Erasmus
School of
Economics

Supervised Machine Learning
Week 1

Patrick J.F. Groenen

2020-2021

Table of Contents

Table of Contents

Introduction

What is machine learning?

- Wikipedia (2019):

  Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence.

Introduction

Categorizations of machine learning techniques:

1. Unsupervised versus supervised learning
   - In unsupervised learning all variables have the same status. Interest lies in relations between variables (or objects).
   - In supervised learning there is a set of independent variables $\mathbf{X}$ to predict one or more dependent variables $\mathbf{Y}$. Thus, use $\mathbf{X}$ to explain $\mathbf{Y}$
     - Regression problems have a dependent variable Y that is quantitative (linear)
     - Classification problems have a dependent variable Y that is categorical (nonmetric)
2. Used for exploration or confirmation.
3. Modeling is linear or nonlinear in the data.

Introduction

Incomplete overview of machine learning techniques:

	Method	Un-/supervised	Exploratory vs. Confirmatory	Linear vs. nonlinear	Objective
1	Multiple Regression	Supervised	Both	Linear	Linear prediction.
2	Ridge Regression	Supervised	Exploratory	Linear	Linear prediction with many variables.
3	Lasso	Supervised	Exploratory	Linear	Linear prediction with variable selection.
4	Elastic net	Supervised	Exploratory	Linear	Linear prediction with variable selection.
5	Analysis of Variance (ANOVA)	Supervised	Confirmatory	(Non)linear	Analyzing differences between group means.
6	Neural nets	Supervised	Exploratory	Nonlinear	Prediction by black box (neural net)
7	Regression trees	Supervised	Exploratory	Nonlinear	Fitting tree structure to classifying individuals into groups.
8	Random forest	Supervised	Exploratory	Nonlinear	Flexible nonlinear prediction.
9	Logistic Regression	Supervised	Confirmatory	Linear	Predicting two groups.
10	Support vector machine (SVM)	Supervised	Exploratory	(Non)linear	Predicting two groups.
11	Principal Components Analysis (PCA)	Unsupervised	Exploratory	Linear	Extracting most important components.
12	Multiple correspondence analysis (MCA)	Unsupervised	Exploratory	Nonlinear	Exploration of categorical relations.
13	Nonlinear PCA	Unsupervised	Exploratory	Nonlinear	Extracting components of ordinal variables.
14	Cluster Analysis	Unsupervised	Exploratory	(Non)linear	Create groupings from object similarities.

Introduction

Goals of this course:

1. Thorough technical understanding of selected supervised machine learning techniques.
2. Implement the technique in the high level language R.
3. Being able to apply the technique sensibly to empirical data, and write a short report about it.
4. Use R-markdown for reproducible research.

Table of Contents

Set Up of the Course

How to graduate from this course?

- Active participation (you can miss at most two lectures).
- Group exercises (programming, not graded, but commented).
- Group assignments (grade counts for 15%).
- Group presentation of code discussion (one group per week presents).
- Individual final assignment (counts for 85%).

Set Up of the Course

Set up of the course:

- 2 hour lecture on Tuesday (from Week 2, first hour is presentation).
- 2 hour lecture on Thursday (one hour group presentation/discussion)
- 1 hour tutorial/question hour by TA.
- Group assignments/exercises for Weeks 1–5 (deadline next Tuesday, 9:00 am).
- Individual assignment (deadline two weeks after publication, Friday, 24:00, 18-12-2020).

Set Up of the Course

Online lecture etiquette:

- Mute your microphones (unless you are invited to speak).
- Use the chat to pose a question or a remark.
- One student moderates the chat and (s)he can interrupt me.
- Make only on topic chats and be constructive.
- For a visual feedback to me, it is good to leave your camera open.

Set Up of the Course

Lecturers:

Prof. Patrick J.F. Groenen
Erasmus School of Economics

https://www.eur.nl/en/ese/people/patrick-groenen

Dr. Pieter Schoonees
Rotterdam School of Management

https://www.rsm.nl/people/pieter-schoonees/

Set Up of the Course

Recommended prior knowledge:

- Programming in R
- Use of R Markdown or knitr
- Mathematics
- Statistics
- Optimization
- Linear Algebra

Set Up of the Course

Material and Literature:

- We will make extensive use of
  - R-Studio (https://www.rstudio.com/)
  - R (http://www.r-project.org/)

  Students are strongly encouraged to install this on their laptop and bring it to the class.
- R-code used in the lecture will be made available.

Set Up of the Course

Literature

- Hastie, Tibshirani, and Friedman [2009]. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.
- Pdf of the book is freely available at `http://www.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf`
- Selected papers.

Set up of the course

Schedule:

Week	Day	Lecturer	Topics	Material	Assignment/Exercise[1]
1	Tuesday	Groenen	Introduction; Linear methods for regression	3.1, 3.2, 3.3	
1	Thursday	Groenen	Model selection and assessment	Xiong (2014)	Group Exerc. Week 1
1	Friday	de Jong	Tutorial/question hour		
2	Tuesday	Groenen	Regularized regression and *k*-fold cross validation	3.4.1-3.4.3, 3.8.4, 7.10	Deadline Assign. Week 1
2	Thursday	Groenen	Remainder regularized regression		Pres. Assign. Week 1
2	Friday	de Jong	Tutorial/question hour		
3	Tuesday	Groenen	Basis funct. expansions, kernels, bias-var. trade-off	5.1-5.2.1, 5.8, 7.3	Deadline Group Assign. Week 2
3	Thursday	Groenen	bias-var. trade-off		Pres. Group Assign. Week 2
3	Friday	de Jong	Tutorial/question hour		
4	Tuesday	Groenen	Support vector machines	Groenen et al. (2009)	Deadline Group Exerc. Week 3
4	Thursday	Groenen		12.1-12.3	Pres. Group Exerc. Week 3
4	Friday	de Jong	Tutorial/question hour		
5	Tuesday	Schoonees	Class. and regr. trees, random forests, bootstrap	7.11, 9.2, 15	Deadline Group Assign. Week 4
5	Thursday	Schoonees			Pres. Group Assign. Week 5
5	Friday	de Jong	Tutorial/question hour		
6	Tuesday	Schoonees	Boosting	10	Deadline Group Exerc. Week 5
6	Thursday	Schoonees	Handing out Individual Assignment		Pres. Group Exerc. Week 5
6	Friday	de Jong	Tutorial/question hour		

