Inf2 - Foundations of Data Science: Regression and inference - From the maximum likelihood principle to linear regression

Announcements

- Project - spec to be released by Wednesday
  - Appleton tower energy data
  - Steam games data
  - UoE course data

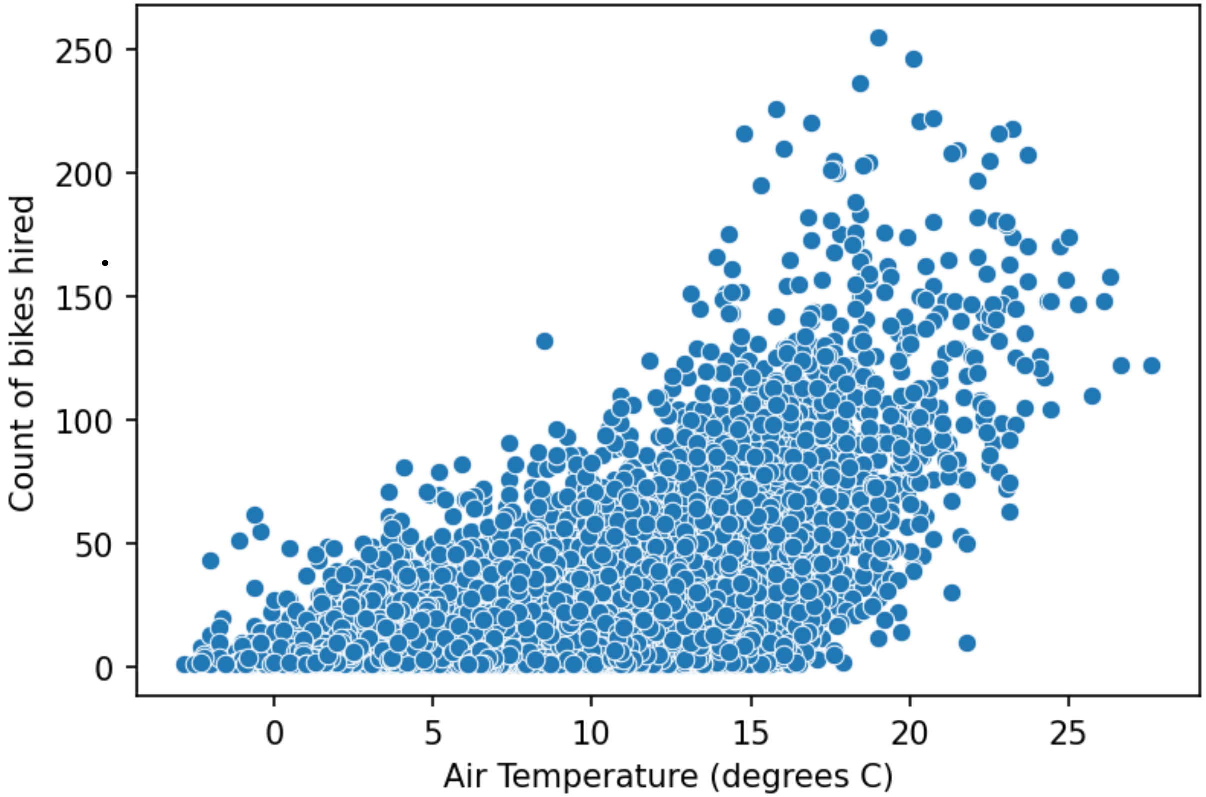
- Badges! Please let me know if I owe you one

We want to investigate the relationship between the number of bikes hired in an hour and the mean temperature during that hour

