Data 401
Maximum Likelihood

Dennis Sun

October 17, 2016

1. The Class So Far
2. Probability Models and Maximum Likelihood
3. Linear Regression and Maximum Likelihood

The Class So Far

The Class So Far

- We've focused on fitting linear regression, which finds the $\boldsymbol{\beta}$ that minimizes

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}))^2.$$

The Class So Far

- We've focused on fitting linear regression, which finds the $\boldsymbol{\beta}$ that minimizes

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}))^2.$$

- In today's lecture, we will kill two birds with one stone:

The Class So Far

- We've focused on fitting linear regression, which finds the $\boldsymbol{\beta}$ that minimizes

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}))^2.$$

- In today's lecture, we will kill two birds with one stone:
  - We will justify why we minimize the sum of squared differences, instead of some other metric.

The Class So Far

- We've focused on fitting linear regression, which finds the $\boldsymbol{\beta}$ that minimizes

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}))^2.$$

- In today's lecture, we will kill two birds with one stone:
  - We will justify why we minimize the sum of squared differences, instead of some other metric.
  - We will see how probability enters into linear regression.

The Class So Far

- We've focused on fitting linear regression, which finds the $\boldsymbol{\beta}$ that minimizes

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}))^2.$$

- In today's lecture, we will kill two birds with one stone:
  - We will justify why we minimize the sum of squared differences, instead of some other metric.
  - We will see how probability enters into linear regression.
- The unifying link will be the principle of **maximum likelihood**.

Probability Models

Probability Models

- In statistics, we assume that our data comes from some probability model.

Probability Models

- In statistics, we assume that our data comes from some probability model.
- This model captures either the randomness inherent in our data collection procedure (e.g., simple random sample, randomized experiments) or noise in our data.

Probability Models

- In statistics, we assume that our data comes from some probability model.
- This model captures either the randomness inherent in our data collection procedure (e.g., simple random sample, randomized experiments) or noise in our data.
- The probabilistic model usually specifies the general **family of distributions**, but not the exact parameters.

Probability Models

- In statistics, we assume that our data comes from some probability model.
- This model captures either the randomness inherent in our data collection procedure (e.g., simple random sample, randomized experiments) or noise in our data.
- The probabilistic model usually specifies the general **family of distributions**, but not the exact parameters.

  Example: Suppose we might model the lifetime of a lightbulb as an exponential distribution, with p.d.f.

$$p_\lambda(x) = \lambda e^{-\lambda x}.$$

Probability Models

- In statistics, we assume that our data comes from some probability model.
- This model captures either the randomness inherent in our data collection procedure (e.g., simple random sample, randomized experiments) or noise in our data.
- The probabilistic model usually specifies the general **family of distributions**, but not the exact parameters.

  Example: Suppose we might model the lifetime of a lightbulb as an exponential distribution, with p.d.f.

  $$p_\lambda(x) = \lambda e^{-\lambda x}.$$

  We still have to estimate λ.

The Estimation Problem

Example: Suppose we might model the lifetime of a lightbulb as an exponential distribution, with p.d.f.

$$p_\lambda(x) = \lambda e^{-\lambda x}.$$

We have a lightbulb and we observe how long it lasts. That lifetime X is a random variable drawn from this distribution.

The Estimation Problem

Example: Suppose we might model the lifetime of a lightbulb as an exponential distribution, with p.d.f.

