Likelihood and maximum likelihood estimation

The ASTA team

Contents

1 Maximum likelihood estimation of a probability

1.1 Estimating a probability

- Assume that we want to estimate a probability p of a certain event, e.g.
  - the probability that a bank customer will default their loan
  - the probability that a customer will buy a certain product
- We take a sample of n observations $Y_1, \ldots, Y_n$, where
  - $Y_i = 1$ if the event happens,
  - $Y_i = 0$ if the event does not happen,
  - The Y_i, $i = 1, \ldots, n$, are independent random variables with $P(Y_i = 1) = p$.
- Let $X = \sum_i Y_i$ be the number of ones in our sample. The natural estimate for p is

$$\hat{P} = \frac{X}{n}.$$

  - Theoretical justification?

1.2 The likelihood function

- Idea: choose $\hat{P}$ to be the value of p that makes our observations *as likely as possible.*
- Suppose we have observed $Y_1 = y_1, \ldots, Y_n = y_n$. The probability of observing this is

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = P(Y_1 = y_1) \cdot \cdots \cdot P(Y_n = y_n).$$

- Note that

$$P(Y_i = y_i) = \begin{cases} p, & y_i = 1, \\ (1-p), & y_i = 0. \end{cases}$$

- Therefore, if we let $x = \sum_i y_i$ be the number of 1's in our sample,

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = p^x \cdot (1-p)^{(n-x)}.$$

- This probability depends on the value of p. We may think of it as a function

$$L(p) = P(Y_1 = y_1, \ldots, Y_n = y_n) = p^x \cdot (1-p)^{(n-x)}.$$

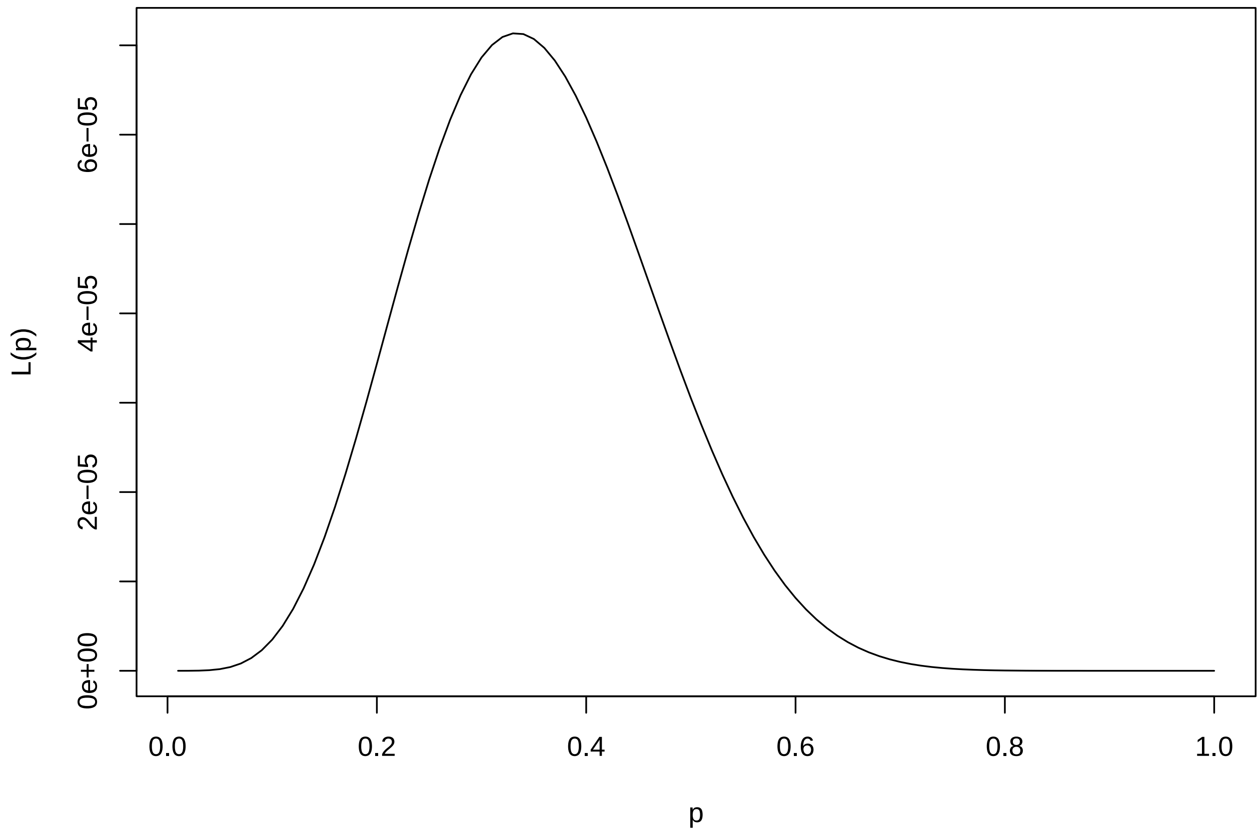
  - This is called the **likelihood function**.
- The **maximum likelihood estimate** $\hat{p}$ is the value of p that maximizes the likelihood function.

