Machine Learning
Bias–Variance Tradeoff

Devdatt Dubhashi
dubhashi@chalmers.se

Dept. of Computer Science and Engg.
Chalmers University

March 26, 2018

Optimum parameters

- We have point estimates of our parameters.
- How confident should we be in them?
  - If we changed them a little bit, would the model still be good?

Confidence in parameter estimates

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

Confidence in parameter estimates

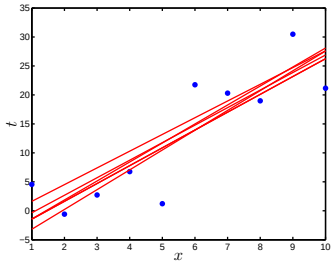
- Imagine there are **true** parameters, $\mathbf{w}$ and σ^2.
- How good our our estimates $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$?
  - Are they correct (on average)?
  - If we could keep adding data, would we converge on the true value?

Confidence in parameter estimates

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

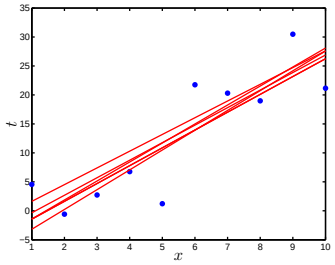
Confidence in parameter estimates

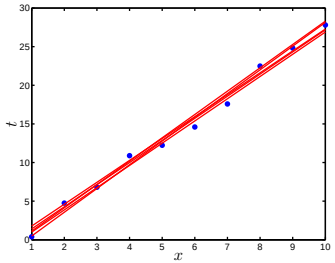
- Imagine there are **true** parameters, $\mathbf{w}$ and σ^2.
- How good our our estimates $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$?
  - Are they correct (on average)?
  - If we could keep adding data, would we converge on the true value?
- How confident should we be in our estimates?
  - Could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



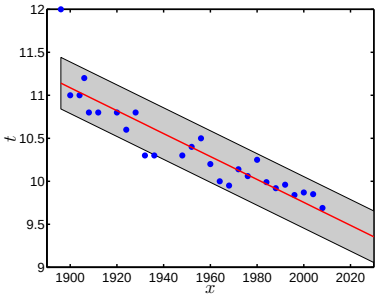


Predictions

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

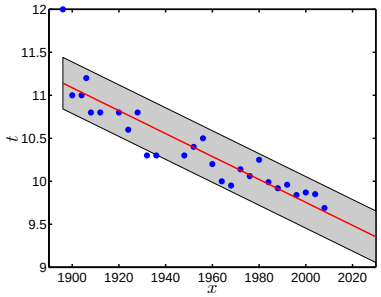
Predictions

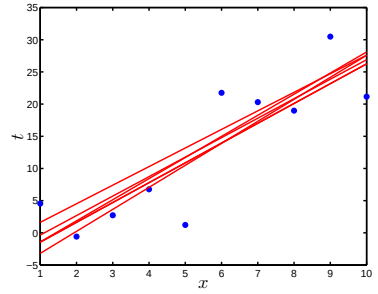
- Our aim is to make predictions (e.g. London 2012)
- The noise in our data tells us that we can't predict exactly.



Predictions

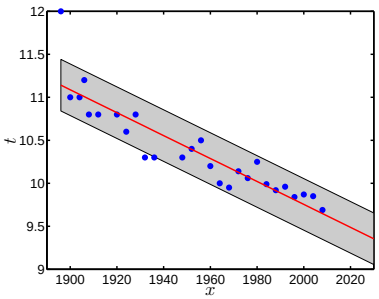
- Our aim is to make predictions (e.g. London 2012)
- The noise in our data tells us that we can't predict exactly.
- The uncertainty in the parameters $\text{cov}\{\widehat{\mathbf{w}}\}$ should make them even less certain.

