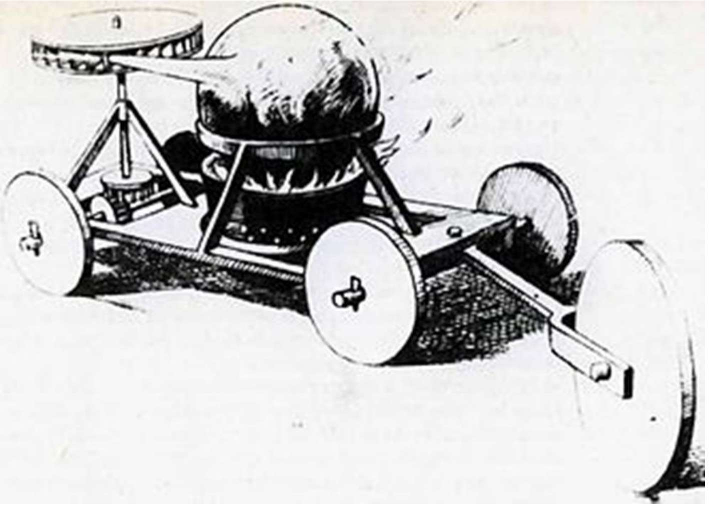
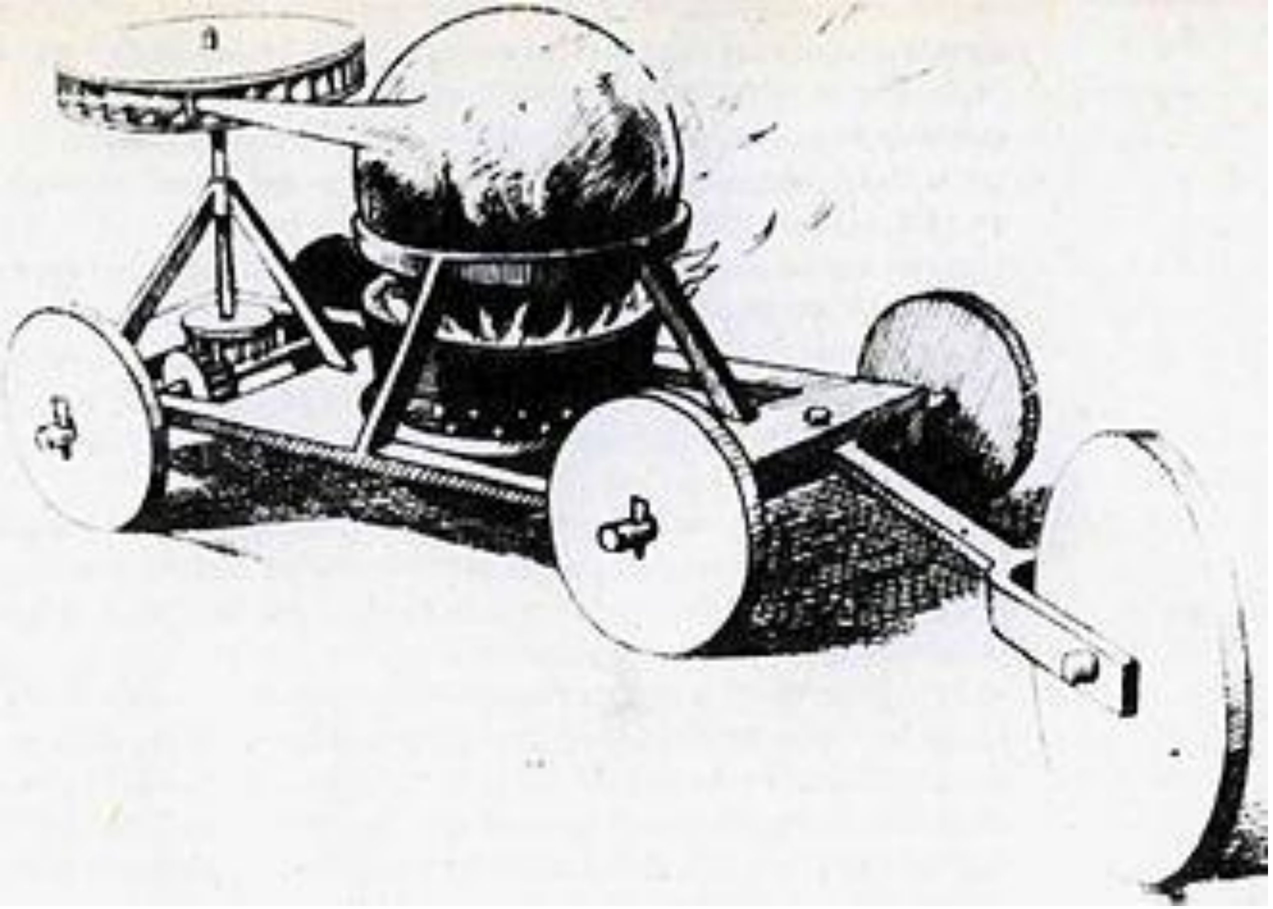
Regression models for prediction (Part I Mixed model and regularization methods)

Vincent Garin



Introduction - Objective



Introduction - Objective

Derivation with respect to u

$$\begin{aligned}\frac{\mathcal{L}}{\partial u} = 0 \Longrightarrow & y'R^{-1}Z + (X\beta)'R^{-1}Z - Z'R^{-1}y + (Z)'R^{-1}X\beta + (Z)'R^{-1}XZu + (Zu)'R^{-1}Z + 2G^{-1}u = 0 \\ & -2y'R^{-1}Z + 2Z'R^{-1}X\beta + 2Z'R^{-1}Zu + 2G^{-1}u = 0 \\ & Z'R^{-1}X\beta + Z'R^{-1}Zu + G^{-1}u = Z'R^{-1}y\end{aligned}$$

The two solution can be reorganized in a matrix system of equations

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

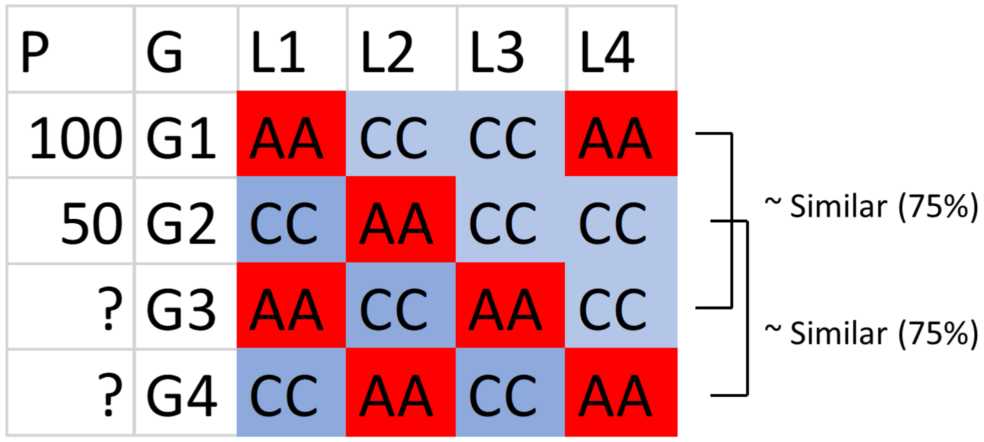
Introduction - Objective

Al-Khwârizmî (789)

« il faut que $4x^2 - 5x + 5x + 7 = 15 + 5x$, donc $4x^2 + 7 = 15 + 5x$ »

Introduction - Objective

K: Kinship matrix, square matrix positive definite



P	G	L1	L2	L3	L4
100	G1	AA	CC	CC	AA
50	G2	CC	AA	CC	CC
?	G3	AA	CC	AA	CC
?	G4	CC	AA	CC	AA

~ Similar (75%)

~ Similar (75%)

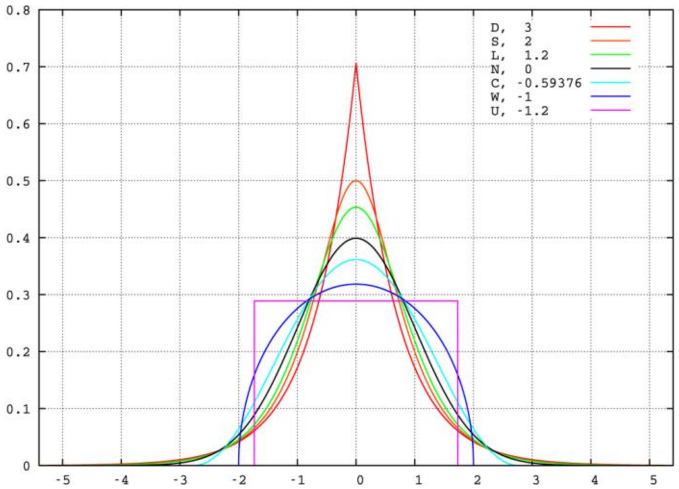
Introduction - Objective

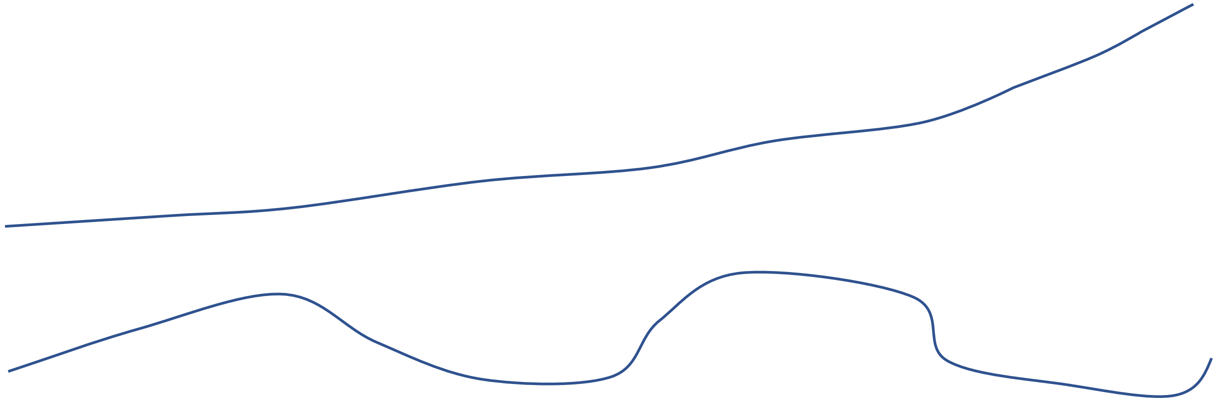
« Ceux qui savent se taisent. Ceux qui ne savent pas prédisent » (Lao-Tseu)

« La prédiction est un art difficile …
Surtout lorsqu'il s'agit de l'avenir» (Pierre Dac)

« La prédiction c'est du pognon» (Picsou)

Introduction – Philosophie : étaler l’apprentissage





Introduction

- Basics: Element of probability
- Vocabulary: Matrix algebra
- Recapitulation: linear model
- Generalized linear model
- Linear mixed model: BLUE, BLUP
- Mixed model equation
- Animal (pedigree) BLUP model
- G BLUP model
- Regularization methods: Ridge regression, LASSO

Elements of probability

Probability space

« *Mathematics is the logic of certainty. Probability is the logic of uncertainty* »

(an Harvard professor)

Probability space (Ω, F, P)

Ω: Univers (set of all possible values) Ex $\Omega = \{1, 2, 3, 4, 5, 6\}$

F: Event space: set of events, a set of outcome in the sample space Ex $\mathrm{F} = \{2, 4, 6\}$ (die lands on an even number) or $\mathrm{F} = \{5\}$ (die land on 5)

P : Probability function. A function that assign a probability (number betwe in the event space.

P($\mathrm{F} = \{5\}$) = 1/6

P($\mathrm{F} = \{2, 4, 6\}$) = P(2 U 4 U 6) = 1/2

Expectation of a random variable

X is a random variable defined on a proability space (Ω, F, P)

If X takes values $x_1, x_2, \dots, x_n$ with probability $p_1, p_2, \dots, p_n$

Then,

$E(X) = x_1 * p_1 + x_2 * p_2 + \cdots + x_n * p_n = \frac{x_1 * p_1 + x_2 * p_2 + \cdots + x_n * p_n}{p_1 + p_2 + \ldots + p_n}$ (Denominator = 1)

$E(X) = \sum_{i=1}^{n} x_i * p\_i$

Correspond to the average in a large (infinite sample)

$E(x) = \frac{1}{n} \sum_{i=1}^{n} x_i$

Expectation of a random variable - Exercise

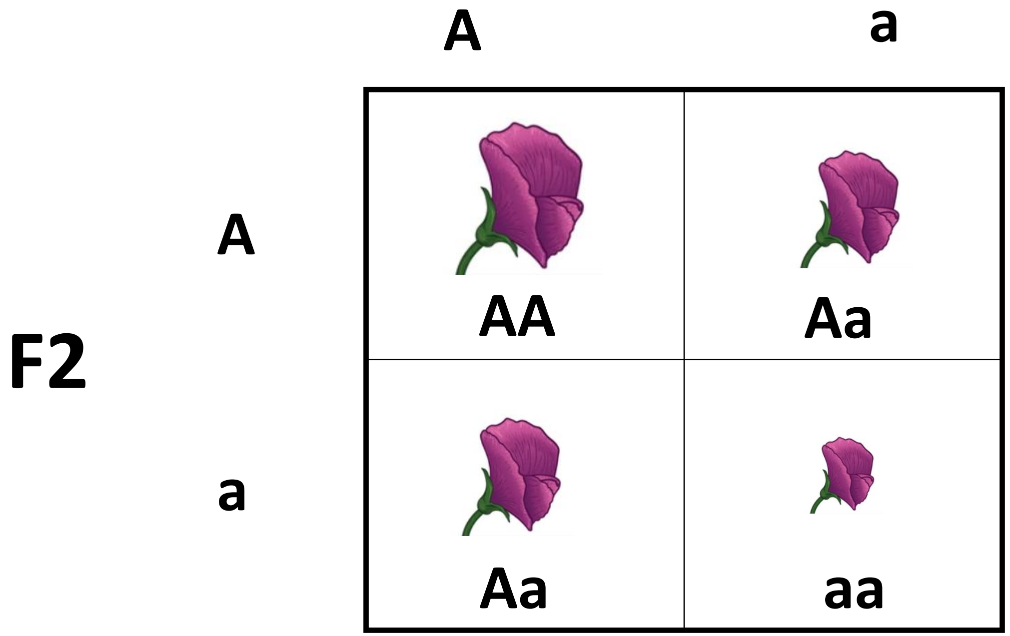
Let us assume that

P(A) = p
P(a) = q = 1 – p

In a F2, we have the following frequencies and phenotype

P(AA) = p^2 $(\text{Ph} = \alpha)$
P(Aa) = $2pq$ $(\text{Ph} = d)$
P(aa) = q^2 $(\text{Ph} = -\alpha)$

Show that E[Ph] = $\alpha * (p - q) + 2pqd$



Expectation of a random variable – Exercise (Answer)

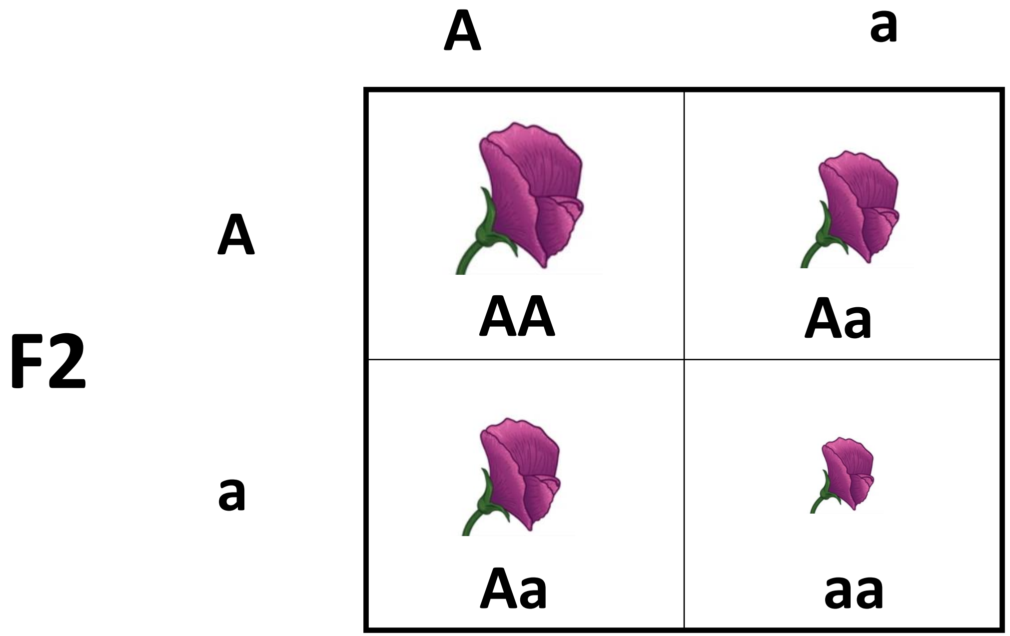
Let us assume that

P(A) = p
P(a) = q = 1 – p

In a F2, we have the following frequencies and phenotype

P(AA) = p^2 $(\mathrm{Ph} = \alpha)$
P(Aa) = $2pq$ $(\mathrm{Ph} = d)$
P(aa) = q^2 $(\mathrm{Ph} = -\alpha)$

E[Ph] = $\alpha * p^2 + 2pqd - \alpha * q^2$
E[Ph] = $\alpha * (p^2 - q^2) + 2pqd$
E[Ph] = $\alpha * (p + q)(p - q) + 2pqd$
E[Ph] = $\alpha * 1 * (p - q) + 2pqd$
E[Ph] = $\alpha * (p - q) + 2pqd$



Variance of a random variable

X is a random variable defined on a proability space (Ω, F, P)

The variance $V(X)$ is the expected value of the squared deviation from the mean

$$\mathrm{V}(X) = E[(X - \mu)^2]$$

The variance expression can be expanded as follow:

$$\begin{aligned}\operatorname{Var}(X) &= \operatorname{E}\left[(X - \operatorname{E}[X])^2\right] \\ &= \operatorname{E}\left[X^2 - 2X\operatorname{E}[X] + \operatorname{E}[X]^2\right] \\ &= \operatorname{E}\left[X^2\right] - 2\operatorname{E}[X]\operatorname{E}[X] + \operatorname{E}[X]^2 \\ &= \operatorname{E}\left[X^2\right] - \operatorname{E}[X]^2\end{aligned}$$

$$\operatorname{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

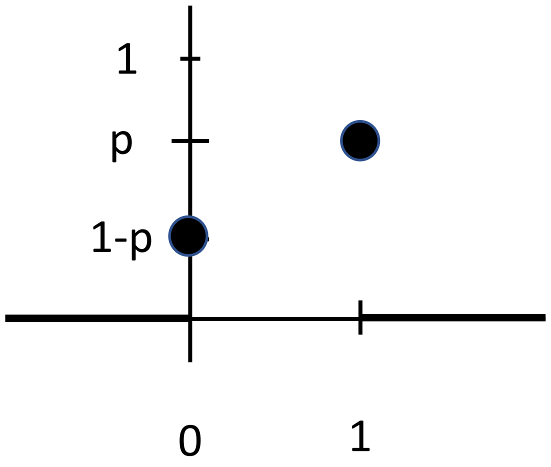
$$\operatorname{Var}(X) = \operatorname{Cov}(X, X).$$

$$\operatorname{cov}(X, Y) = \operatorname{E}\left[(X - \operatorname{E}[X])(Y - \operatorname{E}[Y])\right]$$

Expectation and variance of a Bernoulli random variable

Let us assume a Bernoulli variable with the Following probability density function

$$P(X = x) = \begin{cases} 1 - p & if\ x = 0 \\ p & if\ x = 1 \\ 0 & Otherwise \end{cases}$$



Show that the expectation and variance of such variable are equal to p and p(1-p), respectively

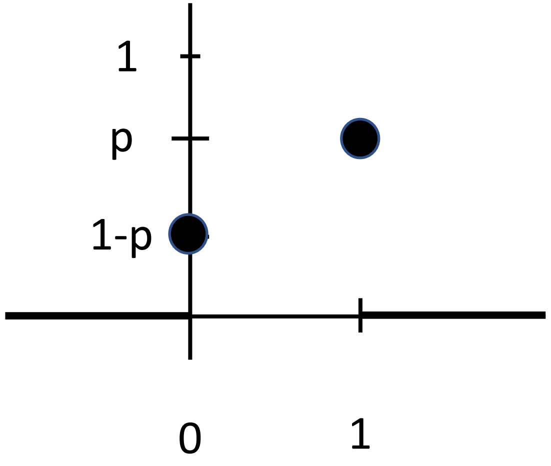
E[X] = p

V[X] = p (1 -p) = pq

Expectation and variance of a Bernoulli random variable

Let us assume a Bernoulli variable with the Following probability density function

$$P(X = x) = \begin{cases} 1 - p & if\ x = 0 \\ p & if\ x = 1 \\ 0 & Otherwise \end{cases}$$



$$\mathrm{E}[X] = \Pr(X = 1) \cdot 1 + \Pr(X = 0) \cdot 0 = p \cdot 1 + q \cdot 0 = p.$$

$$\mathrm{E}[X^2] = \Pr(X = 1) \cdot 1^2 + \Pr(X = 0) \cdot 0^2 = p \cdot 1^2 + q \cdot 0^2 = p = \mathrm{E}[X]$$

From this follows

$$\operatorname{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = \mathrm{E}[X] - \mathrm{E}[X]^2 = p - p^2 = p(1 - p) = pq$$[3]

Probability density function – discrete variable

Probability space (Ω, F, P)

P : Probability function. A function that assign a probability (number between 0 and 1) to each events in the event space.

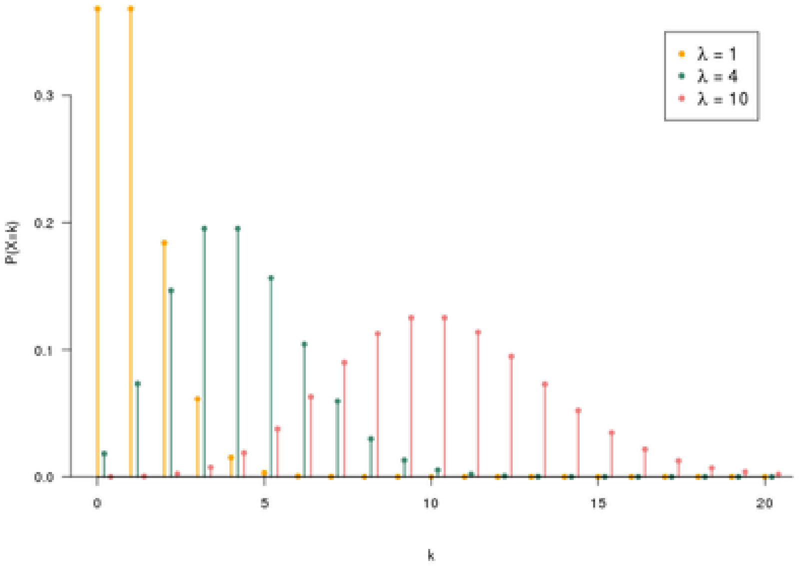
Ex: ***Poisson distribution***: describe the number of observation of a certain phenomenon with known expectation (λ) over a certain period of time independently of the previous measurement.

Probability density function (pdf)

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

In R: dpois(x = k, lambda = λ)

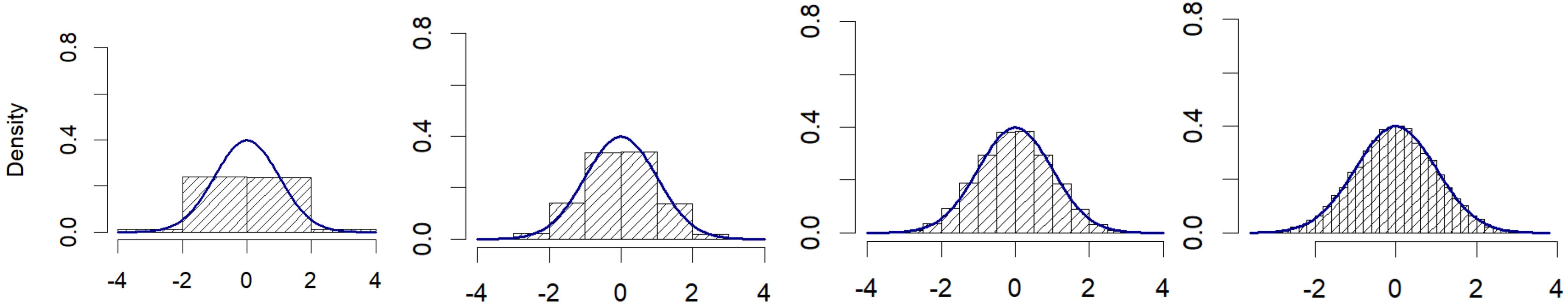
$$\mathrm{E}[\mathrm{X}] = \mathrm{V}(X) = \lambda$$



Probability density function – continuous variable

Normal distribution

From discrete to continuous (go to limit)



Probability density function (pdf)

$$P(X = x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{1}{2} [\left(\frac{x\ - \mu}{\sigma}\right)^2]$$

$$\mathrm{E[X]} = \int_{LB}^{UB} xf(x)dx = \mu$$

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mu^2 = \left(\int_{-\infty}^{\infty} x^2 \cdot f(x)\, dx\right) - \mu^2 \ = \sigma^2$$

Probability density function – continuous variable

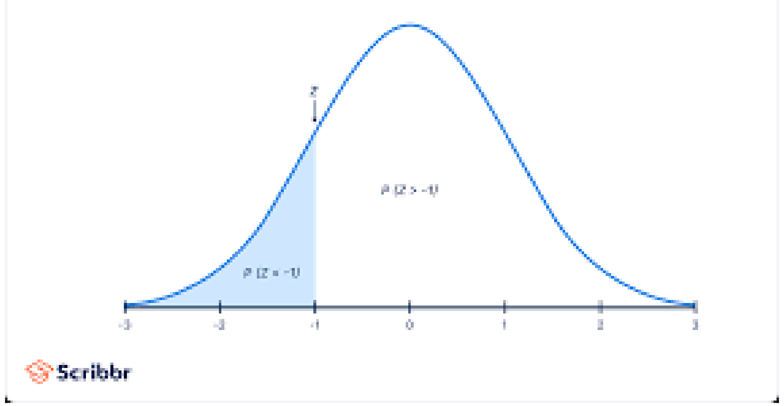
Normal distribution

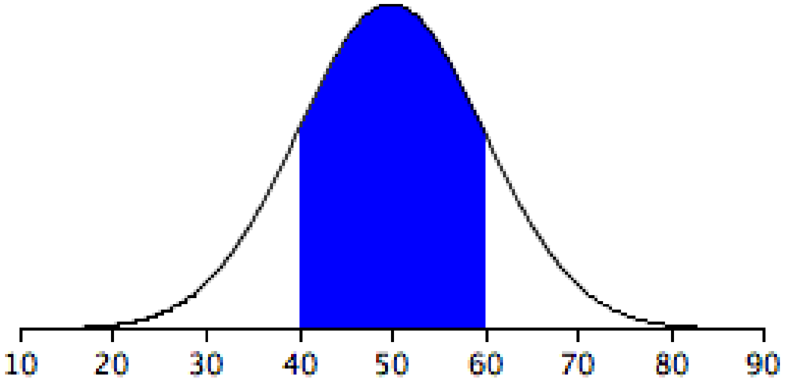
Cumulative distribution function (cdf)

$$F_X(x) = P(X \leq x|\mu, \sigma^2) = \phi\left(\frac{x \; - \mu}{\sigma}\right)$$

$$P(a \leq X \leq b) = \; F_X(b) \; - F_X(a)$$

In R: pnorm()





Expectation and variance properties

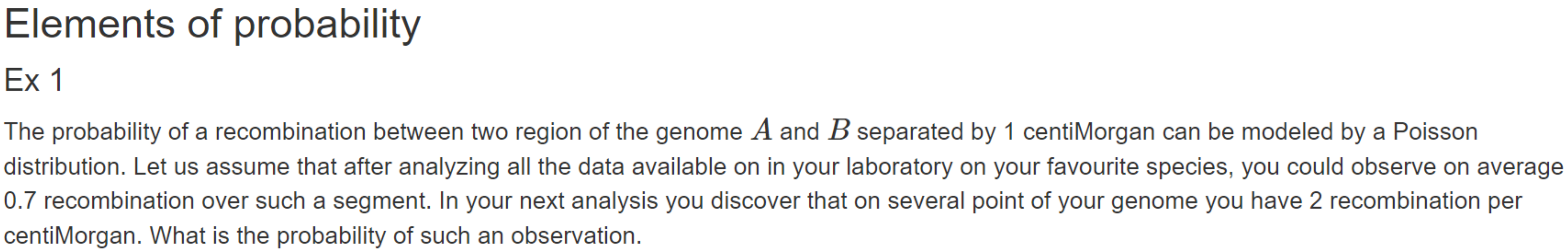
$$E(aX) = aE(X)$$

$$E(aX + bY) = aE(X) + bE(Y)$$

$$\mathrm{V}(aX) = a^2V(X)$$

$$\mathrm{V}(AX) = AV(X)A'$$

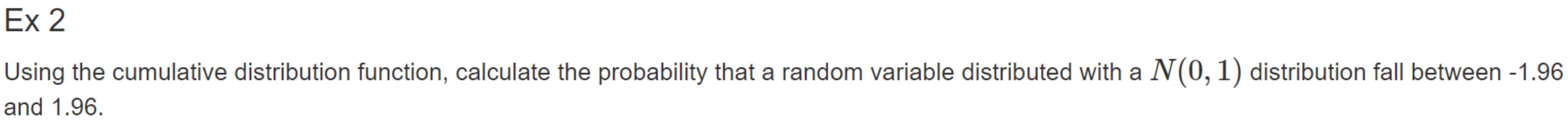
Exercises



Elements of probability

Ex 1

The probability of a recombination between two region of the genome A and B separated by 1 centiMorgan can be modeled by a Poisson distribution. Let us assume that after analyzing all the data available on in your laboratory on your favourite species, you could observe on average 0.7 recombination over such a segment. In your next analysis you discover that on several point of your genome you have 2 recombination per centiMorgan. What is the probability of such an observation.



Ex 2

Using the cumulative distribution function, calculate the probability that a random variable distributed with a $N(0,1)$ distribution fall between -1.96 and 1.96.