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

Image copyright Pashley Cycles

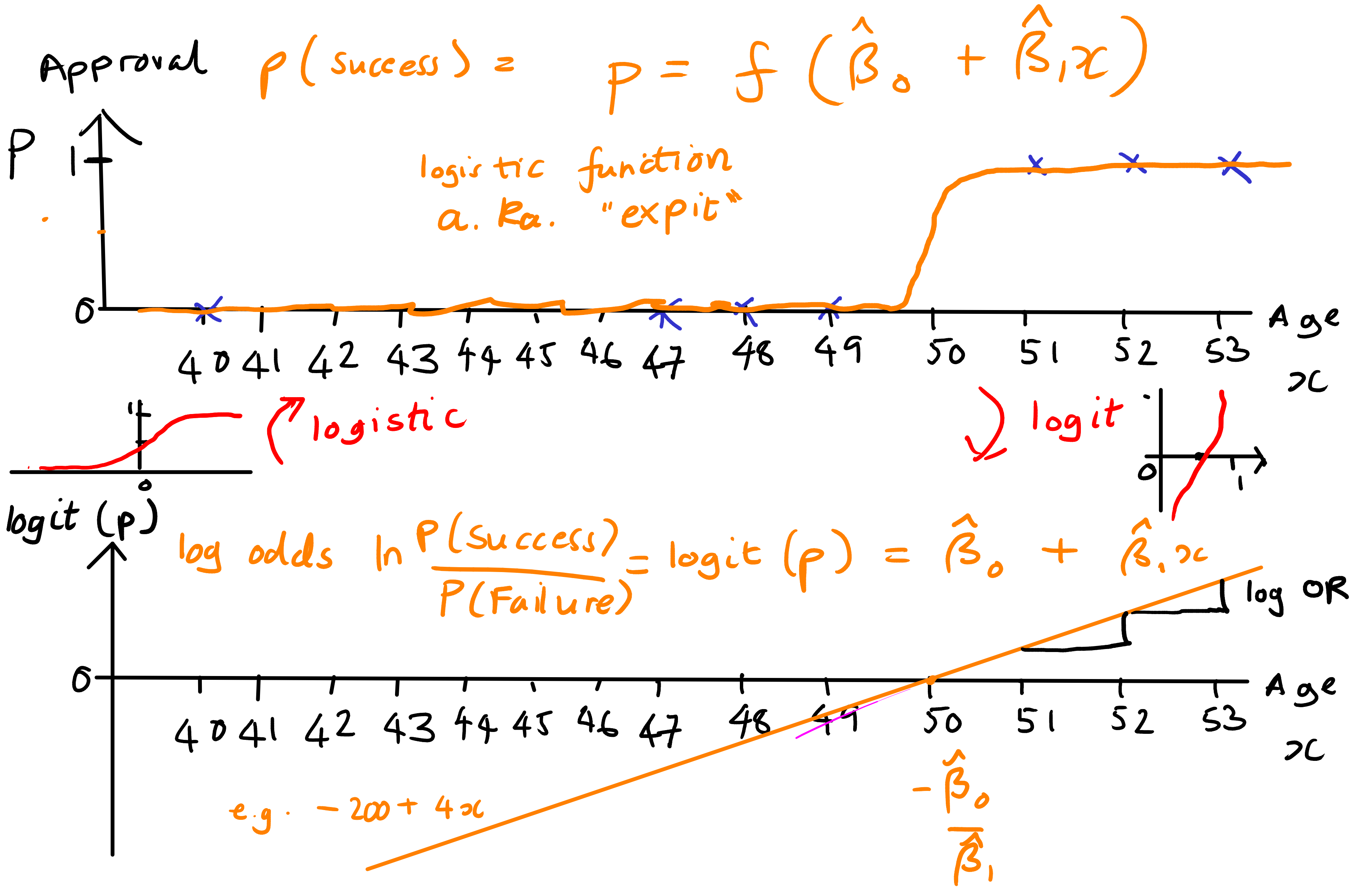
Data sources:
- Edinburgh Just Eat Bikes data 2020, on on OpenData Scotland
- Edinburgh temperature observations, Met Office via MIDAS



Where we're at in the Maximum Likelihood Principle and Regression

Logistic regression

Linear regression



Overview

Today

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

Wednesday

- Max likelihood with non-normal distributions
- Generalised linear regresion, including logistic regression

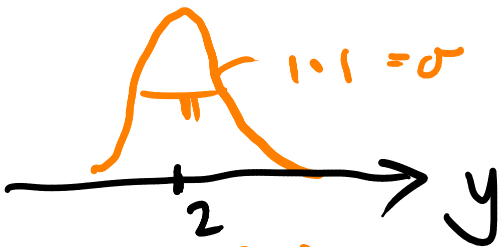
Inf2 - Foundations of Data Science: Regression and inference - The maximum likelihood principle

THE UNIVERSITY of EDINBURGH informatics

FOUNDATIONS OF DATA SCIENCE

Intuition for maximum likelihood principle

Statistical model, e.g. normal dist.



