LINEAR CLASSIFIERS

Prof. Dan A. Simovici

UMB

Matrices as Data Organizers

Let a data set consist of a sequence of m vectors of $\mathbb{R}^n$,

$$\mathcal{E} = (\mathbf{u}_1, \ldots, \mathbf{u}_m).$$

The j^{th} components $(\mathbf{u}_i)_j$ of these vectors correspond to the values of a random variable $\mathcal{V}_j$, where $1 \leqslant j \leqslant n$.
This data is represented as a matrix having m rows $\mathbf{u}'_1, \ldots, \mathbf{u}'_m$ and n columns $\mathbf{v}_1, \ldots, \mathbf{v}_n$. We refer to matrices obtained in this manner as *sample matrices*.
The number m is the *size* of the sample.

Each vector $\mathbf{u}_i'$ corresponds to an experiment E_i in the series of experiments $\mathcal{E} = (\mathcal{E}_1, \ldots, \mathcal{E}_m)$; the experiment $\mathcal{E}_i$ consists of measuring the n components of $\mathbf{u}_i' = (x_{i1}, \ldots, x_{in})$, as shown below.

	$\mathbf{v}_1$	$\cdots$	$\mathbf{v}_n$
$\mathbf{u}_1'$	x_{11}	$\cdots$	x_{1n}
$\mathbf{u}_2'$	x_{21}	$\cdots$	x_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{u}_m'$	x_{m1}	$\cdots$	x_{mn}

The column vector

$$\mathbf{v}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{mj} \end{pmatrix}$$

represents the measurements of the j^{th} variable (attribute) $\mathcal{V}_j$ of the experiment, for $1 \leqslant j \leqslant n$.

Definition

The *sample matrix* of $\mathcal{E}$ is the matrix $X \in \mathbb{C}^{m\times n}$ given by

$$X = \begin{pmatrix} \mathbf{u}_1' \\ \vdots \\ \mathbf{u}_m' \end{pmatrix} = (\mathbf{v}_1 \ \cdots \ \mathbf{v}_n).$$

We have $(\mathbf{v}_j)_i = (\mathbf{u}_i')_j = x_{ij}$ for $1 \leqslant i \leq m$ and $1 \leqslant j \leqslant n$.
We write

$$X = \begin{pmatrix} \mathbf{u}_1' \\ \vdots \\ \mathbf{u}_m' \end{pmatrix},$$

when we are interested in the vectors that represent results of experiments and $X = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$,when we need to work with vectors that represent the values of variables.

A *linear data mapping* for a data sequence $(\mathbf{u}_1, \ldots, \mathbf{u}_m) \in \mathbf{Seq}_m(\mathbb{R}^n)$ is the morphism $\mathsf{r} : \mathbb{R}^n \longrightarrow \mathbb{R}^q$. If $R \in \mathbb{R}^{n \times q}$ is the matrix that represents this mapping, then $\mathsf{r}(\mathbf{u}_i) = R\mathbf{u}_i$ for $1 \leqslant i \leqslant m$.
If $q < n$, we refer to r as a *linear dimensionality-reduction mapping*.
The *reduced data matrix* is given by

$$\mathsf{r}(X_{\mathcal{E}}) = \begin{pmatrix} \mathsf{r}(\mathbf{u}_1)' \\ \vdots \\ \mathsf{r}(\mathbf{u}_m)' \end{pmatrix} = \begin{pmatrix} (R\mathbf{u}_1)' \\ \vdots \\ (R\mathbf{u}_m)' \end{pmatrix} = X_{\mathcal{E}} R \in \mathbb{R}^{m \times q}$$

The reduced data set $\mathsf{r}(X_{\mathcal{E}})$ has new variables $\mathcal{Y}_1, \ldots, \mathcal{Y}_q$. We denote this by writing

$$(\mathcal{Y}_1, \ldots, \mathcal{Y}_q) = \mathsf{r}(\mathcal{V}_1, \ldots, \mathcal{V}_n).$$

The mapping r is a *linear feature selection mapping* if $R \in \{0,1\}^{q \times n}$ is a 0/1-matrix having exactly one unit in every row and at most one unit in every column.

