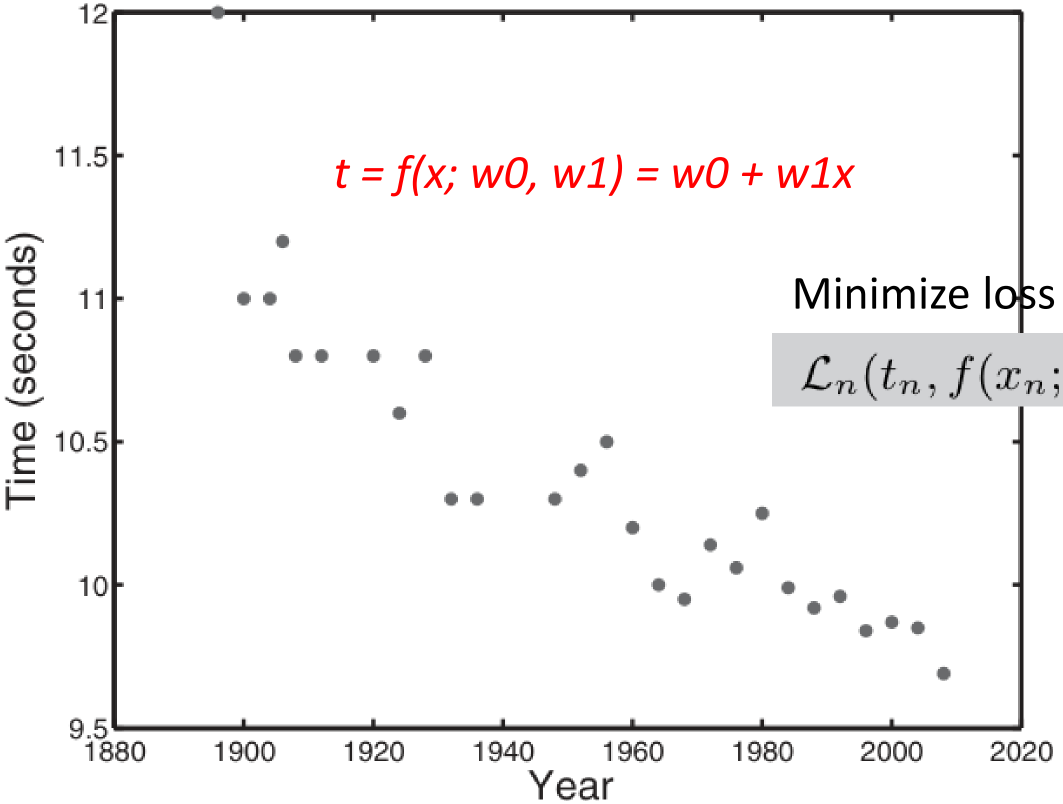
Maximum Likelihood Approach

Prof. Asim Tewari
IIT Bombay

LINEAR MODELLING



t = f(x; w0, w1) = w0 + w1x

Minimize loss function

$$\mathcal{L}_n(t_n, f(x_n; w_0, w_1)) = (t_n - f(x_n; w_0, w_1))^2$$

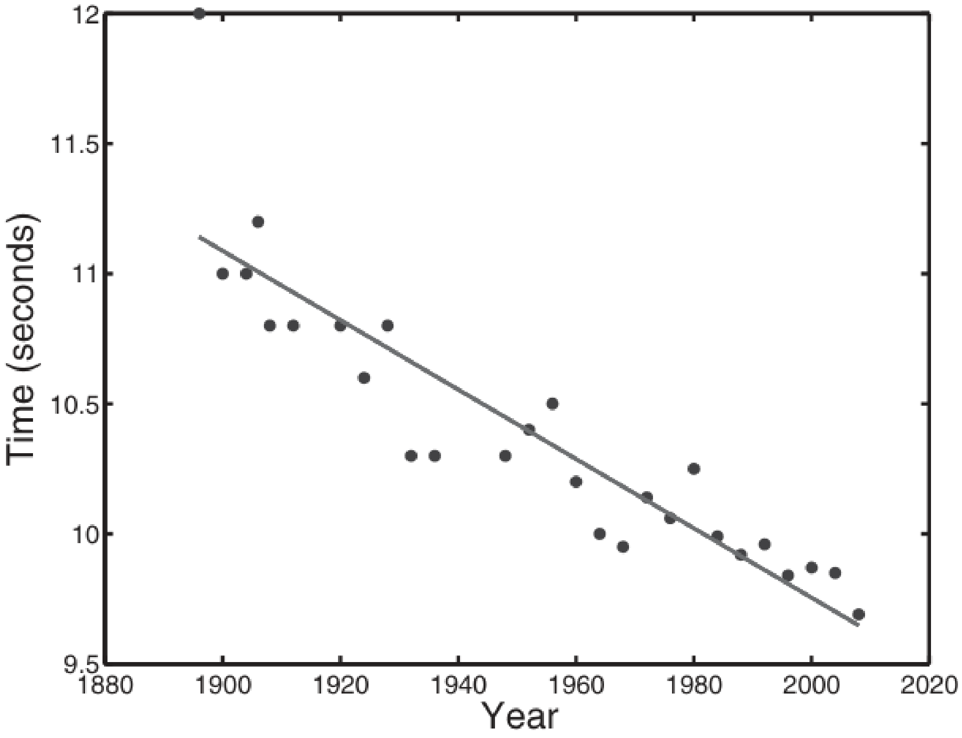
Winning men's 100 m times at the Summer Olympics since 1896

