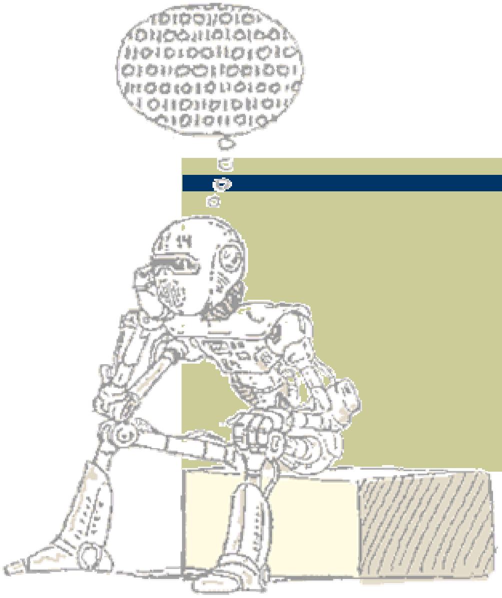


Lecture VI– Regression
(Linear Methods for Regression)

Contents:

- Linear Methods for Regression
    - Least Squares, Gauss Markov Theorem
    - Recursive Least Squares

Linear Regression Model

$$y = f(\mathbf{x}) = w_0 + \sum_{j=1}^{M} x_j w_j = \mathbf{x}^T \mathbf{w} + \varepsilon : \; Linear \;\; Model$$

$$where \;\; \mathbf{x} = (x_1, x_2, \ldots, x_m, 1)^T \equiv Input\, vector,$$

$$\mathbf{w} = (w_1, w_2, \ldots, w_m, w_0)^T \equiv \; regression \; parameters$$

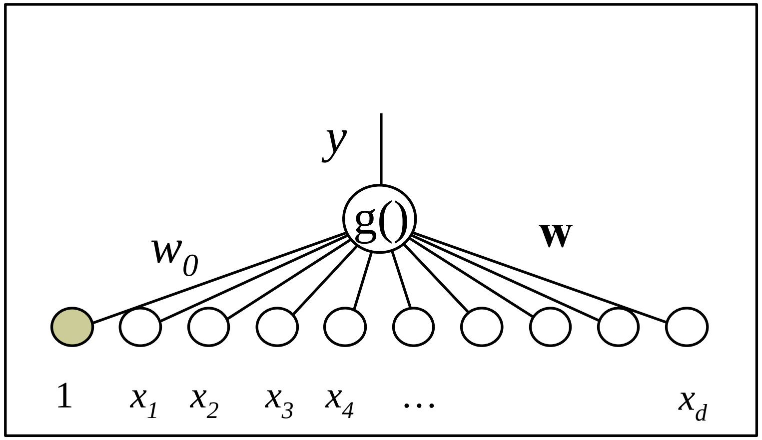
The ***linear model*** either assumes that the regression function f(x) is linear , or that the linear model is a reasonable approximation.

The inputs **x** can be :

- Quantitative inputs
- Transformations of quantitative inputs such as log, square root etc.
- Basis expansions (e.g. polynomial representation) : $x_2 = x_1^2,\; x_3 = x_1^3, \ldots$
- Interaction between variables : $x_3 = x_1 \cdot x_2$
- Dummy coding of levels of qualitative input

In all these cases, the model is ***linear in the parameters***, even though the final function itself may not be linear.

Power of Linear Models



$$y(\mathbf{x}) = f(\mathbf{x},\mathbf{w}) = g\left(\mathbf{w}^T\mathbf{x} + w_0\right)$$

- if g() is linear: only linear functions can be modeled
- however, if **x** is actually preprocessed, complicated functions can be realized

$$\mathbf{x} = \Phi(\mathbf{z}) = \begin{bmatrix} \phi_1(\mathbf{z}) \\ \phi_2(\mathbf{z}) \\ \cdots \\ \phi_d(\mathbf{z}) \end{bmatrix} \quad \text{example:} \quad \mathbf{x} = \Phi(\mathbf{z}) = \begin{bmatrix} z \\ z^2 \\ \cdots \\ z^d \end{bmatrix}$$

