3F8: Inference

Regression

José Miguel Hernández–Lobato and Richard E. Turner
Department of Engineering
University of Cambridge

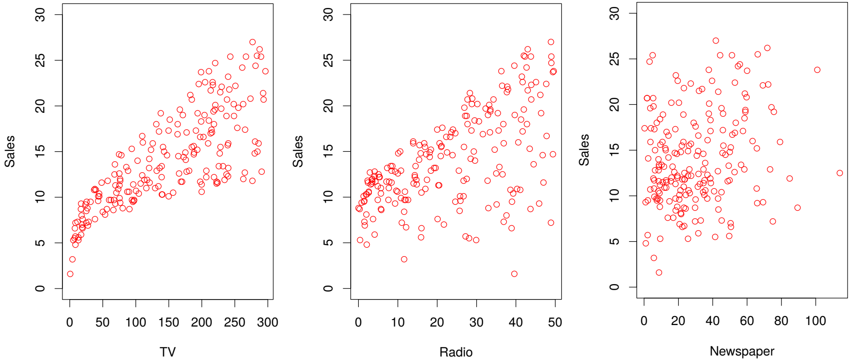
Lent Term

What is regression?

A **type of problem** in machine learning requiring

- to identify **patterns** and **regularities** between **input variables** and a corresponding continuous **output variable**,
- from a **training data set** $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ formed by pairs of input vectors $\mathbf{x}_n = (x_{n,1}, \ldots, x_{n,D})^{\mathsf{T}}$ and corresponding output values $y_n \in \mathbb{R}$.

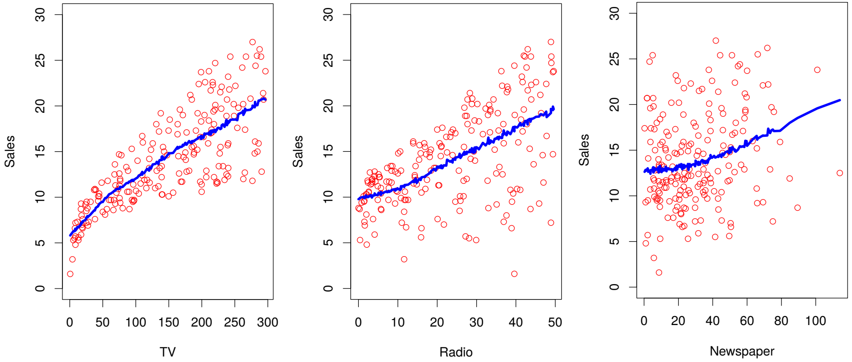
Example



Advertising data: sales, in thousands of units, as a function of TV, radio ,and newspaper budgets for 200 different markets.

G. James, D. Witten, T. Hastie and R. Tibshirani. *An Introduction to statistical learning*, 2013.

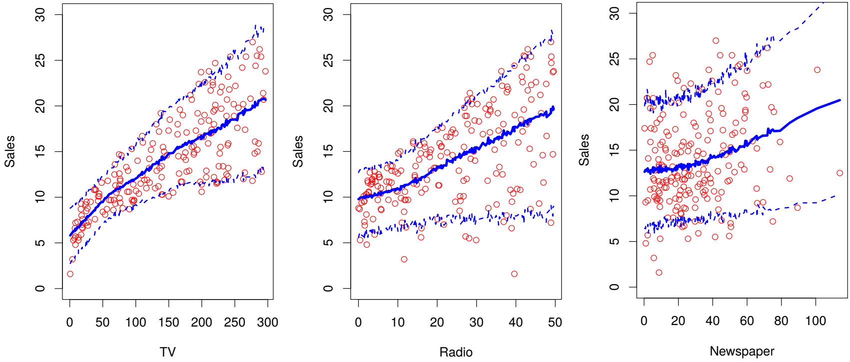
Example



Advertising data: sales, in thousands of units, as a function of TV, radio ,and newspaper budgets for 200 different markets.

G. James, D. Witten, T. Hastie and R. Tibshirani. *An Introduction to statistical learning*, 2013.

Example

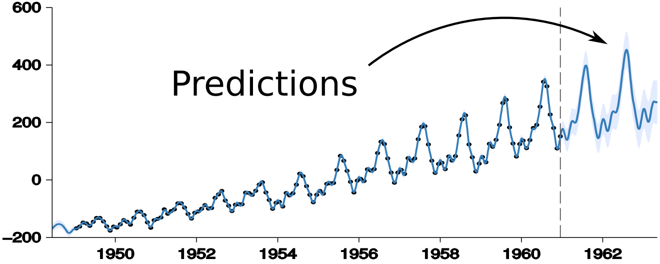


Advertising data: sales, in thousands of units, as a function of TV, radio ,and newspaper budgets for 200 different markets.

G. James, D. Witten, T. Hastie and R. Tibshirani. *An Introduction to statistical learning*, 2013.

Goals of regression

- **Prediction:** we want to predict the output for a new input vector or **test data point**, e.g., airline passenger data:



Structure Discovery in Nonparametric Regression through Compositional Kernel Search
D. Duvenaud, J. R. Lloyd, R. Grosse, J. B. Tenenbaum, Z. Ghahramani. In ICML, 2013

- **Interpretation:** what is the relationship between inputs and outputs?, e.g., which media has the strongest effect on sales?

Interpretation example

2.2 Component 2 : An approximately periodic function with a period of 1.0 years and with linearly increasing amplitude

This component is approximately periodic with a period of 1.0 years and varying amplitude. Across periods the shape of this function varies very smoothly. The amplitude of the function increases linearly. The shape of this function within each period has a typical lengthscale of 6.0 weeks.

This component explains 89.9% of the residual variance; this increases the total variance explained from 85.4% to 98.5%. The addition of this component reduces the cross validated MAE by 63.45% from 34.03 to 12.44.

