CS724: Topics in Algorithms
Least Square Approximations
Slide Set 11

Prof. Dan A. Simovici

UMASS BOSTON

1. Linear Regression

2. The Least Square Approximation and QR Decomposition

The least square method is used in data mining as a method of estimating the parameters of a model by adopting the values that minimize the sum of the squared differences between the predicted and the observed values of data.
This estimation process is also known as *regression*, and several types of regression exist depending on the nature of the assumed model of dependency between the predicted and the observed data.

UMASS BOSTON

The aim of *linear regression* is to explore the existence of a linear relationship between the outcome of an experiment and values of variables that are measured during the experiment.
Experimental data often is presented as a data sample matrix $X \in \mathbb{R}^{m \times n}$, where m is the number of experiments and n is the number of variables measured. The results of the experiments are the components of a vector

$$\boldsymbol{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}.$$

Linear regression amounts to determining $\boldsymbol{r} \in \mathbb{R}^n$ such that $X\boldsymbol{r} = \boldsymbol{b}$.
Knowing the components of $\boldsymbol{r}$ allows us to express the value of the result as a linear combination of the values of the variables. Unfortunately, since m is usually much larger than n, this system is overdetermined and, in general, is inconsistent.

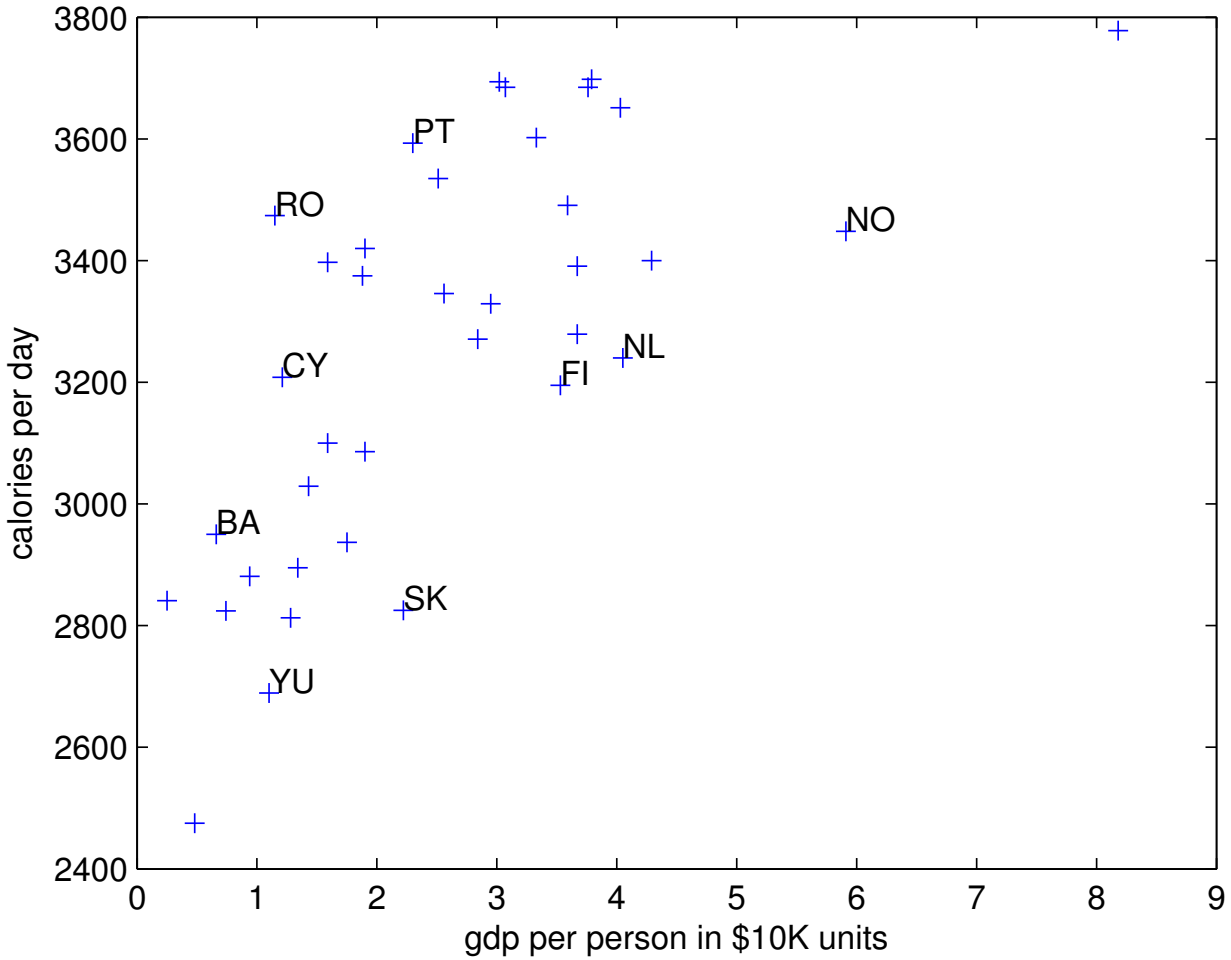
The columns $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ of the matrix X are referred to as the *regressors*; the linear combination $r_1\boldsymbol{v}_1 + \cdots + r_n\boldsymbol{v}_n$ is the *regression of* $\boldsymbol{b}$ *onto the regressors* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$.

