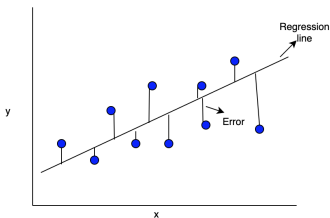


Algorithms and Data Structures

Matrix Decomposition

Overdetermined Systems & Regression

Example



Learning goals

- Overdetermined systems
- Normal equations
- QR decomposition and ridge regression

OVERDETERMINED SYSTEMS

A system of linear equations $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A} \in \mathbb{R}^{m \times n}$, $m \geq n$ with more equations than unknowns, is called **overdetermined**.

In general such a system has no (exact) solution.

A (compromise) solution using **least squares** is the vector \mathbf{x} which minimizes the squared sum of the **residual vector** $\mathbf{r} = \mathbf{b} - \mathbf{Ax}$:

$$\mathbf{x} = \arg \min \|\mathbf{b} - \mathbf{Ax}\|_2^2$$



EXAMPLE: THE REGRESSION MODEL

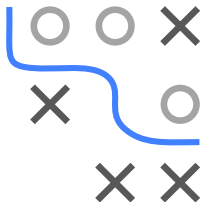
Aim: Solve $\mathbf{X}\beta = \mathbf{y}$ with

- \mathbf{X} : $n \times (p + 1)$, Design matrix
- \mathbf{y} : $n \times 1$, n observations
- β : $(p + 1) \times 1$, p regressors plus intercept

Since the linear system is usually overdetermined (more observations than variables) and has no solution, we minimize the residual sum of squares:

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)$$

Questions: How can the problem be solved in a numerically stable way? Which algorithms are fast?



CONDITION OF NORMAL EQUATIONS

The solution of the optimization problem is (mathematically) equivalent to the solution of the **normal equation**

$$\mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^\top \mathbf{y}$$

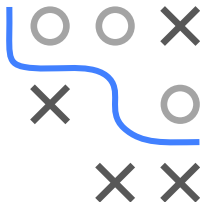
(Derivation: differentiate with respect to $\boldsymbol{\beta}$ and set to 0).

If the matrix \mathbf{X} has full column rank, then the matrix $\mathbf{X}^\top \mathbf{X}$ is symmetric positive-definite and the following holds

$$\kappa(\mathbf{X}^\top \mathbf{X}) = \kappa(\mathbf{X})^2$$

using the spectral norm.

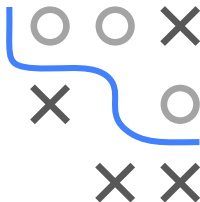
Consequently, the error amplification is $\kappa(\mathbf{X})^2$ when using normal equations.



CONDITION OF NORMAL EQUATIONS / 2

Note:

- Mathematically speaking, the solution of the normal equations is equivalent to the minimization of the residual sum of squares
- However, from a numerical point of view a distinction must be made between the two of them
- A solution using the normal equations requires the calculation of $\mathbf{X}^T \mathbf{X}$, an error in \mathbf{X} is therefore amplified
- Better: Find an efficient method that operates directly on \mathbf{X}



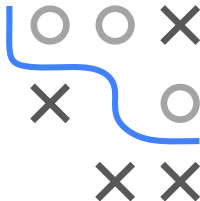
SOLUTION OF THE NORMAL EQUATIONS

Model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Normal equations: $\mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^\top \mathbf{y}$

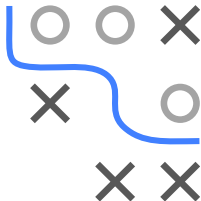
If \mathbf{X} is of full rank then $\mathbf{X}^\top \mathbf{X}$ is positive-definite and the Cholesky decomposition applicable.

- 1 Calculate $\mathbf{X}^\top \mathbf{X}$ and $\mathbf{X}^\top \mathbf{y}$,
- 2 Cholesky decomposition $\mathbf{X}^\top \mathbf{X} = \mathbf{L}\mathbf{L}^\top$,
- 3 Solve $\mathbf{L}\mathbf{w} = \mathbf{X}^\top \mathbf{y}$ for \mathbf{w} ,
- 4 Calculate $RSS = \mathbf{y}^\top \mathbf{y} - \mathbf{w}^\top \mathbf{w}$,
- 5 Solve $\mathbf{L}^\top \boldsymbol{\beta} = \mathbf{w}$ for $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$,
- 6 $(\mathbf{X}^\top \mathbf{X})^{-1} = \mathbf{L}^{-\top} \mathbf{L}^{-1}$.



