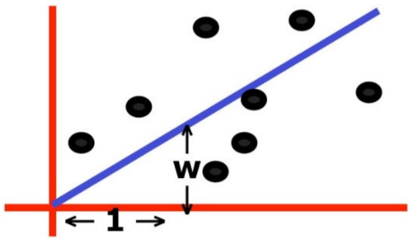
Introduction to Regression

Hee-il Hahn

Professor
Department of Information and Communications Engineering
Hankuk University of Foreign Studies
hihahn@hufs.ac.kr

Linear Regression

- 지도학습의 한 분야로 연속적인 숫자(실수)를 예측하는 것
  - 어떤 사람의 교육수준, 나이, 주거지를 바탕으로 연간소득 예측하는 문제
  - 측정된 점들의 열로부터 가장 근사한 방정식을 구하는 문제



inputs	outputs
$x_1 = 1$	$y_1 = 1$
$x_2 = 3$	$y_2 = 2.2$
$x_3 = 2$	$y_3 = 2$
$x_4 = 1.5$	$y_4 = 1.9$
$x_5 = 4$	$y_5 = 3.1$

- Linear regression assumes that the expected value of the output given an input, $E[Y|X]$, is linear.

  i.e., $E[Y|X] = \alpha + \beta X$ or $E[Y|X] = \alpha + \beta_1 X_1 + \cdots + \beta_p X_p$ where $E[Y|X] = \int yf(y/x)dy$

  - Simplest case: $Out(x) = wx$ for some unknown w.
  - Given the data, we can estimate w.

2-parameter linear regression

Observable dataset : $\mathbf{d}_1(x_1, y_1), \mathbf{d}_2(x_2, y_2) \ldots \mathbf{d}_n(x_n, y_n)$

Model : $y = wx + b$

Compute mean squared error of the model on the dataset

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - wx_i - b)^2$$

To minimize MSE

$$\begin{cases} \frac{\partial}{\partial w} MSE = \frac{\partial}{\partial w}\frac{1}{n}\sum_{i=1}^{n}(y_i - wx_i - b)^2 = 0 & \rightarrow \sum_{i=1}^{n} x_i\,(y_i - wx_i - b) = 0 \\ \frac{\partial}{\partial b} MSE = \frac{\partial}{\partial b}\frac{1}{n}\sum_{i=1}^{n}(y_i - wx_i - b)^2 = 0 & \rightarrow \sum_{i=1}^{n}(y_i - wx_i - b) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} w\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ w\sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i \end{cases}$$

if $b = 0$, $w = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ (즉, $y = wx$ 로 모델링하면)

Bayesian linear regression

- Assume that the data is formed by $y_i = wx_i + noise_i$
  - the noise signals are independent
  - the noise has a normal distribution with mean 0 and unknown variance σ^2
  - p(y|w,x) has a normal distribution with mean wx and variance σ^2
- $y \sim N(wx, \sigma^2)$
- We have a set of data $\mathbf{d}_1(x_1, y_1)$,$\mathbf{d}_2(x_2, y_2)\ldots\mathbf{d}_n(x_n, y_n)$.
- We want to infer w from the data.

$$P(w|\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n) = \mathrm{P}(w|\boldsymbol{D})$$

- We can use BAYES rule to work out a posterior distribution for w given the data.
- Or, we could do Maximum Likelihood Estimation.

Maximum likelihood estimation of w

- Choose the parameter w that maximizes the probability of the data, given that parameter.
- MLE asks: "For which value of w is this data most likely to have happened?"

For what w, is $P(\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n | w)$ maximized?

$\equiv$ For what w, is $\prod_{i=1}^{n} P(\mathbf{d}_i | w)$ maximized?

$\equiv$ For what w, is $\prod_{i=1}^{n} exp\left(-\frac{1}{2}\left(\frac{y_i - wx_i}{\sigma}\right)^2\right)$ maximized?