The mean of a vector

For $\mathbf{u} \in \mathbb{R}^n$ the mean is the number

$$\tilde{\mathbf{u}} = \frac{1}{n}\sum_{i=1}^{n} u_i = \frac{1}{n}\sum_{i=1}^{n} \mathbf{u}'\mathbf{1}_n.$$

- $\mathbf{u}$ is centered if $\tilde{\mathbf{u}} = 0$;
- $\mathbf{u} - \tilde{\mathbf{u}}\mathbf{1}_n$ is always centered.

Sample Mean

Definition

Let $U = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$ be a series of observations in $\mathbb{R}^n$. The sample mean of this sequence is the vector

$$\tilde{U} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{u}_i \in \mathbb{R}^n.$$

Note that the series $(\mathbf{u}_1 - \tilde{U}, \ldots, \mathbf{u}_m - \tilde{U})$ is always centered and that

$$\tilde{U} = \frac{1}{m}(\mathbf{u}_1, \ldots, \mathbf{u}_m)'\mathbf{1}_n = \frac{1}{m}X\mathbf{1}_n \tag{1}$$

Thus, the data matrix $\hat{X}$ that corresponds to the centered sequence of vectors is

$$\hat{(X)} = X - \mathbf{1}_m\tilde{U}'.$$

If $m = 1$, the series of observation is reduced to a vector $\mathbf{v} \in \mathbb{R}^n$.

Standard Deviation

Definition

The standard deviation of a vector $\mathbf{v} \in \mathbb{R}^n$ is the number

$$s_{\mathbf{v}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(v_i - v)^2},$$

where v is the mean of $\mathbf{v}$.

The standard deviation of sample matrix $X \in \mathbb{R}^{m \times n}$, where $X = (\mathbf{v}_1 \ \cdots \ \mathbf{v}_n)$ is the row $\mathbf{s} = (s_{\mathbf{v}_1}, \ldots, s_{\mathbf{v}_n})$.

If the measurement scale for the variables $\mathcal{V}_1, \ldots, \mathcal{V}_n$ involved in the experiment are very different due to different measurement units, some variables may inappropriately influence the analysis process. Therefore, the columns of the data sample matrix need to be *scaled* in order to make. To scale a matrix we need to replace each column $\mathbf{v}_i$ by $\frac{1}{s_{\mathbf{v}_i}}\mathbf{v}_i$. This will yield a matrix having the standard deviation of each column equal to 1.

Centered Sample Matrix

Theorem

Let $X \in \mathbb{R}^{m \times n}$ is a sample matrix

$$X = \begin{pmatrix} \mathbf{u}_1' \\ \vdots \\ \mathbf{u}_m' \end{pmatrix}.$$

The sample matrix that corresponds to the centered sequence is

$$\hat{X} = \left(I_m - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m' \right) X.$$

Proof

The matrix that corresponds to the centered sequence is

$$\hat{X} = \begin{pmatrix} \mathbf{u}_1' - \tilde{U}' \\ \vdots \\ \mathbf{u}_m' - \tilde{U}' \end{pmatrix} = X - \mathbf{1}_m \tilde{U}'.$$

By Equality 1 it follows that

$$\hat{X} = X - \mathbf{1}_m \tilde{U}' = X - \frac{1}{m}\mathbf{1}_m \mathbf{1}_m' X = \left(I_m - \frac{1}{m}\mathbf{1}_m \mathbf{1}_m' \right) X,$$

which yields the desired equality.

Theorem shows that to center a data matrix $X \in \mathbb{R}^{m \times n}$ we need to multiply it at the left by the *centering matrix* H_m defined by

$$H_m = I_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}'_m \in \mathbb{R}^{m \times m},$$

that is, $\hat{X} = H_m X$. Note that $H_m = I_m - \frac{1}{m}J_m$. It is easy to see that H_m is both symmetric and idempotent. Since

$$H_m\mathbf{1}_m = \mathbf{1}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}'_m\mathbf{1}_m = \mathbf{0},$$

it follows that H_m has the eigenvalue 0.