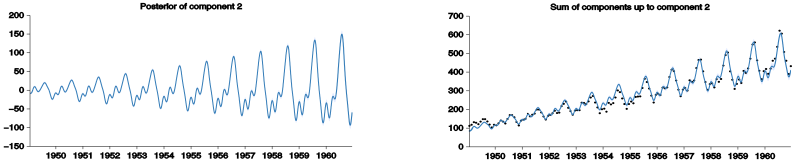
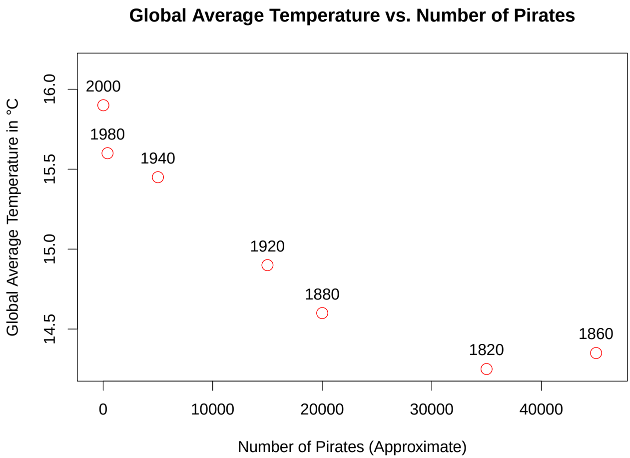


Figure 4: Pointwise posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

Automatic Construction and Natural-Language Description of Nonparametric Regression Models
J. R. Lloyd, D. Duvenaud, R. Grosse, J. B. Tenenbaum, Z. Ghahramani. In AAAI, 2014

Statistical dependence $\neq$ causation

Machine learning methods for regression find **statistical dependencies** and not **causal relationships**.

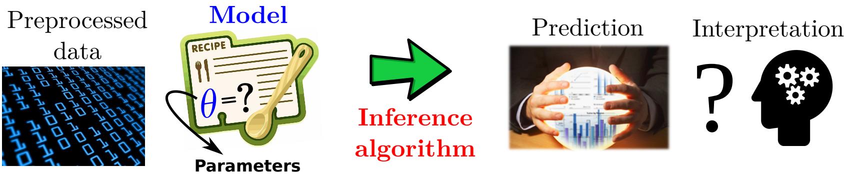


How to solve a regression problem?

A general approach:

- A **model** is used to encode, as function of some **parameters** $\boldsymbol{\theta}$, the type of patterns that we expect to find in the data.

- The **inference algorithm** combines model and data to make predictions and produce interpretations.



The model is a **probability distribution** which we assume to have generated the data.

Linear regression model

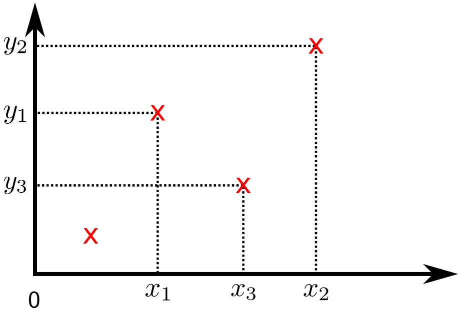
The relationship between inputs $\mathbf{x}_n$ and outputs y_n in $p(y_n|\mathbf{x}_n, \boldsymbol{\theta})$ is **linear**.

Why linear? Simple, easy to understand, widely used, easily generalized.

Linear regression model

The relationship between inputs $\mathbf{x}_n$ and outputs y_n in $p(y_n|\mathbf{x}_n, \boldsymbol{\theta})$ is **linear**.

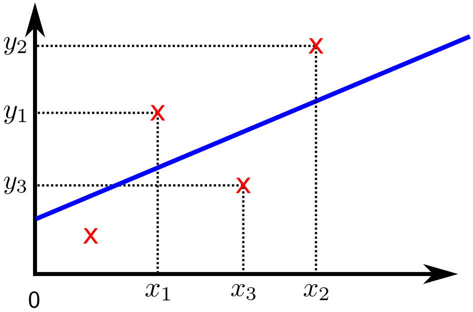
Why linear? Simple, easy to understand, widely used, easily generalized.



Linear regression model

The relationship between inputs $\mathbf{x}_n$ and outputs y_n in $p(y_n|\mathbf{x}_n, \boldsymbol{\theta})$ is **linear**.

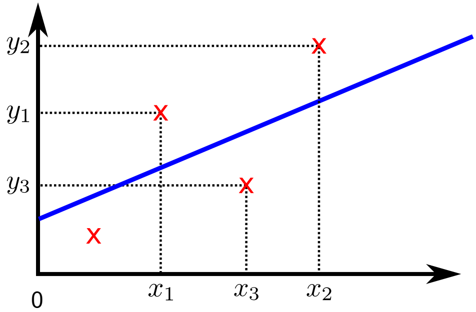
Why linear? Simple, easy to understand, widely used, easily generalized.



Linear regression model

The relationship between inputs $\mathbf{x}_n$ and outputs y_n in $p(y_n|\mathbf{x}_n, \boldsymbol{\theta})$ is **linear**.

Why linear? Simple, easy to understand, widely used, easily generalized.

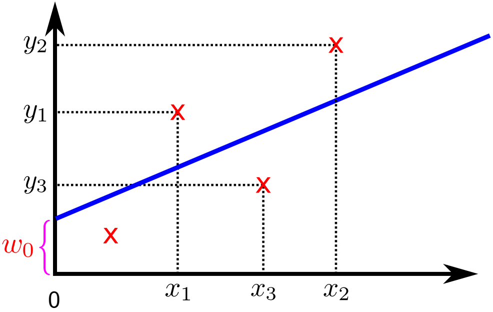


$y_n = w_0 + w_1 x_n$

Linear regression model

The relationship between inputs $\mathbf{x}_n$ and outputs y_n in $p(y_n|\mathbf{x}_n, \boldsymbol{\theta})$ is **linear**.

Why linear? Simple, easy to understand, widely used, easily generalized.

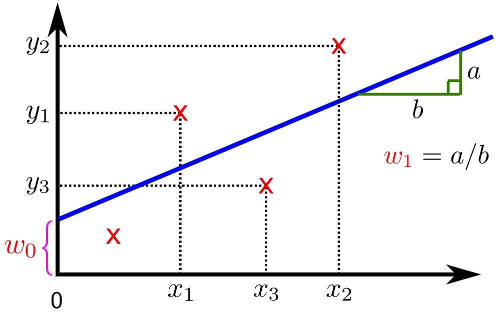


$$y_n = w_0 + w_1 x_n$$

Linear regression model

The relationship between inputs $\mathbf{x}_n$ and outputs y_n in $p(y_n|\mathbf{x}_n, \boldsymbol{\theta})$ is **linear**.

Why linear? Simple, easy to understand, widely used, easily generalized.

