Probabilistic Approach to Linear Regression

Morteza H. Chehreghani
morteza.chehreghani@chalmers.se

Department of Computer Science and Engineering
Chalmers University of Technology

January 23, 2024

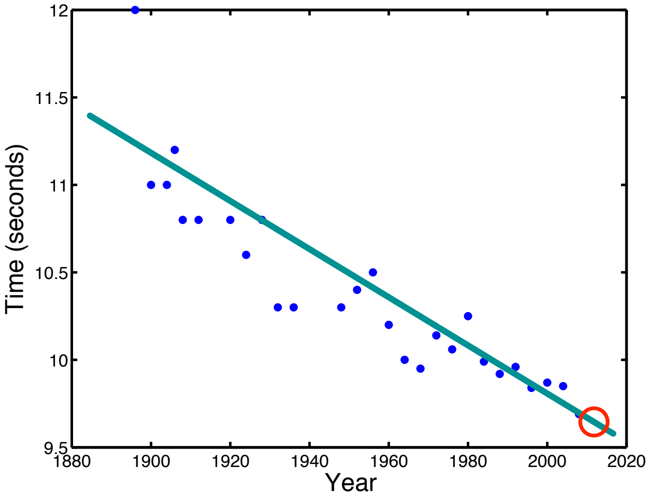
Reference

The content and the slides are adapted from

S. Rogers and M. Girolami, A First Course in Machine Learning (FCML), 2nd edition, Chapman & Hall/CRC 2016, ISBN: 9781498738484

Some data and a problem

Use the model (line) to *predict* the winning time in 2012.



Recipe for a linear model

$$t_n = w_0 + w_1 x_{n,1} + w_2 x_{n,2} + w_3 x_{n,3} + \ldots + w_D x_{n,D}$$

$$\mathbf{x}_n = \begin{bmatrix} 1 \\ x_{n,1} \\ x_{n,2} \\ \vdots \\ x_{n,D} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,D} \\ 1 & x_{2,1} & x_{2,2} & \ldots & x_{2,D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \ldots & x_{N,D} \end{bmatrix} \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix},$$

Recipe for a linear model

$$t_n = w_0 + w_1 x_{n,1} + w_2 x_{n,2} + w_3 x_{n,3} + \ldots + w_D x_{n,D}$$

$$\mathbf{x}_n = \begin{bmatrix} 1 \\ x_{n,1} \\ x_{n,2} \\ \vdots \\ x_{n,D} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,D} \\ 1 & x_{2,1} & x_{2,2} & \ldots & x_{2,D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \ldots & x_{N,D} \end{bmatrix} \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix},$$

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix}, \quad Model: t_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n, \quad or \quad \mathbf{t} = \underbrace{\mathbf{X}\mathbf{w}}_{w^T x_n}$$

What about the errors?

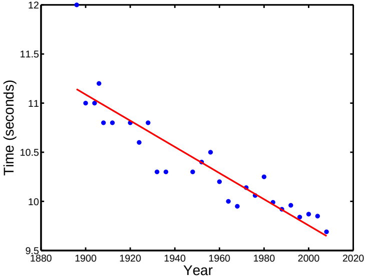
$$t_n = w_0 + w_1 x_{n,1} + w_2 x_{n,2} + w_3 x_{n,3} + \ldots + w_D x_{n,D} = \sum_{d=0}^{D} w_d x_{n,d} = \mathbf{w}^\mathsf{T}\mathbf{x}_n$$

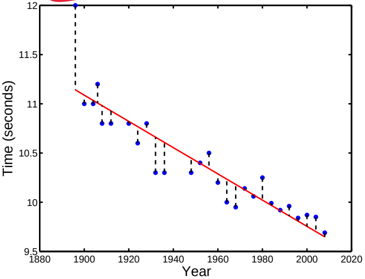
$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n\right)^2 = \frac{1}{N}(t - Xw)^T(t - Xw)$$

$$\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$$

$\left\{ \begin{matrix} w_0 \\ w_1 \end{matrix} \right.$

$$\boxed{t_{new} = \hat{w}^T x_{new}}$$



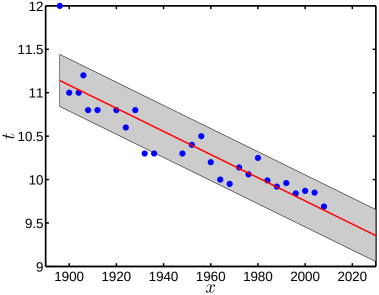


We **should** model the errors

- We know they're there - shouldn't ignore them.

We **should** model the errors

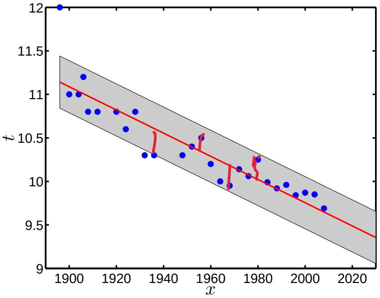
- We know they're there - shouldn't ignore them.
- They tell us how confident our predictions should be:

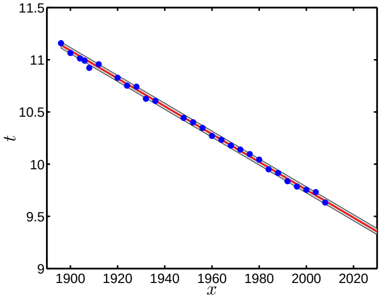


We **should** model the errors

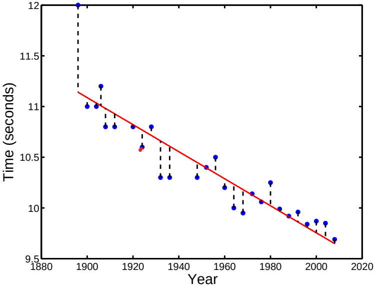
- We know they're there - shouldn't ignore them.

- They tell us how confident our predictions should be:





Additive errors

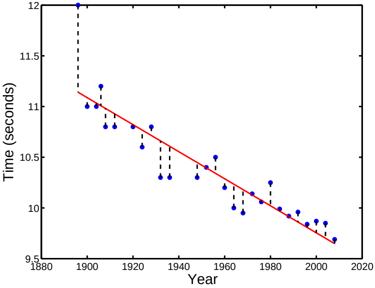


$t_n = w^T x_n$

We'll assume that the noise is an additive term in the model:

$$t_n = \underbrace{\mathbf{w}^\mathsf{T}\mathbf{x}_n}_{w_0 + w_1 x_n} + \epsilon_n$$

Additive errors

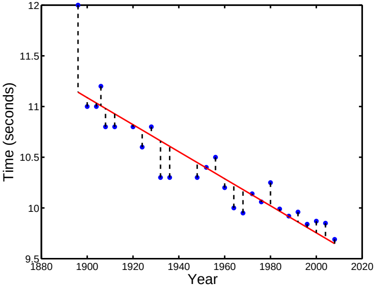


We'll assume that the noise is an additive term in the model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n + \epsilon_n$$

What assumptions can we make about ϵ_n?

Additive errors



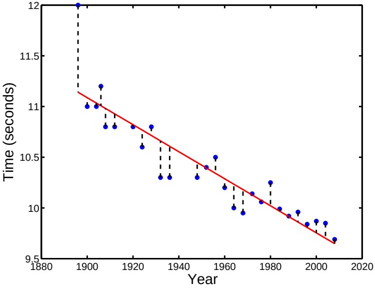
We'll assume that the noise is an additive term in the model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n + \epsilon_n$$

What assumptions can we make about ϵ_n?

- It's different for each n.

Additive errors



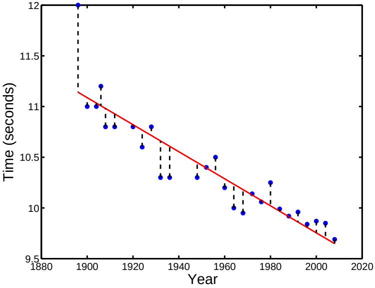
We'll assume that the noise is an additive term in the model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n + \epsilon_n$$

What assumptions can we make about ϵ_n?

- It's different for each n.
- It's positive and negative.

Additive errors



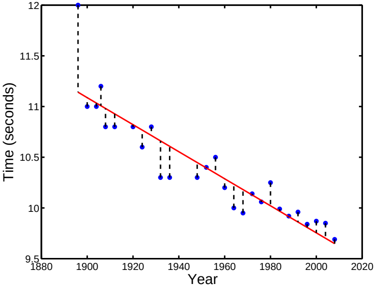
We'll assume that the noise is an additive term in the model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n + \epsilon_n$$

What assumptions can we make about ϵ_n?

- It's different for each n.
- It's positive and negative.
- There doesn't seem to be any relationship between ϵ at different n.

Additive errors



We'll assume that the noise is an additive term in the model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n + \epsilon_n$$

What assumptions can we make about ϵ_n?

- It's different for each n.
- It's positive and negative. $\longrightarrow$ $E[\varepsilon_n] = 0$
- There doesn't seem to be any relationship between ϵ at different n. $\longrightarrow$ independence of t_n's
- Looks very hard to model exactly (if it were, it wouldn't be noise!)

Gaussian noise model

- Our model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n + \epsilon_n$$

Gaussian noise model

- Our model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n + \epsilon_n$$

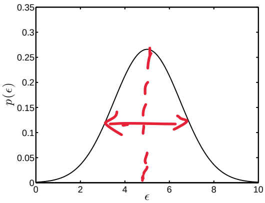
- ϵ_n is continuous.
- We need to choose $p(\epsilon)$.

