Machine Learning

Lecture. 2.

Mark Girolami

girolami@dcs.gla.ac.uk

Department of Computing Science
University of Glasgow

Linear Regression

UNIVERSITY
of
GLASGOW

- *Learning* or *Inferring* a functional relationship between a set of attribute variables and associated response or target variables

Linear Regression

UNIVERSITY
of
GLASGOW

- *Learning* or *Inferring* a functional relationship between a set of attribute variables and associated response or target variables
- Motivation to use model of relationship to predict unknown target values given new values of attributes

Linear Regression

UNIVERSITY *of* GLASGOW

- *Learning* or *Inferring* a functional relationship between a set of attribute variables and associated response or target variables
- Motivation to use model of relationship to predict unknown target values given new values of attributes
- How to learn the relationship from finite set of observations?

Linear Regression

- *Learning* or *Inferring* a functional relationship between a set of attribute variables and associated response or target variables
- Motivation to use model of relationship to predict unknown target values given new values of attributes
- How to learn the relationship from finite set of observations?
- How to assess how good model is as a predictor?

Example Prediction Problem

UNIVERSITY
of
GLASGOW

- Predict Long Jump Gold Medal distance based on previous winning performances

Example Prediction Problem

UNIVERSITY *of* GLASGOW

- Predict Long Jump Gold Medal distance based on previous winning performances
- Data available corresponds to distance and year of games

Example Prediction Problem

UNIVERSITY
of
GLASGOW

- Predict Long Jump Gold Medal distance based on previous winning performances
- Data available corresponds to distance and year of games
- Many other attributes also available which are indicative of target variable

Example Prediction Problem

UNIVERSITY
of
GLASGOW

- Predict Long Jump Gold Medal distance based on previous winning performances
- Data available corresponds to distance and year of games
- Many other attributes also available which are indicative of target variable
- However lets see what sort of predictions, if any, can be made taking account only of time elapsed from first games

Example Prediction Problem

UNIVERSITY
of
GLASGOW

- Look at data available by plotting distance against time elapsed

Example Prediction Problem

UNIVERSITY of GLASGOW

- Look at data available by plotting distance against time elapsed

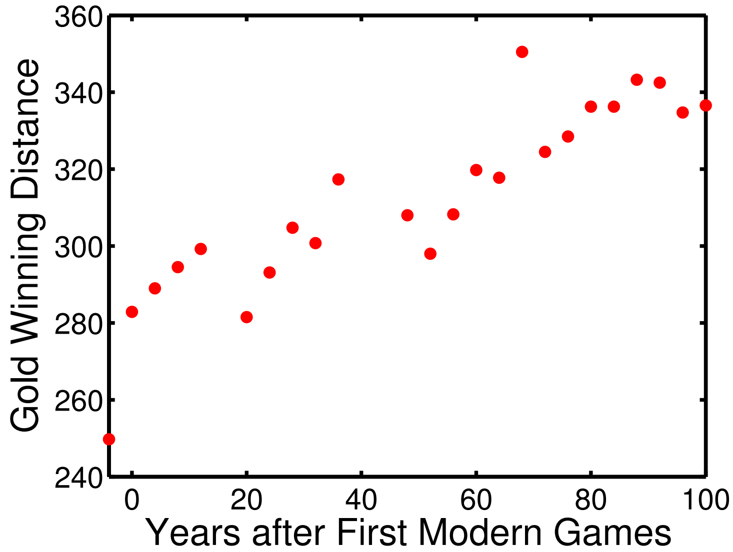


Figure 1: Gold Medal Distance for the long jump from 1896 to 2004 plotted against the number of years since the first modern games were held with 1900 being 0 and 1896 being -4. Note that the two world wars interrupt the games in 1914, 1940 & 1944.

Linear Model

UNIVERSITY
of
GLASGOW

- Visually there appears to be a functional relationship between attributes and targets

Linear Model

UNIVERSITY *of* GLASGOW

- Visually there appears to be a functional relationship between attributes and targets
- A class of functionals which maps integers ($\mathbb{Z}$) to the Real line ($\mathbb{R}$) has to be considered such that

$$f : \mathbb{Z} \rightarrow \mathbb{R}$$

Linear Model

UNIVERSITY of GLASGOW

- Visually there appears to be a functional relationship between attributes and targets
- A class of functionals which maps integers ($\mathbb{Z}$) to the Real line ($\mathbb{R}$) has to be considered such that

$$f : \mathbb{Z} \rightarrow \mathbb{R}$$

- It seems reasonable that a linear relationship exits so assume that

$$f(x; w_0, w_1) = w_1 x + w_0$$

defines our model. The slope w_1 and the intercept w_0 are the *free parameters* of our model which have to be assigned

Loss Functions

- We *identify the model parameters* by considering a *Loss Function* defining the miss-match between model output $f(x; w_0, w_1)$ and target value t

Loss Functions

UNIVERSITY of GLASGOW

- We *identify the model parameters* by considering a *Loss Function* defining the miss-match between model output $f(x; w_0, w_1)$ and target value t
- Loss defined for *all* available input-output example pairs (x_n, t_n) where $n = 1, \cdots, N$ and in this case $N = 25$, the number of game results recorded.

