Statistical Theory and Modeling (ST2601)
Lecture 8 - Linear regression in vector form

Mattias Villani

Department of Statistics
Stockholm University

mattiasvillani.com @matvil @matvil mattiasvillani

Overview

- **Vectors and matrices - minimal intro to linear algebra**
- **Linear regression in vector form**
- **Multivariate normal distribution**

Goals of the lecture

- **Linear regression in vector form**

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$$

- **Least squares estimate** of regression coefficients

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

- **Multivariate normal distribution** $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with pdf

$$f(\boldsymbol{x}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)$$

- What's the deal with all the bold letters? **Linear algebra**.

- Worth the trip. Very useful for linear regression and more.

Vectors

- Linear algebra: a **vector** is an object containing **real numbers**

$$\boldsymbol{a} = \begin{pmatrix} 1 \\ 3 \\ 5 \\ 3 \end{pmatrix}$$

- Common default: a vector is a **column vector**.
- The **transpose** of a vector is a row vector

$$\boldsymbol{a}^\top = \begin{pmatrix} 1 & 3 & 5 & 3 \end{pmatrix}$$

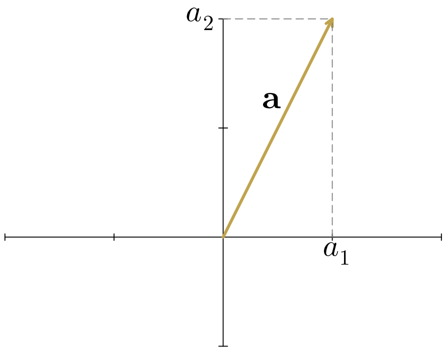
- R:

```
> a = c(1,2,5,3)
> t(a) # transpose
```

Visualizing vectors in 2D

- **2D vector**. Directed line (arrow) in $\mathbb{R}^2$.

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

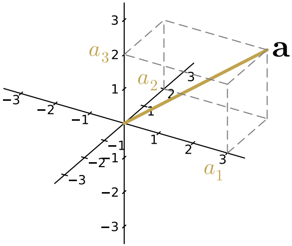


Visualizing vectors in 3D

- **3D vector**. Directed line (arrow) in $\mathbb{R}^3$.

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ 2 \end{pmatrix}$$

$$\mathbf{a} = (3, 2, 2)^\top$$



Vector addition and subtraction

- **Adding two vectors** with the same number of elements

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \quad \boldsymbol{a} + \boldsymbol{b} = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \\ a_3 + b_3 \end{pmatrix}$$

- **Substracting** a vector from another vector

$$\boldsymbol{a} - \boldsymbol{b} = \begin{pmatrix} a_1 - b_1 \\ a_2 - b_2 \\ a_3 - b_3 \end{pmatrix}$$

- Both these operations can be visualized geometrically.

Vector multiplication

- In R a*b will do **elementwise multiplication**

$$\boldsymbol{a} * \boldsymbol{b} = \begin{pmatrix} a_1 b_1 \\ a_2 b_2 \\ a_3 b_3 \end{pmatrix}$$

- In a%*%b will compute the **dot product**

$$\boldsymbol{a} \cdot \boldsymbol{b} = \boldsymbol{a}^\top \boldsymbol{b} = \begin{pmatrix} a_1 & a_2 & a_3 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = a_1 b_1 + a_2 b_2 + a_3 b_3$$

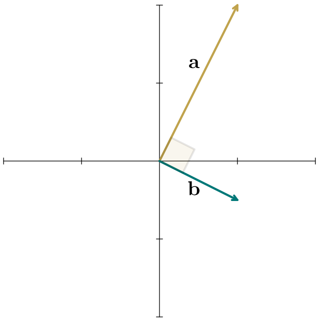
- In general the **dot product** is

$$\boldsymbol{a} \cdot \boldsymbol{b} = \sum_{i=1}^{n} a_i b_i$$

Orthogonal vectors

- Two vectors are **orthogonal** if their dot product is zero

$$\boldsymbol{a} \cdot \boldsymbol{b} = 0$$



- Example in 3D:

$$\boldsymbol{a} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} -6 \\ 2 \\ 2 \end{pmatrix}$$

Matrices

- A **matrix** is like a table, it has **rows** and **columns**

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

- This is a 2×3 matrix since it has 2 rows and 3 columns.
- View a $p \times q$ matrix as q column vector stacked horizontally

$$\boldsymbol{X} = \begin{pmatrix} | & | & & | \\ \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_q \\ | & | & & | \end{pmatrix}$$