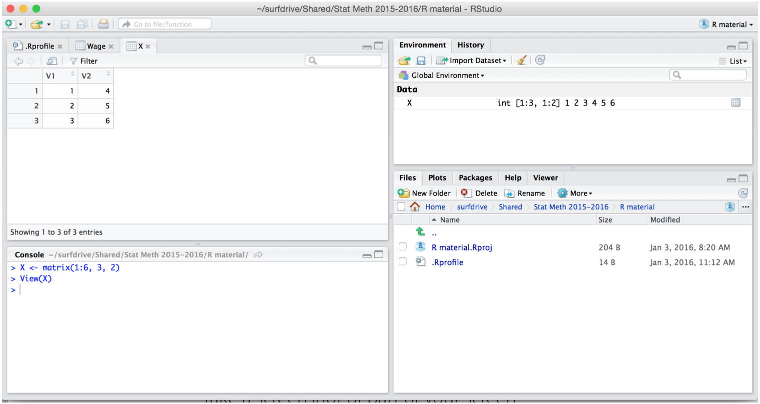
[1] Assignments are graded, exercises not.

Table of Contents

R and R-Studio

- R is a free software environment for statistical computing and graphics: `http://www.r-project.org/`
- Strong points R:
  - State-of-the-art statistical software available.
  - Many contributions by leading scientists in so-called packages.
  - Syntax-based programming allows reproducing your results.
  - R-Studio is a good interface for R.
  - New York Times: Data Analysts Are Mesmerized by the Power of Program R `http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html`

R and R-Studio



R and R-Studio

R-workspace:

- A workspace is the collection of all things currently stored in R's memory:
  - data object
  - function
  - variable with a scalar
  - a variable with a matrix
  - etc.
- Workspace to be saved to a file by: `save.image("MyWorkSpace.RData")`
- Loading a workspace from a file by double-clicking on the file or by: `load("wave5NL.RData")`
- To see what is stored, use the `ls()` command.

R and R-Studio

R-Resources:

- A good resource and introduction to R is Quick-R: `http://www.statmethods.net/`
- For MatLab users who want to switch to R: `http://www.math.umaine.edu/~hiebeler/comp/matlabR.html`

Table of Contents

Linear Algebra in R

- Purpose: Helps you to understand certain properties in MVA techniques in an intuitive way
- Why using matrix algebra?
  - Because the notation is very compact and powerful.
  - Efficient notation compared to element notation.
  - Notation gives more insight.
  - Very useful for linear and quadratic expressions.
  - Understanding of details is less important.

Linear Algebra in R

- A matrix is a emphasizerectangular array of numbers of, say, n rows and m columns.
- Example of a 3×2 matrix $\mathbf{A}$ is $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$
- Properties of a matrix:
  - Matrices are indicated by a boldface uppercase letter: e.g., $\mathbf{A}, \mathbf{B}, \mathbf{X}$, etc.
  - The order of a matrix is the number of rows $\times$ the number of columns: e.g., 3×2.

Linear Algebra in R

Defining a matrix in R:

```
R> ## Matrix A
R> A <- matrix(c(1, 3, 4, 2, 5, 7), nrow = 3, ncol = 2)
R> A

     [,1] [,2]
[1,]    1    2
[2,]    3    5
[3,]    4    7
```

Special Matrices

Type	Description	Order	Example
Square	n is equal to m	$n \times n$	$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 4 \\ 2 & 3 & 3 \end{bmatrix}$
Symmetric	elements ij and ji are equal	$n \times n$	$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 2 \end{bmatrix}$
Diagonal	off-diagonal elements are zero	$n \times n$	$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix}$
Identity $\mathbf{I}$	diagonal matrix with 1 on the diagonal	$n \times n$	$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

Linear Algebra in R

- A vector is a matrix with only one row or column and denoted by a boldface lowercase letter, for example $\mathbf{a}, \mathbf{b}, \mathbf{x}$, etc.
- Examples of column vectors:

  size 3×1, $\mathbf{a} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}$, vector of ones for $n = 3$, $\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$
- Examples of row vectors:

  size 1×2, $\mathbf{b} = \begin{bmatrix} 1 & 3 \end{bmatrix}$, vector of ones for $m = 3$, $\mathbf{1}' = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$

Linear Algebra in R

The transpose of a matrix (or a vector) is denoted by the superscript $\top$ and flips the matrix along the diagonal, examples:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 4 & 7 \end{bmatrix} \quad \Longrightarrow \quad \mathbf{A}^\top = \begin{bmatrix} 1 & 3 & 4 \\ 2 & 5 & 7 \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} 1 & 2 \end{bmatrix} \quad \Longrightarrow \quad \mathbf{b}^\top = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$\mathbf{a} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix} \quad \Longrightarrow \quad \mathbf{a}^\top = \begin{bmatrix} 1 & 3 & 4 \end{bmatrix}$$

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad \Longrightarrow \quad \mathbf{1}^\top = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$$

Linear Algebra in R

Transpose and vectors in R:

```
R> ## Examples of transpose and vectors in R
R> A

     [,1] [,2]
[1,]    1    2
[2,]    3    5
[3,]    4    7

R> # Transpose of A
R> t(A)

     [,1] [,2] [,3]
[1,]    1    3    4
[2,]    2    5    7

R> # Vector a
R> a <- c(3, 0, 2)
R> a

[1] 3 0 2

R> # Create vector of ones
R> ones <- rep(1, length.out = 3)
R> ones
```

