INFO 6105
Data Science Engineering Methods and Tools

Lecture 2
Linear Regression: A Probabilistic Approach

Ebrahim Nasrabadi
nasrabadi@northeastern.edu

College of Engineering
Northeastern University

Fall 2019

Recall from last lecture

Training examples:

$$(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, y_n)$$

where $\vec{x}_i = (x_{i1}, \ldots, x_{im})^T \in \mathbb{R}^m$ is the feature vector and $y_i \in \mathbb{R}$ is the target value for $i^{th} example.$

We seek a function/hypothesis $h : \mathbb{R}^m \longrightarrow \mathbb{R}$ such that $h(x)$ is a good predictor for corresponding values of y.

Recall from last lecture

Training examples:

$$(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, y_n)$$

where $\vec{x}_i = (x_{i1}, \ldots, x_{im})^T \in \mathbb{R}^m$ is the feature vector and $y_i \in \mathbb{R}$ is the target value for $i^{th} example$.

We seek a function/hypothesis $h : \mathbb{R}^m \longrightarrow \mathbb{R}$ such that $h(x)$ is a good predictor for corresponding values of y.

Modeling Choice: We assume that dependency of y on $\vec{x}$ is linear and approximate y as a linear function of x:

$$h_\beta(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m = \vec{\beta}^T \vec{x}$$

where

$$\vec{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)^T \in \mathbb{R}^{m+1}$$

$$\vec{x} = (1, x_1, \ldots, x_m)^T \in \mathbb{R}^{m+1}$$

How to determine β?

Find $\vec{\beta}$ such that the linear model fits the training examples well

input	actual	predicted	residual/error
x_1	y_1	$\hat{y}_1 = \vec{\beta}^T \vec{x}_1$	$e_1 = y_1 - \bar{y}_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	y_i	$\hat{y}_i = \vec{\beta}^T \vec{x}_i$	$e_i = y_i - \bar{y}_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	y_n	$\hat{y}_n = \vec{\beta}^T \vec{x}_n$	$e_n = y_n - \bar{y}_n$

How to determine β?

We define Residual Sum of Squares (RSS) and Mean Square Error (MSE) as

$$\begin{aligned}
\text{RSS} :=& e_1^2 + e_2^2 + \ldots + e_n^2 = \sum_{i=1}^{n} e_i^2 \\
=& \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \\
\text{MSE} :=& \text{RSS}/n
\end{aligned}$$

This is sometimes called the loss function and denoted by $L(\beta)$.

How to determine β?

We define Residual Sum of Squares (RSS) and Mean Square Error (MSE) as

$$\begin{aligned} \text{RSS} :=& e_1^2 + e_2^2 + \ldots + e_n^2 = \sum_{i=1}^{n} e_i^2 \\ =& \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \\ \text{MSE} :=& \text{RSS}/n \end{aligned}$$

This is sometimes called the loss function and denoted by $L(\beta)$.

Choose β so as to minimize RSS.

Ordinary Least Squares (OLS)

Simple Linear regression

$$
\begin{aligned}
h(x) &= \beta_0 + \beta_1 x \\
\frac{\partial L}{\partial \beta_0} &= 0 \quad \frac{\partial L}{\partial \beta_1} = 0 \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}
\end{aligned}
$$

Ordinary Least Squares (OLS)

Simple Linear regression

$$h(x) = \beta_0 + \beta_1 x$$

$$\frac{\partial L}{\partial \beta_0} = 0 \quad \frac{\partial L}{\partial \beta_1} = 0$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Multiple Linear regression

$$h(\vec{x}) = X\vec{\beta}$$

$$\begin{aligned} L(\vec{\beta}) &= ||\vec{y} - X\vec{\beta}||^2 \\ &= (\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta}) \\ &= \vec{y}^T\vec{y} - 2\vec{y}X\vec{\beta} + \vec{\beta}X^TX\vec{\beta} \end{aligned}$$

$$\frac{\partial L}{\partial \vec{\beta}} = 0 - 2\vec{X}^T\vec{y} + \vec{2}X^TX\vec{\beta}$$

$$\frac{\partial L}{\partial \vec{\beta}} = 0 \implies \vec{\beta} = (\vec{X}^TX)^{-1}\vec{X}^T\vec{y}$$

This lecture

Last time

- Discussed OLS as minimizing sum of square errors
- Gave explicit formula to compute parameters for the minimum

This lecture

Last time

- Discussed OLS as minimizing sum of square errors
- Gave explicit formula to compute parameters for the minimum

This time

- We rephrase the goals of OLS in a probabilistic light
  - Quick review of probability
  - Probability distributions
  - Likelihood function
  - Maximum likelihood estimation
  - Likelihood function and maximum likelihood estimates for linear regression

Quick review of probability

- Random Experiments
- Outcomes
- Probabilities
- Random Variables
- Expectation, Variance of a random variable
- Continuous random variables
- Probability density

Basic Definitions

Random Experiment: An experiment whose outcome is uncertain/non-deterministic.