Gaussian noise model

- Our model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n + \epsilon_n$$

- ϵ_n is continuous.
- We need to choose $p(\epsilon)$.
- Gaussian:



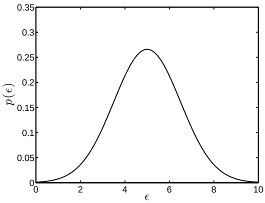
$$p(\epsilon|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(\epsilon-\mu)^2\right\}$$

Gaussian noise model

- Our model:

$$t_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n + \epsilon_n$$

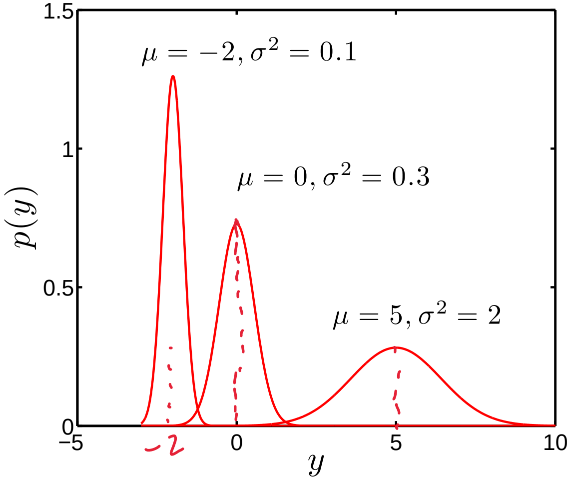
- ϵ_n is continuous.
- We need to choose $p(\epsilon)$.
- Gaussian:



$$p(\epsilon|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(\epsilon - \mu)^2\right\}$$

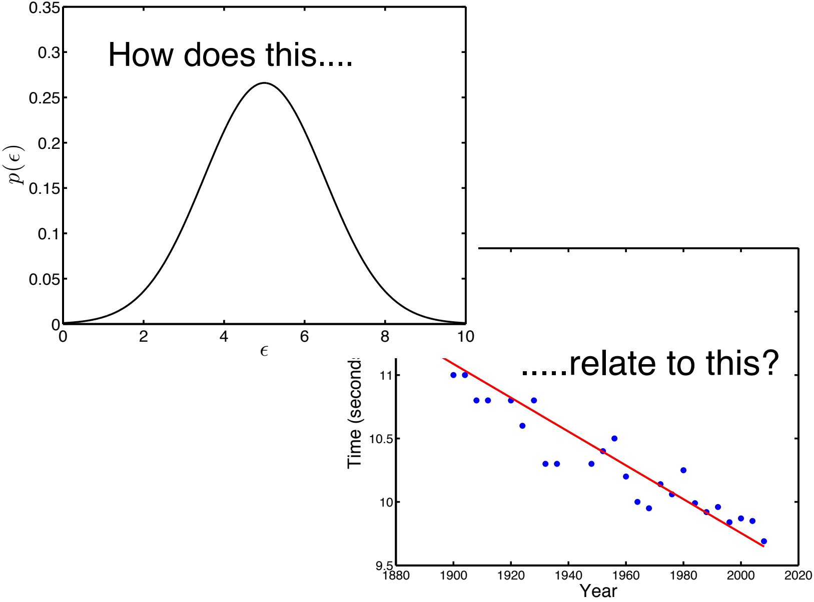
- 2 parameters: Mean μ and Variance σ^2.

Gaussian examples



Effect of varying the mean (μ) and variance (σ^2) parameters of the Gaussian.

Generating data



Likelihood

$t = w^T x + \varepsilon$

$\mu = w^T x, \sigma^2$

$\mu = 0, \sigma^2$

- t is a random variable too!

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

Likelihood

- t is a random variable too!

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

- Evaluate the density at $t = t_n$!
- At $t = t_n$ it is called the Likelihood, i.e., the quantity obtained when evaluating the density.
- The higher the value, the more likely t_n is given the model....

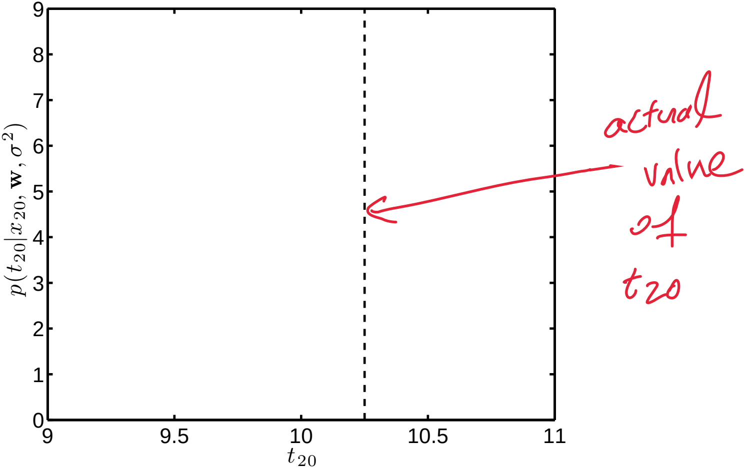
Likelihood

- t is a random variable too!

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

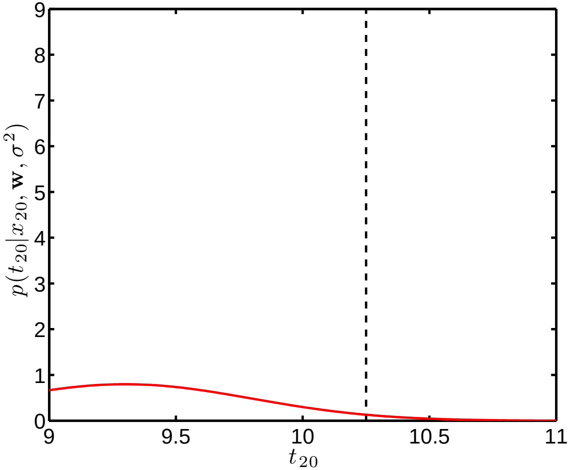
- Evaluate the density at $t = t_n$!
- At $t = t_n$ it is called the Likelihood, i.e., the quantity obtained when evaluating the density.
- The higher the value, the more likely t_n is given the model....
  - ....the better the model is.

Likelihood



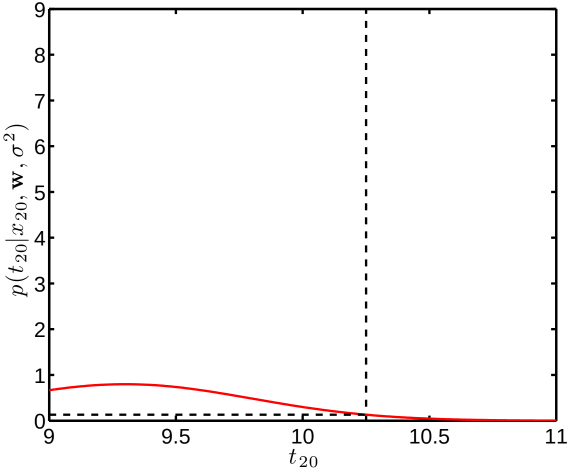
Lets look at the 1980 Olympics ($n = 20$).
Dashed line shows t_{20}.

Likelihood



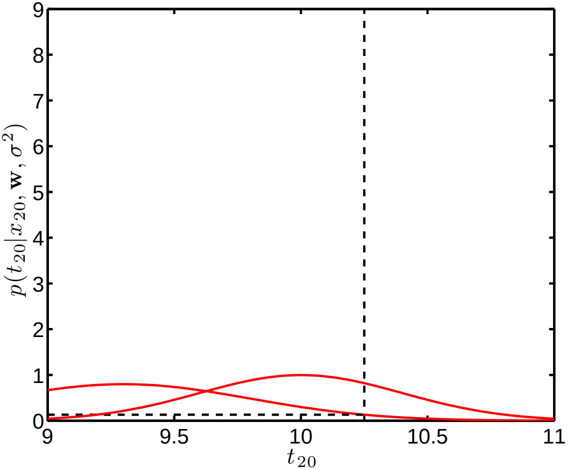
Model 1. Red line shows $\mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$

Likelihood



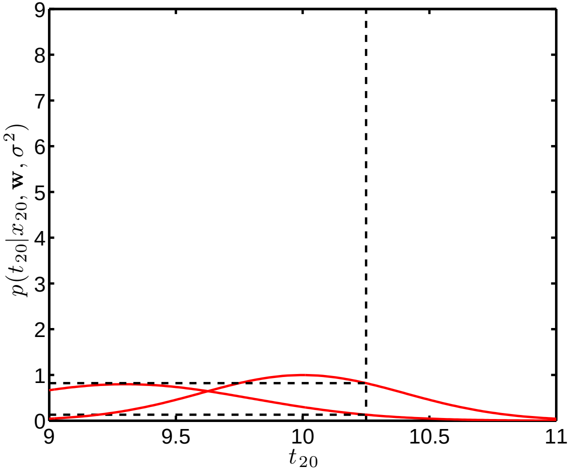
$$p(t_{20}|\ldots) \approx 0.1.$$

Likelihood



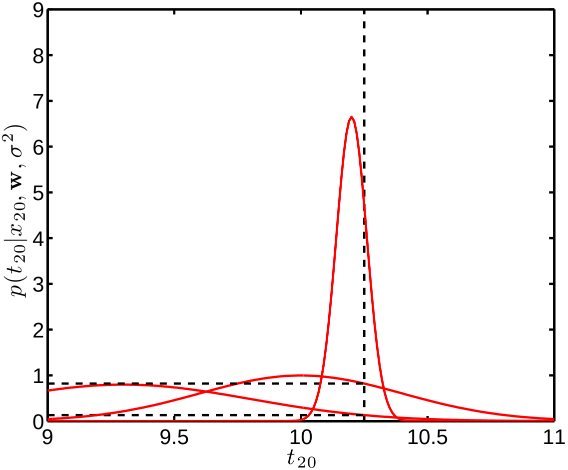
Model 2. Red line shows $\mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$ for a different $\mathbf{w}$