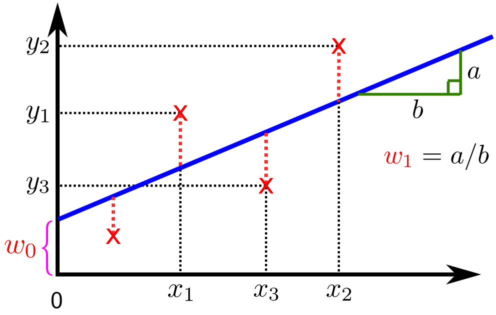


$$y_n = w_0 + w_1 x_n$$

Linear regression model

The relationship between inputs $\mathbf{x}_n$ and outputs y_n in $p(y_n|\mathbf{x}_n, \boldsymbol{\theta})$ is **linear**.

Why linear? Simple, easy to understand, widely used, easily generalized.

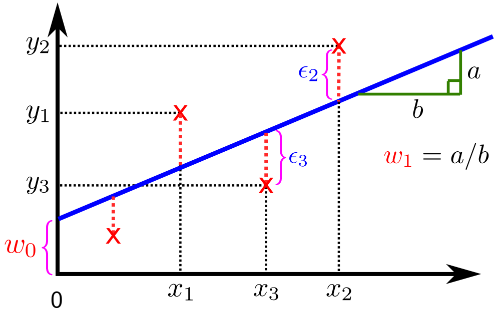


$$y_n = w_0 + w_1 x_n$$

Linear regression model

The relationship between inputs $\mathbf{x}_n$ and outputs y_n in $p(y_n|\mathbf{x}_n, \boldsymbol{\theta})$ is **linear**.

Why linear? Simple, easy to understand, widely used, easily generalized.

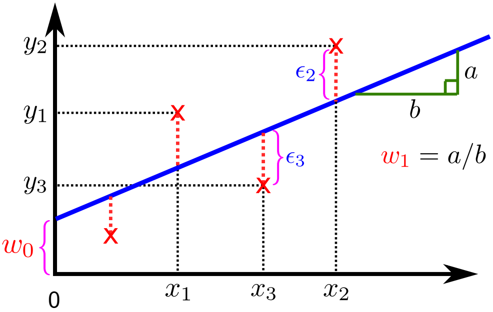


$y_n = w_0 + w_1 x_n$

Linear regression model

The relationship between inputs $\mathbf{x}_n$ and outputs y_n in $p(y_n|\mathbf{x}_n, \boldsymbol{\theta})$ is **linear**.

Why linear? Simple, easy to understand, widely used, easily generalized.

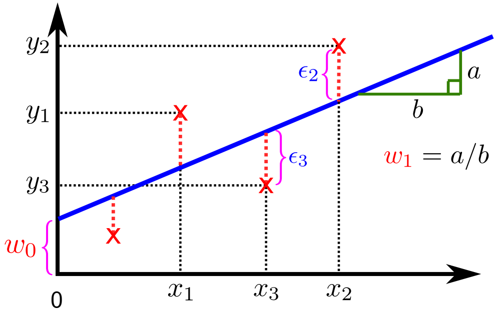


$$y_n = w_0 + w_1 x_n + \epsilon_n$$

Linear regression model

The relationship between inputs $\mathbf{x}_n$ and outputs y_n in $p(y_n|\mathbf{x}_n, \boldsymbol{\theta})$ is **linear**.

Why linear? Simple, easy to understand, widely used, easily generalized.



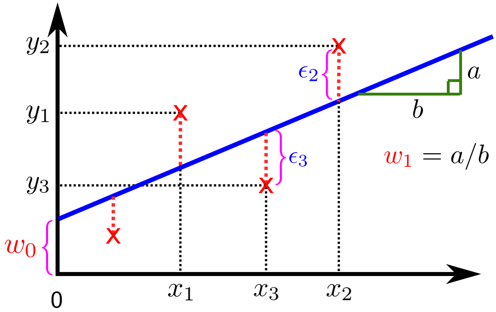
$y_n = w_0 + w_1 x_n + \epsilon_n$

What should the distribution of ϵ_n be?

Linear regression model

The relationship between inputs $\mathbf{x}_n$ and outputs y_n in $p(y_n|\mathbf{x}_n, \boldsymbol{\theta})$ is **linear**.

Why linear? Simple, easy to understand, widely used, easily generalized.



$$y_n = w_0 + w_1 x_n + \epsilon_n$$

What should the distribution of ϵ_n be?

$\epsilon_n \sim \mathcal{N}(0, \sigma^2)$ allows for tractable inference.

Linear regression model

Assuming

$$y_n = w_0 + w_1 x_n + \epsilon_n \,,$$
$$\epsilon_n \sim \mathcal{N}(0, \sigma^2) \,.$$

What is the form of $p(y_n | \mathbf{x}_n, \boldsymbol{\theta})$ with $\boldsymbol{\theta} = \{\sigma^2, w_0, w_1\}$?

Linear regression model

Assuming

$$y_n = w_0 + w_1 x_n + \epsilon_n \,,$$

$$\epsilon_n \sim \mathcal{N}(0, \sigma^2) \,.$$

What is the form of $p(y_n | \mathbf{x}_n, \boldsymbol{\theta})$ with $\boldsymbol{\theta} = \{\sigma^2, w_0, w_1\}$?

We have that

$$\mathbf{E}[y_n] = w_0 + w_1 x_n + \mathbf{E}[\epsilon_n] = w_0 + w_1 x_n \,,$$

$$\mathrm{Var}[y_n] = \mathbf{E}\left[(y_n - \mathbf{E}[y_n])^2\right] = \mathbf{E}\left[{\epsilon_n}^2\right] = \sigma^2 \,.$$

Linear regression model

Assuming

$$y_n = w_0 + w_1 x_n + \epsilon_n \,,$$

$$\epsilon_n \sim \mathcal{N}(0, \sigma^2)\,.$$

What is the form of $p(y_n|\mathbf{x}_n, \boldsymbol{\theta})$ with $\boldsymbol{\theta} = \{\sigma^2, w_0, w_1\}$?