LINEAR MODELLING

t = f(x; w0, w1) = w0 + w1x

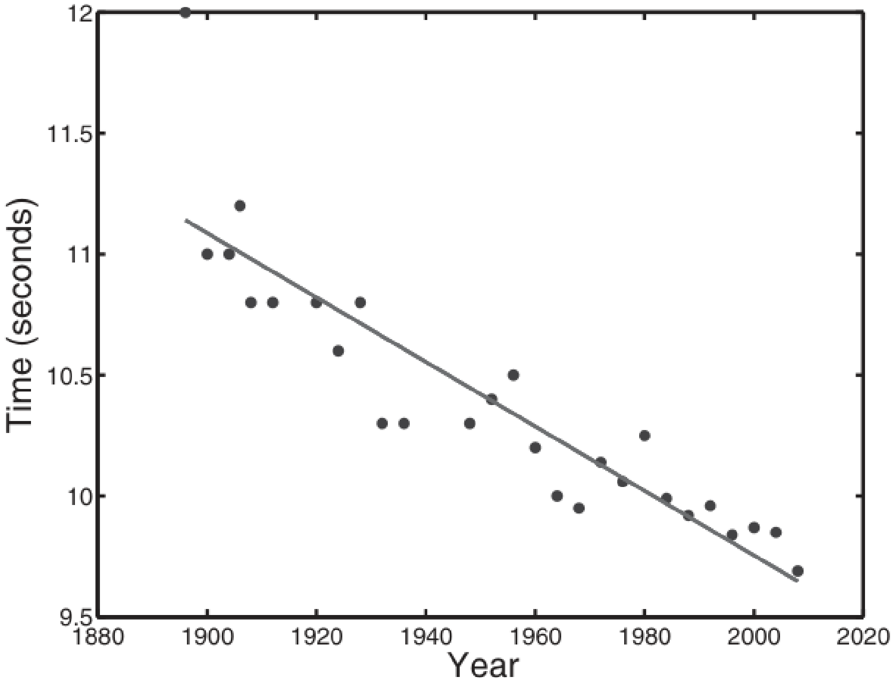
Minimize loss function

$$\mathcal{L}_n(t_n, f(x_n; w_0, w_1)) = (t_n - f(x_n; w_0, w_1))^2$$

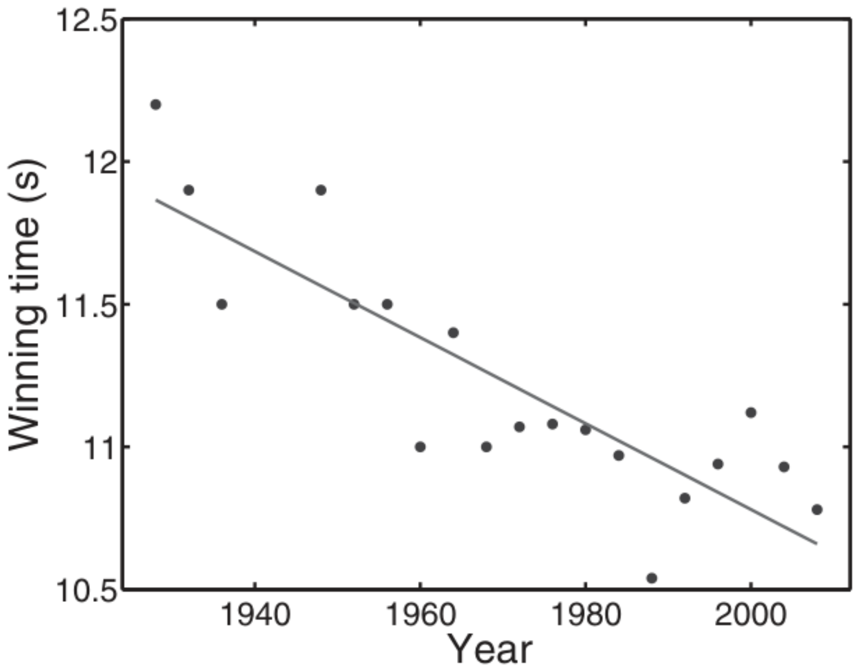


$$f(x; w_0, w_1) = 36.416 - 0.013x$$

LINEAR MODELLING

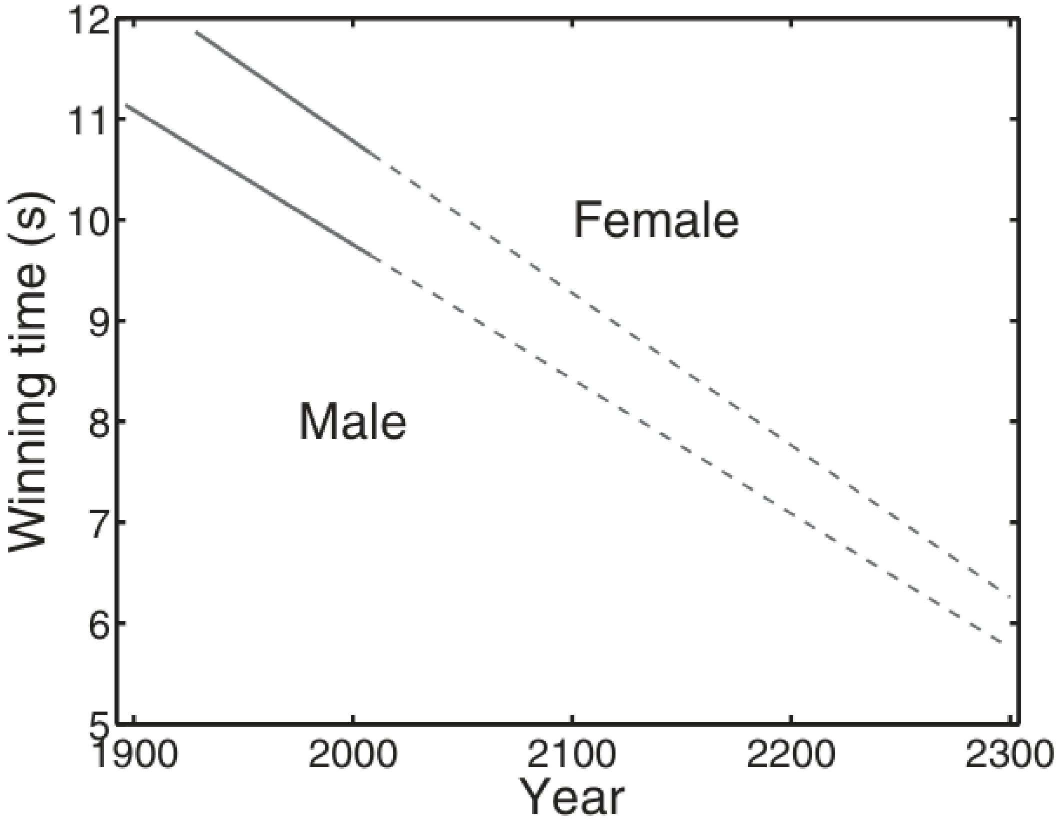