Likelihood



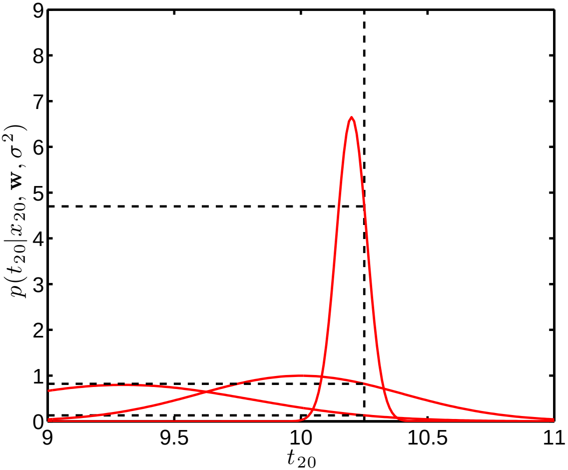
$$p(t_{20}| \ldots) \approx 0.9.$$

Likelihood



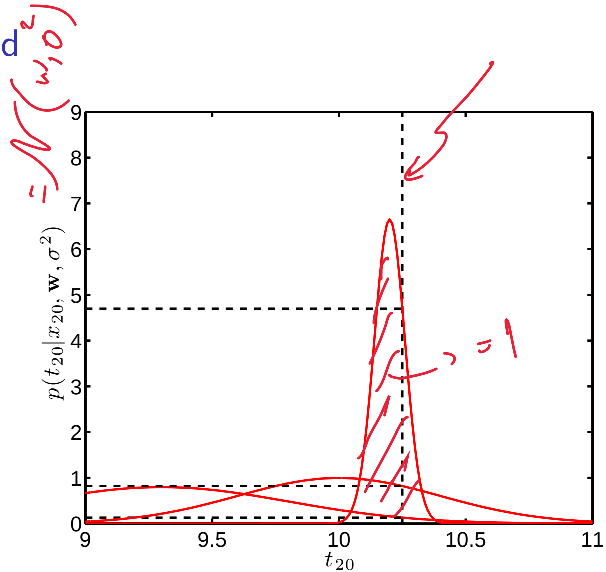
Model 3.

Likelihood



Model 3.

Likelihood



Model 3.

Model 3 looks best.

Likelihood

- The value we get when we evaluate the density function is called the **likelihood**.

Likelihood

- The value we get when we evaluate the density function is called the **likelihood**.
- i.e.
  - The likelihood for model 1 was 0.1.
  - The likelihood for model 2 was 0.9.
  - The likelihood for model 3 was 4.8.
- For continuous random variables, it is **not** a probability!

Likelihood

- The value we get when we evaluate the density function is called the **likelihood**.
- i.e.
  - The likelihood for model 1 was 0.1.
  - The likelihood for model 2 was 0.9.
  - The likelihood for model 3 was 4.8.
- For continuous random variables, it is **not** a probability!
- As t_n is fixed, we can find the values of $\mathbf{w}$ and σ^2 that maximise the likelihood.
  - ...just like we found them that minimised the loss.

Likelihood optimisation

- For each input-response pair, we have a Gaussian likelihood:

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

Likelihood optimisation

- For each input-response pair, we have a Gaussian likelihood:

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

- To combine them all, we want the joint likelihood:

entire dataset

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

Likelihood optimisation

- For each input-response pair, we have a Gaussian likelihood:

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

- To combine them all, we want the joint likelihood:

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

- Assume that the t_n's are independent:

$$p(t_1, \ldots, t_N|\mathbf{w}, \sigma^2, \mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{n=1}^{N} p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2)$$

Likelihood optimisation

Finding the parameters that maximise the likelihood is expressed mathematically as:

$$\underset{\mathbf{w},\sigma^2}{\operatorname{argmax}} \prod_{n=1}^{N} p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2)$$

Likelihood optimisation

$\cancel{\frac{\partial f_1(z) f_2(z)}{\partial z}}$

Finding the parameters that maximise the likelihood is expressed mathematically as:

$$\underset{\mathbf{w},\sigma^2}{\operatorname{argmax}} \prod_{n=1}^{N} p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2)$$

In fact, we'll optimise the (natural) log likelihood because it's easier.

- If we increase z, $\log(z)$ increases, if we decrease z, $\log(z)$ decreases. So, at a maximum of z, $\log(z)$ will also be at a maximum.

Likelihood optimisation

Finding the parameters that maximise the likelihood is expressed mathematically as:

$$\underset{\mathbf{w},\sigma^2}{\operatorname{argmax}} \prod_{n=1}^{N} p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2)$$

In fact, we'll optimise the (natural) log likelihood because it's easier.

- If we increase z, $\log(z)$ increases, if we decrease z, $\log(z)$ decreases. So, at a maximum of z, $\log(z)$ will also be at a maximum.

$$\underset{\mathbf{w},\sigma^2}{\operatorname{argmax}} \log \prod_{n=1}^{N} p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2)$$

$$= \sum_{n=1}^{N} \log p(t_n|w, x_n, \sigma^2)$$

Some re-arranging...

$$
\begin{aligned}
p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2\right\} \\
\log L &= \log \prod_{n=1}^{N} p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2)
\end{aligned}
$$

Some re-arranging...

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

Looks familiar!

Some re-arranging...

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

Looks familiar! To continue (good exercise):

$$
\frac{\partial \log L}{\partial \mathbf{w}} = 0, \quad \frac{\partial \log L}{\partial \sigma^2} = 0
$$

A shortcut

The multi-variate Gaussian

random vector $\leftarrow$ $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \; p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{K/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}$$

$K(=2)$ is number of variables, $|\boldsymbol{\Sigma}|$ is the determinant.

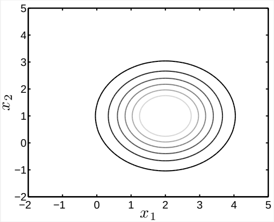
A shortcut

The multi-variate Gaussian

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \; p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{K/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \right\}$$

$K(=2)$ is number of variables, $|\boldsymbol{\Sigma}|$ is the determinant.



$$\boldsymbol{\mu} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \; \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

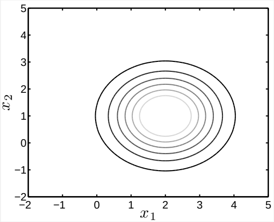
A shortcut

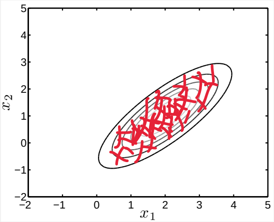
The multi-variate Gaussian

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \;\; p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{K/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right\}$$

$K(= 2)$ is number of variables, $|\boldsymbol{\Sigma}|$ is the determinant.





$$\boldsymbol{\mu} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \;\; \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

A shortcut

The multi-variate Gaussian

A special case:

$$\prod_{n=1}^{N} \mathcal{N}(\mu_n, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

independence of r.v.'s

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

$$\begin{bmatrix} 1 & & 0 \\ & 1 & \\ 0 & & 1 \end{bmatrix}$$

So, in our model:

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

Maximising the multi-variate log-likelihood

- Partial derivative w.r.t. $\mathbf{w}$, set to zero and solve:

$$\begin{aligned}
\log L &= \log \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}) \qquad \mathbf{w}^{\mathsf{T}}\mathbf{x}_n \\
\frac{\partial \log L}{\partial \mathbf{w}} &= \cancel{-\frac{1}{2\sigma^2}}(\cancel{2}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w} - \cancel{2}\mathbf{X}^{\mathsf{T}}\mathbf{t}) = 0 \\
\mathbf{w} &= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t}
\end{aligned}$$

Maximising the multi-variate log-likelihood

- Partial derivative w.r.t. $\mathbf{w}$, set to zero and solve:

$$
\begin{aligned}
\log L &= \log \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}) \\
\frac{\partial \log L}{\partial \mathbf{w}} &= -\frac{1}{2\sigma^2}(2\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} - 2\mathbf{X}^\mathsf{T}\mathbf{t}) = 0 \\
\mathbf{w} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}
\end{aligned}
$$

- This is the same expression we've seen before!

Maximising the multi-variate log-likelihood

- Partial derivative w.r.t. $\mathbf{w}$, set to zero and solve:

$$
\begin{aligned}
\log L &= \log \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}) \\
\frac{\partial \log L}{\partial \mathbf{w}} &= -\frac{1}{2\sigma^2}(2\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} - 2\mathbf{X}^\mathsf{T}\mathbf{t}) = 0 \\
\mathbf{w} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}
\end{aligned}
$$

- This is the same expression we've seen before!
- Same for σ^2:

$$
\begin{aligned}
\frac{\partial \log L}{\partial \sigma^2} &= -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\mathbf{t} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{t} - \mathbf{X}\mathbf{w}) = 0 \\
\sigma^2 &= \frac{1}{N}(\mathbf{t} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{t} - \mathbf{X}\mathbf{w})
\end{aligned}
$$

Optimum parameters

- Compute optimum $\widehat{\mathbf{w}}$ from:

$$\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$$

- Use this to compute optimum $\widehat{\sigma^2}$ from:

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})$$

Optimum parameters

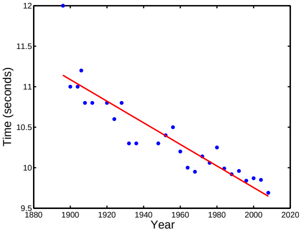
- Compute optimum $\widehat{\mathbf{w}}$ from:

$$\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$$

- Use this to compute optimum $\widehat{\sigma^2}$ from:

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})$$

- e.g. Olympic 100 m data (again!)



$$\widehat{\mathbf{w}} = \begin{bmatrix} 36.416 \\ -0.0133 \end{bmatrix}, \quad \widehat{\sigma^2} = 0.0503$$

Confidence in parameter estimates

- Imagine there are **true** parameters, $\mathbf{w}$ and σ^2.

