Regression - I

Prof. Dan A. Simovici

UMB

Let $A \in \mathbb{R}^{m \times n}$ be a matrix. The *null space of A* is the subspace of $\mathbb{R}^n$ defined by

$$\text{Nullsp}(A) = \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} = \mathbf{0}_m\}.$$

The *range* of A is the subspace of $\mathbb{R}^m$ defined as

$$\text{Ran}(A) = \{\mathbf{y} \in \mathbb{R}^m \mid \mathbf{y} = A\mathbf{x}\}.$$

The rank of A is the number $\text{rank}(A)$ that is dimension of $\text{Ran}(A)$, that is, the size of the largest linearly independent set in $\text{Ran}(A)$.

If $A \in \mathbb{R}^{m \times n}$, the transposed matrix is $A' \in \mathbb{R}^{n \times m}$.
The *inner product* of two vectors $\mathbf{x}, \mathbf{y}$ in $\mathbb{R}^p$ is the number $(\mathbf{x}, \mathbf{y}) = \mathbf{x}'\mathbf{y}$.

Theorem

Let $A \in \mathbb{R}^{m \times n}$, and $\mathbf{x} \in \mathbb{R}^n$, and $\mathbf{y} \in \mathbb{R}^m$. We have $(A\mathbf{x}, \mathbf{y}) = (\mathbf{x}, A'\mathbf{y})$.

Proof: We have $(A\mathbf{x}, \mathbf{y}) = (A\mathbf{x})'\mathbf{y} = \mathbf{x}'A'\mathbf{y}$ and $(\mathbf{x}, A'\mathbf{y}) = \mathbf{x}'(A'\mathbf{y})$ and these numbers are equal by the associativity.

If $a \in \mathbb{R}^{m \times n}$ and $A = BC$, where $B \in \mathbb{R}^{m \times r}$ and $C \in \mathbb{R}^{r \times n}$, then

- the i^{th} row of A is a linear combination of the r rows of C with coefficients from the i^{th} row of B;
- the j^{th} column of A is a linear combination of the r columns of B with coefficients from the j^{th} row of C;

- If any collection of rows $\bar{c}_1, \ldots, \bar{c}_r$ spans the row space of A an $r \times n$ matrix C can be formed by taking these vectors as its rows; then, the i^{th} row of A is a linear combination of the rows of C, say $\bar{a}_i = b_{i1}\bar{c}_1 + \cdots + b_{ir}\bar{c}_r$. This means that $A = BC$, where $B = (b_{ij})$ is the $m \times r$ matrix, where the i^{th} row is $\bar{b}_i = (b_{i1}, \ldots, b_{ir})$;
- similarly, if any r column vectors span the column space of A and B is the $m \times r$ matrix formed by these columns, then the $r \times n$ matrix C formed from appropriate coefficients satisfies $A = BC$.

Theorem

If $A \in \mathbb{R}^{m\times n}$, then the row rank of A is equal to the column rank of A.

Proof: If $A = O_{m\times n}$, then the row rank and the column rank are 0; otherwise, let r be the smallest positive integer such that there exists $B \in \mathbb{R}^{m\times r}$ and $C \in \mathbb{R}^{r\times n}$ such that $A = BC$. Since the r rows of C form a minimal spanning set of the row space of A and the r columns of B form a minimal spanning set of the column space of A, row and column ranks are both r.

Theorem

Let $A \in \mathbb{R}^{m\times n}$. We have

$$\dim(\mathit{Nullsp}(A)) + \dim(\mathrm{Ran}(A)) = n.$$

Suppose that $\{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ is a basis for $\mathrm{Nullsp}(A) \subseteq \mathbb{R}^n$. Extend this base to a base for $\mathbb{R}^n$: $\{\mathbf{e}_1, \ldots, \mathbf{e}_m, \mathbf{e}_{m+1}, \ldots, \mathbf{e}_n\}$. Any $\mathbf{v} \in \mathbb{R}^n$ can be written as $\mathbf{v} = v_1\mathbf{e}_1 + \cdots + v_m\mathbf{e}_m + v_{m+1}\mathbf{e}_{m+1} + \cdots + v_n\mathbf{e}_n$, hence
$A\mathbf{v} = v_{m+1}A\mathbf{e}_{m+1} + \cdots + v_nA\mathbf{e}_n$.
Therefore, $\{A\mathbf{e}_{m+1}, \ldots, A\mathbf{e}_n\}$ spans $\mathrm{Ran}(A)$. This set is linearly independent, so it is a base for $\mathrm{Ran}(A)$ and thus, $\dim(\mathrm{Ran}(A)) = n - m$.

Definition

A matrix A is invertible if there exists a matrix A^{-1} such that $AA^{-1} = A^{-1}A = I_n$.

Theorem

If $A \in \mathbb{R}^{n \times n}$ is invertible, than $rank(A) = n$.