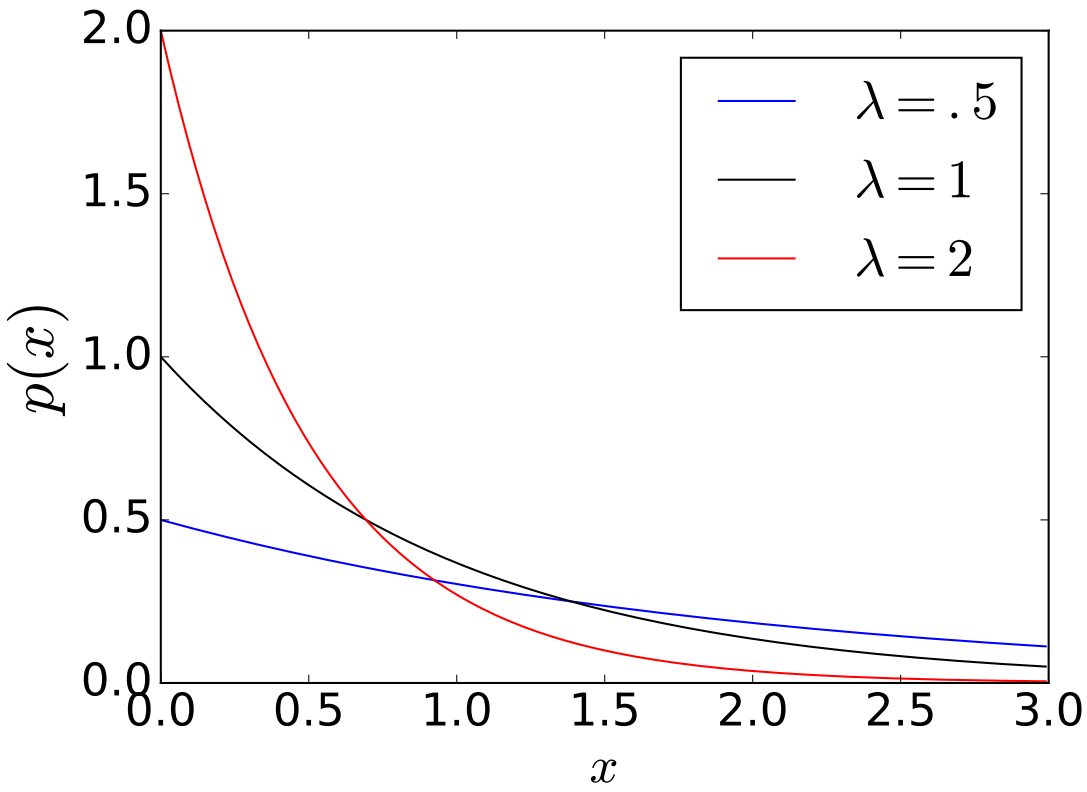
$$p_\lambda(x) = \lambda e^{-\lambda x}.$$

We have a lightbulb and we observe how long it lasts. That lifetime X is a random variable drawn from this distribution.

Suppose $X = 2\,\mathrm{years}$. How would you use this data to estimate λ?

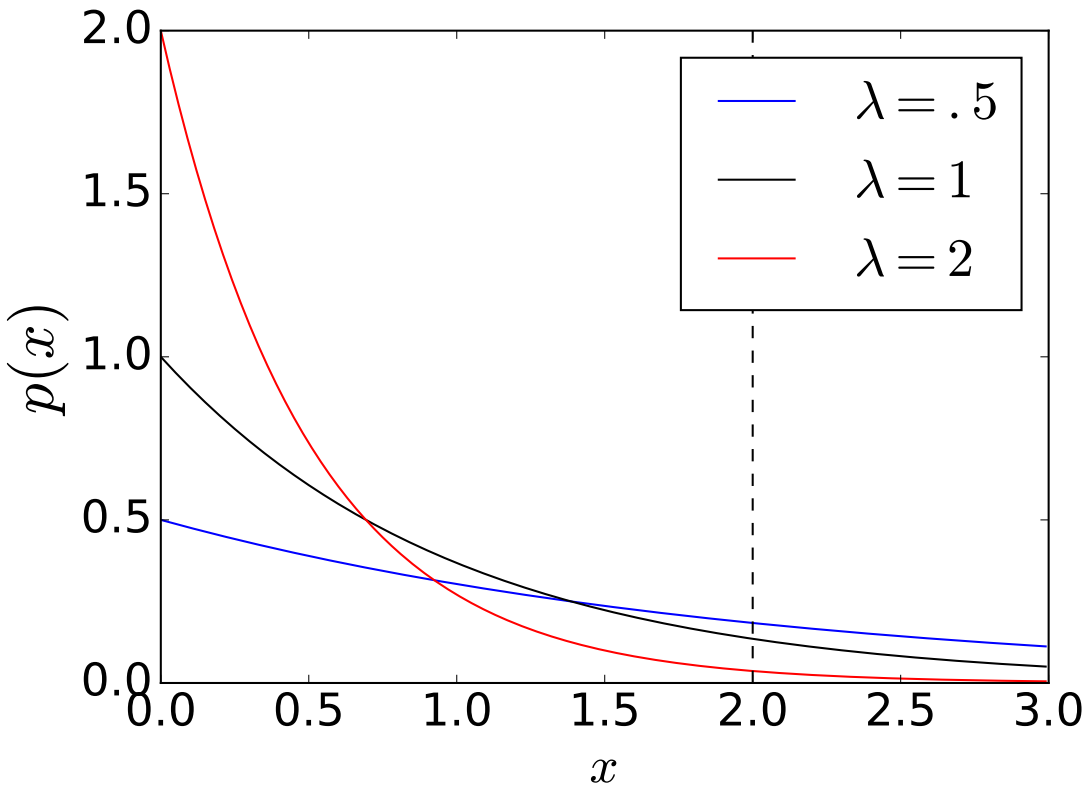
Maximum Likelihood in Pictures

Idea: Find the λ which maximizes the probability of the data.