Confidence in parameter estimates

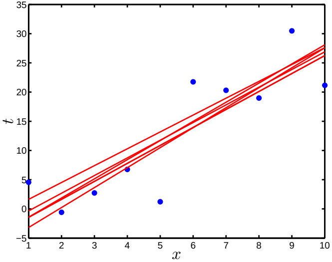
- Imagine there are **true** parameters, $\mathbf{w}$ and σ^2.
- How good are our estimates $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$?
  - Are they correct (on average)?
  - If we could keep adding data, would we converge on the true value?

Confidence in parameter estimates

- Imagine there are **true** parameters, $\mathbf{w}$ and σ^2.
- How good are our estimates $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$?
  - Are they correct (on average)?
  - If we could keep adding data, would we converge on the true value?
- How confident should we be in our estimates?
  - Could we change parameters a little bit and still have a good model?

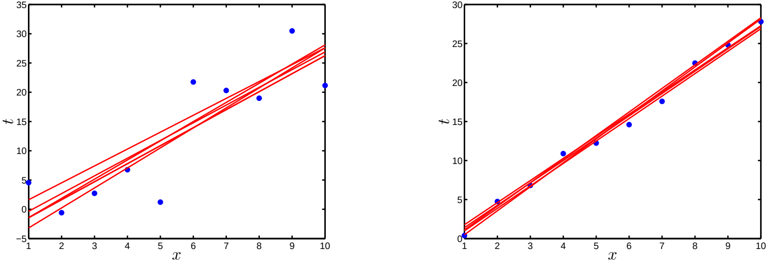
Confidence in parameter estimates

- Imagine there are **true** parameters, $\mathbf{w}$ and σ^2.
- How good are our estimates $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$?
  - Are they correct (on average)?
  - If we could keep adding data, would we converge on the true value?
- How confident should we be in our estimates?
  - Could we change parameters a little bit and still have a good model?



Confidence in parameter estimates

- Imagine there are **true** parameters, $\mathbf{w}$ and σ^2.
- How good are our estimates $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$?
  - Are they correct (on average)?
  - If we could keep adding data, would we converge on the true value?
- How confident should we be in our estimates?
  - Could we change parameters a little bit and still have a good model?

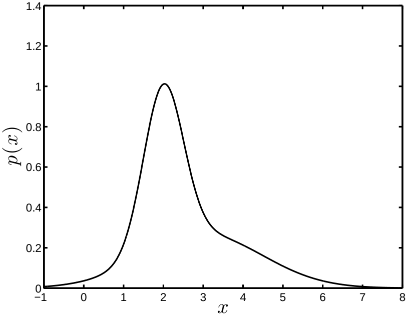


Expectations – refresher

- To progress we need to understand **Expectations**

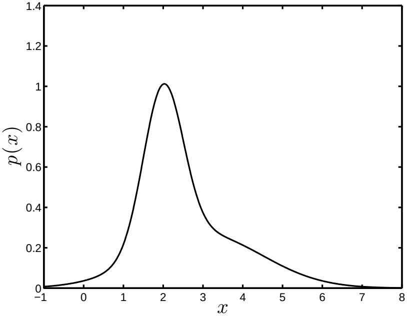
Expectations – refresher

- To progress we need to understand **Expectations**
- Imagine a random variable X with density $p(x)$



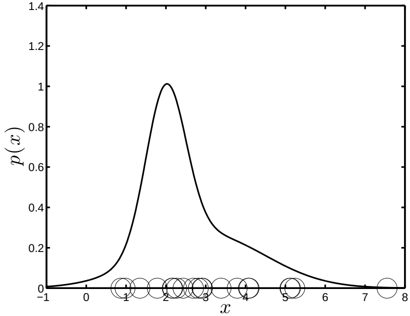
Expectations – refresher

- To progress we need to understand **Expectations**
- Imagine a random variable X with density $p(x)$
- We want to work out the average value of X, $\tilde{x}$.



Expectations – refresher

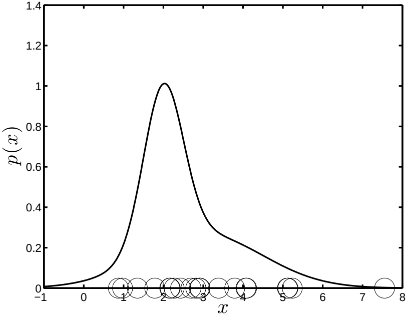
- To progress we need to understand **Expectations**
- Imagine a random variable X with density $p(x)$
- We want to work out the average value of X, $\tilde{x}$.
- Generate S samples, $x_1, \ldots, x_S$



Expectations – refresher

- To progress we need to understand **Expectations**
- Imagine a random variable X with density $p(x)$
- We want to work out the average value of X, $\tilde{x}$.
- Generate S samples, $x_1, \ldots, x_S$
- Average the samples:

$$\tilde{x} \approx \frac{1}{S} \sum_{s=1}^{S} x_s$$



Expectations – refresher

- Our sample based approximation to $\tilde{x}$ will get better as we take more samples.

Expectations – refresher

- Our sample based approximation to $\tilde{x}$ will get better as we take more samples.
- We can also (sometimes) compute it exactly using **expectations**.

Expectations – refresher

- Our sample based approximation to $\tilde{x}$ will get better as we take more samples.
- We can also (sometimes) compute it exactly using **expectations**.
  - Discrete: $\tilde{x} = \mathbf{E}_{p(x)}\{x\} = \sum_x xp(x)$
- Example:
  - X is outcome of rolling die. $P(X = x) = 1/6$
  - $\tilde{x} = \sum_x xP(X = x) = 3.5$

1, 2, 3, 4, 5, 6

Expectations – refresher

- Our sample based approximation to $\tilde{x}$ will get better as we take more samples.
- We can also (sometimes) compute it exactly using **expectations**.
  - Discrete: $\tilde{x} = \mathbf{E}_{p(x)}\left\{x\right\} = \sum_x xp(x)$
  - Continuous: $\tilde{x} = \mathbf{E}_{p(x)}\left\{x\right\} = \int_x xp(x)\ dx$
- Example:
  - X is outcome of rolling die. $P(X = x) = 1/6$
  - $\tilde{x} = \sum_x xP(X = x) = 3.5$
- Example:
  - X is uniform distributed RV between a and b $\rightarrow p(x) = \frac{1}{b-a}$
  - $\tilde{x} = \int_{x=a}^{x=b} xp(x)\ dx = \underline{(b+a)/2}$

Expectations – refresher

- In general:

$$\mathbf{E}_{p(x)}\left\{f(x)\right\} = \int f(x)p(x)\ dx$$

Expectations – refresher

- In general:

$$\mathbf{E}_{p(x)}\left\{f(x)\right\} = \int f(x)p(x)\ dx$$

- Some important things:
  - $\mathbf{E}_{p(x)}\left\{f(x)\right\} \neq f\left(\mathbf{E}_{p(x)}\left\{x\right\}\right)$
  - $\mathbf{E}_{p(x)}\left\{kf(x)\right\} = k\mathbf{E}_{p(x)}\left\{f(x)\right\}$

Expectations – refresher

- In general:

$$\mathbf{E}_{p(x)}\left\{f(x)\right\} = \int f(x)p(x)\ dx$$

- Some important things:
  - $\mathbf{E}_{p(x)}\left\{f(x)\right\} \neq f\left(\mathbf{E}_{p(x)}\left\{x\right\}\right)$
  - $\mathbf{E}_{p(x)}\left\{kf(x)\right\} = k\mathbf{E}_{p(x)}\left\{f(x)\right\}$
  - Mean, $\mu = \mathbf{E}_{p(x)}\left\{x\right\}$

Expectations – refresher

- In general:

$$\mathbf{E}_{p(x)}\left\{f(x)\right\} = \int f(x)p(x)\;dx$$

- Some important things:
  - $\mathbf{E}_{p(x)}\left\{f(x)\right\} \neq f\left(\mathbf{E}_{p(x)}\left\{x\right\}\right)$
  - $\mathbf{E}_{p(x)}\left\{kf(x)\right\} = k\mathbf{E}_{p(x)}\left\{f(x)\right\}$
  - Mean, $\mu = \mathbf{E}_{p(x)}\left\{x\right\}$
  - Variance: $\sigma^2 = \mathbf{E}_{p(x)}\left\{(x-\mu)^2\right\} = \underbrace{\mathbf{E}_{p(x)}\left\{x^2\right\}}_{\neq \mu^2} - \underbrace{\left(\mathbf{E}_{p(x)}\left\{x\right\}\right)^2}_{\mu^2}$