1.3 Likelihood function - example

- **Example:** Suppose we take a sample of $n = 15$ observations. We observe 5 ones and 10 zeros. The likelihoodfunction becomes

$$L(p) = p^5(1-p)^{10}$$

- We plot the graph of $L(p)$:



- The probability of our observations seems to be largest when p is around 1/3.

1.4 The log-likelihood function

- We seek the value of p that maximizes the likelihood function

$$L(p) = p^x \cdot (1-p)^{(n-x)}.$$

- Recall that $\ln(x)$ is a strictly increasing function.
- The value of p that maximizes $L(p)$ also maximizes $\ln(L(p))$.
- This is the **log-likelihood function**

$$l(p) = \ln(L(p)) = x\ln(p) + (n-x)\ln(1-p).$$

- It is often easier to maximize the log-likelihood function.

1.5 Maximum likelihood estimation

- In order to maximize

$$l(p) = x\ln(p) + (n-x)\ln(1-p),$$

we differentiate

$$l'(p) = \frac{x}{p} - \frac{n-x}{1-p}.$$

- The maximum must be found in a point with $l'(p) = 0$. Thus, we solve

$$l'(p) = \frac{x}{p} - \frac{n-x}{1-p} = 0.$$

- Multiply by $p(1-p)$ to get

$$x(1-p) - (n-x)p = 0$$
$$x - xp - np + xp = 0$$
$$x = np$$
$$p = \frac{x}{n}.$$

- Note that this must indeed be a maximum point since

$$\lim_{p \to 0} l(p) = \lim_{p \to 1} l(p) = -\infty.$$

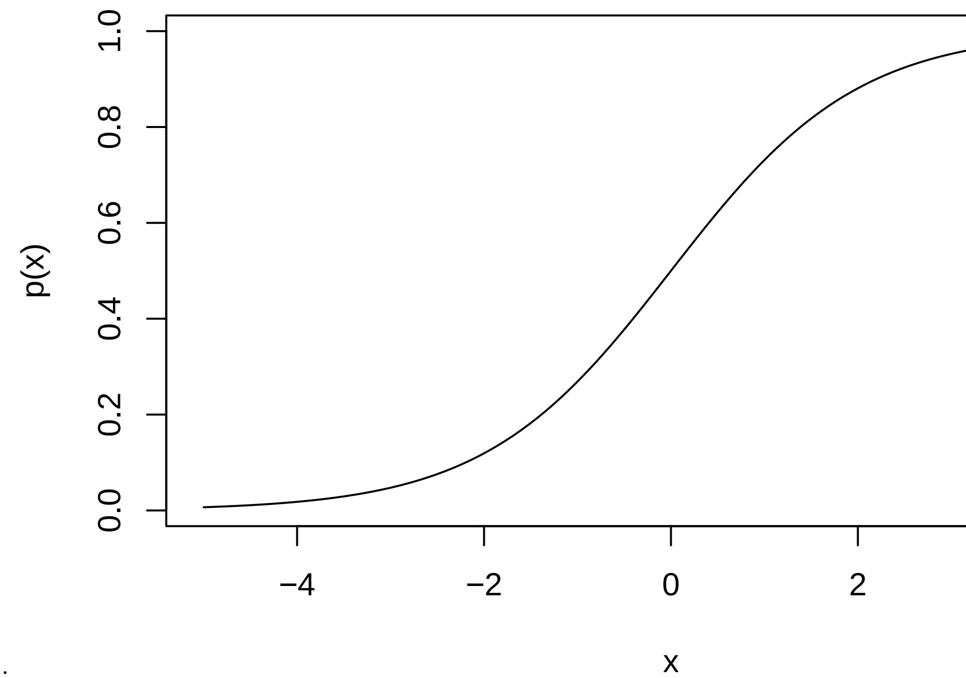
- Our **maximum likelihood estimate** of p is $\hat{p} = \frac{x}{n}$.