Least Squares Optimization

- **Least Squares Cost Function**

$$J = \frac{1}{2}\sum_{i=1}^{N}(t_i - \hat{f}(x_i))^2 \qquad where\ \ N = \#\,of\ \ training\ \ data$$

$$= \frac{1}{2}\sum_{i=1}^{N}(t_i - \mathbf{x}_i^T\mathbf{w})^2 = \frac{1}{2}(\mathbf{t}-\mathbf{X}\mathbf{w})^T(\mathbf{t}-\mathbf{X}\mathbf{w}), \qquad \text{where}\ \ \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \dots \\ t_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \dots \\ \mathbf{x}_n^T \end{bmatrix}$$

- **Minimize Cost**

$$\frac{\partial J}{\partial \mathbf{w}} = 0 = \frac{\partial J}{\partial \mathbf{w}}\left(\frac{1}{2}(\mathbf{t}-\mathbf{X}\mathbf{w})^T(\mathbf{t}-\mathbf{X}\mathbf{w})\right) = -(\mathbf{t}-\mathbf{X}\mathbf{w})^T\mathbf{X}$$

$$= -\mathbf{t}^T\mathbf{X} + (\mathbf{X}\mathbf{w})^T\mathbf{X} = -\mathbf{t}^T\mathbf{X} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}$$

$$\Rightarrow \mathbf{t}^T\mathbf{X} = \mathbf{w}^T\mathbf{X}^T\mathbf{X}$$

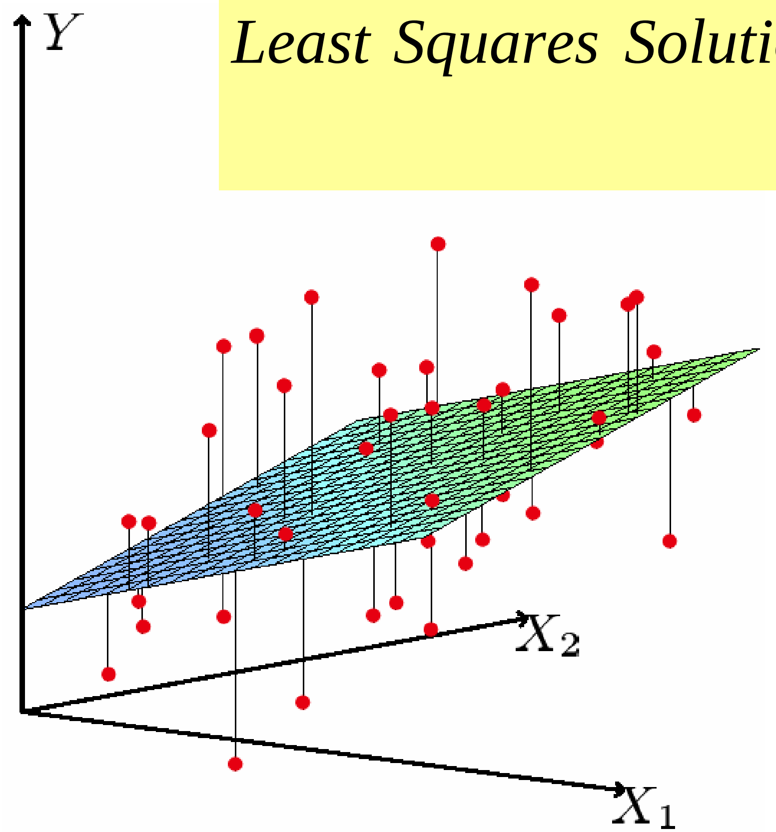
$$\Rightarrow \mathbf{X}^T\mathbf{t} = \mathbf{X}^T\mathbf{X}\mathbf{w}$$

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

What are we really doing ?

$$\textit{Least Squares Solution}:\ \mathbf{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{t}$$

$$y_{pred} = \mathbf{x}^T_{pred}\,\mathbf{w}$$



Linear least squares fitting

We seek the linear function of X that minimizes the sum of the squared residuals from Y