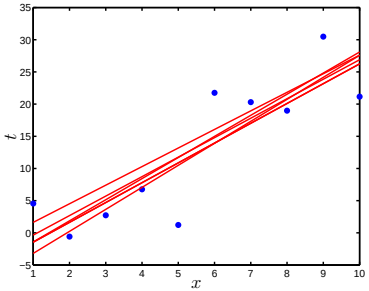




Predictions

- Our aim is to make predictions (e.g. London 2012)
- The noise in our data tells us that we can't predict exactly.
- The uncertainty in the parameters $\text{cov}\{\widehat{\mathbf{w}}\}$ should make them even less certain.





- We'll also (hopefully) clear up some loose ends.

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?

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 $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{t_{\text{new}}\}$?

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

- which is a good thing!

Predictions

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

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

Predictions

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

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

Prediction and variance

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

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

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

- 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

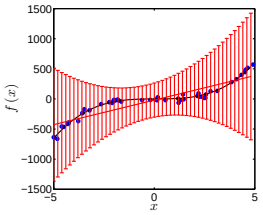
Data sampled from a 3rd order polynomial function:

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

Example

Data sampled from a 3rd order polynomial function:

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



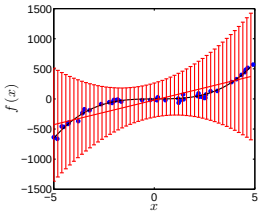
Linear

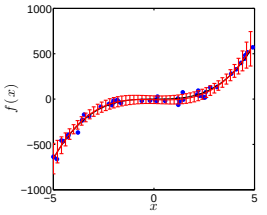
Plots show $t_{\text{new}} \pm \text{var}\{t_{\text{new}}\}$. (Black line is truth).

Example

Data sampled from a 3rd order polynomial function:

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





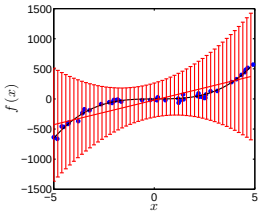
Linear Cubic

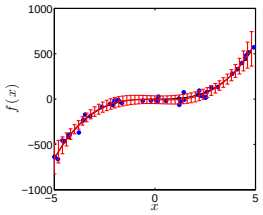
Plots show $t_{\text{new}} \pm \text{var}\{t_{\text{new}}\}$. (Black line is truth).

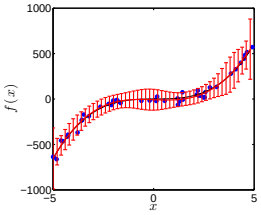
Example

Data sampled from a 3rd order polynomial function:

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







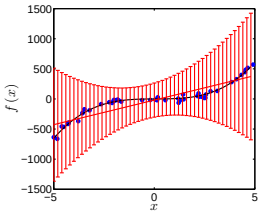
Linear Cubic 6th order

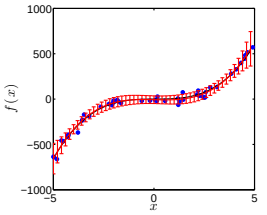
Plots show $t_{\text{new}} \pm \text{var}\{t_{\text{new}}\}$. (Black line is truth).

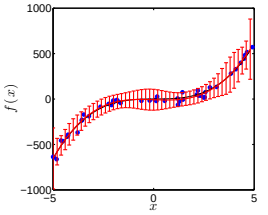
Example

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

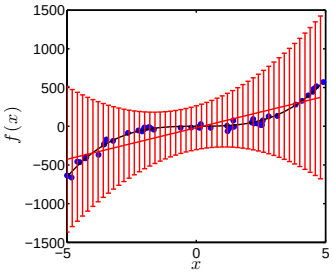
Plots show $t_{\text{new}} \pm \text{var}\{t_{\text{new}}\}$. (Black line is truth).