We have that

$$\mathbf{E}[y_n] = w_0 + w_1 x_n + \mathbf{E}[\epsilon_n] = w_0 + w_1 x_n \,,$$
$$\mathrm{Var}[y_n] = \mathbf{E}\left[(y_n - \mathbf{E}[y_n])^2\right] = \mathbf{E}\left[{\epsilon_n}^2\right] = \sigma^2\,.$$

Therefore,

$$p(y_n|\mathbf{x}_n, \boldsymbol{\theta}) = \mathcal{N}(y_n | w_0 + w_1 x_n, \sigma^2)$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(y_n - w_0 - w_1 x_n)^2}{\sigma^2}\right\}\,.$$

Linear regression in higher dimensions

For a data set $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$, we assume

$$y_n = w_0 + w_1 x_{n,1} + \cdots + w_D x_{n,D} + \epsilon_n$$

where $\boxed{\epsilon_n \sim \mathcal{N}(0, \sigma^2)}$ and $\widetilde{\mathbf{x}}_n = (1, \mathbf{x}_n)^\mathsf{T}$, $\boldsymbol{\theta} = \{\sigma^2, \mathbf{w}\}$.

Linear regression in higher dimensions

For a data set $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$, we assume

$$y_n = w_0 + w_1 x_{n,1} + \cdots + w_D x_{n,D} + \epsilon_n = [w_0, \ldots, w_D] \underbrace{\begin{bmatrix} 1 \\ x_{n,1} \\ \vdots \\ x_{n,D} \end{bmatrix}}_{\widetilde{\mathbf{x}}_n} + \epsilon_n$$

where $\boxed{\epsilon_n \sim \mathcal{N}(0, \sigma^2)}$ and $\widetilde{\mathbf{x}}_n = (1, \mathbf{x}_n)^\mathsf{T}$, $\boldsymbol{\theta} = \{\sigma^2, \mathbf{w}\}$.

Linear regression in higher dimensions

For a data set $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$, we assume

$$y_n = w_0 + w_1 x_{n,1} + \cdots + w_D x_{n,D} + \epsilon_n = [w_0, \ldots, w_D] \underbrace{\begin{bmatrix} 1 \\ x_{n,1} \\ \vdots \\ x_{n,D} \end{bmatrix}}_{\widetilde{\mathbf{x}}_n} + \epsilon_n = \mathbf{w}^\mathsf{T} \widetilde{\mathbf{x}}_n + \epsilon_n \,,$$

where $\boxed{\epsilon_n \sim \mathcal{N}(0, \sigma^2)}$ and $\widetilde{\mathbf{x}}_n = (1, \mathbf{x}_n)^\mathsf{T}$, $\boldsymbol{\theta} = \{\sigma^2, \mathbf{w}\}$.

Linear regression in higher dimensions

For a data set $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$, we assume

$$y_n = w_0 + w_1 x_{n,1} + \cdots + w_D x_{n,D} + \epsilon_n = [w_0, \ldots, w_D] \underbrace{\begin{bmatrix} 1 \\ x_{n,1} \\ \vdots \\ x_{n,D} \end{bmatrix}}_{\widetilde{\mathbf{x}}_n} + \epsilon_n = \mathbf{w}^{\mathsf{T}} \widetilde{\mathbf{x}}_n + \epsilon_n \,,$$

where $\boxed{\epsilon_n \sim \mathcal{N}(0, \sigma^2)}$ and $\widetilde{\mathbf{x}}_n = (1, \mathbf{x}_n)^{\mathsf{T}}$, $\boldsymbol{\theta} = \{\sigma^2, \mathbf{w}\}$.

$$p(y_n | \widetilde{\mathbf{x}}_n, \boldsymbol{\theta}) =$$

Linear regression in higher dimensions

For a data set $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$, we assume

$$y_n = w_0 + w_1 x_{n,1} + \cdots + w_D x_{n,D} + \epsilon_n = [w_0, \ldots, w_D] \underbrace{\begin{bmatrix} 1 \\ x_{n,1} \\ \vdots \\ x_{n,D} \end{bmatrix}}_{\widetilde{\mathbf{x}}_n} + \epsilon_n = \mathbf{w}^\mathsf{T} \widetilde{\mathbf{x}}_n + \epsilon_n\,,$$

where $\boxed{\epsilon_n \sim \mathcal{N}(0, \sigma^2)}$ and $\widetilde{\mathbf{x}}_n = (1, \mathbf{x}_n)^\mathsf{T}$, $\boldsymbol{\theta} = \{\sigma^2, \mathbf{w}\}$.

$$p(y_n | \widetilde{\mathbf{x}}_n, \boldsymbol{\theta}) = \mathcal{N}(y_n | \mathbf{w}^\mathsf{T} \widetilde{\mathbf{x}}_n, \sigma^2)\,,$$

Linear regression in higher dimensions

For a data set $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$, we assume

$$y_n = w_0 + w_1 x_{n,1} + \cdots + w_D x_{n,D} + \epsilon_n = [w_0, \ldots, w_D] \underbrace{\begin{bmatrix} 1 \\ x_{n,1} \\ \vdots \\ x_{n,D} \end{bmatrix}}_{\widetilde{\mathbf{x}}_n} + \epsilon_n = \mathbf{w}^\mathsf{T} \widetilde{\mathbf{x}}_n + \epsilon_n \,,$$

where $\boxed{\epsilon_n \sim \mathcal{N}(0, \sigma^2)}$ and $\widetilde{\mathbf{x}}_n = (1, \mathbf{x}_n)^\mathsf{T}$, $\boldsymbol{\theta} = \{\sigma^2, \mathbf{w}\}$.

$$p(y_n | \widetilde{\mathbf{x}}_n, \boldsymbol{\theta}) = \mathcal{N}(y_n | \mathbf{w}^\mathsf{T} \widetilde{\mathbf{x}}_n, \sigma^2) \,,$$

Jargon for regression:

- $\mathbf{x}_n$ are the inputs, features, covariates, independent variables, etc.
- $\mathbf{y}_n$ are the outputs, responses, targets, dependent variables, etc.
- $\mathbf{w}$ are the coefficients, weights, etc. (w_0 is called the bias or intercept).
- ϵ_n are the errors, disturbances or noise.

Inference: Maximum Likelihood Estimate (MLE)

Maximize the **likelihood** function $p(y_1, \ldots, y_n | \widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

Inference: Maximum Likelihood Estimate (MLE)

Maximize the **likelihood** function $p(y_1, \ldots, y_n | \widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

In practice, it is the **log-likelihood** function what is maximized.

Inference: Maximum Likelihood Estimate (MLE)

Maximize the **likelihood** function $p(y_1, \ldots, y_n | \widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

In practice, it is the **log-likelihood** function what is maximized.

$$\mathcal{L}(\boldsymbol{\theta}) = \log p(y_1, \ldots, y_n | \widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n, \boldsymbol{\theta})$$

Inference: Maximum Likelihood Estimate (MLE)

Maximize the **likelihood** function $p(y_1, \ldots, y_n | \widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

In practice, it is the **log-likelihood** function what is maximized.