Matrix algebra

Matrix algebra

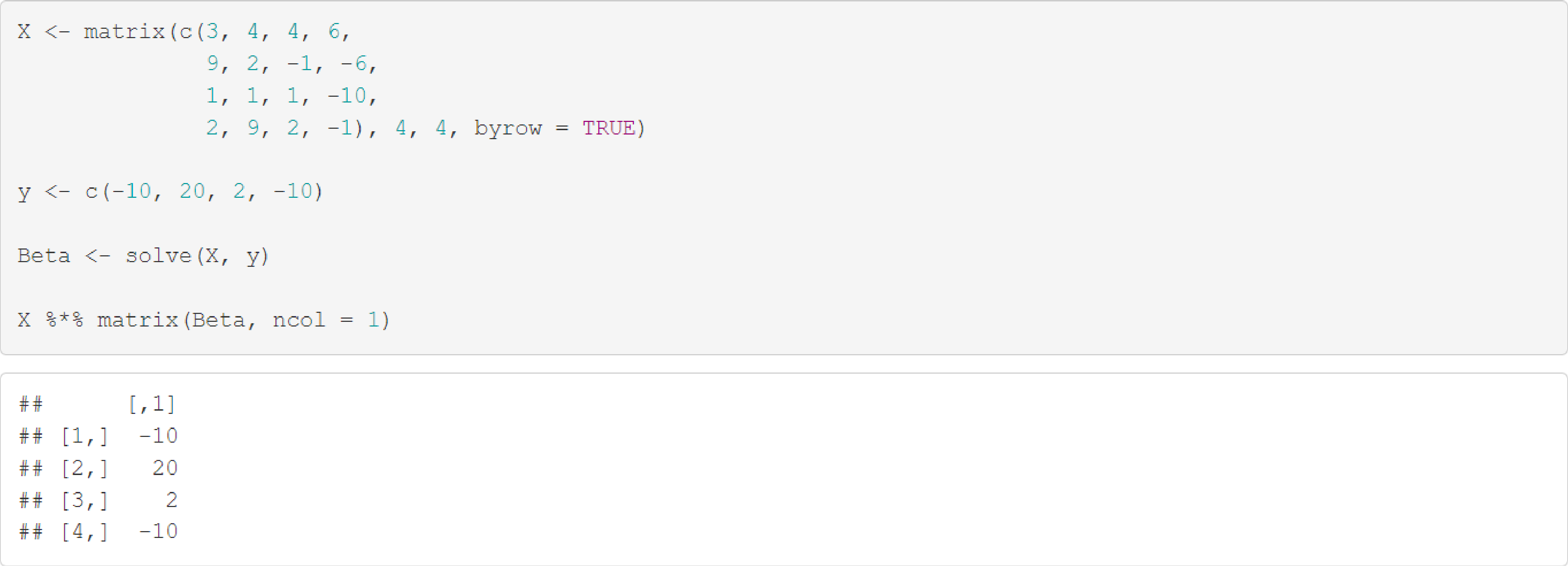
Matrices can be seen as a system of linear equations. For example:

$$
\begin{aligned}
3x_1 + 4x_2 + 4x_3 + 6x_4 &= -10\\
9x_1 + 2x_2 - 1x_3 - 6x_4 &= 20\\
1x_1 + 1x_2 + 1x_3 - 10x_4 &= 2\\
2x_1 + 9x_2 + 2x_3 + 1x_4 &= -10
\end{aligned}
$$

$$
\begin{pmatrix} 3 & 4 & 4 & 6 \\ 9 & 2 & -1 & -6 \\ 1 & 1 & 1 & -10 \\ 2 & 9 & 2 & 1 \end{pmatrix}
\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}
=
\begin{pmatrix} -10 \\ 20 \\ 2 \\ -10 \end{pmatrix}
$$

Matrix algebra

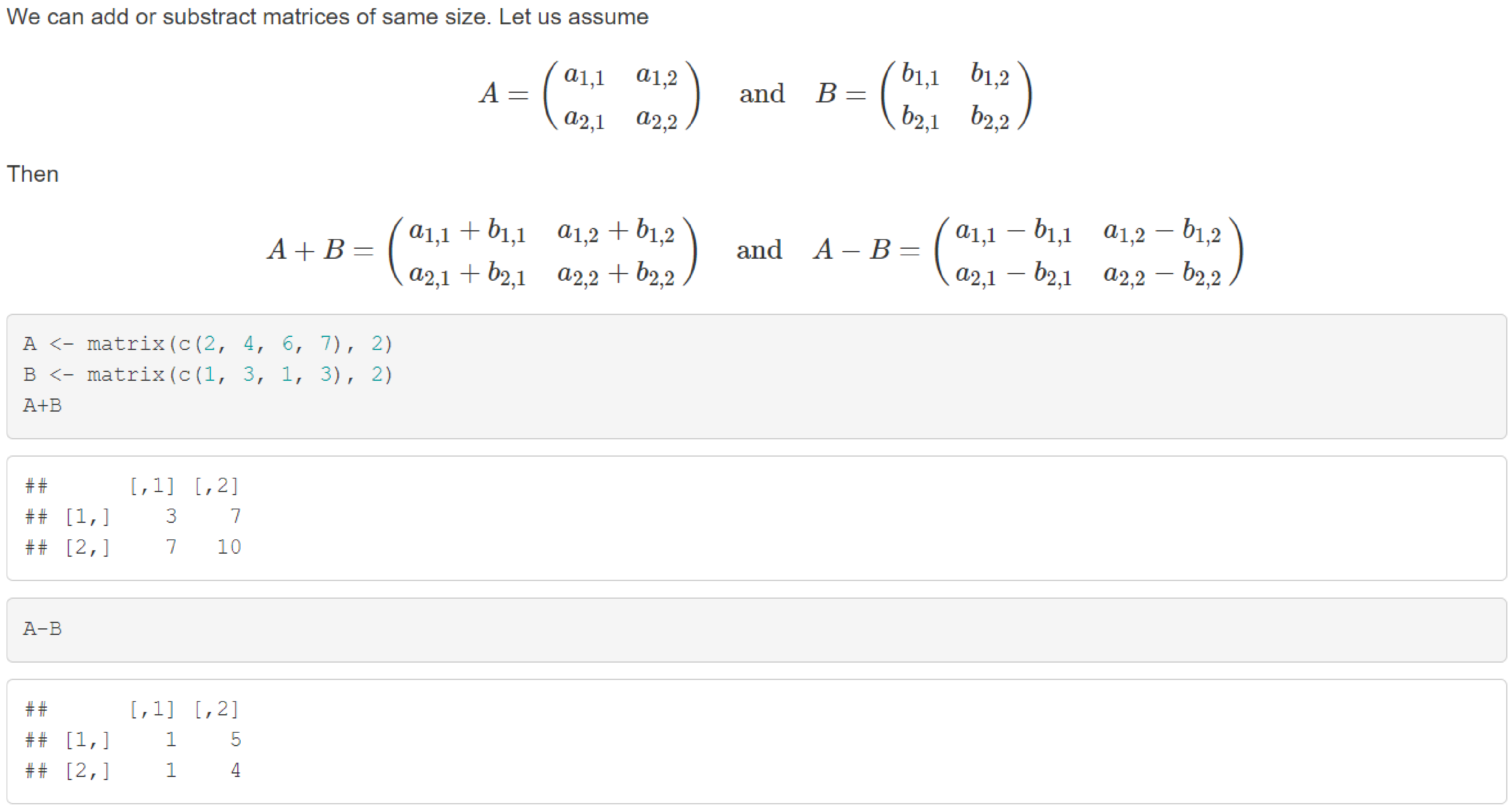
Matrices help to summary information and solve this kind of systems



```r
X <- matrix(c(3, 4, 4, 6,
              9, 2, -1, -6,
              1, 1, 1, -10,
              2, 9, 2, -1), 4, 4, byrow = TRUE)

y <- c(-10, 20, 2, -10)

Beta <- solve(X, y)

X %*% matrix(Beta, ncol = 1)
```

```
##      [,1]
## [1,]  -10
## [2,]   20
## [3,]    2
## [4,]  -10
```

Matrix basic operations – Addition, substraction, transpose



We can add or substract matrices of same size. Let us assume

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{pmatrix}$$

Then

$$A + B = \begin{pmatrix} a_{1,1} + b_{1,1} & a_{1,2} + b_{1,2} \\ a_{2,1} + b_{2,1} & a_{2,2} + b_{2,2} \end{pmatrix} \quad \text{and} \quad A - B = \begin{pmatrix} a_{1,1} - b_{1,1} & a_{1,2} - b_{1,2} \\ a_{2,1} - b_{2,1} & a_{2,2} - b_{2,2} \end{pmatrix}$$

```
A <- matrix(c(2, 4, 6, 7), 2)
B <- matrix(c(1, 3, 1, 3), 2)
A+B
```

```
##      [,1] [,2]
## [1,]    3    7
## [2,]    7   10
```

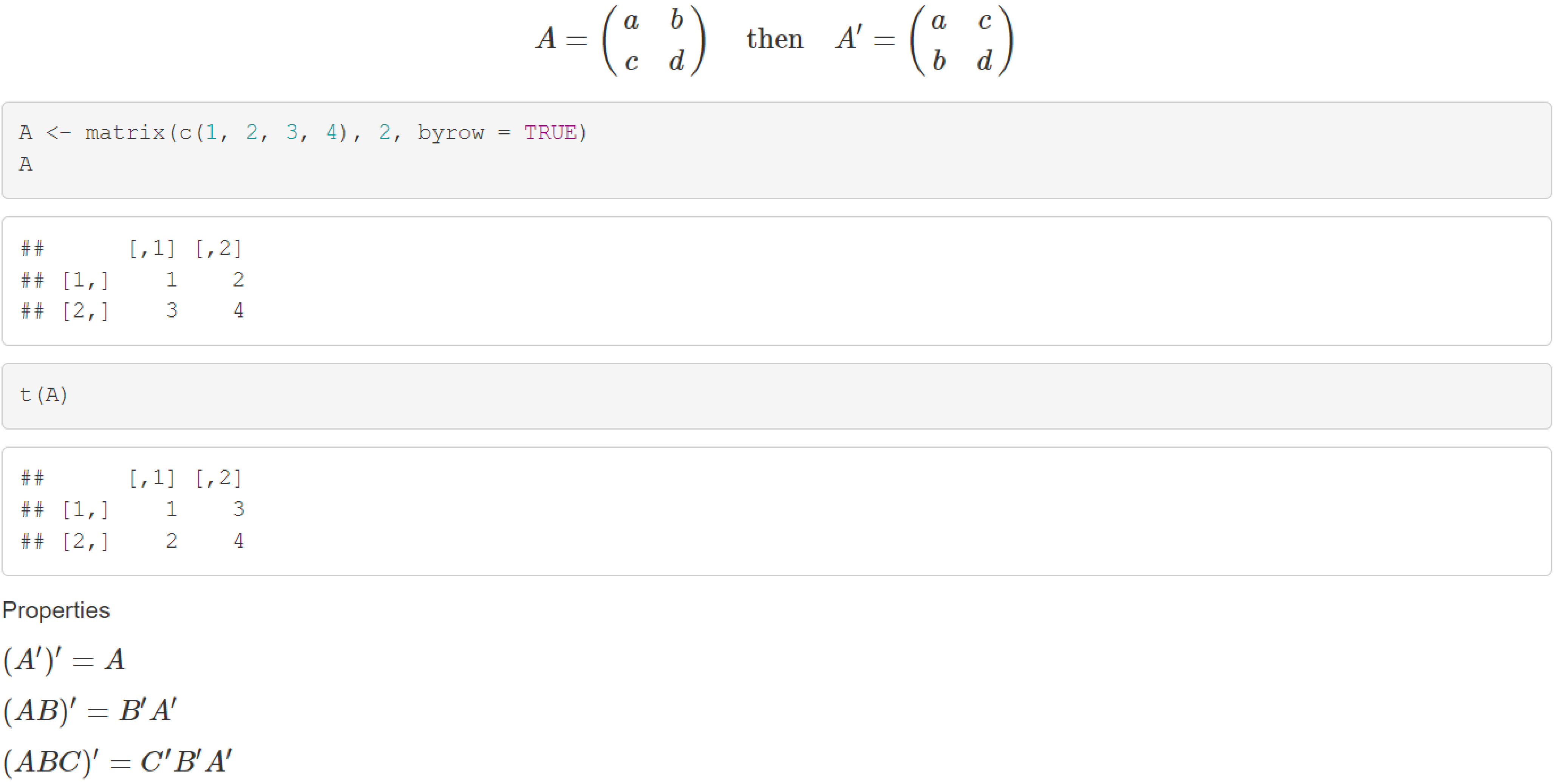
```
A-B
```

```
##      [,1] [,2]
## [1,]    1    5
## [2,]    1    4
```

Properties: $A + B = B + A$; $A - B \neq B - A$

$A + (B + C) = (A + B) + C$; $A - (B - C) \neq (A - B) - C$

Matrix basic operations – Addition, substraction, transpose



$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{then} \quad A' = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

```r
A <- matrix(c(1, 2, 3, 4), 2, byrow = TRUE)
A
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
```

```r
t(A)
```

```
##      [,1] [,2]
## [1,]    1    3
## [2,]    2    4
```

Properties

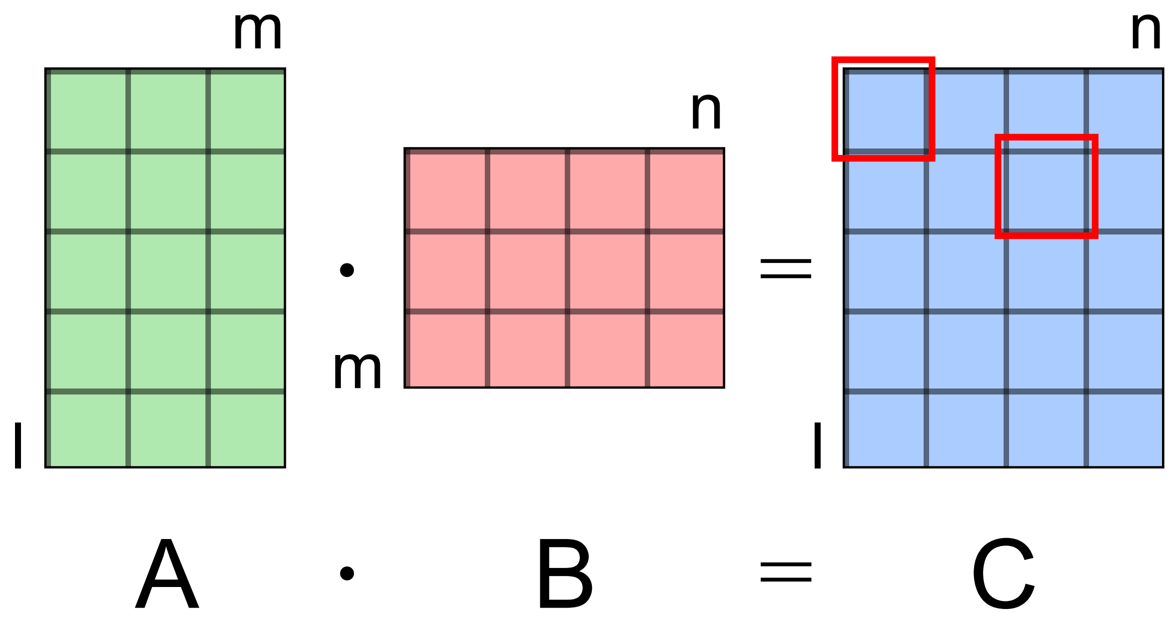
$(A')' = A$

$(AB)' = B'A'$

$(ABC)' = C'B'A'$

Trace: sum of the diagonal elements. In R: sum(diag(A))

Matrix basic operations – Multiplication

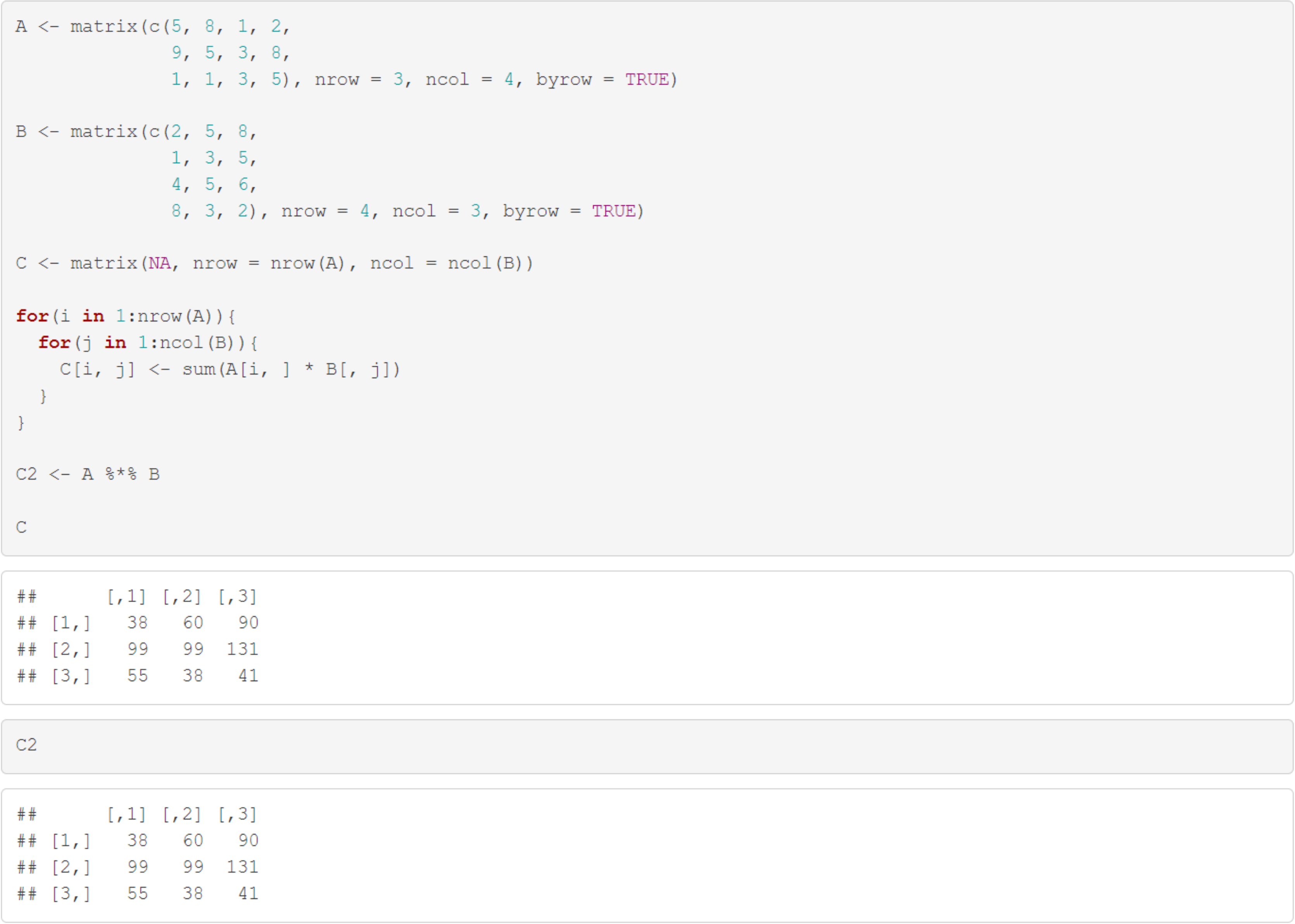


$$C_{11} = A_{1}.' B_{.1} = a_{11} * b_{11} + a_{12} * b_{21} + a_{13} * b_{31}$$

$$C_{23} = A_{2}.' B_{.3} = a_{21} * b_{13} + a_{22} * b_{23} + a_{23} * b_{33}$$

$$C_{ij} = A_{i}.' B_{.j} = \sum_{k=1}^{m} a_{ik} * b_{kj}$$

Matrix basic operations – Multiplication



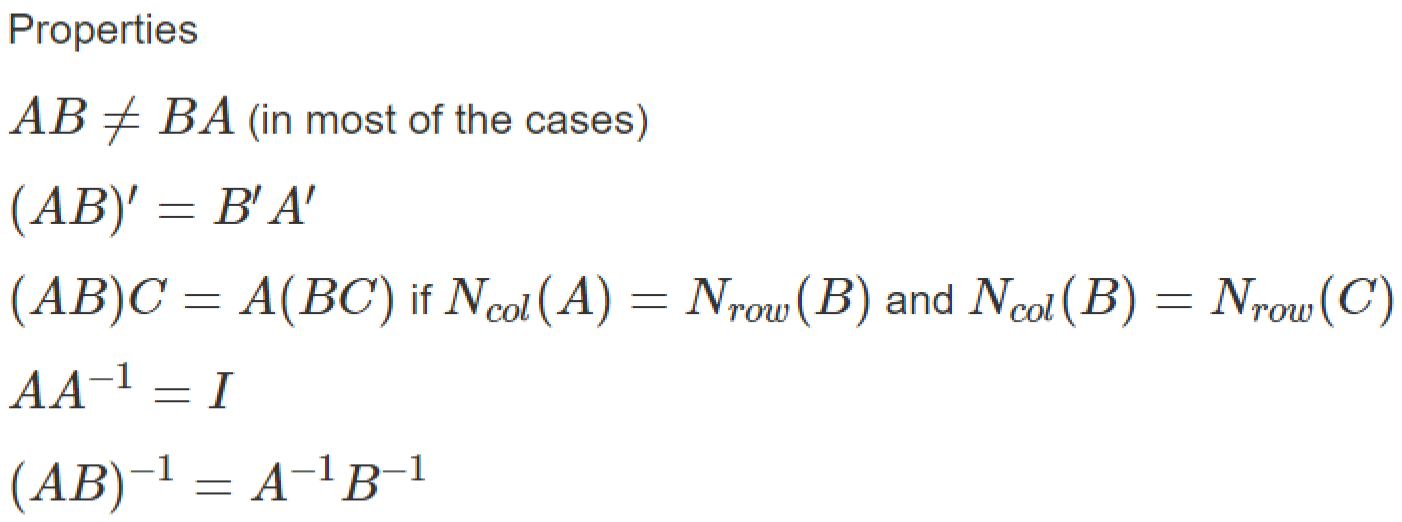
```
A <- matrix(c(5, 8, 1, 2,
              9, 5, 3, 8,
              1, 1, 3, 5), nrow = 3, ncol = 4, byrow = TRUE)

B <- matrix(c(2, 5, 8,
              1, 3, 5,
              4, 5, 6,
              8, 3, 2), nrow = 4, ncol = 3, byrow = TRUE)

C <- matrix(NA, nrow = nrow(A), ncol = ncol(B))

for(i in 1:nrow(A)){
  for(j in 1:ncol(B)){
    C[i, j] <- sum(A[i, ] * B[, j])
  }
}

C2 <- A %*% B

C
```

```
##      [,1] [,2] [,3]
## [1,]   38   60   90
## [2,]   99   99  131
## [3,]   55   38   41
```

```
C2
```

```
##      [,1] [,2] [,3]
## [1,]   38   60   90
## [2,]   99   99  131
## [3,]   55   38   41
```

Matrix basic operations – Multiplication



Properties

$AB \neq BA$ (in most of the cases)

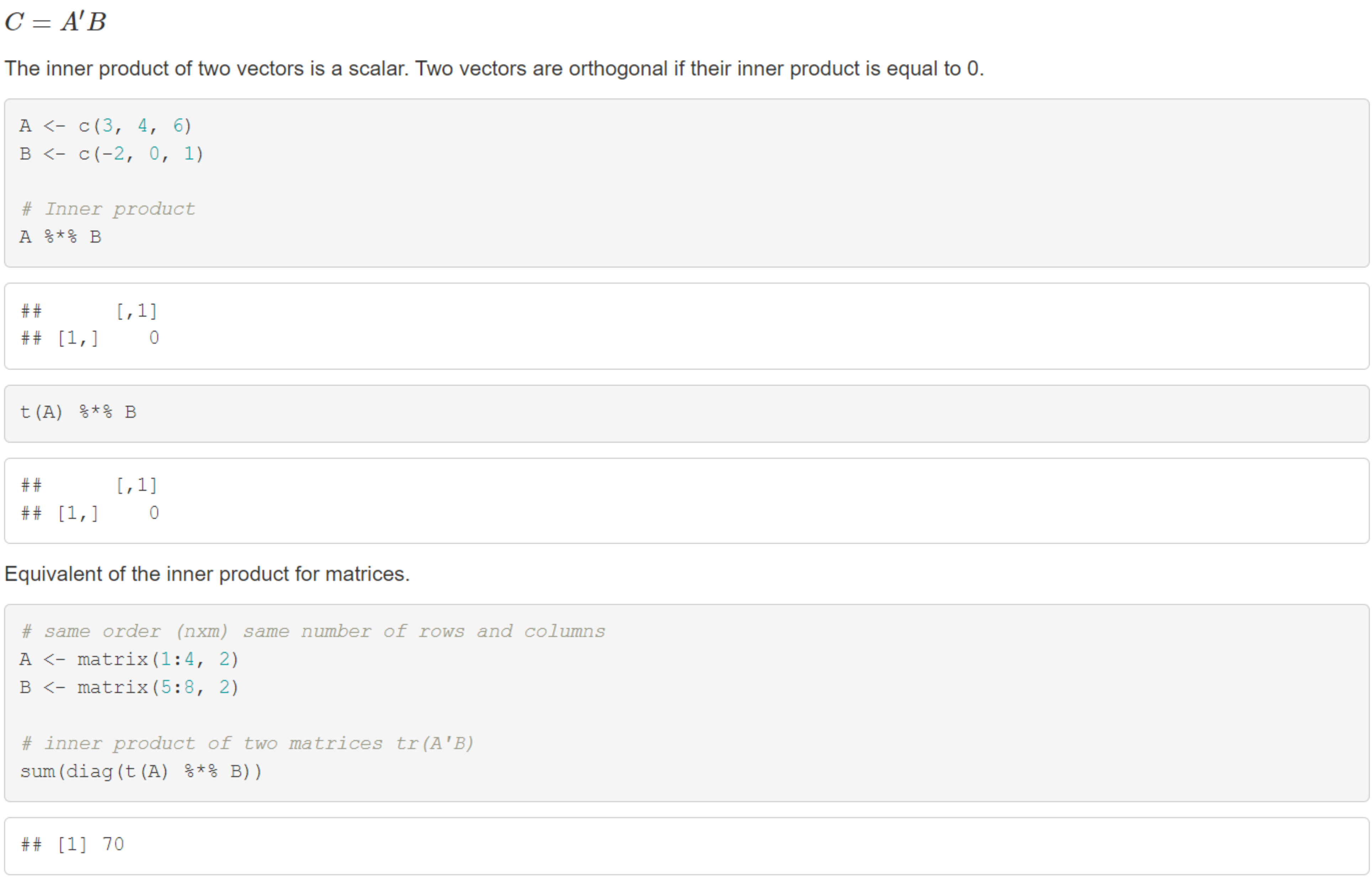
$(AB)' = B'A'$

$(AB)C = A(BC)$ if $N_{col}(A) = N_{row}(B)$ and $N_{col}(B) = N_{row}(C)$

$AA^{-1} = I$

$(AB)^{-1} = A^{-1}B^{-1}$

Matrix basic operations – Multiplication inner product



$C = A'B$

The inner product of two vectors is a scalar. Two vectors are orthogonal if their inner product is equal to 0.

```
A <- c(3, 4, 6)
B <- c(-2, 0, 1)

# Inner product
A %*% B
```

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

```
t(A) %*% B
```

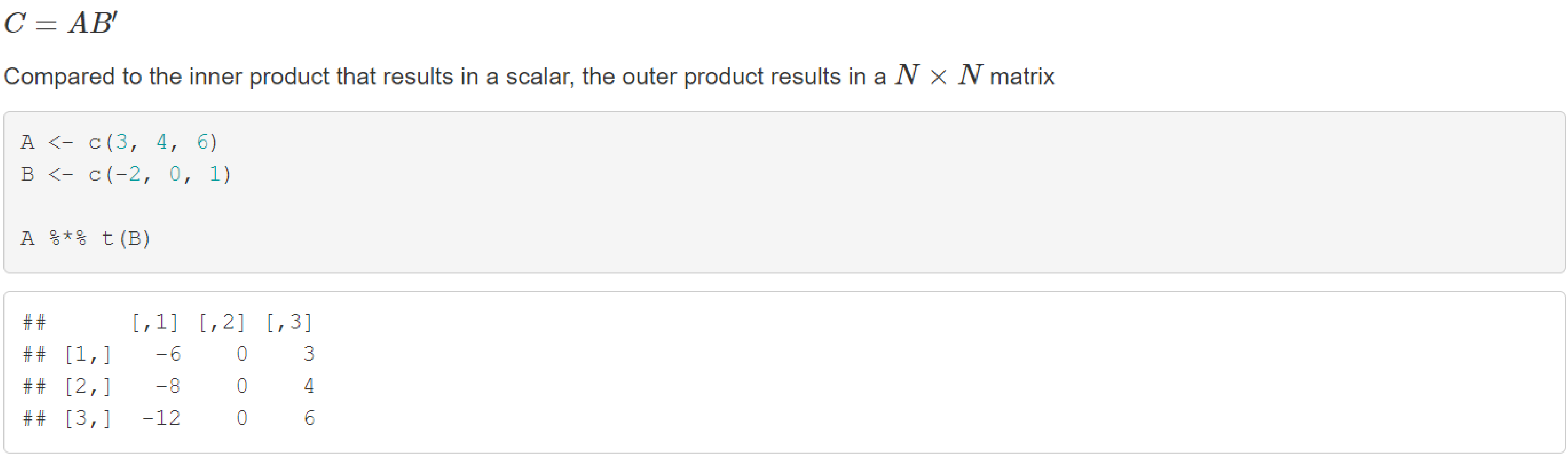
```
##      [,1]
## [1,]    0
```

Equivalent of the inner product for matrices.

```
# same order (nxm) same number of rows and columns
A <- matrix(1:4, 2)
B <- matrix(5:8, 2)

# inner product of two matrices tr(A'B)
sum(diag(t(A) %*% B))
```

```
## [1] 70
```