More insights into the LS solution

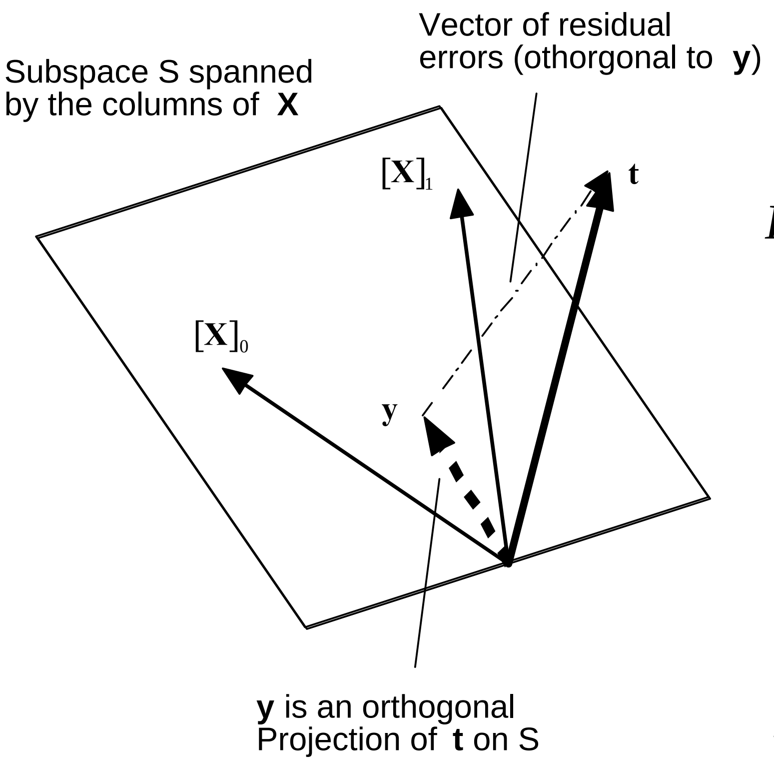
- The Pseudo-Inverse $\mathbf{X}^{+} = \left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T}$
  - pseudo inverses are a special solution to an infinite set of solutions of a non-unique inverse problem (we talked about it in the previous lecture)
  - the matrix inversion above may still be ill-defined if $\mathbf{X}^{T}\mathbf{X}$ is close to singular and so-called ***Ridge Regression*** needs to be applied

- Ridge Regression $\mathbf{X}^{+} = \left(\mathbf{X}^{T}\mathbf{X} + \gamma\mathbf{I}\right)^{-1}\mathbf{X}^{T}$ where $\gamma << 1$

- Multiple Outputs: just like multiple single output regressions

$$\mathbf{W} = \left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T}\mathbf{Y}$$

Geometrical Interpretation of LS



$$Residual \;\; vector: \; \mathbf{t}-\mathbf{y}=\mathbf{t}-\mathbf{X}\mathbf{w}=\mathbf{t}-\sum_i [\mathbf{X}]_i w_i$$

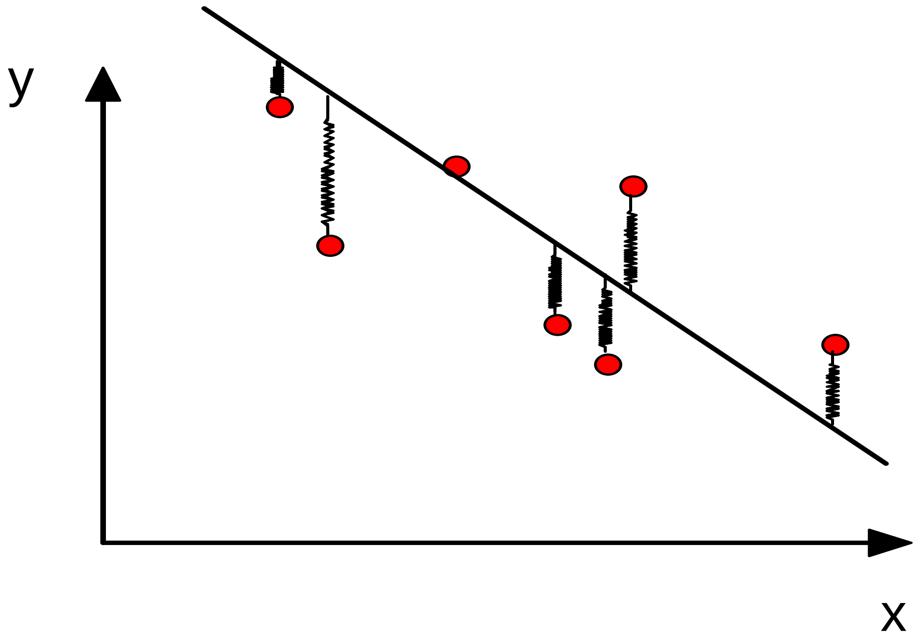
… is orthogonal to the space spanned by columns of X since …

