Linear Regression

Suppose we have a dataset on \( n \) houses with sizes \( x_1, \ldots, x_n \) square feet and prices \( y_1, \ldots, y_n \) dollars. We would like to model the dependency of \( y_i \) on \( x_i \). In the model of linear regression, we assume the data comes from a probability model:

\[
Y_i = aX_i + b + Z_i, \quad Z_i \sim N(0, \sigma^2),
\]

where we assume the \( X_i \)'s and \( Z_i \)'s are mutually independent random variables.

The parameter we would like to estimate is \( \theta = (a, b) \). This is equivalent to finding the slope and the y-intercept of a line that fits the data points. After we estimate the parameters by the maximum likelihood estimate \( \hat{a}_{\text{ML}}, \hat{b}_{\text{ML}} \), the predictor for the data point \( x \) would be \( \hat{y} = \hat{a}_{\text{ML}}x + \hat{b}_{\text{ML}} \). Note that \( \sigma^2 \) is also an unknown in the model that we can estimate. However, since \( \sigma^2 \) does not affect the predictor \( \hat{y} = \hat{a}_{\text{ML}}x + \hat{b}_{\text{ML}} \), we focus on estimating \( a \) and \( b \) here.

Note that the problem we considered in the last lecture of estimating the mean \( \mu \) of \( n \) i.i.d. observations \( Y_1, Y_2, \ldots, Y_n \) is a special case of the linear regression problem, with \( a = 0 \) and \( b = \mu \). However, in typical applications of statistics and machine learning, we observe features (house size in our example) and want to use them to predict a target variable of interest (house price in our example). So the more general linear regression model with \( a \neq 0 \) is of much more interest.

To find the maximum likelihood estimate of the parameters, we find \( (a, b) \) which maximizes the joint density \( f(x_1, y_1, \ldots, x_n, y_n; a, b) \). Since we assume each data point is independent,

\[
f(x_1, y_1, \ldots, x_n, y_n; a, b) = f(x_1, y_1; a, b) \cdots f(x_n, y_n; a, b).
\]

Consider the individual terms \( f(x_i, y_i; a, b) \). Since the distribution of \( X_i \) does not depend on \( a, b \),

\[
f(x_i, y_i; a, b) = f(x_i) f(y_i|x_i; a, b)
\]

\[
= f(x_i) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - ax_i - b)^2}{2\sigma^2}\right).
\]

Hence finding the MLE is equivalent to minimizing over all \( a \) and \( b \):

\[
\sum_{i=1}^{n} (y_i - ax_i - b)^2.
\]

This is called the least squares problem. We are fitting a line through the data points that minimizes the sum of the squared errors. The minimizing \( (a, b) \) is given as

\[
\begin{bmatrix}
  a^* \\
  b^*
\end{bmatrix} = (A^T A)^{-1} A^T \begin{bmatrix}
  y_1 \\
  \vdots \\
  y_n
\end{bmatrix}, \quad \text{where } A = \begin{bmatrix}
  x_1 & x_2 & \cdots & x_n \\
  1 & 1 & \cdots & 1
\end{bmatrix}^T.
\]
Note that the probabilistic model with Gaussian additive noise justifies the squared error as a criterion of measuring the goodness of fit. If the model is different, then the criterion would be different. However, maximum likelihood parameter estimation is a general principle that can be applied to any probabilistic model of how the data is generated and allows us to derive the appropriate data fitting procedure. We will see an example of that in the next lecture.