Let us study the data set that represent (using the function `plot` of MATLAB), the number of calories consumed by a person per day vs. the gross national product per person in European countries.

ccode	gdp	cal	ccode	gdp	cal
'AL'	0.74	2824.00	'IT'	3.07	3685.00
'AT'	4.03	3651.00	'LV'	1.43	3029.00
'BY'	1.34	2895.00	'LT'	1.59	3397.00
'BE'	3.79	3698.00	'LU'	8.18	3778.00
'BA'	0.66	2950.00	'MK'	0.94	2881.00
'BG'	1.28	2813.00	'MT'	2.51	3535.00
'HR'	1.75	2937.00	'MD'	0.25	2841.00
'CY'	1.21	3208.00	'NL'	4.05	3240.00
'CZ'	2.56	3346.00	'NO'	5.91	3448.00
'DK'	3.67	3391.00	'PL'	1.88	3375.00
'EE'	1.90	3086.00	'PT'	2.30	3593.00
'FI'	3.53	3195.00	'RO'	1.15	3474.00
'FR'	3.33	3602.00	'RU'	1.59	3100.00
'GE'	0.48	2475.00	'YU'	1.10	2689.00
'DE'	3.59	3491.00	'SK'	2.22	2825.00
'GR'	3.02	3694.00	'SI'	2.84	3271.00
'HU'	1.90	3420.00	'ES'	2.95	3329.00
'IS'	3.67	3279.00	'CH'	4.29	3400.00
'IE'	3.76	3685.00			

UMASS BOSTON

This data is represented below:



We seek to approximate the calorie intake as a linear function of the gdp of the form

$$\mathrm{cal} = r_1 + r_2\ \mathrm{gdp}.$$

This amounts to solving a linear system that consists of 37 equations and two unknowns:

$$\begin{aligned} r_1 + 0.74 r_2 &= 2824 \\ &\vdots \\ r_1 + 4.29 r_2 &= 3400 \end{aligned}$$

and, clearly such a system is inconsistent.

If the linear system $X\boldsymbol{r} = \boldsymbol{b}$ has no solution, the "next best thing" is to find a vector $\boldsymbol{c} \in \mathbb{R}^n$ such that $\| X\boldsymbol{c} - \boldsymbol{b} \|_2 \leq \| X\boldsymbol{w} - \boldsymbol{b} \|_2$ for every $\boldsymbol{w} \in \mathbb{R}^n$, an approach known as *the least square method*. We will refer to the triple $(X, \boldsymbol{r}, \boldsymbol{b})$ as an *instance of the least square problem*.
Note that $X\boldsymbol{r} \in \text{range}(X)$ for any $\boldsymbol{r} \in \mathbb{R}^n$. Thus, solving this problem amounts to finding a vector $X\boldsymbol{r}$ in the subspace $\text{range}(X)$ such that $X\boldsymbol{r}$ is as close to $\boldsymbol{b}$ as possible.