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

Alternative notation

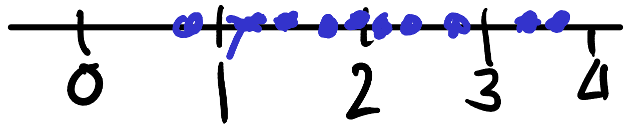
$$y \sim \mathcal{N}(2, 1.1^2)$$

normal dist $\quad \mu \quad \sigma^2$

sample $\longrightarrow$

Data

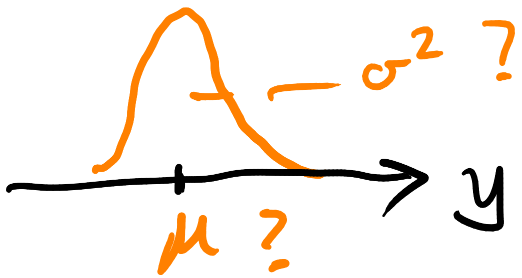
"Fake data"



"y is drawn from a normal dist with $\mu = 2$ and $\sigma^2 = 1.1^2$"

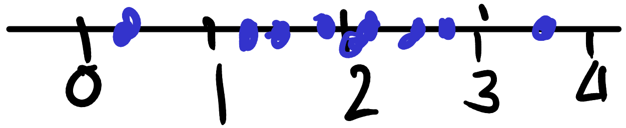
Intuition for maximum likelihood principle

Statistical model, e.g. normal dist.



infer

Data



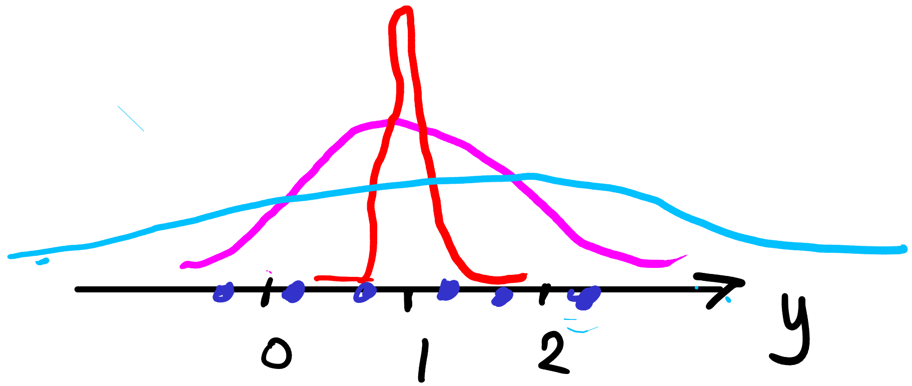
$$p(y \mid \underbrace{\mu, \sigma^2}_{\text{parameters}}) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}}$$

Given a normal distribution, what parameters are most likely to have generated 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. $y_i \sim \mathcal{N}(1, 0.1)$
3. $y_i \sim \mathcal{N}(1, 5)$
4. $y_i \sim \mathcal{N}(1, 1)$

Definition of Principle of Maximum Likelihood

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 : $\{y_1, \dots, y_n\}$

Likelihood $P(\underline{Y} = y_1, \dots, y_n \mid \vartheta_1, \dots, \vartheta_m)$

Normal : $\vartheta_1 = \mu$

$\vartheta_2 = \sigma^2$

Inf2 - 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 a normal distribution

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

Assumption 1 :

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

$$\times P(Y = y_2 \mid \mu, \sigma^2)$$

$$\times \ldots$$

$$\times P(Y = y_n \mid \mu, \sigma^2)$$

More compact notation...