Men’s Olympic 100 m data with a linear model



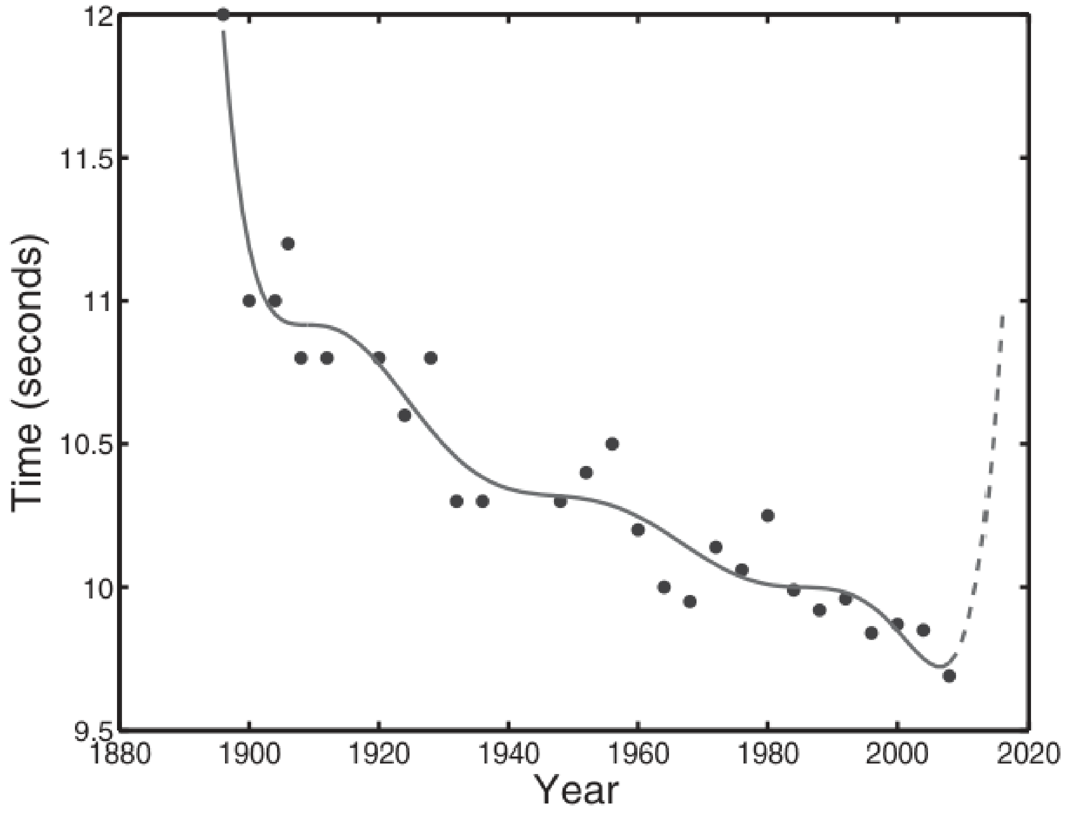
Women’s Olympic 100 m data with a linear model