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \log p(y_1, \ldots, y_n | \widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n, \boldsymbol{\theta}) \\
&= \log \prod_{n=1}^{N} p(y_n | \widetilde{\mathbf{x}}_n, \boldsymbol{\theta}) =
\end{aligned}
$$

Inference: Maximum Likelihood Estimate (MLE)

Maximize the **likelihood** function $p(y_1, \ldots, y_n | \widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

In practice, it is the **log-likelihood** function what is maximized.

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \log p(y_1, \ldots, y_n | \widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n, \boldsymbol{\theta}) \\ &= \log \prod_{n=1}^{N} p(y_n | \widetilde{\mathbf{x}}_n, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log \mathcal{N}(y_n | \mathbf{w}^{\mathsf{T}} \widetilde{\mathbf{x}}_n, \sigma^2)\end{aligned}$$

Inference: Maximum Likelihood Estimate (MLE)

Maximize the **likelihood** function $p(y_1, \ldots, y_n | \widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

In practice, it is the **log-likelihood** function what is maximized.

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \log p(y_1, \ldots, y_n | \widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n, \boldsymbol{\theta}) \\
&= \log \prod_{n=1}^{N} p(y_n | \widetilde{\mathbf{x}}_n, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log \mathcal{N}(y_n | \mathbf{w}^{\mathsf{T}} \widetilde{\mathbf{x}}_n, \sigma^2) \\
&= \sum_{n=1}^{N} \left\{ -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y_n - \mathbf{w}^{\mathsf{T}} \widetilde{\mathbf{x}}_n)^2 \right\}
\end{aligned}
$$

Inference: Maximum Likelihood Estimate (MLE)

Maximize the **likelihood** function $p(y_1, \ldots, y_n|\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

In practice, it is the **log-likelihood** function what is maximized.

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \log p(y_1, \ldots, y_n|\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n, \boldsymbol{\theta}) \\
&= \log \prod_{n=1}^{N} p(y_n|\widetilde{\mathbf{x}}_n, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log \mathcal{N}(y_n|\mathbf{w}^{\mathsf{T}}\widetilde{\mathbf{x}}_n, \sigma^2) \\
&= \sum_{n=1}^{N} \left\{ -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_n - \mathbf{w}^{\mathsf{T}}\widetilde{\mathbf{x}}_n)^2 \right\} \\
&= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\mathbf{y} - \widetilde{\mathbf{X}}\mathbf{w}\right)^{\mathsf{T}}\left(\mathbf{y} - \widetilde{\mathbf{X}}\mathbf{w}\right)
\end{aligned}
$$

Inference: Maximum Likelihood Estimate (MLE)

Maximize the **likelihood** function $p(y_1, \ldots, y_n | \widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

In practice, it is the **log-likelihood** function what is maximized.

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \log p(y_1, \ldots, y_n | \widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n, \boldsymbol{\theta}) \\
&= \log \prod_{n=1}^{N} p(y_n | \widetilde{\mathbf{x}}_n, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log \mathcal{N}(y_n | \mathbf{w}^\mathsf{T} \widetilde{\mathbf{x}}_n, \sigma^2) \\
&= \sum_{n=1}^{N} \left\{ -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y_n - \mathbf{w}^\mathsf{T} \widetilde{\mathbf{x}}_n)^2 \right\} \\
&= -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right)^\mathsf{T} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right)
\end{aligned}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\mathsf{T}$, $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{x}}_1; \ldots; \widetilde{\mathbf{x}}_n)^\mathsf{T}$, and we have used $\mathbf{a}^\mathsf{T}\mathbf{a} = \sum_i a_i$.

Inference: Maximum Likelihood Estimate (MLE)

Maximize the **likelihood** function $p(y_1, \ldots, y_n | \widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

In practice, it is the **log-likelihood** function what is maximized.

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \log p(y_1, \ldots, y_n | \widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n, \boldsymbol{\theta}) \\
&= \log \prod_{n=1}^{N} p(y_n | \widetilde{\mathbf{x}}_n, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log \mathcal{N}(y_n | \mathbf{w}^{\mathsf{T}} \widetilde{\mathbf{x}}_n, \sigma^2) \\
&= \sum_{n=1}^{N} \left\{ -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y_n - \mathbf{w}^{\mathsf{T}} \widetilde{\mathbf{x}}_n)^2 \right\} \\
&= -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left(\mathbf{y} - \widetilde{\mathbf{X}}\mathbf{w} \right)^{\mathsf{T}} \left(\mathbf{y} - \widetilde{\mathbf{X}}\mathbf{w} \right) \\
&= -\frac{N}{2} \log(2\pi\sigma^2) - \frac{\mathbf{y}^{\mathsf{T}}\mathbf{y}}{2\sigma^2} - \underbrace{\frac{\mathbf{w}^{\mathsf{T}} \widetilde{\mathbf{X}}^{\mathsf{T}} \widetilde{\mathbf{X}} \mathbf{w}}{2\sigma^2}}_{\text{Quadratic term}} + \underbrace{\frac{\mathbf{y}^{\mathsf{T}} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2}}_{\text{Linear term}} .
\end{aligned}
$$

where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathsf{T}}$, $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{x}}_1; \ldots; \widetilde{\mathbf{x}}_n)^{\mathsf{T}}$, and we have used $\mathbf{a}^{\mathsf{T}}\mathbf{a} = \sum_i a_i$.