- Example: the following three vectors give the matrix above

$$\boldsymbol{x}_1 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \quad \boldsymbol{x}_2 = \begin{pmatrix} 3 \\ 2 \end{pmatrix}, \quad \boldsymbol{x}_3 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

```
> x1 = c(2,3); x2 = c(3,2); x3=c(1,0);
> cbind(x1,x2,x3) # column bind. Also rbind exists
```

Matrix-Vector multiplication

- $\boldsymbol{A}$ is an $m \times n$ matrix $\boldsymbol{A}$
- $\boldsymbol{b}$ is an n-element vector
- **Matrix-vector product**: dot product of each row in $\boldsymbol{A}$ with $\boldsymbol{b}$

$$\underset{(m\times n)}{\boldsymbol{A}} = \begin{pmatrix} - & \boldsymbol{a}_1^\top & - \\ - & \boldsymbol{a}_2^\top & - \\ & \vdots & \\ - & \boldsymbol{a}_m^\top & - \end{pmatrix} \qquad \underset{(n\times 1)}{\boldsymbol{b}} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

$$\boldsymbol{A}\boldsymbol{b} = \begin{pmatrix} \boldsymbol{a}_1^\top \boldsymbol{b} \\ \boldsymbol{a}_2^\top \boldsymbol{b} \\ \vdots \\ \boldsymbol{a}_m^\top \boldsymbol{b} \end{pmatrix}$$

Matrix-Matrix multiplication

- Matrix product of $\boldsymbol{A}$ and $\boldsymbol{B}$: pairwise dot product of a row in $\boldsymbol{A}$ and a column in $\boldsymbol{B}$

$$\underset{(m\times n)}{\boldsymbol{A}} = \begin{pmatrix} - & \boldsymbol{a}_1^\top & - \\ - & \boldsymbol{a}_2^\top & - \\ & \vdots & \\ - & \boldsymbol{a}_m^\top & - \end{pmatrix} \qquad \underset{(n\times r)}{\boldsymbol{B}} = \begin{pmatrix} | & | & & | \\ \boldsymbol{b}_1 & \boldsymbol{b}_2 & \cdots & \boldsymbol{b}_r \\ | & | & & | \end{pmatrix}$$

$$\boldsymbol{AB} = \begin{pmatrix} \boldsymbol{a}_1^\top \boldsymbol{b}_1 & \boldsymbol{a}_1^\top \boldsymbol{b}_2 & \cdots & \boldsymbol{a}_1^\top \boldsymbol{b}_r \\ \boldsymbol{a}_2^\top \boldsymbol{b}_1 & \boldsymbol{a}_2^\top \boldsymbol{b}_2 & \cdots & \boldsymbol{a}_2^\top \boldsymbol{b}_r \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{a}_m^\top \boldsymbol{b}_1 & \boldsymbol{a}_m^\top \boldsymbol{b}_2 & \cdots & \boldsymbol{a}_m^\top \boldsymbol{b}_r \end{pmatrix}$$

Matrix-Matrix multiplication

- Example

$$\boldsymbol{A} = \begin{pmatrix} 2 & 3 \\ 3 & 2 \end{pmatrix}, \qquad \boldsymbol{B} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$$

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

$$= \begin{pmatrix} 2 \cdot 1 + 3 \cdot 0 = 2 & 2 \cdot 2 + 3 \cdot 1 = 7 \\ 3 \cdot 1 + 2 \cdot 0 = 3 & 3 \cdot 2 + 2 \cdot 1 = 8 \end{pmatrix}$$

```
> A = matrix(c(2,3,3,2), 2, 2, byrow = TRUE)
> B = matrix(c(1,2,0,1), 2, 2, byrow = TRUE)
> A%*%B # A*B would do elementwise multiplication
```

Linear regression - one observation

- One observation

$$y = \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

- In vector form

$$y = \underbrace{\begin{pmatrix} x_1 & \cdots & x_p \end{pmatrix}}_{\boldsymbol{x}^\top} \underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}_{\boldsymbol{\beta}} + \varepsilon = \boldsymbol{x}^\top \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- Add a one for the **intercept**

$$\begin{pmatrix} 1 & x_1 & \cdots & x_p \end{pmatrix} \qquad \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