Outcome : A realization of the random experiment

Sample Space Ω: Set of all possible outcomes of the random experiment

Event E: A subset of Ω is an event

Basic Definitions

Random Experiment: An experiment whose outcome is uncertain/non-deterministic.
Outcome : A realization of the random experiment
Sample Space Ω: Set of all possible outcomes of the random experiment
Event E: A subset of Ω is an event

Random Experiment: Roll a Die
Outcomes: 1,2,3,4,5,6
$\Omega = \{1, 2, 3, 4, 5, 6\}$
Events: $\{1\}, \{1, 2\}, \{1, 3, 5\}, \ldots$

Random Experiment: Throw a dart at a dart-board
Outcomes: Any point (x, y) on the dart-board
$\Omega = \{(x, y) : \ (x, y) \in \text{the dart-board}\}$
Events: $E = \{(x, y) : \ (x, y) \in \text{inner circle}\}, \ldots$

Probability of an Event

Probability: A mathematical quantity that defines how likely it is for an event to occur in a random experiment.

Probability of an Event

Probability: A mathematical quantity that defines how likely it is for an event to occur in a random experiment.
Mathematically,

$$P : \text{Set of events} \longrightarrow [0, 1]$$

that satisfies the following conditions

Laws of probability

(1) If $A_1, A_2, \ldots$ are disjoint events then

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$$

(2)

$$P(\Omega) = 1$$

Interpretation of Probability

Two Interpretations:

Frequentist: Probability equals the long-run frequency of the event to occur, if the experiment was repeated several times.
Examples:

- The long-run frequency of heads in repeated tosses of a fair coin is $1/2$.
- The long-run frequency of 1's in repeated rolling of a fair die is $1/6$.

Subjective: Probability measures degree of subjective belief about uncertainty.
Examples:

- Probability that project X fails is 0.2.

Random Variable

Random variable: a variable whose value depends on possible outcomes of the randomized experiment.
Mathematically, is a mapping from the sample space Ω to a well-defined set.

- Real–valued $X : \Omega \longrightarrow \mathbb{R}$
- Discrete–valued $X : \Omega \longrightarrow \mathbb{N}$
- Binary–valued $X : \Omega \longrightarrow \{0, 1\}$
- Categorical $X : \Omega \longrightarrow \{\text{Excellent, Very Good, Good, Fair, Poor}\}$

Random Variable

Example: Suppose you throw a coin.

Sample space $\Omega = \{H, T\}$ Define a random variable:

$$X = \begin{cases} 1 & \text{if head occurs} \\ 0 & \text{if tail occurs} \end{cases}$$

For a fair coin,

$$P(X = 1) = P(X = 0) = 0.5$$

otherwise,

$$p(X = 0) = p \quad p(X = 1) = 1 - p$$

where

$$p = \frac{\text{number of times you observe a head}}{\text{total number of throws}}$$

will converge to p as the number of throws increases.

Random Variable

Random variables can be discrete or continuous.

- Discrete random variables have a countable number of outcomes Examples:
  -
  - {Excellent, Very Good, Good, Fair, Poor},
  - number of received calls
  - click or non-click
- Continuous random variables have an infinite continuum of possible values Examples:
  - weight, volume of a package
  - Time between calls to the customer service

Probability mass function

A probability mass function (pmf) is a function that gives the probability that a discrete random variable is exactly equal to some value.
Suppose that $X : \Omega \longrightarrow A$ is a discrete random variable defined on a sample space Ω. Then the probability mass function $f_X : A \longrightarrow [0, 1]$ for X is defined as:

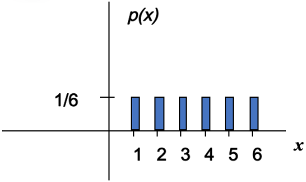
$$f_X(x) = Pr(X = x) = Pr(\{s \in S : X(s) = x\})$$