Let $X \in \mathbb{R}^{m \times n}$ be a full-rank matrix such that $m > n$, so $\text{rank}(X) = n$. The symmetric square matrix $X'X \in \mathbb{R}^{n \times n}$ has the same rank n as the matrix X. Therefore, the system $(X'X)\boldsymbol{r} = X'\boldsymbol{b}$ a unique solution $\boldsymbol{s}$. Moreover, $X'X$ is positive definite because
$\boldsymbol{r}'X'X\boldsymbol{r} = (X\boldsymbol{r})'X\boldsymbol{r} = \| X\boldsymbol{r} \|_2^2 > 0$ for $\boldsymbol{r} \neq \boldsymbol{0}$.

Theorem

Let $X \in \mathbb{R}^{m \times n}$ be a full-rank matrix such that $m > n$ and let $\boldsymbol{b} \in \mathbb{R}^m$. The unique solution of the system $(X'X)\boldsymbol{r} = X'\boldsymbol{b}$ equals the projection of the vector $\boldsymbol{b}$ on the subspace range(X).

UMASS BOSTON

Proof

The n columns of the matrix $X = (\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_n)$ constitute a basis of the subspace range(X). Therefore, we seek the projection $\boldsymbol{c}$ of $\boldsymbol{b}$ on range(X) as a linear combination $\boldsymbol{c} = X\boldsymbol{t}$, which allows us to reduce this problem to a minimization of the function

$$\begin{aligned} f(\boldsymbol{t}) &= \| X\boldsymbol{t} - \boldsymbol{b} \|_2^2 \\ &= (X\boldsymbol{t} - \boldsymbol{b})'(X\boldsymbol{t} - \boldsymbol{b}) = (\boldsymbol{t}'X' - \boldsymbol{b}')(X\boldsymbol{t} - \boldsymbol{b}) \\ &= \boldsymbol{t}'X'X\boldsymbol{t} - \boldsymbol{b}'X\boldsymbol{t} - \boldsymbol{t}'X'\boldsymbol{b} + \boldsymbol{b}'\boldsymbol{b}. \end{aligned}$$

The necessary condition for the minimum is

$$(\nabla f)(\boldsymbol{t}) = 2X'X\boldsymbol{t} - 2X'\boldsymbol{b} = 0,$$

which implies $X'X\boldsymbol{t} = X'\boldsymbol{b}$.

The linear system $(X'X)\boldsymbol{t} = X'\boldsymbol{b}$ is known as the *system of normal equations* of X and $\boldsymbol{b}$.

Example

We augment the data sample matrix by a column that consists of 1s to accommodate a constant term r_1; thus, we work with the data sample matrix $X \in \mathbb{R}^{37\times 2}$ given by

$$X = \begin{pmatrix} 1 & 0.74 \\ \vdots & \vdots \\ 1 & 4.29 \end{pmatrix}$$

whose second column consists of the countries' gross domestic products in \$10K units. The matrix $C = X'X$ is