- The ith observation

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

Linear regression - all observations

- The ith observation

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

- All $i = 1, \dots n$ observations stacked under each other

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^\top \boldsymbol{\beta} \\ \boldsymbol{x}_2^\top \boldsymbol{\beta} \\ \vdots \\ \boldsymbol{x}_n^\top \boldsymbol{\beta} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- With matrix-vector multiplication

$$\begin{pmatrix} \boldsymbol{x}_1^\top \boldsymbol{\beta} \\ \boldsymbol{x}_2^\top \boldsymbol{\beta} \\ \vdots \\ \boldsymbol{x}_n^\top \boldsymbol{\beta} \end{pmatrix} = \underbrace{\begin{pmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix}}_{\boldsymbol{X}} \boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta}$$

- $\boldsymbol{X}$ is the $n \times p$ **covariate matrix** with n observations as rows.

Linear regression

- **Linear regression in vector form**

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

- **Least squares estimate** = maximum likelihood estimate

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

- We now understand that

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{pmatrix} \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} & \cdots & \sum_{i=1}^n x_{1i}x_{pi} \\ \sum_{i=1}^n x_{1i}x_{2i} & \sum_{i=1}^n x_{2i}^2 & \cdots & \sum_{i=1}^n x_{2i}x_{pi} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{1i}x_{pi} & \sum_{i=1}^n x_{2i}x_{pi} & \cdots & \sum_{i=1}^n x_{pi}^2 \end{pmatrix}$$

$$\boldsymbol{X}^\top \boldsymbol{y} = \begin{pmatrix} \sum_{i=1}^n x_{1i}y_i \\ \sum_{i=1}^n x_{2i}y_i \\ \vdots \\ \sum_{i=1}^n x_{pi}y_i \end{pmatrix}$$

- But what does $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ mean? Inverse of a matrix? 🤯

Matrix inverse

- The inverse of regular number x is x^{-1} which is defined by

$$x^{-1}x = xx^{-1} = \frac{x}{x} = 1$$

- **Inverse of $p \times p$ matrix $\boldsymbol{A}$** is denoted by $\boldsymbol{A}^{-1}$ and defined by

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}_p$$

where $\boldsymbol{I}_p$ is the $p \times p$ **identity matrix**

$$\boldsymbol{I}_p = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

```
> A = matrix(c(2,3,3,2), 2, 2, byrow = TRUE)
> invA = solve(A)
> invA %*% A # returns the identity matrix
```

Least squares estimate

- Least squares minimizes the sum of squared residuals

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Find minimum of $Q(\beta_0, \beta_1)$ by solving system of equations

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1) = 0$$

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0$$

  gives the so called normal equations

$$n\bar{y} = n\beta_0 + \beta_1 n\bar{x}$$

$$\sum_{i=1}^{n} x_i y_i = \beta_0 n\bar{x} + \beta_1 \sum_{i=1}^{n} x_i^2$$

- With solution

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Least squares estimate - vector form

- Sum of squared residuals in vector notation

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - x_i^\top \boldsymbol{\beta})^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

- Set **gradient** vector equal to zero

$$\frac{\partial}{\partial \boldsymbol{\beta}} Q(\boldsymbol{\beta}) = -2\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}$$

gives the normal equations

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

- Multiply both sides with the matrix inverse of $\boldsymbol{X}^\top \boldsymbol{X}$

$$\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

gives the least squares solution

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

Gradients

- Bivariate function $z = f(x, y)$.
- **Partial derivative in x**: change in x, *holding y constant*

$$f_x(x, y) = \frac{\partial}{\partial x} f(x, y)$$

- **Partial derivative in y**: change in y, *holding x constant*

$$f_y(x, y) = \frac{\partial}{\partial y} f(x, y)$$

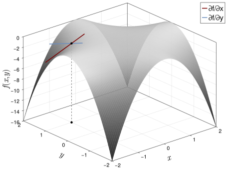
- **Gradient** is the vector of partial derivatives

$$\begin{pmatrix} f_x(x, y) \\ f_y(x, y) \end{pmatrix}$$

- General $f(x_1, \ldots, x_p)$ or $f(\boldsymbol{x})$. **Gradient** is p-dim vector

$$\frac{\boldsymbol{\partial}}{\boldsymbol{\partial x}} = \begin{pmatrix} \frac{\partial}{\partial x_1} f(\boldsymbol{x}) \\ \vdots \\ \frac{\partial}{\partial x_p} f(\boldsymbol{x}) \end{pmatrix}$$