2 Maximum likelihood for logistic regression

2.1 The logistic regression model

- Estimation of a probability was a simple use of maximum likelihood estimation, which could easily have been treated by more direct methods.
- **Logistic regression** is a more complex case, where we want to model a probability $p(x)$ that depends on a predictor variable x.
  - E.g. the probability of a customer buying a certain product as a function of their monthly income.
- In logistic regression, $p(x)$ is modelled by a logistic function

$$p(x) = \frac{1}{1 + e^{-(\alpha + \beta x)}}.$$



- Graph of $p(x)$ when $\alpha = 0$ and $\beta = 1$.
  - α determines how steep the graph is.
  - β shifts the graph along the x-axis.
- How to estimate α and β?

2.2 Maximum likelihood estimation for logistic regression

- A sample consists of $(x_1, y_1), \ldots, (x_n, y_n)$, where x_i is the predictor and y_i is the response, which is either 0 or 1.
- The probability of our observations is

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_i p(Y_i = y_i) = \prod_i p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

  since

$$p(x_i)^{y_i}(1 - p(x_i))^{1-y_i} = \begin{cases} p(x_i), & y_i = 1, \\ 1 - p(x_i), & y_i = 0. \end{cases}$$

- Inserting what $p(x_i)$ is, we obtain a function of the unknown parameters α and β:

$$L(\alpha, \beta) = P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_i \left(\frac{1}{1 + e^{-(\alpha+\beta x_i)}}\right)^{y_i} \left(1 - \frac{1}{1 + e^{-(\alpha+\beta x_i)}}\right)^{1-y_i}$$

- Again, it is easier to maximize the log-likelihood

$$l(\alpha, \beta) = \sum_i \Big(y_i \ln(p(x_i)) + (1 - y_i) \ln(1 - p(x_i))\Big).$$

- However, this maximum can only be found using numerical methods.

2.3 Logistic regression - example

- We consider a dataset from the ISLR package on whether or not 10000 bank costumers will default their loans.
  - Response: default (1=yes, 0=no)
  - Predictor: income

```
library(ISLR)
x<-Default$income/10000 # Annual income in 10000 dollars
y<-as.numeric(Default$default=="Yes") # Loan default, 1 means "Yes"
```

- We want to model the probability of default as a logistic function of income

$$p(x) = \frac{1}{1 + e^{-(\alpha+\beta x)}}.$$

2.4 Logistic regression - example continued

- We make a function in R that computes the log-likelihood function as a function of the vector $\theta = (\alpha, \beta)^T$.
  - We first compute a vector px that contains all the probabilities $p(x_i)$.
  - Then we compute the vector logpy which contains all the $\ln(P(Y_i = y_i)) = y_i \ln(p(x_i)) + (1 - y_i) \ln(1 - p(x_i))$.
  - Finally, we compute the log-likelihood with the formula

$$l(\alpha, \beta) = \sum_i \Big(y_i \ln(p(x_i)) + (1 - y_i) \ln(1 - p(x_i)) \Big)$$

```
loglik <- function(theta) {
alpha=theta[1]
beta=theta[2]
px<-1/(1+exp(-alpha-beta*x))
logpy<-y*log(px) + (1-y)*log(1-px)
sum(logpy)
}
loglik(c(2,2))
```

```
## [1] -84250
```