Vector calculus I

$$\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_D \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix}, \qquad \mathbf{a}^\mathsf{T}\mathbf{w} = \sum_{i=0}^{D} a_i w_i \,,$$

Vector calculus I

$$\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_D \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix}, \qquad \mathbf{a}^\mathsf{T}\mathbf{w} = \sum_{i=0}^{D} a_i w_i \,,$$

$$\frac{d[\mathbf{a}^\mathsf{T}\mathbf{w}]}{d\mathbf{w}} = \begin{bmatrix} \frac{d[\mathbf{a}^\mathsf{T}\mathbf{w}]}{dw_0} \\ \frac{d[\mathbf{a}^\mathsf{T}\mathbf{w}]}{dw_1} \\ \vdots \\ \frac{d[\mathbf{a}^\mathsf{T}\mathbf{w}]}{dw_D} \end{bmatrix}$$

Vector calculus I

$$\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_D \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix}, \qquad \mathbf{a}^\mathsf{T}\mathbf{w} = \sum_{i=0}^{D} a_i w_i \,,$$

$$\frac{d[\mathbf{a}^\mathsf{T}\mathbf{w}]}{d\mathbf{w}} = \begin{bmatrix} \frac{d[\mathbf{a}^\mathsf{T}\mathbf{w}]}{dw_0} \\ \frac{d[\mathbf{a}^\mathsf{T}\mathbf{w}]}{dw_1} \\ \vdots \\ \frac{d[\mathbf{a}^\mathsf{T}\mathbf{w}]}{dw_D} \end{bmatrix} = \mathbf{a}$$

Vector calculus I

$$\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_D \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix}, \qquad \mathbf{a}^\top \mathbf{w} = \sum_{i=0}^{D} a_i w_i \,,$$

$$\frac{d[\mathbf{a}^\top \mathbf{w}]}{d\mathbf{w}} = \begin{bmatrix} \frac{d[\mathbf{a}^\top \mathbf{w}]}{dw_0} \\ \frac{d[\mathbf{a}^\top \mathbf{w}]}{dw_1} \\ \vdots \\ \frac{d[\mathbf{a}^\top \mathbf{w}]}{dw_D} \end{bmatrix} = \mathbf{a} = \frac{d[\mathbf{w}^\top \mathbf{a}]}{d\mathbf{w}} \,.$$

Vector calculus II

Using the previous result

$$\frac{d[\mathbf{a}^\top \mathbf{w}]}{d\mathbf{w}} = \frac{d[\mathbf{w}^\top \mathbf{a}]}{d\mathbf{w}} = \mathbf{a}\,,$$

and the product rule of calculus,

$$\frac{d}{dx}[f(x)g(x)] = \underbrace{[\frac{d}{dx}f(x)]g(x)}_{g(x)\text{ as constant}} + \underbrace{f(x)[\frac{d}{dx}g(x)]}_{f(x)\text{ as constant}}\,, \tag{1}$$

we obtain

$$\frac{d[\mathbf{w}^\top \mathbf{A}\mathbf{w}]}{d\mathbf{w}} = \underbrace{\mathbf{A}^\top \mathbf{w}}_{\mathbf{w}^\top \mathbf{A}\text{ as constant}} + \underbrace{\mathbf{A}\mathbf{w}}_{\mathbf{A}\mathbf{w}\quad\text{as constant}}$$

Vector calculus II

Using the previous result

$$\frac{d[\mathbf{a}^\top \mathbf{w}]}{d\mathbf{w}} = \frac{d[\mathbf{w}^\top \mathbf{a}]}{d\mathbf{w}} = \mathbf{a}\,,$$

and the product rule of calculus,

$$\frac{d}{dx}[f(x)g(x)] = \underbrace{[\frac{d}{dx}f(x)]g(x)}_{g(x) \text{ as constant}} + \underbrace{f(x)[\frac{d}{dx}g(x)]}_{f(x) \text{ as constant}}\,, \tag{1}$$

we obtain

$$\frac{d[\mathbf{w}^\top \mathbf{A}\mathbf{w}]}{d\mathbf{w}} = \underbrace{\mathbf{A}^\top \mathbf{w}}_{\mathbf{w}^\top \mathbf{A} \text{ as constant}} + \underbrace{\mathbf{A}\mathbf{w}}_{\mathbf{A}\mathbf{w} \text{ as constant}} = 2\mathbf{A}\mathbf{w} \quad \text{if } \mathbf{A} \text{ symmetric}\,.$$

Finding the maximum likelihood parameters

The gradient of the log-likelihood at the maximizer is zero.

Finding the maximum likelihood parameters

The gradient of the log-likelihood at the maximizer is zero.

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left(-\frac{\mathbf{w}^\mathsf{T} \widetilde{\mathbf{X}}^\mathsf{T} \widetilde{\mathbf{X}} \mathbf{w}}{2\sigma^2} + \frac{\mathbf{y}^\mathsf{T} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2} \right)$$

.

Finding the maximum likelihood parameters

The gradient of the log-likelihood at the maximizer is zero.

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}} \left(-\frac{\mathbf{w}^\mathsf{T} \widetilde{\mathbf{X}}^\mathsf{T} \widetilde{\mathbf{X}} \mathbf{w}}{2\sigma^2} + \frac{\mathbf{y}^\mathsf{T} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2} \right) \\
&= -\frac{\widetilde{\mathbf{X}}^\mathsf{T} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2} + \frac{\widetilde{\mathbf{X}}^\mathsf{T} \mathbf{y}}{\sigma^2} = 0
\end{aligned}
$$

.

Finding the maximum likelihood parameters

The gradient of the log-likelihood at the maximizer is zero.

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}} \left(-\frac{\mathbf{w}^{\mathsf{T}} \widetilde{\mathbf{X}}^{\mathsf{T}} \widetilde{\mathbf{X}} \mathbf{w}}{2\sigma^2} + \frac{\mathbf{y}^{\mathsf{T}} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2} \right) \\
&= -\frac{\widetilde{\mathbf{X}}^{\mathsf{T}} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2} + \frac{\widetilde{\mathbf{X}}^{\mathsf{T}} \mathbf{y}}{\sigma^2} = 0 \Leftrightarrow
\end{aligned}
$$

.

Finding the maximum likelihood parameters

The gradient of the log-likelihood at the maximizer is zero.

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}} \left(-\frac{\mathbf{w}^\mathsf{T} \widetilde{\mathbf{X}}^\mathsf{T} \widetilde{\mathbf{X}} \mathbf{w}}{2\sigma^2} + \frac{\mathbf{y}^\mathsf{T} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2} \right) \\
&= -\frac{\widetilde{\mathbf{X}}^\mathsf{T} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2} + \frac{\widetilde{\mathbf{X}}^\mathsf{T} \mathbf{y}}{\sigma^2} = 0 \Leftrightarrow \mathbf{w} = \left(\widetilde{\mathbf{X}}^\mathsf{T} \widetilde{\mathbf{X}} \right)^{-1} \widetilde{\mathbf{X}}^\mathsf{T} \mathbf{y}\,.
\end{aligned}
$$

.

Finding the maximum likelihood parameters

The gradient of the log-likelihood at the maximizer is zero.

