COMP6237 Linear Regression and Maximum Likelihood Estimation

February 9, 2021

Abstract

Solutions to the problem sheet for lecture on linear regression and maximum likelihood estimation. Will be discussed in the tutorial session.

1 Linear Regression I

A data set is constructed by taking 100 samples from a normal distribution with mean $\mu = 5$ and standard deviation $\sigma = 2$ to construct a random variable $X_i, i = 1,, 100$. Then, a 2nd random variable $Y_i, i = 1,, 100$ is constructed by taking the values of the corresponding X_i and adding one half of a third random variate drawn from a normal distribution with mean 5 and standard deviation 2 and thus a set of 100 pairs (X_i, Y_i) is obtained. Find the parameters of a linear regression of Y on X (both by doing the numerical experiment and by calculating the result analytically).

solution: The numerical experiment could be run in any software package you like. A program in R to do this could be:

```
> data.X <- rnorm(100,5,2)
> data.Y <- data.X + 0.5 * rnorm (100,5,2)
> reg <- lm (data.Y ~ data.X)
> summary (reg)
```

In my case the output I obtained was

```
Residuals:
     Min       1Q   Median       3Q      Max
-3.12731 -0.80066 -0.05604  0.81004  2.49546

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.66818    0.26573   10.04   <2e-16 ***
data.X       0.97903    0.04972   19.69   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.094 on 98 degrees of freedom
```

```
Multiple R-squared:  0.7983,Adjusted R-squared:  0.7962
F-statistic: 387.8 on 1 and 98 DF,  p-value: < 2.2e-16
```

The model is highly significant and I read a slope of $m = 0.98$ and intercept of $b = 2.67$ from this. One could write a script to repeat this experiment a certain number of times and would then see that the average result is consistent with $m = 1$ and $b = 2.5$.

The result can also be calculated analytically. Recall from the lecture that if we carry out regression of X on Y, then $m = Cov[X, Y]/V[X]$ and $b = E[Y] - mE[X]$ (where $E[.]$ stands for the expectation of the random variable). In the problem setting above, $X\ N(5, 2)$, (say) $Z\ N(5, 2)$, and $Y = X + 1/2Z$. Now we can write

$$
\begin{aligned}
m &= Cov[X, Y]/V[X] && (1)\\
&= Cov[X, X + 1/2Z]/V[X] && (2)\\
&= (Cov[X, X] + 1/2Cov[X, Z])/V[X] && (3)\\
&= V[X]/V[X] = 1, && (4)
\end{aligned}
$$

where I have made use of the linearity of $Cov[., .]$ and have exploited that $Cov[X, Y] = 0$ because X and Z are independent. Further, we have $b = E[X + 1/2Z] - 1 \times E[X] = 1/2E[Z] = 2.5$ (using that $Z\ N(5, 2)$). Both values are consistent with the numerical experiment.

2 Linear Regression II

Some person wants to conduct a least squares regression on a data set of N (X, Y) pairs, but attaches varying importance to deviations of various (X, Y) pairs to the line of best fit. The relative importance of deviations of pair (x_i, y_i) are given by a function $f(i) = f_i$. Find an expression for the line of best fit generated by this procedure.

solution:
Using the weight function, we basically want to minimize

$$E(m, b) = \sum_i f_i(y_i - mx_i - b)^2. \quad (5)$$

We can proceed in the usual way, by calculating derivatives with regard to m and b and equating them to zero. We have

$$\partial E/\partial m = 2\sum_i f_i(y_i - mx_i - b) = 0, \quad (6)$$

or $b = E_f[Y] - mE_f[X]$ (with $E_f[X] = \sum_i f_i x_i / \sum_i f_i$ defined as an f-weighted average). Inserting this into the expression for $E(m, b)$ we obtain:

$$E(m) = \sum_i f_i((y_i - E_f[Y]) - m(x_i - E_f[X]))^2 \quad (7)$$

$$\partial E/\partial b = -2\sum_i f_i(x_i - E_f[X])((y_i - E_f[Y]) - m(x_i - E_f[X])) = 0 \quad (8)$$

The last line can again be read as $Cov_f[X,Y] - mV_f[X] = 0$ or $m = Cov_f[X,Y]/V_f[X]$ (where f-weighted variance and co-variance have been defined as seen). The result essentially shows that: (i) also with an arbitrary weight function, results of linear regression can be calculated analytically, (ii) all one has to do it to replace variance/co-variance with the respective weighted counterparts.