- We maximize the log-likelihood using the `optim()` function in R.
  - It needs an initial guess of θ. Here we use `c(2,2)`.
  - The option `control=list(fnscale=-1)` ensures that we maximize rather than minimize.

```
optim(c(2,2),loglik,control=list(fnscale=-1))
```

```
## $par
## [1] -3.099 -0.081
##
## $value
## [1] -1458
##
## $counts
## function gradient
##       69       NA
##
## $convergence
## [1] 0
##
```

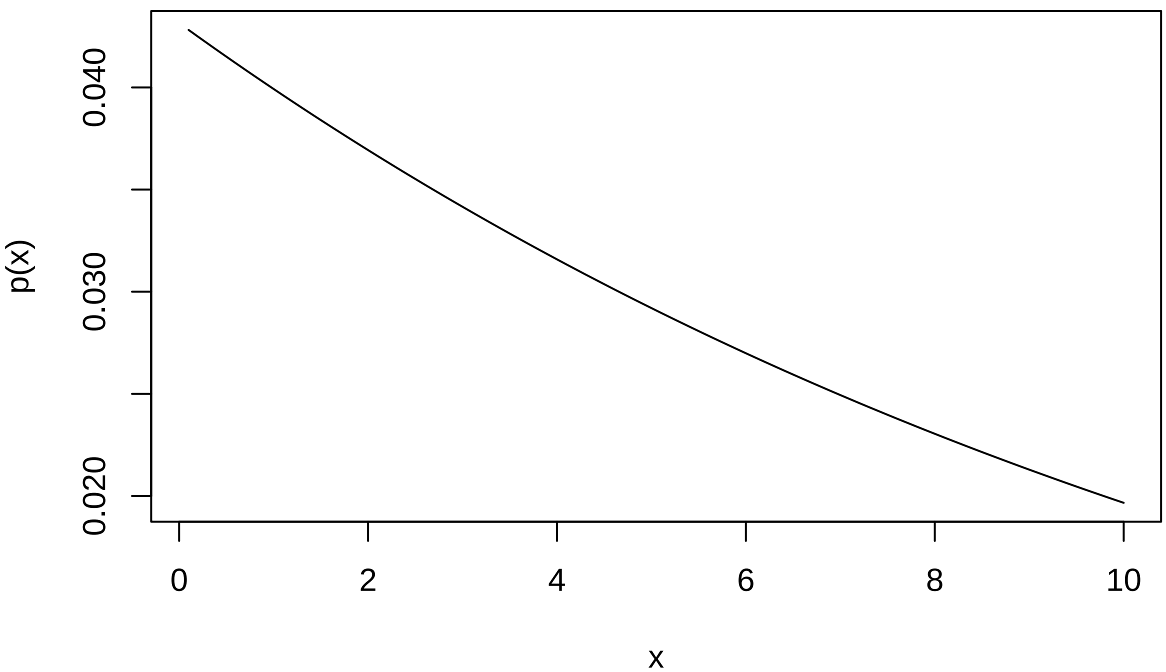
```
## $message
## NULL
```

- We obtain the maximum likelihood estimates $\hat{\alpha} = -3.099$ and $\hat{\beta} = -0.081$.