LINEAR MODELLING



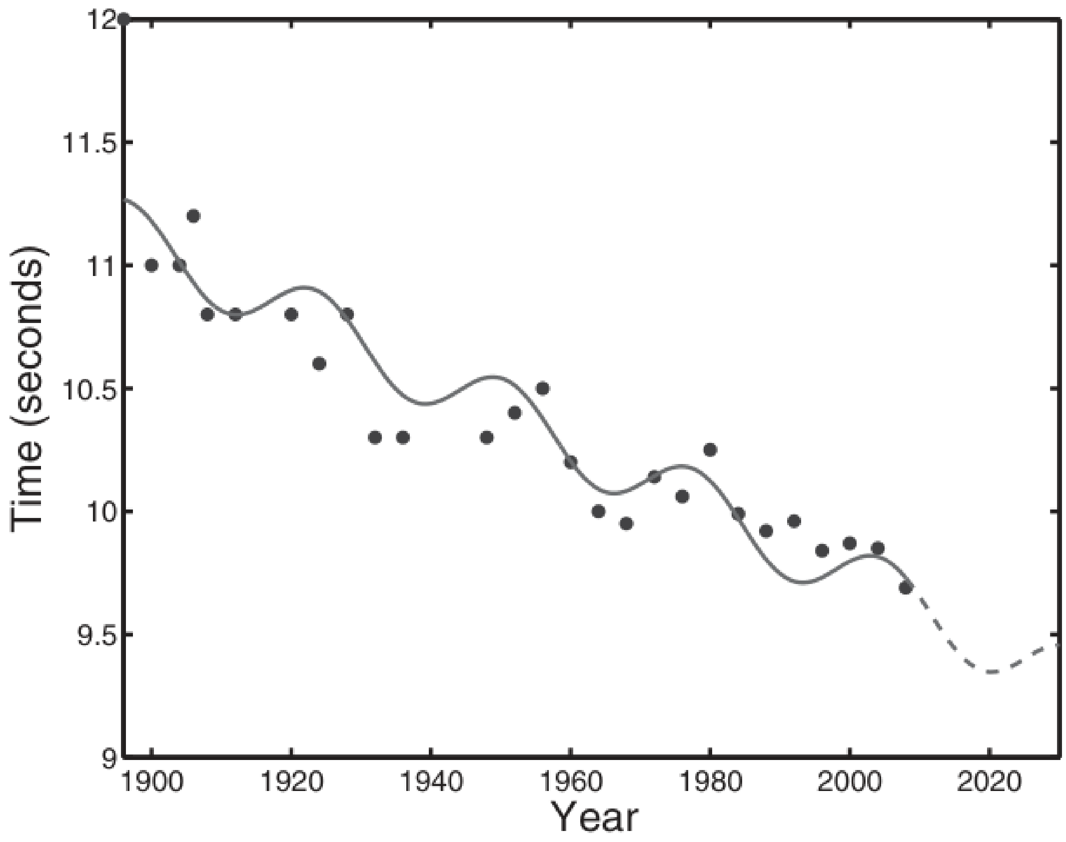
Male and female functions extrapolated into the future

NON-LINEAR MODELLING



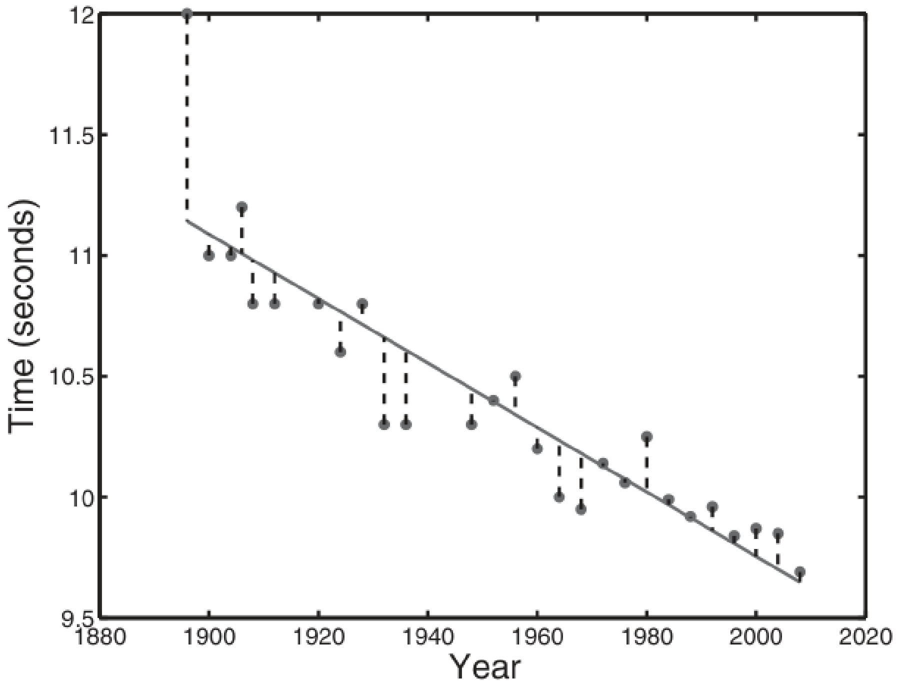
Eighth-order polynomial fitted to the Olympic 100 m men's sprint data