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 $\widehat{\sigma^2}$:

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

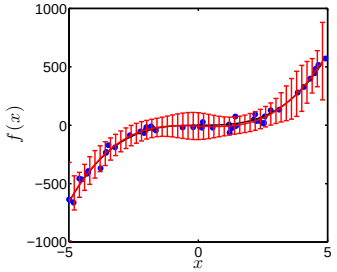


- The model is too simple.
- Some true variability can only be modelled noise.
- $\widehat{\sigma^2}$ is significantly over-estimated.
- Results in high $\text{var}\{t_{\text{new}}\}$.

Too complex model – parameters not well defined

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

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

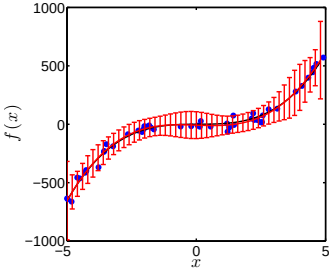


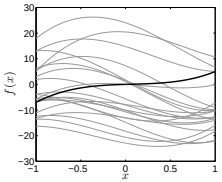
- 6th order model is too flexible.
- Many sets of parameters lead to a good model.
- Means that $\text{cov}\{\widehat{\mathbf{w}}\}$ is high.

Too complex model – parameters not well defined

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

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





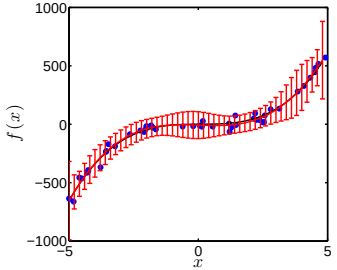
- 6th order model is too flexible.
- Many sets of parameters lead to a good model.
- Means that $\text{cov}\{\widehat{\mathbf{w}}\}$ is high.

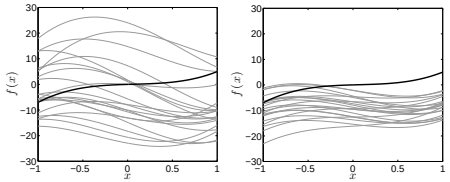
- 'good' 6th order models.

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.

- 'good' 6th order models.
- 'good' 3rd order models.

Olympic prediction

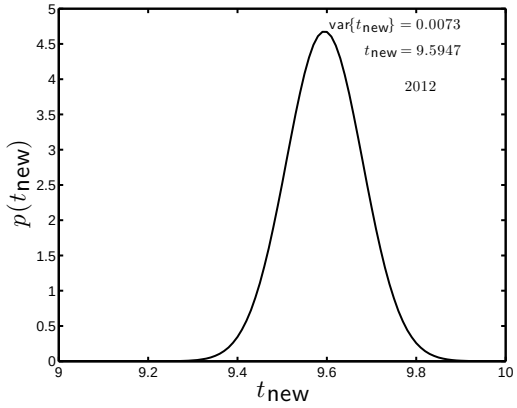
Linear model:

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

Olympic prediction

Linear model:

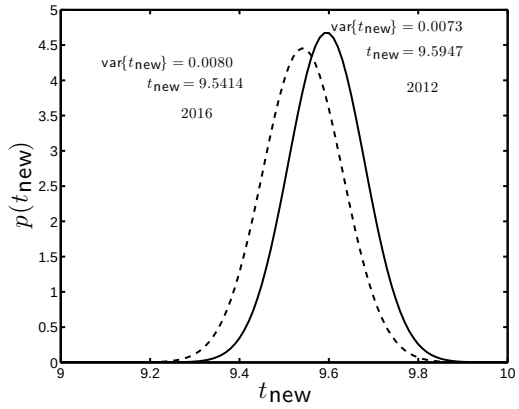
$$t = w_0 + w_1 x + \epsilon$$



Olympic prediction

Linear model:

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

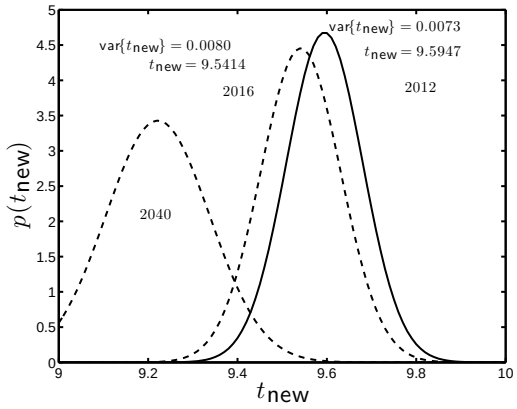


Predictive variance increases as we get further from the training data.