Expectations – refresher

- In general:
$$\mathbf{E}_{p(x)}\left\{f(x)\right\} = \int f(x)p(x)\ dx$$
- Some important things:
  - $\mathbf{E}_{p(x)}\left\{f(x)\right\} \neq f\left(\mathbf{E}_{p(x)}\left\{x\right\}\right)$
  - $\mathbf{E}_{p(x)}\left\{kf(x)\right\} = k\mathbf{E}_{p(x)}\left\{f(x)\right\}$
  - Mean, $\mu = \mathbf{E}_{p(x)}\left\{x\right\}$
  - Variance: $\sigma^2 = \mathbf{E}_{p(x)}\left\{(x-\mu)^2\right\} = \mathbf{E}_{p(x)}\left\{x^2\right\} - \left(\mathbf{E}_{p(x)}\left\{x\right\}\right)^2$
- For vectors of random variables:
  - $\mathbf{E}_{p(\mathbf{x})}\left\{f(\mathbf{x})\right\} = \int f(\mathbf{x})p(\mathbf{x})\ dx$

Expectations – refresher

- In general:
$$\mathbf{E}_{p(x)}\{f(x)\} = \int f(x)p(x)\ dx$$
- Some important things:
  - $\mathbf{E}_{p(x)}\{f(x)\} \neq f\left(\mathbf{E}_{p(x)}\{x\}\right)$
  - $\mathbf{E}_{p(x)}\{kf(x)\} = k\mathbf{E}_{p(x)}\{f(x)\}$
  - Mean, $\mu = \mathbf{E}_{p(x)}\{x\}$
  - Variance: $\sigma^2 = \mathbf{E}_{p(x)}\left\{(x-\mu)^2\right\} = \mathbf{E}_{p(x)}\left\{x^2\right\} - \left(\mathbf{E}_{p(x)}\{x\}\right)^2$
- For vectors of random variables:
  - $\mathbf{E}_{p(\mathbf{x})}\{f(\mathbf{x})\} = \int f(\mathbf{x})p(\mathbf{x})\ dx$
  - Mean: $\boldsymbol{\mu} = \mathbf{E}_{p(\mathbf{x})}\{\mathbf{x}\}$
  - Covariance:

$$\begin{aligned}\underline{\mathrm{cov}\{x\}} &= \mathbf{E}_{p(\mathbf{x})}\left\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\right\} \\ &= \mathbf{E}_{p(\mathbf{x})}\left\{\mathbf{x}\mathbf{x}^{\mathsf{T}}\right\} - \mathbf{E}_{p(\mathbf{x})}\{\mathbf{x}\}\,\mathbf{E}_{p(\mathbf{x})}\left\{\mathbf{x}^{\mathsf{T}}\right\}\end{aligned}$$

Expectations – Gaussians

- Uni-variate
  - $p(x|\mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2)$
  - Mean: $\mathbf{E}_{p(x)}\left\{x\right\} = \mu$
  - Variance: $\mathbf{E}_{p(x)}\left\{(x-\mu)^2\right\} = \sigma^2$

Expectations – Gaussians

- Uni-variate
  - $p(x|\mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2)$
  - Mean: $\mathbf{E}_{p(x)}\left\{x\right\} = \mu$
  - Variance: $\mathbf{E}_{p(x)}\left\{(x-\mu)^2\right\} = \sigma^2$
- Multi-variate
  - $p(\mathbf{x}|\mu, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - Mean: $\mathbf{E}_{p(\mathbf{x})}\left\{\mathbf{x}\right\} = \boldsymbol{\mu}$
  - Covariance: $\mathbf{E}_{p(\mathbf{x})}\left\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\right\} = \boldsymbol{\Sigma}$

Back to the model...

- Parameter estimates:

$$
\begin{aligned}
\widehat{\mathbf{w}} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} \\
\widehat{\sigma^2} &= \frac{1}{N}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})
\end{aligned}
$$

Back to the model...

- Parameter estimates:

$$\begin{aligned}
\widehat{\mathbf{w}} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} \\
\widehat{\sigma^2} &= \frac{1}{N}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})
\end{aligned}$$

- **True** values: $\mathbf{w}$, σ^2

Back to the model...

- Parameter estimates:

$$\begin{aligned} \widehat{\mathbf{w}} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} \\ \widehat{\sigma^2} &= \frac{1}{N}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}}) \end{aligned}$$

- **True** values: $\mathbf{w}$, σ^2
- Our model:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$$

Back to the model...

- Parameter estimates:

$$\begin{aligned}\widehat{\mathbf{w}} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} \\ \widehat{\sigma^2} &= \frac{1}{N}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})\end{aligned}$$

- **True** values: $\mathbf{w}$, σ^2
- Our model:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$$

- What's $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$?

Back to the model...

- Parameter estimates:

$$\begin{aligned} \widehat{\mathbf{w}} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} \\ \widehat{\sigma^2} &= \frac{1}{N}(\mathbf{t}-\mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t}-\mathbf{X}\widehat{\mathbf{w}}) \end{aligned}$$

- **True** values: $\mathbf{w}, \ \sigma^2$
- Our model:

$$p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w},\sigma^2\mathbf{I})$$

- What's $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$?
    - What do we expect our parameter estimate to be?

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$

We'll try and find $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$ in terms of the true value $\mathbf{w}$:

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$

We'll try and find $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$ in terms of the true value $\mathbf{w}$:

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} = \int \widehat{\mathbf{w}} p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\, d\mathbf{t}$$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$

We'll try and find $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$ in terms of the true value $\mathbf{w}$:

$\mathbb{E}_p\{t\}$

$$\begin{aligned}
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} &= \int \widehat{\mathbf{w}}\, p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\ d\mathbf{t} \\
&= \int (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}\, p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\ d\mathbf{t} \\
&= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\right\}
\end{aligned}$$

?

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$

We'll try and find $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$ in terms of the true value $\mathbf{w}$:

$$
\begin{aligned}
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} &= \int \widehat{\mathbf{w}} p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\; d\mathbf{t} \\
&= \int (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}\, p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\; d\mathbf{t} \\
&= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\right\} \\
&= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\underbrace{\mathbf{X}\mathbf{w}}
\end{aligned}
$$

$t_n = \underbrace{w^T x_n} + \varepsilon_n$

$E(E_n) =$

$E(\varepsilon_n) = 0$

Xw

$w^T x_n$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$

We'll try and find $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$ in terms of the true value $\mathbf{w}$:

$$
\begin{aligned}
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} &= \int \widehat{\mathbf{w}}\, p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\; d\mathbf{t} \\
&= \int (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}\, p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\; d\mathbf{t} \\
&= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\right\} \\
&= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} \qquad = I w \\
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} &= \mathbf{I}\mathbf{w} = \mathbf{w}
\end{aligned}
$$

$A^{-1}A = I$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$

We'll try and find $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$ in terms of the true value $\mathbf{w}$:

$$
\begin{aligned}
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} &= \int \widehat{\mathbf{w}}\, p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\ d\mathbf{t} \\
&= \int (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}\, p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\ d\mathbf{t} \\
&= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\right\} \\
&= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} \\
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} &= \mathbf{I}\mathbf{w} = \mathbf{w}
\end{aligned}
$$

bias : diff of $E\{\hat{w}\}$, w

$\widehat{\mathbf{w}}$ is unbiased

On average, we expect our estimate to equal the true value!

$\hat{w} \neq w$

$\hat{z} = \begin{cases} -1 \\ +1 \end{cases}$, $z = 0$

$\mathrm{cov}\{\underbrace{\widehat{\mathbf{w}}\}}_{}$ $\to$ vector

- What does $\mathrm{cov}\{\widehat{\mathbf{w}}\}$ tell us?

$\mathrm{cov}\{\widehat{\mathbf{w}}\}$

- What does $\mathrm{cov}\{\widehat{\mathbf{w}}\}$ tell us?
- Recall the linear model, $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$

$\mathrm{cov}\{\widehat{\mathbf{w}}\}$

- What does $\mathrm{cov}\{\widehat{\mathbf{w}}\}$ tell us?
- Recall the linear model, $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$

$$\mathrm{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

$\text{cov}\{\widehat{\mathbf{w}}\}$

- What does $\text{cov}\{\widehat{\mathbf{w}}\}$ tell us?
- Recall the linear model, $\mathbf{w} = \left[\begin{array}{c} w_0 \\ w_1 \end{array}\right]$

$$\text{cov}\{\widehat{\mathbf{w}}\} = \left[\begin{array}{cc} a & b \\ b & c \end{array}\right]$$

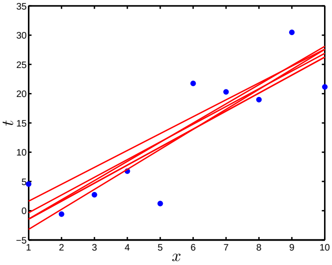
- Tells us how well defined the parameters are by the data. How much can the parameters vary and still give a **good** model.
  - a and c: how much can we change w_0 and w_1.
  - b: how the values should be changed together.

$\text{cov}\{\widehat{\mathbf{w}}\}$

- What does $\text{cov}\{\widehat{\mathbf{w}}\}$ tell us?
- Recall the linear model, $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$

$$\text{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

- Tells us how well defined the parameters are by the data. How much can the parameters vary and still give a **good** model.
  - a and c: how much can we change w_0 and w_1.
  - b: how the values should be changed together.

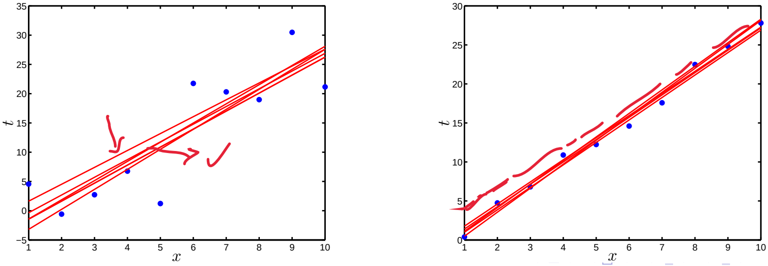


$\text{cov}\{\widehat{\mathbf{w}}\}$

- What does $\text{cov}\{\widehat{\mathbf{w}}\}$ tell us?
- Recall the linear model, $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$

$$\text{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

- Tells us how well defined the parameters are by the data. How much can the parameters vary and still give a **good** model.
  - a and c: how much can we change w_0 and w_1.
  - b: how the values should be changed together.