2.5 Logistic regression - example continued

- We can plot the estimated logistic function

$$\hat{p}(x) = \frac{1}{1 + e^{3.099+0.081x}}.$$



- The maximum likelihood estimates of α and β can be found directly using R:

```
model<-glm(y~x,family="binomial")
summary(model)
```

```
##
## Call:
## glm(formula = y ~ x, family = "binomial")
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.0941     0.1463  -21.16   <2e-16 ***
## x            -0.0835     0.0421   -1.99    0.047 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 2916.7  on 9998  degrees of freedom
## AIC: 2921
```

```
##
## Number of Fisher Scoring iterations: 6
```

3 Maximum likelihood estimation with continuous variables

3.1 The probability density function

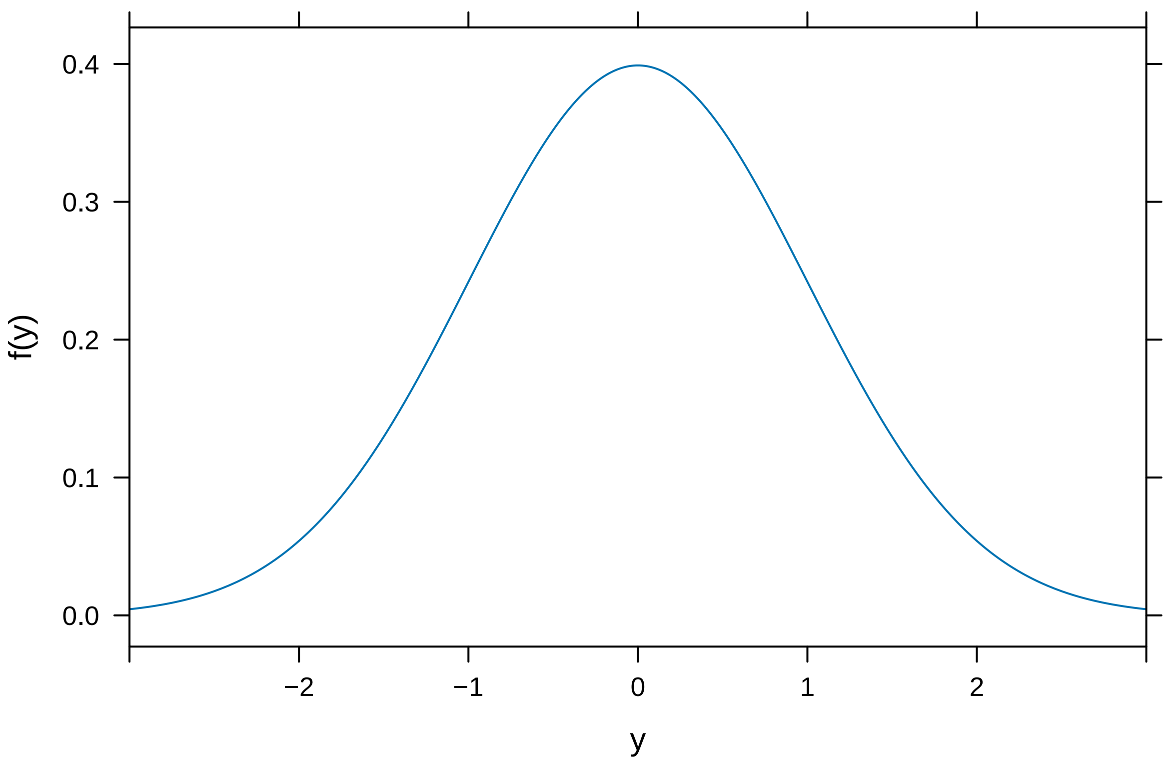
- Suppose we have a sample $Y_1, \dots, Y_n$ of independent variables with

$$Y_i \sim N(\mu, \sigma)$$

- We would like to estimate the unknown parameters μ and σ.
- For a continuous variable Y we have $P(Y = y) = 0$ for all y.
  - We cannot use the probability of observing a given outcome to define the likelihood function.
- Instead we consider the probability density function

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]$$

- E.g. for $\mu = 0$ and $\sigma = 1$:



- The most likely values are the ones where $f(y)$ is large.
- Thus we will use $f(y)$ as a measure of how likely it is to observe $Y = y$.