Loss Functions

UNIVERSITY *of* GLASGOW

- We *identify the model parameters* by considering a *Loss Function* defining the miss-match between model output $f(x; w_0, w_1)$ and target value t
- Loss defined for *all* available input-output example pairs (x_n, t_n) where $n = 1, \cdots, N$ and in this case $N = 25$, the number of game results recorded.
- The sample average loss is given as

$$\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(t_n, f(x_n; w_0, w_1))$$

Squared-Error Loss

- The notion of *Loss* is quite general and now need a specific loss function

Squared-Error Loss

UNIVERSITY *of* GLASGOW

- The notion of *Loss* is quite general and now need a specific loss function
- Squared Error Loss is a sensible choice - historical significance, also has probabilistic basis

Squared-Error Loss

UNIVERSITY
of
GLASGOW

- The notion of *Loss* is quite general and now need a specific loss function
- Squared Error Loss is a sensible choice - historical significance, also has probabilistic basis
- Robust losses based on absolute deviations can also be considered

Squared-Error Loss

- The notion of *Loss* is quite general and now need a specific loss function
- Squared Error Loss is a sensible choice - historical significance, also has probabilistic basis
- Robust losses based on absolute deviations can also be considered
- Sample Mean Squared Error (MSE) Loss

$$\frac{1}{N}\sum_{n=1}^{N}|t_n - f(x_n; w_0, w_1)|^2$$

Matrix Notation

UNIVERSITY of GLASGOW

- We can define the 2×1 dimensional column vector $\mathbf{w}$ and the $N \times 1$ dimensional column vector $\mathbf{t}$ such that

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \quad \& \quad \mathbf{t} = \begin{bmatrix} t_1 \\ \vdots \\ t_N \end{bmatrix}$$

Matrix Notation

- We can define the 2×1 dimensional column vector $\mathbf{w}$ and the $N \times 1$ dimensional column vector $\mathbf{t}$ such that

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \quad \& \quad \mathbf{t} = \begin{bmatrix} t_1 \\ \vdots \\ t_N \end{bmatrix}$$

- The $N \times 2$ dimensional matrix $\mathbf{X}$ is defined as

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}$$

Squared-Error Loss

- Using the defined vector & matrix notation the MSE can be written compactly as

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

Squared-Error Loss

UNIVERSITY *of* GLASGOW

- Using the defined vector & matrix notation the MSE can be written compactly as

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

- Tutorial exercise to show that MSE can be written as above

Squared-Error Loss

UNIVERSITY
of
GLASGOW

- Using the defined vector & matrix notation the MSE can be written compactly as

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

- Tutorial exercise to show that MSE can be written as above
- Now require to find value of vector, $\mathbf{w}$, which minimises MSE

Minimising MSE

- Find stationary point of MSE by setting gradient of all partial derivatives to zero

Minimising MSE

UNIVERSITY *of* GLASGOW

- Find stationary point of MSE by setting gradient of all partial derivatives to zero

$$
\frac{\partial MSE}{\partial \mathbf{w}} = \begin{bmatrix} \frac{\partial MSE}{\partial w_0} \\ \\ \frac{\partial MSE}{\partial w_1} \\ \end{bmatrix}
$$

$$
= \begin{bmatrix} -\frac{2}{N}\sum_{n=1}^{N}(t_n - f(x_n; w_0, w_1)) \\ \\ -\frac{2}{N}\sum_{n=1}^{N}(t_n - f(x_n; w_0, w_1))x_n \\ \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}
$$

Stationary Point

UNIVERSITY *of* GLASGOW

- Employing vector & matrix notation the gradient of MSE can be written neatly as

Stationary Point

UNIVERSITY of GLASGOW

- Employing vector & matrix notation the gradient of MSE can be written neatly as

$$\frac{\partial MSE}{\partial \mathbf{w}} = -\frac{2}{N}\mathbf{X}^{\mathsf{T}}(\mathbf{t} - \mathbf{X}\mathbf{w})$$

Stationary Point

UNIVERSITY of GLASGOW

- Employing vector & matrix notation the gradient of MSE can be written neatly as

$$\frac{\partial MSE}{\partial \mathbf{w}} = -\frac{2}{N}\mathbf{X}^{\mathsf{T}}(\mathbf{t} - \mathbf{X}\mathbf{w})$$

- Tutorial exercise to show this. Matrix Cookbook on Module website.