$\equiv$ For what w, is $\sum_{i=1}^{n} (y_i - wx_i)^2$ minimized?

where $P(\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n | w)$ is called the Likelihood, and

$$P(\mathbf{d}_i | w) = P(y_i | w, x_i) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2}\left(\frac{y_i - wx_i}{\sigma}\right)^2\right).$$

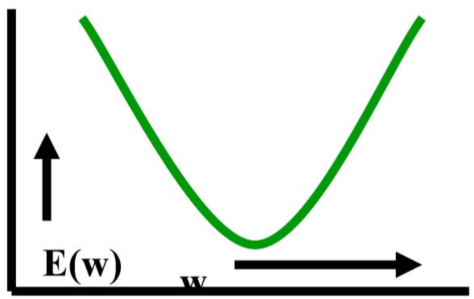
First result

- MLE with Gaussian noise is the same as minimizing the L_2 error

$$\text{argmin}_w \sum_{i=1}^{n} (y_i - wx_i)^2$$

The maximum likelihood w is the one that minimizes sum-of-squares of residuals

$$E = \sum_{i=1}^{n} (y_i - wx_i)^2$$
$$= \sum_{i=1}^{n} y_i^2 - (2 \sum x_i y_i) w + \left(\sum x_i^2\right) w^2$$



We want to minimize a quadratic function of w.

Linear regression

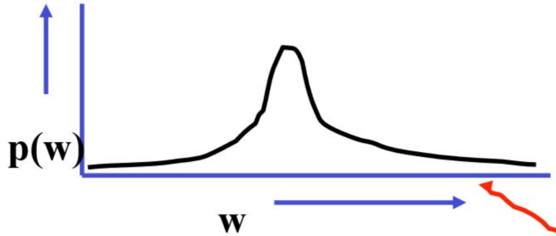
- Easy to show the sum of squares is minimized when

$$w = \frac{\sum x_i y_i}{\sum x_i^2}$$

The maximum likelihood model is:

$$Out(x) = wx$$

We can use it for prediction.



Note: In Bayesian stats you'd have ended up with a prob distribution of w

And predictions would have given a prob disribution of expected output

Often useful to know your confidence. Max likelihood can give some kinds of confidence too.

Maximum a Posteriori estimation of w

- MAP
  - Choose w that maximizes the posteriori probability of w.
  - Posterior probability of w is given by the Bayes Rule:

$$P(w|\boldsymbol{D}) = \frac{P(w)P(\boldsymbol{D}|w)}{P(\boldsymbol{D})}$$

where $P(w)$: Prior probability of w assumed as $w \sim N(0, \gamma^2)$

$P(\boldsymbol{D})$: Probability of data (independent of w)

$$P(\boldsymbol{D}) = \int P(w)P(\boldsymbol{D}|w)dw$$

Maximum a Posteriori estimation - cont'd

- MAP

$$\hat{w}_{MAP} = \text{argmax}_w P(w|\boldsymbol{D})$$

$$= \text{argmax}_w \frac{P(w)P(\boldsymbol{D}|w)}{P(\boldsymbol{D})}$$

$$\cong \text{argmax}_w \; P(w)P(\boldsymbol{D}|w)$$

$$= \text{argmax}_w \prod_{i=1}^{n} P(\mathbf{d}_i|w)\, P(w)$$

$$= \text{argmax}_w \sum_{i=1}^{n} logP(\mathbf{d}_i|w) + logP(w)$$

$$(cf\!: \quad \hat{w}_{MLE} = \text{argmax}_w P(\boldsymbol{D}|w)$$

$$= \text{argmax}_w \prod_{i=1}^{n} P(\mathbf{d}_i|w))$$

Maximum a Posteriori estimation - cont'd

For what w, is $\prod_{i=1}^{n} P(\mathbf{d}_i|w\ )P(w)$ maximized?