Maximum Likelihood in Pictures

Idea: Find the λ which maximizes the probability of the data.



From this plot, it looks like $\lambda = .5$ is the best. But how do we know that there isn't some other value of λ that's even better?

Maximum Likelihood in Algebra

$$p_\lambda(x) = \lambda e^{-\lambda x}$$

You are used to fixing a value of λ, say $\lambda = 1$, and calculating how likely we are to observe a particular x.

Maximum Likelihood in Algebra

$$p_\lambda(x) = \lambda e^{-\lambda x}$$

You are used to fixing a value of λ, say $\lambda = 1$, and calculating how likely we are to observe a particular x.

Maximum likelihood involves inverting this perspective. We fix the value of x depending on the observed data and vary λ.

Maximum Likelihood in Algebra

$$p_\lambda(x) = \lambda e^{-\lambda x}$$

You are used to fixing a value of λ, say $\lambda = 1$, and calculating how likely we are to observe a particular x.

Maximum likelihood involves inverting this perspective. We fix the value of x depending on the observed data and vary λ.

In other words, maximum likelihood requires that we view the p.d.f. as a function of λ, not as a function of x!

$$L_X(\lambda) = p_\lambda(X)$$

Maximum Likelihood in Algebra

$$p_\lambda(x) = \lambda e^{-\lambda x}$$

You are used to fixing a value of λ, say $\lambda = 1$, and calculating how likely we are to observe a particular x.

Maximum likelihood involves inverting this perspective. We fix the value of x depending on the observed data and vary λ.

In other words, maximum likelihood requires that we view the p.d.f. as a function of λ, not as a function of x!

$$L_X(\lambda) = p_\lambda(X)$$

L_X is called the **likelihood function**.

Maximum Likelihood in Algebra

How do we find the λ that maximizes the likelihood:

$$L_X(\lambda) = \lambda e^{-\lambda X}?$$

Maximum Likelihood in Algebra

How do we find the λ that maximizes the likelihood:

$$L_X(\lambda) = \lambda e^{-\lambda X}?$$

Easy! Take the derivative and set it equal to 0.

Maximum Likelihood in Algebra

How do we find the λ that maximizes the likelihood:

$$L_X(\lambda) = \lambda e^{-\lambda X}?$$

Easy! Take the derivative and set it equal to 0.

Ugh! We'll have to use the *product rule*. In general, when you have probabilities, there will be lots of multiplication.

Maximum Likelihood in Algebra

How do we find the λ that maximizes the likelihood:

$$L_X(\lambda) = \lambda e^{-\lambda X}?$$

Easy! Take the derivative and set it equal to 0.

Ugh! We'll have to use the *product rule*. In general, when you have probabilities, there will be lots of multiplication.

To make derivatives more convenient to calculate, statisticians typically maximize the **log-**likelihood instead.

$$\log L_X(\lambda) = \log \lambda - \lambda X.$$

Maximum Likelihood in Algebra

How do we find the λ that maximizes the likelihood:

$$L_X(\lambda) = \lambda e^{-\lambda X}?$$

Easy! Take the derivative and set it equal to 0.

Ugh! We'll have to use the *product rule*. In general, when you have probabilities, there will be lots of multiplication.

To make derivatives more convenient to calculate, statisticians typically maximize the **log-**likelihood instead.

$$\log L_X(\lambda) = \log \lambda - \lambda X.$$

So $\hat{\lambda}_{\text{MLE}} = \frac{1}{X}$. When $X = 2$, we get $\hat{\lambda}_{\text{MLE}} = .5$.

Review of Last Class

There are about 10 special cases where you can solve for the MLE by setting the derivative equal to 0.

Review of Last Class

There are about 10 special cases where you can solve for the MLE by setting the derivative equal to 0.

What do we do for all the other cases where we can't solve this equation?

Review of Last Class

There are about 10 special cases where you can solve for the MLE by setting the derivative equal to 0.

