MATH5743M: Statistical Learning - Lecture 3

Dr Seppo Virtanen, s.virtanen@leeds.ac.uk

Semester 2: 7 Feb 2021

Outline

Linear regression

- Fundamental statistical tool
- Simple, yet powerful

Linear model in one dimension

- Likelihood and MLE
- glm command in R

Linear regression model

For a one dimensional linear regression problem, we assume that the output, Y is a sum of three components:

- A scalar constant – β_0
- The input, X multiplied by a second constant – $\beta_1 x$
- A residual ϵ, which is unknown, but assumed to come from a Normal distribution with zero mean and unknown variance: $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Therefore we may write the model in full as:

$$Y = \beta_0 + \beta_1 X + \epsilon, \ \epsilon \sim \mathcal{N}(0, \sigma) \tag{1}$$

The likelihood function

Each output value is generated *independently* from a normal distribution with some variance σ^2, and with a mean that is a linear function of the associated *input*

$$\begin{aligned}
\mathcal{L}(\beta_0, \beta_1, \sigma) &= P(\mathrm{Data} \mid \beta_0, \beta_1, \sigma) \\
&= P(Y = y \mid X = x, \beta_0, \beta_1, \sigma) \\
&= \mathcal{N}(y; \beta_0 + \beta_1 x, \sigma^2) \\
&= \prod_{i=1}^{n} \mathcal{N}(y_i; \beta_0 + \beta_1 x_i, \sigma^2)
\end{aligned} \tag{2}$$

Note, importantly, that the final product over the n data points is possible because we assumed that each output was generated *independently*. Therefore the probability for all of the data points is simply the product of the probability of observing each, individually.

The log-likelihood function

Log-likelihood:

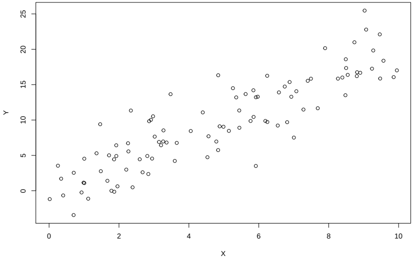
$$\log \mathcal{L} = \sum_{i=1}^{n} \log \mathcal{N}(y_i; \beta_0 + \beta_1 x_i, \sigma^2) \tag{3}$$

Easier to work with.

Example

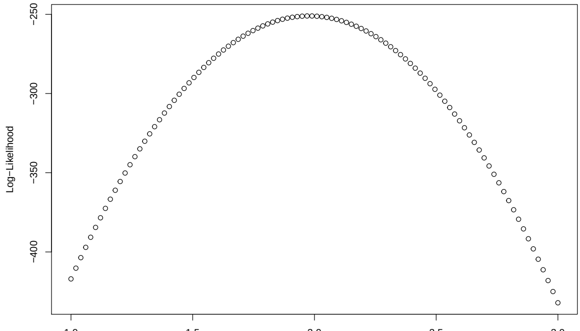
$$Y = \beta X + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2 = 3^2),\ \beta = 2 \tag{4}$$

```r
#Define the number of data points
N = 100
X = runif(n=N, min=0, max=10)
Y = 2*X + rnorm(n=N, mean=0, sd= 3)
plot(X, Y)
```



Numeric optimisation

```
#Define the guesses
N = 100; LL = rep(NA, N); beta=seq(1, 3, length.out=N)
#Calculate the log-likelihood for each beta
for (i in 1:N){
  LL[i] = sum(dnorm(x=Y, mean=beta[i]*X, sd=3, log=TRUE))
}
#This time plot the log-likelihood as points
plot(beta, LL, xlab="Beta", ylab="Log-Likelihood")
```



optim in R

Minimise the *negative* log-likelihood function.

```
neg_LL <- function(param){-sum(dnorm(x=Y, mean=param*X,
                                     sd=3, log=TRUE))}
optim_result = optim(par = 1, fn=neg_LL)
```

```
## Warning in optim(par = 1, fn = neg_LL): one-dimensional
## use "Brent" or optimize() directly
```

```
print(paste("MLE estimate of beta is: ", optim_result$par,
            collapse=""))
```

```
## [1] "MLE estimate of beta is:  1.9783203125"
```

