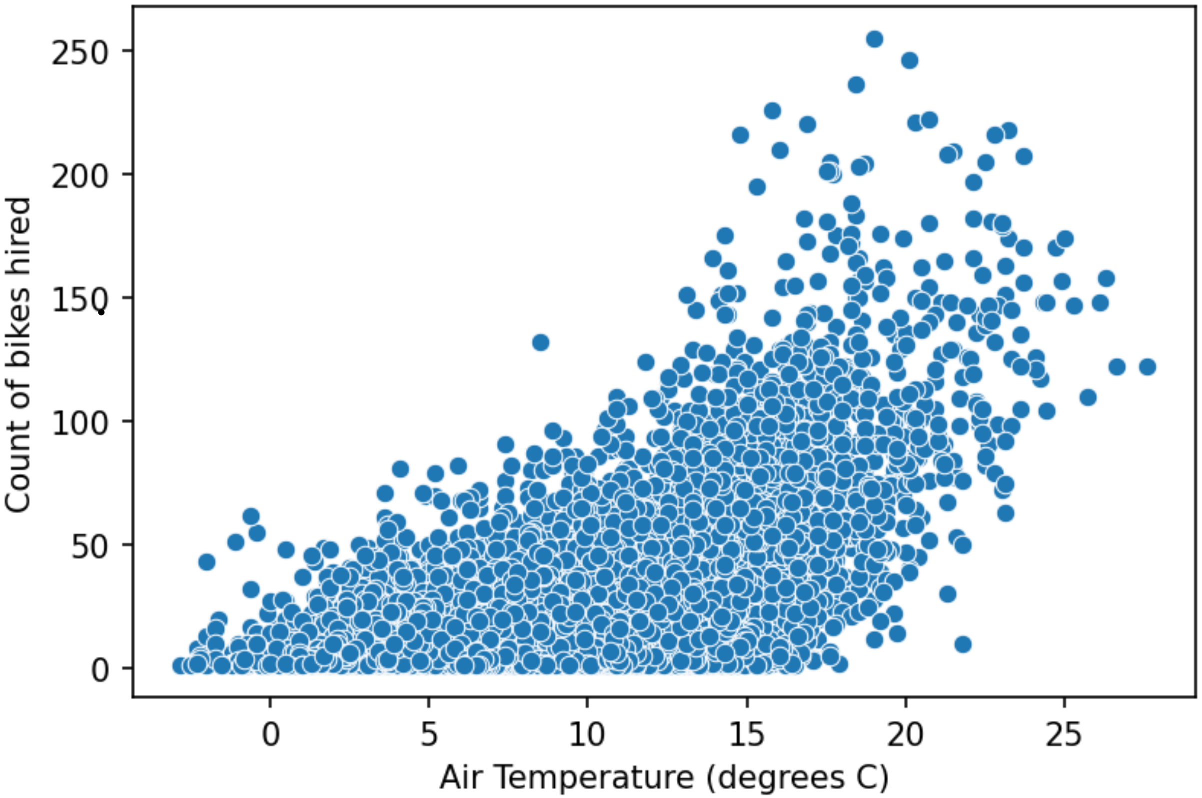
Foundations of Data Science: Regression and inference – From the maximum likelihood principle to linear regression

THE UNIVERSITY of EDINBURGH
informatics
FOUNDATIONS OF DATA SCIENCE

We want to investigate the relationship between the number of bikes hired in a day and the temperature on that day

Is there a problem with using ordinary least squares linear regression to do this?

Data sources:
- Edinburgh Just Eat Bikes data 2020
- Edinburgh temperature observations, Met Office via MIDAS



Overview

Today

1. The maximum likelihood principle
2. Application of maximum likelihood principle
   to a simple example
3. Application of maximum likelihood principle
   to linear regression

Wednesday

- Max likelihood with non-normal distributions
- Generalised linear regresion

Foundations of Data Science: Regression and inference – The maximum likelihood principle