Linear Algebra in R

- Matrix addition (or subtraction) is done elementwise

$$
\begin{aligned}
\mathbf{A} + \mathbf{B} &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + a_{22} \end{bmatrix} \\
&= \begin{bmatrix} 3 & 6 \\ 7 & 2 \end{bmatrix} + \begin{bmatrix} 1 & -6 \\ 4 & -3 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 11 & -1 \end{bmatrix}
\end{aligned}
$$

Linear Algebra in R

Matrix addition in R:

```
R> ## Matrix addition:
R> A <- matrix(c(3, 7, 6, 2), nrow = 2, ncol = 2)
R> A

     [,1] [,2]
[1,]    3    6
[2,]    7    2

R> B <- matrix(c(1, 4, -6, -3), nrow = 2, ncol = 2)
R> B

     [,1] [,2]
[1,]    1   -6
[2,]    4   -3

R> A + B

     [,1] [,2]
[1,]    4    0
[2,]   11   -1
```

Linear Algebra in R

- Matrix multiplication is not elementwise.
- Let $\mathbf{AB} = \mathbf{C}$. Then $\mathbf{C}$ has elements $c_{ij} = \sum_k a_{ik} b_{kj}$.

$$\mathbf{a}^\top \mathbf{b} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} = a_{11} b_{11} + a_{12} b_{21} + a_{13} b_{31} = c$$

$$\begin{bmatrix} 3 & 0 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} = 3 \times 2 + 0 \times 0 + 2 \times 1 = 8$$

Linear Algebra in R

- Matrix multiplication is not elementwise.
- Let $\mathbf{AB} = \mathbf{C}$. Then $\mathbf{C}$ has elements $c_{ij} = \sum_k a_{ik} b_{kj}$.

$$\mathbf{a}^\top \mathbf{b} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} = a_{11} b_{11} + a_{12} b_{21} + a_{13} b_{31} = c$$

$$\begin{bmatrix} 3 & 0 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} = 3 \times 2 + 0 \times 0 + 2 \times 1 = 8$$

Matrix multiplication in R:

```
R> ## Inner product (vector product)
R> a <- c(3, 0, 2)
R> b <- c(2, 0, 1)
R> a %*% b
```

[,1] [1,] 8

Linear Algebra in R

- Matrix multiplication is not elementwise.
- Let $\mathbf{AB} = \mathbf{C}$. Then $\mathbf{C}$ has elements $c_{ij} = \sum_k a_{ik} b_{kj}$.

$$\mathbf{AB} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{bmatrix} = \mathbf{C}$$

$$\begin{bmatrix} 3 & 0 & 2 \\ 1 & 2 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 8 & 5 \\ 2 & 3 \\ -1 & -1 \end{bmatrix}$$

- Important: Order in matrix multiplication must be right!
- Multiplication with the identity matrix $\mathbf{I}$ has no effect: $\mathbf{AI} = \mathbf{A}$

Linear Algebra in R

Matrix multiplication in R:

```
R> ## Matrix multiplication (matrix product)
R> A <- matrix(c(3, 1, 0, 0, 2, 0, 2, 0, -1), nrow = 3, ncol = 3)
R> A

     [,1] [,2] [,3]
[1,]    3    0    2
[2,]    1    2    0
[3,]    0    0   -1

R> B <- matrix(c(2, 0, 1, 1, 1, 1), nrow = 3, ncol = 2)
R> B

     [,1] [,2]
[1,]    2    1
[2,]    0    1
[3,]    1    1

R> A %*% B

     [,1] [,2]
[1,]    8    5
[2,]    2    3
[3,]   -1   -1
```

Linear Algebra in R

- Let $\mathbf{A}$ be a square $n \times n$ matrix
- Then $\mathbf{A}^{-1}$ is the matrix inverse of $\mathbf{A}$ such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

Linear Algebra in R

Matrix inverse in R:

```
R> ## Matrix inverse
R> A <- matrix(c(3, 1, 0, 2), nrow = 2, ncol = 2)
R> Ainv <- solve(A, diag(2))                    # diag(2) creates the 2 x 2 identity matrix
R> A

     [,1] [,2]
[1,]    3    0
[2,]    1    2

R> Ainv

       [,1] [,2]
[1,]  0.333  0.0
[2,] -0.167  0.5

R> Ainv %*% A

     [,1] [,2]
[1,]    1    0
[2,]    0    1

R> A %*% Ainv

     [,1] [,2]
```

Linear Algebra in R

Example of matrix inverse that does not exist in R:

```
R> ## The inverse does not always exist
R> A <- matrix(1, nrow = 2, ncol = 2)
R> A

     [,1] [,2]
[1,]    1    1
[2,]    1    1

R> Ainv <- solve(A, diag(2))

Error in solve.default(A, diag(2)):  Lapack routine dgesv:  system is exactly singular:
U[2,2] = 0
```

Table of Contents

Multiple Regression

Material this lecture:

Topic		To read
1.	Introduction	Multiple Regression 3.1, 3.2
2.	Subset selection	3.3, Xiong (2014)

Multiple Regression

Regression

- Regression is the work horse in statistics.
- Careful understanding is needed
- Later on, adaptations, extensions are introduced.

Multiple Regression

Advertising data set

- Advertising budget in thousands of dollars.
- $n = 200$ markets.
- Response variable is sales.
- Three predictor variables: TV, radio, and newspaper.