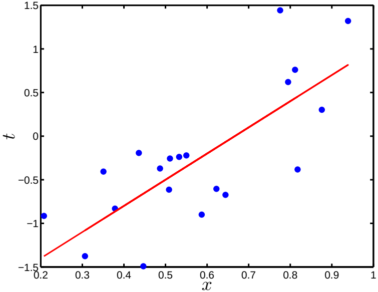
$\mathrm{cov}\{\widehat{\mathbf{w}}\}$

$$\begin{aligned}\mathrm{cov}\{\widehat{\mathbf{w}}\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\right\} \\ &\quad -\underbrace{\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}}_{w}\,\underbrace{\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}^{\mathsf{T}}}_{w^T}\end{aligned}$$

$\mathrm{cov}\{\widehat{\mathbf{w}}\}$

$$
\begin{aligned}
\mathrm{cov}\{\widehat{\mathbf{w}}\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\right\} \\
&\quad -\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\widehat{\mathbf{w}}\}\,\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\widehat{\mathbf{w}}\}^{\mathsf{T}} \\
&= \mathbf{E}\left\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\right\} - \mathbf{w}\mathbf{w}^{\mathsf{T}} \\
&= \vdots \\
\mathrm{cov}\{\widehat{\mathbf{w}}\} &= \sigma^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}
\end{aligned}
$$

Example

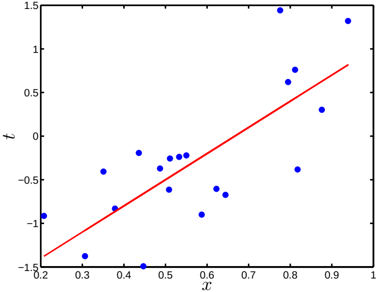


$$t_n = \overbrace{-2}^{w_0} + \overbrace{3}^{w_1} x_n + \epsilon_n$$

$$p(\epsilon_n) = \mathcal{N}(0, \sigma^2)$$

$$\sigma^2 = \underline{0.5^2}$$

Example



$$t_n = -2 + 3x_n + \epsilon_n$$

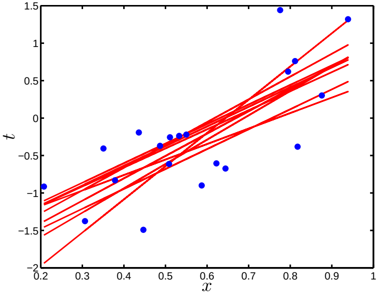
$$p(\epsilon_n) = \mathcal{N}(0, \sigma^2)$$

$$\sigma^2 = 0.5^2$$

$$\widehat{\mathbf{w}} = \begin{bmatrix} -1.95 \\ 2.94 \end{bmatrix}, \quad \mathrm{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} 0.1195 & -0.1847 \\ -0.1847 & 0.3190 \end{bmatrix}$$

Example

$\hat{w} \longrightarrow \begin{cases} \text{bias} \\ \text{variance/co-var} \end{cases}$



$$t_n = -2 + 3x_n + \epsilon_n$$

$$p(\epsilon_n) = \mathcal{N}(0, \sigma^2)$$

$$\sigma^2 = 0.5^2$$

$$\widehat{\mathbf{w}} = \begin{bmatrix} -1.95 \\ 2.94 \end{bmatrix}, \quad \mathrm{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} 0.1195 & -0.1847 \\ -0.1847 & 0.3190 \end{bmatrix}$$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\}$ – beyond this class

We saw that $\widehat{\mathbf{w}}$ was unbiased, what about $\widehat{\sigma^2}$?

$$
\begin{aligned}
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} &= \frac{1}{N}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{(\mathbf{t}-\mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t}-\mathbf{X}\widehat{\mathbf{w}})\right\} \\
&= \sigma^2\left(1-\frac{D}{N}\right).
\end{aligned}
$$

Useful identity

$$
\begin{aligned}
p(\mathbf{t}) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\mathbf{E}_{p(\mathbf{t})}\left\{\mathbf{t}^\mathsf{T}\mathbf{A}\mathbf{t}\right\} &= \mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\mathsf{T}\mathbf{A}\boldsymbol{\mu} \\
\mathrm{Tr}(\mathbf{A}) &= \sum_i A_{ii}
\end{aligned}
$$

Another useful identity

$$
\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})
$$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\}$ – beyond this class

$$
\begin{aligned}
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} &= \frac{1}{N}(\mathrm{Tr}(\sigma^2\mathbf{I}) + \mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}) \\
&\quad -\frac{1}{N}(\mathrm{Tr}(\sigma^2\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}) + \mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}) \\
&= \sigma^2 - \frac{\sigma^2}{N}\mathrm{Tr}(\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}) \\
&= \sigma^2 - \frac{\sigma^2}{N}\mathrm{Tr}(\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}) \\
&= \sigma^2\left(1 - \frac{D}{N}\right)
\end{aligned}
$$

Where D is the number of columns in $\mathbf{X}$ (the number of elements in $\mathbf{w}$.

Another useful identity

$$\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\}$ – beyond this class

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2\left(1 - \frac{D}{N}\right)$$

- In general $D < N$.
- So $1 - D/N < 1$.
- So $\widehat{\sigma^2} < \sigma^2$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\}$ – beyond this class

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2\left(1 - \frac{D}{N}\right)$$

- In general $D < N$.
- So $1 - D/N < 1$.
- So $\widehat{\sigma^2} < \sigma^2$
- $\widehat{\sigma^2}$ is biased and will generally be too low.

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\}$ – beyond this class

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2\left(1 - \frac{D}{N}\right)$$

- In general $D < N$.
- So $1 - D/N < 1$.
- So $\widehat{\sigma^2} < \sigma^2$
- $\widehat{\sigma^2}$ is biased and will generally be too low.
- Why?
  - Because it is based on $\widehat{\mathbf{w}}$ which will, in general, be closer to the data than $\mathbf{w}$.

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\}$ – beyond this class

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2\left(1 - \frac{D}{N}\right)$$

- In general $D < N$.
- So $1 - D/N < 1$.
- So $\widehat{\sigma^2} < \sigma^2$
- $\widehat{\sigma^2}$ is biased and will generally be too low.
- Why?
  - Because it is based on $\widehat{\mathbf{w}}$ which will, in general, be closer to the data than $\mathbf{w}$.
- As N increases, $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} \to \sigma^2$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\}$ – beyond this class

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2\left(1 - \frac{D}{N}\right)$$

- In general $D < N$.
- So $1 - D/N < 1$.
- So $\widehat{\sigma^2} < \sigma^2$
- $\widehat{\sigma^2}$ is biased and will generally be too low.
- Why?
  - Because it is based on $\widehat{\mathbf{w}}$ which will, in general, be closer to the data than $\mathbf{w}$.
- As N increases, $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} \rightarrow \sigma^2$
- To think about – what if $D = N$ or $D > N$?

Example – beyond this class

Generate 100 datasets from the following model:

$$t_n = w_0 + w_1 x_n + \epsilon_n, \;\; p(\epsilon_n) = \mathcal{N}(0, 0.25)$$

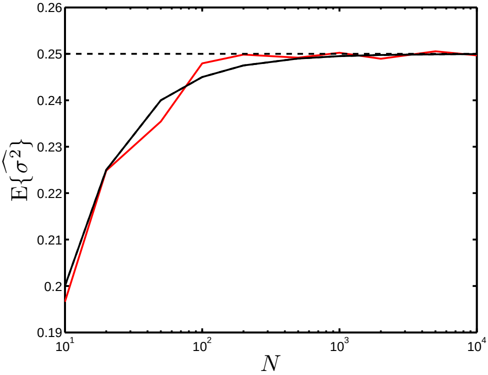
For $N = [10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000]$

Example – beyond this class

Generate 100 datasets from the following model:

$$t_n = w_0 + w_1 x_n + \epsilon_n, \;\; p(\epsilon_n) = \mathcal{N}(0, 0.25)$$

For $N = [10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000]$



Red curve – average $\widehat{\sigma^2}$ over 100 datasets. Black curve – theoretical value. Dashed line – true value.

Summary

- Computed $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} = \mathbf{w}$
  - $\widehat{\mathbf{w}}$ is **unbiased**.

Summary

- Computed $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} = \mathbf{w}$
  - $\widehat{\mathbf{w}}$ is **unbiased**.
- Computed $\text{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$
  - Tells us how much slack there is in our parameters.

Summary

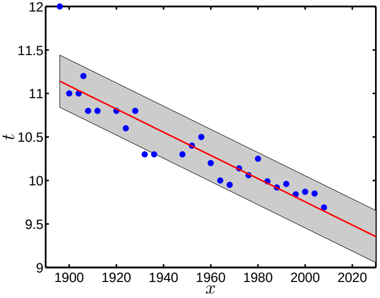
- Computed $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} = \mathbf{w}$
  - $\widehat{\mathbf{w}}$ is **unbiased**.
- Computed $\text{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$
  - Tells us how much slack there is in our parameters.
- Computed $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2(1 - D/N)$ [*beyond this class!*]
  - $\widehat{\sigma^2}$ is **biased**.
  - Gets better and better as we get more data.