The Effect of Centering on a Vector

For $\mathbf{w} \in \mathbb{R}^m$ we have

$$\begin{aligned} H_m \mathbf{w} &= \left(I_m - \frac{1}{m} \mathbf{1}_m \mathbf{1}'_m \right) \mathbf{w} \\ &= \mathbf{w} - \mathbf{1}_m \left(\frac{1}{m} \mathbf{1}'_m \mathbf{w} \right) \\ &= \mathbf{w} - \mathbf{1}_m \tilde{\mathbf{w}} \\ &= \begin{pmatrix} w_1 - \tilde{w} \\ \vdots \\ w_m - \tilde{w} \end{pmatrix} . \end{aligned}$$

Example

Let $X \in \mathbb{R}^{3\times 2}$ be the data matrix

$$X = \begin{pmatrix} 5 & 0 \\ 3 & 5 \\ 1 & 7 \end{pmatrix}.$$

We have $m = 3$ and $n = 2$. The centering matrix is

$$\begin{aligned} H_3 &= I_3 - \frac{1}{3}\mathbf{1}_3\mathbf{1}_3' \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \frac{1}{3}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}(1 \;\; 1 \;\; 1) \\ &= \begin{pmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{pmatrix} \end{aligned}$$

Example (cont'd)

The centered matrix is

$$H_3X = \begin{pmatrix} 2 & -4 \\ 0 & 1 \\ -2 & 3 \end{pmatrix}$$

Inertia of a Sequence of Vectors

Definition

Let $U = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$ be a sequence of vectors in $\mathbb{R}^n$. The *inertia* of this sequence relative to a vector $\mathbf{z} \in \mathbb{R}^n$ is the number

$$I_{\mathbf{z}}(U) = \sum_{j=1}^{m} \| \mathbf{u}_j - \mathbf{z} \|_2^2 .$$

Huygens' Inertia Theorem

Theorem

Let $U = (\mathbf{u}_1, \ldots, \mathbf{u}_m) \in \mathbf{Seq}_m(\mathbb{R}^n)$. *We have*

$$I_{\mathbf{z}}(U) - I_{\tilde{U}}(U) = m \parallel \tilde{U} - \mathbf{z} \parallel_2^2,$$

for every $\mathbf{z} \in \mathbb{R}^n$.

Proof

The inertia of U relative to $\tilde{U}$ is

$$\begin{aligned} I_{\tilde{U}}(U) &= \sum_{j=1}^{m} \| \mathbf{u}_j - \tilde{U} \|_2^2 \\ &= \sum_{j=1}^{m} (\mathbf{u}_j - \tilde{U})'(\mathbf{u}_j - \tilde{U}) \\ &= \sum_{j=1}^{m} (\mathbf{u}_j'\mathbf{u}_j - \tilde{U}'\mathbf{u}_j - \mathbf{u}_j'\tilde{U} + \tilde{U}'\tilde{U}). \end{aligned}$$

Similarly, we have $I_{\mathbf{z}}(U) = \sum_{j=1}^{m} (\mathbf{u}_j'\mathbf{u}_j - \mathbf{z}'\mathbf{u}_j - \mathbf{u}_j'\mathbf{z} + \mathbf{z}'\mathbf{z})$.

Proof (cont'd)

This allows us to write

$$\begin{aligned} I_{\mathbf{z}}(U) - I_{\tilde{U}}(U) &= \sum_{j=1}^{m} (\tilde{U} - \mathbf{z})' \mathbf{u}_j + \sum_{j=1}^{m} \mathbf{u}_j' (\tilde{U} - \mathbf{z}) + \mathbf{z}'\mathbf{z} - \tilde{U}'\tilde{U} \\ &= (\tilde{U} - \mathbf{z})' \sum_{i=1}^{m} \mathbf{u}_j + \left(\sum_{j=1}^{m} \mathbf{u}_j \right)' (\tilde{U} - \mathbf{z}) + m(\mathbf{z}'\mathbf{z} - \tilde{U}'\tilde{U}) \\ &= m(\tilde{U} - \mathbf{z})'\tilde{U} + m\tilde{U}'(\tilde{U} - \mathbf{z}) + m(\mathbf{z}'\mathbf{z} - \tilde{U}'\tilde{U}) \\ &= m \parallel \tilde{U} - \mathbf{z} \parallel_2^2, \end{aligned}$$

which is the equality of the theorem.