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left(-\frac{\mathbf{w}^{\mathsf{T}} \widetilde{\mathbf{X}}^{\mathsf{T}} \widetilde{\mathbf{X}} \mathbf{w}}{2\sigma^2} + \frac{\mathbf{y}^{\mathsf{T}} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2} \right)$$

$$= -\frac{\widetilde{\mathbf{X}}^{\mathsf{T}} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2} + \frac{\widetilde{\mathbf{X}}^{\mathsf{T}} \mathbf{y}}{\sigma^2} = 0 \Leftrightarrow \mathbf{w} = \left(\widetilde{\mathbf{X}}^{\mathsf{T}} \widetilde{\mathbf{X}} \right)^{-1} \widetilde{\mathbf{X}}^{\mathsf{T}} \mathbf{y}\,.$$

This is the Linear Least Squares Solution (LLSS).

.

Finding the maximum likelihood parameters

The gradient of the log-likelihood at the maximizer is zero.

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left(-\frac{\mathbf{w}^\mathsf{T} \widetilde{\mathbf{X}}^\mathsf{T} \widetilde{\mathbf{X}} \mathbf{w}}{2\sigma^2} + \frac{\mathbf{y}^\mathsf{T} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2} \right)$$

$$= -\frac{\widetilde{\mathbf{X}}^\mathsf{T} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2} + \frac{\widetilde{\mathbf{X}}^\mathsf{T} \mathbf{y}}{\sigma^2} = 0 \Leftrightarrow \mathbf{w} = \left(\widetilde{\mathbf{X}}^\mathsf{T} \widetilde{\mathbf{X}} \right)^{-1} \widetilde{\mathbf{X}}^\mathsf{T} \mathbf{y} \, .$$

This is the Linear Least Squares Solution (LLSS).

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2} \left(-\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right)^\mathsf{T} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right) \right)$$

.

Finding the maximum likelihood parameters

The gradient of the log-likelihood at the maximizer is zero.

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left(-\frac{\mathbf{w}^\mathsf{T} \widetilde{\mathbf{X}}^\mathsf{T} \widetilde{\mathbf{X}} \mathbf{w}}{2\sigma^2} + \frac{\mathbf{y}^\mathsf{T} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2} \right)$$
$$= -\frac{\widetilde{\mathbf{X}}^\mathsf{T} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2} + \frac{\widetilde{\mathbf{X}}^\mathsf{T} \mathbf{y}}{\sigma^2} = 0 \Leftrightarrow \mathbf{w} = \left(\widetilde{\mathbf{X}}^\mathsf{T} \widetilde{\mathbf{X}} \right)^{-1} \widetilde{\mathbf{X}}^\mathsf{T} \mathbf{y}\,.$$

This is the Linear Least Squares Solution (LLSS).

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2} \left(-\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right)^\mathsf{T} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right) \right)$$
$$= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right)^\mathsf{T} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right) = 0$$

.

Finding the maximum likelihood parameters

The gradient of the log-likelihood at the maximizer is zero.

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}} \left(-\frac{\mathbf{w}^{\mathsf{T}} \widetilde{\mathbf{X}}^{\mathsf{T}} \widetilde{\mathbf{X}} \mathbf{w}}{2\sigma^2} + \frac{\mathbf{y}^{\mathsf{T}} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2} \right) \\
&= -\frac{\widetilde{\mathbf{X}}^{\mathsf{T}} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2} + \frac{\widetilde{\mathbf{X}}^{\mathsf{T}} \mathbf{y}}{\sigma^2} = 0 \Leftrightarrow \mathbf{w} = \left(\widetilde{\mathbf{X}}^{\mathsf{T}} \widetilde{\mathbf{X}} \right)^{-1} \widetilde{\mathbf{X}}^{\mathsf{T}} \mathbf{y} \,.
\end{aligned}
$$

This is the Linear Least Squares Solution (LLSS).

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2} \left(-\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right)^{\mathsf{T}} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right) \right) \\
&= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right)^{\mathsf{T}} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right) = 0 \Leftrightarrow
\end{aligned}
$$

.

Finding the maximum likelihood parameters

The gradient of the log-likelihood at the maximizer is zero.

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left(-\frac{\mathbf{w}^\mathsf{T} \widetilde{\mathbf{X}}^\mathsf{T} \widetilde{\mathbf{X}} \mathbf{w}}{2\sigma^2} + \frac{\mathbf{y}^\mathsf{T} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2} \right)$$

$$= -\frac{\widetilde{\mathbf{X}}^\mathsf{T} \widetilde{\mathbf{X}} \mathbf{w}}{\sigma^2} + \frac{\widetilde{\mathbf{X}}^\mathsf{T} \mathbf{y}}{\sigma^2} = 0 \Leftrightarrow \mathbf{w} = \left(\widetilde{\mathbf{X}}^\mathsf{T} \widetilde{\mathbf{X}} \right)^{-1} \widetilde{\mathbf{X}}^\mathsf{T} \mathbf{y}\,.$$

This is the Linear Least Squares Solution (LLSS).

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2} \left(-\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right)^\mathsf{T} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right) \right)$$

$$= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right)^\mathsf{T} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right) = 0 \Leftrightarrow$$

$$\sigma^2 = \frac{1}{N} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right)^\mathsf{T} \left(\mathbf{y} - \widetilde{\mathbf{X}} \mathbf{w} \right)\,.$$

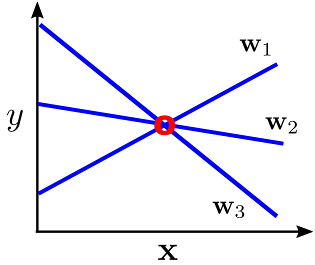
Problems of MLE

When $N < D + 1$ the MLE

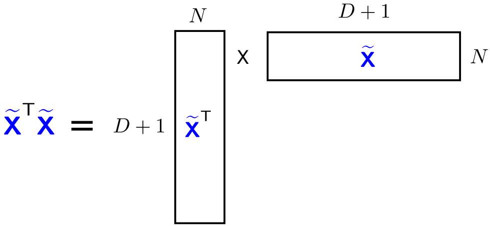
$$\mathbf{w} = \left(\widetilde{\mathbf{X}}^{\mathsf{T}}\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}^{\mathsf{T}}\mathbf{y}$$

is not defined. In this case...

Many values of $\mathbf{w}$ fit the training data equally well, achieving **zero error**.