3.2 The likelihood function for n observations

- Since $Y_1, \ldots, Y_n$ are independent observations, the joint density function becomes a product of marginal densities:
$$f_{(Y_1,\ldots,Y_n)}(y_1, \ldots, y_n) = \prod_i f_{Y_i}(y_i).$$
- If we have observed a sample $Y_1 = y_1, \ldots, Y_n = y_n$, our likelihood function is defined as

$$L(\mu, \sigma) = f_{(Y_1,\ldots,Y_n)}(y_1, \ldots, y_n) = \prod_i f_{Y_i}(y_i) = \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}. \tag{1}$$

- The maximum likelihood estimate $(\hat{\mu}, \hat{\sigma})$ is the value of (μ, σ) that maximizes the likelihood function.

3.3 Log-likelihood function in the normal case

- We found the log-likelihood function

$$L(\mu, \sigma) = \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}. \tag{2}$$

- Again it is easier to maximize the log-likelihood function.

$$l(\mu, \sigma) = \ln(L(\mu, \sigma)) = \sum_i \ln\Big(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}\Big)$$
$$= \sum_i \Big(-\ln(\sigma\sqrt{2\pi}) - \frac{(y_i-\mu)^2}{2\sigma^2}\Big) = -n\ln(\sigma\sqrt{2\pi}) - \sum_i \frac{(y_i-\mu)^2}{2\sigma^2}$$

- We find the partial derivatives and set them equal to 0. First with respect to μ:

$$\frac{\partial}{\partial\mu} l(\mu, \sigma) = \sum_i \frac{2(y_i-\mu)}{2\sigma^2} = \frac{1}{\sigma^2}\sum_i (y_i - \mu) = \frac{1}{\sigma^2}\Big(\sum_i y_i - n\mu\Big) = 0$$

- Vi får at $n\mu = \sum_i y_i$, så $\mu = \frac{1}{n}\sum_i y_i = \bar{y}$.
- Then with respect to σ:

$$\frac{\partial}{\partial\sigma} l(\mu, \sigma) = -\frac{n}{\sigma} + \sum_i \frac{(y_i-\mu)^2}{\sigma^3} = 0$$

- We multiply by σ and insert $\mu = \bar{y}$:

$$-n + \sum_i \frac{(y_i-\bar{y})^2}{\sigma^2} = 0$$
$$n = \frac{1}{\sigma^2}\sum_i (y_i-\bar{y})^2$$
$$\sigma^2 = \frac{1}{n}\sum_i (y_i-\bar{y})^2$$

- In total we get the maximum likelihood estimates:

$$\hat{\mu} = \frac{1}{n}\sum_i y_i = \bar{y}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i (y_i - \bar{y})^2$$