Intuition for maximum likelihood principle

Statistical model

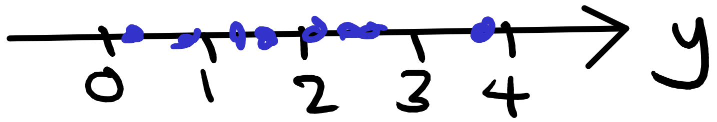
e.g Normal dist.

$$p(Y=y \mid 2, 1.1^2) = \frac{1}{\sqrt{2\pi}\,1.1} e^{-\frac{1}{2}\left(\frac{y-2}{1.1}\right)^2}$$

$\uparrow$ μ $\quad$ $\uparrow$ σ^2

Generates $\longrightarrow$

"Fake data"



Alternative notation:

$$y \sim \mathcal{N}(2, 1.1^2)$$

$\uparrow$ normal dist $\quad$ μ $\quad$ σ^2

"y is drawn from normal dist with mean 2 and variance 1.1^2"

Intuition for maximum likelihood principle

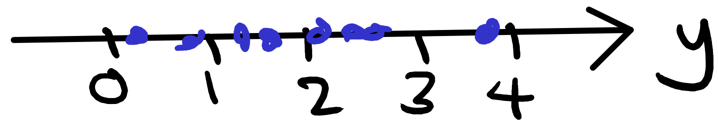
Statistical model

e.g Normal dist.

$$p(Y=y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,|\sigma|} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$

$\uparrow$ μ $\quad$ $\uparrow$ σ^2

"data"



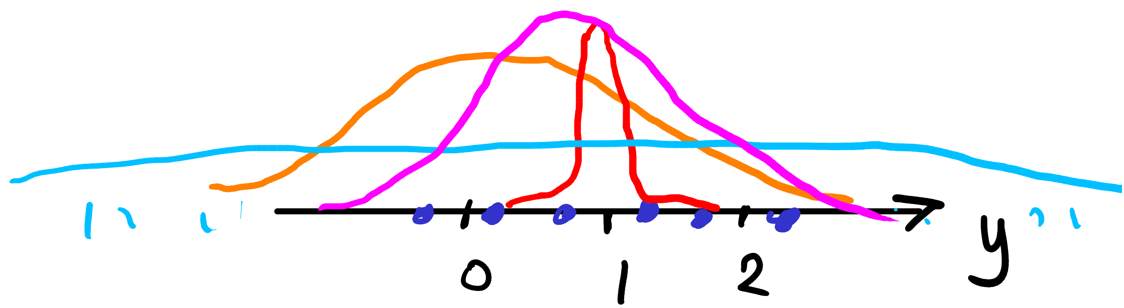
$\longleftarrow$ Infer μ and σ^2

Find parameters (here μ, σ^2) that maximise likelihood of the model having produced the observed data.

Exercise

Which model is most likely to have generated this data?

Data



$y_1, \ldots, y_n$

① $y_i \sim \mathcal{N}(\overset{\mu}{0}, \overset{\sigma^2}{1})$

② $y_i \sim \mathcal{N}(1, 0.1^2)$

③ $y_i \sim \mathcal{N}(1, 5^2)$

④ $y_i \sim \mathcal{N}(1, 1)$

Definition of the maximum likelihood principle

For a set of observed data and
a given statistical model
the principle of maximum likelihood states that
the parameters of the model are adjusted so as to maximise
the likelihood that the model generated the observed data.

Data : $\{y_1, \ldots, y_n\}$

Model : $p(\underline{Y} = [y_1, \ldots, y_n] \mid \vartheta_1, \ldots, \vartheta_m)$

$\uparrow$

Likelihood of data given model

Foundations of Data Science: Regression and inference – Application of the maximum likelihood principle to a simple example

Application to 1-variable example

1. Assume samples are drawn independently
2. Assume each sample is drawn from an normal distribution

$$P(Y = y_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{y_i - \mu}{\sigma^2}\right)^2} \quad (2)$$

Assumption (1) $\Rightarrow$

$$P(\underline{Y} = [y_1, \ldots, y_n] \mid \mu, \sigma^2) = p(Y = y_1 \mid \mu, \sigma^2) \times$$
$$p(Y = y_2 \mid \mu, \sigma^2) \times$$
$$\cdots\cdots$$
$$p(Y = y_n \mid \mu, \sigma^2)$$