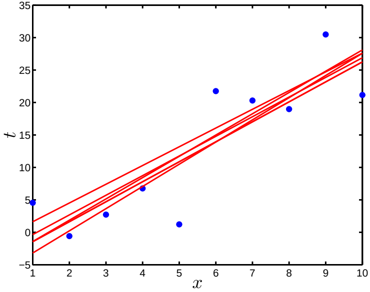
Predictions

- Our aim is to make predictions (e.g. London 2012)

Predictions

- Our aim is to make predictions (e.g. London 2012)
- The noise in our data tells us that we can't predict exactly.





Predictions

- Our model is defined as:

$$t = \mathbf{w}^{\mathsf{T}}\mathbf{x} + \epsilon$$

- Given our estimate of the parameters, $\widehat{\mathbf{w}}$ and a new input, $\mathbf{x}_{\text{new}}$, if we had to predict a single value:

$$\underline{t_{\text{new}}} = \widehat{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_{\text{new}} \quad + \cancel{\epsilon_{\text{new}}}$$

$$E(\text{new}) = 0$$

- Is this sensible?

bias

variance

Predictions

- Our model is defined as:

$$t = \mathbf{w}^\mathsf{T}\mathbf{x} + \epsilon$$

- Given our estimate of the parameters, $\widehat{\mathbf{w}}$ and a new input, $\mathbf{x}_{\text{new}}$, if we had to predict a single value:

$$t_{\text{new}} = \widehat{\mathbf{w}}^\mathsf{T}\mathbf{x}_{\text{new}}$$

- Is this sensible? What is $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t_{\text{new}}\right\}$?

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t_{\text{new}}\right\} = \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}^\mathsf{T}\mathbf{x}_{\text{new}}\right\} = \mathbf{w}^\mathsf{T}\mathbf{x}_{\text{new}}$$

- which is a good thing!

$E(\hat{w}^T) = w^T$

Predictions

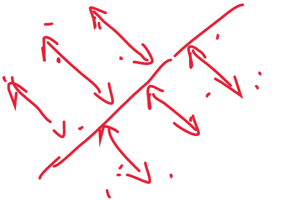
- What about $\text{var}\{t_{\text{new}}\}$?

$$\text{var}\{t_{\text{new}}\} = \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t_{\text{new}}^2\right\} - \underline{\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t_{\text{new}}\right\}^2}$$

Predictions

- What about $\mathrm{var}\{t_{\text{new}}\}$?

$$
\begin{aligned}
\mathrm{var}\{t_{\text{new}}\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t_{\text{new}}^2\right\} - \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t_{\text{new}}\right\}^2 \\
&= \mathbf{E}\left\{(\widehat{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_{\text{new}})^2\right\} - (\mathbf{w}^{\mathsf{T}}\mathbf{x}_{\text{new}})^2 \\
&= \mathbf{x}_{\text{new}}^{\mathsf{T}}\mathbf{E}\left\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\right\}\mathbf{x}_{\text{new}} - \mathbf{x}_{\text{new}}^{\mathsf{T}}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{x}_{\text{new}} \\
&= \vdots \\
\mathrm{var}\{t_{\text{new}}\} &= \sigma^2\mathbf{x}_{\text{new}}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_{\text{new}}
\end{aligned}
$$



Prediction and variance

$$
\begin{aligned}
t_{\text{new}} &= \widehat{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_{\text{new}} \\
\text{var}\{t_{\text{new}}\} &= \sigma^2 \mathbf{x}_{\text{new}}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_{\text{new}}
\end{aligned}
$$

Prediction and variance

$$
\begin{aligned}
t_{\text{new}} &= \widehat{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_{\text{new}} \\
\text{var}\{t_{\text{new}}\} &= \sigma^2 \mathbf{x}_{\text{new}}^{\mathsf{T}} (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} \mathbf{x}_{\text{new}}
\end{aligned}
$$

- Recall the expression for the covariance of the parameter estimate:

$$
\text{cov}\{\widehat{\mathbf{w}}\} = \sigma^2 (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}
$$

Prediction and variance

$$
\begin{aligned}
t_{\text{new}} &= \widehat{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_{\text{new}} \\
\text{var}\{t_{\text{new}}\} &= \sigma^2 \mathbf{x}_{\text{new}}^{\mathsf{T}} (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_{\text{new}}
\end{aligned}
$$

- Recall the expression for the covariance of the parameter estimate:
$$\text{cov}\{\widehat{\mathbf{w}}\} = \sigma^2 (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$$
- Appears in the variance of the prediction:

$$\text{var}\{t_{\text{new}}\} = \mathbf{x}_{\text{new}}^{\mathsf{T}} \text{cov}\{\widehat{\mathbf{w}}\} \mathbf{x}_{\text{new}}$$

Prediction and variance

$$
\begin{aligned}
t_{\text{new}} &= \widehat{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_{\text{new}} \\
\text{var}\{t_{\text{new}}\} &= \sigma^2 \underline{\mathbf{x}_{\text{new}}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_{\text{new}}}
\end{aligned}
$$

- Recall the expression for the covariance of the parameter estimate:
$$\text{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$$
- Appears in the variance of the prediction:
$$\text{var}\{t_{\text{new}}\} = \mathbf{x}_{\text{new}}^{\mathsf{T}}\text{cov}\{\widehat{\mathbf{w}}\}\mathbf{x}_{\text{new}}$$
- If the variance in the parameters is high, so is the variance in the predictions.

Example

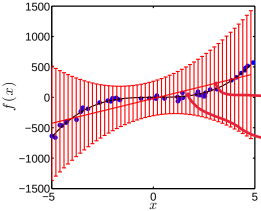
Data sampled from a 3rd order polynomial function:

$$t = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \epsilon$$

Example

Data sampled from a 3rd order polynomial function:

$$t = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \epsilon$$



estimated model

true model

Linear

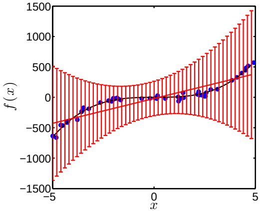
Plots show $t_{\text{new}} \pm \text{var}\{t_{\text{new}}\}$. (Black line is truth).

$w_0 + w_1 x$

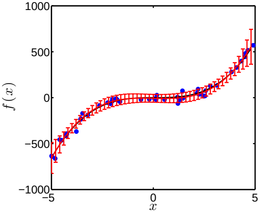
Example

Data sampled from a 3rd order polynomial function:

$$t = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \epsilon$$



Linear



Cubic

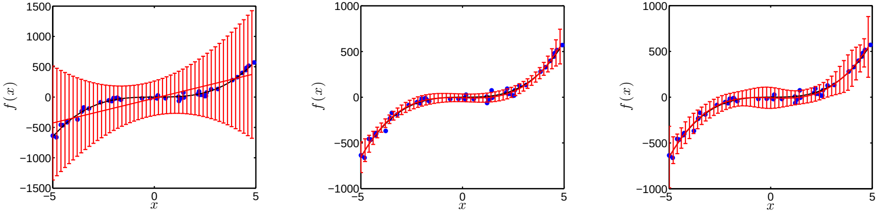
Plots show $t_{\text{new}} \pm \text{var}\{t_{\text{new}}\}$. (Black line is truth).

$w_0 + w_1 x + w_2 x^2 + w_3 x^3$

Example

Data sampled from a 3rd order polynomial function:

$$t = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \epsilon$$



Linear Cubic 6th order

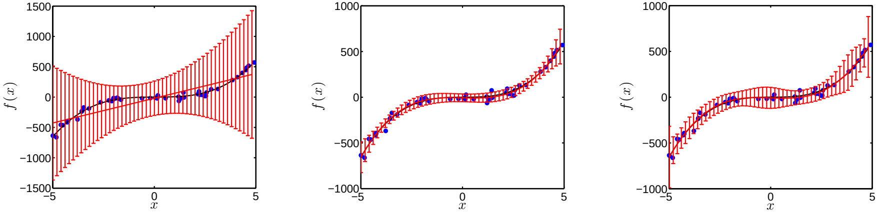
Plots show $t_{\text{new}} \pm \text{var}\{t_{\text{new}}\}$. (Black line is truth).

$w_0 + \cdots x w_6 x^{\phi}$

Example

Data sampled from a 3rd order polynomial function:

$$t = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \epsilon$$



Linear | Cubic | 6th order

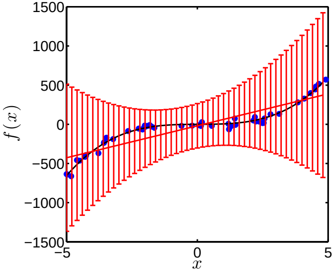
Plots show $t_{\text{new}} \pm \text{var}\{t_{\text{new}}\}$. (Black line is truth).

Why does the predictive variance increase above and below the correct order?

Not complex enough model – more 'noise'

In practice we don't know σ^2 so substitute $\widehat{\sigma^2}$:

$$\text{var}\{t_{\text{new}}\} = \widehat{\sigma^2}\mathbf{x}_{\text{new}}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_{\text{new}}$$

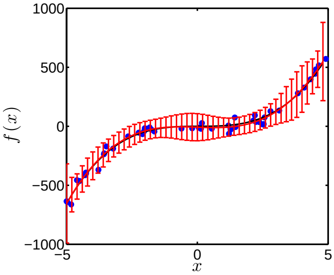


- The model is too simple.
- Some true variability can only be modelled noise.
- $\widehat{\sigma^2}$ is significantly over-estimated.
- Results in high $\text{var}\{t_{\text{new}}\}$.