The matrix $\widetilde{\mathbf{X}}^{\mathsf{T}}\widetilde{\mathbf{X}}$ is low rank and not invertible:



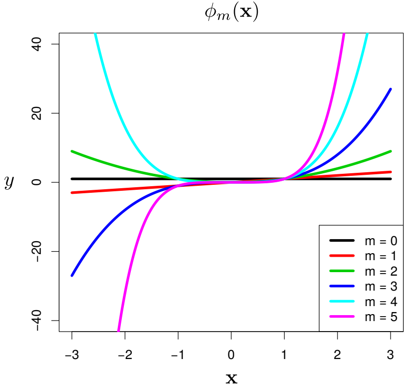
Non-linear (basis function) regression

Linear regression can model non-linear relationships by replacing $\mathbf{x}$ with some non-linear function of the inputs $\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_M(\mathbf{x}))^\mathsf{T}$.

Inference does not change, just replace each $\mathbf{x}_n$ with the new $\phi(\mathbf{x}_n)$.

Example, **polynomials** for 1D data $\phi_m(x) = x^m$, $m = 1, \ldots, M$:

$$
\begin{aligned}
M = 0\,, &\quad \phi(x) = [1]^\mathsf{T}\,,\\
M = 1\,, &\quad \phi(x) = [1, x]^\mathsf{T}\,,\\
M = 2\,, &\quad \phi(x) = [1, x, x^2]^\mathsf{T}\,,\\
M = 3\,, &\quad \phi(x) = [1, x, x^2, x^3]^\mathsf{T}\,,\\
M = 4\,, &\quad \phi(x) = [1, x, x^2, x^3, x^4]^\mathsf{T}\,,\\
M = 5\,, &\quad \phi(x) = [1, x, x^2, x^3, x^4, x^5]^\mathsf{T}\,,
\end{aligned}
$$



What should the value of M be?

1D example with polynomials

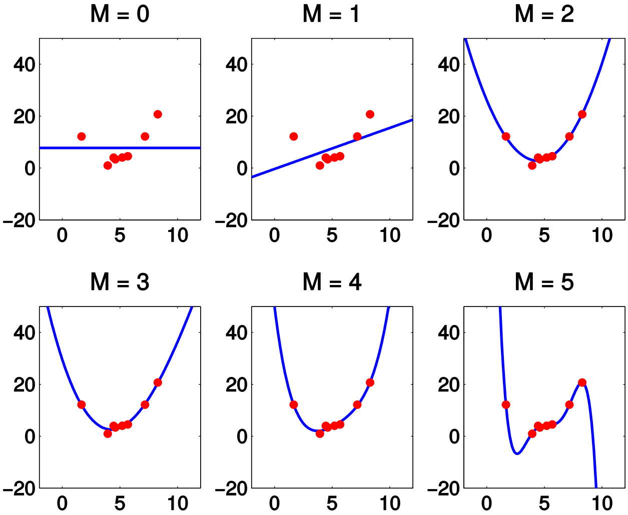
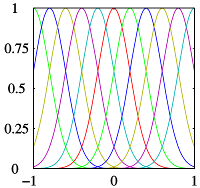


Figure: Z. Ghahramani.

Other basis functions

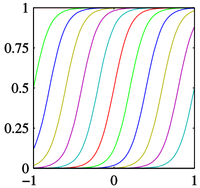
Gaussian radial basis functions with center $\mathbf{c}_m$ and width s:

$$\phi_m(\mathbf{x}) = \exp\left\{-\frac{1}{2}s(\mathbf{x},\mathbf{c}_m,s)^2\right\}$$
$$s(\mathbf{x},\mathbf{c}_m,s) = \sqrt{(\mathbf{x}-\mathbf{c}_m)^\mathsf{T}(\mathbf{x}-\mathbf{c}_m)/s^2}\,.$$



Sigmoidal basis functions:

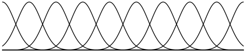
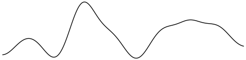
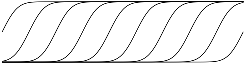
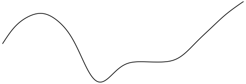
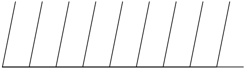
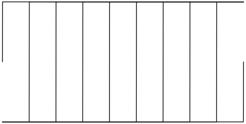
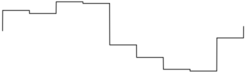
$$\phi_m(\mathbf{x}) = \sigma\left(s(\mathbf{x},\mathbf{c}_m,s)\right)$$
$$\sigma(x) = \frac{1}{1+\exp(-x)}\,.$$



They are uniformly spread in input space to capture non-linearities everywhere.

Figure: C. Bishop. *Pattern Recognition and Machine Learning*, 2006.

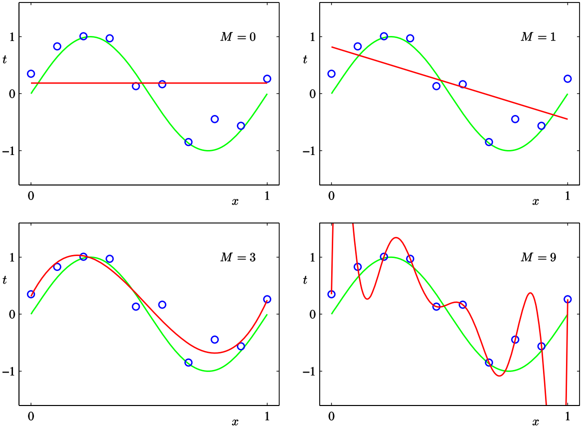
Examples

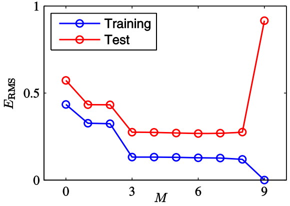
$\phi_m(\mathbf{x})$	Type of basis function	$\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$
$\mathcal{N}(x \mid c_m, 1)$	Gaussian 	
$\dfrac{1}{1+\exp(-x+c_m)}$	Sigmoidal 	
$\max(0, x-c_m)$	Truncated linear 	
$\operatorname{sign}(x-c_m)$	Step functions 	

$c_1 = -10, c_2 = -8, \ldots, c_9 = 8, c_{10} = 10, s = 1.$

Overfitting

A large number of basis functions can lead to **over-fitting**: the model fits the **training data** well but it performs poorly on new **test data**.





	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

Solution: use a prior distribution to enforce the entries of $\mathbf{w}$ to be small.

Figures and table: C. Bishop. *Pattern Recognition and Machine Learning*, 2006.