Stationary Point

- Employing vector & matrix notation the gradient of MSE can be written neatly as

$$\frac{\partial MSE}{\partial \mathbf{w}} = -\frac{2}{N}\mathbf{X}^{\mathsf{T}}(\mathbf{t} - \mathbf{X}\mathbf{w})$$

- Tutorial exercise to show this. Matrix Cookbook on Module website.
- Is stationary point a minimum, maximum or saddle point?

Stationary Point

- Schoolboy calculus for single variable functions if second-derivatives at stationary point strictly positive, then point is minimum of function

Stationary Point

UNIVERSITY of GLASGOW

- Schoolboy calculus for single variable functions if second-derivatives at stationary point strictly positive, then point is minimum of function
- Multi-parameter function use generalisation of above rule

Stationary Point

UNIVERSITY *of* GLASGOW

- Schoolboy calculus for single variable functions if second-derivatives at stationary point strictly positive, then point is minimum of function
- Multi-parameter function use generalisation of above rule
- Matrix of all partial second-derivatives, $\mathbf{H}$, requires to be *positive-definite* i.e. $\mathbf{a}^{\mathsf{T}}\mathbf{H}\mathbf{a} > 0$ for any $\mathbf{a}$

Stationary Point

- Schoolboy calculus for single variable functions if second-derivatives at stationary point strictly positive, then point is minimum of function
- Multi-parameter function use generalisation of above rule
- Matrix of all partial second-derivatives, $\mathbf{H}$, requires to be *positive-definite* i.e. $\mathbf{a}^{\mathsf{T}}\mathbf{H}\mathbf{a} > 0$ for any $\mathbf{a}$
- Require expression for *Hessian* matrix

Stationary Point

- Can obtain matrix of second-partial derivatives of MSE

Stationary Point

UNIVERSITY *of* GLASGOW

- Can obtain matrix of second-partial derivatives of MSE

$$
\frac{\partial^2 MSE}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}} = \begin{bmatrix} \frac{\partial^2 MSE}{\partial w_0 \partial w_0} & \frac{\partial^2 MSE}{\partial w_0 \partial w_1} \\ \\ \frac{\partial^2 MSE}{\partial w_1 \partial w_0} & \frac{\partial^2 MSE}{\partial w_1 \partial w_1} \\ \end{bmatrix}
$$

$$
= \begin{bmatrix} 2 & \frac{2}{N} \sum_{n=1}^{N} x_n \\ \\ \frac{2}{N} \sum_{n=1}^{N} x_n & \frac{2}{N} \sum_{n=1}^{N} x_n^2 \\ \end{bmatrix}
$$

Stationary Point

UNIVERSITY *of* GLASGOW

- As will become usual in this course we can write the matrix of second-derivatives succinctly as

$$\frac{\partial^2 MSE}{\partial \mathbf{w} \partial \mathbf{w}^{\mathsf{T}}} = \frac{2}{N} \mathbf{X}^{\mathsf{T}} \mathbf{X}$$

Stationary Point

UNIVERSITY *of* GLASGOW

- As will become usual in this course we can write the matrix of second-derivatives succinctly as

$$\frac{\partial^2 MSE}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}} = \frac{2}{N} \mathbf{X}^\mathsf{T} \mathbf{X}$$

- If $\mathbf{X}^\mathsf{T}\mathbf{X}$ can be inverted it is positive definite

Stationary Point

- As will become usual in this course we can write the matrix of second-derivatives succinctly as

$$\frac{\partial^2 MSE}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}} = \frac{2}{N}\mathbf{X}^\mathsf{T}\mathbf{X}$$

- If $\mathbf{X}^\mathsf{T}\mathbf{X}$ can be inverted it is positive definite
- Providing $N \geq D$ then hessian is p.d. and can be inverted

Stationary Point

- As will become usual in this course we can write the matrix of second-derivatives succinctly as

$$\frac{\partial^2 MSE}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}} = \frac{2}{N} \mathbf{X}^\mathsf{T} \mathbf{X}$$

- If $\mathbf{X}^\mathsf{T}\mathbf{X}$ can be inverted it is positive definite
- Providing $N \geq D$ then hessian is p.d. and can be inverted
- So stationary point of MSE is indeed a minimum... phew..