Multiple Regression

- Example:

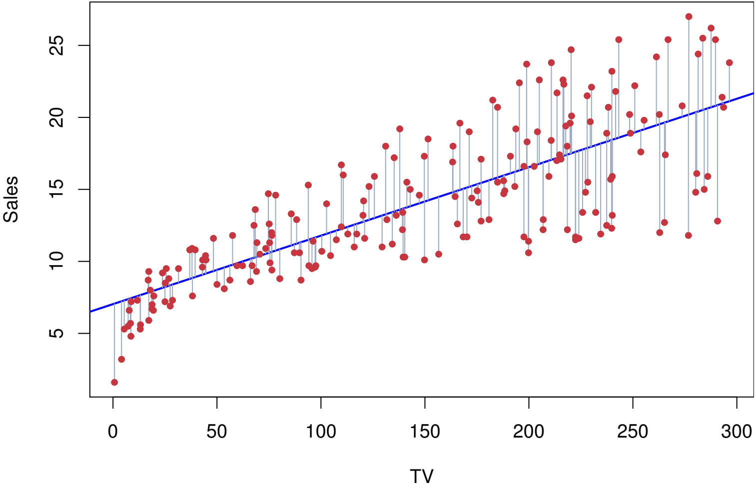
$$\text{sales} \approx \beta_0 + \beta_1 \times \text{TV}$$

- For prediction we need estimates by using training data denoted by the symbol ˆ.
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates of β_0 and β_1 based on the training data.
- A future prediction of sales $\hat{y}$ for a particular value x of TV advertising is:

$$\begin{aligned} \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x \\ \hat{\mathbf{y}} &= \mathbf{X}\hat{\beta} \end{aligned}$$

with $\mathbf{X} = [\mathbf{1}, \mathbf{x}_{\mathrm{TV}}]$

Multiple Regression



Multiple Regression

- Consider the error

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

- Estimation is done by minimizing the sum of squared errors over all observations $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$:
  residual sum of squares (RSS)

$$\begin{aligned} \text{RSS}(\beta) &= \mathbf{e}^\top \mathbf{e} = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) \\ \hat{\beta} &= \text{argmin}_{\beta} \text{RSS}(\beta) \end{aligned}$$

.

Multiple Regression

Least-squares estimates of weights $\hat{\beta}$:

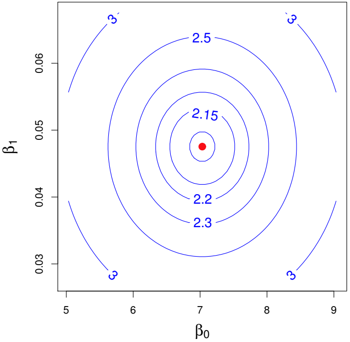
```
R> ## Least-squares estimates of weights for simple regression model
R> load("Advertising.Rdata")         # Load the Advertsing data set
R> result <- lm(Sales ~ TV, Advertising)   # Call the linear regression model (lm
R>                                   # Sales dependent, TV as predictor
R> round(coef(result), digits = 3)  # Give weights

(Intercept)          TV
      7.033       0.048
```

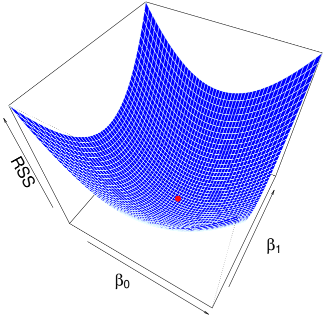
Thus,

$$\begin{aligned} \hat{\beta}_0 &= 7.033 \\ \hat{\beta}_1 &= 0.048 \end{aligned}$$

Multiple Regression

Residual sum of squares (RSS) as a function of β_0 and β_1





The red dot corresponds to the (β_0, β_1) with the lowest RSS: $\hat{\boldsymbol{\beta}} = [7.03, 0.048]^\top$

Multiple Regression

Multiple regression: make a linear combination of predictors to predict Y

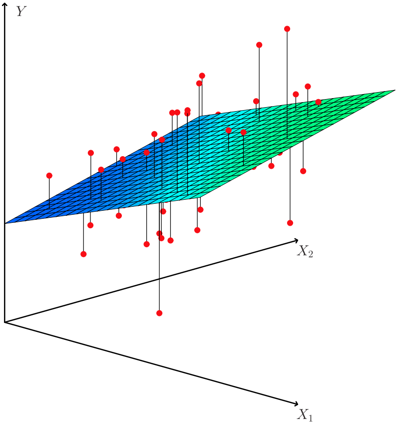
$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \\
\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}
\end{aligned}
$$

with

- Y: random response variable
- X: random vector of p predictor variables
- $\mathbf{y}$ observed n vector of response variable
- $\mathbf{X}$ $n \times (p+1)$ prediction variables with first column of ones for the intercept
- $\boldsymbol{\beta}$ $(p+1)$ vector of weights $[\beta_0, \beta_1, \beta_2, \ldots, \beta_p]^\top$

Multiple Regression

Residuals for multiple regression of Y on X_1 and X_2.



Multiple Regression: Advertising Example

```
R> load("Advertising.Rdata")  # Load the Advertsing data set
R> # Call the linear regression model (lm):
R> result <- lm(Sales ~ TV + Radio + Newspaper, Advertising)
R> summary(result)            # Give a summary of the results object



Call:
lm(formula = Sales ~ TV + Radio + Newspaper, data = Advertising)

Residuals:
   Min     1Q Median     3Q    Max
-8.828 -0.891  0.242  1.189  2.829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.93889    0.31191    9.42   <2e-16 ***
TV           0.04576    0.00139   32.81   <2e-16 ***
Radio        0.18853    0.00861   21.89   <2e-16 ***
Newspaper   -0.00104    0.00587   -0.18     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.69 on 196 degrees of freedom
Multiple R-squared:  0.897,Adjusted R-squared:  0.896
F-statistic:  570 on 3 and 196 DF,  p-value: <2e-16
```

Multiple Regression: Advertising Example

Interpretation

- Prediction goes very well: $R^2 = 90\%$.
- TV and Radio budgets constribute significantly. Newspaper does not contribute.
- Interpretation TV coefficient:
  $\hat{\beta}_1 = 0.0457$, all other predicters remaining the same, then \$ 1,000,000 more TV advertising is expected to increase sales by 45.7 units.
- Interpretation Radio coefficient:
  $\hat{\beta}_2 = 0.1885$, all other predicters remaining the same, then \$1,000,000 more radio advertising is expected to increase sales by 188.5 units.