SOLUTION OF THE NORMAL EQUATIONS / 2

```
X = matrix(c(rep(1, 6), c(1.01, 1.01)), ncol = 2)
X
## [,1] [,2]
## [1,] 1 1.00
## [2,] 1 1.00
## [3,] 1 1.01
## [4,] 1 1.01
XX = t(X) %*% X
XX
## [,1] [,2]
## [1,] 4.00 4.0200000000000000
## [2,] 4.02 4.0402000000000000
```



SOLUTION OF THE NORMAL EQUATIONS / 3

```
XX2 = round(XX, 3)
```

```
XX2
```

```
## [,1] [,2]
```

```
## [1,] 4.00 4.02
```

```
## [2,] 4.02 4.04
```

```
cholesky(XX)
```

```
## [,1] [,2]
```

```
## [1,] 2.00 0.000000000000000000
```

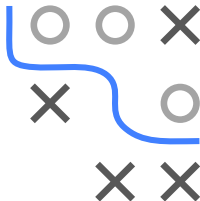
```
## [2,] 2.01 0.0100000000000007716
```

```
cholesky(XX2)
```

```
## [,1] [,2]
```

```
## [1,] 2.00 0
```

```
## [2,] 2.01 NaN
```



⇒ Number of decimal digits matters, matrix no longer positive-definite!

SOLUTION OF THE NORMAL EQUATIONS / 4

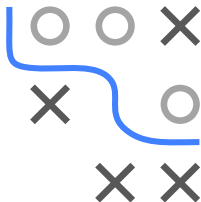
In general a solution using normal equations is to be avoided

$$\mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^\top \mathbf{y}$$

since:

- **High computational effort:** First calculation of $\mathbf{X}^\top \mathbf{X}$, then matrix decomposition of $\mathbf{X}^\top \mathbf{X}$, then forward and back substitution
- **Numeric instability:** In all these individual steps there is a risk that errors will be amplified.

A further problem occurs if we want to solve the normal equations in case of **collinearity** in the design matrix \mathbf{X} . The reason for this is the singularity of the product of $\mathbf{X}^\top \mathbf{X}$ which results from collinearity.



SOLUTION OF THE NORMAL EQUATIONS / 5

It is often more suitable to operate directly on \mathbf{X} by using QR decomposition $\mathbf{X} = \mathbf{QR}$:

$$\mathbf{X}^\top \mathbf{X} = (\mathbf{QR})^\top (\mathbf{QR}) = \mathbf{R}^\top \mathbf{Q}^\top \mathbf{QR} = \mathbf{R}^\top \mathbf{R}$$

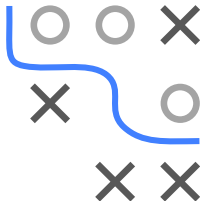
The normal equations can then be written as

$$\mathbf{R}^\top \mathbf{R} \boldsymbol{\beta} = \mathbf{R}^\top \mathbf{Q}^\top \mathbf{y}$$

and since \mathbf{R}^\top is nonsingular it follows

$$\mathbf{R} \boldsymbol{\beta} = \mathbf{Q}^\top \mathbf{y}$$

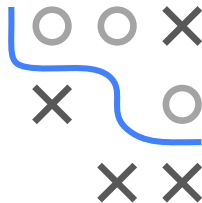
Since \mathbf{R} is an upper triangular matrix, the equation system (after multiplying $\mathbf{Q}^\top \mathbf{y}$) can be solved using back substitution in $\mathcal{O}(n^2)$.



SOLUTION OF THE NORMAL EQUATIONS / 6

The steps to solve a linear regression problem using QR decomposition are therefore as follows:

- 1 Calculate the QR decomposition $\mathbf{X} = \mathbf{QR}$,
- 2 Calculate $\mathbf{z} = \mathbf{Q}^\top \mathbf{y}$,
- 3 Solve the equation system $\mathbf{R}\boldsymbol{\beta} = \mathbf{z}$ using back substitution.