NON-LINEAR MODELLING

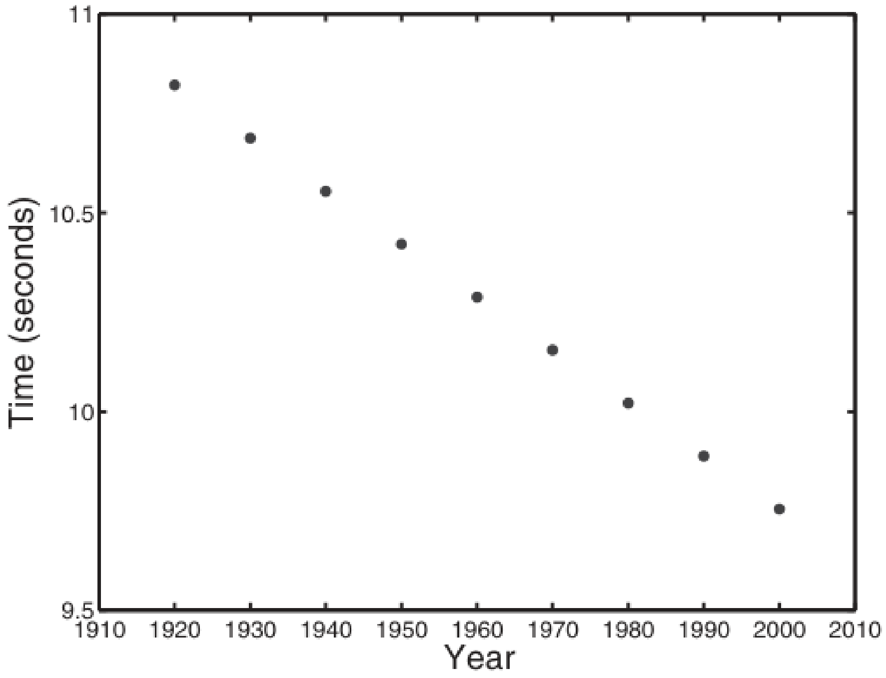


Least squares fit of $f(x; w) = w0 + w1x + w2 \sin((x-a)/b)$
to the 100 m sprint data ($a = 2660$, $b = 4.3$)

MODELING ERRORS AS NOISE



Linear fit to the Olympic men's 100 m data with errors highlighted



Dataset generated from the linear model

LIKELIHOOD

$$t_n = f(\mathbf{x}_n; \mathbf{w}) + \epsilon_n, \;\; \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$

Therefore, the random variable t_n has the density function

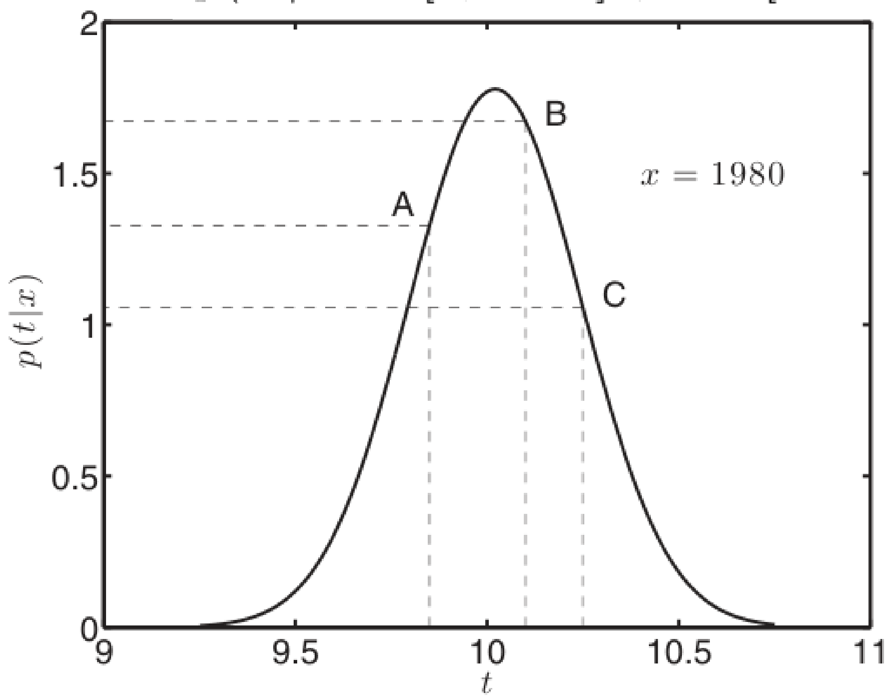
$$p(t_n | \mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$$

To see how we can use this to find optimal values of $\mathbf{w}$ and σ^2, consider one of the years from our dataset – 1980. Based on the model (w_0, w_1) found in the previous chapter and assuming again that $\sigma^2 = 0.05$, we can plot $p(t_n | x_n = 1980, \mathbf{w}, \sigma^2)$ as a function of t_n, shown