Olympic prediction

Linear model:

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



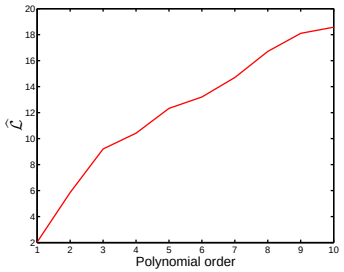
Predictive variance increases as we get further from the training data.

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 $\log L$?

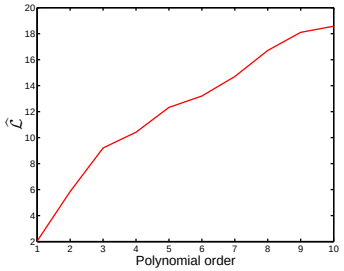


Data from 3rd order polynomial.

- 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 $\log L$?

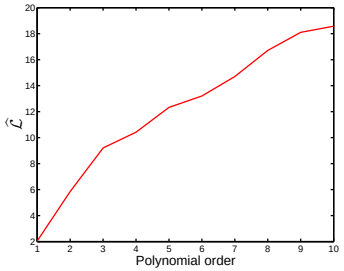


Data from 3rd order polynomial.

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

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.
  - Results in lower $\widehat{\sigma^2}$ and higher likelihood.

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 RV:

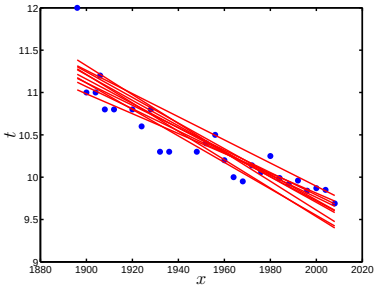
$$p(\mathbf{q}) = \mathcal{N}(\widehat{\mathbf{w}}, \text{cov}\{\widehat{\mathbf{w}}\})$$

- 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.

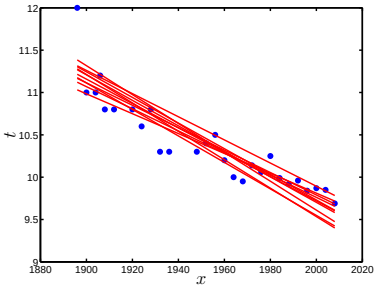


- Sample lots of $\mathbf{q}$ from:

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

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

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



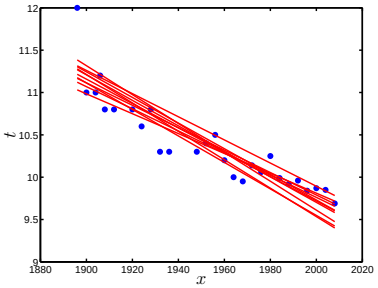
- Sample lots of $\mathbf{q}$ from:

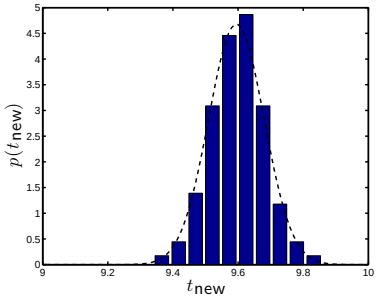
$$p(\mathbf{q}) = \mathcal{N}(\widehat{\mathbf{w}}, \text{cov}\{\widehat{\mathbf{w}}\})$$

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

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

- Look at the distribution of predictions:





Do we need to take samples at all?

- Take an expectation...

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

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