Least Squares

Dr François Pitié

Problem 1. Which of the following models with input x_1, x_2, parameters w_1, w_2 and noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, are linear in the parameters and can be used as such for Least Squares:

1. $y = w_0 + w_1 x^2 + \epsilon$
2. $y = w_0 x^{w_1} + w_2 + \epsilon$
3. $y = \exp(w_0 + w_1 x) + \epsilon$
4. $\log(y) = w_0 + w_1 x + \epsilon$

answer:

1. yes, it is linear in the parameters
2. no, it is not linear in the parameters as the function $z \mapsto x^z$ is not linear.
3. no, it is not linear in the parameters
4. This one is a bit misleading and wouldn't be put this way in an exam. This model is technically not linear as the output is expressed as $\log(y)$ and not y, and by convention I said that y is the outcome. However, it is interesting to note that model (3) can be expressed as a linear model. Indeed, not only you can transform the features in a non-linear way, but you can also transform the outcome y as it is a constant. Thus if we change the outcome from y to $y' = \log(y)$, the model $y' = w_0 + w_1 x + \epsilon'$ becomes linear if you consider that the outcome is y' instead of y. Note however that the error term ϵ' is also now different from the error term in (3) and is not necessary Gaussian anymore.

Problem 2. Which of the following models with input x_1, x_2, parameters w_1, w_2 and noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, are linear in the parameters and can be used as such for Least Squares:

1. $y = \sin(x_1 w_1 + w_2) + \epsilon$
2. $y = \log(x_1) w_1 + \log(x_2) w_2 + \epsilon$
3. $y = w_1 x_1^2 + \epsilon$
4. $y = {w_1}^2 x_1 + \epsilon$

answer:

1. not linear
2. linear
3. linear
4. not linear

Problem 3. For n real numbers $x_1, \cdots, x_n$, what is the value $\hat{x}$ that minimises the sum of squared distances from x to each x_i:

$$\hat{x} = \arg\min_x \sum_{i=1}^{n} (x_i - x)^2$$

answer:

Let's express this problem as a least squares problem. Consider that the outcome is x_i (I know, notations get a bit counter-intuitive here, I am just trying to map the problem to the LS formulation), the input feature is the constant 1 and the linear model can be written as follows:

$$x_i = w_0 \times 1 + \epsilon_i$$

The mean squared error is

$$E(w_0) = \frac{1}{n} \sum_{i=1}^{n} (w_0 - x_i)^2$$

and we can see that the LS estimate $\hat{w}_0$ is the value $\hat{x}$ that we are looking for.
The design matrix is

$$\mathsf{X} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad \text{and the outcome vector is} \quad \mathsf{y} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

The normal equations tell us

$$\hat{w} = \hat{x} = (\mathsf{X}^\top \mathsf{X})^{-1} \mathsf{X}^\top \mathsf{y}$$

$$\mathsf{X}^\top \mathsf{X} = \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = n$$

$$\mathsf{X}^\top \mathsf{y} = \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \sum_{i=1}^{n} x_i$$

$$\hat{w} = \hat{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Problem 4. For a linear model $\mathsf{y} = \mathsf{X}\mathsf{w} + \boldsymbol{\epsilon}$, derive, in a matrix form, the expression of the least square error. That is, for $E(\mathsf{w}) = \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}$ derive the expression of $\min_{\mathsf{w}} E(\mathsf{w})$.

answer:

we know the minimum is reached at the LS estimate.

$$min_{\mathsf{w}} E(\mathsf{w}) = E(\hat{\mathsf{w}}) = \left(\mathsf{y} - \mathsf{X}\hat{\mathsf{w}}\right)^\top \left(\mathsf{y} - \mathsf{X}\hat{\mathsf{w}}\right)$$

with $\hat{\mathsf{w}}$ given by:

$$\hat{\mathsf{w}} = \left(\mathsf{X}^\top \mathsf{X}\right)^{-1} \mathsf{X}^\top \mathsf{y}$$