Note:

$$\sum_{x \in A} f_X(x) = 1$$

Probability mass function

Probability Distribution Function (pdf)



$$\sum_{\text{all } x} P(x) = 1$$

x	$p(x)$
1	$p(x=1)=1/6$
2	$p(x=2)=1/6$
3	$p(x=3)=1/6$
4	$p(x=4)=1/6$
5	$p(x=5)=1/6$
6	$p(x=6)=1/6$
	1.0

Probability density function

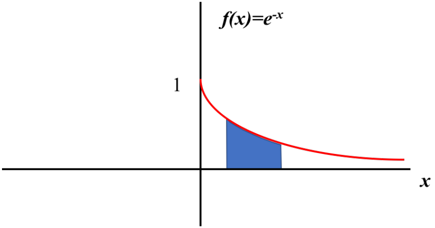
A probability density function (pdf) of a continuous random variable is a function whose value at any given sample (or point) in the sample space can be interpreted as providing a relative likelihood that the value of the random variable would equal that sample.

The probability density function ("p.d.f.") of a continuous random variable X with sample space Ω is an integrable function $f(x)$ satisfying the following:

1. $f(x) \geq 0$ for all $x \in \Omega$
2. $\int_{\Omega} f(x)\ dx = 1$
3. $P(a \leq X \leq b) = \int_a^b f(x)dx$

The Exponential Distribution

Generally the exponential distribution describes waiting time between Poisson occurrences.



$$P(1 \le x \le 2) = \int_1^2 e^{-x} = -e^{-x} \Big|_1^2 = -e^{-2} - -e^{-1} = -.135 + .368 = .23$$

$$P(1 \le x \le 2) = P(x \le 2) - P(x \le 1) = F(2) - F(1) = 2.3$$

Expectation

The expected value (or mean)of arandom variable, intuitively, is the long-run average value of repetitions of the experiment it represent.

- Discrete random variable X,

$$E(X) = \sum_{i}^{n} x_i P(x_i)$$

- Continuous random variable X,

$$E(X) = \int_{\Omega} x f(x)$$

Expectation

- Example: 6 faces fair dice, r.v. X is the value of one experiment

$$E(X) = \sum_{i=1}^{n} x_i P(x_i) = \frac{1}{6}(1 + 2 + 3 + \cdots + 6) = 3.5$$

- Example: exponential distribution X, $f(x) = e^{-x}$ $(x > 0)$

$$E(X) = \int xf(x)dx = \int_0^\infty xe^{-x} = 1$$

Variance

Variance is the expectation of the squared deviation of a random variable from itsmean, and it informally measures how far a set of (random) numbers are spread out from their mean.

$$Var(X) = E\left[(X - E(X))^2\right] = E(X^2) - E^2(X)$$

Variance

- Example: 6 faces fair dice, r.v. X is the value of one experiment

$$E(X^2) = \frac{1}{6}(1^2 + 2^2 + \cdots 6^2) = 15.16, E(X) = 3.5$$

$$Var(X) = E(X^2) - E^2(X) = 15.16 - 3.5^2 = 2.92$$

- Example: exponential distribution

$$E(X^2) = \int_0^\infty x^2 f(x)dx = \int_0^\infty x^2 e^{-x}\, dx = 2$$

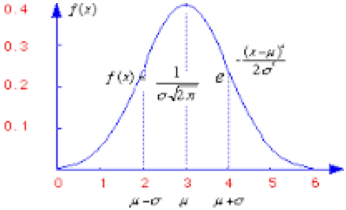
$$Var(X) = E(X^2) - E^2(X) = 2 - 1 = 1$$

Normal distribution

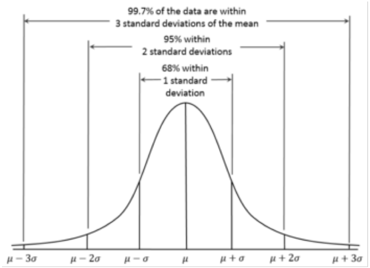
The probability density function of a Gaussian distribution is given by

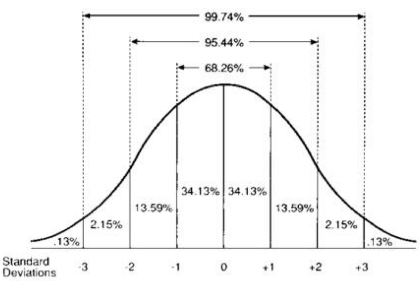
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Where μ is the mean and σ^2 is the variance.



Standard Normal distribution





Maximum Likelihood Estimation

Question Why might linear regression, and specifically why might the least-squares cost function be a reasonable choice?