Matrix basic operations – Multiplication outer product



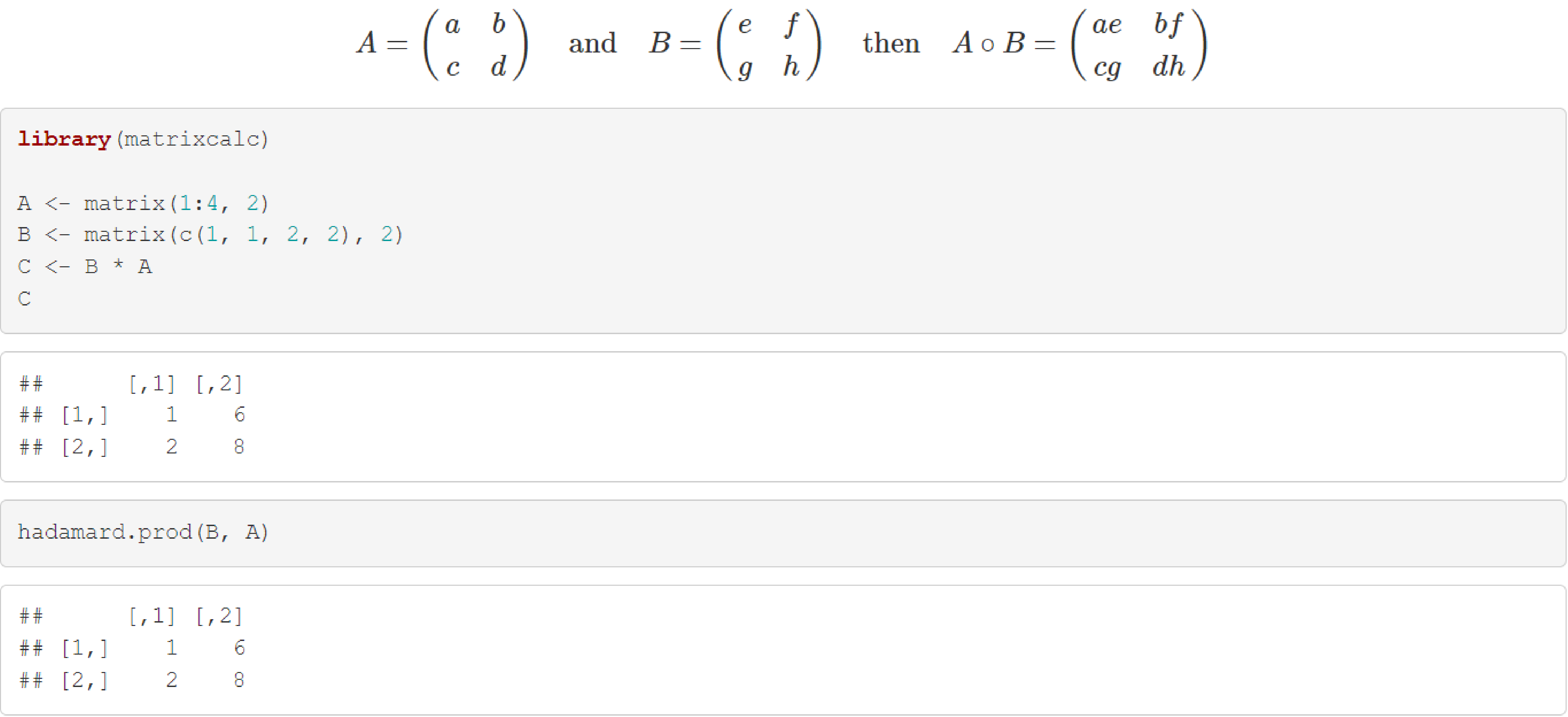
$C = AB'$

Compared to the inner product that results in a scalar, the outer product results in a $N \times N$ matrix

```
A <- c(3, 4, 6)
B <- c(-2, 0, 1)

A %*% t(B)
```

```
##      [,1] [,2] [,3]
## [1,]   -6    0    3
## [2,]   -8    0    4
## [3,]  -12    0    6
```

Matrix basic operations – Multiplication Hadamar product



$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} e & f \\ g & h \end{pmatrix} \quad \text{then} \quad A \circ B = \begin{pmatrix} ae & bf \\ cg & dh \end{pmatrix}$$

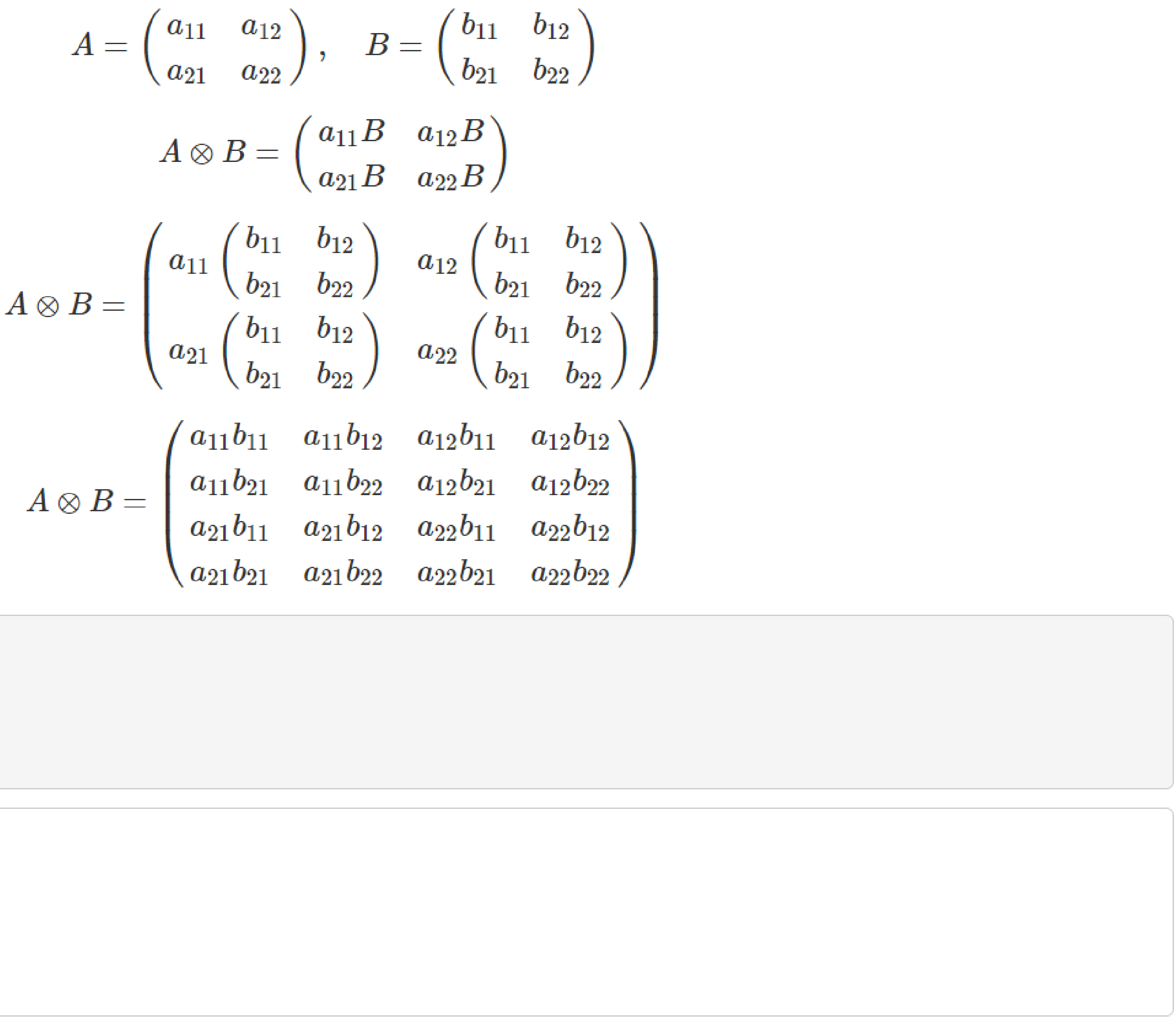
```r
library(matrixcalc)

A <- matrix(1:4, 2)
B <- matrix(c(1, 1, 2, 2), 2)
C <- B * A
C
```

```
##      [,1] [,2]
## [1,]    1    6
## [2,]    2    8
```

```r
hadamard.prod(B, A)
```

```
##      [,1] [,2]
## [1,]    1    6
## [2,]    2    8
```

Matrix basic operations – Multiplication Kronecker product



$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B \\ a_{21}B & a_{22}B \end{pmatrix}$$

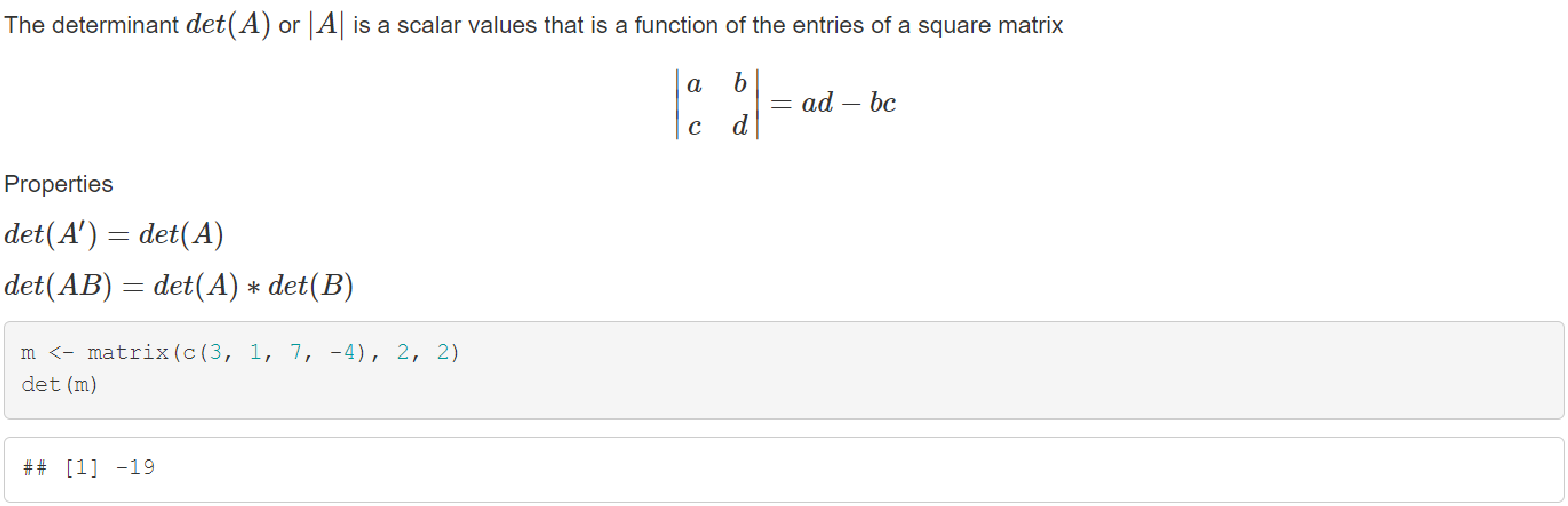
$$A \otimes B = \begin{pmatrix} a_{11}\begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} & a_{12}\begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \\ a_{21}\begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} & a_{22}\begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \end{pmatrix}$$

$$A \otimes B = \begin{pmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{12}b_{11} & a_{12}b_{12} \\ a_{11}b_{21} & a_{11}b_{22} & a_{12}b_{21} & a_{12}b_{22} \\ a_{21}b_{11} & a_{21}b_{12} & a_{22}b_{11} & a_{22}b_{12} \\ a_{21}b_{21} & a_{21}b_{22} & a_{22}b_{21} & a_{22}b_{22} \end{pmatrix}$$

```r
A <- matrix(1:4, 2)
B <- matrix(c(1, 1, 2, 2), 2)
C <- B %x% A
C
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    3    2    6
## [2,]    2    4    4    8
## [3,]    1    3    2    6
## [4,]    2    4    4    8
```

Matrix basic operations – Determinant



The determinant $det(A)$ or $|A|$ is a scalar values that is a function of the entries of a square matrix

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

Properties

$$det(A') = det(A)$$

$$det(AB) = det(A) * det(B)$$

```
m <- matrix(c(3, 1, 7, -4), 2, 2)
det(m)
```

```
## [1] -19
```

Matrix basic operations – Determinant

let assume the following block matrix with sub-matrices A, B, C, D of dimensions $m \times m$, $m \times n$, $n \times m$, and $n \times n$

$$det\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

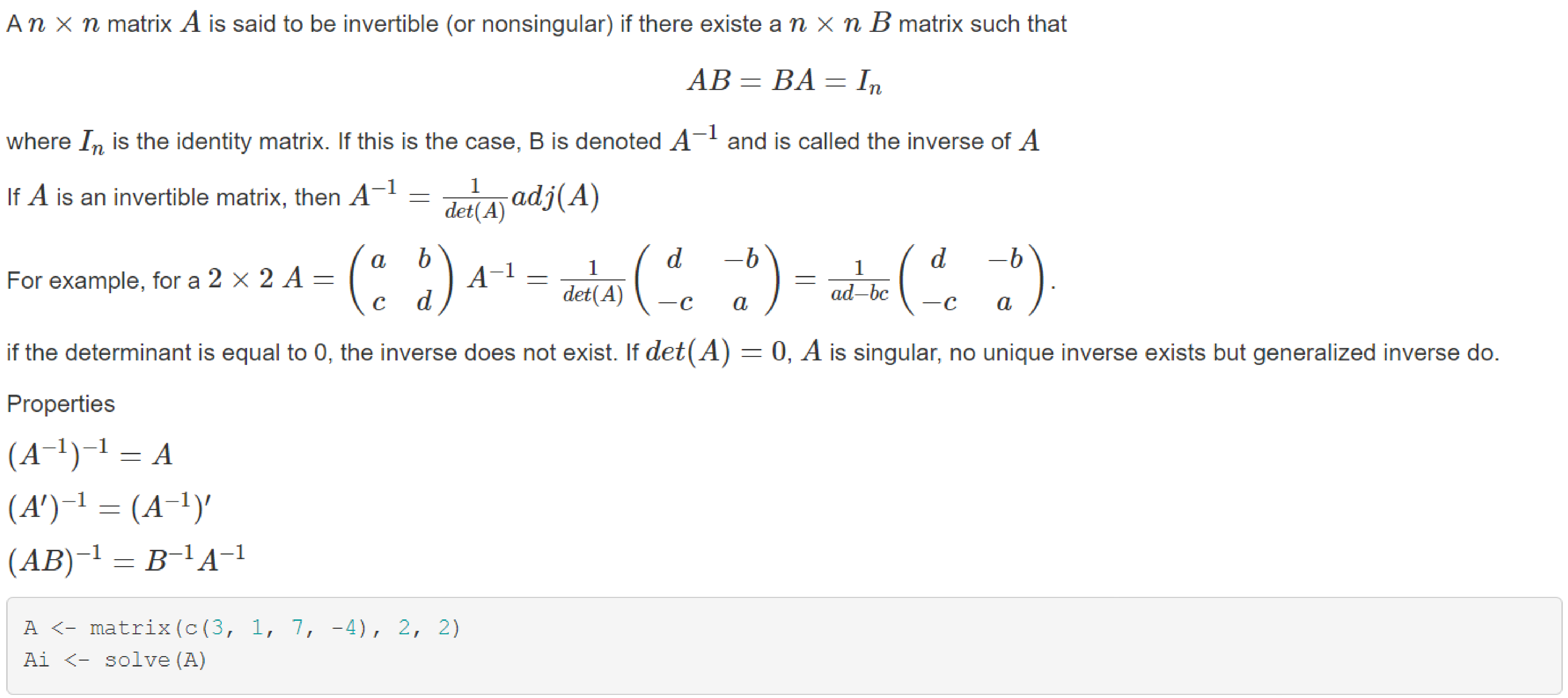
If A is invertible.

$$det\begin{pmatrix} A & B \\ C & D \end{pmatrix} = det(A) * det(D - CA^{-1}B)$$

If D is invertible

$$det\begin{pmatrix} A & B \\ C & D \end{pmatrix} = det(D) * det(A - BD^{-1}C)$$

Matrix basic operations – Inversion



A $n \times n$ matrix A is said to be invertible (or nonsingular) if there existe a $n \times n$ B matrix such that

$$AB = BA = I_n$$

where I_n is the identity matrix. If this is the case, B is denoted A^{-1} and is called the inverse of A

If A is an invertible matrix, then $A^{-1} = \frac{1}{det(A)} adj(A)$

For example, for a 2×2 $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ $A^{-1} = \frac{1}{det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

if the determinant is equal to 0, the inverse does not exist. If $det(A) = 0$, A is singular, no unique inverse exists but generalized inverse do.

Properties

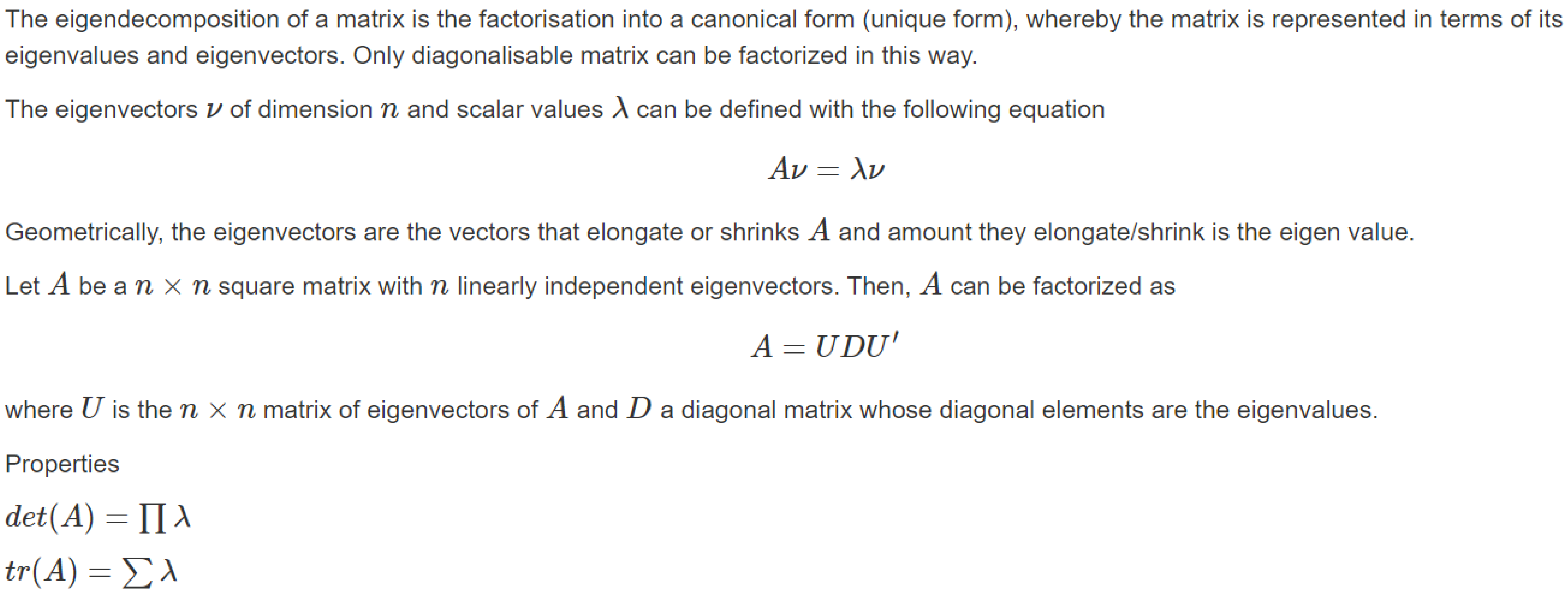
$(A^{-1})^{-1} = A$

$(A')^{-1} = (A^{-1})'$

$(AB)^{-1} = B^{-1}A^{-1}$

```
A <- matrix(c(3, 1, 7, -4), 2, 2)
Ai <- solve(A)
```

Matrix basic operations – Eigenvalue decomposition



The eigendecomposition of a matrix is the factorisation into a canonical form (unique form), whereby the matrix is represented in terms of its eigenvalues and eigenvectors. Only diagonalisable matrix can be factorized in this way.

The eigenvectors ν of dimension n and scalar values λ can be defined with the following equation

$$A\nu = \lambda\nu$$

Geometrically, the eigenvectors are the vectors that elongate or shrinks A and amount they elongate/shrink is the eigen value.

Let A be a $n \times n$ square matrix with n linearly independent eigenvectors. Then, A can be factorized as

$$A = UDU'$$

where U is the $n \times n$ matrix of eigenvectors of A and D a diagonal matrix whose diagonal elements are the eigenvalues.

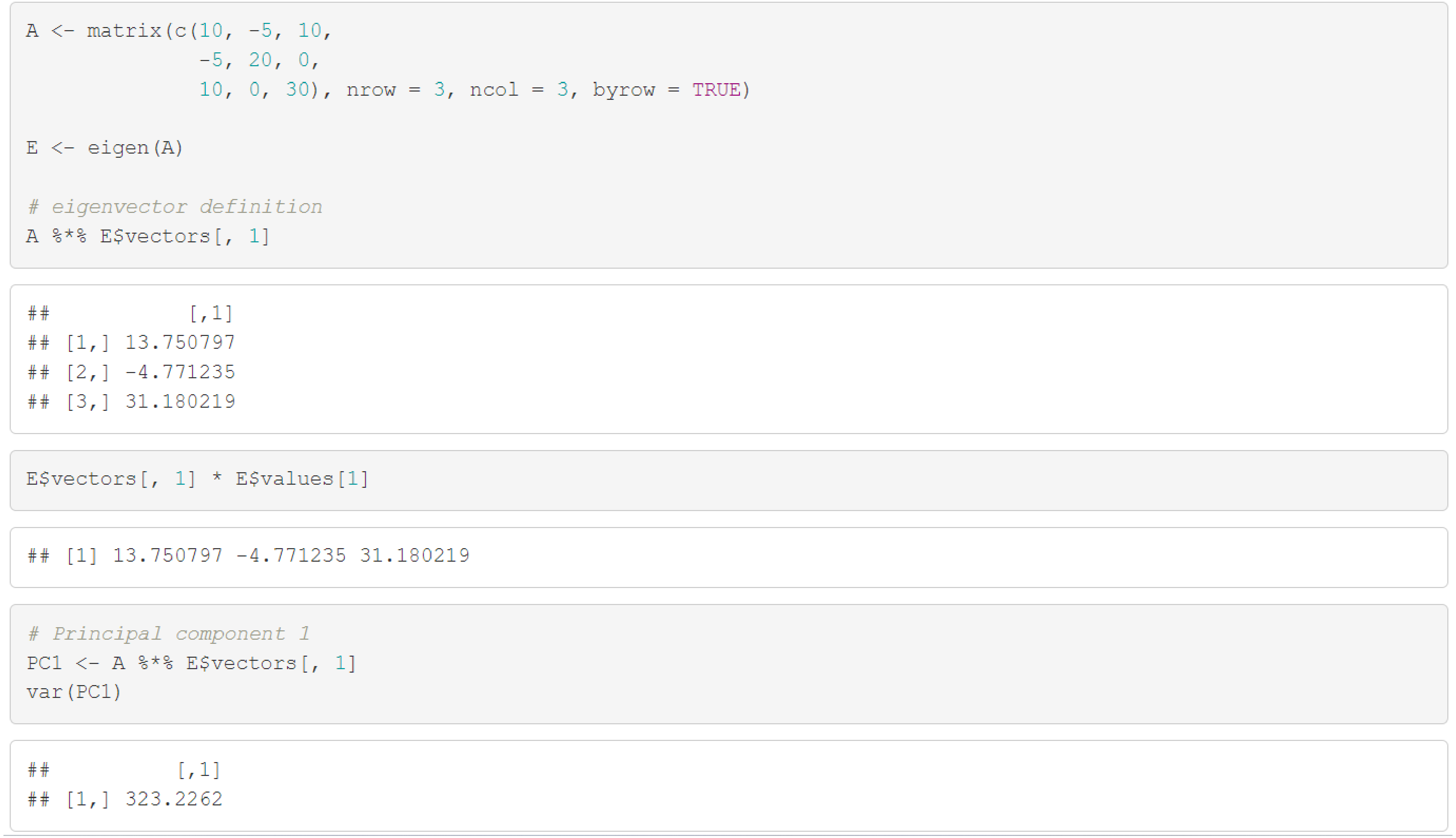
Properties

$det(A) = \prod \lambda$

$tr(A) = \sum \lambda$

Usage: Reduce the matrix dimensionality
simplified and faster computation

Matrix basic operations – Eigenvalue decomposition



```r
A <- matrix(c(10, -5, 10,
              -5, 20, 0,
              10, 0, 30), nrow = 3, ncol = 3, byrow = TRUE)

E <- eigen(A)

# eigenvector definition
A %*% E$vectors[, 1]
```

```
##           [,1]
## [1,] 13.750797
## [2,] -4.771235
## [3,] 31.180219
```

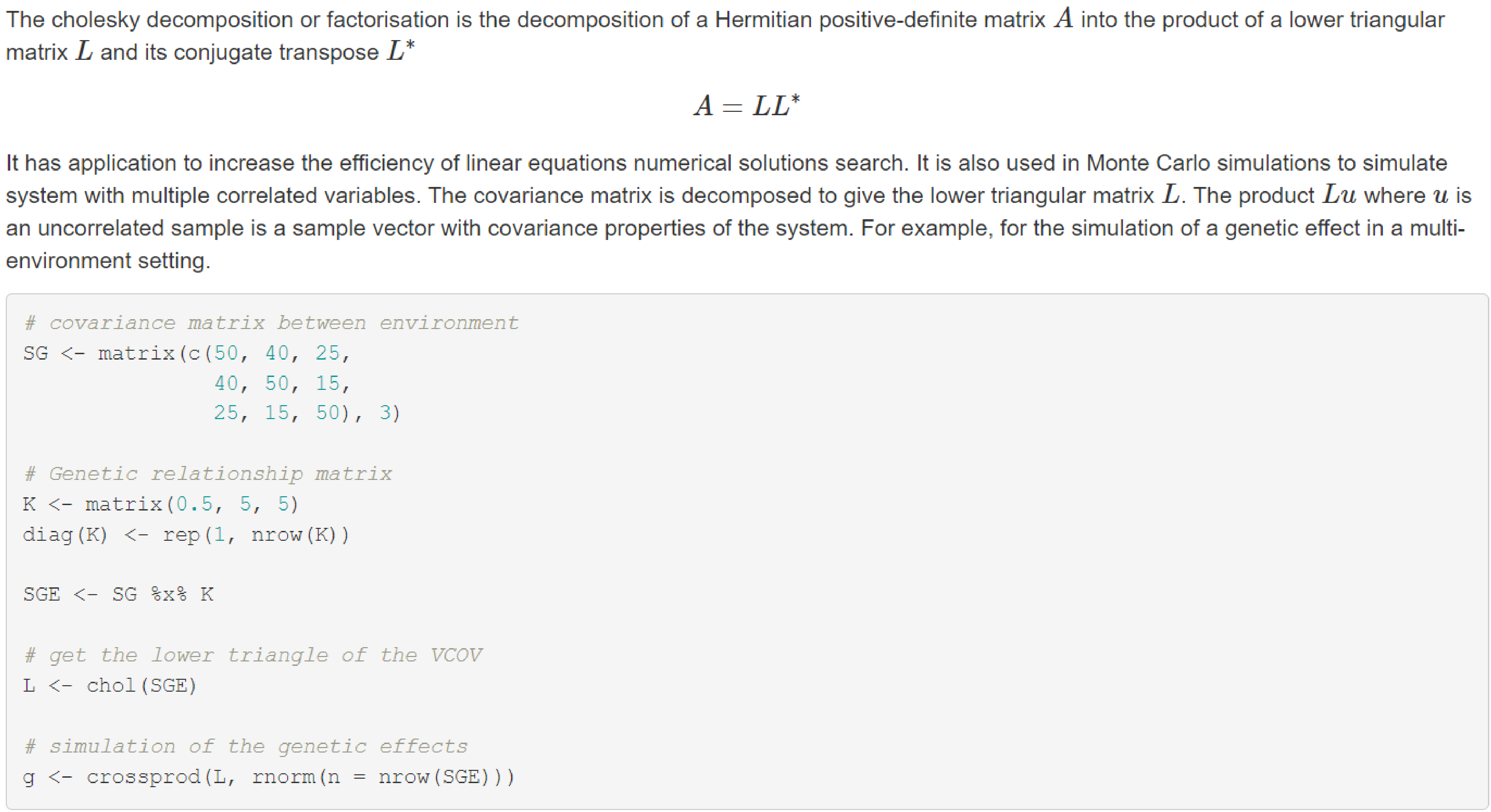
```r
E$vectors[, 1] * E$values[1]
```

```
## [1] 13.750797 -4.771235 31.180219
```

```r
# Principal component 1
PC1 <- A %*% E$vectors[, 1]
var(PC1)
```

```
##          [,1]
## [1,] 323.2262
```

Matrix basic operations – Cholesky decomposition



The cholesky decomposition or factorisation is the decomposition of a Hermitian positive-definite matrix A into the product of a lower triangular matrix L and its conjugate transpose L^*

$$A = LL^*$$

It has application to increase the efficiency of linear equations numerical solutions search. It is also used in Monte Carlo simulations to simulate system with multiple correlated variables. The covariance matrix is decomposed to give the lower triangular matrix L. The product Lu where u is an uncorrelated sample is a sample vector with covariance properties of the system. For example, for the simulation of a genetic effect in a multi-environment setting.

```r
# covariance matrix between environment
SG <- matrix(c(50, 40, 25,
               40, 50, 15,
               25, 15, 50), 3)

# Genetic relationship matrix
K <- matrix(0.5, 5, 5)
diag(K) <- rep(1, nrow(K))

SGE <- SG %x% K

# get the lower triangle of the VCOV
L <- chol(SGE)

# simulation of the genetic effects
g <- crossprod(L, rnorm(n = nrow(SGE)))
```

Matrix algebra - Exercises

Ex 1

Calculate the following matrix $B = ((A')^{-1})^2$ with

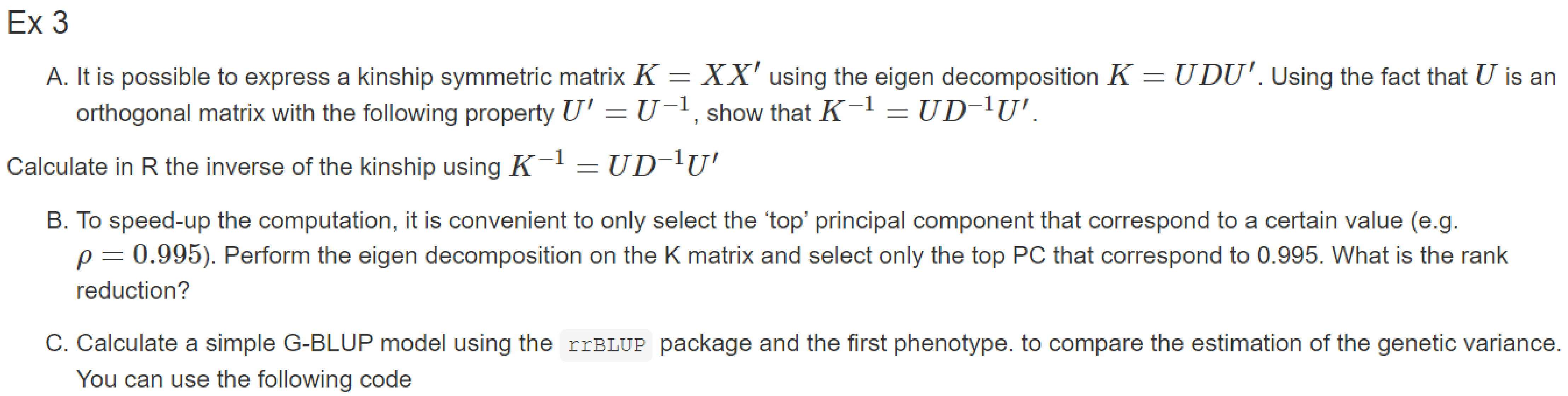
$$A = \begin{pmatrix} 1 & 0 \\ 2 & 3 \end{pmatrix}$$

Ex 2

A. Using the provided marker matrix X, calculate a kinship matrix using the van Raden method $K = \frac{ZZ'}{2\sum p_i(1-p_i)}$, where Z is the centered marker matrix with $z_{ij} = x_{ij} - 2p_i$ and p_i frequency of the i^{th} marker.