$\equiv$

For what w, is $\prod_{i=1}^{n} exp\left(-\frac{1}{2}\left(\frac{y_i-wx_i}{\sigma}\right)^2\right) exp\left(-\frac{1}{2}\left(\frac{w}{\gamma}\right)^2\right)$maximized?

$\equiv$

For what w, is $\sum_{i=1}^{n} -\frac{1}{2}\left(\frac{y_i-wx_i}{\sigma}\right)^2 - \frac{1}{2}\left(\frac{w}{\gamma}\right)^2$ maximized?

$\equiv$

For what w, is $\sum_{i=1}^{n}(y_i - wx_i)^2 + \left(\frac{\sigma w}{\gamma}\right)^2$ minimized?

Second result

- MAP with a Gaussian prior on w is the same as minimizing the L_2 error plus an L_2 penalty on w

$$\text{argmin}_w \sum_{i=1}^{n} (y_i - wx_i)^2 + \rho w^2$$

$$\rho = \frac{\sigma}{\gamma}$$

  - MLE estimation of a parameter leads to unregularized solutions.
  - MAP estimation of a parameter leads to regularized solutions.
  - The prior distribution $P(w)$ acts as a regularizer in MAP estimation.

Multivariate regression

- What if the inputs are vectors?

  Write matrix $\boldsymbol{X}$ and $\mathbf{y}$:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix}$$

  where $\mathbf{x}_1 = (x_{11}, \cdots, x_{1p}\,), \mathbf{x}_2 = (x_{21}, \cdots, x_{1p}\,), \cdots$

  $\mathbf{x}_n = (x_{n1}, \cdots, x_{np}\,)$

- Assume that the data is formed by $y_i = \boldsymbol{w}^T \mathbf{x}_i + noise_i$

$$y \sim N(\boldsymbol{w}^T \mathbf{x}, \sigma^2)$$

Multivariate regression - cont'd

- Probability of each response variable

$$P(\mathbf{d}_i|\mathbf{w}) = P(y_i|\mathbf{w},\mathbf{x}_i) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2}\left(\frac{y_i-\mathbf{w}^T\mathbf{x}_i}{\sigma}\right)^2\right).$$

- Given data $\boldsymbol{D} = \{\mathbf{d}_1(\mathbf{x}_1, y_1), \cdots, \mathbf{d}_n(\mathbf{x}_n, y_n)\}$, we want to estimate the weight vector $\mathbf{w}$.

Likelihood:

$$L(\mathbf{w}) = P(\boldsymbol{D}|\mathbf{w}) = P(y|\mathbf{w},\mathbf{X}) = \prod_{i=1}^{n} P(\mathbf{d}_i|\mathbf{w})$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2}\left(\frac{y_i-\mathbf{w}^T\mathbf{x}_i}{\sigma}\right)^2\right)$$

Log-likelihood:

$$logL(\mathbf{w}) = \sum_{i=1}^{n}\left\{-\frac{1}{2}log(2\pi\sigma^2) - \frac{1}{2}\left(\frac{y_i-\mathbf{w}^T\mathbf{x}_i}{\sigma}\right)^2\right\}$$

Multivariate regression - cont'd

- Maximum likelihood solution:

$$\hat{\mathbf{w}}_{MLE} = \text{argmax}_{\mathbf{w}} \prod_{i=1}^{n} P(\mathbf{d}_i|\mathbf{w})$$

$$= \text{argmax}_{\mathbf{w}} \sum_{i=1}^{n} -\frac{1}{2}\left(\frac{y_i-\mathbf{w}^T\mathbf{x}_i}{\sigma}\right)^2$$

$$= \text{argmin}_{\mathbf{w}} \sum_{i=1}^{n} \left(\frac{y_i-\mathbf{w}^T\mathbf{x}_i}{\sigma}\right)^2$$

$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{y}$$

(from $\frac{d}{d\mathbf{w}}(\boldsymbol{y}-\boldsymbol{X}\mathbf{w})^T(\boldsymbol{y}-\boldsymbol{X}\mathbf{w}) = \mathbf{0}$)