$B \in \mathbb{R}^{m \times n}$ is a *full-rank matrix* if $\text{rank}(B) = \min\{m, n\}$.
Let $B \in \mathbb{R}^{m \times n}$ be a full-rank matrix such that $m > n$, so $\text{rank}(B) = n$.
The symmetric square matrix $B'B \in \mathbb{R}^{n \times n}$ has the same rank n as the matrix B because $\text{Nullsp}(B'B) = \text{Nullsp}(B)$. This makes $B'B$ an invertible matrix, that is, there exists $(B'B)^{-1}$.

Experimental Setting

Suppose that the results of a series of m experiments are the components of a vector $\mathbf{y} \in \mathbb{R}^m$. For the i^{th} experiment, the values of the n input variables $x_1, \ldots, x_n$ are placed in the i^{th} row of a matrix $B \in \mathbb{R}^{m \times n}$ known as the *design matrix*, and we assume that the outcome of the i^{th} experiment y_i is a linear function of the values $b_{i1}, \ldots, b_{in}$ of $x_1, \ldots, x_n$, that is

$$y_i = b_{i1} r_1 + \cdots + b_{in} r_n.$$

The variables $x_1, \ldots, x_n$ are referred to as the *regressors*. Note that the values assummed by the variable x_j in the series of m experiments, $b_{1j}, \ldots, b_{mj}$ have been placed in the j^{th} column $\mathbf{b}_j$ of the matrix B.

Linear regression assumes the existence of a linear relationship between the outcome of an experiment and values of variables that are measured during the experiment.
In general there are more experiments than variables, that is, we have $n < m$. In matrix form we have $\mathbf{y} = B\mathbf{r}$, where $B \in \mathbb{R}^{m \times n}$ and $\mathbf{r} \in \mathbb{R}^n$. The problem is to determine $\mathbf{r}$, when B and $\mathbf{y}$ are known. Since $n < m$, this linear system is inconsistent, but is is possible to obtain an approximative solution by determining $\mathbf{r}$ such that $\| \mathbf{y} - B\mathbf{r} \|$ is minimal. This amounts to approximating $\mathbf{y}$ by a vector in the subspace $\mathrm{Ran}(B)$ generated by the columns of the matrix B.

The columns $\mathbf{b}_1, \ldots, \mathbf{b}_n$ of the matrix B are referred to as the *regressors*; the linear combination $r_1\mathbf{b}_1 + \cdots + r_n\mathbf{b}_n$ is the *regression of* $\mathbf{y}$ *onto the regressors* $\mathbf{b}_1, \ldots, \mathbf{b}_n$.
A variant of the previous model is to asumme that $\mathbf{y}$ is affinely dependent on $\mathbf{b}_1, \ldots, \mathbf{b}_q$, that is,

$$\mathbf{y} = r_0 + r_1\mathbf{b}_1 + \cdots + r_n\mathbf{b}_n,$$

and we seek to determine the coefficients $r_0, r_1, \ldots, r_n$. The term r_0 is the *bias* of the model. The dependency of $\mathbf{y}$ on $\mathbf{b}_1, \ldots, \mathbf{b}_n$ can be homogenized by introducing a dummy vector $\mathbf{b}_0$ having all components equal to 1, which gives

$$\mathbf{y} = r_0\mathbf{b}_0 + r_1\mathbf{b}_1 + \cdots + r_n\mathbf{b}_n,$$

as the defining assumption of the model.

As we stated before, if the linear system $B\mathbf{r} = \mathbf{y}$ has no solution $\mathbf{r}$, the "next best thing" is to find a vector $\mathbf{r} \in \mathbb{R}^n$ such that

$$\| B\mathbf{r} - \mathbf{y} \|_2 \leqslant \| B\mathbf{w} - \mathbf{y} \|_2$$

for every $\mathbf{w} \in \mathbb{R}^n$. This approach is known as *the least square method*. We will refer to the triple $(B, \mathbf{r}, \mathbf{y})$ as an *instance of the least square problem*.

- Note that $B\mathbf{r} \in$ **range of** (**is** B) for any $\mathbf{r} \in \mathbb{R}^n$. Thus, solving this problem amounts to finding a vector $B\mathbf{r}$ in the subspace **range of** (**is** B) such that $B\mathbf{r}$ is as close to $\mathbf{y}$ as possible.
- Let $B \in \mathbb{R}^{m \times n}$ be a full-rank matrix such that $m > n$, so $\text{rank}(B) = n$. The symmetric square matrix $B'B \in \mathbb{R}^{n \times n}$ has the same rank n as the matrix B. Therefore, the system $(B'B)\mathbf{r} = B'\mathbf{y}$ a unique solution $\mathbf{r} = (B'B)^{-1}B'\mathbf{y}$. Moreover, $B'B$ is positive definite because $\mathbf{r}'B'B\mathbf{r} = (B\mathbf{r})'B\mathbf{r} = \| B\mathbf{r} \|_2^2 > 0$ for $\mathbf{r} \neq \mathbf{0}_n$.