Gradients



Determinant of a square matrix

- Let $\boldsymbol{A}$ be a 2×2 matrix

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

- The **determinant** is the number

$$|\boldsymbol{A}| = a_{11}a_{22} - a_{12}a_{21}$$

- Better intuition soon on why the determinant is important.
- Formulas for larger matrices are complicated. Use a computer.

```
> A = matrix(c(2,3,3,2), 2, 2)
> det(A) # returns -5
```

Bivariate normal distribution

- X and Y follow a **bivariate normal distribution**

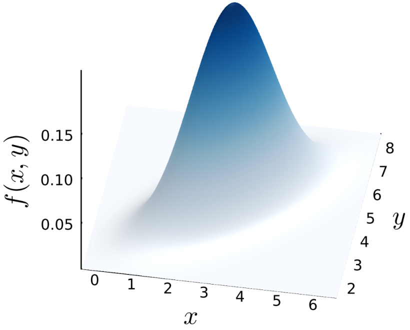
$$(X, Y) \sim N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$$

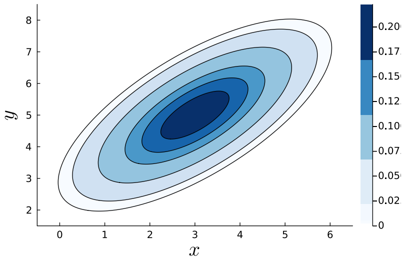
with **joint pdf**

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]\right)$$

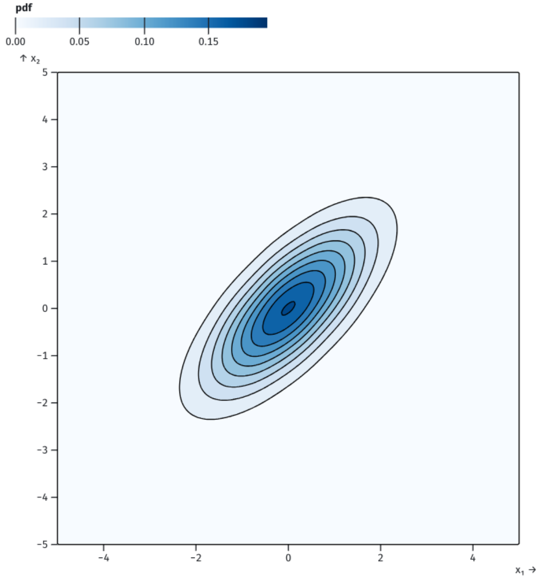
- Parameters:
  - μ_x the mean of X
  - μ_y the mean of Y
  - σ_x the standard deviation of X
  - σ_y the standard deviation of Y
  - ρ the correlation between X and Y

Bivariate normal distribution





Bivariate normal - widget



Properties bivariate normal distribution

- Let X and Y follow a bivariate normal distribution

$$(X, Y) \sim N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$$

- The **marginal density for X** is also normal

$$X \sim N(\mu_x, \sigma_x^2)$$

with the same parameters as those in the bivariate normal.

- The **marginal density for Y** is also normal

$$Y \sim N(\mu_y, \sigma_y^2)$$

- **Conditional densities** $f_{Y|X}(y)$ and $f_{X|Y}(x)$ are normal, see wikipedia, if you are curious.

Multivariate normal distribution

- $\boldsymbol{x} = (X_1, X_2, \ldots, X_p)^\top$ and follows a **multivariate normal distribution**

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

  with **joint pdf**

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- Clash in notation: *small* bold letters for random vectors.
- Parameters when $p = 2$:
  - **Mean vector**

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

  - **Covariance matrix**

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Multivariate normal distribution

- Determinant measures **total variance**

$$|\mathbf{\Sigma}| = \sigma_1^2\sigma_2^2(1-\rho^2)$$

  - No correlation: $|\mathbf{\Sigma}| = \sigma_1^2\sigma_2^2$
  - Strong positive correlation: $|\mathbf{\Sigma}|$ small
  - Strong negative correlation: $|\mathbf{\Sigma}|$ small

- The quadratic form

$$(\boldsymbol{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

is the vector version of a squared standardized variable

$$\left(\frac{X-\mu}{\sigma}\right)^2$$