3.4 Numerical solution - normal distribution

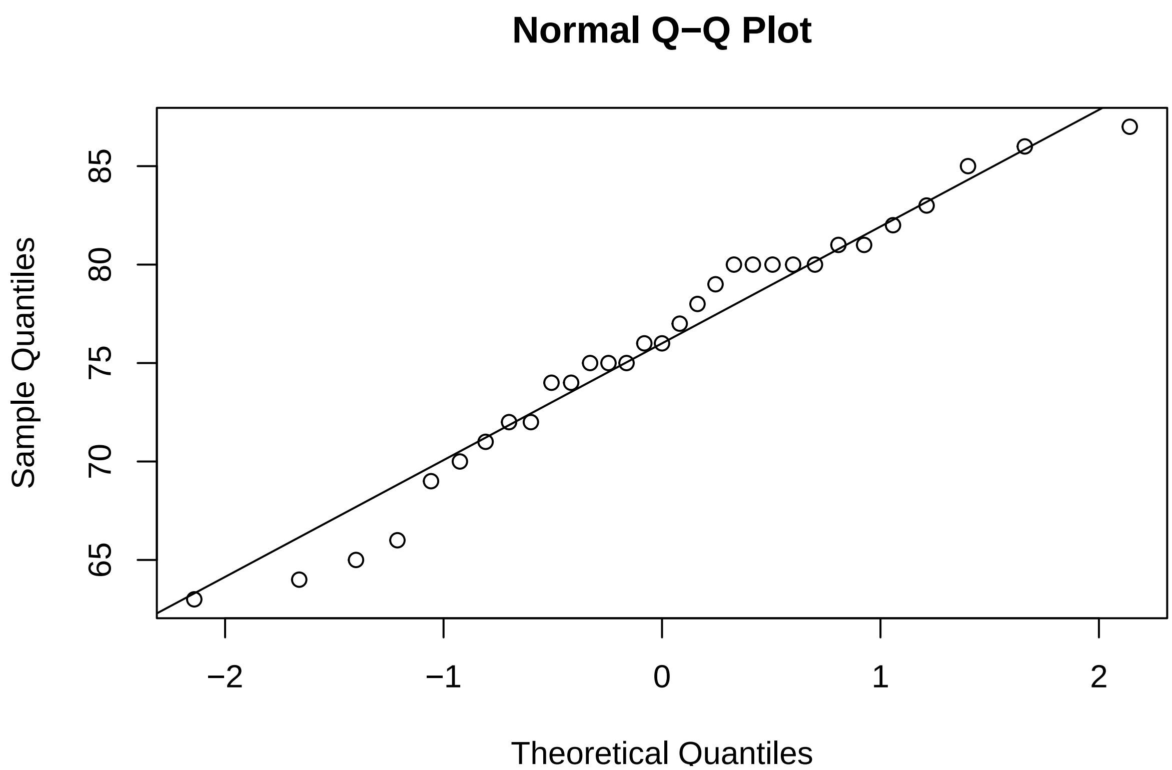
- The maximum likelihood estimates can also be found numerically. We consider again the `trees` data.

```
trees <- read.delim("https://asta.math.aau.dk/datasets?file=trees.txt")
head(trees)
```

```
##   Girth Height Volume
## 1   8.3     70     10
## 2   8.6     65     10
## 3   8.8     63     10
## 4  10.5     72     16
## 5  10.7     81     19
## 6  10.8     83     20
```

- We will assume that the variable `Height` is normally distributed.

```
qqnorm(trees$Height)
qqline(trees$Height)
```



3.5 Numerical solution - normal distribution

- We define the log-likelihood as a function of the parameter vector $\theta = (\mu, \sigma)^T$.

- `dnorm(y, mean = mu, sd = sigma)` gives the normal density $f(y)$ with mean μ and standard deviation σ evaluated at y.

```
loglik_normal <- function(theta) {
  mu <- theta[1]
  sigma <- theta[2]
  y<-trees$Height
  fy<-dnorm(y , mean = mu, sd = sigma)
  sum(log(fy))
}
loglik_normal(c(1,5))
```

```
## [1] -3590
```

- We maximize again using `optim()`:

```
optim(c(1, 5), loglik_normal,control=list(fnscale=-1))
```

```
## $par
## [1] 76.0  6.3
##
## $value
## [1] -101
##
## $counts
## function gradient
##      103       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

- We can compare this to the theoretical formulas for the maximum likelihood estimates:

$$\hat{\mu} = \bar{y},$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_i (y_i - \bar{y})^2} = \sqrt{\frac{n-1}{n}}s.$$

```
mean(trees$Height)
```

```
## [1] 76
```

```
sd(trees$Height)
```

```
## [1] 6.4
```

```
n <- length(trees$Height)
sd(trees$Height)*sqrt((n-1)/n)
```

```
## [1] 6.3
```

4 Properties of maximum likelihood estimators

- Suppose $\theta \in \mathbb{R}$ is a parameter that we estimate by $\hat{\theta}$ using maximum likelihood estimation. Then (under suitable conditions) one may show the following mathematically.

- **Consistency:** For all $\varepsilon > 0$,
$$\lim_{n\to\infty} P(|\theta - \hat{\theta}| > \varepsilon) = 0$$
- **Central limit theorem:** When $n \to \infty$,
$$\sqrt{n}(\hat{\theta} - \theta) \to N(0, \sigma_\theta^2).$$
That is, for large n,
$$\sqrt{n}(\hat{\theta} - \theta) \approx N(0, \sigma_\theta^2),$$
or equivalently,
$$\hat{\theta} \approx N\Big(\theta, \frac{\sigma_\theta^2}{n}\Big).$$