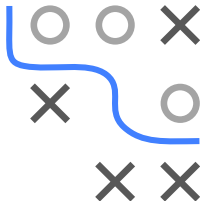
SOLUTION OF THE NORMAL EQUATIONS / 7

Advantage of QR decomposition: As already stated, the solution when using QR decomposition can be written as

$$\mathbf{R}\beta = \mathbf{Q}^\top \mathbf{y}$$

- \mathbf{X} (and not $\mathbf{X}^\top \mathbf{X}$) is decomposed directly.
- $\mathbf{X}^\top \mathbf{X}$ does not have to be calculated (thus avoiding numerical instability from collinearity).
- If a stable algorithm is used to calculate the QR decomposition (e.g. Householder), the method is stable.
- Extreme runtime advantages if a regression is to be performed for constant design matrix \mathbf{X} , but different \mathbf{y} .

Note: For linear models `lm()` in R the QR decomposition is applied when calculating β .



QR DECOMPOSITION AND RIDGE REGRESSION

In order to avoid a high variance, large parameters are often penalized by a penalty term. We minimize a penalized version of the residual sum of squares

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2$$

with regularization parameter $\lambda > 0$. If the L_2 norm is selected for the penalty, the procedure is known as **ridge regression**.

We obtain a general version of the normal equations by setting the first derivative to 0

$$(\mathbf{X}^\top \mathbf{X} + \lambda I) \beta = \mathbf{X}^\top \mathbf{y}$$



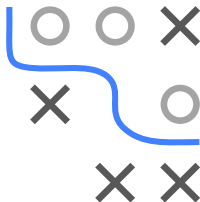
QR DECOMPOSITION AND RIDGE REGRESSION / 2

With $\mathbf{A} \text{oneqq} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda} \mathbf{I} \end{pmatrix} \in \mathbb{R}^{n+p,p}$ the normal equations for ridge regression can be rewritten as

$$\begin{aligned} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}) \boldsymbol{\beta} &= \mathbf{X}^\top \mathbf{y} \\ \mathbf{A}^\top \mathbf{A} \boldsymbol{\beta} &= \mathbf{A}^\top \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} \end{aligned}$$

We use the QR decomposition of $\mathbf{A} = \mathbf{Q}_\lambda \mathbf{R}_\lambda$ depending on λ :

$$\begin{aligned} \mathbf{A}^\top \mathbf{A} \boldsymbol{\beta} &= \mathbf{A}^\top \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} \\ \mathbf{R}_\lambda^\top \mathbf{Q}_\lambda^\top \mathbf{Q}_\lambda \mathbf{R}_\lambda \boldsymbol{\beta} &= \mathbf{R}_\lambda^\top \mathbf{Q}_\lambda^\top \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} \\ \mathbf{R}_\lambda^\top \mathbf{R}_\lambda \boldsymbol{\beta} &= \mathbf{R}_\lambda^\top \mathbf{Q}_\lambda^\top \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} \end{aligned}$$



QR DECOMPOSITION AND RIDGE REGRESSION / 3

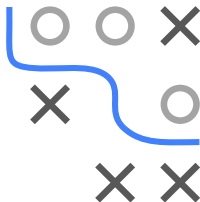
- If \mathbf{R}_λ is singular, we have to solve two LES in echelon form
- If \mathbf{R}_λ is nonsingular and thus invertible, the equation simplifies to one linear system in echelon form

The regularization parameter λ is a hyperparameter that must be selected by the user. Often the linear system has to be solved several times for different λ to find a sensible degree of regularization.

In such situations the QR decomposition

$$\mathbf{R}_\lambda^\top \mathbf{R}_\lambda \boldsymbol{\beta} = \mathbf{R}_\lambda^\top \mathbf{Q}_\lambda^\top \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}$$

has to be calculated anew for each λ . The QR decomposition of the matrix \mathbf{A} is calculated in $\mathcal{O}(n^3)$. Forward and back substitution are operations of $\mathcal{O}(n^2)$, so in total the runtime is given by $\mathcal{O}(n^3)$.



COMPARISON OF METHODS FOR REGRESSION

Method	Runtime	General Numerical Stability	Stability in Collinearity
Naive approach	- -	no	no
LU	+ +	yes, with pivotisation	no
QR (Householder)	-	yes	yes



⇒ **Note:** QR decomposition is not the fastest method regarding runtime, but it is always numerically stable in a regression context.

COMPARISON OF METHODS FOR REGRESSION / 2