Likelihood

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

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

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

Likelihood as a function of parameters

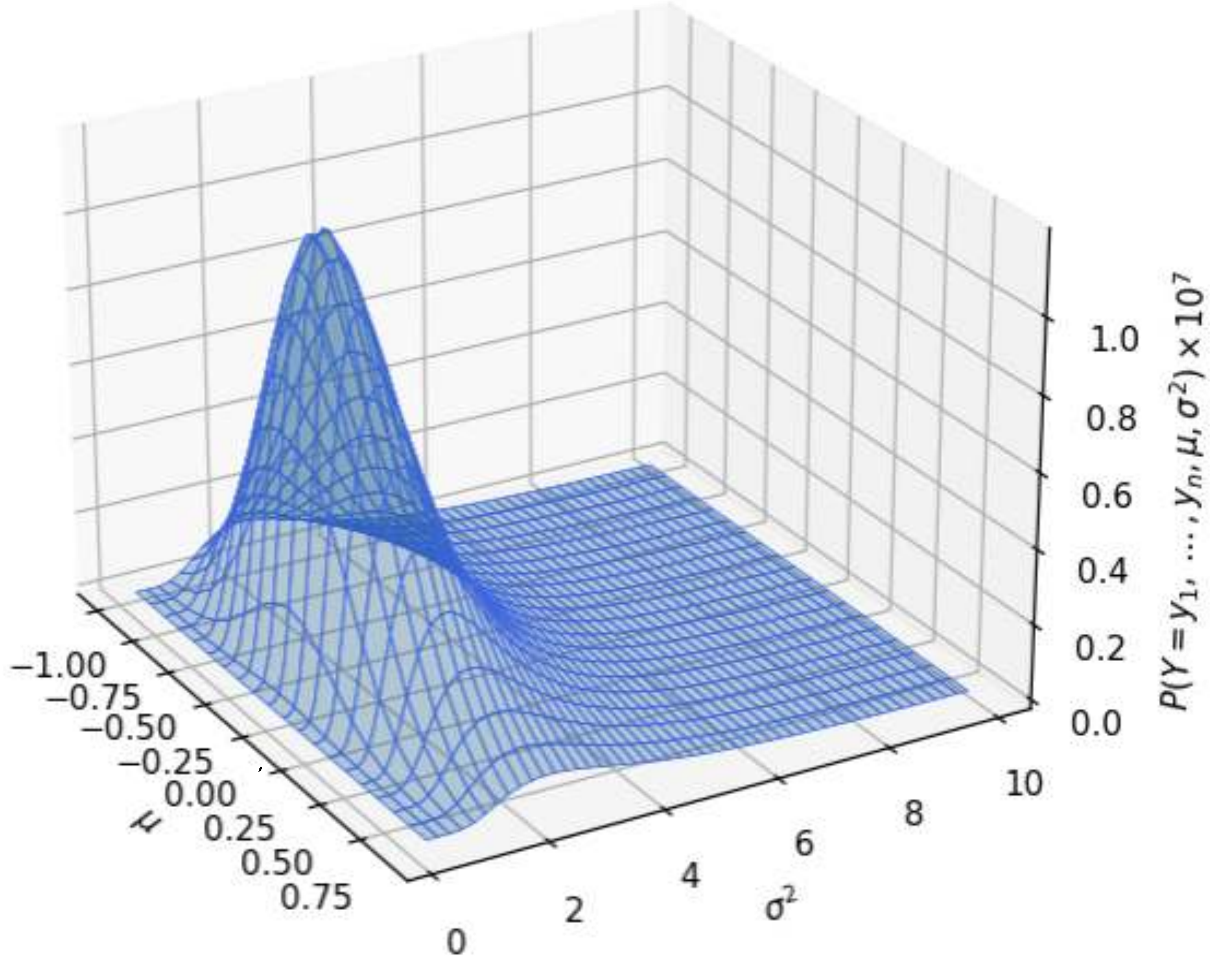
Data:

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

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

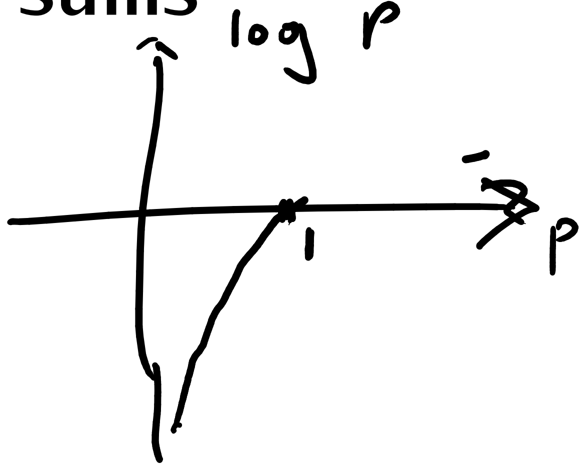
	Data
y_1	1.624345
y_2	-0.611756
y_3	-0.528172
y_4	-1.072969
y_5	0.865408
y_6	-2.301539
y_7	1.744812
y_8	-0.761207
y_9	0.319039
y_{10}	-0.249370

[code]