Corollary

Let $U = (\mathbf{u}_1, \ldots, \mathbf{u}_m) \in \mathbf{Seq}_m(\mathbb{R}^n)$. *The minimal value of the inertia* $I_{\mathbf{z}}(U)$ *is achieved for* $\mathbf{z} = \tilde{U}$.

Let $\mathbf{u}$ and $\mathbf{w}$ be two vectors in $\mathbb{R}^m$, where $m > 1$, having the means u and w, and the standard deviations s_u and s_v, respectively.

Definition

The *covariance coefficient* of $\mathbf{u}$ and $\mathbf{w}$ is the number

$$cov(\mathbf{u}, \mathbf{w}) = \frac{1}{m-1} \sum_{i=1}^{m-1} (u_i - u)(w_i - w)$$

The *correlation coefficient* of $\mathbf{u}$ and $\mathbf{w}$ is the number

$$\rho(\mathbf{u}, \mathbf{w}) = \frac{cov(\mathbf{u}, \mathbf{w})}{s_u s_w}$$

$$\left| \sum_{i=1}^{m} (u_i - u)(w_i - w) \right| \leq \sqrt{\sum_{i=1}^{m} (u_i - u)^2} \cdot \sqrt{\sum_{i=1}^{m} (w_i - w)^2},$$

which implies

Definition

Let $X \in \mathbb{R}^{m \times n}$ be a sample matrix and let $\hat{X}$ be the centered sample matrix corresponding to X. The *sample covariance matrix* is the matrix

$$cov(X) = \frac{1}{m-1}\hat{X}'\hat{X} \in \mathbb{R}^{n \times n}.$$

If X is centered, $cov(X) = \frac{1}{m-1}X'X$.

Example

We saw that for

$$X = \begin{pmatrix} 5 & 0 \\ 3 & 5 \\ 1 & 7 \end{pmatrix} \text{ we have } \hat{X} = \begin{pmatrix} 2 & -4 \\ 0 & 1 \\ -2 & 3 \end{pmatrix}.$$

The covariance matrix is

$$cov(X) = \frac{1}{2}\hat{X}'\hat{X} = \begin{pmatrix} 4 & -7 \\ -7 & 4 \end{pmatrix}.$$

Note the negative value of the covariance $cov(X)_{12}$; this suggests that the values associated to the two variables vary in oposite directions: when one increases the other decreases.

Variance

If $n = 1$ the matrix is reduced to one column $X = (\mathbf{v})$. If $\mathbf{v}$ is centered, then

$$cov(\mathbf{v}) = \frac{1}{m-1}\mathbf{v}'\mathbf{v} = \frac{1}{m-1} \parallel \mathbf{v} \parallel^2 \in \mathbb{R}.$$

In this case we refer to $cov(\mathbf{v})$ as the variance of $\mathbf{v}$; this number is denoted by $var(\mathbf{v})$.
If $\mathbf{v}$ is not centered we have

$$\begin{aligned} cov(\mathbf{v}) &= \frac{1}{m-1}\sum_{i=1}^{m}(v_i - \bar{v})^2 \\ &= \frac{1}{m-1}\left(\sum_{i=1}^{m} v_i^2 - m(\bar{v})^2\right). \end{aligned}$$

Covariance and Correlation Matrices

If $X = (\mathbf{v}_1 \ \cdots \ \mathbf{v}_n)$, then $(cov(X))_{ij} = cov(\mathbf{v}_i, \mathbf{v}_j)$ for $1 \leqslant i, j \leqslant n$. The covariance matrix can be written also as

$$cov(X) = \frac{1}{m-1} X' H_m H_m X = \frac{1}{m-1} X' H_m X.$$

The sample correlation matrix is the matrix $corr(X)$ given by

$$(corr(X))_{ij} = \rho(\mathbf{v}_i, \mathbf{v}_j)$$

for $1 \leqslant i, j \leqslant n$. If X is centered, then $cov(X) = \frac{1}{m-1} X'X$. Clearly, the covariance matrix is a symmetric, positive semidefinite matrix.
If $(cov(X))_{ij} = 0$ we say that features $\mathcal{V}_i$ and $\mathcal{V}_j$ are uncorrelated.