Multiple Regression

How to minimize $\text{RSS}(\boldsymbol{\beta})$ over $\boldsymbol{\beta}$ with

$$\text{RSS}(\boldsymbol{\beta}) = \mathbf{e}^\top \mathbf{e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})?$$

Multiple Regression

Linear algebra: multiple regression estimation of two or more predictors:

- In linear algebra, $\hat{\beta}$ is given by

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

  with $\mathbf{X}$ is the vector of ones appended with the matrix of predictors.
- R code:

```
X <- as.matrix(Advertising[, 1:3])      # Select first 3 columns of
                                        # Advertising as predictors and
                                        # make it a matrix instead of a
                                        # dataframe
y <- as.vector(Advertising[, 4])        # Make Sales the dependent variable
X <- cbind(rep(1, nrow(X)), X)          # Add a column of ones for the interce
XXinv <- solve(t(X) %*% X, diag(ncol(X))) # Compute (X'X)^-1
beta <- XXinv %*% t(X) %*% y            # Compute OLS beta = (X'X)^-1 X'y
```

Table of Contents

An MM Algorithm for Multiple Regression

- Multiple regression involves solving the linear system $\mathbf{X}^\top \mathbf{X}\beta = \mathbf{X}^\top \mathbf{y}$ or the computation of $(\mathbf{X}^\top \mathbf{X})^{-1}$.
- Advantage: the minimum is analytical.
- What if your computer language does not have a linear system solver?
- Here we derive an MM-based iterative algorithm.
- Will be useful for better subset selection.

An MM Algorithm for Multiple Regression

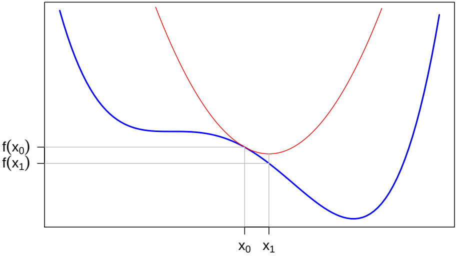
- Majorization is also known under other names:
- MM algorithm [Hunter and Lange, 2004]
  - Minimization by majorization
  - Maximization by minorization
- Machine learning literature, concave convex procedure (CCCP) [Yuille and Rangarajan, 2003]
- Generalized Weiszfeld's method [Voss and Eckhardt, 1980].

An MM Algorithm for Multiple Regression

- Some references of iterative majorization (IM):
  - De Leeuw [1993]; numerical aspects
  - Heiser [1995]; overview article
  - Borg and Groenen [2005]; simple, step by step explanation
  - Lange et al. [2000]; discussion article
  - Kiers [2002]; matrix optimization problems

An MM Algorithm for Multiple Regression

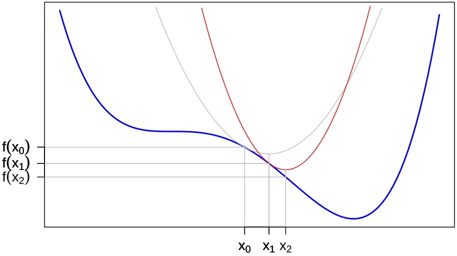
- Replace original function $f(x)$ by a simpler function, the majorizing function $g(x, y)$.



- Requirements majorizing function $g(x, y)$:
  1. $f(y) = g(y, y)$ touch at supporting point y
  2. $f(x) \leq g(x, y)$
  3. $g(x, y)$ must be simple (usually linear or quadratic)

An MM Algorithm for Multiple Regression

- Replace original function $f(x)$ by a simpler function, the majorizing function $g(x, y)$.



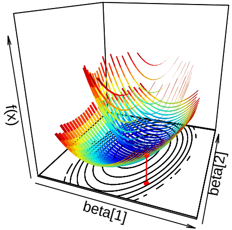
- Requirements majorizing function $g(x, y)$:
  1. $f(y) = g(y, y)$ touch at supporting point y
  2. $f(x) \leq g(x, y)$
  3. $g(x, y)$ must be simple (usually linear or quadratic)

An MM Algorithm for Multiple Regression

- Minimize

$$
\begin{aligned}
\text{RSS}(\boldsymbol{\beta}) &= \mathbf{e}^\top\mathbf{e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{y}^\top\mathbf{y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y}
\end{aligned}
$$

- Difficult part lies in $\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}$.
- Solution: find majorizing function of the form $\lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}$



An MM Algorithm for Multiple Regression

Goal Minimize $RSS(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{y} + \mathbf{y}^\top \mathbf{y}$.

Step 1a Find a majorizing function for $\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}$ of the form $\lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\mathbf{b} + c$.

Step 1b To do so, we need $\mathbf{X}^\top\mathbf{X} - \lambda\mathbf{I}$ to be negative semidefinite (nsd) so that $\mathbf{r}^\top(\mathbf{X}^\top\mathbf{X} - \lambda\mathbf{I})\mathbf{r} \leq 0$ for any vector $\mathbf{r}$.

Step 1c Fact: $\mathbf{X}^\top\mathbf{X} - \lambda\mathbf{I}$ is nsd if $\lambda \geq \lambda_{\max}$ with $\lambda_{\max}$ the largest eigenvalue of $\mathbf{X}^\top\mathbf{X}$.

An MM Algorithm for Multiple Regression

Step 2a Then $(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top(\mathbf{X}^\top\mathbf{X} - \lambda\mathbf{I})(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \leq 0$ because $\mathbf{X}^\top\mathbf{X} - \lambda\mathbf{I}$ is nsd.