$P(Y = y_1, \ldots, y_n \mid \mu, \sigma^2)$

Log-likelihood equations: products to sums



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

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

$$\ln \left(\prod_{i=1}^{n} p_i \right) = \sum_{i=1}^{n} \ln p_i$$

Log likelihood

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

The log of the normal distribution

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

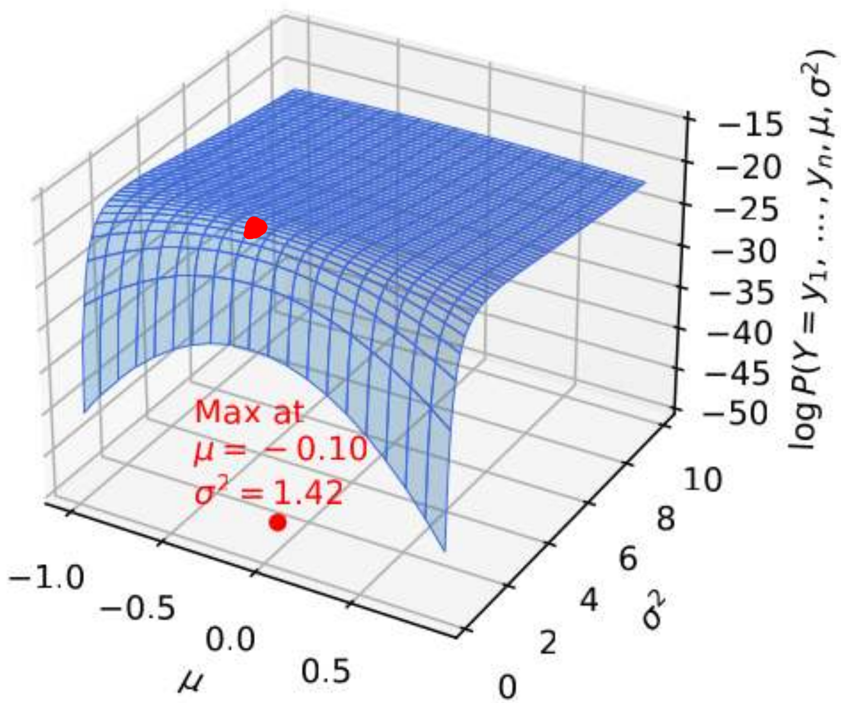
$$= \ln\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right) + \left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right)$$

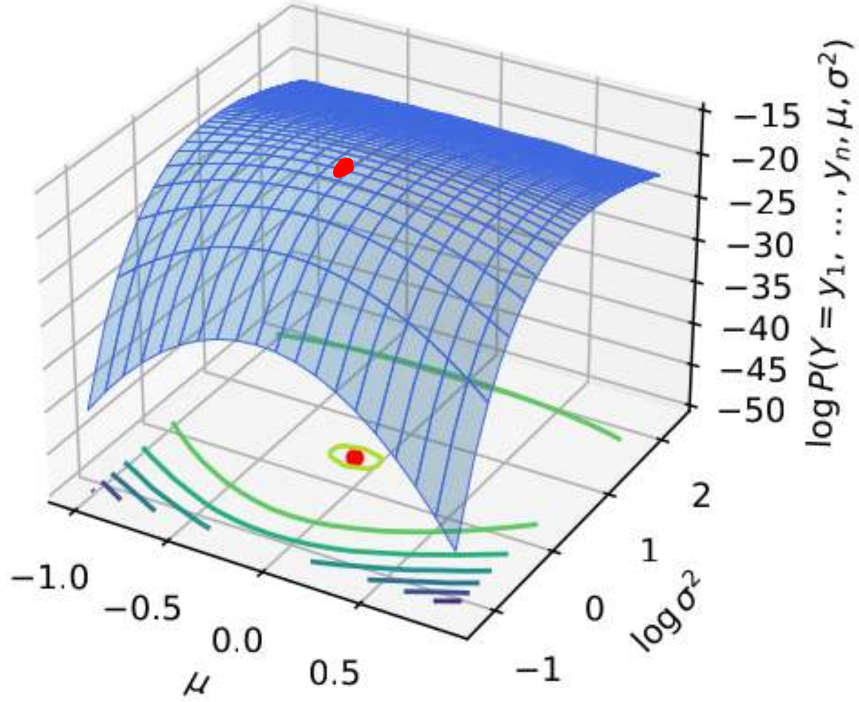
$$= -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2$$

Log-likelihood of data drawn from a normal distribution as a function of parameters

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

$$= \sum_{i=1}^{n} \Big(\underbrace{-\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2} \frac{(y_i - \mu)^2}{\sigma^2}}_{\text{previous}} \Big)$$





Maximum likelihood estimates of parameters

$$\ln P(\underline{Y} = y_1, \ldots, y_n \mid \mu, \sigma^2)$$
$$= \sum_{i=1}^{n} \left(-\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2} \frac{(y_i - \mu)^2}{\sigma^2} \right)$$

Maximise w.r.t. μ and σ^2 to give max 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

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

.