- the rank of $cov(X)$ is the same as the rank of $\hat{X}$ and, since m, the size of the sample is usually much larger than n we are often justified in asumming that $rank(cov(X)) = n$;
- for a sample matrix $X = (\mathbf{v}_1 \ \cdots \ \mathbf{v}_n) \in \mathbb{R}^{m \times n}$ we have

$$H_m \mathbf{v}_p = (I_m - \frac{1}{m}\mathbf{1}_m \mathbf{1}'_m)\mathbf{v}_p = \mathbf{v}_p - \frac{1}{m}\mathbf{1}_m \mathbf{1}'_m \mathbf{v}_p = \mathbf{v}_p - a_p \mathbf{1}_m,$$

because $\frac{1}{m}\mathbf{1}'_m \mathbf{v}_p = a_p$ for $1 \leqslant p \leqslant n$, where $\tilde{\mathbf{u}}' = (a_1, \ldots, a_n)$.

The Covariance Matrix

The covariance matrix can be written as

$$\begin{aligned} cov(X) &= \frac{1}{m-1}(\mathbf{v}_1 \ \cdots \ \mathbf{v}_n)' H_m' H_m (\mathbf{v}_1 \ \cdots \ \mathbf{v}_n) \\ &= \frac{1}{m-1}(H_m\mathbf{v}_1 \ \cdots \ H_m\mathbf{v}_n)'(H_m\mathbf{v}_1 \ \cdots \ H_m\mathbf{v}_n), \end{aligned}$$

which implies that the (p, q)-entry of this matrix is

$$cov(X)_{pq} = \frac{1}{m-1}(H_m\mathbf{v}_p)'(H_m\mathbf{v}_q) = \frac{1}{m-1}(\mathbf{v}_p - a_p\mathbf{1}_m)'(\mathbf{v}_q - a_q\mathbf{1}_m).$$

The Covariance Matrix

- For a diagonal element we have

$$cov(X)_{pp} = \frac{1}{m-1}\sum_{i=1}^{m}(\mathbf{v}_q - a_q\mathbf{1}_m)_i^2,$$

  which shows that $cov(X)_{pp}$ measures the scattering of the values of the p^{th} variable around the corresponding component a_i of the mean sample.
- $cov(X)_{pp}$ is known as the p^{th} *variance* and is denoted by σ_p^2 for $1 \leq p \leqslant n$.
- The total variance TVAR(X) of X is $trace(cov(X))$.

For $p \neq q$ the element c_{pq} of the matrix $C = cov(X)$ is referred to as the (p, q)-*covariance*. We have:

$$\begin{aligned}(cov(X))_{pq} &= \frac{1}{m-1}(\mathbf{v}_p - a_p\mathbf{1}_m)'(\mathbf{v}_q - a_q\mathbf{1}_m) \\ &= \frac{1}{m}\left(\mathbf{v}_p'\mathbf{v}_q - a_p\mathbf{1}_m'\mathbf{v}_q - a_q\mathbf{v}_p'\mathbf{1}_m + ma_pa_q\right) \\ &= \mathbf{v}_p'\mathbf{v}_q - a_pa_q.\end{aligned}$$

If $cov(X)_{pq} = 0$, then we say that the variables $\mathcal{V}_p$ and $\mathcal{V}_q$ are *uncorrelated*.

Problem Formulation

Let $X \in \mathbb{R}^{m \times n}$ be a data matrix that contains results of m experiments involving n variables.
Suppose that the experiments produce a result vector $\mathbf{y} \in \mathbb{R}^m$.
The purpose of linear regression is to learn a vector $\mathbf{r} \in \mathbb{R}^n$ and a scalar a such that

$$\mathbf{y} = X\mathbf{r} + a\mathbf{1}_m, \tag{2}$$

which allows us to express the value of the result as a linear combination of the values of the variables.