$$\frac{\partial J}{\partial \mathbf{w}} = 0 = -(\mathbf{t}-\mathbf{X}\mathbf{w})^T \mathbf{X}$$

And hence, …

y **is the optimal reconstruction of** t **in the range of** X

Physical Interpretation of LS



- all springs have the same spring constant
- points far away generate more “force” (danger of outliers)
- springs are vertical
- solution is the minimum energy solution achieved by the springs

Minimum variance unbiased estimator

Gauss-Markov Theorem

Least Squares estimate of the parameters **w** has the smallest variance among all ***linear unbiased*** estimates.

Least Squares are also called BLUE estimates – **B**est **L**inear **U**nbiased **E**stimators

$$\begin{aligned}\hat{\mathbf{w}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t} \quad : \textit{Least Squares Estimate} \\ &= \mathbf{Ht} \quad \textit{where } \mathbf{H} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\end{aligned}$$

In other words, Gauss-Markov theorem says that there is no other matrix **C** such that the estimator formed by

$\tilde{\mathbf{w}} = \mathbf{Ct}$ will be both unbiased and have a smaller variance than $\hat{\mathbf{w}}$.

$\hat{\mathbf{w}}$ (*Least Squares Estimate*) is an Unbiased Estimate since $E(\hat{\mathbf{w}}) = \mathbf{w}$

(Homework !!)

Gauss-Markov Theorem (Proof)

$$\begin{aligned} E(\tilde{\mathbf{w}}) &= E(\mathbf{Ct}) \\ &= E(\mathbf{C}(\mathbf{Xw}+\boldsymbol{\varepsilon})) \\ &= E(\mathbf{CXw}+\mathbf{C}\boldsymbol{\varepsilon}) \\ &= \mathbf{CXw}+\mathbf{C}E(\boldsymbol{\varepsilon}) \\ &= \mathbf{CXw} \end{aligned}$$

$$\textit{For Unbiased Estimate}: E(\tilde{\mathbf{w}}) = \mathbf{w}$$
$$\Rightarrow \mathbf{CXw} = \mathbf{w} \Rightarrow \mathbf{CX} = \mathbf{I}$$

$$\begin{aligned} Var(\tilde{\mathbf{w}}) &= E[(\tilde{\mathbf{w}} - E(\tilde{\mathbf{w}}))(\tilde{\mathbf{w}} - E(\tilde{\mathbf{w}}))^T] \\ &= E[(\tilde{\mathbf{w}} - \mathbf{w})(\tilde{\mathbf{w}} - \mathbf{w})^T] \\ &= E[(\mathbf{Ct} - \mathbf{w})(\mathbf{Ct} - \mathbf{w})^T] \\ &= E[(\mathbf{CXw} + \mathbf{C}\boldsymbol{\varepsilon} - \mathbf{w})(\mathbf{CXw} + \mathbf{C}\boldsymbol{\varepsilon} - \mathbf{w})^T] \\ &= E[(\mathbf{C}\boldsymbol{\varepsilon})(\mathbf{C}\boldsymbol{\varepsilon})^T] \quad \textit{...since } \mathbf{CX} = \mathbf{I} \\ &= \mathbf{C}E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T]\mathbf{C}^T \\ &= \sigma^2 \mathbf{CC}^T \end{aligned}$$

Gauss-Markov Theorem (Proof)