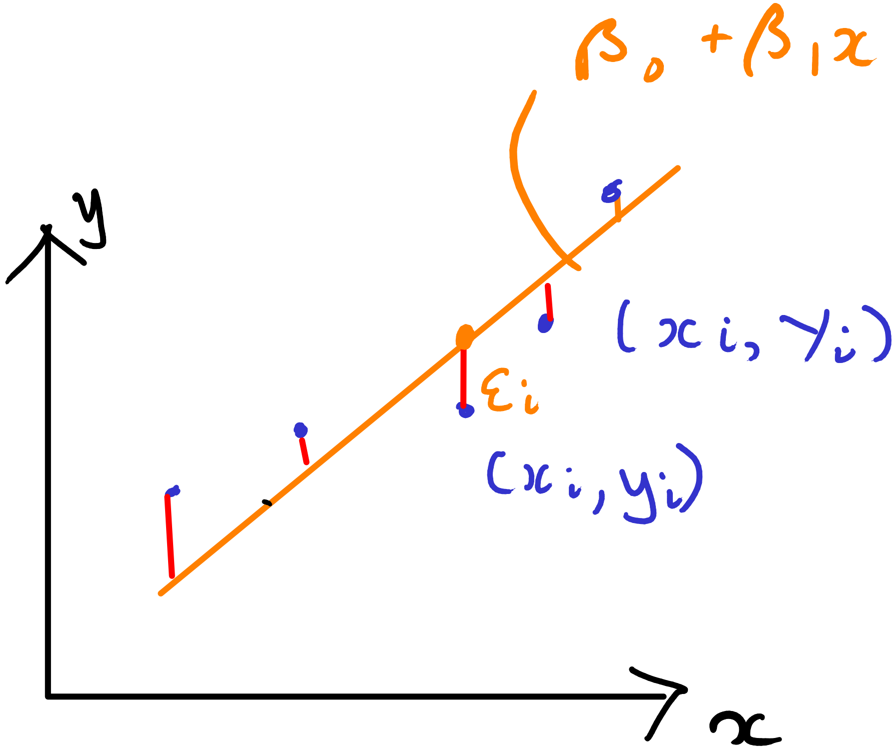
.

Inf2 - Foundations of Data Science: Regression and inference - Application of the maximum likelihood principle to linear regression

Application of max likelihood to linear regression

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

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



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

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

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

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

Relationship to ordinary least squares

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

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

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

Estimate of coefficients

- Analytical solutions for $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$ that maximise likelihood
- $\hat{\beta}_0$ and $\hat{\beta}_1$: as per ordinary least squares
- Variance of residuals:

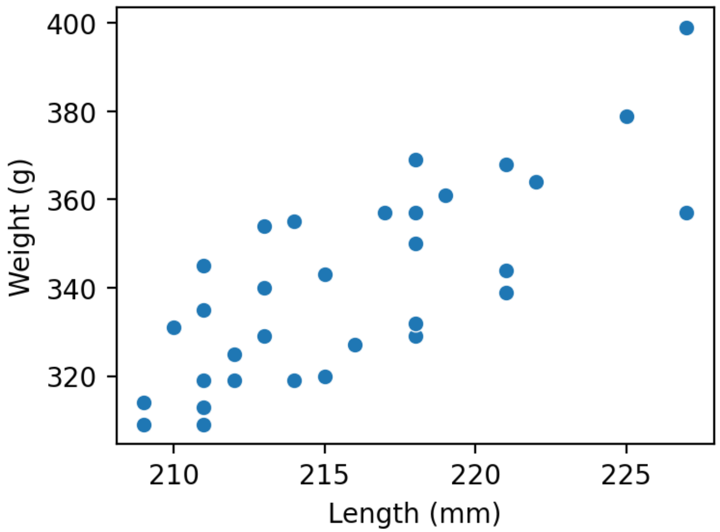
$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