$$C = \begin{pmatrix} 37.0000 & 94.4600 \\ 94.4600 & 333.6592 \end{pmatrix}.$$

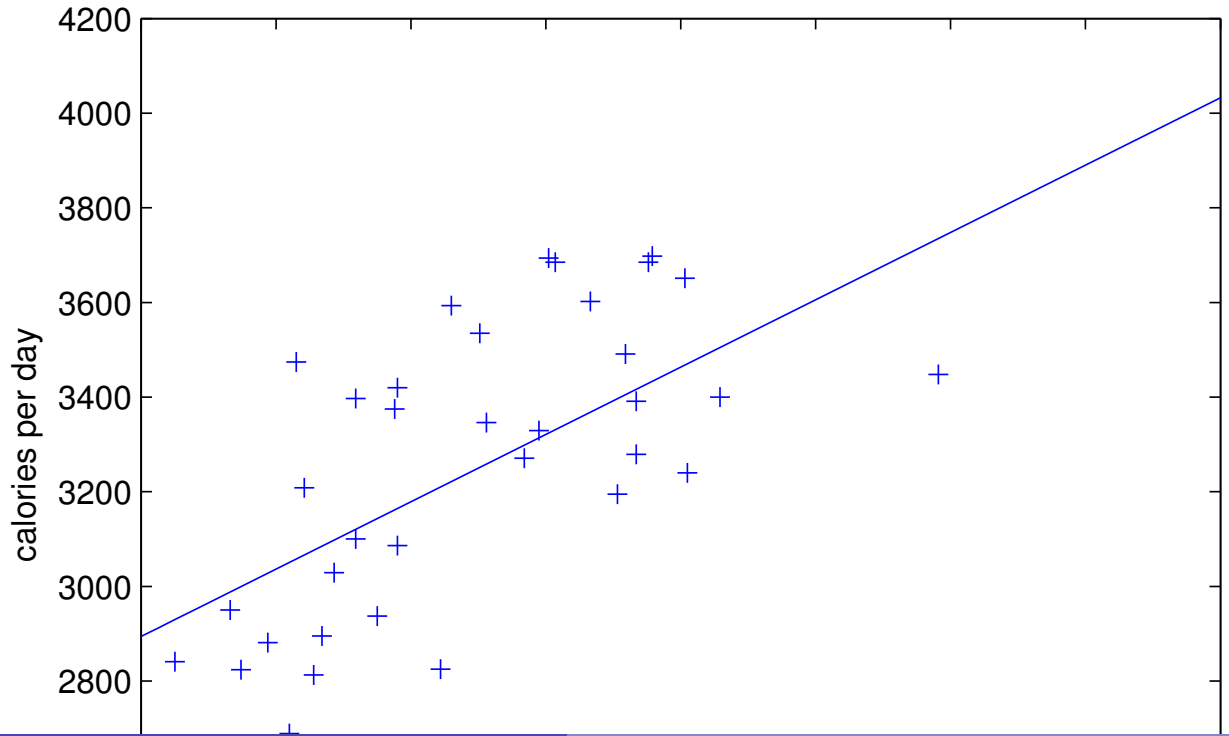
Example cont'd

Example

Solving the normal system using the MATLAB statement $\boldsymbol{r} = C \setminus (X' * b)$ yields

$$\boldsymbol{r} = \begin{pmatrix} 2894.2 \\ 142.3 \end{pmatrix},$$

so the regression line is $\text{cal} = 142.3 * \text{gdp} + 2894.2$, shown in next:



Suppose now that $X \in \mathbb{R}^{m \times n}$ has rank k, where $k < \min\{m, n\}$, and $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ are orthonormal matrices such that X can be factored as $X = UMV'$, where

$$M = \begin{pmatrix} R & O_{k,n-k} \\ O_{m-k,k} & O_{m-k,n-k} \end{pmatrix} \in \mathbb{R}^{m \times n},$$

$R \in \mathbb{R}^{k \times k}$, and $\text{rank}(R) = k$.

For $\boldsymbol{b} \in \mathbb{R}^m$ define $\boldsymbol{c} = U'\boldsymbol{b} \in \mathbb{R}^m$ and let $\boldsymbol{c} = \begin{pmatrix} \boldsymbol{c}_1 \\ \boldsymbol{c}_2 \end{pmatrix}$, where $\boldsymbol{c}_1 \in \mathbb{R}^k$ and $\boldsymbol{c}_2 \in \mathbb{R}^{m-k}$. Since $\text{rank}(R) = k$, the linear system $R\boldsymbol{z} = \boldsymbol{c}_1$ has a unique solution $\boldsymbol{z}_1$.

Theorem

All vectors $\boldsymbol{r}$ *that minimize* $\| X\boldsymbol{r} - \boldsymbol{b} \|_2$ *have the form*

$$\boldsymbol{r} = V \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{w} \end{pmatrix},$$

for an arbitrary $\boldsymbol{w}$.

UMASS BOSTON

Proof

We have

$$\begin{aligned} \| Xr - b \|_2^2 &= \| UMV'r - UU'b \|_2^2 \\ &= \| U(MV'r - U'b) \|_2^2 = \| MV'r - U'b \|_2^2 \\ &\quad \text{(because multiplication by an orthonormal matrix} \\ &\quad \text{is norm-preserving)} \\ &= \| MV'r - c \|_2^2 = \| My - c \|_2^2 \\ &= \| Rz - c_1 \|_2^2 + \| c_2 \|_2^2, \end{aligned}$$

where z consists of the first r components of y. This shows that the minimal value of $\| Xr - b \|_2^2$ is achieved by the solution of the system $Rz = c_1$ and is equal to $\| c_2 \|_2^2$. Therefore, the vectors r that minimize $\| Xr - b \|_2^2$ have the form $\begin{pmatrix} z \\ w \end{pmatrix}$ for an arbitrary $w \in \mathbb{R}^{n-r}$.