We give a set of probabilistic assumptions, under which least-squares regression is derived as a very natural algorithm.

Linear Regression Assumptions

Given training examples:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

we assume that $y_i = f(x_i) + \epsilon_i$ where

- $f(x_i) = \beta_0 + \beta_1 x_i$ (linear assumption)
- ϵ_i is a random term that captures either unmodeled effects, or random noise
  - $\epsilon_1, \ldots, \epsilon_n$ are distributed IID (independently and identically distributed) according to a Gaussian distribution mean zero and some variance σ^2, i.e.,
  $$\epsilon_i \sim N(0, \sigma^2)$$

Recall that the density of ϵ_i is given by

$$p(\epsilon_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\epsilon_i^2}{\sigma^2}\right)$$

In other words, we assume that each label y_i is Gaussian distributed with mean $\vec{\beta}^T \vec{x}_i$ and variance σ:

$$y_i \sim N(\vec{\beta}^T \vec{x}_i, \sigma^2)$$

This implies that

$$p(y_i | \vec{x}_i; \vec{\beta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \vec{\beta}^T \vec{x}_i)^2}{2\sigma^2} \right)$$

The notation "$p(y_i | \vec{x}_i; \vec{\beta})$" indicates that this is the distribution of given $\vec{x}_i$ and parameterized by $\vec{\beta}$

Sales vs Price

Price	Avg. Sales
\$10.99	5.89
\$11.99	5.29
\$12.99	4.76
\$13.99	4.61
\$14.99	3.82
\$15.99	3.34
\$16.99	3.35
\$17.99	3.21
\$18.99	3.08
\$19.99	3.01
\$21.99	2.93

Sales vs Price

Given price, sales is a random variable with normal distribution parameterized by β.

Likelihood

Likelihood: probability of data given the parameters of the distribution
The probability of the data is given by

$$
\begin{aligned}
p(\vec{y}|X;\vec{\beta}) &= p((y_1,\ldots,y_n)|\vec{X};\vec{\beta}) \\
&= \prod_{i=1}^{n} p(y_i|\vec{x}_i;\vec{\beta})
\end{aligned}
$$

This quantity is typically viewed a function of (and perhaps X), for a fixed value of β
When we wish to explicitly view this as a function of β, we will instead call it the likelihood function:

$$L(\beta) := L(\beta; X, \vec{y}) = p(\vec{y}|\vec{X};\vec{\beta})$$

Maximum Likelihood

Maximum Likelihood Estimation: Find the parameters $\vec{\beta}$ such that, under normal distribution parameterized by $\vec{\beta}$, the observed data is most likely to occur. i.e., we should choose $\vec{\beta}$ to maximize $L(\vec{\beta})$

$$
\begin{aligned}
L(\vec{\beta}) &:= p(\vec{y}|X;\vec{\beta}) = p((y_1,\ldots,y_n)|\vec{X};\vec{\beta}) \\
&= \prod_{i=1}^{n} p(y_i|\vec{x}_i;\vec{\beta}) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i-\vec{\beta}^T\vec{x}_i)^2}{2\sigma^2}\right) \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(y_i-\vec{\beta}^T\vec{x}_i)^2}{2\sigma^2}\right) \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{(\vec{y}-X\vec{\beta})^T(\vec{y}-X\vec{\beta})}{2\sigma^2}\right)
\end{aligned}
$$

Maximum Likelihood

Maximum Likelihood Estimation: choose $\vec{\beta}$ to maximize $L(\vec{\beta})$. Instead of maximizing $L(\vec{\beta})$, we can also maximize any strictly increasing function of $L(\vec{\beta})$.
The derivations will be a bit simpler if we instead maximize the log likelihood $\ell(\vec{\beta})$

$$
\begin{aligned}
\log L(\vec{\beta}) &:= \log\left(\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{(\vec{y}-X\vec{\beta})^T(\vec{y}-X\vec{\beta})}{\sigma^2}\right)\right) \\
&= \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n + \log\left(\exp\left(-\frac{(\vec{y}-X\vec{\beta})^T(\vec{y}-X\vec{\beta})}{\sigma^2}\right)\right) \\
&= n\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}(\vec{y}-X\vec{\beta})^T(\vec{y}-X\vec{\beta})
\end{aligned}
$$

Maximum Likelihood

Maximum Likelihood Estimation: choose $\vec{\beta}$ to maximize $L(\vec{\beta})$ or equivalently $\log L(\vec{\beta})$.