Multivariate regression - cont'd

- Maximum-a-Posteriori Solution:

  - Assume a Gaussian prior distribution over the weight vector $\mathbf{w}$.

$$P(\mathbf{w}) \sim N(0, \lambda^{-1}\boldsymbol{I}\,) = \frac{1}{(2\pi)^{p/2}} exp\left(-\frac{\lambda}{2}\mathbf{w}^T\mathbf{w}\right)$$

  - Posteriori probability:

$$P(\mathbf{w}|\boldsymbol{D}\,) = \frac{P(\mathbf{w})P(\boldsymbol{D}|\mathbf{w}\,)}{P(\boldsymbol{D})}$$

  - Log Posteriori probability:

$$logP(\mathbf{w}|\boldsymbol{D}\,) = log\frac{P(\mathbf{w})P(\boldsymbol{D}|\mathbf{w}\,)}{P(\boldsymbol{D})}$$

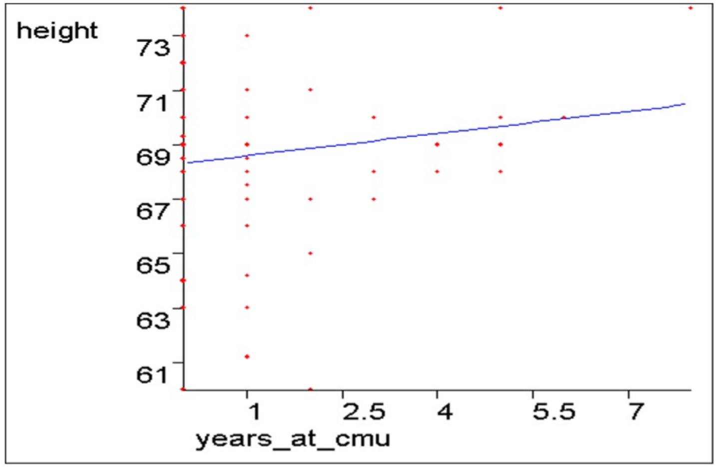
Multivariate regression - cont'd

- Maximum-a-Posteriori Solution:

$$
\begin{aligned}
\hat{\mathbf{w}}_{MAP} &= \text{argmax}_{\mathbf{w}}\, logP(\mathbf{w}|\boldsymbol{D}\,) \\
&= \text{argmax}_{\mathbf{w}}\, \{logP(\boldsymbol{D}|\mathbf{w}) + logP(\mathbf{w})\} \\
&= \text{argmax}_{\mathbf{w}} \{logP(\mathbf{w}) + \textstyle\sum_{i=1}^{n} logP(\mathbf{d}_i|\mathbf{w})\} \\
&= \text{argmax}_{\mathbf{w}} \left\{-\frac{\lambda}{2}\mathbf{w}^T\mathbf{w} - \textstyle\sum_{i=1}^{n} \frac{1}{2}\left(\frac{y_i - \mathbf{w}^T\mathbf{x}_i}{\sigma}\right)^2\right\} \\
&= \text{argmin}_{\mathbf{w}} \left\{\frac{\lambda}{2}\mathbf{w}^T\mathbf{w} + \textstyle\sum_{i=1}^{n} \frac{1}{2}\left(\frac{y_i - \mathbf{w}^T\mathbf{x}_i}{\sigma}\right)^2\right\} \\
&= \text{argmin}_{\mathbf{w}} \left\{\frac{\lambda}{2}\mathbf{w}^T\mathbf{w} + \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\mathbf{w})^T(\boldsymbol{y} - \boldsymbol{X}\mathbf{w})\right\} \\
&= \left(\boldsymbol{X}^T\boldsymbol{X} + \frac{\sigma^2}{2}\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\mathbf{y}
\end{aligned}
$$

Constant term in linear regression

- We may expect linear data that does not go through the origin.
- Statisticians and Neural Net Folks all agree on a simple obvious hack. Can you guess??