B. Using the microbenchmark package compare the speed of inversion of the kinship matrix using standard function `slove()` and cholesky decomposition `cholinv2()`. Sometimes, the inversion of a kinship matrix is problematic because it is singular. It is possible to solve the problem by adding a small digit (e.g. 10E-4) to the diagonal.

Matrix algebra - Exercises



Ex 3

A. It is possible to express a kinship symmetric matrix $K = XX'$ using the eigen decomposition $K = UDU'$. Using the fact that U is an orthogonal matrix with the following property $U' = U^{-1}$, show that $K^{-1} = UD^{-1}U'$.

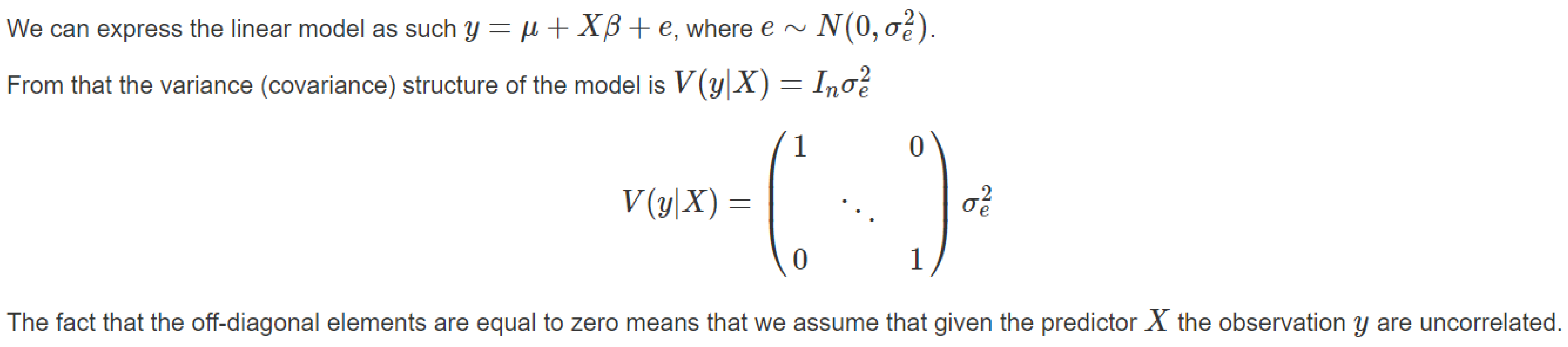
Calculate in R the inverse of the kinship using $K^{-1} = UD^{-1}U'$

B. To speed-up the computation, it is convenient to only select the 'top' principal component that correspond to a certain value (e.g. $\rho = 0.995$). Perform the eigen decomposition on the K matrix and select only the top PC that correspond to 0.995. What is the rank reduction?

C. Calculate a simple G-BLUP model using the `rrBLUP` package and the first phenotype. to compare the estimation of the genetic variance. You can use the following code

Linear models

Linear model



We can express the linear model as such $y = \mu + X\beta + e$, where $e \sim N(0, \sigma_e^2)$.

From that the variance (covariance) structure of the model is $V(y|X) = I_n\sigma_e^2$

$$V(y|X) = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \sigma_e^2$$

The fact that the off-diagonal elements are equal to zero means that we assume that given the predictor X the observation y are uncorrelated.

Linear model – Ordinary least squares

We derive an expression for the β using the ordinary least squared method (OLS).

Let us assume the following model

$$y = \alpha + \beta x + \epsilon$$

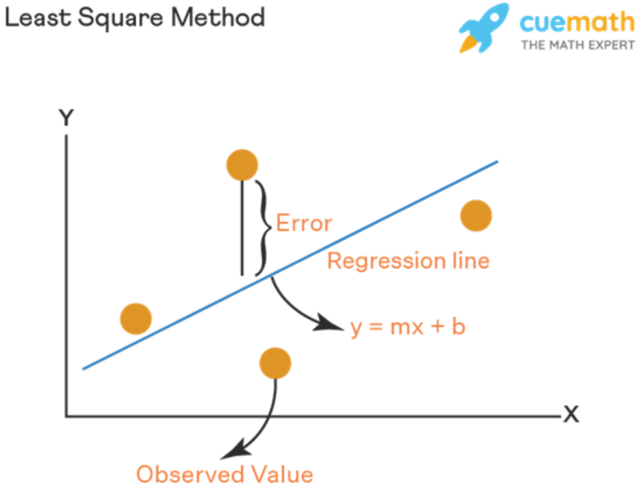
To find the ordinary least square estimates, we want to find the regression line that minimize the error between the observation y and the prediction $\hat{y} = \hat{\alpha} + \hat{\beta} x$

So we want to minimize the error

$$\epsilon = y - \hat{y} = y - \hat{\alpha} - \hat{\beta} x$$

This correspond to the following optimization problem

$$\underset{\hat{\alpha},\hat{\beta}}{\operatorname{argmin}}\, SS_e = \underset{\hat{\alpha},\hat{\beta}}{\operatorname{argmin}} \sum \epsilon^2 = \underset{\hat{\alpha},\hat{\beta}}{\operatorname{argmin}} \sum (y - \hat{\alpha} - \hat{\beta} x)^2$$



Linear model – Ordinary least squares

Before deriving the previous expression, let us introduce the following relationships that will be useful for the demonstration

$$\begin{aligned}
SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) &= \sum_{i=1}^{n}[(x_i - \bar{x})x_i - (x_i - \bar{x})\bar{x}] \\
&= \sum_{i=1}^{n}(x_i - \bar{x})x_i - \sum_{i=1}^{n}(x_i - \bar{x})\bar{x} \\
&= \sum_{i=1}^{n}(x_i - \bar{x})x_i - \bar{x}\sum_{i=1}^{n}(x_i - \bar{x}) \\
&= \sum_{i=1}^{n}(x_i - \bar{x})x_i - \bar{x} * 0 \\
&= \sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i^2 - \bar{x}n * \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2
\end{aligned}$$

Similarly we can show that

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

Linear model – Ordinary least squares

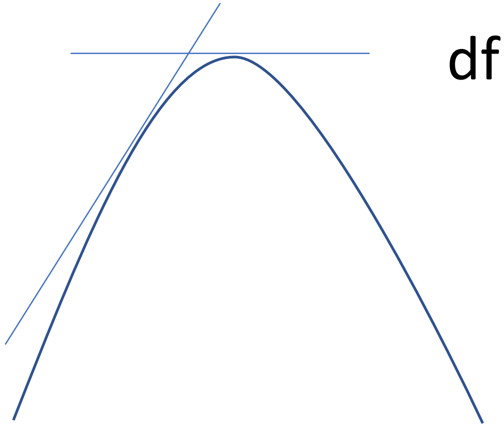
Then we can optimize SS_e with respect to $\hat{\alpha}$.

$$\frac{\partial SS_e}{\partial \hat{\alpha}} = \sum_{i=1}^{n} 2(y_i - \hat{\alpha} - \hat{\beta}x_i) * (-1) = \sum_{i=1}^{n} y_i - n\hat{\alpha} - \hat{\beta}\sum_{i=1}^{n} x_i$$

$$\frac{\partial SS_e}{\partial \hat{\alpha}} = 0 \Longrightarrow \sum_{i=1}^{n} y_i - n\hat{\alpha} - \hat{\beta}\sum_{i=1}^{n} x_i = 0$$

$$n\hat{\alpha} = \sum_{i=1}^{n} y_i - \hat{\beta}\sum_{i=1}^{n} x_i$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$



df(x)/x = 0

Linear model – Ordinary least squares

Then we can derive $\hat{\beta}$

$$
\begin{aligned}
\frac{\partial SS_e}{\partial \hat{\beta}} &= \sum_{i=1}^{n} 2(y_i - \hat{\alpha} - \hat{\beta} x_i) * (-x_i) = -2(\sum_{i=1}^{n} y_i xi - n\hat{\alpha} \sum_{i=1}^{n} x_i - \hat{\beta} \sum_{i=1}^{n} x_i^2) \\
&= -2(\sum_{i=1}^{n} y_i xi - n(\bar{y} - \hat{\beta}\bar{x}) \sum_{i=1}^{n} x_i - \hat{\beta} \sum_{i=1}^{n} x_i^2) \\
&= -2(\sum_{i=1}^{n} y_i xi - n\bar{y} \sum_{i=1}^{n} x_i - n\hat{\beta}\bar{x} \sum_{i=1}^{n} x_i - \hat{\beta} \sum_{i=1}^{n} x_i^2) \\
&= -2(\sum_{i=1}^{n} y_i xi - n\bar{y}\bar{x} - \hat{\beta} * (\sum_{i=1}^{n} x_i^2 - n\bar{x} \sum_{i=1}^{n} x_i)) \\
&= -2(SS_{xy} - \hat{\beta} SS_{xx})
\end{aligned}
$$

$$
\frac{\partial SS_e}{\partial \hat{\beta}} = 0 \Longrightarrow SS_{xy} - \hat{\beta} SS_{xx} = 0
$$

$$
\hat{\beta} = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})} = \frac{\sigma_{xy}}{\sigma_x^2}
$$

Linear model – Ordinary least squares Matrix notation

Let us start from the same linear model written in matrix notation

$y = X\beta + \epsilon$

We want to find the solution $\hat{\beta}$ that minimize $\sum e_i^2 = e'e = (y - X\beta)'(y - X\beta)$

$$\begin{aligned} e'e &= (y - X\beta)'(y - X\beta) \\ &= (y' - \beta'X')(y - X\beta) \\ &= y'y - y'X\beta - \beta'X'y + \beta'X'X\beta \\ &= y'y - 2\beta'X'y + \beta'X'X\beta \end{aligned}$$

In the last line we use the fact that $y'X\beta$ is a scalar because $y'_{1\times n}X_{n\times p}\beta_{p\times 1}$ and the property that the transpose of a scalar is the scalar itself so $y'X\beta = (y'X\beta)' = \beta'X'y$

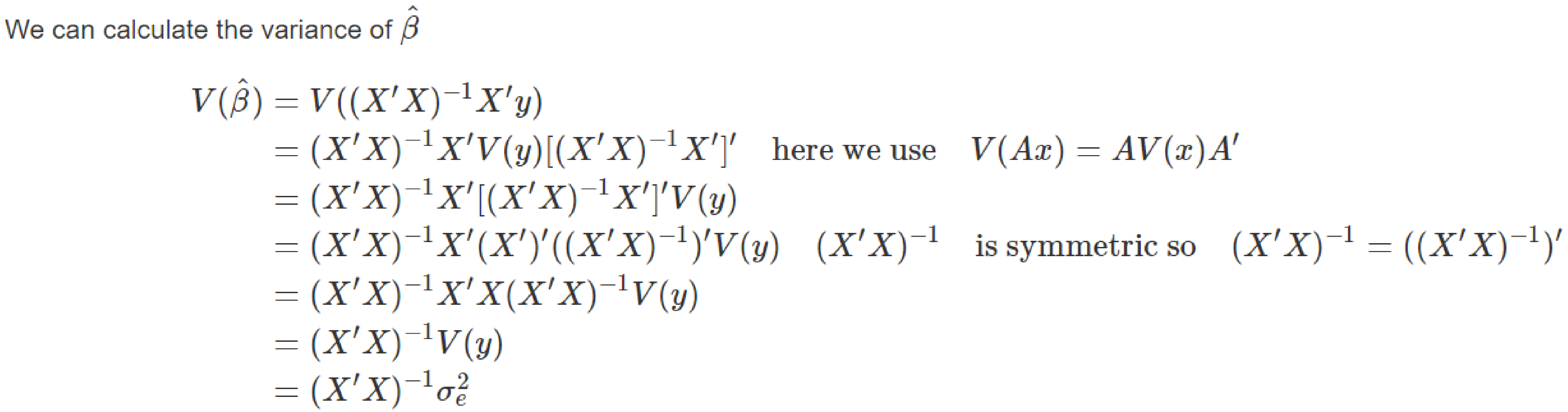
We can derive with respect to β and equal to zero

$$\frac{\partial e'e}{\partial \beta} = 0 - 2X'y + 2X'X\beta = 0 \Longrightarrow X'X\beta = X'y$$

Which give the following solution

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

Linear model – Ordinary least squares Matrix notation



We can calculate the variance of $\hat{\beta}$

$$\begin{aligned}
V(\hat{\beta}) &= V((X'X)^{-1}X'y) \\
&= (X'X)^{-1}X'V(y)[(X'X)^{-1}X']' \quad \text{here we use} \quad V(Ax) = AV(x)A' \\
&= (X'X)^{-1}X'[(X'X)^{-1}X']'V(y) \\
&= (X'X)^{-1}X'(X')'((X'X)^{-1})'V(y) \quad (X'X)^{-1} \quad \text{is symmetric so} \quad (X'X)^{-1} = ((X'X)^{-1})' \\
&= (X'X)^{-1}X'X(X'X)^{-1}V(y) \\
&= (X'X)^{-1}V(y) \\
&= (X'X)^{-1}\sigma_e^2
\end{aligned}$$

Maximum likelihood principle

$$P(X = x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{1}{2} [\left(\frac{x \ - \mu}{\sigma}\right)^2]$$

$$P(X = x_1) = 0{,}5$$
$$P(X = x_2 \ ) = 0{,}5$$
$$P(X = x_1 \ \cap X = x_2) = 0{,}5 * 0{,}5$$

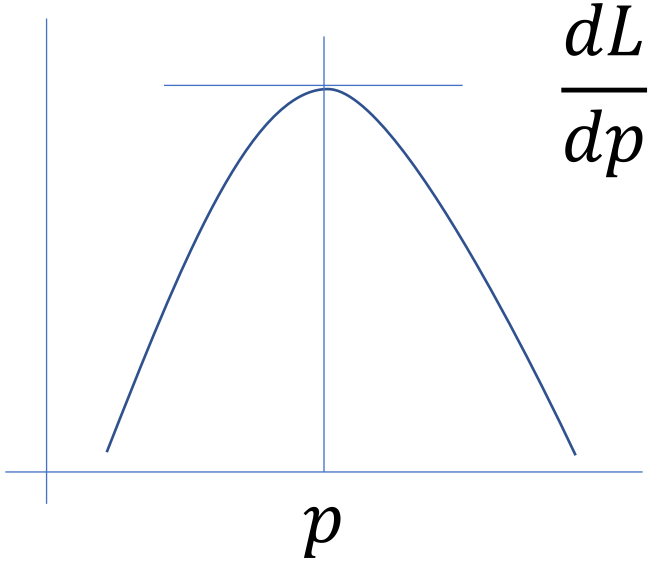
Likelihood function

$$L = P(X = x_1 \ \cap X = x_2 \cap \cdots \cap X = x_n) = \prod_{i=1}^{n} P(X = x_i|\mu, \sigma^2)$$

Maximum likelihood principle

Likelihood function

$$L = P(X = x_1 \cap X = x_2 \cap \cdots \cap X = x_n) = \prod_{i=1}^{n} P(X = x_i | \mu, \sigma^2)$$

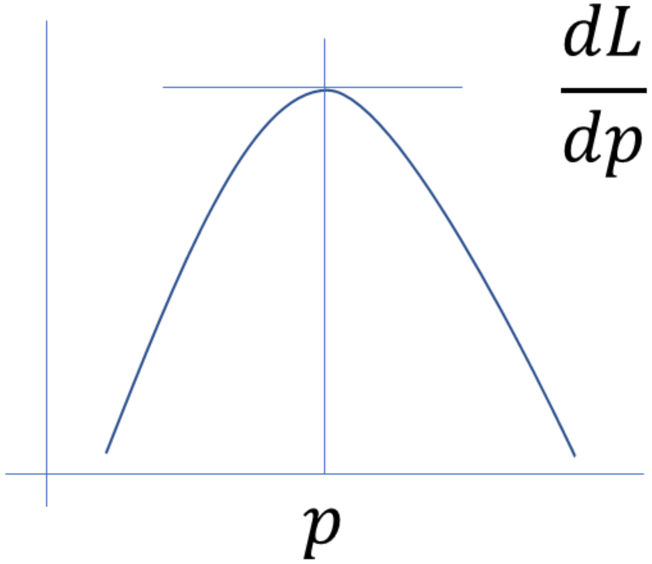


$$\frac{dL}{dp} = 0$$

Maximum likelihood principle -illustration

Likelihood function

$$P(X = 19 - 0 | p = football\ match)$$



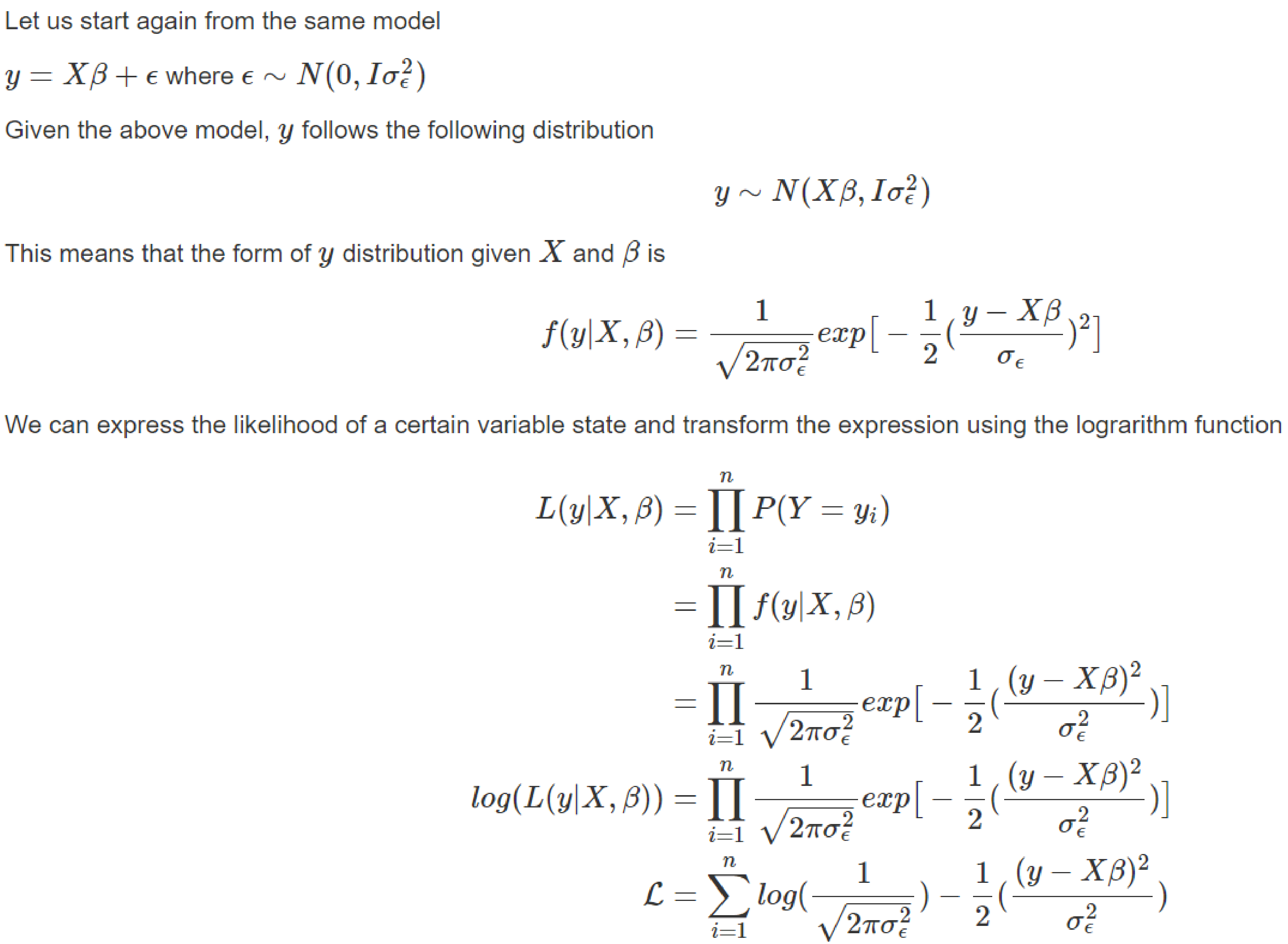
$$\frac{dL}{dp} = 0$$

$$p = PSG\ - Marseille$$

$$p = PSG\ - \text{Montceau–Les–Mines}$$

p= parameter value that maximise the Likelihood given the assumed distribution

Linear model – Maximum likelihood estimation



Let us start again from the same model

$y = X\beta + \epsilon$ where $\epsilon \sim N(0, I\sigma_\epsilon^2)$

Given the above model, y follows the following distribution

$$y \sim N(X\beta, I\sigma_\epsilon^2)$$

This means that the form of y distribution given X and β is

$$f(y|X,\beta) = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} exp[-\frac{1}{2}(\frac{y - X\beta}{\sigma_\epsilon})^2]$$

We can express the likelihood of a certain variable state and transform the expression using the lograrithm function

$$\begin{aligned} L(y|X,\beta) &= \prod_{i=1}^{n} P(Y = y_i) \\ &= \prod_{i=1}^{n} f(y|X,\beta) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} exp[-\frac{1}{2}(\frac{(y - X\beta)^2}{\sigma_\epsilon^2})] \\ log(L(y|X,\beta)) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} exp[-\frac{1}{2}(\frac{(y - X\beta)^2}{\sigma_\epsilon^2})] \\ \mathcal{L} &= \sum_{i=1}^{n} log(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}) - \frac{1}{2}(\frac{(y - X\beta)^2}{\sigma_\epsilon^2}) \end{aligned}$$

Linear model – Maximum likelihood estimation

To obtain the expression for β we can derive $\mathcal{L}$ with respect to β and find an optimum

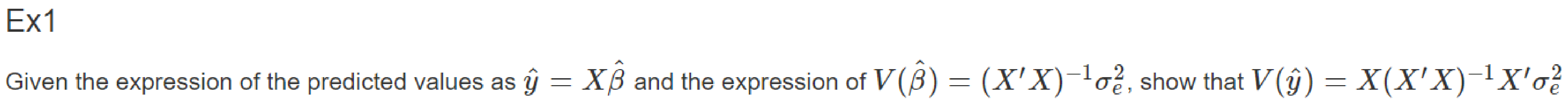
$$\begin{aligned}
\frac{\mathcal{L}}{\partial\beta} &= \sum_{i=1}^{n} 0 - \frac{1}{2}\frac{\partial}{\partial\beta}\frac{(y - X\beta)^2}{\sigma_\epsilon^2} \\
&= -\frac{1}{2}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\frac{y^2 - 2X\beta y + X^2\beta^2}{\sigma_\epsilon^2} \\
&= -\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma_\epsilon^2} - 2X'y + 2\beta X'X
\end{aligned}$$

$$\frac{\mathcal{L}}{\partial\beta} = 0 \implies \sum_{i=1}^{n} -2X'y + 2\beta X'X = 0$$

$$\beta X'X - X'y = 0 \implies \beta_{ML} = (X'X)^{-1}X'y$$

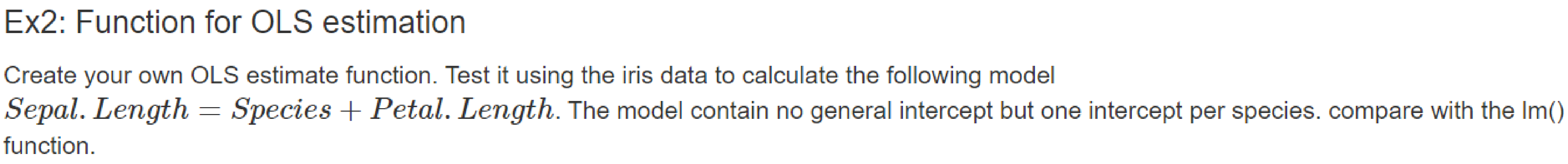
We can notice that $\beta_{ML} = \beta_{OLS}$

Linear models - Exercises



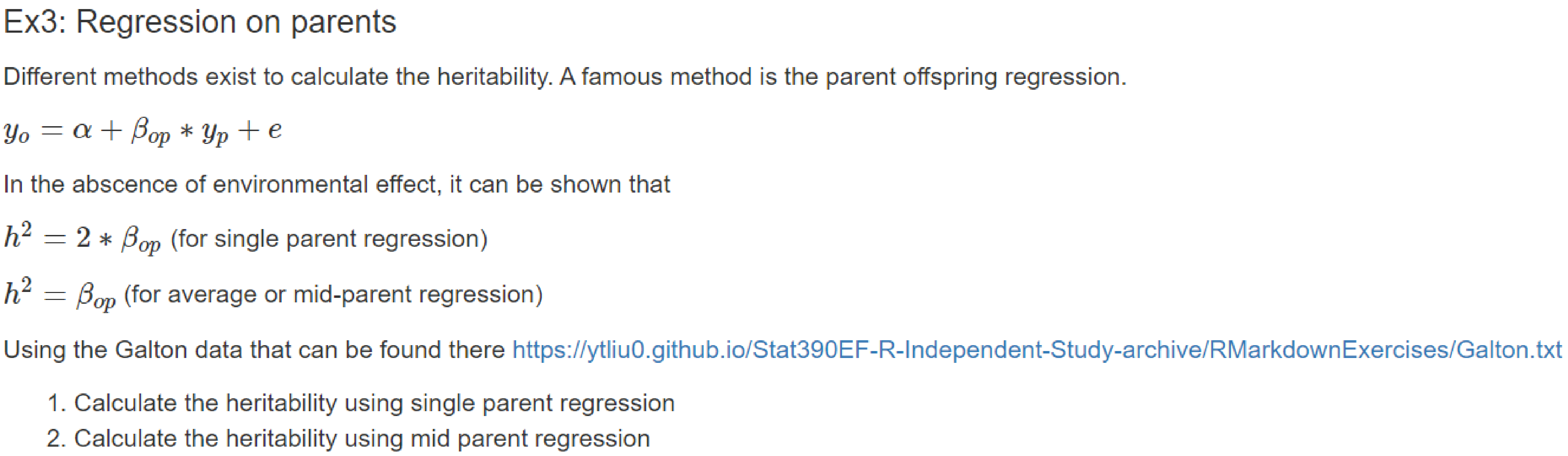
Ex1

Given the expression of the predicted values as $\hat{y} = X\hat{\beta}$ and the expression of $V(\hat{\beta}) = (X'X)^{-1}\sigma_e^2$, show that $V(\hat{y}) = X(X'X)^{-1}X'\sigma_e^2$



Ex2: Function for OLS estimation

Create your own OLS estimate function. Test it using the iris data to calculate the following model $Sepal.Length = Species + Petal.Length$. The model contain no general intercept but one intercept per species. compare with the lm() function.



Ex3: Regression on parents

Different methods exist to calculate the heritability. A famous method is the parent offspring regression.

$y_o = \alpha + \beta_{op} * y_p + e$

In the abscence of environmental effect, it can be shown that

$h^2 = 2 * \beta_{op}$ (for single parent regression)

$h^2 = \beta_{op}$ (for average or mid-parent regression)

Using the Galton data that can be found there https://ytliu0.github.io/Stat390EF-R-Independent-Study-archive/RMarkdownExercises/Galton.txt

1. Calculate the heritability using single parent regression
2. Calculate the heritability using mid parent regression

Recap

Probability (P space, pdf, E, V)

Linear mixed model

Matrix (+-*Inv, eigen, Cholesky)

Linear model

Exercises

Probabilité

Ex: 1

Matrix:

Ex: 2

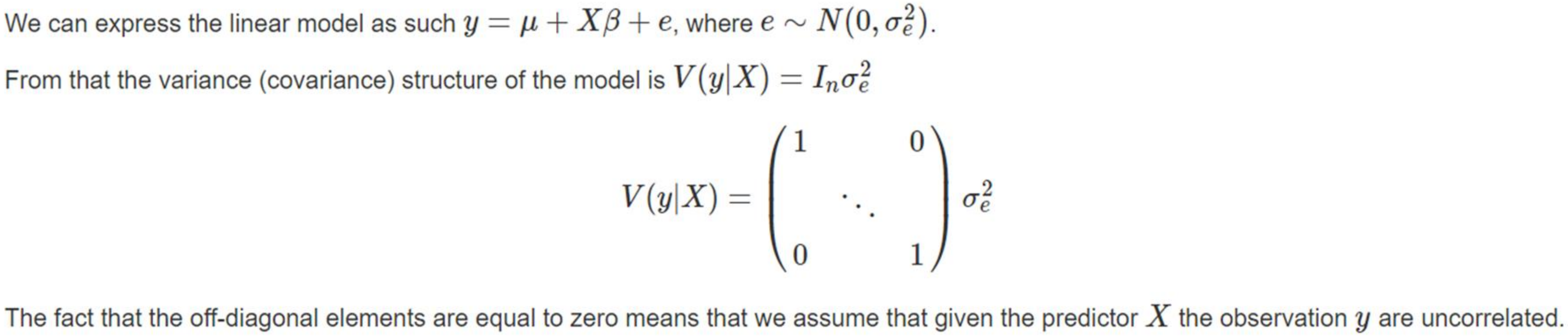
Ex 3: très motivé

Model linéaire

Ex 1: formule

Genearlized Least squares

Genarlized least square model

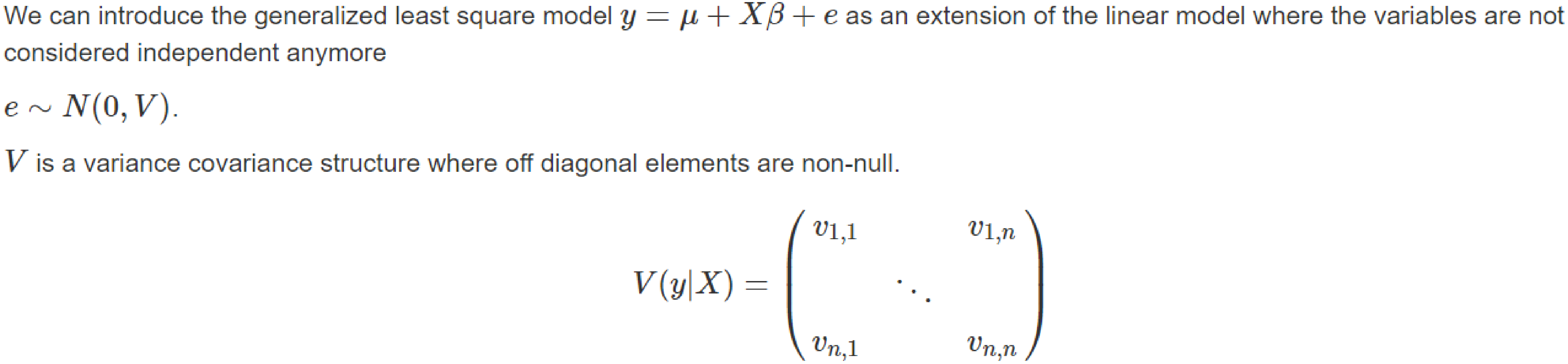


We can express the linear model as such $y = \mu + X\beta + e$, where $e \sim N(0, \sigma_e^2)$.