$$\log L(\vec{\beta}) = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta})$$

Maximum Likelihood

Maximum Likelihood Estimation: choose $\vec{\beta}$ to maximize $L(\vec{\beta})$ or equivalently $\log L(\vec{\beta})$.

$$\log L(\vec{\beta}) = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta})$$

$$\begin{aligned}\frac{\partial \log L(\vec{\beta})}{\partial \vec{\beta}} &= 0 - \frac{1}{2\sigma^2}\frac{\partial}{\partial \vec{\beta}}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta}) \\ &= \frac{1}{2\sigma^2}\frac{\partial}{\partial \vec{\beta}}\left(-\vec{X}^T\vec{y} + \vec{X}^T X\vec{\beta}\right)\end{aligned}$$

Maximum Likelihood

Maximum Likelihood Estimation: choose $\vec{\beta}$ to maximize $L(\vec{\beta})$ or equivalently $\log L(\vec{\beta})$.

$$\log L(\vec{\beta}) = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta})$$

$$\begin{aligned}\frac{\partial \log L(\vec{\beta})}{\partial \vec{\beta}} &= 0 - \frac{1}{2\sigma^2}\frac{\partial}{\partial \vec{\beta}}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta}) \\ &= \frac{1}{2\sigma^2}\frac{\partial}{\partial \vec{\beta}}\left(-\vec{X}^T\vec{y} + \vec{X}^T X\vec{\beta}\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial \log L(\vec{\beta})}{\partial \vec{\beta}} = 0 &\implies -\vec{X}^T\vec{y} + \vec{X}^T X\vec{\beta} = 0 \\ &\implies \vec{X}^T X\vec{\beta} = \vec{X}^T\vec{y} \\ &\implies \vec{\beta} = (\vec{X}^T X)^{-1}\vec{X}^T\vec{y}\end{aligned}$$

ML estimate of σ

Maximum Likelihood Estimation: choose σ to maximize $L(\vec{\beta}, \sigma)$ or equivalently $\log L(\vec{\beta}, \sigma)$.

$$\log L(\vec{\beta}, \sigma) = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta})$$

ML estimate of σ

Maximum Likelihood Estimation: choose σ to maximize $L(\vec{\beta}, \sigma)$ or equivalently $\log L(\vec{\beta}, \sigma)$.

$$\log L(\vec{\beta}, \sigma) = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta})$$

$$\begin{aligned}\frac{\partial \log L(\vec{\beta}, \sigma)}{\partial \sigma} &= -\frac{n}{2} \cdot \frac{1}{2\pi\sigma^2} \cdot 4\pi\sigma - \frac{1}{2} \cdot \frac{-2}{\sigma^3}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta}) \\ &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta})\end{aligned}$$

ML estimate of σ

Maximum Likelihood Estimation: choose σ to maximize $L(\vec{\beta}, \sigma)$ or equivalently $\log L(\vec{\beta}, \sigma)$.

$$\log L(\vec{\beta}, \sigma) = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta})$$

$$\begin{aligned} \frac{\partial \log L(\vec{\beta}, \sigma)}{\partial \sigma} &= -\frac{n}{2} \cdot \frac{1}{2\pi\sigma^2} \cdot 4\pi\sigma - \frac{1}{2} \cdot \frac{-2}{\sigma^3}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta}) \\ &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta}) \end{aligned}$$

$$\begin{aligned} \frac{\partial \log L(\vec{\beta})}{\partial \vec{\beta}} = 0 &\implies -\frac{n}{\sigma} + \frac{1}{\sigma^3}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta}) = 0 \\ &\implies \sigma^2 = \frac{1}{n}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta}) = \frac{1}{n}||\vec{y} - X\vec{\beta}||^2 \end{aligned}$$

Making predictions

Given ML estimates $\vec{\beta}, \sigma^2$, for a new input $\vec{x}^*$,

$$p(y_i \ \vec{x}^*; \beta, \sigma^2) = N(\vec{\beta}^T \vec{x}^*, \sigma^2)$$

The expected value of y_i is

$$E(y_i) = \vec{\beta}^T \vec{x}^*$$

and 95% confidence interval is given by

$$(\vec{\beta}^T \vec{x}^* - 2\sigma, \vec{\beta}^T \vec{x}^* + 2\sigma)$$

that is,

$$p(\vec{\beta}^T \vec{x}^* - 2\sigma \leq y_i \leq \vec{\beta}^T \vec{x}^* + 2\sigma) = 0.95$$