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

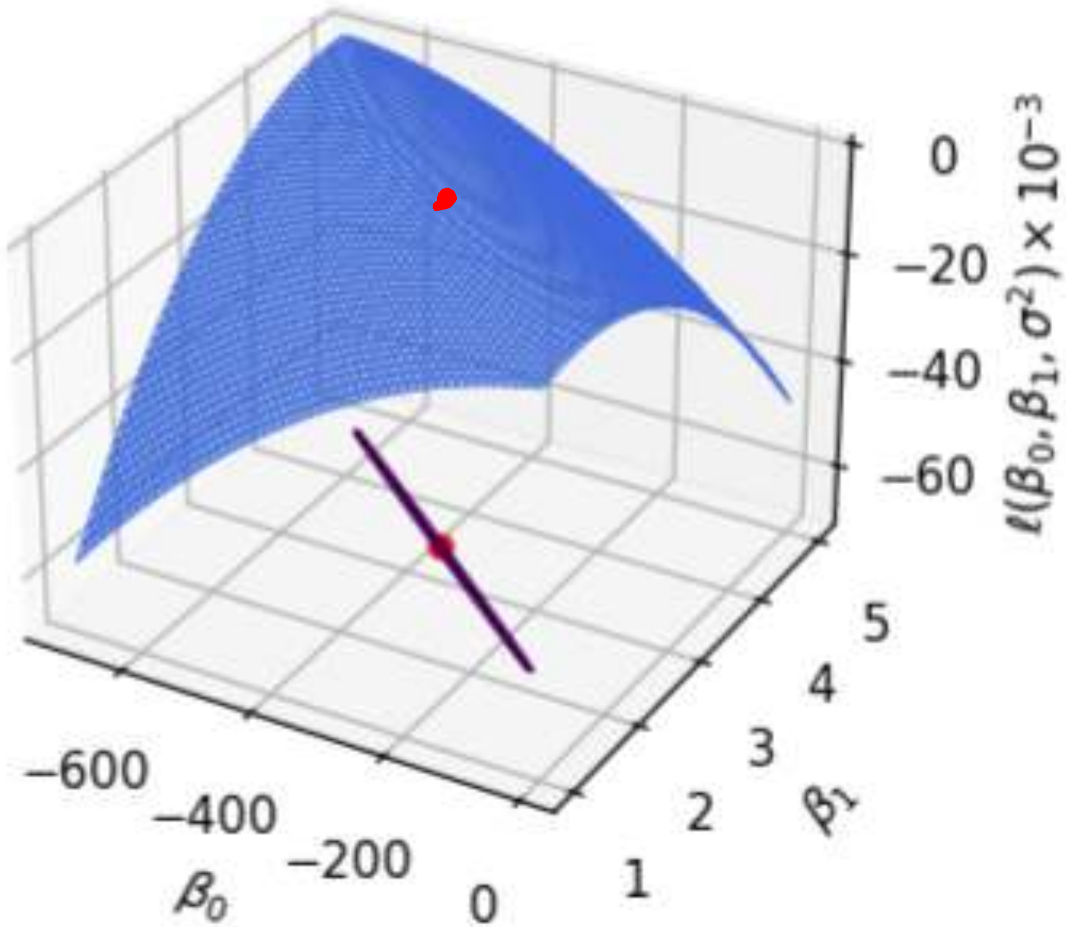
Sampling theory $\hat{\sigma}^2 = \frac{SSE}{n-2} \leftarrow$ Unbiased