LIKELIHOOD

To see how we can use this to find optimal values of $\mathbf{w}$ and σ^2, consider one of the years from our dataset – 1980. Based on the model (w_0, w_1) found in the previous chapter and assuming again that $\sigma^2 = 0.05$, we can plot $p(t_n|x_n = 1980, \mathbf{w}, \sigma^2)$ as a function of t_n, shown

$$p(t_n|\mathbf{x}_n = [1,\ 1980]^\mathsf{T}, \mathbf{w} = [36.416, -0.0133]^\mathsf{T}, \sigma^2 = 0.05)$$



Likelihood function for the year 1980

For this the Gaussian density has a mean μ = 36.416 – 0.0133 × 1980 = 10.02 sec.

The actual winning time in the 1980 Olympics is C (10.25 seconds)

LIKELIHOOD

Dataset likelihood

In general, we are not interested in the likelihood of a single data point but that of all of the data. If we have N data points, we are interested in the joint conditional density:

$$p(t_1, \ldots, t_N | \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{w}, \sigma^2).$$

The assumption that the noise at each data point is independent ($p(\epsilon_1, \ldots, \epsilon_N) = \prod_n p(\epsilon_n)$) enables us to factorise this density into something more manageable. In particular, this joint conditional density can be factorised into N separate terms, one for each data object:

$$L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2).$$

LIKELIHOOD

Maximum likelihood

$$L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\top \mathbf{x}_n, \sigma^2).$$

$$\begin{aligned}
\log L &= \sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(t_n - f(\mathbf{x}_n; \mathbf{w}))^2\right\}\right) \\
&= \sum_{n=1}^{N} \left(-\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}(t_n - f(\mathbf{x}_n, \mathbf{w}))^2\right) \\
&= -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - f(\mathbf{x}_n; \mathbf{w}))^2.
\end{aligned}$$

LIKELIHOOD

Substituting our particular deterministic component $f(\mathbf{x}_n; \mathbf{w}) = \mathbf{w}^{\mathsf{T}}\mathbf{x}_n$ gives us the log-likelihood expression that we will work with:

$$\log L = -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^{\mathsf{T}}\mathbf{x}_n)^2.$$

$$\begin{aligned}\frac{\partial \log L}{\partial \mathbf{w}} &= \frac{1}{\sigma^2}\sum_{n=1}^{N}\mathbf{x}_n(t_n - \mathbf{x}_n^{\mathsf{T}}\mathbf{w}) \\ &= \frac{1}{\sigma^2}\sum_{n=1}^{N}\mathbf{x}_n t_n - \mathbf{x}_n\mathbf{x}_n^{\mathsf{T}}\mathbf{w} = \mathbf{0}.\end{aligned}$$

LIKELIHOOD

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}, \quad \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}$$

In this notation, $\sum_{n=1}^{N} \mathbf{x}_n t_n$ can be written as $\mathbf{X}^\top \mathbf{t}$ and similarly $\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w}$ as $\mathbf{X}^\top \mathbf{X} \mathbf{w}$. This allows us to write the derivative in the more convenient vector/matrix form:

$$\frac{\partial \log L}{\partial \mathbf{w}} = \frac{1}{\sigma^2}(\mathbf{X}^\top \mathbf{t} - \mathbf{X}^\top \mathbf{X}\mathbf{w}) = \mathbf{0}.$$

Solving this expression for $\mathbf{w}$ will lead to an expression for the optimal value:

$$\begin{aligned} \frac{1}{\sigma^2}(\mathbf{X}^\top \mathbf{t} - \mathbf{X}^\top \mathbf{X}\mathbf{w}) &= 0 \\ \mathbf{X}^\top \mathbf{t} - \mathbf{X}^\top \mathbf{X}\mathbf{w} &= 0 \\ \mathbf{X}^\top \mathbf{X}\mathbf{w} &= \mathbf{X}^\top \mathbf{t} \\ \mathbf{w} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{t}. \end{aligned}$$

$$\widehat{\sigma^2} = \frac{1}{N} \sum_{n=1}^{N} (t_n - \mathbf{x}^\top \widehat{\mathbf{w}})^2.$$