More compact notation...

$$p(\underline{Y} = [y_1, \ldots, y_n] \mid \mu, \sigma^2) = p(Y = y_1 \mid \mu, \sigma^2) \times$$
$$p(Y = y_2 \mid \mu, \sigma^2) \times$$
$$\cdots\cdots$$
$$p(Y = y_n \mid \mu, \sigma^2)$$

$$= \prod_{i=1}^{n} p(Y = y_i \mid \mu, \sigma^2)$$

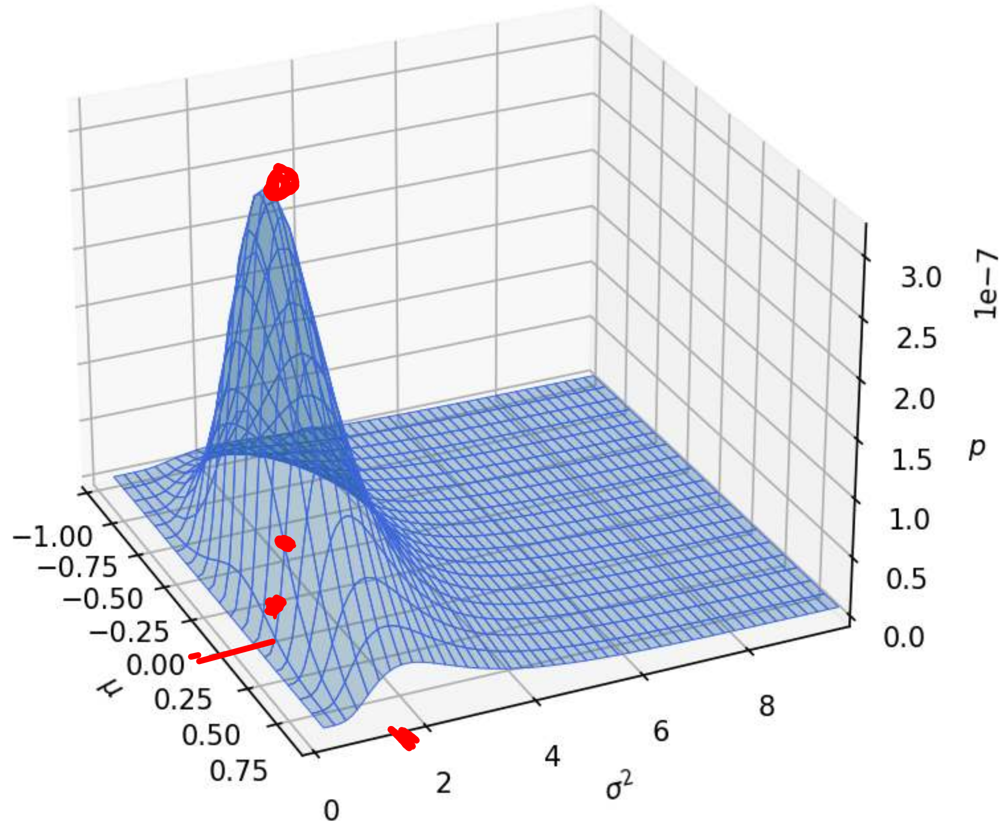
$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\frac{(y_i - \mu)^2}{\sigma^2}}$$

Likelihood as a function of parameters

Data:

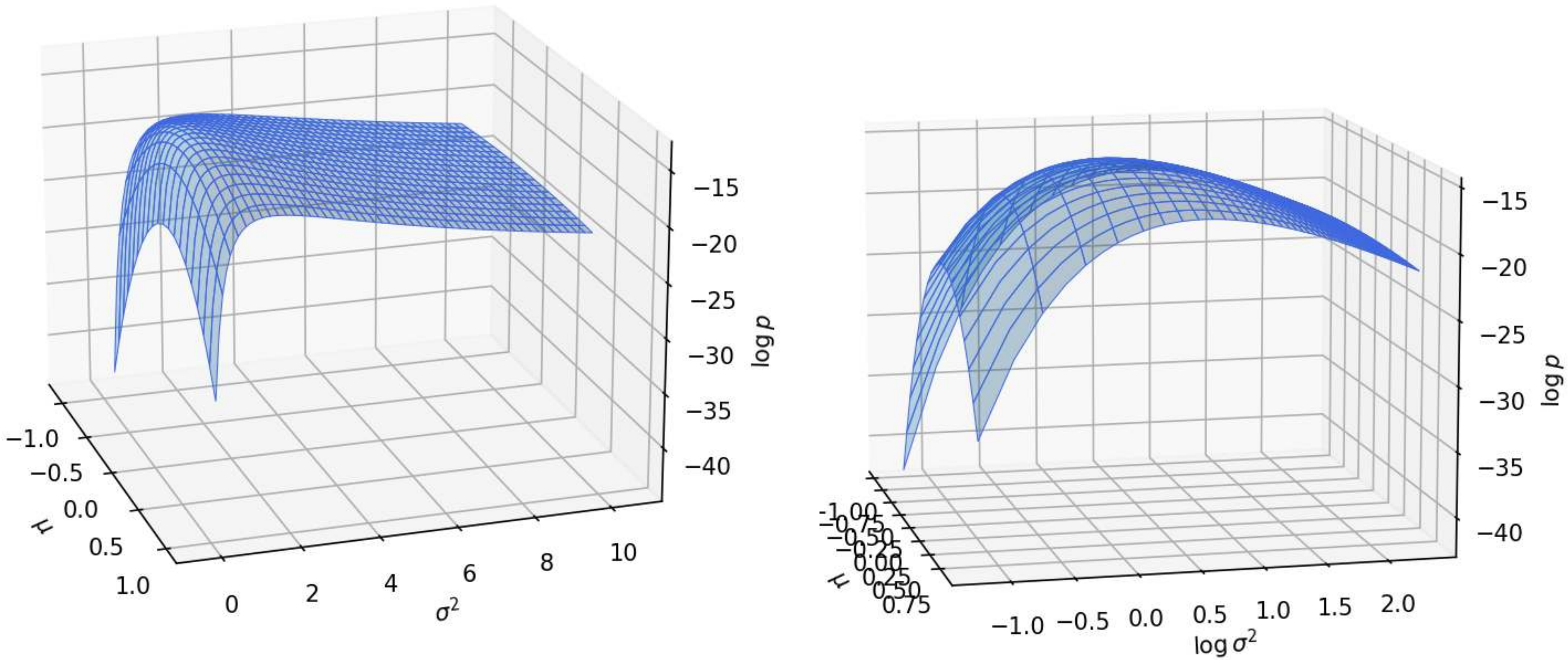
$y_1, \ldots, y_{10}$ drawn from

$\mathcal{N}(0, 1)$



[code]

Log-likelihood as a function of parameters



Log-likelihood equations: products to sums

$$\ln ab = \ln a + \ln b$$

So $\ln(p_1 \times p_2 \times \dots \times p_n) = \ln p_1 + \ln p_2 + \dots + \ln p_n$

So $\ln \prod_{i=1}^{n} p_i = \sum_{i=1}^{n} \ln p_i$

$$\ln \prod_{i=1}^{n} P(Y = y_i \mid \mu, \sigma^2) = \sum_{i=1}^{n} \ln P(Y = y_i \mid \mu, \sigma^2)$$

The beauty of logs and sums

- Sum of logs is easy to represent within limits of floating point arithmetic

- Log likelihood function is smoother than likelihood function

- Sums are easy to differentiate; products are not

The log of the normal distribution

$$\ln \left\{ \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} \right\}$$

$$= -\ln\left(\sqrt{2\pi}\,\sigma\right) - \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2$$

$$= -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2$$

Final expression for log likelihood

$$\ln p(\underline{Y} = y_1, \ldots y_n \mid \mu, \sigma^2)$$

$$= \sum_{i=1}^{n} \left(-\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right)$$

Maximise w.r.t μ and σ^2

$\Rightarrow$ Maximum likelihood estimates (MLE)