Too complex model – parameters not well defined

Similarly, we substitute $\widehat{\sigma^2}$ into expression for $\text{cov}\{\widehat{\mathbf{w}}\}$:

$$\text{cov}\{\widehat{\mathbf{w}}\} = \widehat{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$$



- 6th order model is too flexible.
- Many sets of parameters lead to a good model.
- Means that $\text{cov}\{\widehat{\mathbf{w}}\}$ is high.

Olympic prediction

Linear model:

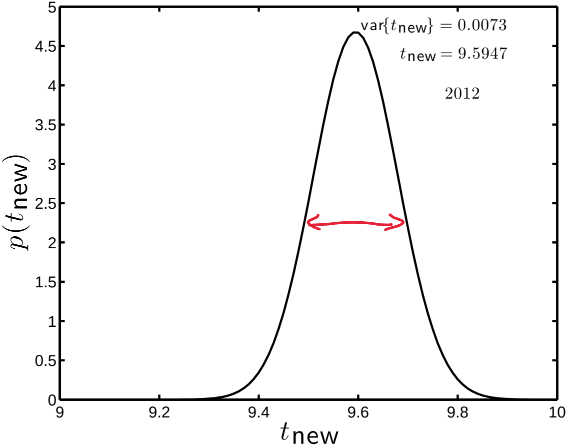
$$t = w_0 + w_1 x + \epsilon$$

Olympic prediction

$\hat{w}^T x_{new}$

Linear model:

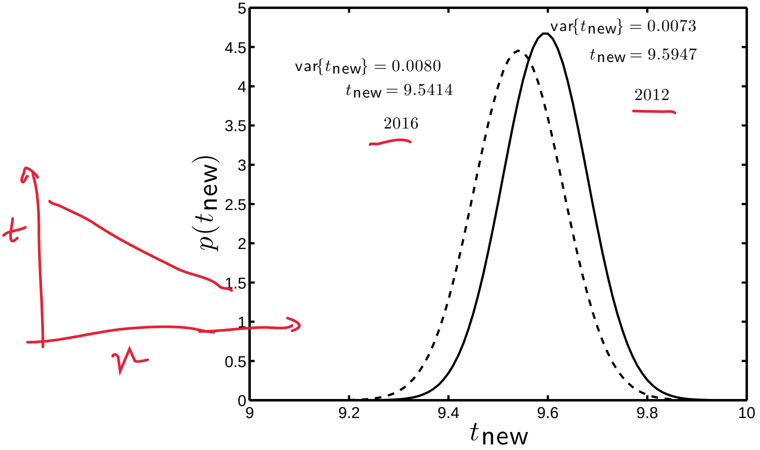
$$t = w_0 + w_1 x + \epsilon$$



Olympic prediction

Linear model:

$$t = w_0 + w_1 x + \epsilon$$

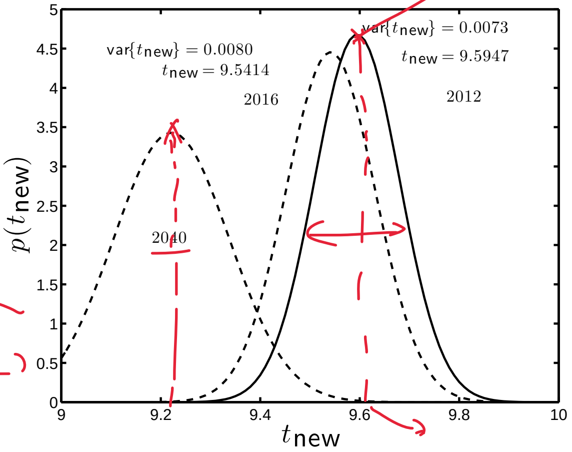


Predictive variance increases as we get further from the training data.

Olympic prediction

Linear model:

$$t = w_0 + w_1 x + \epsilon$$



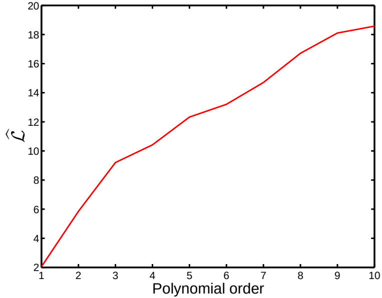
Predictive variance increases as we get further from the training data.

Can we use likelihood to choose models?

- We've already seen that training loss is no good for model choice.
- Described cross-validation as an alternative.
- Can we use the likelihood L or $\log L$?

Can we use likelihood to choose models?

- We've already seen that training loss is no good for model choice.
- Described cross-validation as an alternative.
- Can we use the likelihood L or $\log L$?

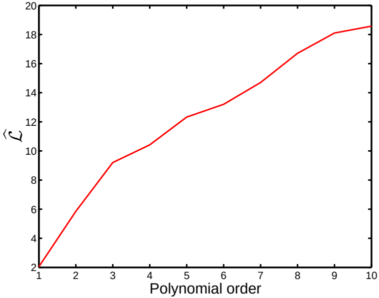


Data from 3rd order polynomial.

- No.

Can we use likelihood to choose models?

- We've already seen that training loss is no good for model choice.
- Described cross-validation as an alternative.
- Can we use the likelihood L or $\log L$?



Data from 3rd order polynomial.

- No.
  - More complex models can always get closer to the data.

Summary

- Decided to model the noise.
- Recapped random variables.
- Introduced likelihood and maximised it to find $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$.
- What did it buy us?

Summary

- Decided to model the noise.
- Recapped random variables.
- Introduced likelihood and maximised it to find $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$.
- What did it buy us?
- We can now:
  - Quantify the uncertainty in our parameters.
  - Quantify the uncertainty in our predictions.
  - This is very important in all applications....

Summary

- Decided to model the noise.
- Recapped random variables.
- Introduced likelihood and maximised it to find $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$.
- What did it buy us?
- We can now:
  - Quantify the uncertainty in our parameters.
  - Quantify the uncertainty in our predictions.
  - This is very important in all applications....
- What next?
  - Going Bayesian.
  - Got to forget about single parameter values - parameters are random variables too.

Aside - from one model to many

- All of our efforts so far have been to find the 'best' model:
  - The one that minimises the loss.
  - The one that maximises the likelihood.
- Given the uncertainty, maybe we shouldn't trust one on its own?
- Consider the following random variable (RV):

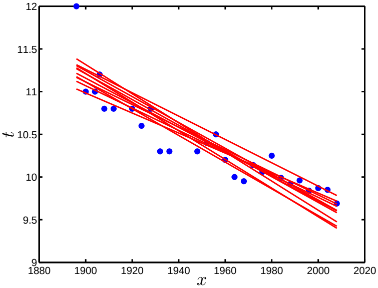
$$p(\mathbf{q}) = \mathcal{N}(\widehat{\mathbf{w}}, \mathrm{cov}\{\widehat{\mathbf{w}}\})$$

- Samples of this RV $\mathbf{q}_s$ are **models** (assume $\widehat{\sigma^2}$ is fixed)
- We can generate lots of good models...

- Sample lots of $\mathbf{q}$ from:

$$p(\mathbf{q}) = \mathcal{N}(\widehat{\mathbf{w}}, \text{cov}\{\widehat{\mathbf{w}}\})$$

- Each corresponds to a model.

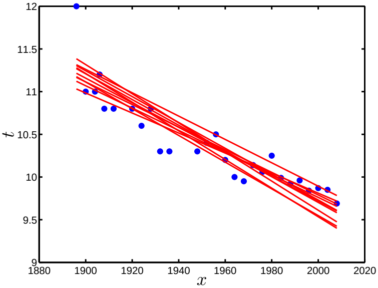


- Sample lots of $\mathbf{q}$ from:

$$p(\mathbf{q}) = \mathcal{N}(\widehat{\mathbf{w}}, \mathrm{cov}\{\widehat{\mathbf{w}}\})$$

- Each corresponds to a model.
- Compute a prediction from each one:

$$t_s = \mathbf{q}_s^{\mathsf{T}} \mathbf{x}_{\mathrm{new}}$$



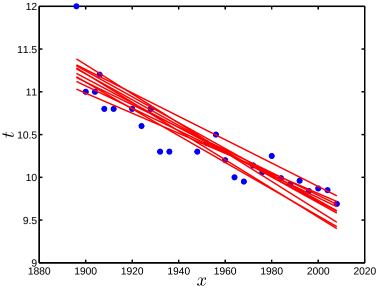
- Sample lots of $\mathbf{q}$ from:

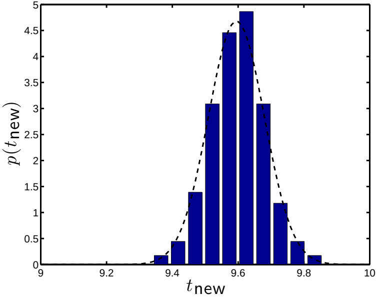
$$p(\mathbf{q}) = \mathcal{N}(\widehat{\mathbf{w}}, \mathrm{cov}\{\widehat{\mathbf{w}}\})$$

- Each corresponds to a model.
- Compute a prediction from each one:

$$t_s = \mathbf{q}_s^{\mathsf{T}} \mathbf{x}_{\mathrm{new}}$$

- Look at the distribution of predictions:





Do we need to take samples at all?

- Take an expectation...

$$\mathbf{E}_{p(\mathbf{q})}\left\{t_{\text{new}}\right\} = \int t_{\text{new}} \mathcal{N}(\widehat{\mathbf{w}}, \text{cov}\{\widehat{\mathbf{w}}\})\ dt_{\text{new}}$$

- We'll see more of this in the next lecture....