We want to show that $Var(\hat{\mathbf{w}}) \leq Var(\tilde{\mathbf{w}})$

$$Let\ \mathbf{C} = \mathbf{D} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

$$(\mathbf{D} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{X} = \mathbf{I} \quad since\ \ \mathbf{CX} = \mathbf{I} \Rightarrow \mathbf{DX} + \mathbf{I} = \mathbf{I} \Rightarrow \mathbf{DX} = \mathbf{0}$$

$$\begin{aligned}
Var(\tilde{\mathbf{w}}) &= \sigma^2\mathbf{CC}^T = \sigma^2(\mathbf{D} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(\mathbf{D} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\
&= \sigma^2(\mathbf{DD}^T + (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1} + 2\mathbf{DX}(\mathbf{X}^T\mathbf{X})^{-1}) \\
&= \sigma^2\mathbf{DD}^T + \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \qquad ....since\ \mathbf{DX} = 0 \\
Var(\tilde{\mathbf{w}}) &= \sigma^2\mathbf{DD}^T + Var(\hat{\mathbf{w}})
\end{aligned}$$

It is sufficient to show that diagonal elements of $\sigma^2\mathbf{DD}^T$ *are non – negative.*
This is true by definition. Hence, proved.

Biased vs unbiased

Bias-Variance decomposition of error

$$E\left\{\hat{f}(\mathbf{x}_i)\right\} = \sigma_\varepsilon^2 + \left(E\{\hat{y}_i\} - f(\mathbf{x}_i)\right)^2 + E\left\{\left(\hat{y}_i - E\{\hat{y}_i\}\right)^2\right\}$$
$$= \mathrm{var}(noise) + bias^2 + \mathrm{var}(estimate)$$

Gauss-Markov Theorem says that Least Squares achieves the estimate with the ***minimum variance*** (and hence, the minimum Mean Squared Error) among all the ***unbiased estimates*** (bias=0).

Does that mean that we should always work with unbiased estimators ??

No !! since there may exists some biased estimators with a smaller net mean squared error – ***they trade a little bias for a larger reduction in variance***.

Variable Subset Selection and Shrinkage are methods (which we will explore soon) that introduce bias and try to reduce the variance of the estimate.

Recursive Least Squares

- **The Sherman-Morrison-Woodbury Theorem**

$$\left(\mathbf{A} - \mathbf{z}\mathbf{z}^T\right)^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}\mathbf{z}\mathbf{z}^T\mathbf{A}^{-1}}{1 - \mathbf{z}^T\mathbf{A}^{-1}\mathbf{z}}$$

- **More General: The Matrix Inversion Theorem**

$$(\mathbf{A} - \mathbf{BC})^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{I} + \mathbf{CA}^{-1}\mathbf{B}\right)^{-1}\mathbf{CA}^{-1}$$

- **Recursive Least Squares Update**

$$\text{Initialize}: \mathbf{P}^n = \mathbf{I}\frac{1}{\gamma} \text{ where } \gamma << 1 \text{ (note } \mathbf{P} \equiv \left(\mathbf{X}^T\mathbf{X}\right)^{-1})$$

For every new data point $(\mathbf{x}, \mathbf{t})$ (note that $\mathbf{x}$ includes the bias term) :

$$\mathbf{P}^{n+1} = \frac{1}{\lambda}\left(\mathbf{P}^n - \frac{\mathbf{P}^n\mathbf{x}\mathbf{x}^T\mathbf{P}^n}{\lambda + \mathbf{x}^T\mathbf{P}^n\mathbf{x}}\right) \text{ where } \lambda = \begin{cases} 1 & \text{if no forgetting} \\ <1 & \text{if forgetting} \end{cases}$$

$$\mathbf{w}^{n+1} = \mathbf{W}^n + \mathbf{P}^{n+1}\mathbf{x}\left(\mathbf{t} - \mathbf{w}^{nT}\mathbf{x}\right)^T$$

Recursive Least Squares (cont'd)

- **Some amazing facts about recursive least squares**
  - Results for W are EXACTLY the same as for normal least squares update (batch update) after every data point was added once! (no iterations)
  - NO matrix inversion necessary anymore
  - NO learning rate necessary
  - Guaranteed convergence to optimal **W** (linear regression is an optimal estimator under many conditions)
  - Forgetting factor λ allows to forget data in case of changing target functions
  - Computational load is larger than batch version of linear regression
  - But don't get fooled: if data is singular, you still will have problems!