Step 2b From the inequality of Step 2a:

$$\begin{aligned} \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} &\leq \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} - 2\lambda\boldsymbol{\beta}^\top(\boldsymbol{\beta}_0 - \lambda^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_0) + \lambda\boldsymbol{\beta}_0^\top(\lambda\mathbf{I} - \mathbf{X}^\top\mathbf{X})\boldsymbol{\beta}_0 \\ &= \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} - 2\lambda\boldsymbol{\beta}^\top(\boldsymbol{\beta}_0 - \lambda^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_0) + c_1 \end{aligned}$$

▸ Details

Step 3 Substitute majorising function for $\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}$ in $RSS(\boldsymbol{\beta})$:

$$\begin{aligned} RSS(\boldsymbol{\beta}) &\leq \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} - 2\lambda\boldsymbol{\beta}^\top(\boldsymbol{\beta}_0 - \lambda^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_0 + \lambda^{-1}\mathbf{X}^\top\mathbf{y}) + c_1 + \mathbf{y}^\top\mathbf{y} \\ &= \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} - 2\lambda\boldsymbol{\beta}^\top\mathbf{u} + c_2 = g(\boldsymbol{\beta}, \boldsymbol{\beta}_0) \end{aligned}$$

Step 4 Set the gradient of $g(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$ to zero:

$$\nabla g(\boldsymbol{\beta}, \boldsymbol{\beta}_0) = 2\lambda(\boldsymbol{\beta} - \mathbf{u}) = \mathbf{0}$$

Step 5 Update $\boldsymbol{\beta}^+$ without the need of an inverse of $\mathbf{X}^\top\mathbf{X}$ is:

$$\boldsymbol{\beta}^+ = \mathbf{u} = \boldsymbol{\beta}_0 - \lambda^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_0 + \lambda^{-1}\mathbf{X}^\top\mathbf{y}$$

An MM Algorithm for Multiple Regression

- An MM algorithm for the elastic net:

> Choose with some initial $\beta_0 \in \mathbb{R}^p$ and small ϵ
> Compute $RSS(\beta_0)$
> Compute λ as the largest eigenvalue of $\mathbf{X}^\top\mathbf{X}$
> Set $k \leftarrow 1$
> **while** $k = 1$ or $(RSS(\beta_{k-1}) - RSS(\beta_k))\,/RSS(\beta_{k-1}) > \epsilon$ **do**
> $k \leftarrow k + 1$
>   The update $\beta^{(k)} = \beta_{k-1} - \lambda^{-1}\mathbf{X}^\top\mathbf{X}\beta_{k-1} + \lambda^{-1}\mathbf{X}^\top\mathbf{y}$
>   As a check, print $k, RSS(\beta_k)$, and $RSS(\beta_{k-1}) - RSS(\beta_k)$
> **end**

- Useful test for programming: in each iteration $RSS(\beta_k)$ must go down.

An MM Algorithm for Multiple Regression

Group exercise for next Thursday (not graded):

- I have formed groups of three (some four) on Canvas.
- Read Sections 3.1, 3.2, 3.3.
- The data set `Airq.RData` comes from the `Ecdat` package in R and describes airquality. Look at the help file on `Airq` in the `Ecdat` package for more details on the variables.
- The response variable is `airq` and the remaining variables are predictors.
- Program a multiple regression including R^2, the $\boldsymbol{\beta}$, their standard deviation, and their z-values.
- Make a numeric comparison of your own results and those oobtained from to the standard output of R through `lm()`. How can you can you make a numeric comparison?
- Write a general R-function that implements the MM-algorithm for multiple regression. At the very least, it should accept as input $\mathbf{y}$ and $\mathbf{X}$ and return the vector of weights $\hat{\boldsymbol{\beta}}$.
- Do the above before the coming Thursday lecture.

Multiple Regression

How well does the model fit the data?

- We assume that $\mathbf{y}$ and $\mathbf{X}$ have column mean 0 (thus no intercept).
- The multiple R^2 is the squared correlation of $\mathbf{y}$ and $\hat{\mathbf{y}}$;

$$R^2 = \left(\frac{(\mathbf{y}^\top \hat{\mathbf{y}})}{(\mathbf{y}^\top \mathbf{y})^{1/2} (\hat{\mathbf{y}}^\top \hat{\mathbf{y}})^{1/2}} \right)^2 = \frac{(\mathbf{y}^\top \hat{\mathbf{y}})^2}{(\mathbf{y}^\top \mathbf{y})(\hat{\mathbf{y}}^\top \hat{\mathbf{y}})}$$

- R^2 always increases when a predictor variable is added.
- R^2 is closely related to the proportion of Variance Accounted For (VAF):

$$R^2 = 1 - \frac{\hat{\mathbf{y}}^\top \hat{\mathbf{y}}}{\mathbf{y}^\top \mathbf{y}} = 1 - \text{VAF}$$

- As an exercise, prove the relation between R^2 and VAF.
  (Hint: use $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ and $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$).

Table of Contents

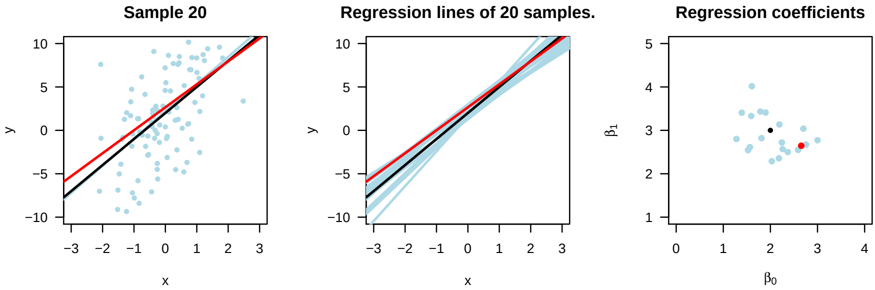
Inference for Multiple Regression

Important questions:

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful to predict the response?
2. Is a subset of predictors enough to explain Y or do we need all?
3. How accurate can we determine the regression weights?