Log likelihood of coefficients

Peter Trimming, Wikimedia Commons, CC BY 2.0

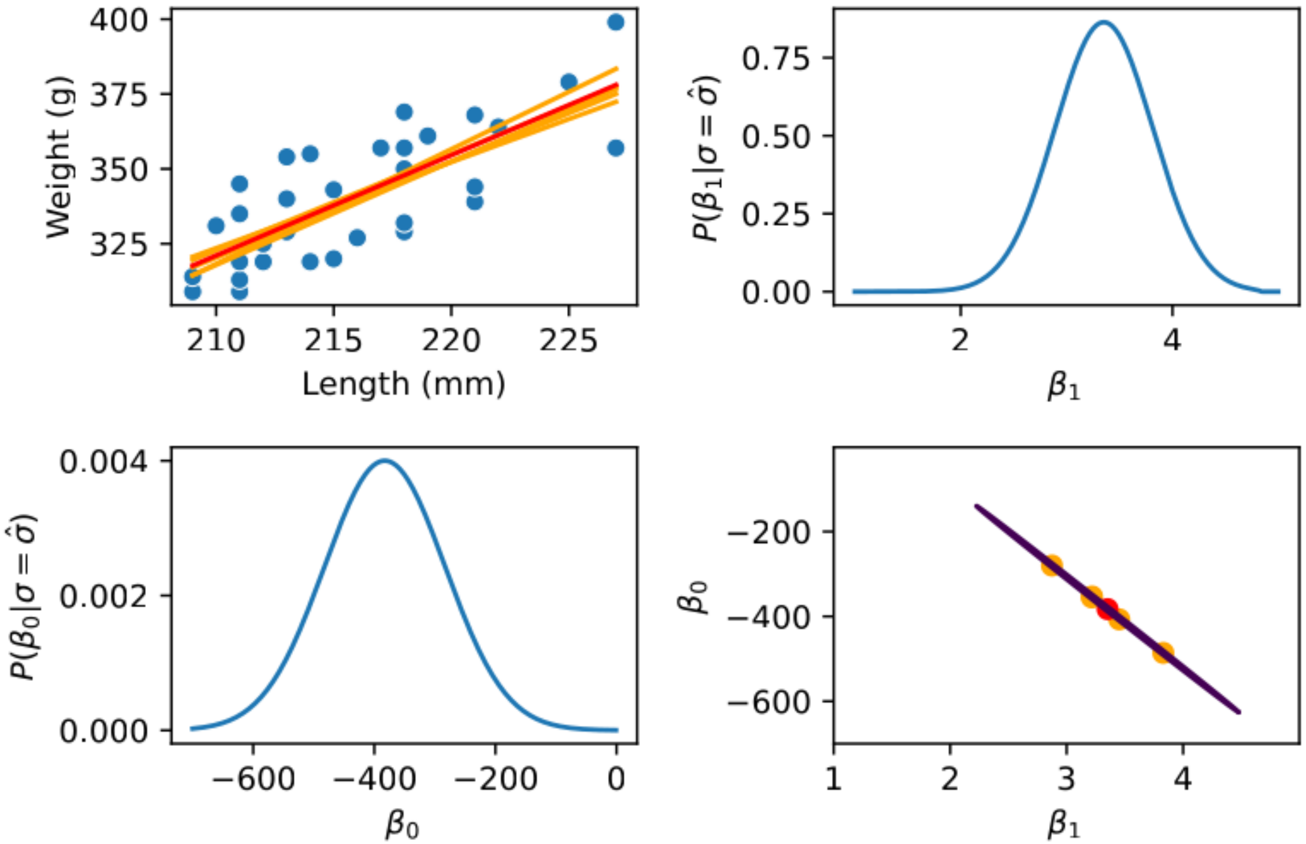


Data from Wauters and Dhondt 1989

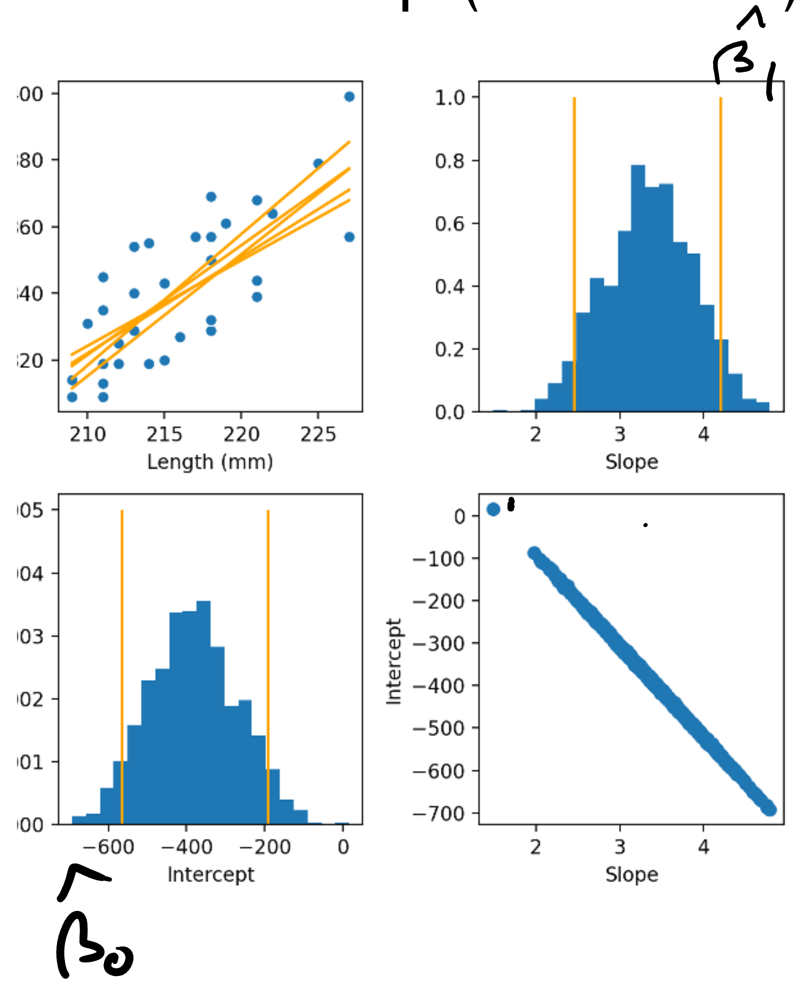


Parameter uncertainty

Max likelihood



cf Bootstrap (last week)



Data from Wauters and Dhondt 1989

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 an hour and the mean temperature during that hour

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