From that the variance (covariance) structure of the model is $V(y|X) = I_n \sigma_e^2$

$$V(y|X) = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \sigma_e^2$$

The fact that the off-diagonal elements are equal to zero means that we assume that given the predictor X the observation y are uncorrelated.



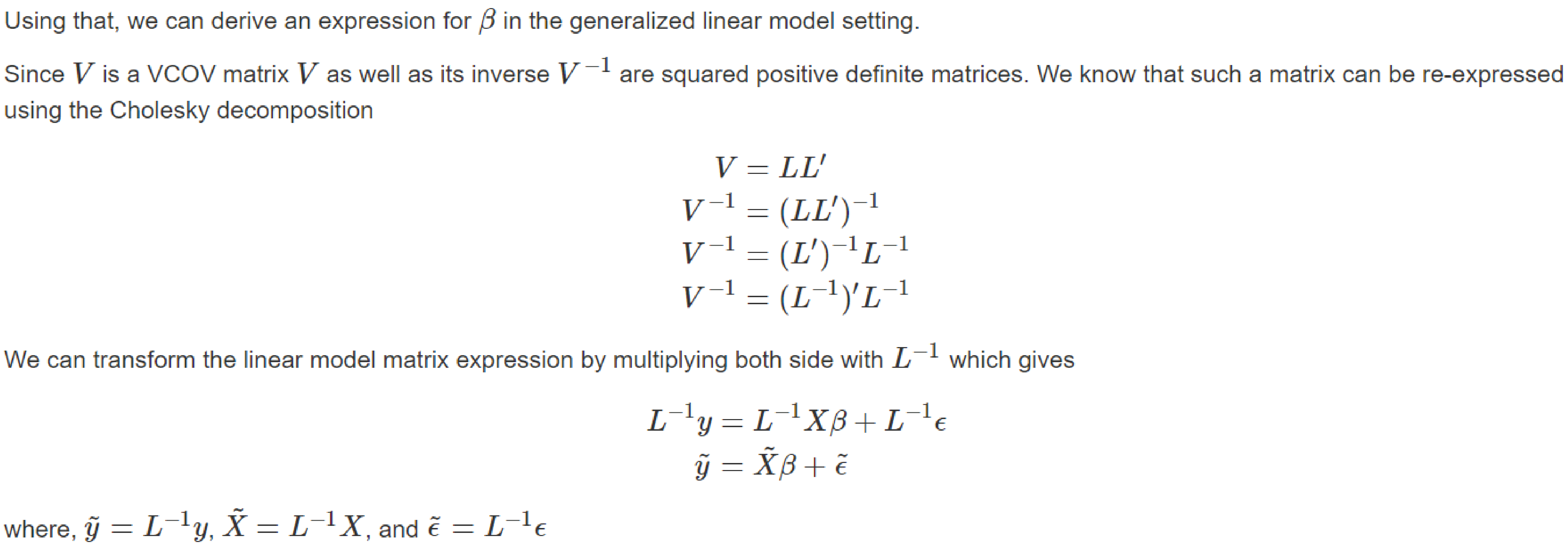
We can introduce the generalized least square model $y = \mu + X\beta + e$ as an extension of the linear model where the variables are not considered independent anymore

$e \sim N(0, V)$.

V is a variance covariance structure where off diagonal elements are non-null.

$$V(y|X) = \begin{pmatrix} v_{1,1} & & v_{1,n} \\ & \ddots & \\ v_{n,1} & & v_{n,n} \end{pmatrix}$$

Genarlized least square model



Using that, we can derive an expression for β in the generalized linear model setting.

Since V is a VCOV matrix V as well as its inverse V^{-1} are squared positive definite matrices. We know that such a matrix can be re-expressed using the Cholesky decomposition

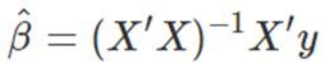
$$\begin{aligned} V &= LL' \\ V^{-1} &= (LL')^{-1} \\ V^{-1} &= (L')^{-1}L^{-1} \\ V^{-1} &= (L^{-1})'L^{-1} \end{aligned}$$

We can transform the linear model matrix expression by multiplying both side with L^{-1} which gives

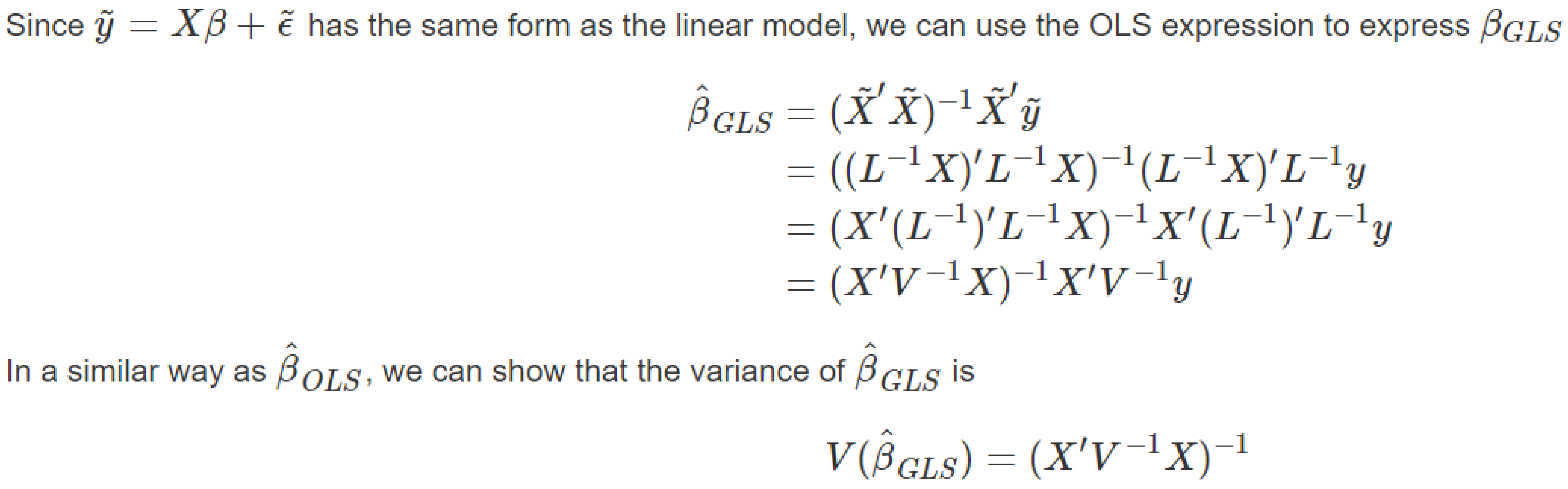
$$\begin{aligned} L^{-1}y &= L^{-1}X\beta + L^{-1}\epsilon \\ \tilde{y} &= \tilde{X}\beta + \tilde{\epsilon} \end{aligned}$$

where, $\tilde{y} = L^{-1}y$, $\tilde{X} = L^{-1}X$, and $\tilde{\epsilon} = L^{-1}\epsilon$

Genarlized least square model



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



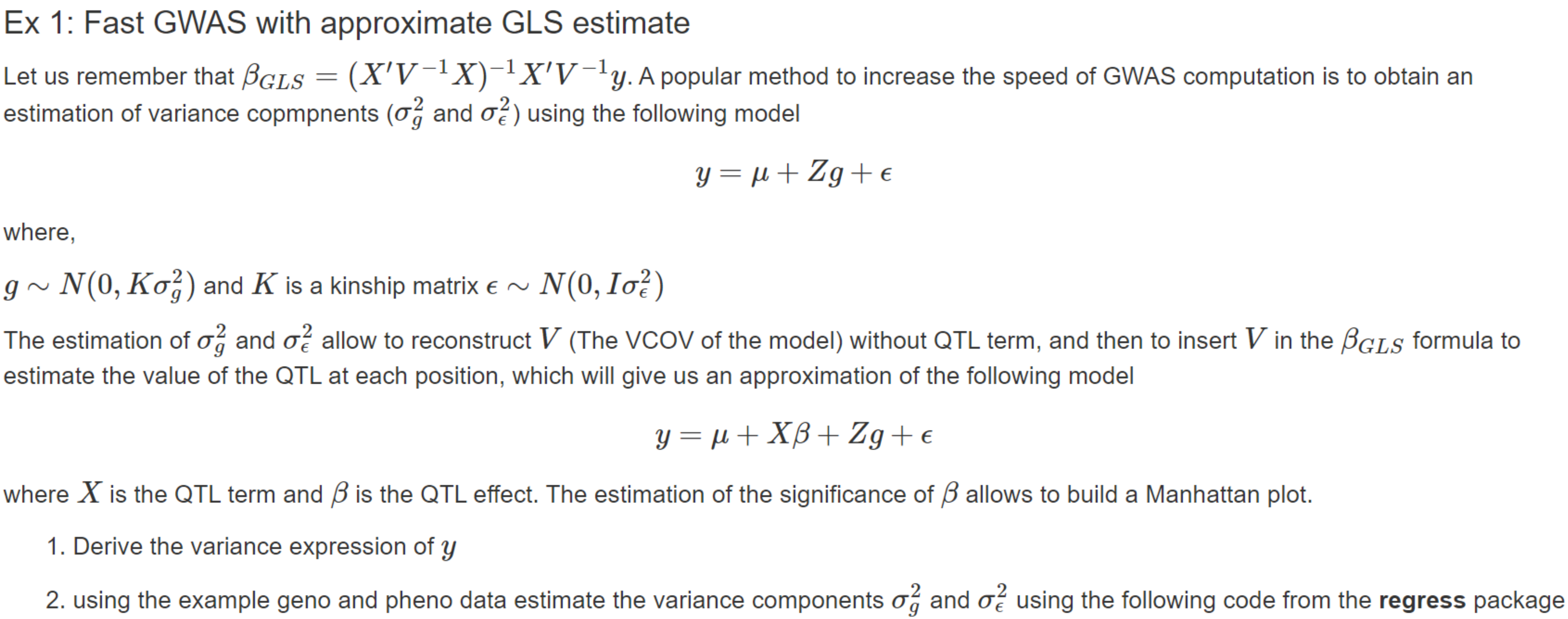
Since $\tilde{y} = X\beta + \tilde{\epsilon}$ has the same form as the linear model, we can use the OLS expression to express β_{GLS}

$$\begin{aligned}
\hat{\beta}_{GLS} &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} \\
&= ((L^{-1}X)'L^{-1}X)^{-1}(L^{-1}X)'L^{-1}y \\
&= (X'(L^{-1})'L^{-1}X)^{-1}X'(L^{-1})'L^{-1}y \\
&= (X'V^{-1}X)^{-1}X'V^{-1}y
\end{aligned}$$

In a similar way as $\hat{\beta}_{OLS}$, we can show that the variance of $\hat{\beta}_{GLS}$ is

$$V(\hat{\beta}_{GLS}) = (X'V^{-1}X)^{-1}$$

Exercise



Ex 1: Fast GWAS with approximate GLS estimate

Let us remember that $\beta_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}y$. A popular method to increase the speed of GWAS computation is to obtain an estimation of variance copmpnents (σ_g^2 and σ_ϵ^2) using the following model

$$y = \mu + Zg + \epsilon$$

where,

$g \sim N(0, K\sigma_g^2)$ and K is a kinship matrix $\epsilon \sim N(0, I\sigma_\epsilon^2)$

The estimation of σ_g^2 and σ_ϵ^2 allow to reconstruct V (The VCOV of the model) without QTL term, and then to insert V in the β_{GLS} formula to estimate the value of the QTL at each position, which will give us an approximation of the following model

$$y = \mu + X\beta + Zg + \epsilon$$

where X is the QTL term and β is the QTL effect. The estimation of the significance of β allows to build a Manhattan plot.

1. Derive the variance expression of y
2. using the example geno and pheno data estimate the variance components σ_g^2 and σ_ϵ^2 using the following code from the **regress** package

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

Since $\tilde{y} = X\beta + \tilde{\epsilon}$ has the same form as the linear model, we can use the OLS expression to express β_{GLS}

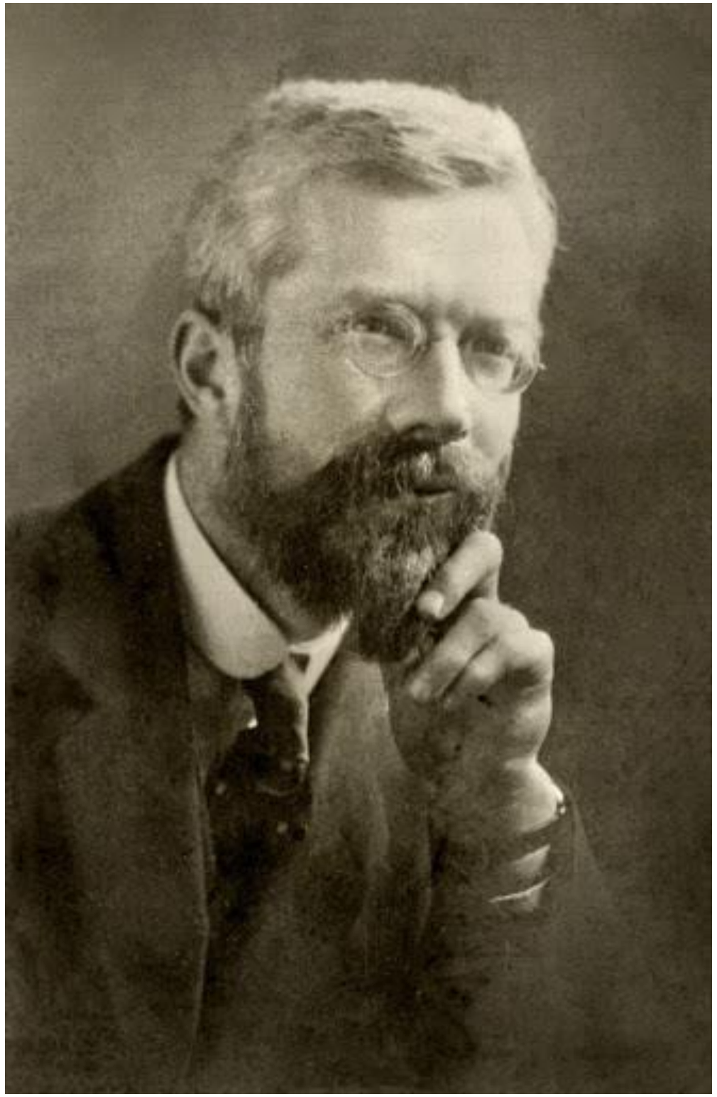
$$\begin{aligned}
\hat{\beta}_{GLS} &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} \\
&= ((L^{-1}X)'L^{-1}X)^{-1}(L^{-1}X)'L^{-1}y \\
&= (X'(L^{-1})'L^{-1}X)^{-1}X'(L^{-1})'L^{-1}y \\
&= (X'V^{-1}X)^{-1}X'V^{-1}y
\end{aligned}$$

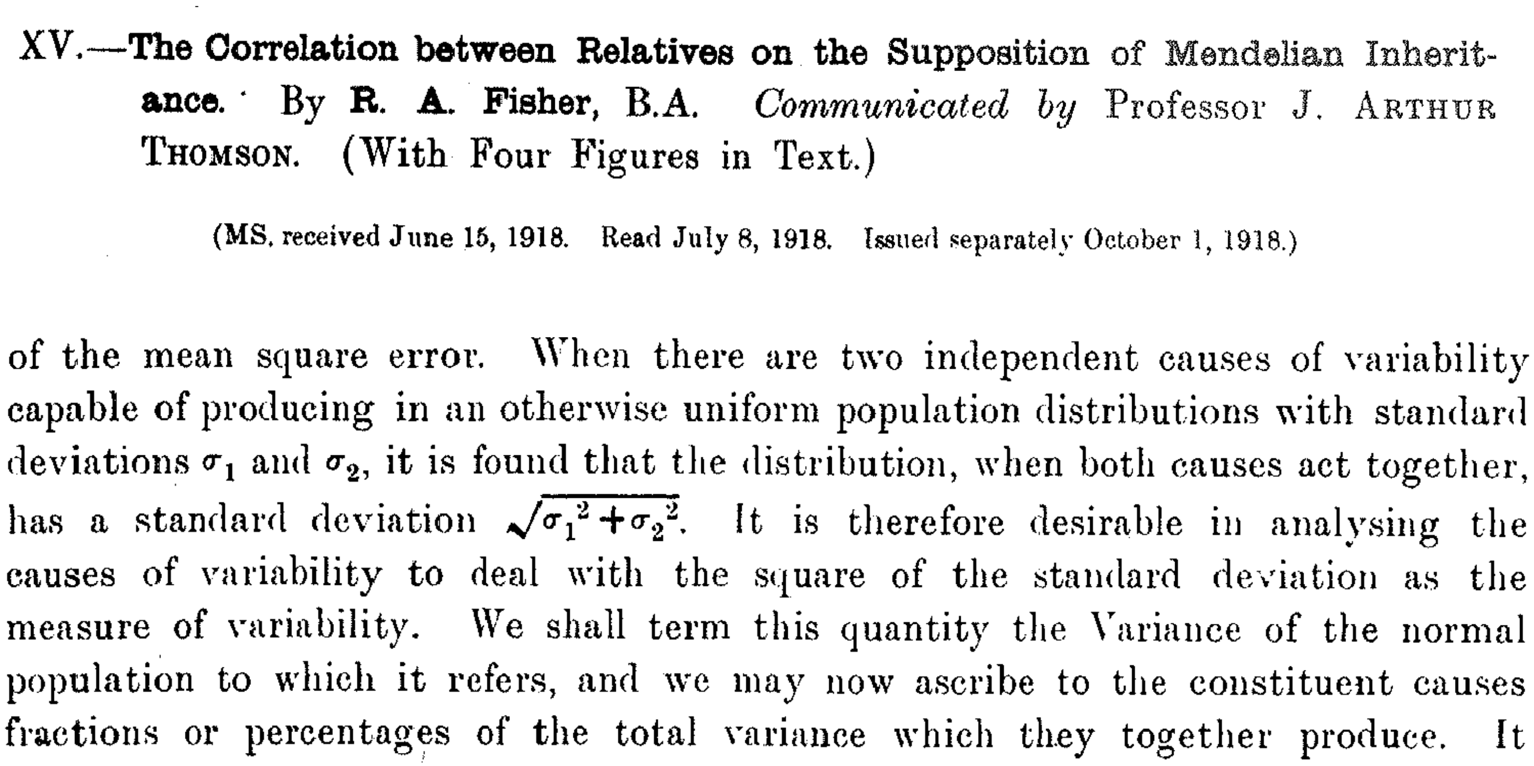
In a similar way as $\hat{\beta}_{OLS}$, we can show that the variance of $\hat{\beta}_{GLS}$ is

$$V(\hat{\beta}_{GLS}) = (X'V^{-1}X)^{-1}$$

Mixed models

Introduction





XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. By **R. A. Fisher**, B.A. *Communicated by* Professor J. ARTHUR THOMSON. (With Four Figures in Text.)

(MS. received June 15, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

of the mean square error. When there are two independent causes of variability capable of producing in an otherwise uniform population distributions with standard deviations σ_1 and σ_2, it is found that the distribution, when both causes act together, has a standard deviation $\sqrt{\sigma_1{}^2+\sigma_2{}^2}$. It is therefore desirable in analysing the causes of variability to deal with the square of the standard deviation as the measure of variability. We shall term this quantity the Variance of the normal population to which it refers, and we may now ascribe to the constituent causes fractions or percentages of the total variance which they together produce. It

Concept of analysis of variance (ANOVA)

Partition of the variance into different term: genetic, environmental, GxE

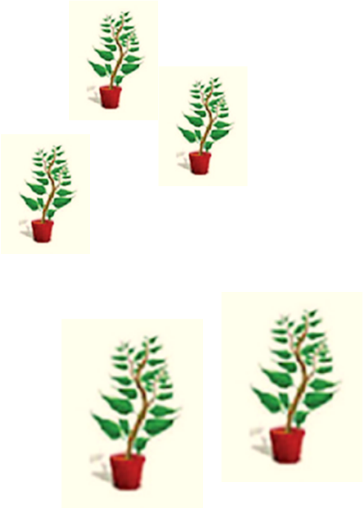
Seminal thoughts about the concept of mixed model: modelling second moment (variance) of the normal distribution.

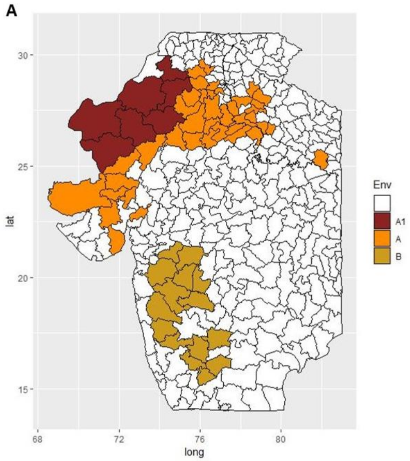
Mixed models justification

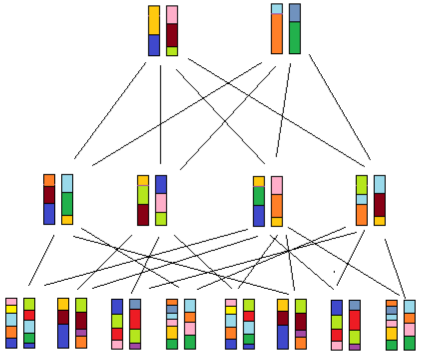
Take the dependency between observations into consideration

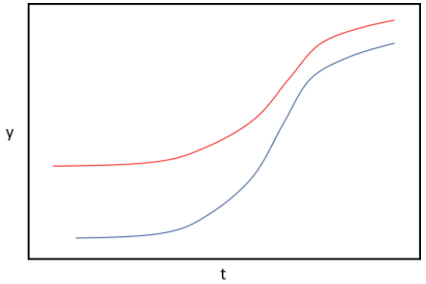
$$V(y|X) = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \qquad V(y|X) = \begin{pmatrix} v_{1,1} & & v_{1,n} \\ & \ddots & \\ v_{n,1} & & v_{n,n} \end{pmatrix}$$

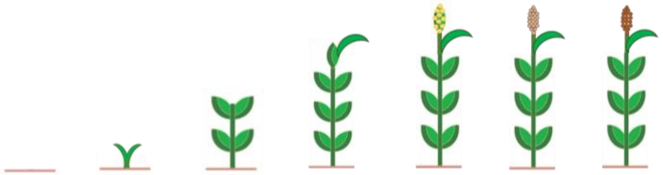
Multiple source of dependency between observations: Shared environment/conditions, time (longitudinal data), (genetic) relatedness, etc.





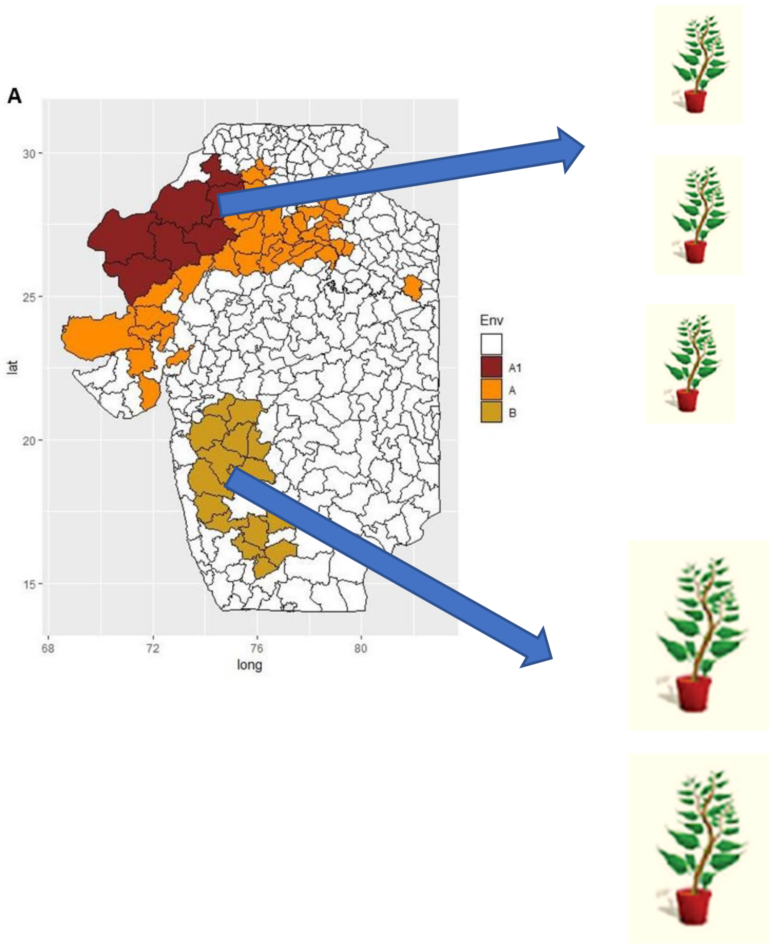


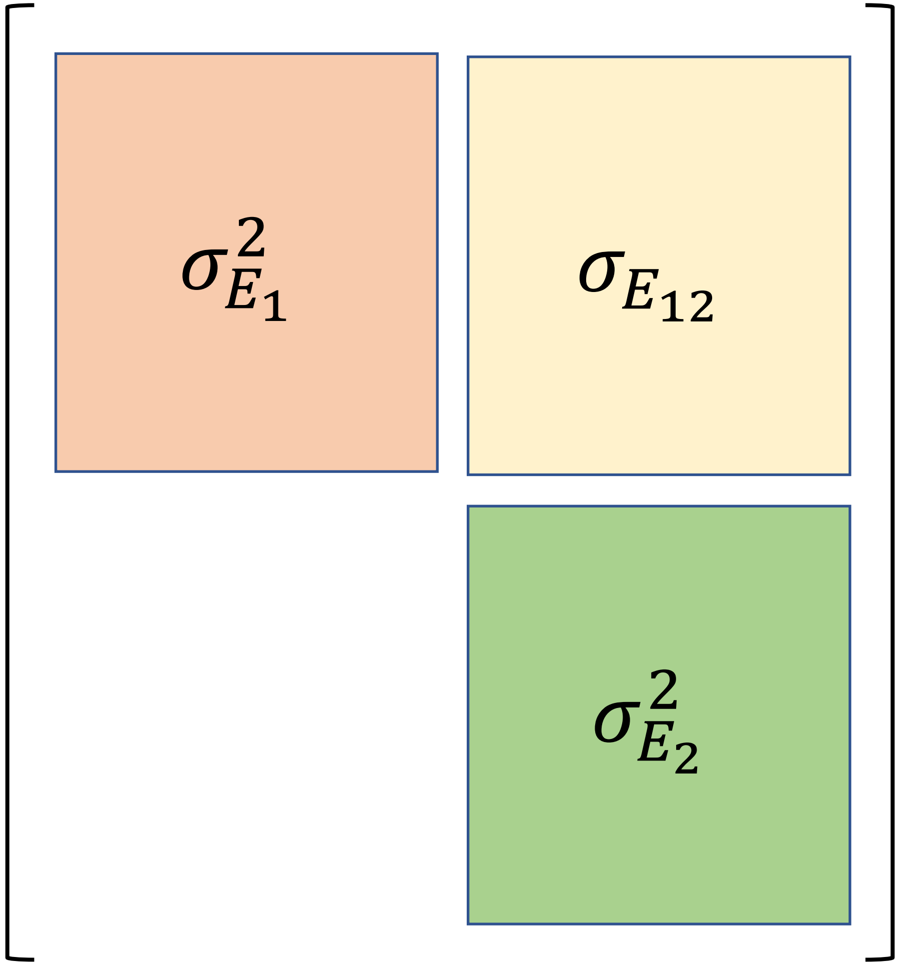




Mixed models justification

Take the dependency between observations into consideration





$$\begin{bmatrix} \sigma^2_{E_1} & \sigma_{E_{12}} \\ & \sigma^2_{E_2} \end{bmatrix}$$

Mixed models justification

Advantages

- Reduce the residual variability (error)
- Important to form test statistic

$$V(y|X) = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$$

$$V(y|X) = \begin{pmatrix} v_{1,1} & & v_{1,n} \\ & \ddots & \\ v_{n,1} & & v_{n,n} \end{pmatrix}$$