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu})^2$$

Exercise: prove these statements by differentiating w.r.t. μ and σ^2

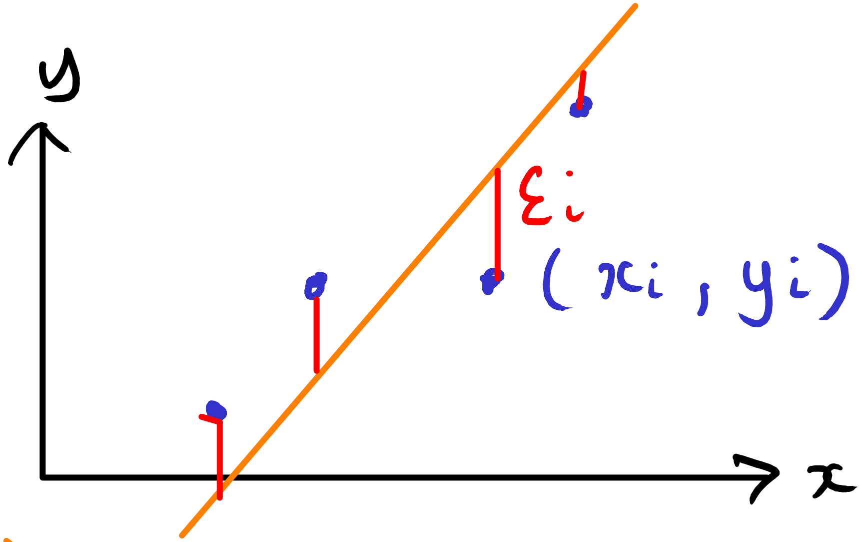
Foundations of Data Science: Regression and inference – Application of the maximum likelihood principle to linear regression

Application of max likelihood to linear regression

$$Y_i = \beta_0 + \beta_1 x_i + \underbrace{\varepsilon_i}_{\text{Error term}}$$

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

↑ residual



OR

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$\ln p(\underline{Y} = y_1, \ldots y_n \mid \beta_0, \beta_1, \sigma^2)$$

$$= \sum_{i=1}^{n} \left(-\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2} \right)$$

Relationship to ordinary least squares

$$\ln p(\underline{Y} = y_1, \dots, y_n ; x_1, \dots, x_n \mid \beta_0, \beta_1, \sigma^2)$$

$$= \sum_{i=1}^{n} \left(-\frac{1}{2} \ln 2\pi\sigma^2 \underbrace{- \frac{1}{2} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2}}_{\text{SSE}} \right)$$

Estimates of coefficients

Analytical solutions for $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$ that maximise likelihood

$\hat{\beta}_0$ and $\hat{\beta}_1$: as per ordinary least squares

Variance of residuals :

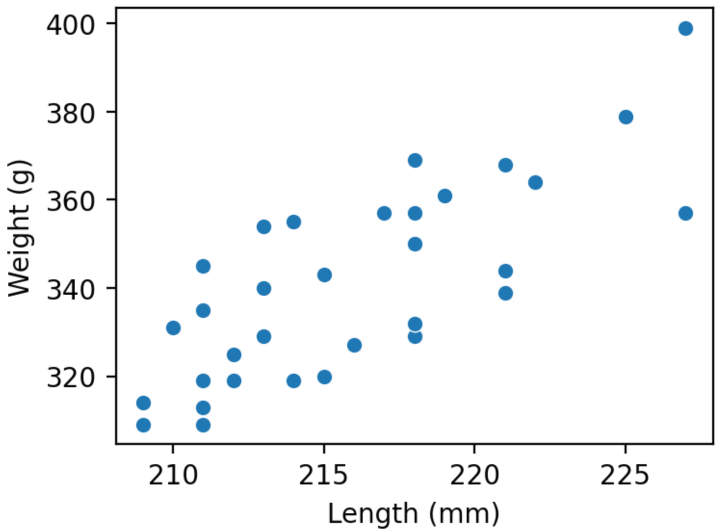
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \underbrace{\hat{\beta}_0 - \hat{\beta}_1 x_i}\right)^2$$