The constant term

- The trick is to create a fake input "X_0" that always takes the value 1 .

X_1	X_2	Y
2	4	16
3	4	17
5	5	20

Before:

$$Y = w_1 X_1 + w_2 X_2$$

" Poor model "

X_0	X_1	X_2	Y
1	2	4	16
1	3	4	17
1	5	5	20

After:

$$Y = w_0 X_0 + w_1 X_1 + w_2 X_2$$

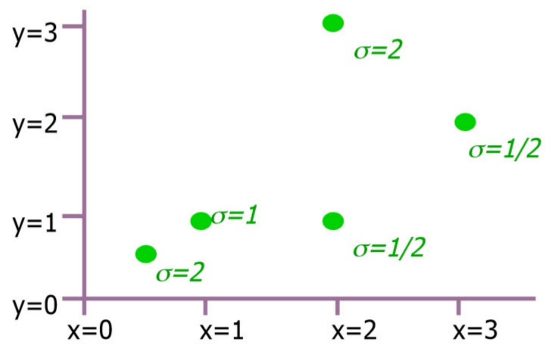
" has a fine constant term "

you Should be able to see the MLE w_0, w_1, w_2 by inspection.

Linear regression with varying noise

- Suppose you know the variance of the noise that was added to each data point.

x_i	y_i	σ_i^2
½	½	4
1	1	1
2	1	1/4
2	3	4
3	2	1/4



Assume $y_i \sim N(wx_i, \sigma_i^2)$

What is the MLE estimate of w?

MLE estimation with varying noise

$$\operatorname{argmax}_w \log P(\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n | w, \sigma_1^2, \dots, \sigma_n^2)$$

$$= \operatorname{argmin}_w \sum_{i=1}^{n} \frac{(y_i - wx_i)^2}{\sigma_i^2}$$

$$\rightarrow w = \frac{\sum_{i=1}^{n} \frac{x_i y_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{x_i{}^2}{\sigma_i^2}}$$

Nonlinear regression

- Suppose you know that y is related to a function of x in such a way that the predicted values have a non-linear dependence on w, e.g. :

Assume $y_i \sim N(\sqrt{w + x_i}, \sigma^2)$

What is the MLE estimate of w?

Nonlinear regression - cont'd

$$\mathrm{argmax}_w \, logP(\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n | w, \sigma_1^2, \dots, \sigma_n^2)$$

$$= \mathrm{argmin}_w \sum_{i=1}^{n} \frac{(y_i - \sqrt{w + x_i})^2}{\sigma_i^2}$$