Equality (2) can be also written as

$$\mathbf{y} = \begin{pmatrix} X & \mathbf{1}_m \end{pmatrix} \begin{pmatrix} \mathbf{r} \\ a \end{pmatrix}$$

(the homogeneous form of the regression equation). The homogeneous form allows us to deal with the regression problem as solving in $\mathbf{r}$ an overdetermined linear system $Y \begin{pmatrix} \mathbf{r} \\ a \end{pmatrix} = \mathbf{y}$, where

$$Y = \begin{pmatrix} X & \mathbf{1}_m \end{pmatrix} \in \mathbb{R}^{m \times (n+1)}.$$

Since m is usually much larger than n, this system is over-determined and, in general, is inconsistent.

- the columns $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of the matrix X are referred to as the *regressors*;
- the linear combination $r_1\mathbf{v}_1 + \cdots + r_n\mathbf{v}_n + a\mathbf{1}_m$ is the *regression of* $\mathbf{y}$ *onto the regressors* $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

If the linear system

$$Y\begin{pmatrix}\mathbf{r}\\ a\end{pmatrix}=\mathbf{y}$$

has no solution, the "next best thing" is to find a vector $\mathbf{c}\in\mathbb{R}^{n+1}$ such that minimizes $\|\ Y\mathbf{c}-\mathbf{y}\ \|_2$, that is,

$$\|\ Y\mathbf{c}-\mathbf{y}\ \|_2\leqslant\|\ Y\mathbf{w}-\mathbf{y}\ \|_2$$

for every $\mathbf{w}\in\mathbb{R}^{n+1}$, an approach known as *the least square method*. Note that

$$Y\begin{pmatrix}\mathbf{r}\\ a\end{pmatrix}\in\text{Ran}(Y)$$

Thus, solving this problem amounts to finding a vector $Y\begin{pmatrix}\mathbf{r}\\ a\end{pmatrix}$ in the subspace $\text{Ran}(Y)$ that is as close to $\mathbf{y}$ as possible.

Let $Y \in \mathbb{R}^{m\times(n+1)}$ be a full-rank matrix such that $m \geqslant n+1$, so $rank(Y) = n+1$. The symmetric square matrix $Y'Y \in \mathbb{R}^{(n+1)\times(n+1)}$ has the same rank $n+1$ as the matrix Y. Therefore, the system $(Y'Y)\mathbf{r} = Y'\mathbf{y}$ has a unique solution $\mathbf{r}$.
$Y'Y$ is positive definite because $\mathbf{r}'Y'Y\mathbf{r} = (Y\mathbf{r})'Y\mathbf{r} = \| Y\mathbf{r} \|_2^2 > 0$ for $\mathbf{r} \neq \mathbf{0}$.

Theorem

Let $Y \in \mathbb{R}^{m\times(n+1)}$ *be a full-rank matrix such that* $m \geqslant n+1$ *and let* $\mathbf{y} \in \mathbb{R}^m$. *The unique solution of the system*

$$(Y'Y)\mathbf{z} = Y'\mathbf{y}$$

equals the projection of the vector $\mathbf{y}$ *on the subspace* $\mathrm{Ran}(Y)$.

Proof

The columns of the matrix $Y = (\mathbf{v}_1 \ \cdots \ \mathbf{v}_n \ \mathbf{1}_m)$ constitute a basis of the subspace Ran(Y). Therefore, we seek the projection $\mathbf{z}$ of $\mathbf{y}$ on Ran(Y) as a linear combination $\mathbf{c} = Y\mathbf{t}$, which allows us to reduce this problem to a minimization of the function

$$\begin{aligned} f(\mathbf{t}) &= \| \ Y\mathbf{t} - \mathbf{y} \ \|_2^2 \\ &= (Y\mathbf{t} - \mathbf{y})'(Y\mathbf{t} - \mathbf{y}) = (\mathbf{t}'Y' - \mathbf{y}')(Y\mathbf{t} - \mathbf{y}) \\ &= \mathbf{t}'Y'Y\mathbf{t} - \mathbf{y}'Y\mathbf{t} - \mathbf{t}'Y'\mathbf{y} + \mathbf{y}'\mathbf{y}. \end{aligned}$$