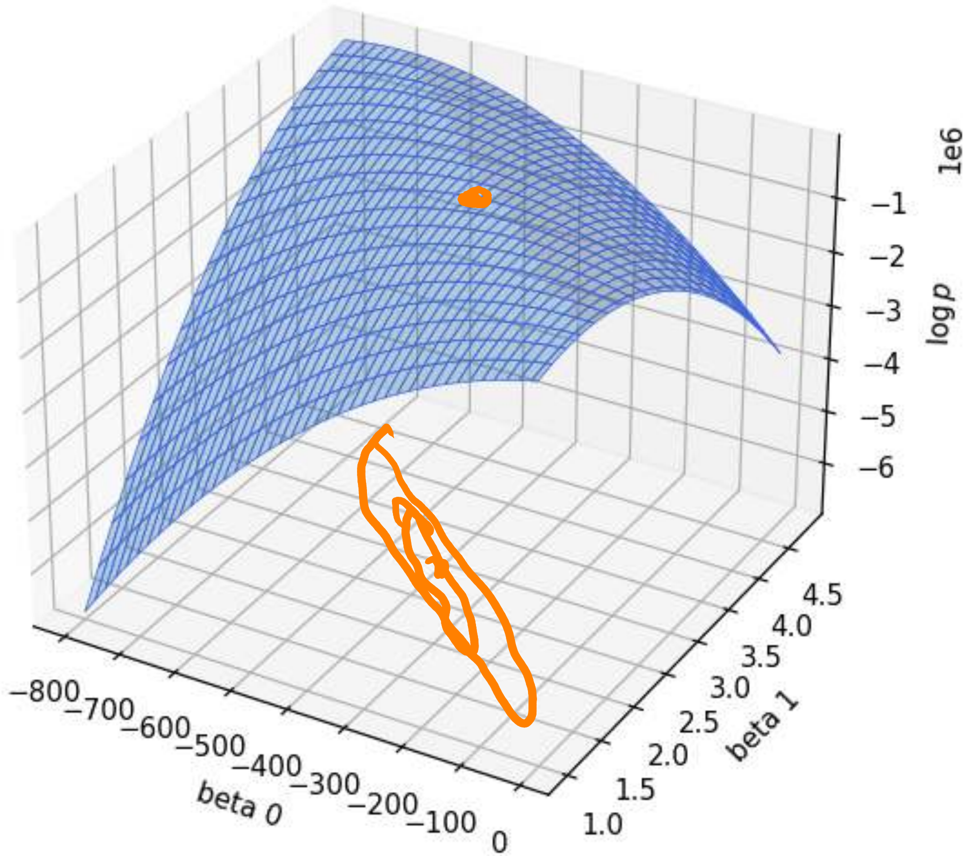
$$= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \frac{SSE}{n} \leftarrow \text{Biassed}$$