Theorem

Let $B \in \mathbb{R}^{m \times n}$ be a full-rank matrix such that $m > n$ and let $\mathbf{y} \in \mathbb{R}^m$. The unique solution $\mathbf{r} = (B'B)^{-1}B'\mathbf{y}$ of the system $(B'B)\mathbf{r} = B'\mathbf{y}$ equals the projection of the vector $\mathbf{y}$ on the subspace $\mathrm{Ran}(B)$.

Proof

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

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

The necessary condition for the minimum is

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

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

The linear system $(B'B)\mathbf{t} = B'\mathbf{y}$ is known as the *system of normal equations* of B and $\mathbf{y}$.

The Case of non-full rank matrix B

Suppose now that $B \in \mathbb{R}^{m\times n}$ 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 B can be factored as $B = UMV'$, 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 $\mathbf{y} \in \mathbb{R}^m$ define $\mathbf{c} = U'\mathbf{y} \in \mathbb{R}^m$ and let $\mathbf{c} = \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{pmatrix}$, where $\mathbf{c}_1 \in \mathbb{R}^k$ and $\mathbf{c}_2 \in \mathbb{R}^{m-k}$. Since $\text{rank}(R) = k$, the linear system $R\mathbf{z} = \mathbf{c}_1$ has a unique solution $\mathbf{z}_1$.

Theorem

All vectors $\mathbf{r}$ *that minimize* $\| B\mathbf{r} - \mathbf{y} \|_2$ *have the form*

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

for an arbitrary $\mathbf{w}$.

Proof

We have

$$\begin{aligned} \| B\mathbf{r} - \mathbf{y} \|_2^2 &= \| UMV'\mathbf{r} - UU'\mathbf{y} \|_2^2 \\ &= \| U(MV'\mathbf{r} - U'\mathbf{y}) \|_2^2 = \| MV'\mathbf{r} - U'\mathbf{y} \|_2^2 \\ &\quad \text{(because multiplication by an orthonormal matrix} \\ &\quad \text{is norm-preserving)} \\ &= \| MV'\mathbf{r} - \mathbf{c} \|_2^2 = \| M\mathbf{y} - \mathbf{c} \|_2^2 \\ &= \| R\mathbf{z} - \mathbf{c}_1 \|_2^2 + \| \mathbf{c}_2 \|_2^2, \end{aligned}$$

where $\mathbf{z}$ consists of the first r components of $\mathbf{y}$. This shows that the minimal value of $\| B\mathbf{r} - \mathbf{y} \|_2^2$ is achieved by the solution of the system $R\mathbf{z} = \mathbf{c}_1$ and is equal to $\| \mathbf{c}_2 \|_2^2$. Therefore, the vectors $\mathbf{r}$ that minimize $\| B\mathbf{r} - \mathbf{y} \|_2^2$ have the form $\begin{pmatrix} \mathbf{z} \\ \mathbf{w} \end{pmatrix}$ for an arbitrary $\mathbf{w} \in \mathbb{R}^{n-r}$.

Instead of the Euclidean norm we can use the $\|\cdot\|_\infty$. Note that we have $t = \| B\mathbf{r} - \mathbf{y} \|_\infty$ if and only if $-t\mathbf{1} \leq B\mathbf{r} - \mathbf{y} \leq t\mathbf{1}$, so finding $\mathbf{r}$ that minimizes $\|\cdot\|_\infty$ amounts to solving a linear programming problem: minimize t subjected to the restrictions $-t\mathbf{1} \leq B\mathbf{r} - \mathbf{y} \leq t\mathbf{1}$.

An Equivalent Formulation

An optimization approach to linear regression seeks $\mathbf{r} \in \mathbb{R}^n$ that minimizes the *square loss function* $L : \mathbb{R}^n \longrightarrow \mathbb{R}_{\geqslant 0}$ defined as

$$L(\mathbf{r}) = \frac{1}{n} \sum_{j=1}^{n} ((\mathbf{b}_j, \mathbf{r}) - y_j)^2.$$

Since

$$\frac{\partial L}{\partial r_k} = \frac{2}{n} \sum_{j=1}^{n} ((\mathbf{b}_j, \mathbf{r}) - y_j) b_k,$$

it follows that the gradient of L is

$$(\nabla L)(\mathbf{r}) = \frac{2}{n} \sum_{j=1}^{n} ((\mathbf{b}_j, \mathbf{r}) - y_j) \mathbf{b}_j.$$

The condition $(\nabla L)(\mathbf{r}) = 0$ that is necessary for the optimum amounts now to $(B'B)\mathbf{r} = B'\mathbf{y}$, that is to the system of normal equations of B and $\mathbf{y}$.