Least Squares Solution

UNIVERSITY *of* GLASGOW

- As the matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is positive-definite it can be inverted and so we obtain the Least-Squares estimate $\widehat{\mathbf{w}}$

Least Squares Solution

UNIVERSITY *of* GLASGOW

- As the matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$ is positive-definite it can be inverted and so we obtain the Least-Squares estimate $\widehat{\mathbf{w}}$

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

Least Squares Solution

UNIVERSITY *of* GLASGOW

- As the matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$ is positive-definite it can be inverted and so we obtain the Least-Squares estimate $\widehat{\mathbf{w}}$

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

- The Least-Squares solution for Long-Jump Data is

$$\widehat{\mathbf{w}} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 276.78 \\ 0.748 \end{bmatrix}$$

Least Squares Solution

UNIVERSITY of GLASGOW

- As the matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$ is positive-definite it can be inverted and so we obtain the Least-Squares estimate $\widehat{\mathbf{w}}$

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

- The Least-Squares solution for Long-Jump Data is

$$\widehat{\mathbf{w}} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 276.78 \\ 0.748 \end{bmatrix}$$

- Can now employ this model to make predictions

Stationary Point

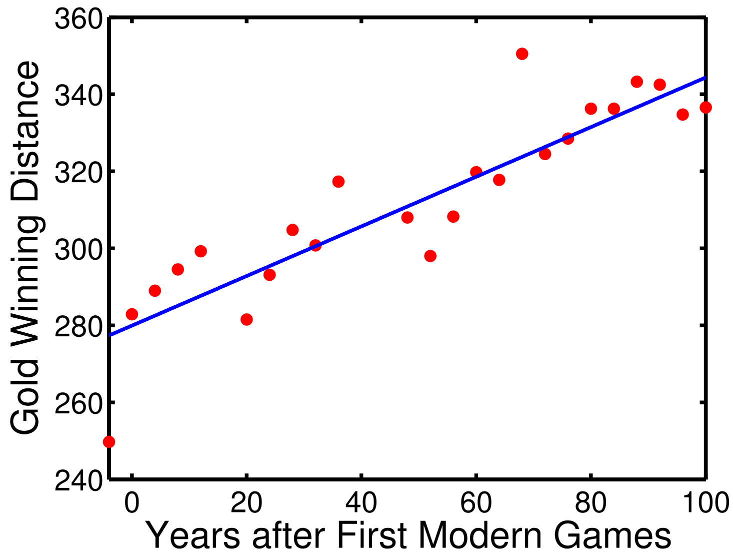
- With this parameter estimate our predictions for the given target values $\widehat{\mathbf{t}}$ follow as

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

Stationary Point

- With this parameter estimate our predictions for the given target values $\widehat{\mathbf{t}}$ follow as

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



Prediction

UNIVERSITY
of
GLASGOW

- What will be the winning distance at the London 2012 Olympic Games?

Prediction

UNIVERSITY
of
GLASGOW

- What will be the winning distance at the London 2012 Olympic Games?

$$\widehat{t}_{2012} = \mathbf{x}_{2012}^{\mathsf{T}}\widehat{\mathbf{w}} = [1 \;\; 112]\widehat{\mathbf{w}} = [1 \;\; 112]\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t}$$

Prediction

UNIVERSITY *of* GLASGOW

- What will be the winning distance at the London 2012 Olympic Games?

$$\widehat{t}_{2012} = \mathbf{x}_{2012}^{\mathsf{T}}\widehat{\mathbf{w}} = [1 \;\; 112]\widehat{\mathbf{w}} = [1 \;\; 112]\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t}$$

- Linear regression model predicts a gold medal winning distance of $276.78 + 0.748 \times 112 = 360.5$ inches in London.

Prediction

- What will be the winning distance at the London 2012 Olympic Games?

$$\widehat{t}_{2012} = \mathbf{x}_{2012}^{\mathsf{T}}\widehat{\mathbf{w}} = [1 \;\; 112]\widehat{\mathbf{w}} = [1 \;\; 112]\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t}$$

- Linear regression model predicts a gold medal winning distance of $276.78 + 0.748 \times 112 = 360.5$ inches in London.
- Current Olympic record stands at 350.39 inches and the current World Record was set in 1991 a distance of 352.36 inches.

