Linear Systems and Linear Regression

From Linear Systems to Regression

- In practice, data come as **observations**
- Each observation may include:
  - a single input x_i, or
  - a vector of inputs (features)
- The goal: explain or predict an output variable

From Linear Systems to Regression

- In practice, data come as **observations**
- Each observation may include:
  - a single input x_i, or
  - a vector of inputs (features)
- The goal: explain or predict an output variable

Key idea

Linear regression is the extension of linear systems to data with noise and many observations.

A Biological Example

We want to understand how several cellular quantities influence gene expression.

A Biological Example

We want to understand how several cellular quantities influence gene expression.
For each biological sample i, we measure:

- transcription factor TF1,
- transcription factor TF2,
- ATP level (metabolic state),
- target gene expression y_i.

A Biological Example

We want to understand how several cellular quantities influence gene expression.
For each biological sample i, we measure:

- transcription factor TF1,
- transcription factor TF2,
- ATP level (metabolic state),
- target gene expression y_i.

We represent each sample as a feature vector:

$$x_i = (\mathrm{TF1}_i,\ \mathrm{TF2}_i,\ \mathrm{ATP}_i)$$

Linear Regression Model

We assume that gene expression depends *approximately linearly* on the features:

$$y_i \approx c_1\,\mathsf{TF1}_i + c_2\,\mathsf{TF2}_i + c_3\,\mathsf{ATP}_i + c_4$$

Linear Regression Model

We assume that gene expression depends *approximately linearly* on the features:

$$y_i \approx c_1 \, \mathsf{TF1}_i + c_2 \, \mathsf{TF2}_i + c_3 \, \mathsf{ATP}_i + c_4$$

- c_1, c_2, c_3: influence of biological factors
- c_4: baseline expression

Linear Regression Model

We assume that gene expression depends *approximately linearly* on the features:

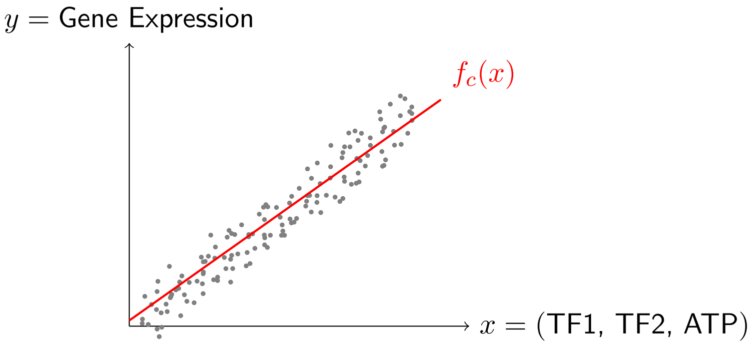
$$y_i \approx c_1\,\mathsf{TF1}_i + c_2\,\mathsf{TF2}_i + c_3\,\mathsf{ATP}_i + c_4$$

- c_1, c_2, c_3: influence of biological factors
- c_4: baseline expression

Interpretation

Linear regression estimates how much each factor contributes to the output.

Schematic View of Linear Regression



Linear Model with Nonlinear Features

A model is **linear** if it is linear in the parameters, even if features are nonlinear.

Linear Model with Nonlinear Features

A model is **linear** if it is linear in the parameters, even if features are nonlinear.
Example:

$$f_c(x) = c_1 \sin(x) + c_2 x^2$$

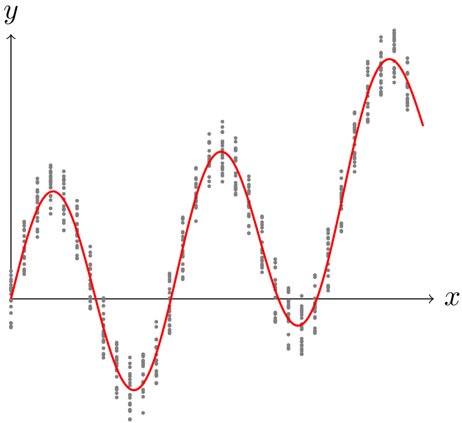
Linear Model with Nonlinear Features

A model is **linear** if it is linear in the parameters, even if features are nonlinear.
Example:

$$f_c(x) = c_1 \sin(x) + c_2 x^2$$

- $\sin(x)$ and x^2 are nonlinear
- coefficients c_1, c_2 enter linearly

Example: Nonlinear Features, Linear Regression



Problem Setup

We consider:

- m observations (x_i, y_i)
- each input vector:

$$x_i = (x_{i1}, \dots, x_{ip}) \in \mathbb{R}^p$$

- output $y_i \in \mathbb{R}$

Problem Setup

We consider:

- m observations (x_i, y_i)
- each input vector:

$$x_i = (x_{i1}, \dots, x_{ip}) \in \mathbb{R}^p$$

- output $y_i \in \mathbb{R}$

Goal

Find coefficients that best predict y_i from x_i.

Feature Transformations

Features can be arbitrary transformations:

$$\phi_1(x_i),\ \phi_2(x_i),\ \ldots,\ \phi_n(x_i)$$

Feature Transformations

Features can be arbitrary transformations:

$$\phi_1(x_i),\ \phi_2(x_i),\ \ldots,\ \phi_n(x_i)$$

Examples:

- raw variables: x_{i1}, x_{i2}
- polynomials: $x_{i1}^2, x_{i1}x_{i2}$
- nonlinear: $\log(x_{i2}+1), \sin(x_{i1})$

Feature Transformations

Features can be arbitrary transformations:

$$\phi_1(x_i),\ \phi_2(x_i),\ \ldots,\ \phi_n(x_i)$$

Examples:

- raw variables: x_{i1}, x_{i2}
- polynomials: $x_{i1}^2, x_{i1}x_{i2}$
- nonlinear: $\log(x_{i2}+1), \sin(x_{i1})$

Key message

Nonlinear behavior can be modeled with linear regression by choosing suitable features.

Objective of Linear Regression

Model:

$$f_c(x_i) = \sum_{j=1}^{n} c_j \, \phi_j(x_i)$$

Objective of Linear Regression

Model:

$$f_c(x_i) = \sum_{j=1}^{n} c_j \, \phi_j(x_i)$$

Choose c to approximate y_i:

$$y_i \approx f_c(x_i)$$

Objective of Linear Regression

Model:

$$f_c(x_i) = \sum_{j=1}^{n} c_j \, \phi_j(x_i)$$

Choose c to approximate y_i:

$$y_i \approx f_c(x_i)$$

Least squares criterion

$$SSE = \sum_{i=1}^{m} (y_i - f_c(x_i))^2$$

Matrix Formulation

Design matrix:

$$A = \begin{pmatrix} \phi_1(x_1) & \cdots & \phi_n(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_m) & \cdots & \phi_n(x_m) \end{pmatrix}$$

Matrix Formulation

Design matrix:

$$A = \begin{pmatrix} \phi_1(x_1) & \cdots & \phi_n(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_m) & \cdots & \phi_n(x_m) \end{pmatrix}$$

Vectors:

$$c = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}, \qquad y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$$

Matrix Formulation

Design matrix:

$$A = \begin{pmatrix} \phi_1(x_1) & \cdots & \phi_n(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_m) & \cdots & \phi_n(x_m) \end{pmatrix}$$

Vectors:

$$c = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}, \qquad y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$$

Compact form

$$\boxed{Ac \approx y}$$

Interpretation of $Ac \approx y$

- Each row of A corresponds to one observation
- Each column corresponds to one feature
- c weights the importance of each feature

Interpretation of $Ac \approx y$

- Each row of A corresponds to one observation
- Each column corresponds to one feature
- c weights the importance of each feature

Special cases:

- $Ac = y$: exact fit
- $m > n$: overdetermined (least squares)
- $m < n$: underdetermined (many solutions)

Summary

Key ideas

- Linear regression generalizes linear systems to data
- Models are linear in parameters, not necessarily in inputs
- Feature engineering controls model expressiveness

Summary

Key ideas

- Linear regression generalizes linear systems to data
- Models are linear in parameters, not necessarily in inputs
- Feature engineering controls model expressiveness

Matrix view

Regression reduces to solving

$$Ac \approx y$$

which connects directly to numerical linear algebra.