Analytic optimisation

From the definition of the log-likelihood we get

$$
\begin{aligned}
\log \mathcal{L}(\beta_0, \beta_1, \sigma) &= \sum_{i=1}^{n} \log \mathcal{N}(y_i; \beta_0 + \beta_1 x_i, \sigma^2) \\
&= \sum_{i=1}^{n} \log \left(\frac{1}{\sqrt{2\pi}\sigma} \exp \left(\frac{-(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right) \right) \\
&= \sum_{i=1}^{n} -\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} - n \log(\sigma) - \frac{n}{2} \log 2\pi
\end{aligned}
\tag{5}
$$

Analytic optimisation

First, lets look at β_0. The condition for a maximum is:

$$\frac{\partial \log \mathcal{L}}{\partial \beta_0}|_{\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}} = 0 \quad (6)$$

$$\sum_{i=1}^{n} \frac{(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)}{\hat{\sigma}^2} = 0$$

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (7)$$

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

Analytic optimisation

Now lets consider β_1. The condition for a maximum is:

$$\frac{\partial \log \mathcal{L}}{\partial \beta_1}\Big|_{\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}} = 0 \tag{8}$$

therefore

$$\begin{aligned}
\sum_{i=1}^{n} \frac{(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i}{\hat{\sigma}^2} &= 0 \\
\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i &= 0 \\
\overline{yx} - \hat{\beta}_0 \bar{x} - \hat{\beta}_1 \overline{x^2} &= 0 \\
\overline{yx} - \bar{y}\bar{x} + \hat{\beta}_1 \bar{x}^2 - \hat{\beta}_1 \overline{x^2} &= 0, \ [\text{using } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}] \\
\Rightarrow \hat{\beta}_1 &= \frac{\overline{yx} - \bar{y}\bar{x}}{\overline{x^2} - \bar{x}^2} \\
&= \frac{\mathrm{cov}(y, x)}{}
\end{aligned} \tag{9}$$

Analytic optimisation

Finally we consider σ^2. The condition for a maximum is

$$\frac{\partial \log \mathcal{L}}{\partial \sigma}\Big|_{\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma}} = 0 \tag{10}$$

substituting the expression for $\mathcal{L}$ we have:

$$\begin{aligned} \sum_{i=1}^{n} \frac{(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\hat{\sigma}^3} - \frac{n}{\hat{\sigma}} = 0 \\ \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \end{aligned} \tag{11}$$

Analytic optimisation

1. $\hat{\beta}_1 = \frac{\overline{yx} - \bar{y}\bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{\mathrm{COV}(y, x)}{\mathrm{VAR}(x)}$
2. $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
3. $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

glm in R

- GLM stands for Generalised Linear Model, which we will learn more about later in the course.
- tool for fitting linear models

Key steps

- set up your data in the correct way
- interpret the resulting model that is produced
- use it to make predictions

Data frames in R

Create two columns labelled input and output and assign them the values of X and Y.

```
mydata = data.frame(input=X, output=Y)
head(mydata)
```

```
##      input     output
## 1 2.840145  2.3665880
## 2 2.862926  9.8296964
## 3 5.846076 14.2017542
## 4 2.389436  0.4847947
## 5 1.715763  5.0062828
## 6 3.475958 13.6540596
```

```
mydata[3, 2]
```

```
## [1] 14.20175
```

Using glm

```r
mymodel = glm(output ~ input, data = mydata)
summary(mymodel)
```

```
##
## Call:
## glm(formula = output ~ input, data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.2118  -2.0721  -0.4281   1.7864   7.6801
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1742     0.5940   0.293     0.77
## input         1.9514     0.1063  18.357   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
```

Confidence intervals

The 95% C.I. for any regression coefficient in this one-dimensional model is:

estimate $\pm$ **standard error** $\times t_{2.5\%, n-2}$

A confidence interval for a model coefficient that does not contain zero indicates that we can reject the null hypothesis that the true coefficient is zero, just as with estimates for a population mean.

A confidence interval that places a coefficient far from zero indicates that the input in question provides some information about the output.