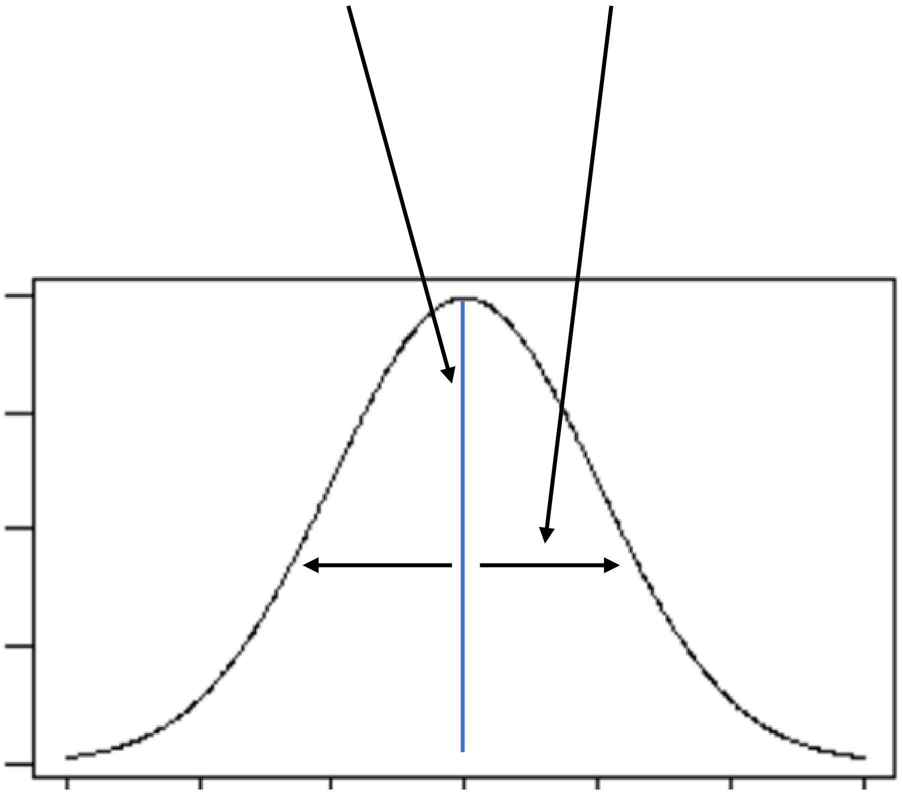
$$\frac{\hat{\beta}}{\sqrt{V(\hat{\beta})}} = \frac{(X'X)^{-1}X'y}{\sqrt{(X'I\sigma_e^2X)^{-1}}} \sim t$$

$$\frac{\hat{\beta}^2}{V(\hat{\beta})} = \frac{(X'V^{-1}X)^{-1}X'V^{-1}y}{(X'V^{-1}X)^{-1}} \sim \chi^2$$

Mixed models justification

Model the two moments of the distribution (expectation and variance)

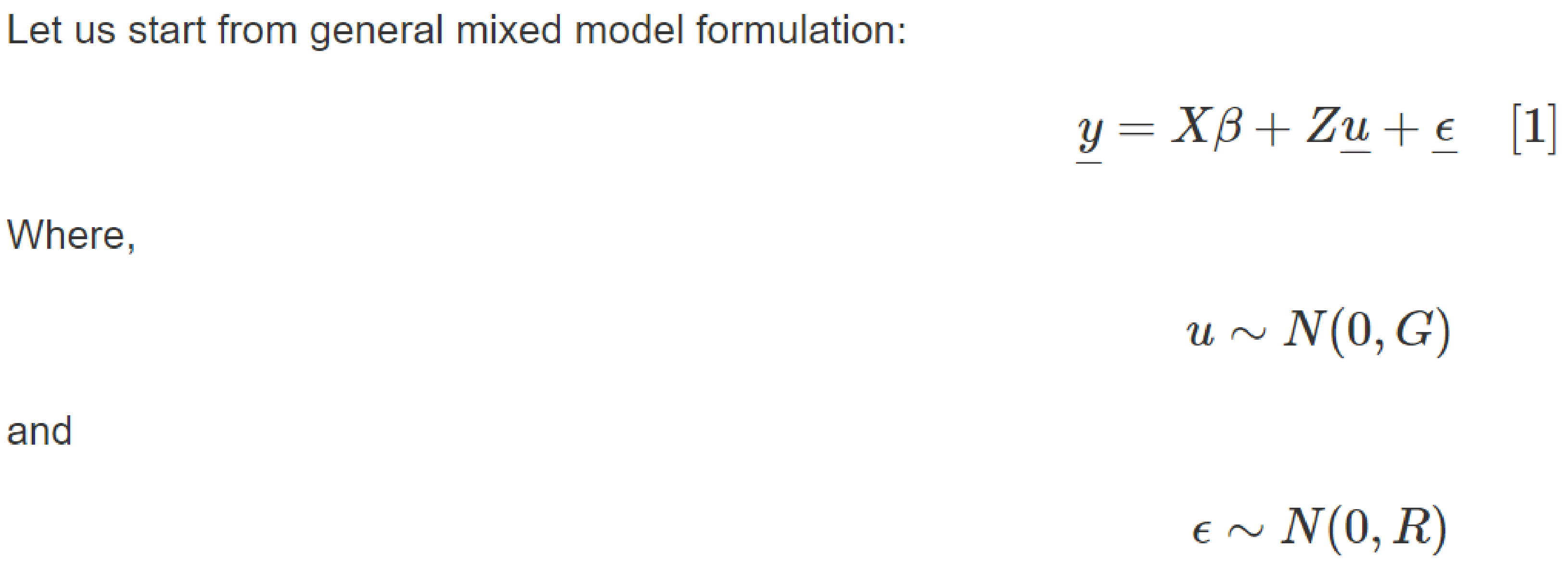
$$\underline{y} = X\beta + Z\underline{u} + \underline{e}$$



Fixed terms (β): fixed constant affect the first moment of the distribution (Expectation, central tendency)

Random terms (u): term defined by a distribution (e.g. $u \sim N(0, \sigma_u^2)$) affect the second moment of the distribution (Variation, dispersion)

Mixed model general formulation



Let us start from general mixed model formulation:

$$\underline{y} = X\beta + Z\underline{u} + \underline{\epsilon} \quad [1]$$

Where,

$$u \sim N(0, G)$$

and

$$\epsilon \sim N(0, R)$$

$u \perp \epsilon$: The two random terms are independent (cov(u, e) = 0)

Mixed model general formulation

$$\underline{y} = X\beta + Z\underline{u} + \underline{\epsilon} \quad [1]$$

$$u \sim N(0, G)$$

$$\epsilon \sim N(0, R)$$

Properties of y distribution

$$\begin{aligned} E[y] &= E[X\beta] + E[Zu] + E[\epsilon] \\ &= X\beta + ZE[u] + 0 \\ &= X\beta + Z * 0 + 0 \\ &= X\beta \end{aligned}$$

$$\begin{aligned} V[y] &= V[X\beta] + V[Zu] + V[\epsilon] \\ &= 0 + ZV[u]Z' + R \\ &= 0 + ZGZ' + R \\ V &= ZGZ' + R \end{aligned}$$

Hence,

$$y \sim MVN(X\beta, V)$$

Multivariate normal distribution - Intermezzo

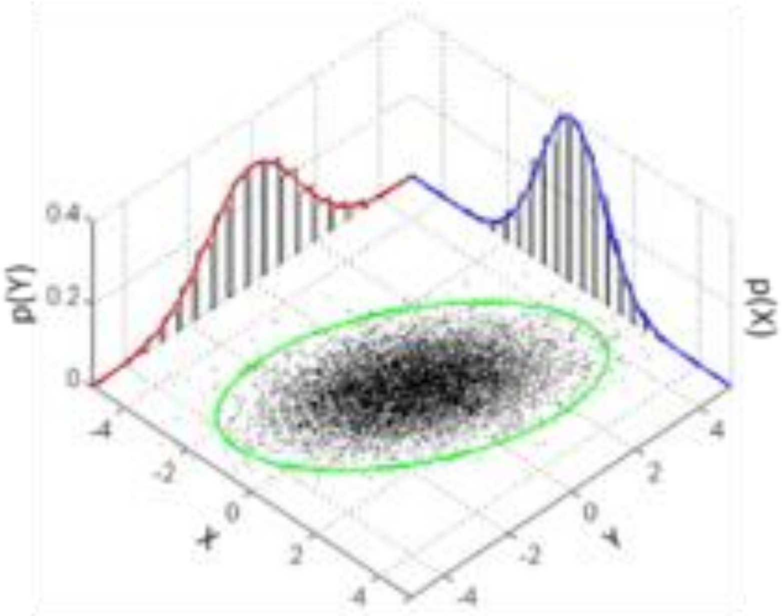
Let us start to introduce the multivariate normal distribution (MVN). Let us assume a random vector $(x_1, x_2, \ldots, x_n)$ of independent variables that are univariate normal. For example for two random variables

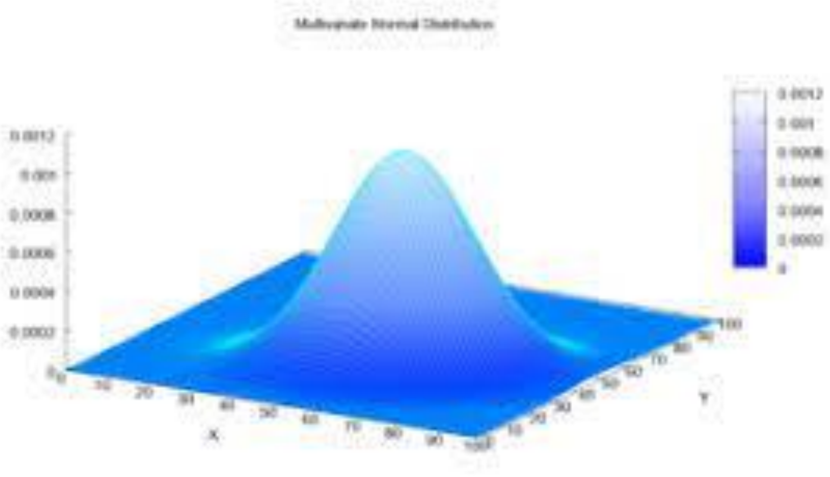
$$x_1 \sim N(0, \sigma_1^2)$$

$$x_2 \sim N(0, \sigma_2^2)$$

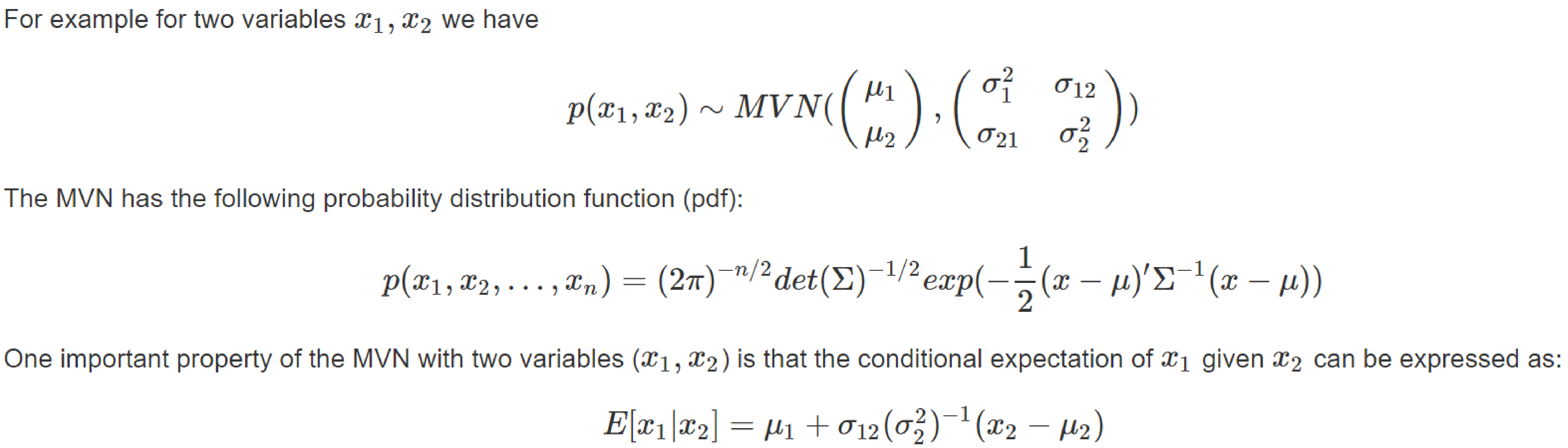
The MVN describe the joint distribution of those variables: $MVN(\mu, \Sigma)$ where μ is the vector of means $(\mu_1, \mu_2, \ldots, \mu_n)$ and Σ is the variance covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & & \sigma_{1n} \\ \sigma_{21} & & & \\ \vdots & & \ddots & \vdots \\ \sigma_{n1} & & & \sigma_n^2 \end{pmatrix}$$





Multivariate normal distribution - Intermezzo



For example for two variables x_1, x_2 we have

$$p(x_1, x_2) \sim MVN(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix})$$

The MVN has the following probability distribution function (pdf):

$$p(x_1, x_2, \ldots, x_n) = (2\pi)^{-n/2} det(\Sigma)^{-1/2} exp(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu))$$

One important property of the MVN with two variables (x_1, x_2) is that the conditional expectation of x_1 given x_2 can be expressed as:

$$E[x_1|x_2] = \mu_1 + \sigma_{12}(\sigma_2^2)^{-1}(x_2 - \mu_2)$$

Mixed models – BLUE estimation

Pdf: $$p(x_1, x_2, \ldots, x_n) = (2\pi)^{-n/2} det(\Sigma)^{-1/2} exp(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu))$$

We can find an expression for the fixed term coefficient β by optimizing the likelihood function with respect to β. Let us start from the pdf

$$f(y|X, \beta, V) = (2\pi)^{-n/2} |V|^{-1/2} exp\Big[-\frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta)\Big]$$

We can express the likelihood of a certain variable state and transform the expression using the lograrithm function

$$\begin{aligned} L(y|X, \beta) &= \prod_{i=1}^{n} P(Y = y_i) \\ &= \prod_{i=1}^{n} f(y|X, \beta, V) \\ &= \prod_{i=1}^{n} (2\pi)^{-n/2} |V|^{-1/2} exp\Big[-\frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta)\Big] \end{aligned}$$

$$log(L(y|X, \beta)) = \mathcal{L} \propto \sum_{i=1}^{n} log(C) - \frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta)$$

Mixed models – BLUE estimation

To obtain the expression for β we can derive $\mathcal{L}$ with respect to β and find an optimum

$$\frac{\mathcal{L}}{\partial\beta} = \sum_{i=1}^{n} 0 - \frac{1}{2}\frac{\partial}{\partial\beta}(y - X\beta)'V^{-1}(y - X\beta)$$

$$= -\frac{1}{2}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}y'V^{-1}y - 2X\beta'V^{-1}y + \beta'X'V^{-1}X\beta$$

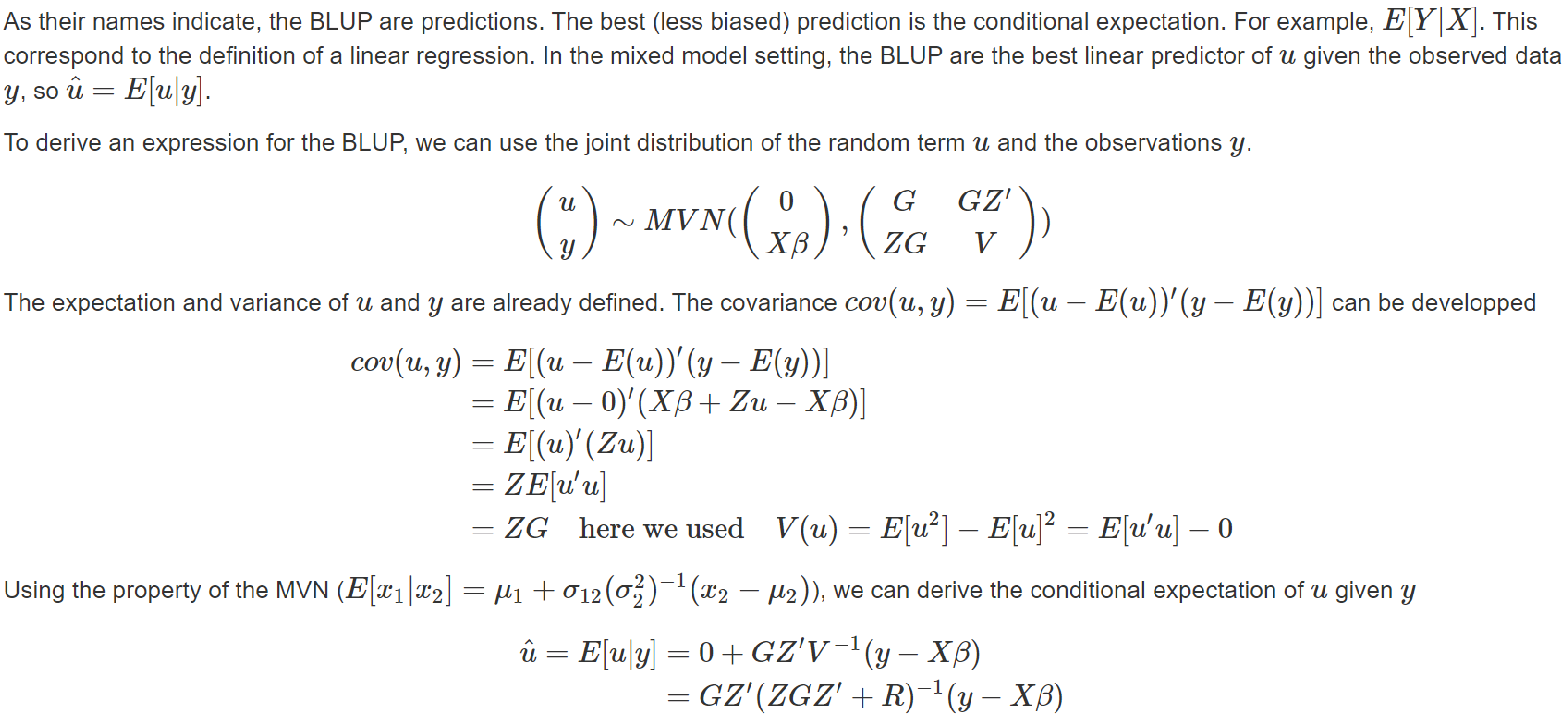
$$= -\frac{1}{2}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}0 - 2X'V^{-1}y + 2\beta X'V^{-1}X$$

$$\frac{\mathcal{L}}{\partial\beta} = 0 \Longrightarrow \sum_{i=1}^{n} -2X'V^{-1}y + 2\beta X'V^{-1}X = 0$$

$$\beta X'V^{-1}X - X'V^{-1}y = 0 \Longrightarrow \beta_{MM} = (X'V^{-1}X)^{-1}X'V^{-1}y$$

We can notice that $\beta_{MM} = \beta_{GLS}$

Mixed models – BLUP prediction



As their names indicate, the BLUP are predictions. The best (less biased) prediction is the conditional expectation. For example, $E[Y|X]$. This correspond to the definition of a linear regression. In the mixed model setting, the BLUP are the best linear predictor of u given the observed data y, so $\hat{u} = E[u|y]$.

To derive an expression for the BLUP, we can use the joint distribution of the random term u and the observations y.

$$\begin{pmatrix} u \\ y \end{pmatrix} \sim MVN(\begin{pmatrix} 0 \\ X\beta \end{pmatrix}, \begin{pmatrix} G & GZ' \\ ZG & V \end{pmatrix})$$

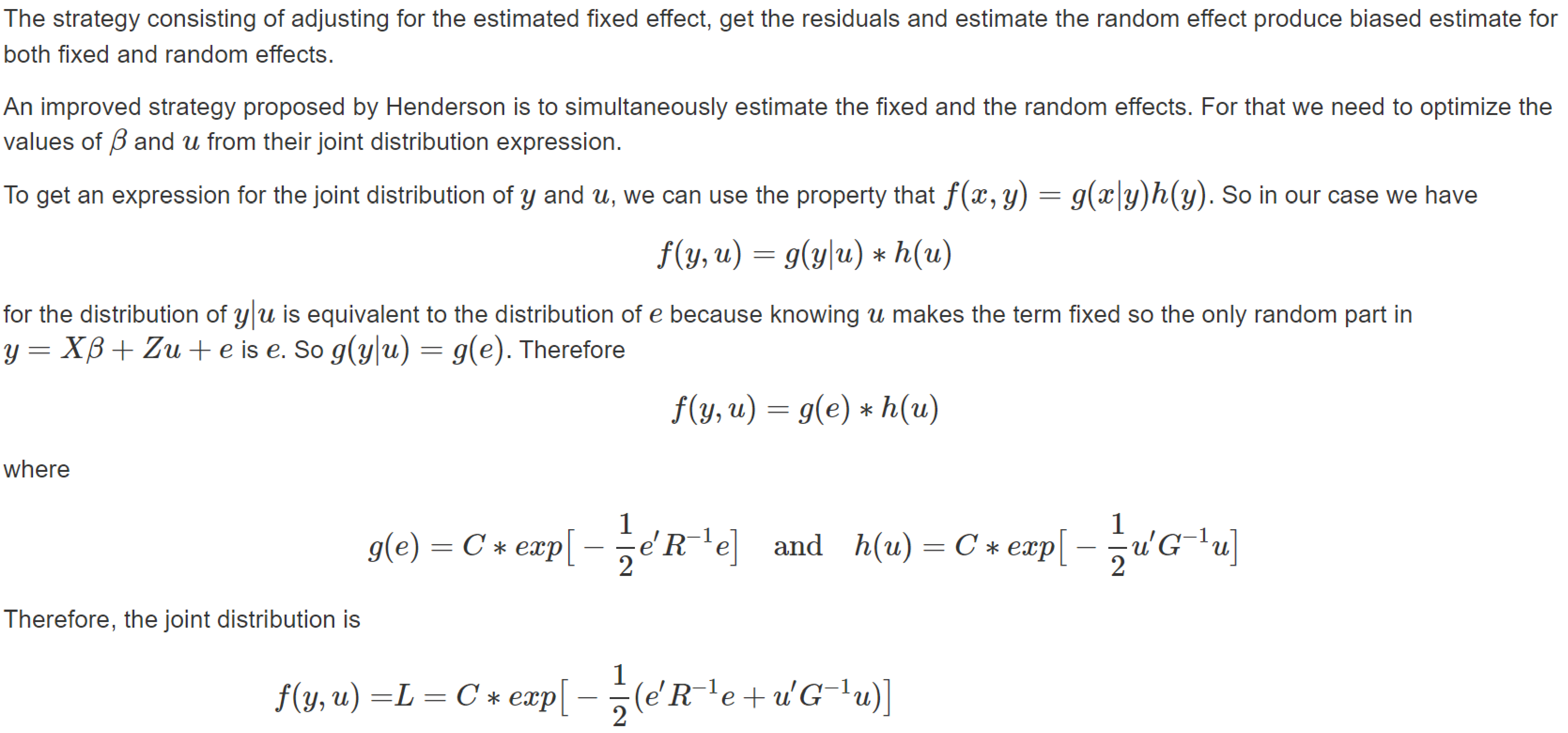
The expectation and variance of u and y are already defined. The covariance $cov(u, y) = E[(u - E(u))'(y - E(y))]$ can be developped

$$\begin{aligned} cov(u, y) &= E[(u - E(u))'(y - E(y))] \\ &= E[(u - 0)'(X\beta + Zu - X\beta)] \\ &= E[(u)'(Zu)] \\ &= ZE[u'u] \\ &= ZG \quad \text{here we used} \quad V(u) = E[u^2] - E[u]^2 = E[u'u] - 0 \end{aligned}$$

Using the property of the MVN ($E[x_1|x_2] = \mu_1 + \sigma_{12}(\sigma_2^2)^{-1}(x_2 - \mu_2)$), we can derive the conditional expectation of u given y

$$\begin{aligned} \hat{u} = E[u|y] &= 0 + GZ'V^{-1}(y - X\beta) \\ &= GZ'(ZGZ' + R)^{-1}(y - X\beta) \end{aligned}$$

Mixed models equations derivation



The strategy consisting of adjusting for the estimated fixed effect, get the residuals and estimate the random effect produce biased estimate for both fixed and random effects.

An improved strategy proposed by Henderson is to simultaneously estimate the fixed and the random effects. For that we need to optimize the values of β and u from their joint distribution expression.

To get an expression for the joint distribution of y and u, we can use the property that $f(x,y) = g(x|y)h(y)$. So in our case we have

$$f(y,u) = g(y|u) * h(u)$$

for the distribution of $y|u$ is equivalent to the distribution of e because knowing u makes the term fixed so the only random part in $y = X\beta + Zu + e$ is e. So $g(y|u) = g(e)$. Therefore

$$f(y,u) = g(e) * h(u)$$

where

$$g(e) = C * exp[-\frac{1}{2}e'R^{-1}e] \quad \text{and} \quad h(u) = C * exp[-\frac{1}{2}u'G^{-1}u]$$

Therefore, the joint distribution is

$$f(y,u) = L = C * exp[-\frac{1}{2}(e'R^{-1}e + u'G^{-1}u)]$$

Mixed models equations derivation

$$
\begin{aligned}
&\mathcal{L} = log(L) = log(C) - \frac{1}{2}(e'R^{-1}e + u'G^{-1}u) \\
&\mathcal{L} \propto e'R^{-1}e + u'G^{-1}u \\
&\mathcal{L} \propto (y - X\beta - Zu)'R^{-1}(y - X\beta - Zu) + u'G^{-1}u \quad (e = y - X\beta - Zu) \\
&\mathcal{L} \propto (y - X\beta - Zu)'R^{-1}(y - X\beta - Zu) + u'G^{-1}u
\end{aligned}
$$

If we develop the expression

$$
\begin{aligned}
\mathcal{L} \propto & (y - X\beta - Zu)'R^{-1}(y - X\beta - Zu) + u'G^{-1}u \\
& y'R^{-1}y - y'R^{-1}X\beta - y'R^{-1}Zu \\
& - (X\beta)R^{-1}y + (X\beta)R^{-1}X\beta + (X\beta)R^{-1}Zu \\
& - (Zu)R^{-1}y + (Zu)R^{-1}X\beta + (Zu)R^{-1}Zu + u'G^{-1}u
\end{aligned}
$$

Mixed models equations derivation

$$
\begin{aligned}
\mathcal{L} \propto & (y - X\beta - Zu)'R^{-1}(y - X\beta - Zu) + u'G^{-1}u \\
& y'R^{-1}y - y'R^{-1}X\beta - y'R^{-1}Zu \\
& - (X\beta)R^{-1}y + (X\beta)R^{-1}X\beta + (X\beta)R^{-1}Zu \\
& - (Zu)R^{-1}y + (Zu)R^{-1}X\beta + (Zu)R^{-1}Zu + u'G^{-1}u
\end{aligned}
$$

Derivation with respect to β

$$
\begin{aligned}
\frac{\mathcal{L}}{\partial\beta} = 0 \Longrightarrow & - y'R^{-1}X - X'R^{-1}y + (X\beta)'R^{-1}X + XR^{-1}X\beta + X'R^{-1}Zu + (Zu)'R^{-1}X = 0 \\
& - 2X'R^{-1}y + 2X'R^{-1}X\beta + 2X'R^{-1}Zu = 0 \\
& X'R^{-1}X\beta + X'R^{-1}Zu = X'R^{-1}y
\end{aligned}
$$

Derivation with respect to u

$$
\begin{aligned}
\frac{\mathcal{L}}{\partial u} = 0 \Longrightarrow & y'R^{-1}Z + (X\beta)'R^{-1}Z - Z'R^{-1}y + (Z)'R^{-1}X\beta + (Z)'R^{-1}XZu + (Zu)'R^{-1}Z + 2G^{-1}u = 0 \\
& - 2y'R^{-1}Z + 2Z'R^{-1}X\beta + 2Z'R^{-1}Zu + 2G^{-1}u = 0 \\
& Z'R^{-1}X\beta + Z'R^{-1}Zu + G^{-1}u = Z'R^{-1}y
\end{aligned}
$$

The two solution can be reorganized in a matrix system of equations

$$
\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}
$$

Mixed models equations

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z+G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

Estimation of the variance component R and G using for example REML

Then estimation of the BLUE/BLUP by inserting the variance estimates in the MM equations

Fixed/random?

$$\underline{y} = X\beta + Z\underline{u} + \underline{\epsilon} \quad [1]$$

y = u + design (rep, block) + treatment (fertilisation) + geno + e

Fixed terms

Question: le terme (env, rep, géno) est-il divisé en un nombre (réduit) de catégories clairement identifiables pour lesquels le chercheur/sélectionneur a un intérêt particulier

Ex: level of fertilization: low nitrogen vs high nitrogen: Experiment is specifically designed to know this constrast

Estimation of the BLUE (Best linear unbiased effect)

Fixed/random?

y = u + **design (rep, block)** + treatment (fertilisation) + **geno** + e

Random terms

Question*:*

- Les différents niveaux du termes peuvent-ils être considérés comme provenant d'une population plus générale ou une distribution probabiliste (ex geno1, geno2, … genoN sont des entités d'une population plus large représentant la diversité génétique)

- le terme est-il divisé en un (grand) nombre de catégories peu identifiées (ex: block 1 ~ block 2 ~ block 3 petites différences entre les blocks).
- « Nuisance » terme to control the variability associated to a large number of unidentified effects : ex: experimental design terms.
- Use random term if you specifically target to model dependence between observation (e.g. hierarchichal or clustered analyses).
- Control for the randomness associated with the distribution/population + interest in the realized values -> Estimation of the BLUP (Best linear unbiased predictor)

Fixed/random?