Sampling theory $\hat{\sigma}^2 = \frac{SSE}{n-2} \leftarrow$ Unbiassed

Log likelihood of coefficients

Peter Trimming, Wikimedia Commons, CC BY 2.0

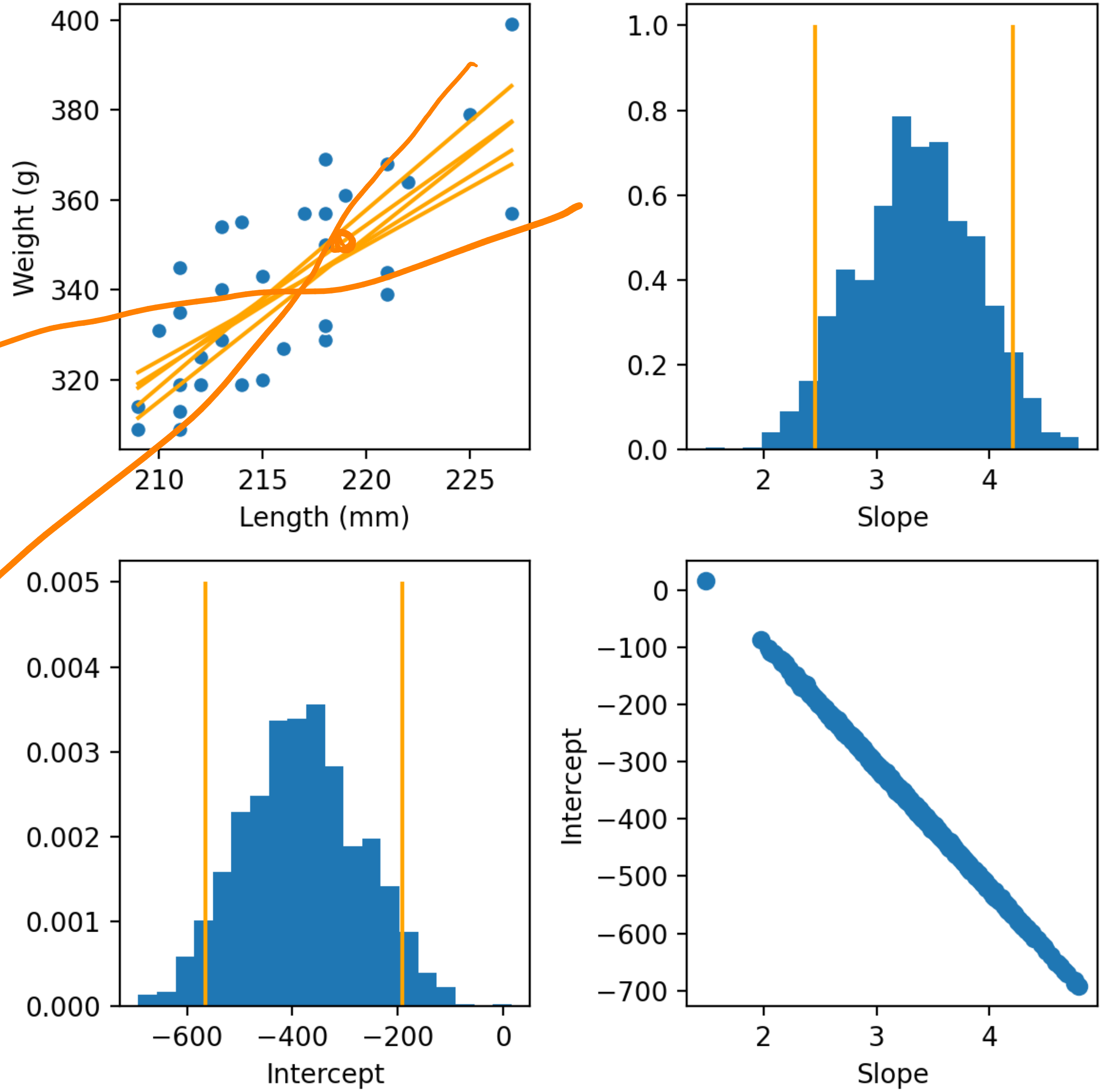




Data from Wauters and Dhondt 1989

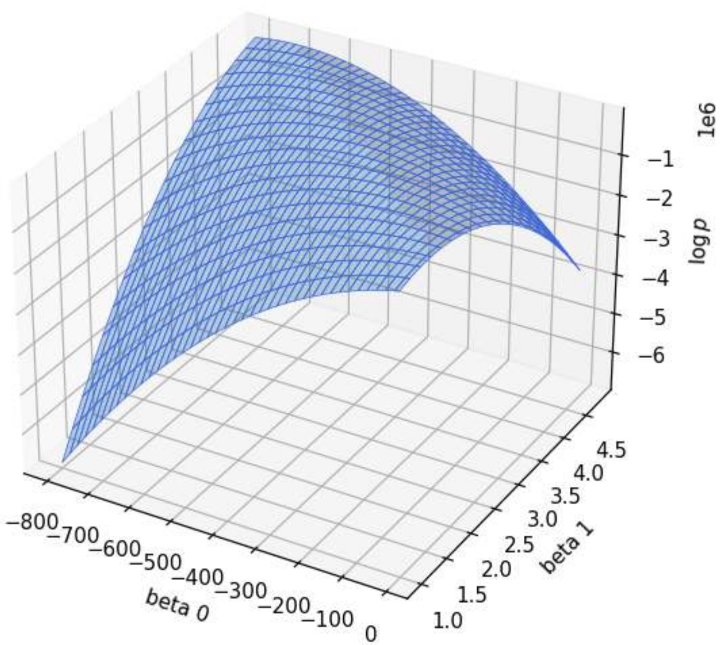
Bootstrap inference of coefficients

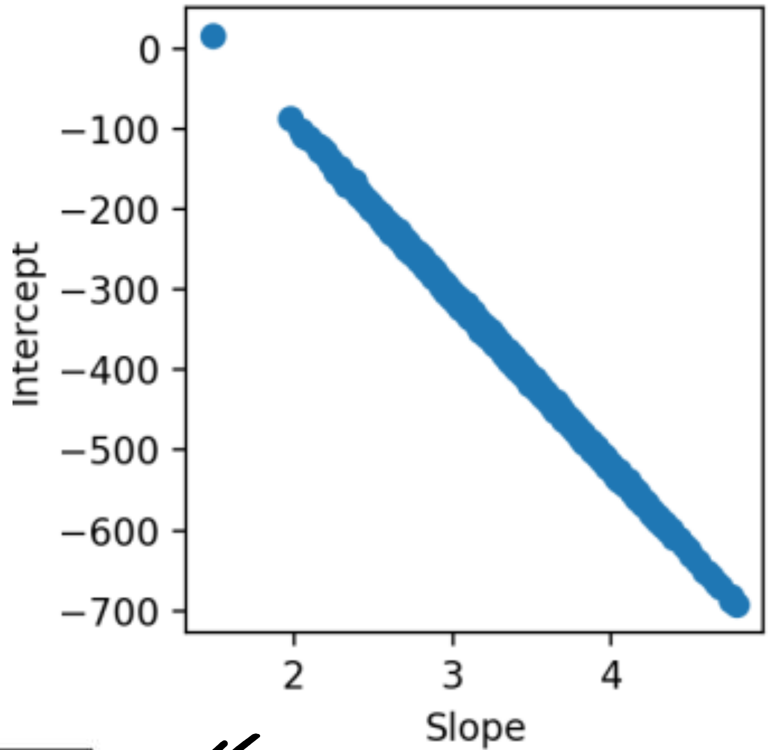
Peter Trimming, Wikimedia Commons, CC BY 2.0

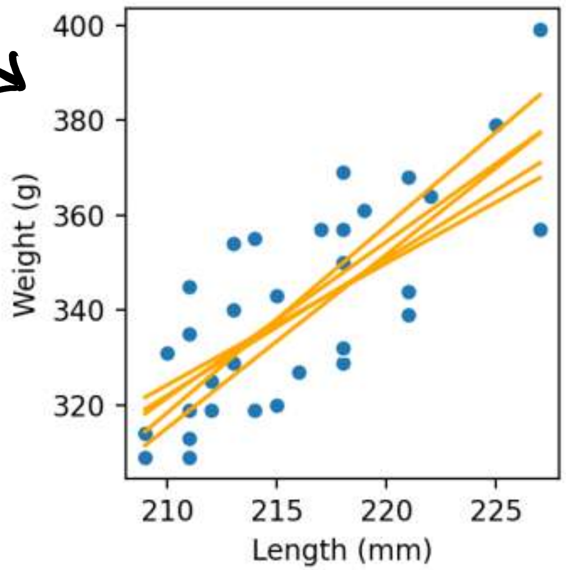


(Log) Likelihood function

Bootstrap samples







Overview

1. Maximum likelihood principle
   - What model was most likely to have generated the data

2. Maximum likelihood principle applied to simple example
   - Log likelihood turns out to be useful
   - Gives rise to familiar estimates for mean and variance

3. Maximum likelihood principle applied to linear regression
   - Turns out to give ordinary least squares
   - Link with coefficient uncertainty and the bootstrap
     estimates of parameter uncertainty

We want to investigate the relationship between the number of bikes hired in a day and the temperature on that day

Is there a problem with using ordinary least squares linear regression to do this?