Inference for Multiple Regression

Generated samples with $n = 100$, $\beta_0 = 2$ and $\beta_1 = 3$



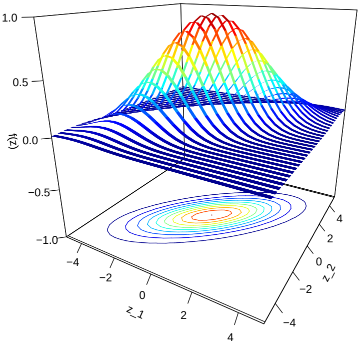
Inference for Multiple Regression

Multivariate normal distribution:

$$\mathbf{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with p dimensional normal density function:

$$f(\mathbf{z}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{z}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})/2}$$



Inference for Multiple Regression

- The multiple regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

  has the following assumptions on the errors in $\boldsymbol{\epsilon}$:
  - the ϵ_i are independent (and thus have correlation 0);
  - all ϵ_i are identically distributed;
  - ϵ_i is normally distributed with mean 0 and variance σ^2: $\epsilon_i \sim N(0, \sigma^2)$.
- Consequently, $\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\epsilon}$ is multivariate normally distributed with $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$:

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

Inference for Multiple Regression

The covariance matrix of $\hat{\beta}$ equals

$$\text{var}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

with

$$\hat{\sigma}^2 = \frac{(\mathbf{y}-\mathbf{X}\hat{\beta})^\top(\mathbf{y}-\mathbf{X}\hat{\beta})}{n-p-1}$$

```
          Intercept  TV         Radio      Newspaper
Intercept  0.0972867 -0.0002657 -0.0011155 -0.0005910
TV        -0.0002657  0.0000019 -0.0000004 -0.0000003
Radio     -0.0011155 -0.0000004  0.0000742 -0.0000178
Newspaper -0.0005910 -0.0000003 -0.0000178  0.0000345
```

Inference for Multiple Regression

The standard deviation of the elements of $\hat{\boldsymbol{\beta}}$ is

$$\mathrm{Diag}(\hat{\sigma}^2(\mathbf{X}^\top \mathbf{X})^{-1})^{1/2}$$

```
            beta    sdev    z
Intercept  2.9389  0.3119  9.4223
TV         0.0458  0.0014 32.8086
Radio      0.1885  0.0086 21.8935
Newspaper -0.0010  0.0059  0.1767
```

$z_j = \hat{\beta}_j/\mathrm{sdev}_j$ and is t-distributed with $n - p - 1$ degrees of freedom.

Inference for Multiple Regression

```
R> ## Test for simultaneous contribution of Radio and Newspaper
R> summary(result)


Call:
lm(formula = Sales ~ TV + Radio + Newspaper, data = Advertising)

Residuals:
   Min     1Q Median     3Q    Max
-8.828 -0.891  0.242  1.189  2.829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.93889    0.31191    9.42   <2e-16 ***
TV           0.04576    0.00139   32.81   <2e-16 ***
Radio        0.18853    0.00861   21.89   <2e-16 ***
Newspaper   -0.00104    0.00587   -0.18     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.69 on 196 degrees of freedom
Multiple R-squared:  0.897,Adjusted R-squared:  0.896
F-statistic:  570 on 3 and 196 DF,  p-value: <2e-16
```

Table of Contents

Subset Selection

- Which predictors X_j are important? (variable selection)
- Incomplete list of subset selection methods:
  1. Exhaustive search: try out all possible combinations of predictor variables.
  2. Forward selection
  3. Backward elimination
  4. Mixed selection: combine forward selection and backward elimination.
  5. Best subset: search for the best subset of K variables.
  6. Lasso: shrinkage method with automatic variable selection (discussed next week).

Subset Selection

Ad 1. Exhaustive search

- Try out all possible combinations of predictor variables.
  With two predictors X_1 and X_2, the three models are

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 \\
Y &= \beta_0 + \beta_2 X_2 \\
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2
\end{aligned}
$$

- Evaluate a fit statistic (Mallows C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R^2).
- Disadvantage: 2^p- models need to be evaluated (cannot be done for p large: $p > 30$ leads to $2^{30} = 1,073,741,824$ models).

Subset Selection

Ad 2. Forward selection:

- Start with intercept only.
- Try p simple regression and select the model with the lowest RSS.
- Expand selected model by trying out combinations with $p-1$ remaining predictors and select the one with the lowest RSS.
- Keep on adding a predictor until some stopping rule is satisfied.

Subset Selection

Ad 3. Backward elimination:

- Start with all p predictors in the model.
- Eliminate the predictor with the largest p-value (contributes the least).
- Keep on removing a predictor until some stopping rule is satisfied.

Ad 4. Mixed selection:

- Start with foward selection until p-value rises above a threshold value.
- Then, remove this variable.
- Repeat these steps until all variables have a p-value below a threshold value.

Subset Selection

Example stepwise selection predictors:

```
R> ## Stepwise regression (automatic selection of predictors)
R> result <- lm(Sales ~ TV + Radio + Newspaper, data = Advertising)
R> step <- stepAIC(result, direction="both")


Start:  AIC=213
Sales ~ TV + Radio + Newspaper

            Df Sum of Sq  RSS AIC
- Newspaper  1         0  557 211
<none>                    557 213
- Radio      1      1362 1919 458
- TV         1      3058 3615 585

Step:  AIC=211
Sales ~ TV + Radio

            Df Sum of Sq  RSS AIC
<none>                    557 211
+ Newspaper  1         0  557 213
- Radio      1      1546 2103 475
- TV         1      3062 3618 583
```

Subset Selection

Ad 5. Best subset selection:

- Goal: try to find best subset of m nonzero weights in $\boldsymbol{\beta}$.
- This is an NP-hard combinatorial problem.
- For fixed m, we use an approximate algorithm based on MM proposed by Xiong [2014] called better subsets regression.
- One can try various values of m.

