054      368.2        161.6    109.773 1949   60.171
1950         89.5 284.599      335.1        165.0    110.929 1950   61.187
```


Notes

Linear Regression Fit

```
Call:
lm(formula = response ~ ., data = trainingData)

Residuals:
   1960    1948    1953    1949    1947    1959    1954    1962    1958
-0.2393  0.9650  0.6495 -0.7423 -0.3187 -0.3387  0.1607 -0.1808  1.2922
   1956    1957    1955
 0.3738  0.3889 -2.0104

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.232e+04  2.332e+04  -0.957    0.382
GNP          -1.596e-01  4.535e-01  -0.352    0.739
Unemployed   -8.768e-02  1.138e-01  -0.770    0.476
Armed.Forces -5.346e-02  5.626e-02  -0.950    0.386
Population   -1.331e+00  1.322e+00  -1.007    0.360
Year          1.173e+01  1.210e+01   0.970    0.377
Employed     -3.918e+00  3.498e+00  -1.120    0.314

Residual standard error: 1.284 on 5 degrees of freedom
Multiple R-squared:  0.9939, Adjusted R-squared:  0.9866
F-statistic: 136.2 on 6 and 5 DF,  p-value: 2.251e-05
```


Notes

The Predictor Variables are Highly Correlated

```
             GNP Unemployed Armed.Forces Population Year Employed
GNP          1.00       0.60         0.45       0.99 1.00     0.98
Unemployed   0.60       1.00        -0.18       0.69 0.67     0.50
Armed.Forces 0.45      -0.18         1.00       0.36 0.42     0.46
Population   0.99       0.69         0.36       1.00 0.99     0.96
Year         1.00       0.67         0.42       0.99 1.00     0.97
Employed     0.98       0.50         0.46       0.96 0.97     1.00
```

```
        GNP   Unemployed Armed.Forces   Population         Year
14350.70398    601.69137     98.18754    558.11084  22897.44840
   Employed
 1064.78369
```


Notes

Ridge Regression as Multicollinearity Remedy

- Recall least squares suffers because $(\boldsymbol{X}^T\boldsymbol{X})$ is almost singular thereby resulting in highly unstable parameter estimates
- Modification of least squares that overcomes multicollinearity problem

$$\hat{\beta}_{\text{ridge}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\tilde{\boldsymbol{Y}} - \boldsymbol{Z}\boldsymbol{\beta}\right)^T \left(\tilde{\boldsymbol{Y}} - \boldsymbol{Z}\boldsymbol{\beta}\right) \quad \text{s.t.} \sum_{j=1}^{p-1} \beta_j^2 \leq t,$$

  where $\boldsymbol{Z}$ is assumed to be standardized and $\tilde{\boldsymbol{Y}}$ is assumed to be centered

- Ridge regression results in (slightly) biased but more stable estimates

Notes

Ridge Regression Fit

```
Call:
linearRidge(formula = response ~ ., data = trainingData)


Coefficients:
               Estimate Scaled estimate Std. Error (scaled)
(Intercept)  -1.337e+03              NA                  NA
GNP           2.997e-02       1.016e+01           1.973e+00
Unemployed    1.614e-02       4.465e+00           2.033e+00
Armed.Forces  8.106e-03       1.833e+00           1.835e+00
Population    4.732e-02       1.086e+00           4.174e+00
Year          6.940e-01       1.114e+01           1.356e+00
Employed      8.821e-01       1.056e+01           3.988e+00
             t value (scaled) Pr(>|t|)
(Intercept)                NA       NA
GNP                     5.151 2.60e-07 ***
Unemployed              2.196  0.02807 *
Armed.Forces            0.999  0.31800
Population              0.260  0.79480
Year                    8.215 2.22e-16 ***
Employed                2.648  0.00809 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Ridge parameter: 0.01640472, chosen automatically, computed using 2 PCs

Degrees of freedom: model 3.474 , variance 3.104 , residual 3.844
```


Notes