3 Linear Regression III

Consider the regression problem in which we have pairs of (x_i, y_i), $x_i \in R^d$ and $y_i \in R$ (see page 27/28 of the lecture slides). Show that the (augmented) parameter vector $\tilde{w}$ can be obtained from the data matrix $\tilde{X}$ and the vector of y-values y via $\tilde{w} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T y$.

solution:
We are using the notation from the lecture slides, i.e. the augmented data matrix is $\tilde{X}$, the augmented parameter vector $\tilde{w}$, and the regression problem translates into minimizing

$$
\begin{aligned}
E(\tilde{w}) &= ||y - \tilde{X}\tilde{w}||^2 && (9)\\
&= (y - \tilde{X}\tilde{w})^T(y - \tilde{X}\tilde{w}) && (10)\\
&= y^T y - \tilde{w}^T\tilde{X}^T y - y^T\tilde{X}\tilde{w} + \tilde{w}^T\tilde{X}^T\tilde{X}\tilde{w} && (11)\\
&= y^T y - 2\tilde{w}^T\tilde{X}^T y + \tilde{w}^T\tilde{X}^T\tilde{X}\tilde{w}, && (12)
\end{aligned}
$$

where we have used rules for transposition and the symmetry of the scalar product (in the last line). Proceed to calculate gradients with regard to $\tilde{w}$ and equate them to zero:

$$\partial E/\partial\tilde{w} = -2\tilde{X}^T y + 2\tilde{X}^T\tilde{X}\tilde{w} = 0. \tag{13}$$

It follows $\tilde{w} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T y$, i.e. the above result.

4 Maximum Likelihood Estimation I

Repeated coin tossing of an (unfair) coin produces 100 heads up and 120 tails up. Find a maximum likelihood estimate for the probability that a coin toss will result in heads up.
solution:
Say we have $N_u = 100$ coin tosses with heads up and $N_d = 120$ coin tosses with heads down. Each coin toss will produce (say) heads up with probability p; this is the probability we want to estimate using MLE. Let us define a random variable X_i, such that $X_i = 1$ if we find heads up in coin toss number i and $X_i = 0$ otherwise. Then, the probability of observing X_i at coin toss i is

$$Pr\{X_i; p\} = p^{X_i}(1-p)^{1-X_i}. \tag{14}$$

The likelihood function of having observed a sequence $X_1, ..., X_N$ then is

$$L(X_1, ..., X_N; p) = \prod_i p^{X_i}(1-p)^{1-X_i} \quad (15)$$

$$= p^{N_u}(1-p)^{N_d}. \quad (16)$$

Thus,

$$\ln L(N_u, N_d; p) = N_u \ln p + N_d \ln(1-p). \quad (17)$$

We can now proceed as usual by calculating the derivative with regard to p and setting it to zero. We obtain:

$$\partial \ln L/\partial p = N_u/p - N_d/(1-p) \quad (18)$$

$$= 0. \quad (19)$$

We see that $N_u - pN_u = pN_d$ or $N_u = pN$ resulting in $p = N_u/N$ (which is probably what you would have guessed anyway).

5 Maximum Likelihood Estimation II

A number of observations x_i, $i = 1, ..., N$ are known to have been sampled from an exponential distribution $P(x) \sim \exp(-\lambda x)$. Find a maximum likelihood estimate for λ.
solution:
First, we should find a normalization for the exponential distribution. We require that $\int_0^\infty p(x)dx = 1$ or $\int_0^\infty \exp(-\lambda x)dx = -1/\lambda[\exp(-\lambda x)]_0^\infty = 1/\lambda = 1$. If follows that $P(x) = \lambda \exp(-\lambda x)$. All observations have been drawn from this distribution, thus:

$$L(x_1, ..., x_N; \lambda) = \prod_i \lambda \exp(-\lambda x_i) \quad (20)$$

$$= \lambda^N \exp(-\lambda \sum_i x_i) \quad (21)$$

$$\ln L(x_1, ..., x_N; \lambda) = N \ln \lambda - \lambda \sum_i x_i. \quad (22)$$

We can now proceed as usual, i.e. calculate $\partial \ln L/\partial \lambda$ and equate it to zero. We obtain:

$$\partial \ln L/\partial \lambda = N/\lambda - \sum_i x_i = 0, \quad (23)$$

i.e. we have $\lambda = N/\sum_i x_i = 1/E[X]$, which gives the desired maximum likelihood estimator.