The necessary condition for the minimum is

$$(\nabla f)(\mathbf{t}) = 2Y'Y\mathbf{t} - 2Y'\mathbf{y} = 0,$$

which implies $Y'Y\mathbf{t} = Y'\mathbf{y}$.

If $n = 1$, then we have the univariate regression.

- $Y \in \mathbb{R}^{m\times 2}$ is reduced to a matrix $(\mathbf{v}\ \mathbf{1}_m) \in \mathbb{R}^{m\times 2}$;
- we seek $r \in \mathbb{R}$ such that

$$(\mathbf{v}\ \mathbf{1}_m)\begin{pmatrix} r \\ a \end{pmatrix} = \mathbf{v}r + \mathbf{1}_m a = \mathbf{y},$$

The compatible system $Y'Y\mathbf{t} = Y'\mathbf{y}$ amounts to

$$\begin{pmatrix} \mathbf{v}' \\ \mathbf{1}'_m \end{pmatrix}(\mathbf{v}\ \mathbf{1}_m)\begin{pmatrix} \mathbf{r} \\ a \end{pmatrix} = \begin{pmatrix} \mathbf{v}' \\ \mathbf{1}'_m \end{pmatrix}\mathbf{y}$$

An equivalent form of the system is

$$\begin{pmatrix} \mathbf{v}'\mathbf{v} & \mathbf{v}'\mathbf{1}_m \\ \mathbf{1}'_m\mathbf{v} & m \end{pmatrix} \begin{pmatrix} \mathbf{r} \\ a \end{pmatrix} = \begin{pmatrix} \mathbf{v}'\mathbf{y} \\ \mathbf{1}'_m\mathbf{y} \end{pmatrix}.$$

If $\mathbf{v}$ is centered, then $\mathbf{v}'\mathbf{1}_m = 0$ and this yields

$$\begin{pmatrix} \mathbf{v}'\mathbf{v} & 0 \\ 0 & m \end{pmatrix} \begin{pmatrix} \mathbf{r} \\ a \end{pmatrix} = \begin{pmatrix} \mathbf{v}'\mathbf{y} \\ \mathbf{1}'_m\mathbf{y} \end{pmatrix},$$