Extra considerations

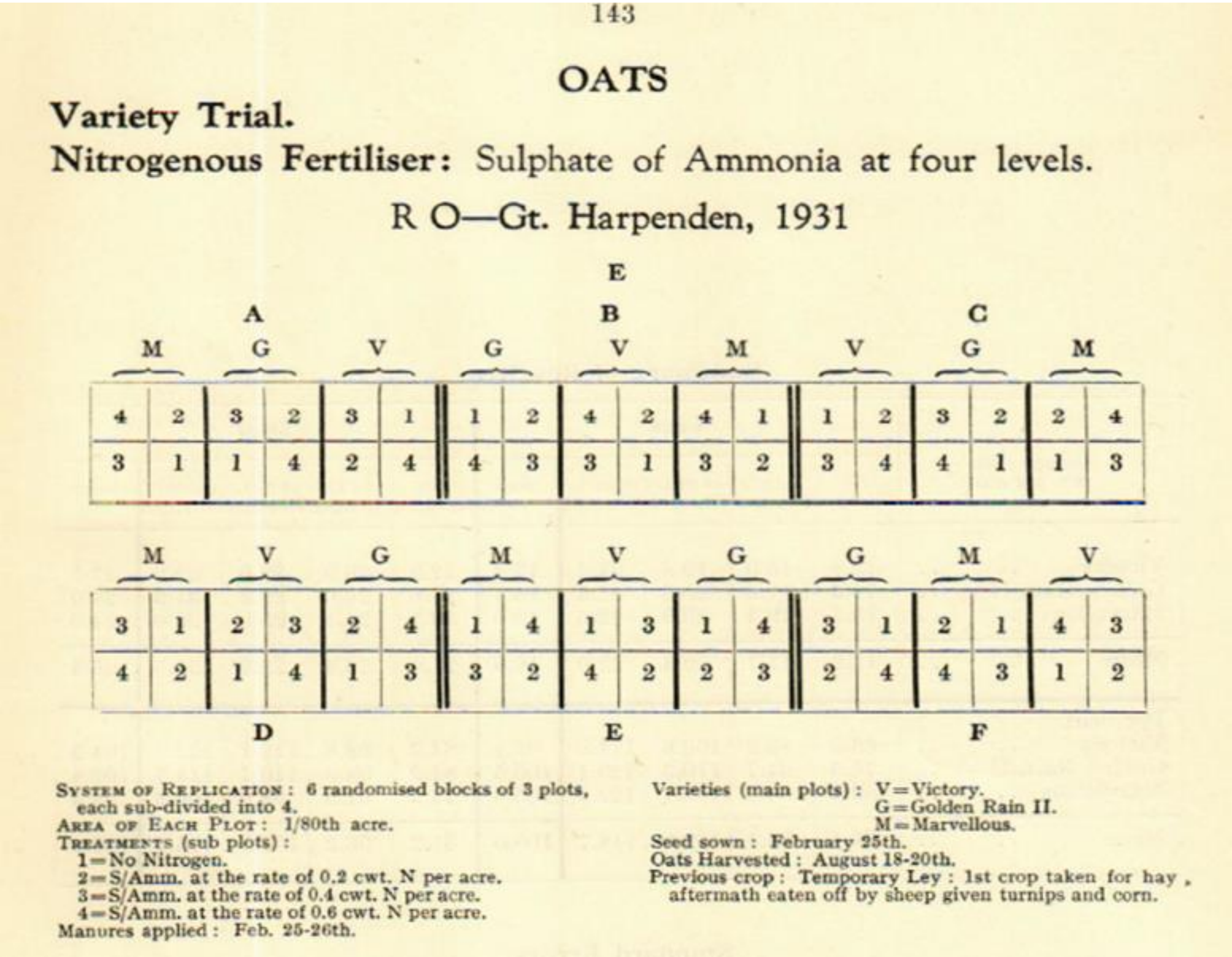
- Y (dependent variable) and e (error term) are always random

- The definition of a terms can depends of the next usage. For example, in a two-stage analysis, if the second stage model use a random (fixed) genetic term, in the first stage (adjusted mean computation) you should calculate random BLUP (fixed BLUE).

- Random terms use less degree of freedom. Fixed term k levels degrees of freedom. Random term 1 degree of freedom.

- Computationally easier to estimate fixed effects. So if your model fail to converge consider passing some terms as fixed.

Exercise



143

OATS

Variety Trial.

Nitrogenous Fertiliser: Sulphate of Ammonia at four levels.

R O—Gt. Harpenden, 1931

E

A						B						C					
M		G		V		G		V		M		V		G		M	
4	2	3	2	3	1	1	2	4	2	4	1	1	2	3	2	2	4
3	1	1	4	2	4	4	3	3	1	3	2	3	4	4	1	1	3

M		V		G		M		V		G		G		M		V	
3	1	2	3	2	4	1	4	1	3	1	4	3	1	2	1	4	3
4	2	1	4	1	3	3	2	4	2	2	3	2	4	4	3	1	2
D						E						F					

System of Replication: 6 randomised blocks of 3 plots, each sub-divided into 4.
Area of Each Plot: 1/80th acre.
Treatments (sub plots):
1=No Nitrogen.
2=S/Amm. at the rate of 0.2 cwt. N per acre.
3=S/Amm. at the rate of 0.4 cwt. N per acre.
4=S/Amm. at the rate of 0.6 cwt. N per acre.
Manures applied: Feb. 25-26th.

Varieties (main plots): V=Victory.
G=Golden Rain II.
M=Marvellous.
Seed sown: February 25th.
Oats Harvested: August 18-20th.
Previous crop: Temporary Ley: 1st crop taken for hay, aftermath eaten off by sheep given turnips and corn.

Mixed models equations - simplification

The following mixed model

$$\underline{y} = X\beta + Z\underline{u} + \underline{\epsilon} \quad u \sim N(0, G) \quad \epsilon \sim N(0, R)$$

is associated with the follosing MM equations

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

Let us assume now that $R = I\sigma_\epsilon^2$, Then the MM reduce to that expression (multiplication by R on both sides)

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \sigma_\epsilon^2 G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Animal or pedigree BLUP model

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \sigma_\epsilon^2 G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Let us rewrite the random term of the mixed model to consider that the covariance between two individual is proportional to their (genetic) relatedness

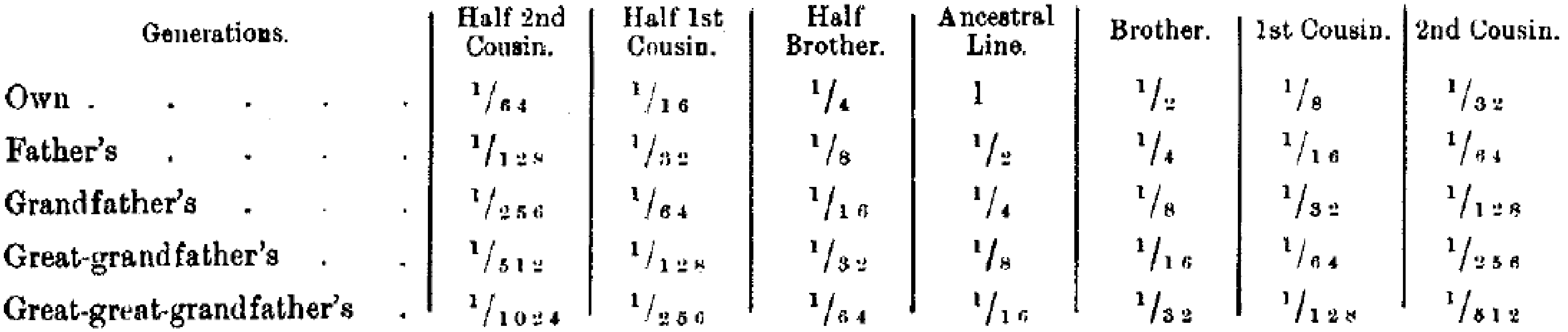
$$\underline{y} = X\beta + Z\underline{u} + \underline{\epsilon} \quad u \sim N(0, G = A\sigma_a^2) \quad \epsilon \sim N(0, I\sigma_\epsilon^2)$$

Where the A matrix represent the expected (genetic) relatedness between individual infered from pedigree records.

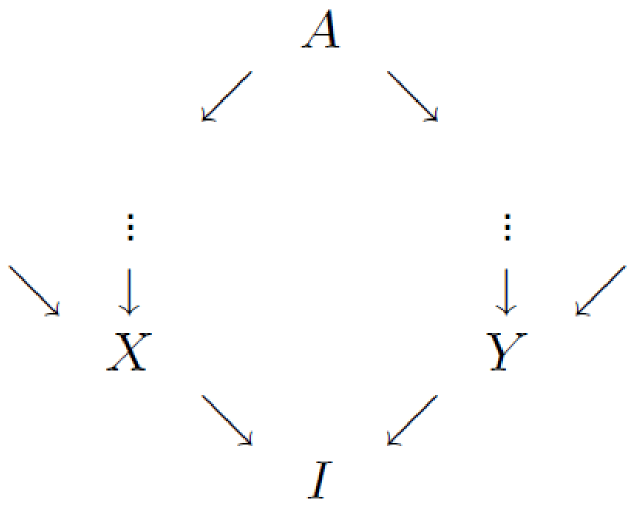
We can replace the G term in the MM equation

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \frac{\sigma_\epsilon^2}{\sigma_a^2} A^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Animal or pedigree BLUP model



Generations.	Half 2nd Cousin.	Half 1st Cousin.	Half Brother.	Ancestral Line.	Brother.	1st Cousin.	2nd Cousin.
Own	1/64	1/16	1/4	1	1/2	1/8	1/32
Father's	1/128	1/32	1/8	1/2	1/4	1/16	1/64
Grandfather's . . .	1/256	1/64	1/16	1/4	1/8	1/32	1/128
Great-grandfather's . .	1/512	1/128	1/32	1/8	1/16	1/64	1/256
Great-great-grandfather's .	1/1024	1/256	1/64	1/16	1/32	1/128	1/512

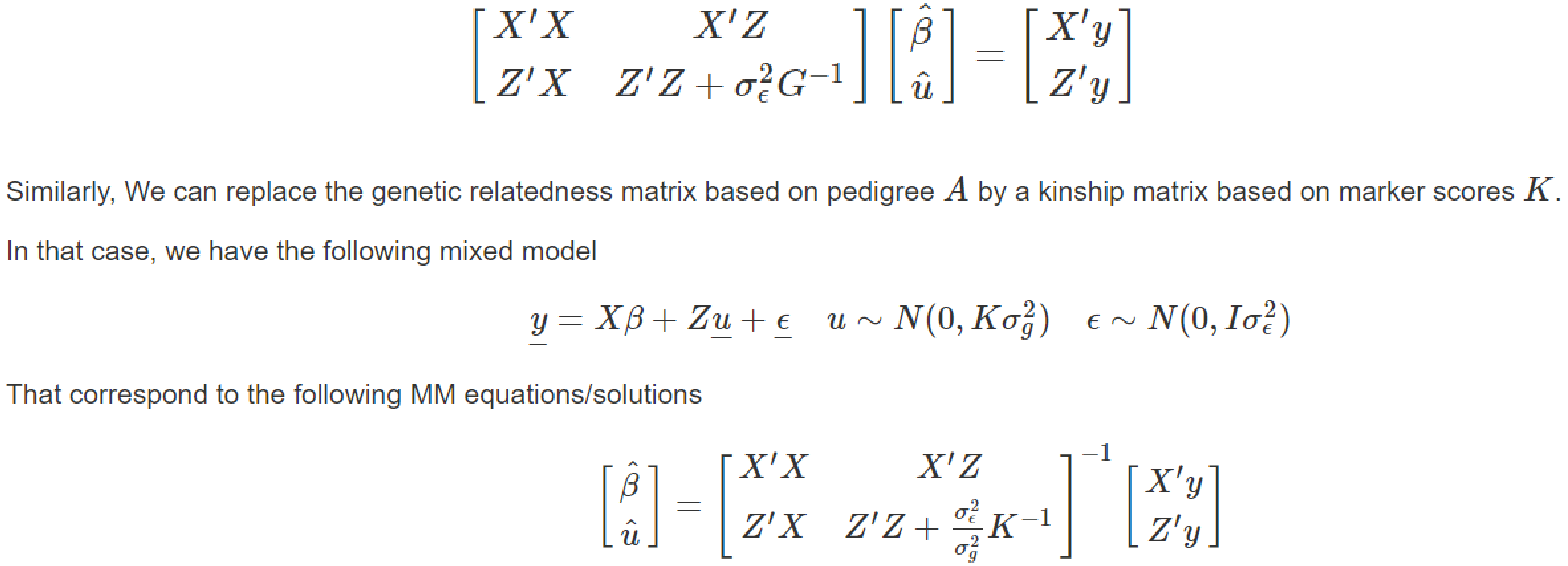


$$F_I = \theta_{XY} = \left(\frac{1}{2}\right)^n (1 + F_A)$$

	G1	G2	[...]	Gn
G1	1	0,5	0,5	0,5
G2		1	0,25	0,25
[...]			1	0,25
Gn				1

σ_g^2

Genomic G-BLUP model



$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \sigma_\epsilon^2 G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Similarly, We can replace the genetic relatedness matrix based on pedigree A by a kinship matrix based on marker scores K.

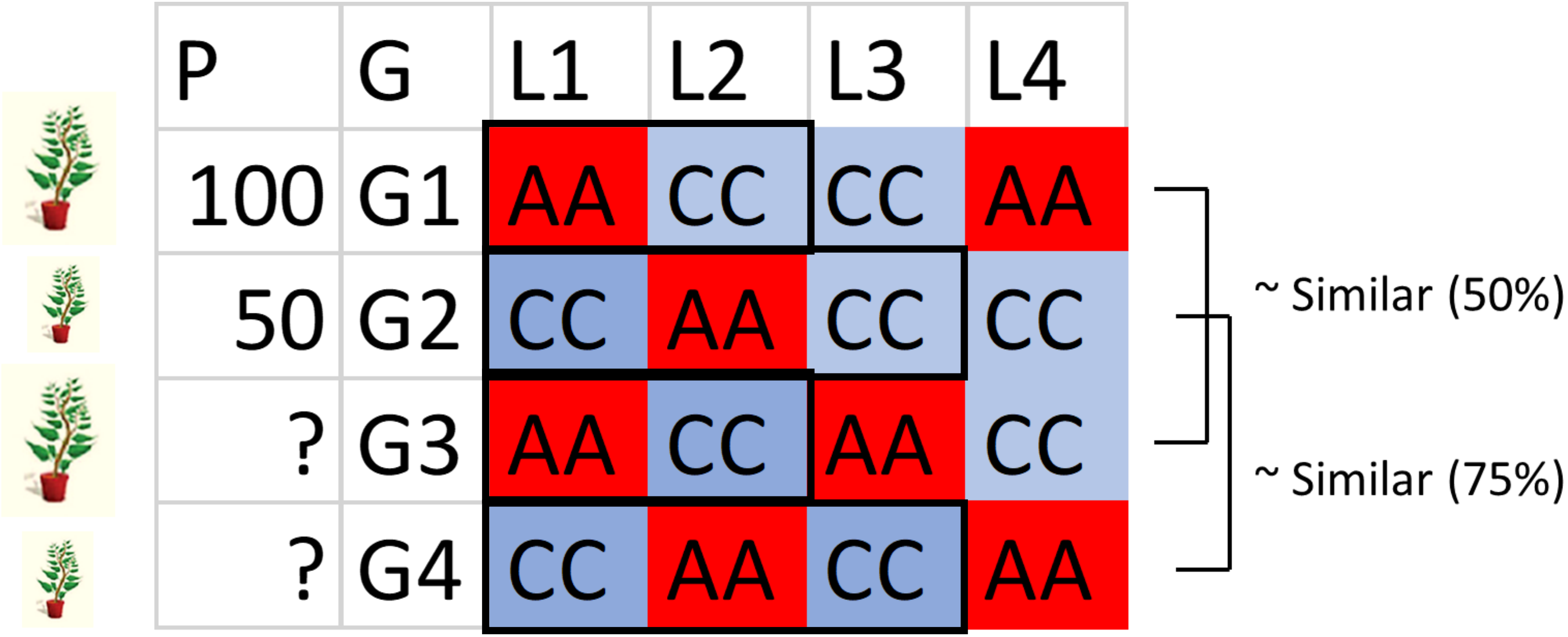
In that case, we have the following mixed model

$$\underline{y} = X\beta + Z\underline{u} + \underline{\epsilon} \quad u \sim N(0, K\sigma_g^2) \quad \epsilon \sim N(0, I\sigma_\epsilon^2)$$

That correspond to the following MM equations/solutions

$$\begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \frac{\sigma_\epsilon^2}{\sigma_g^2} K^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Genomic G-BLUP model



P	G	L1	L2	L3	L4
100	G1	AA	CC	CC	AA
50	G2	CC	AA	CC	CC
?	G3	AA	CC	AA	CC
?	G4	CC	AA	CC	AA

~ Similar (50%)

~ Similar (75%)

Genotypic (marker) similarity should cause phenotypic similarity

Genomic G-BLUP model

Compared to the A matrix, K allows to differentiate between individual with the same expectd co-ancestry coefficient (e.g. two sibs) due to the allele Mendelian sampling

A: Expected co-ancestry

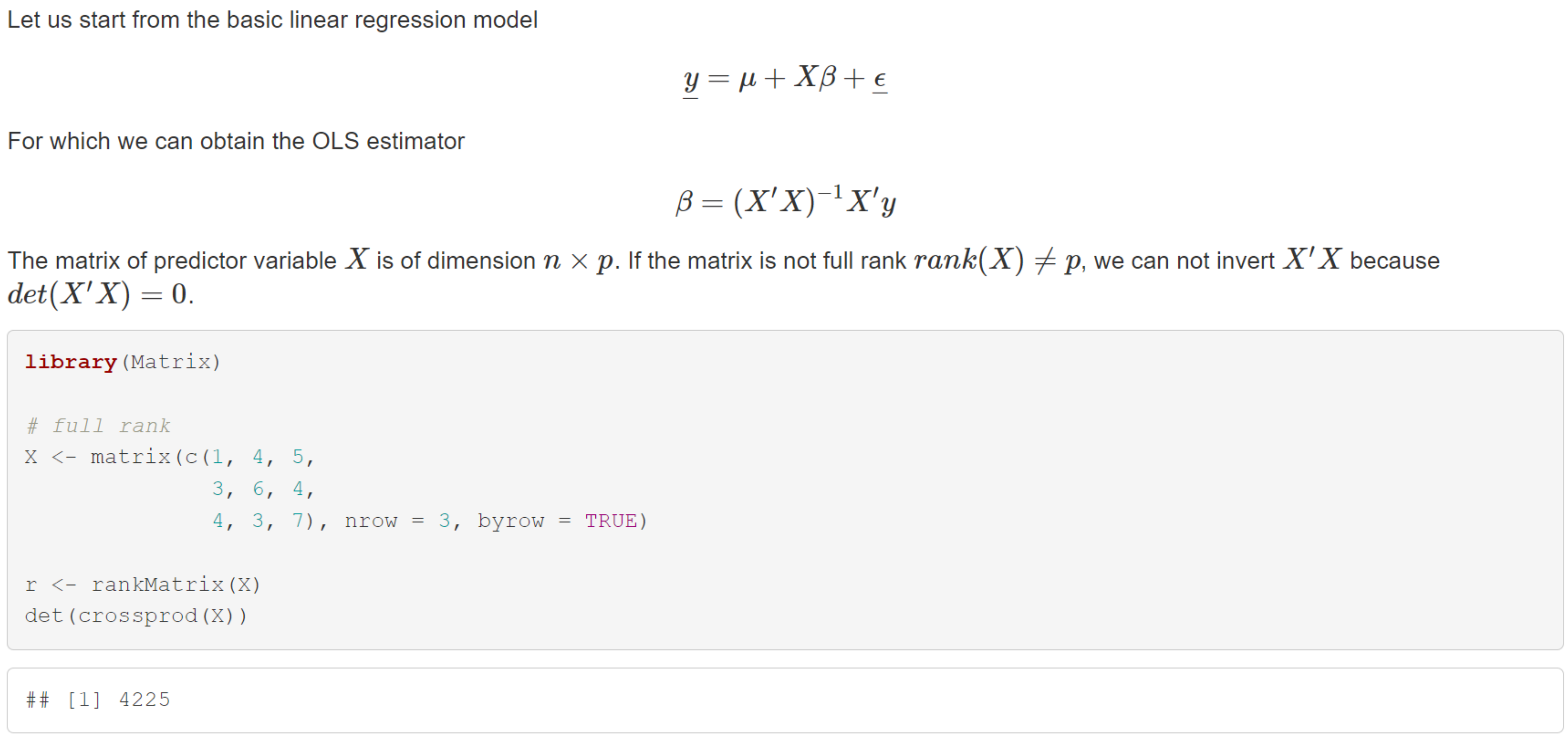
	G1	G2	[…]	Gn
G1	1	0,5	0,5	0,5
G2		1	0,25	0,25
[…]			1	0,25
Gn				1

K: Realized or molecular co-ancestry

	G1	G2	[…]	Gn
G1	1	0,42	0,58	0,48
G2		1	0,2	0,28
[…]			1	0,24
Gn				1

Regularization methods

Rank deficiency



Let us start from the basic linear regression model

$$\underline{y} = \mu + X\beta + \underline{\epsilon}$$

For which we can obtain the OLS estimator

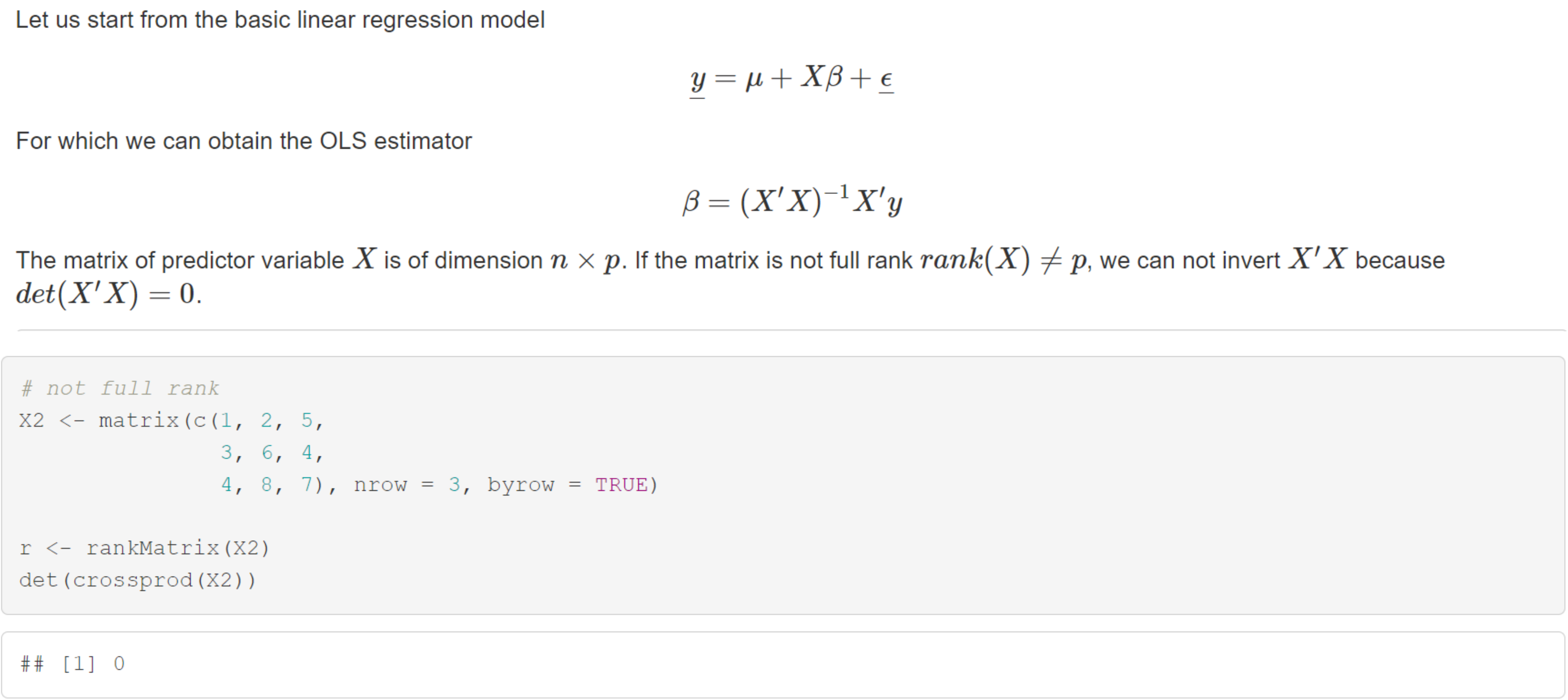
$$\beta = (X'X)^{-1}X'y$$

The matrix of predictor variable X is of dimension $n \times p$. If the matrix is not full rank $rank(X) \neq p$, we can not invert $X'X$ because $det(X'X) = 0$.

```
library(Matrix)

# full rank
X <- matrix(c(1, 4, 5,
              3, 6, 4,
              4, 3, 7), nrow = 3, byrow = TRUE)

r <- rankMatrix(X)
det(crossprod(X))
```

```
## [1] 4225
```

Rank deficiency



Let us start from the basic linear regression model

$$\underline{y} = \mu + X\beta + \underline{\epsilon}$$

For which we can obtain the OLS estimator

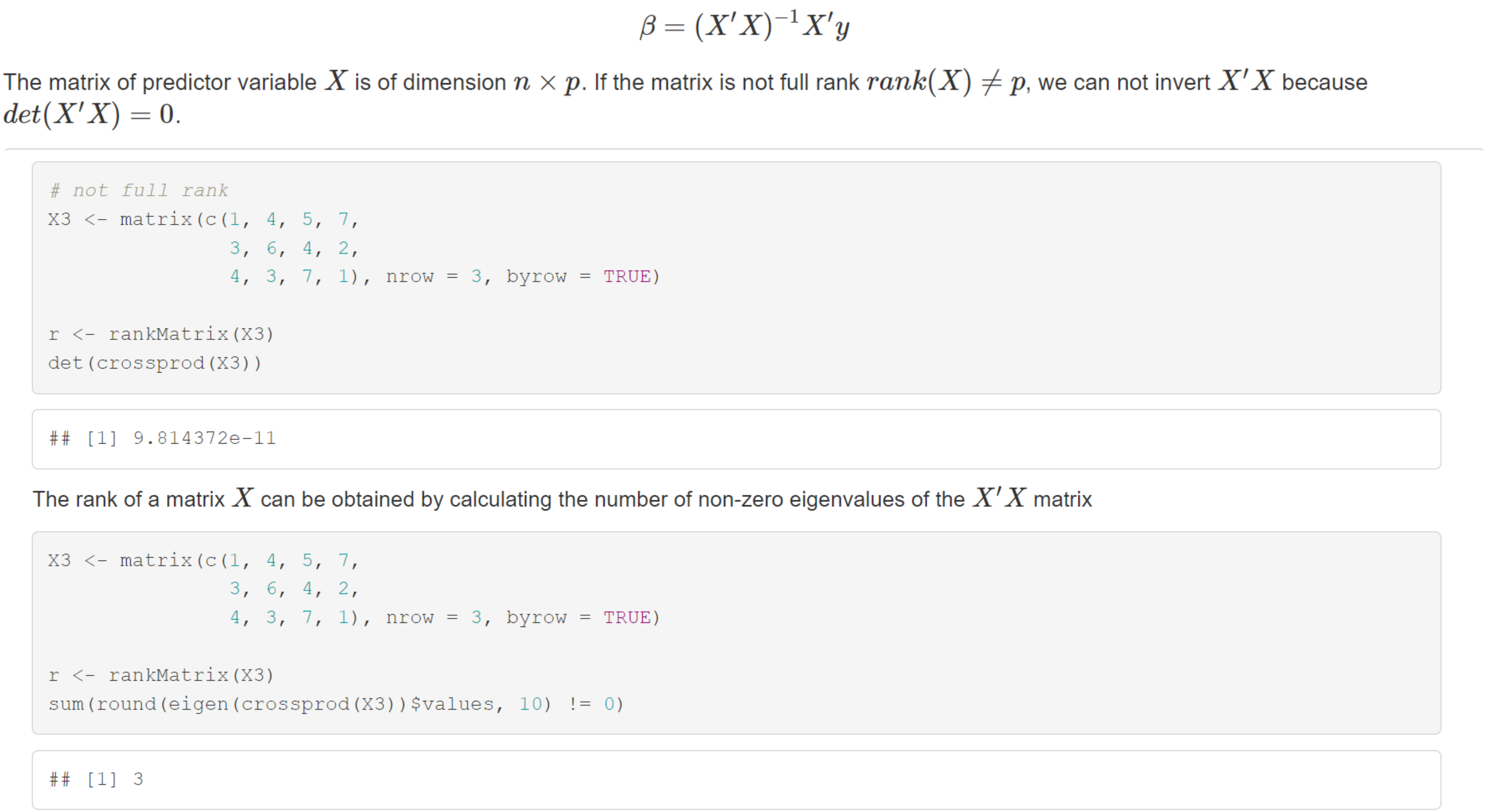
$$\beta = (X'X)^{-1}X'y$$

The matrix of predictor variable X is of dimension $n \times p$. If the matrix is not full rank $rank(X) \neq p$, we can not invert $X'X$ because $det(X'X) = 0$.

```
# not full rank
X2 <- matrix(c(1, 2, 5,
               3, 6, 4,
               4, 8, 7), nrow = 3, byrow = TRUE)

r <- rankMatrix(X2)
det(crossprod(X2))
```

```
## [1] 0
```

Rank deficiency (n << p)



$$\beta = (X'X)^{-1}X'y$$

The matrix of predictor variable X is of dimension $n \times p$. If the matrix is not full rank $rank(X) \neq p$, we can not invert $X'X$ because $det(X'X) = 0$.

```
# not full rank
X3 <- matrix(c(1, 4, 5, 7,
               3, 6, 4, 2,
               4, 3, 7, 1), nrow = 3, byrow = TRUE)

r <- rankMatrix(X3)
det(crossprod(X3))
```

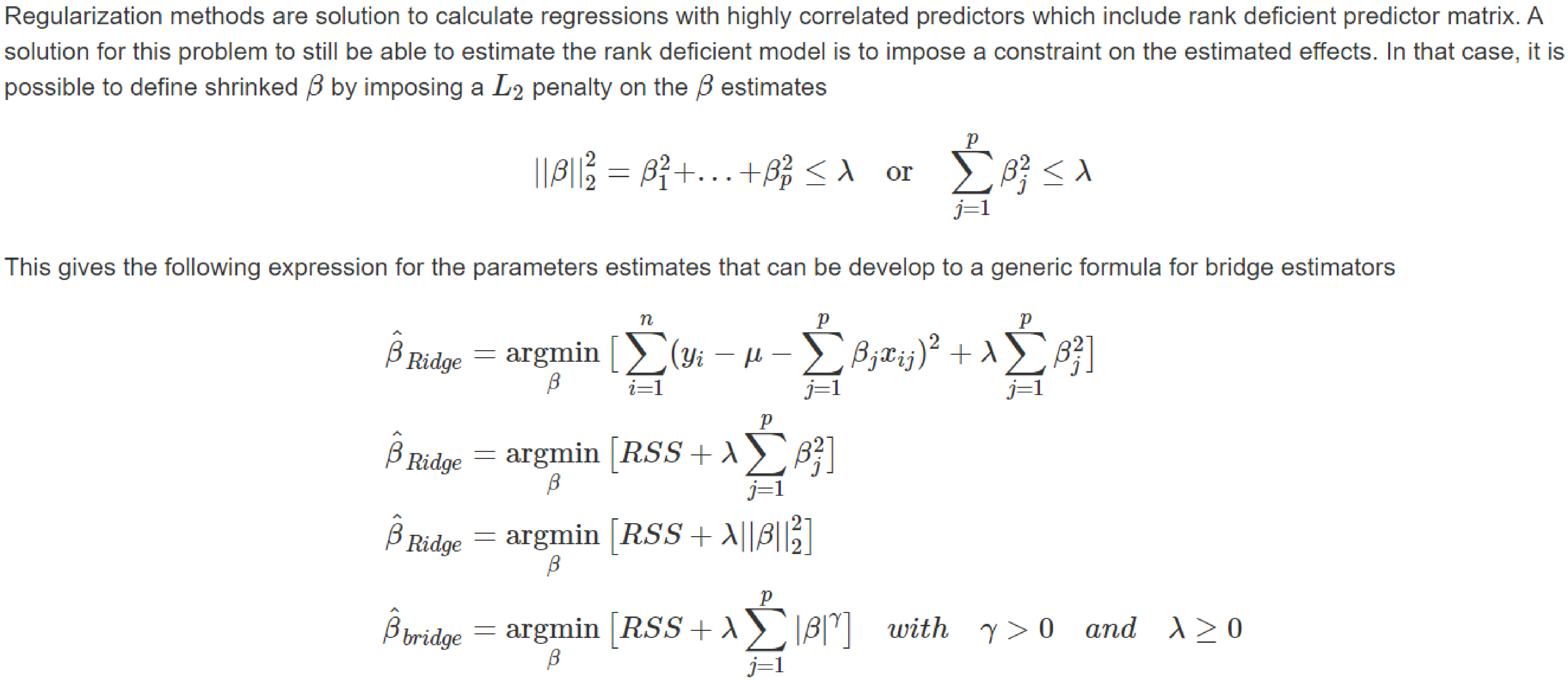
```
## [1] 9.814372e-11
```

The rank of a matrix X can be obtained by calculating the number of non-zero eigenvalues of the $X'X$ matrix

```
X3 <- matrix(c(1, 4, 5, 7,
               3, 6, 4, 2,
               4, 3, 7, 1), nrow = 3, byrow = TRUE)

r <- rankMatrix(X3)
sum(round(eigen(crossprod(X3))$values, 10) != 0)
```

```
## [1] 3
```