Prediction

- What will be the winning distance at the London 2012 Olympic Games?

$$\widehat{t}_{2012} = \mathbf{x}_{2012}^{\mathsf{T}}\widehat{\mathbf{w}} = [1 \;\; 112]\widehat{\mathbf{w}} = [1 \;\; 112]\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t}$$

- Linear regression model predicts a gold medal winning distance of $276.78 + 0.748 \times 112 = 360.5$ inches in London.
- Current Olympic record stands at 350.39 inches and the current World Record was set in 1991 a distance of 352.36 inches.
- Our prediction seems somewhat optimistic!!!!

Nonlinear Model

UNIVERSITY *of* GLASGOW

- Model is linear in parameters

Nonlinear Model

- Model is linear in parameters
- Can apply nonlinear transformation to inputs providing more flexible model

Nonlinear Model

UNIVERSITY of GLASGOW

- Model is linear in parameters
- Can apply nonlinear transformation to inputs providing more flexible model
- But still linear in parameters - provided no additional parameters associated with transform

Nonlinear Model

UNIVERSITY of GLASGOW

- Model is linear in parameters
- Can apply nonlinear transformation to inputs providing more flexible model
- But still linear in parameters - provided no additional parameters associated with transform
- For example if a cubic polynomial assumed

$$f(x; \mathbf{w}) = w_3 x^3 + w_2 x^2 + w_1 x + w_0$$

or more generally an arbitrary K'th order polynomial holds

$$f(x; \mathbf{w}) = \sum_{i=0}^{K} w_i x^i$$

Nonlinear Model

UNIVERSITY *of* GLASGOW

- It should be straightforward to see that by now defining the $N \times (K+1)$ dimensional matrix $\mathbf{X}$ such that

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^K \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^K \end{bmatrix}$$

Nonlinear Model

UNIVERSITY *of* GLASGOW

- It should be straightforward to see that by now defining the $N \times (K+1)$ dimensional matrix $\mathbf{X}$ such that

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^K \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^K \end{bmatrix}$$

- Least Squares solution still holds where now $\widehat{\mathbf{w}}$ will be a $(K+1) \times 1$ column vector

Nonlinear Model

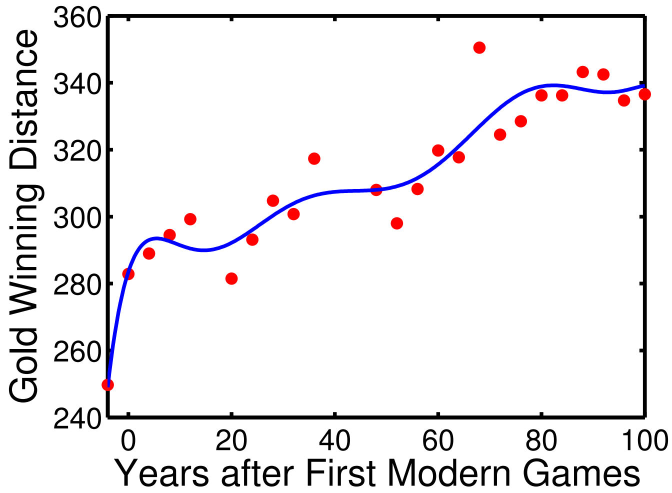
UNIVERSITY of GLASGOW

- Nonlinear Model (Linear regression model!!) of order $K = 9$

Nonlinear Model

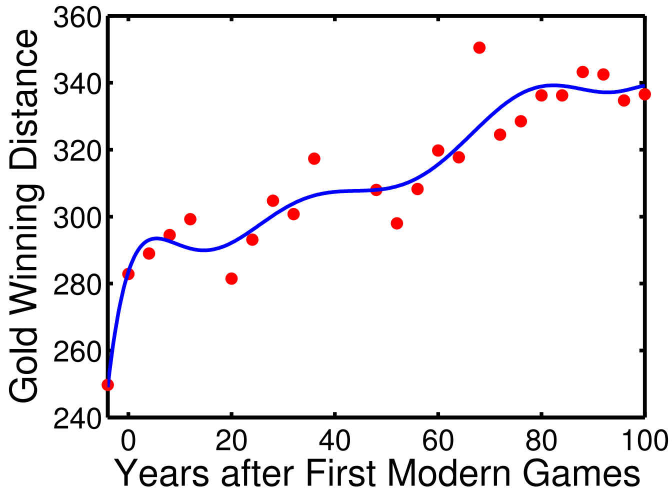
UNIVERSITY *of* GLASGOW

- Nonlinear Model (Linear regression model!!) of order $K = 9$



Nonlinear Model

- Nonlinear Model (Linear regression model!!) of order $K = 9$



- Is this a better model??... Stay tuned till next week