Subset Selection

Ad 5. Main ideas better subset algorithm:

- The same majorizing function is used as in the MM algorithm for multiple regression.
- Then we have:

$$
\begin{aligned}
RSS(\boldsymbol{\beta}) &\leq \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} - 2\lambda \boldsymbol{\beta}^\top \mathbf{u} + c_2 \\
&= \lambda(\boldsymbol{\beta} - \mathbf{u})^\top (\boldsymbol{\beta} - \mathbf{u}) - \lambda \mathbf{u}^\top \mathbf{u} + c_2 \\
&= \lambda \sum_{j=1}^{p} (\beta_j - u_j)^2 + c_3 = g(\boldsymbol{\beta}, \beta_0)
\end{aligned}
$$

- Only m values of β_j can be chosen different from 0.
- Then, sort $|u_j|$ from large to small and set $\beta_j^+ = u_j$ for the first m elements.
- This choice ensures that $g(\boldsymbol{\beta}, \beta_0)$ is minimal.

Subset Selection

Ad 5. Main ideas better subset algorithm:

- Better subset regression algorithm:

> Choose some initial $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ with at most m values nonzero
> Choose a small ϵ
> Compute $RSS(\boldsymbol{\beta}_0)$
> Compute λ as the largest eigenvalue of $\mathbf{X}^\top\mathbf{X}$
> Set $k \leftarrow 1$
> **while** $k = 1$ or $(RSS(\boldsymbol{\beta}_{k-1}) - RSS(\boldsymbol{\beta}_k))\,/RSS(\boldsymbol{\beta}_{k-1}) > \epsilon$ **do**
> $k \leftarrow k+1$
>   Compute $\mathbf{u} = \boldsymbol{\beta}_{k-1} - \lambda^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_{k-1} + \lambda^{-1}\mathbf{X}^\top\mathbf{y}$
>   Sort $|u_j|$ and return index vector $\boldsymbol{\psi}$ such that
> $|u_{\psi_1}| \leq |u_{\psi_2}| \leq \ldots \leq |u_{\psi_m}|$
>   Update $\beta^{(k)}_{\psi_\ell} = u_{\psi_\ell}$ for $\ell = 1,\ldots,m$ and $\beta^{(k)}_{\psi_\ell} = 0$ otherwise
>   As a check, print $k, RSS(\boldsymbol{\beta}_k)$, and $RSS(\boldsymbol{\beta}_{k-1}) - RSS(\boldsymbol{\beta}_k)$
> **end**

- Because of MM: in each iteration $RSS(\boldsymbol{\beta}_k)$ must go down.

Table of Contents

Summary and Assignment

Summary:

Week	Topics	Material
1	Introduction; Introduction to R; Linear methods for regression, model selection, and assessment	3.1, 3.2, 3.3, Xiong (2014)
2	Regularized regression and k-fold cross validation	3.4.1-3.4.3, 3.8.4, 7.10
3	Basis function expansions, kernels, bias-variance trade-off	5.1-5.2.1, 5.8, 7.3
4	Support vector machines	Groenen, Nalbantov, Bioch (2009); 12.1-12.3
5	Classification and regression trees, random forests, bootstrap	7.11, 9.2, 15
6	Boosting	10

Assignment Week 1

To do before first lecture next week:

- Read Sections 3.1, 3.2, 3.3 and use the slides (or Xiong 2014) for better subset regression.
- The data set `Airq.RData` comes from the `Ecdat` package in R and describes airquality. Look at the help file on `Airq` in the `Ecdat` package for more details on the variables.
- The response variable is `airq` and the remaining variables are predictors.
- Implement the better subset regression MM algorithm and use it.
- Write a 4 page report. For the report template for requirements. Write the report in R-markdown (as the template does, deadline Tuesday, 09:00).

References I

I. Borg and P. J. F. Groenen. Modern multidimensional scaling. Springer, New York, 2. edition, 2005.

J. De Leeuw. Fitting distances by least squares. Technical Report 130, Interdivisional Program in Statistics, UCLA, Los Angeles, CA, 1993.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media, 2009.

Willem J. Heiser. Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis. In W. J. Krzanowski, editor, Recent advances in descriptive multivariate analysis, pages 157–189, Oxford, 1995. Oxford University Press.

D. R. Hunter and K. Lange. A tutorial on MM algorithms. The American Statistician, 39: 30–37, 2004.

Henk A. L. Kiers. Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems. Computational Statistics and Data Analysis, 41:157–170, 2002.

K. Lange, D. R. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. Journal of Computational and Graphical Statistics, 9:1–20, 2000.

H. Voss and U. Eckhardt. Linear convergence of generalized Weiszfeld's method. Computing, 25(3):243–251, 1980.

Shifeng Xiong. Better subset regression. Biometrika, 101(1):71–84, 2014.

Alan L Yuille and Anand Rangarajan. The concave-convex procedure. Neural computation, 15 (4):915–936, 2003.

Acknowledgement

Some of the figures in this presentation are taken from “An Introduction to Statistical Learning, with applications in R” (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Derivation Majorization Inequality

$$
\begin{aligned}
(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top (\mathbf{X}^\top \mathbf{X} - \lambda \mathbf{I})(\boldsymbol{\beta} - \boldsymbol{\beta}_0) &\leq 0 \text{ because } \lambda \mathbf{I} - \mathbf{X}^\top \mathbf{X} \text{ is nsd} \\
(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \mathbf{X}^\top \mathbf{X} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) &\leq \lambda (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\beta}_0{}^\top \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta}_0 - 2 \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta}_0 &\leq \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} + \lambda \boldsymbol{\beta}_0{}^\top \boldsymbol{\beta}_0 - 2\lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}_0 \\
\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} &\leq \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} - 2\lambda \boldsymbol{\beta}^\top (\boldsymbol{\beta}_0 - \lambda^{-1} \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta}_0) \\
&\quad + \boldsymbol{\beta}_0^\top (\lambda \mathbf{I} - \mathbf{X}^\top \mathbf{X}) \boldsymbol{\beta}_0
\end{aligned}
$$

Back