Instead of the Euclidean norm we can use the $\| \cdot \|_\infty$. Note that we have $t = \| Xr - b \|_\infty$ if and only if $-t\mathbf{1} \leq Xr - b \leq t\mathbf{1}$, so finding r that minimizes $\| \cdot \|_\infty$ amounts to solving a linear programming problem: minimize t subjected to the restrictions $-t\mathbf{1} \leq Xr - b \leq t\mathbf{1}$.
Similarly, we can use the norm $\| \cdot \|_p$. If $y = Xr - b$, then we need to minimize $\| y \|_p^p = |y_1|^p + \cdots + |y_m|^p$, subjected to the restrictions $-y \leq Ar - b \leq y$.

Solving the system of normal equation presents numeric difficulties because the condition number of the matrix $X'X$ is the square of the condition number of X. An alternative approach to finding $\boldsymbol{r} \in \mathbb{R}^n$ that minimizes $f(\boldsymbol{u}) = \| X\boldsymbol{u} - \boldsymbol{b} \|_2^2$ is to use a full QR decomposition of the matrix X, where $X \in \mathbb{R}^{m \times n}$ is a full-rank matrix and $m > n$.

Suppose that

$$X = Q \begin{pmatrix} R \\ O_{m-n,n} \end{pmatrix},$$

where $Q \in \mathbb{R}^{m \times m}$ is an orthonormal matrix and $R \in \mathbb{R}^{n \times n}$ is an upper triangular matrix such that

$$\begin{pmatrix} R \\ O_{m-n,n} \end{pmatrix} \in \mathbb{R}^{m \times n}.$$

We have

$$\begin{aligned} X\boldsymbol{u} - \boldsymbol{b} &= Q \begin{pmatrix} R \\ O_{m-n,n} \end{pmatrix} \boldsymbol{u} - \boldsymbol{b} \\ &= Q \begin{pmatrix} R \\ O_{m-n,n} \end{pmatrix} \boldsymbol{u} - QQ'\boldsymbol{b} \\ &\quad \text{(because } Q \text{ is orthonormal and therefore } QQ' = I_m\text{)} \\ &= Q \left(\begin{pmatrix} R \\ O_{m-n,n} \end{pmatrix} \boldsymbol{u} - Q'\boldsymbol{b} \right). \end{aligned}$$

multiplication by an orthogonal matrix preserves the Euclidean norm of vectors. Thus,

$$\| X\boldsymbol{u} - \boldsymbol{b} \|_2^2 = \left\| \begin{pmatrix} R \\ O_{m-n,n} \end{pmatrix} \boldsymbol{u} - Q'\boldsymbol{b} \right\|_2^2$$

If we write $Q = (L_1 \; L_2)$, where $L_1 \in \mathbb{R}^{m \times n}$ and $L_2 \in \mathbb{R}^{m \times (m-n)}$, then

$$\begin{aligned} \| X\boldsymbol{u} - \boldsymbol{b} \|_2^2 &= \left\| \begin{pmatrix} R \\ O_{m-n,n} \end{pmatrix} \boldsymbol{u} - \begin{pmatrix} L_1'\boldsymbol{b} \\ L_2'\boldsymbol{b} \end{pmatrix} \right\|_2^2 \\ &= \left\| \begin{pmatrix} R\boldsymbol{u} - L_1'\boldsymbol{b} \\ -L_2'\boldsymbol{b} \end{pmatrix} \right\|_2^2 \\ &= \| R\boldsymbol{u} - L_1'\boldsymbol{b} \|_2^2 + \| L_2'\boldsymbol{b} \|_2^2 . \end{aligned}$$

Observe that the system $R\boldsymbol{u} = L_1'\boldsymbol{b}$ can be solved and its solution minimizes $\| X\boldsymbol{u} - \boldsymbol{b} \|_2$.