or $(\mathbf{v}'\mathbf{v})r = \mathbf{v}'\mathbf{y}$, that is,

$$r = \frac{\sum_{i=1}^m v_i y_i}{\sum_{i=1}^m v_i^2}$$

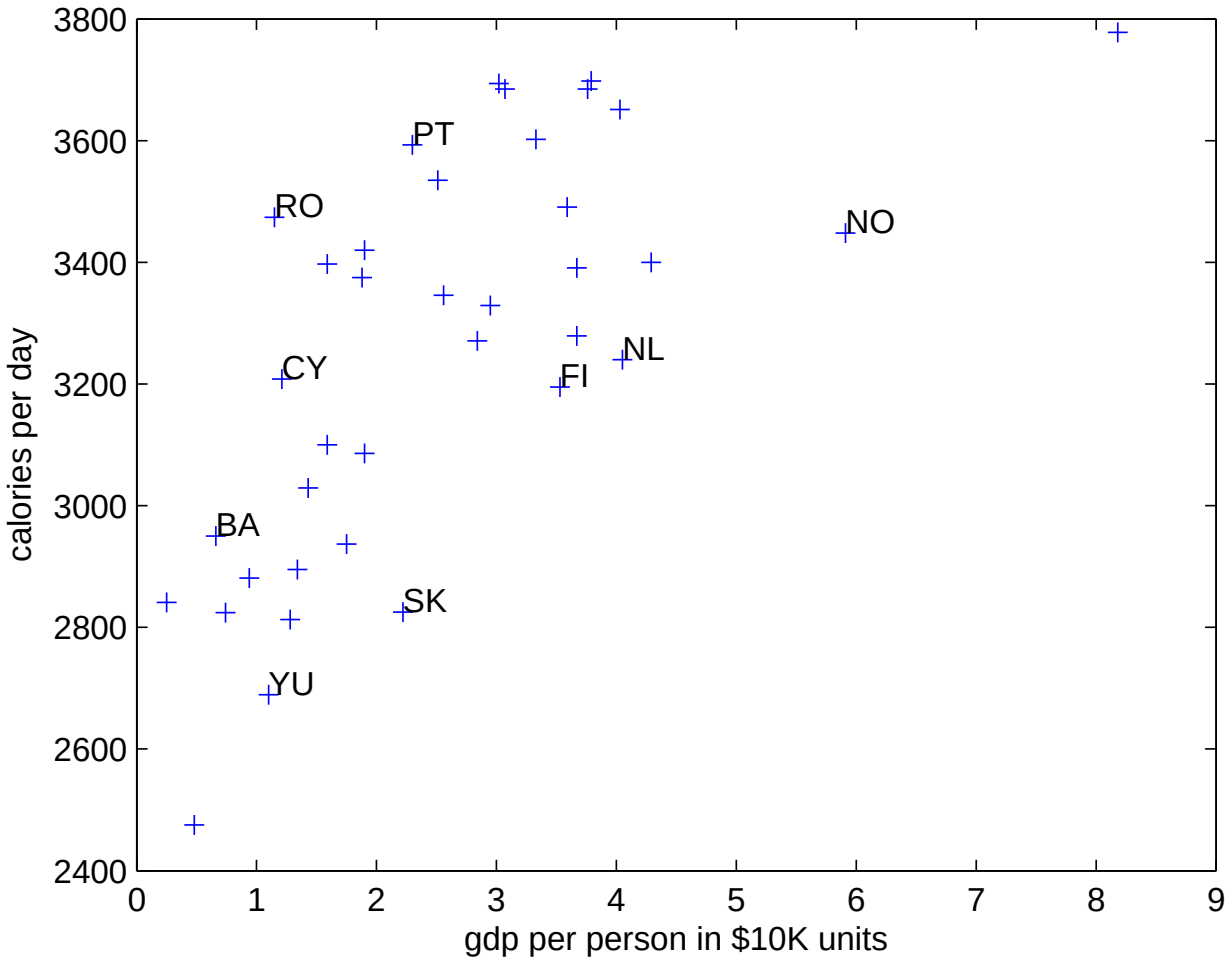
and $a = \frac{1}{m}\mathbf{1}'_m\mathbf{y} = \frac{1}{m}\sum_{i=1}^m y_i$.

Example

Next table represents the number of calories consumed by a person per day vs. the gross national product per person in European countries (in 10K$ units).

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			

Calories vs. GDP



Example (cont'd)

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

$$\mathrm{cal} = r\ \mathrm{gdp} + a.$$

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

$$\begin{aligned} 0.74r + a &= 2824 \\ &\vdots \\ 4.29r + a &= 3400 \end{aligned}$$

and, clearly such a system is inconsistent.
The matrix Y and the vector $\mathbf{y}$ are

$$Y = \begin{pmatrix} 0.74 & 1 \\ \vdots & \vdots \\ 4.29 & 1 \end{pmatrix} \text{ and } \mathbf{y} = \begin{pmatrix} 2824 \\ \vdots \\ 3400 \end{pmatrix}.$$

Example (cont'd)

The system we need to solve is

$$Y'Y \begin{pmatrix} r \\ a \end{pmatrix} = Y'\mathbf{y}.$$

We have

$$Y'Y = \begin{pmatrix} 333.6592 & 94.46 \\ 94.46 & 37.00 \end{pmatrix} \text{ and } Y'\mathbf{y} = \begin{pmatrix} 320880 \\ 120530 \end{pmatrix}$$

and the system yields $r = 142.3$ and $a = 2894.2$.

Regression Line