What do we do for all the other cases where we can't solve this equation?

Gradient descent / ascent!

The Maximum Likelihood Recipe

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{X}).$$

The Maximum Likelihood Recipe

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{X}).$$

- Calculate $\log p_{\boldsymbol{\theta}}(\mathbf{X})$.

The Maximum Likelihood Recipe

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{X}).$$

- Calculate $\log p_{\boldsymbol{\theta}}(\mathbf{X})$.
- Take the derivative (gradient) with respect to $\boldsymbol{\theta}$.

The Maximum Likelihood Recipe

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{X}).$$

- Calculate $\log p_{\boldsymbol{\theta}}(\mathbf{X})$.
- Take the derivative (gradient) with respect to $\boldsymbol{\theta}$.
- If you can solve for $\hat{\boldsymbol{\theta}}$ by setting the gradient equal to $\mathbf{0}$, do it. Otherwise, use gradient ascent.

Probability Model for Linear Regression

Probability Model for Linear Regression

- Treat X's as fixed.

Probability Model for Linear Regression

- Treat X's as fixed.
- $Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ are independent.

Probability Model for Linear Regression

- Treat X's as fixed.
- $Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ are independent.
- So $Y_i \sim N(\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}, \sigma^2)$ are independent.

Probability Model for Linear Regression

- Treat X's as fixed.
- $Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ are independent.
- So $Y_i \sim N(\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}, \sigma^2)$ are independent.

The p.d.f. of a single Y_i is:

$$p_{\boldsymbol{\beta}}(Y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2},$$

Probability Model for Linear Regression

- Treat X's as fixed.
- $Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ are independent.
- So $Y_i \sim N(\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}, \sigma^2)$ are independent.

The p.d.f. of a single Y_i is:

$$p_{\boldsymbol{\beta}}(Y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2},$$

so the p.d.f. of all the Y_is is:

$$\begin{aligned} p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) &= p_{\boldsymbol{\beta}}(Y_1) p_{\boldsymbol{\beta}}(Y_2) ... p_{\boldsymbol{\beta}}(Y_n) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2}. \end{aligned}$$

Maximum Likelihood Estimator of β

$$p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2}.$$

Maximum Likelihood Estimator of β

$$p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2}.$$

Step 1: Take the log.

Maximum Likelihood Estimator of β

$$p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2}.$$

Step 1: Take the log.

$$\log p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \sum_{i=1}^{n} \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2}\right)$$

Maximum Likelihood Estimator of β

$$p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2}.$$

Step 1: Take the log.

$$\begin{aligned}
\log p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) &= \sum_{i=1}^{n} \log \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2} \right) \\
&= \sum_{i=1}^{n} \log \frac{1}{\sigma\sqrt{2\pi}} + \\
&\qquad \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2
\end{aligned}$$

Maximum Likelihood Estimator of β

$$p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2}.$$

Maximum Likelihood Estimator of β

$$p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2}.$$

Step 2: Maximize with respect to $\boldsymbol{\beta}$

Maximum Likelihood Estimator of β

$$p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2}.$$

Step 2: Maximize with respect to $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \log \frac{1}{\sigma\sqrt{2\pi}} - \sum_{i=1}^{n} \frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2$$

Maximum Likelihood Estimator of β

$$p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2}.$$

Step 2: Maximize with respect to $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \log \frac{1}{\sigma\sqrt{2\pi}} - \sum_{i=1}^{n} \frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2$$

$$= \arg\max_{\boldsymbol{\beta}} - \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2$$

Maximum Likelihood Estimator of β

$$p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2}.$$

Step 2: Maximize with respect to $\boldsymbol{\beta}$

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{MLE}} &= \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \log \frac{1}{\sigma\sqrt{2\pi}} - \sum_{i=1}^{n} \frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2 \\
&= \arg\max_{\boldsymbol{\beta}} - \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2 \\
&= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2
\end{aligned}$$

Maximum Likelihood Estimator of β

$$p_{\boldsymbol{\beta}}(Y_1, ..., Y_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2}.$$

Step 2: Maximize with respect to $\boldsymbol{\beta}$

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{MLE}} &= \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \log \frac{1}{\sigma\sqrt{2\pi}} - \sum_{i=1}^{n} \frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2 \\
&= \arg\max_{\boldsymbol{\beta}} - \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2 \\
&= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}))^2
\end{aligned}$$

So the linear regression estimate is the MLE when we assume that the errors are normally distributed.