That would be enough as an answer. Below we continue the derivation as a few simplifications occur:

$$\begin{aligned}
E(\hat{\mathsf{w}}) &= \left(\mathsf{y} - \mathsf{X}\hat{\mathsf{w}}\right)^\top \left(\mathsf{y} - \mathsf{X}\hat{\mathsf{w}}\right) \\
&= \mathsf{y}^\top \mathsf{y} + \hat{\mathsf{w}}^\top \mathsf{X}^\top \mathsf{X}\hat{\mathsf{w}} - 2\mathsf{y}^\top \mathsf{X}\hat{\mathsf{w}} \\
&= \mathsf{y}^\top \mathsf{y} + \left(\left(\mathsf{X}^\top \mathsf{X}\right)^{-1} \mathsf{X}^\top \mathsf{y}\right)^\top \mathsf{X}^\top \mathsf{X} \left(\mathsf{X}^\top \mathsf{X}\right)^{-1} \mathsf{X}^\top \mathsf{y} - 2\mathsf{y}^\top \mathsf{X} \left(\mathsf{X}^\top \mathsf{X}\right)^{-1} \mathsf{X}^\top \mathsf{y} \\
&= \mathsf{y}^\top \mathsf{y} + \mathsf{y}^\top \mathsf{X} \left(\left(\mathsf{X}^\top \mathsf{X}\right)^{-1}\right)^\top \mathsf{X}^\top \mathsf{X} \left(\mathsf{X}^\top \mathsf{X}\right)^{-1} \mathsf{X}^\top \mathsf{y} - 2\mathsf{y}^\top \mathsf{X} \left(\mathsf{X}^\top \mathsf{X}\right)^{-1} \mathsf{X}^\top \mathsf{y} \\
&= \mathsf{y}^\top \mathsf{y} + \mathsf{y}^\top \mathsf{X} \left(\mathsf{X}^\top \mathsf{X}\right)^{-1} \mathsf{X}^\top \mathsf{y} - 2\mathsf{y}^\top \mathsf{X} \left(\mathsf{X}^\top \mathsf{X}\right)^{-1} \mathsf{X}^\top \mathsf{y} \\
&= \mathsf{y}^\top \mathsf{y} - \mathsf{y}^\top \mathsf{X} \left(\mathsf{X}^\top \mathsf{X}\right)^{-1} \mathsf{X}^\top \mathsf{y}
\end{aligned}$$

here we use the fact that $\mathsf{X}^\top \mathsf{X}$ is a symmetric matrix, $\mathsf{X}^\top \mathsf{X} = \left(\mathsf{X}^\top \mathsf{X}\right)^\top$ and this is also true of its inverse: $\left(\mathsf{X}^\top \mathsf{X}\right)^{-1} = \left(\left(\mathsf{X}^\top \mathsf{X}\right)^{-1}\right)^\top$

Problem 5. An autoregressive model is when a value from a time series is regressed on previous values from that same time series.

$$x_t = w_0 + \sum_{i=1}^{p} w_i x_{t-i} + \varepsilon_t$$

write the design matrix for this problem.

answer:

Here we have to be careful that for $t < p$ the values for x may not be defined. For instance x_{-3} may not be defined. In the following, we consider that we collect n consecutive observations from the available time series history. Say we start collecting data from time t, the n observations will be $x_t, x_{t+1}, \cdots, x_{t+n-1}$. The design matrix is then:

$$\mathsf{X} = \begin{pmatrix} 1 & x_{t-1} & x_{t-2} & x_{t-3} & \cdots & x_{t-p} \\ 1 & x_{t+1-1} & x_{t+1-2} & x_{t+1-3} & \cdots & x_{t+1-p} \\ 1 & x_{t+2-1} & x_{t+2-2} & x_{t+2-3} & \cdots & x_{t+2-p} \\ 1 & x_{t+3-1} & x_{t+3-2} & x_{t+3-3} & \cdots & x_{t+3-p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{t+n-1-1} & x_{t+n-1-2} & x_{t+n-1-3} & \cdots & x_{t+n-1-p} \end{pmatrix}$$

We could try to extrapolate the values of x and for instance set that $x_{-1}, \cdots, x_{-p} = 0$. That way we wouldn't have to worry about out of range access.

Problem 6. Consider the linear model $y = w_0 + w_1 x$. We want to bias w_1 towards the value $\hat{w}_1$. Write a loss function that achieves this.

answer:

The original LS loss function is

$$E(w_0, w_1) = \sum_{i=1}^{n} (w_0 + w_1 x_i - y_i)^2$$

We can achieve the bias by for instance adding a L2 penalty on w_1 deviating from $\hat{w}_1$:

$$E'(w_0, w_1) = E(w_0, w_1) + \lambda (w_1 - \hat{w}_1)^2$$