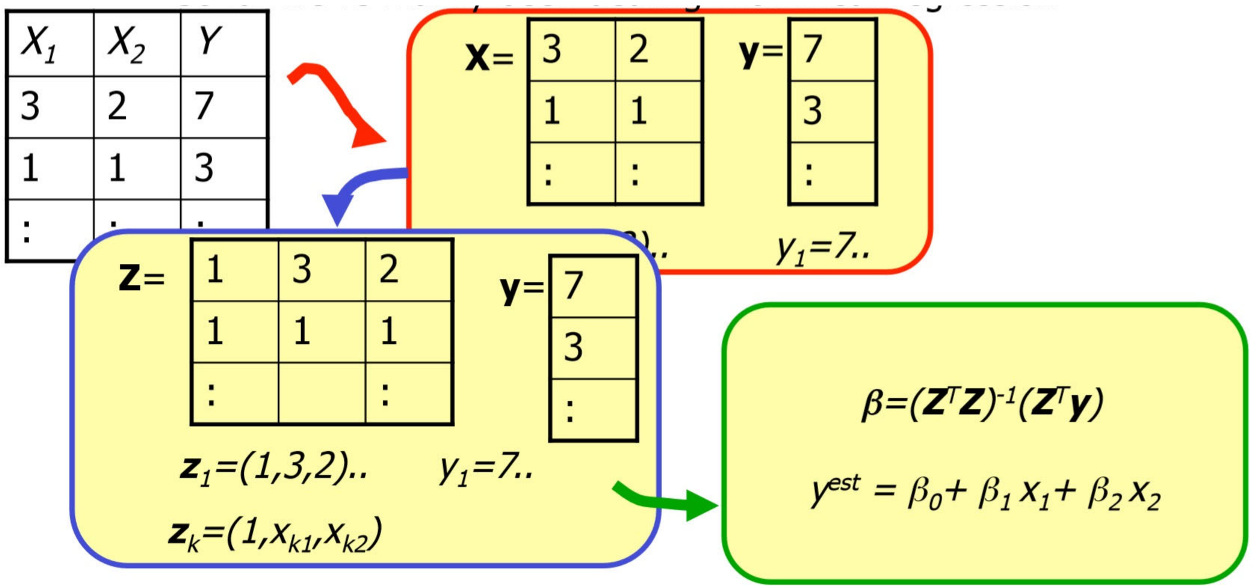
$$\rightarrow w \quad such\ that \ \sum_{i=1}^{n} \frac{y_i - \sqrt{w + x_i}}{\sigma_i^2 \sqrt{w + x_i}} = 0$$

Nonlinear regression - cont'd

- Common (but not only) approach:
- Numerical Solutions:
  - Line Search
  - Simulated Annealing
  - Gradient Descent
  - Conjugate Gradient
  - Levenberg Marquart
  - Newton's Method
  - Also, special purpose statistical-optimization-specific tricks such as E.M.

Polynomial regression

- So far we've mainly been dealing with linear regression



Quadratic regression

- It’s trivial to do linear fits of fixed nonlinear basis functions.

X_1	X_2	Y
3	2	7
1	1	3
:	:	:

$$\mathbf{X}=\begin{bmatrix} 3 & 2 \\ 1 & 1 \\ : & : \end{bmatrix} \qquad \mathbf{y}=\begin{bmatrix} 7 \\ 3 \\ : \end{bmatrix}$$

$y_1=7..$

$$\mathbf{Z}=\begin{bmatrix} 1 & 3 & 2 & 9 & 6 & 4 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ : & & & & & : \end{bmatrix} \qquad \mathbf{y}=\begin{bmatrix} 7 \\ 3 \\ : \end{bmatrix}$$

$$\mathbf{z}=(1,\ x_1,\ x_2,\ x_1^2,\ x_1x_2, x_2^2)$$

$$\beta=(\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{y})$$

$$y^{est} = \beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1^2+\beta_4 x_1x_2+\beta_5 x_2^2$$

Quadratic regression - cont'd

Each component of a z vector is called a term.

Each column of the Z matrix is called a term column

How many terms in a quadratic regression with *m* inputs?

•1 constant term

•m linear terms

•(m+1)-choose-2 = m(m+1)/2 quadratic terms

(m+2)-choose-2 terms in total = $O(m^2)$

Note that solving $\beta=(\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{y})$ is thus $O(m^6)$

Q^{th}-degree polynomial regression

X_1	X_2	Y
3	2	7
1	1	3
:	:	:

$$\mathbf{X} = \begin{bmatrix} 3 & 2 \\ 1 & 1 \\ \vdots & \vdots \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 7 \\ 3 \\ \vdots \end{bmatrix}$$

$$\mathbf{Z} = \begin{bmatrix} 1 & 3 & 2 & 9 & 6 & \ldots \\ 1 & 1 & 1 & 1 & 1 & \ldots \\ \vdots & & & & & \ldots \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 7 \\ 3 \\ \vdots \end{bmatrix}$$

z=(all products of powers of inputs in which sum of powers is q or less)

$$\beta = (\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{y})$$

$$y^{est} = \beta_0 + \beta_1 x_1 + \ldots$$