Ridge regression a solution to rank deficiency (n << p)



Regularization methods are solution to calculate regressions with highly correlated predictors which include rank deficient predictor matrix. A solution for this problem to still be able to estimate the rank deficient model is to impose a constraint on the estimated effects. In that case, it is possible to define shrinked β by imposing a L_2 penalty on the β estimates

$$||\beta||_2^2 = \beta_1^2 + \ldots + \beta_p^2 \leq \lambda \quad \text{or} \quad \sum_{j=1}^{p} \beta_j^2 \leq \lambda$$

This gives the following expression for the parameters estimates that can be develop to a generic formula for bridge estimators

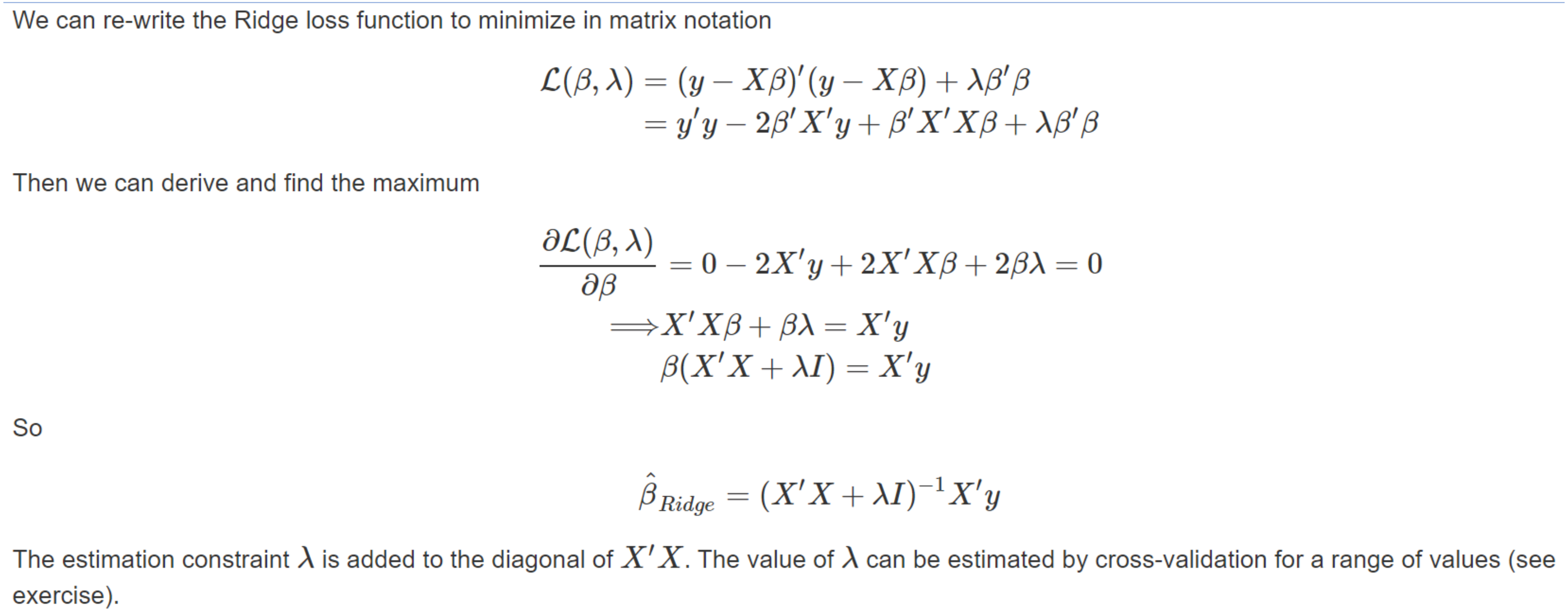
$$\hat{\beta}_{Ridge} = \underset{\beta}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (y_i - \mu - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2\right]$$

$$\hat{\beta}_{Ridge} = \underset{\beta}{\operatorname{argmin}} \left[RSS + \lambda \sum_{j=1}^{p} \beta_j^2\right]$$

$$\hat{\beta}_{Ridge} = \underset{\beta}{\operatorname{argmin}} \left[RSS + \lambda ||\beta||_2^2\right]$$

$$\hat{\beta}_{bridge} = \underset{\beta}{\operatorname{argmin}} \left[RSS + \lambda \sum_{j=1}^{p} |\beta|^{\gamma}\right] \quad with \quad \gamma > 0 \quad and \quad \lambda \geq 0$$

Ridge regression a solution to rank deficiency (n << p)



We can re-write the Ridge loss function to minimize in matrix notation

$$\begin{aligned}\mathcal{L}(\beta,\lambda) &= (y - X\beta)'(y - X\beta) + \lambda\beta'\beta \\ &= y'y - 2\beta'X'y + \beta'X'X\beta + \lambda\beta'\beta\end{aligned}$$

Then we can derive and find the maximum

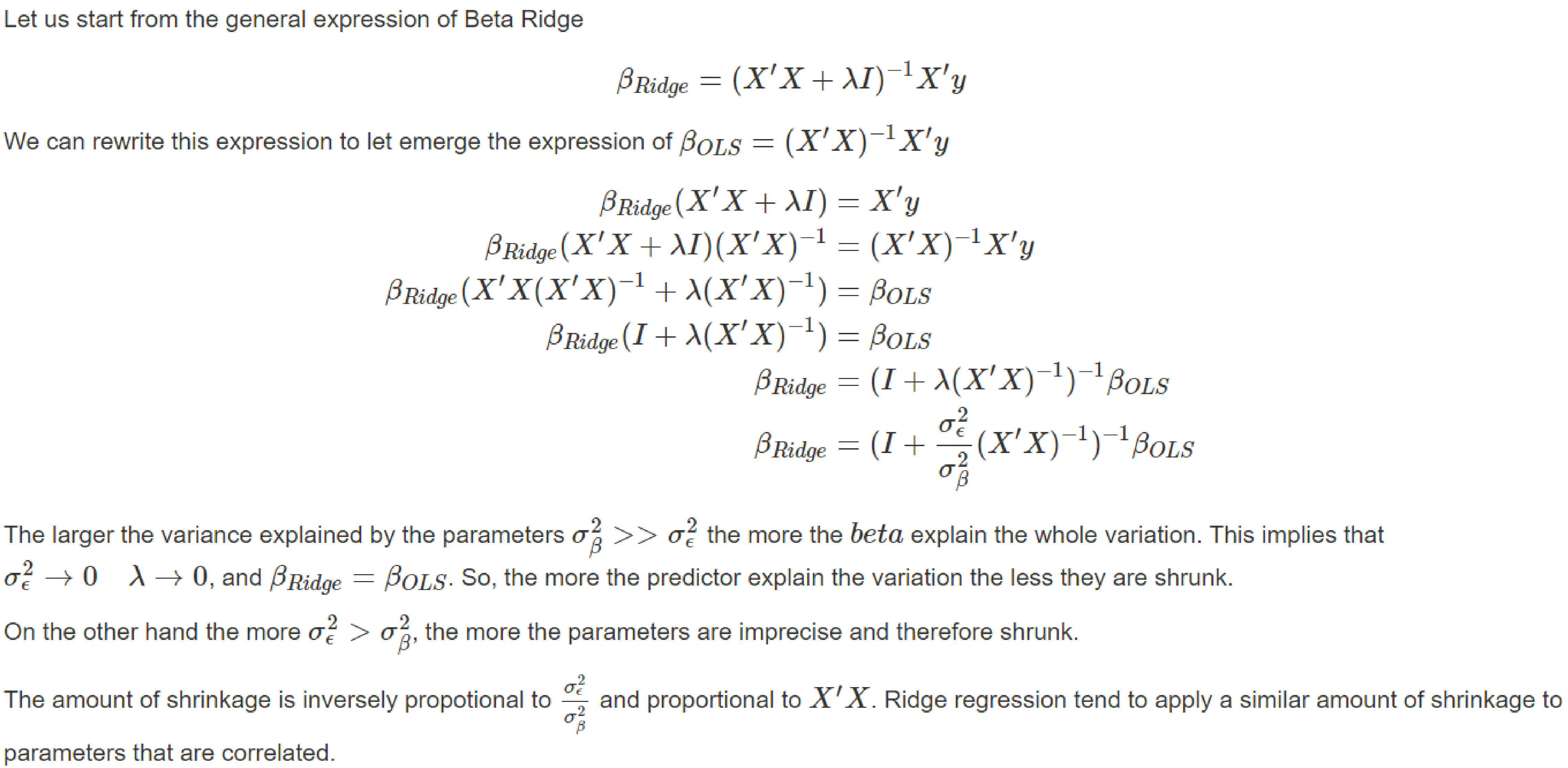
$$\begin{aligned}\frac{\partial\mathcal{L}(\beta,\lambda)}{\partial\beta} &= 0 - 2X'y + 2X'X\beta + 2\beta\lambda = 0 \\ \Longrightarrow & X'X\beta + \beta\lambda = X'y \\ & \beta(X'X + \lambda I) = X'y\end{aligned}$$

So

$$\hat{\beta}_{Ridge} = (X'X + \lambda I)^{-1}X'y$$

The estimation constraint λ is added to the diagonal of $X'X$. The value of λ can be estimated by cross-validation for a range of values (see exercise).

Relationship beta Ridge beta OLS



Let us start from the general expression of Beta Ridge

$$\beta_{Ridge} = (X'X + \lambda I)^{-1} X'y$$

We can rewrite this expression to let emerge the expression of $\beta_{OLS} = (X'X)^{-1} X'y$

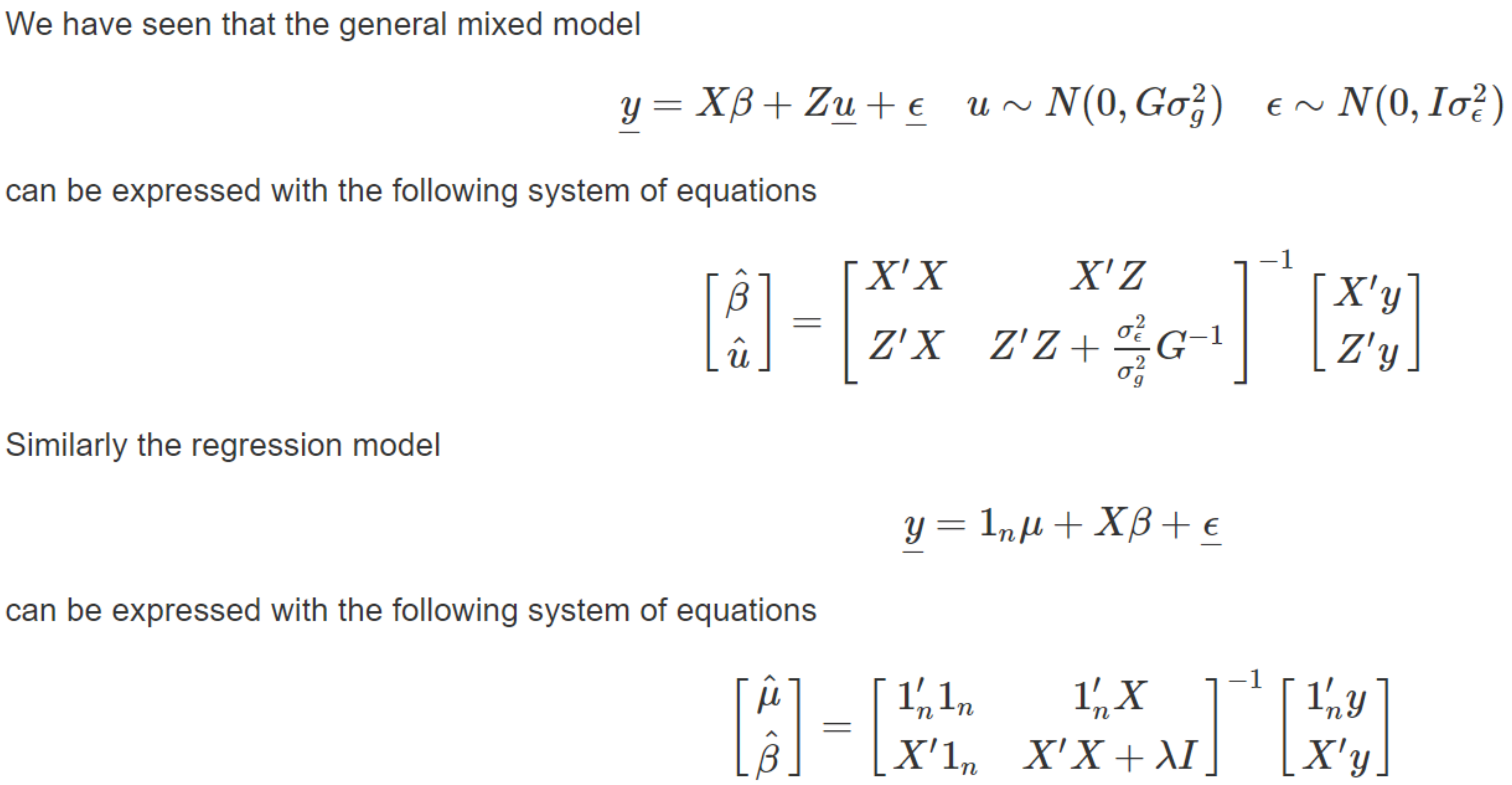
$$\begin{aligned} \beta_{Ridge}(X'X + \lambda I) &= X'y \\ \beta_{Ridge}(X'X + \lambda I)(X'X)^{-1} &= (X'X)^{-1} X'y \\ \beta_{Ridge}(X'X(X'X)^{-1} + \lambda (X'X)^{-1}) &= \beta_{OLS} \\ \beta_{Ridge}(I + \lambda (X'X)^{-1}) &= \beta_{OLS} \\ \beta_{Ridge} &= (I + \lambda (X'X)^{-1})^{-1} \beta_{OLS} \\ \beta_{Ridge} &= (I + \frac{\sigma_\epsilon^2}{\sigma_\beta^2} (X'X)^{-1})^{-1} \beta_{OLS} \end{aligned}$$

The larger the variance explained by the parameters $\sigma_\beta^2 >> \sigma_\epsilon^2$ the more the $beta$ explain the whole variation. This implies that $\sigma_\epsilon^2 \to 0 \quad \lambda \to 0$, and $\beta_{Ridge} = \beta_{OLS}$. So, the more the predictor explain the variation the less they are shrunk.

On the other hand the more $\sigma_\epsilon^2 > \sigma_\beta^2$, the more the parameters are imprecise and therefore shrunk.

The amount of shrinkage is inversely propotional to $\frac{\sigma_\epsilon^2}{\sigma_\beta^2}$ and proportional to $X'X$. Ridge regression tend to apply a similar amount of shrinkage to parameters that are correlated.

Relationship Ridge regression and mixed mode estimation



We have seen that the general mixed model

$$\underline{y} = X\beta + Z\underline{u} + \underline{\epsilon} \quad u \sim N(0, G\sigma_g^2) \quad \epsilon \sim N(0, I\sigma_\epsilon^2)$$

can be expressed with the following system of equations

$$\begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \frac{\sigma_\epsilon^2}{\sigma_g^2} G^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

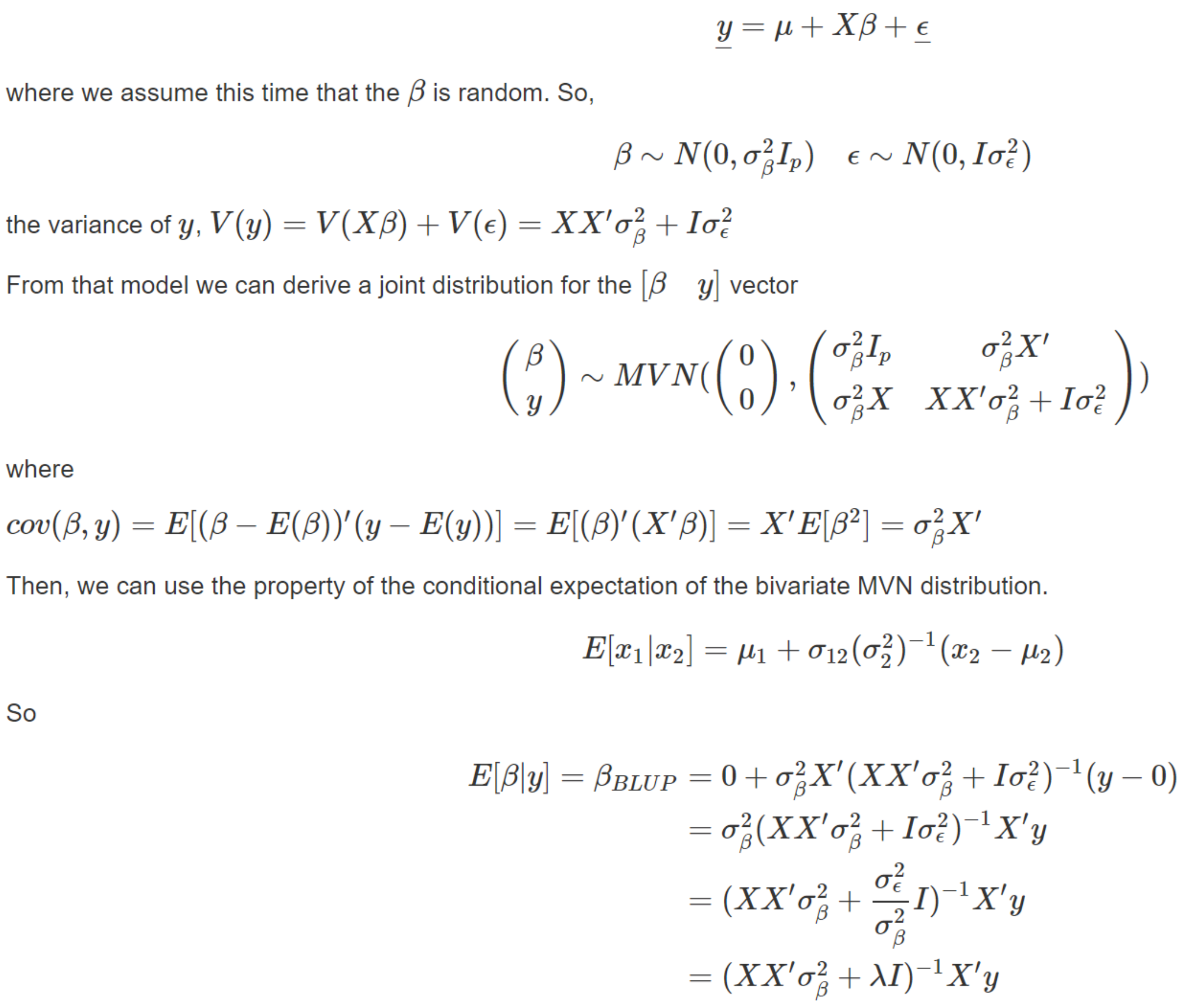
Similarly the regression model

$$\underline{y} = 1_n\mu + X\beta + \underline{\epsilon}$$

can be expressed with the following system of equations

$$\begin{bmatrix} \hat{\mu} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} 1_n' 1_n & 1_n' X \\ X' 1_n & X'X + \lambda I \end{bmatrix}^{-1} \begin{bmatrix} 1_n' y \\ X'y \end{bmatrix}$$

Relationship Ridge regression and mixed mode estimation



$$\underline{y} = \mu + X\beta + \underline{\epsilon}$$

where we assume this time that the β is random. So,

$$\beta \sim N(0, \sigma_\beta^2 I_p) \quad \epsilon \sim N(0, I\sigma_\epsilon^2)$$

the variance of y, $V(y) = V(X\beta) + V(\epsilon) = XX'\sigma_\beta^2 + I\sigma_\epsilon^2$

From that model we can derive a joint distribution for the $[\beta \quad y]$ vector

$$\begin{pmatrix} \beta \\ y \end{pmatrix} \sim MVN(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\beta^2 I_p & \sigma_\beta^2 X' \\ \sigma_\beta^2 X & XX'\sigma_\beta^2 + I\sigma_\epsilon^2 \end{pmatrix})$$

where

$cov(\beta, y) = E[(\beta - E(\beta))'(y - E(y))] = E[(\beta)'(X'\beta)] = X'E[\beta^2] = \sigma_\beta^2 X'$

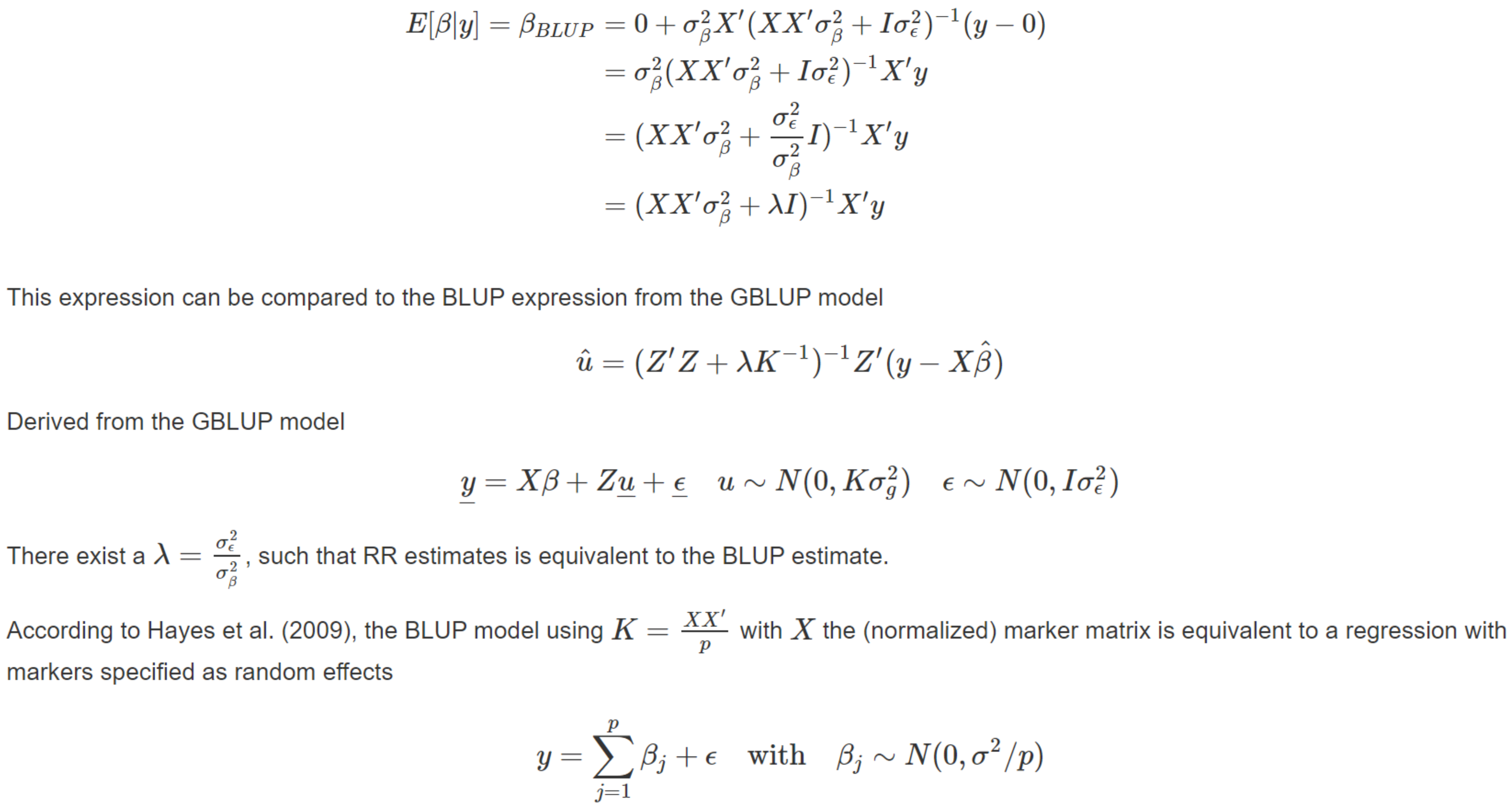
Then, we can use the property of the conditional expectation of the bivariate MVN distribution.

$$E[x_1|x_2] = \mu_1 + \sigma_{12}(\sigma_2^2)^{-1}(x_2 - \mu_2)$$

So

$$\begin{aligned} E[\beta|y] = \beta_{BLUP} &= 0 + \sigma_\beta^2 X'(XX'\sigma_\beta^2 + I\sigma_\epsilon^2)^{-1}(y - 0) \\ &= \sigma_\beta^2 (XX'\sigma_\beta^2 + I\sigma_\epsilon^2)^{-1} X'y \\ &= (XX'\sigma_\beta^2 + \frac{\sigma_\epsilon^2}{\sigma_\beta^2} I)^{-1} X'y \\ &= (XX'\sigma_\beta^2 + \lambda I)^{-1} X'y \end{aligned}$$

Relationship Ridge regression and mixed mode estimation



$$\begin{aligned} E[\beta|y] = \beta_{BLUP} &= 0 + \sigma^2_\beta X'(XX'\sigma^2_\beta + I\sigma^2_\epsilon)^{-1}(y-0) \\ &= \sigma^2_\beta(XX'\sigma^2_\beta + I\sigma^2_\epsilon)^{-1}X'y \\ &= (XX'\sigma^2_\beta + \frac{\sigma^2_\epsilon}{\sigma^2_\beta}I)^{-1}X'y \\ &= (XX'\sigma^2_\beta + \lambda I)^{-1}X'y \end{aligned}$$

This expression can be compared to the BLUP expression from the GBLUP model

$$\hat{u} = (Z'Z + \lambda K^{-1})^{-1}Z'(y - X\hat{\beta})$$

Derived from the GBLUP model

$$\underline{y} = X\beta + Z\underline{u} + \underline{\epsilon} \quad u \sim N(0, K\sigma^2_g) \quad \epsilon \sim N(0, I\sigma^2_\epsilon)$$

There exist a $\lambda = \frac{\sigma^2_\epsilon}{\sigma^2_\beta}$, such that RR estimates is equivalent to the BLUP estimate.

According to Hayes et al. (2009), the BLUP model using $K = \frac{XX'}{p}$ with X the (normalized) marker matrix is equivalent to a regression with markers specified as random effects

$$y = \sum_{j=1}^{p} \beta_j + \epsilon \quad \text{with} \quad \beta_j \sim N(0, \sigma^2/p)$$

LASSO: Least absolute shrinkage and selection operator

The Ridge estimator does perform any variable selection, which means that all estimates get a non-zero value. The logic of the Ridge regression can be extended to perform variable selection by using a L_1 penality

$$||\beta||_1 = |\beta_1| + \ldots + |\beta_p| \leq \lambda \quad \text{or} \quad \sum_{j=1}^{p} |\beta_j| \leq \lambda$$

Therefore we can define least absolute shrinkage and selection operator (LASSO)

$$\hat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \left[RSS + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

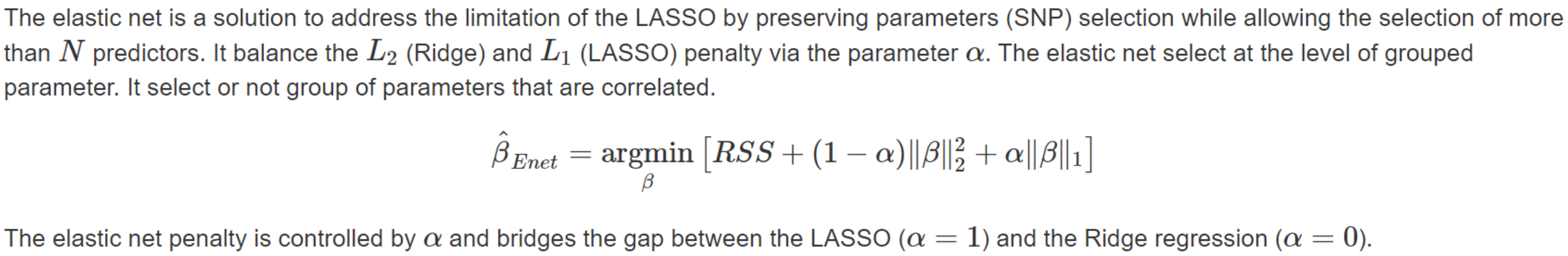
$$\hat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \left[RSS + \lambda ||\beta||_1 \right]$$

It correspond to the following matrix notation loss function

$$\mathcal{L}(\beta, \lambda) = (y - X\beta)'(y - X\beta) + \lambda ||\beta||_1$$

Interpretation: The higher the value of λ the fewer the non-zero parameters. The LASSO cannot select a larger number of predictors p than the sample size N if $p >> N$. The LASSO parameter select one parameter among the higly correlated ones and discard the rest.

Elastic net



The elastic net is a solution to address the limitation of the LASSO by preserving parameters (SNP) selection while allowing the selection of more than N predictors. It balance the L_2 (Ridge) and L_1 (LASSO) penalty via the parameter α. The elastic net select at the level of grouped parameter. It select or not group of parameters that are correlated.

$$\hat{\beta}_{Enet} = \underset{\beta}{\operatorname{argmin}} \left[RSS + (1-\alpha)||\beta||_2^2 + \alpha||\beta||_1 \right]$$

The elastic net penalty is controlled by α and bridges the gap between the LASSO ($\alpha = 1$) and the Ridge regression ($\alpha = 0$).